A Comprehensive Analysis of Breast Cancer Clinical Trials: Trends and Patterns

Simona Cana Ungureanu

Master's Degree in Data Science

Project supervisor
Susana Pérez Álvarez

Coordinating professor
Albert Solé Ribalta

Date of submission
January 2024

Universitat Oberta de Catalunya

SUMMARY OF THE FINAL PROJECT

Title of the project:	*A Comprehensive Study of Breast Cancer Clinical Trials: Trends and Patterns*
Author name:	*Simona Cana*
Project supervisor:	*Susana Pérez Álvarez*
Coordinating professor:	*Albert Solé*
Date of submission (MM/YYYY):	*01/2024*
Name of the degree:	Master's Degree in Data Science
Topic of the final project:	*Breast Cancer Clinical Trials*
Language:	*English*
Keywords:	*Clinical trial, breast cancer, data analysis*

Abstract

Breast cancer remains a global health challenge, accounting for 30% of all new female cancer cases in 2023 (1). Beyond the plain statistics, it profoundly affects patients and their families, emphasizing the importance of ongoing research for prevention, early detection, and improved treatment methods.

Clinical trials are crucial in cancer care, setting the standards for cancer treatment. *Clinicaltrials.gov*, an online database, provides access to more than 460.000 clinical trials from more than 220 countries.

This project aimed to explore the breast cancer clinical trials included in the *ClinicalTrials.gov* database and perform a comprehensive analysis of the data to understand their evolution and current state. The study included 13,524 breast cancer clinical trials, performing an Exploratory Data Analysis (EDA), followed by the application of clustering and Principal Component Analysis (PCA).

The results highlighted a shift in the focus of breast research cancer with the introduction of behavioral interventions and advanced screening techniques in addition to testing of drugs. The study revealed a gender imbalance in breast cancer research with only 0.3% of studies focusing on male patients, while 1% of the total cases are diagnosed in men. The geographical distribution of the studies showed that almost 50% are conducted from the United States, while Africa or Latin America have little presence in breast cancer research.

In conclusion, the results of the study emphasize the need of a more inclusive approach in breast cancer clinical trials that is aligned with the Sustainable Development Goals in terms of gender equality and race/ethnicity representation.

TABLE OF CONTENTS

TABLE OF FIGURES

1. Introduction

1.1 Context and motivation

Breast cancer is the most frequently diagnosed cancer in women and the second cause of cancer-related death in women, only after lung cancer, as illustrated in Figure 1.1. Despite significant progress in breast cancer research and treatment, there are still ongoing gaps and uncertainties in our understanding of the disease and its treatment. Early diagnosis has been a crucial factor in the increase of survival rates, seeing nowadays 5-year survival rates in the range of 90% and 10-year survival rates being about 80% (2). Apart from strategies for early diagnosis and treatment, these higher survival rates require a new approach in managing cancer to increase patients' quality of life.

Estimated New Cases

Males			Females		
Prostate	288,300	29%	Breast	297,790	31%
Lung & bronchus	117,550	12%	Lung & bronchus	120,790	13%
Colon & rectum	81,860	8%	Colon & rectum	71,160	8%
Urinary bladder	62,420	6%	Uterine corpus	66,200	7%
Melanoma of the skin	58,120	6%	Melanoma of the skin	39,490	4%
Kidney & renal pelvis	52,360	5%	Non-Hodgkin lymphoma	35,670	4%
Non-Hodgkin lymphoma	44,880	4%	Thyroid	31,180	3%
Oral cavity & pharynx	39,290	4%	Pancreas	30,920	3%
Leukemia	35,670	4%	Kidney & renal pelvis	29,440	3%
Pancreas	33,130	3%	Leukemia	23,940	3%
All Sites	**1,010,310**	**100%**	**All Sites**	**948,000**	**100%**

Estimated Deaths

Males			Females		
Lung & bronchus	67,160	21%	Lung & bronchus	59,910	21%
Prostate	34,700	11%	Breast	43,170	15%
Colon & rectum	28,470	9%	Colon & rectum	24,080	8%
Pancreas	26,620	8%	Pancreas	23,930	8%
Liver & intrahepatic bile duct	19,000	6%	Ovary	13,270	5%
Leukemia	13,900	4%	Uterine corpus	13,030	5%
Esophagus	12,920	4%	Liver & intrahepatic bile duct	10,380	4%
Urinary bladder	12,160	4%	Leukemia	9,810	3%
Non-Hodgkin lymphoma	11,780	4%	Non-Hodgkin lymphoma	8,400	3%
Brain & other nervous system	11,020	3%	Brain & other nervous system	7,970	3%
All Sites	**322,080**	**100%**	**All Sites**	**287,740**	**100%**

Figure 1.1. Ten leading cancer types for the estimated new cancer cases and deaths by sex, United States, 2023. Source (1)

Further progress depends on clinical trials that evaluate new diagnostic techniques, innovative treatments, and ways to enhance the quality of life for breast cancer patients. Clinical trials require significant resources and collaboration from enrolled patients; thus, it is crucial to focus the efforts efficiently and ensure that no significant research areas are left neglected.

Clinical trials can be **interventional**, where a treatment or drug is tested, **observational**, where researchers observe a group of people in different situations without trying to alter the course of natural events, and **expanded access**, also called 'compassionate use', employed for those patients that need to gain access to an investigational medical product in a life-threatening situation and usually have a very small number of participants, many times just one (3).

As described by National Institutes of Health (4), clinical trials usually go through the following phases:

- **Early Phase 1**, formerly known as Phase 0, refers to an exploratory trial conducted before the traditional Phase 1 to investigate the impact of a drug on the participants. This phase involves a very limited human exposure to the drug.
- **Phase 1** trials - Researchers test a drug or treatment on a small group (20-80) to study its safety and identify possible side effects.
- **Phase 2** trials - The new drug or treatment is administered to a larger group (100-300) to determine its effectiveness.
- **Phase 3** trials include a larger group of people (1,000 – 3,000) and are aimed at confirming effectiveness, monitoring side effects and comparing it to standard or similar treatments.
- **Phase 4** trials occur after the responsible institutions approve the treatment or drug and its aim is to track the treatments' safety in the general population.

The *clinicaltrials.gov* web site inaugurated in 2000 and was established by the Department of Health and Human Services as part of the Food and Drug Administration Modernization Act of 1997. While initially created to increase public awareness of clinical trials, it is nowadays a mandatory repository for information on most clinical studies, especially those conducted under US regulations. Since its launch, it has expanded to cover other laws and regulations and the number of clinical trials has grown from 1000 to more than 460.000 currently.

In this project, we will extract data pertaining to breast cancer clinical trials from clinicaltrials.gov web site and explore the dataset to understand the evolution of research on breast cancer between 2000 and 2023. We will start by applying descriptive statistical techniques to the metadata included in the clinical trials database regarding the composition, size, design, and types of trials being funded as well as patient demographics. Subsequently, we will analyze the evolution on clinical trials over time and apply machine learning techniques to identify trends and patterns.

The motivation of this work comes from a profound commitment to improving the lives of individuals affected directly or indirectly by breast cancer. The main aim is to advance cancer research concentrating efforts on the appropriate areas with the highest potential to influence survival and enhance the life quality for patients, ultimately ensuring that they receive the most effective and compassionate care possible.

1.1.1 Personal motivation

The personal motivation behind this project is deeply rooted in a desire to make a tangible impact on people's lives, particularly in the public health domain. I have so far dedicated my career to cybersecurity, and while fulfilling, it left me feeling disconnected from the real-world impact of my work. While I understood the importance of keeping digital spaces safe, such as hospitals or airlines, I needed a more human purpose that was more connected to the well-being of individuals.

It was back in 2021 when I encountered news about a specific mouthwash showing a positive impact on COVID-19 virus spread and I was awe-struck by how this discovery was made through research and data analysis. It was then when I realized the immense potential of applying my technical background and analytical skills to the research of health issues. This led me to pursue this master's degree, where I was able to develop my data analysis skills and combine the learnings with my passion for languages in different natural language processing projects.

My aspiration is to bridge the gap between technology and real-world health challenges, contributing this way to the well-being of individuals facing critical health issues. The current research represents my commitment to making a more terrestrial impact on the lives of people by addressing health concerns and identifying opportunities for future breast cancer research.

1.2 Goals

The **main goals** of this paper are:

- **Analyze the evolution of breast cancer clinical trials to identify trends and patterns.**
    - Quantify the annual number of trials over the years to identify trends in the volume of clinical trials.
    - Assess changes in clinical trials over time based on specific criteria (e.g., status, phase, intervention type, primary purpose, funder type, and availability of results)
- **Correlate the results with the publicly available breast cancer statistics to identify research gaps and opportunities.**
    - Compare clinical trials proportion with epidemiological data to confirm if the representation of breast cancer in clinical trials matches its proportion in the totality of cancer cases.
    - Map the geographical distribution of trials against the prevalence of breast cancer in each country.
    - Analyze the alignment of trials distribution with participants' demographic data in breast cancer statistics.

To achieve these main goals, it is necessary to define the below **secondary goals**:

- **Explore patient demographics across different breast cancer clinical trials.**
  - Calculate the distribution of participants by gender and age.
- **Examine geographic diversity of breast cancer clinical trials and its impact on research results.**
  - Map the locations of trials and compare the trials distribution with the actual population distribution.
- **Assess the impact of funding sources on the design of clinical trials.**
  - Categorize clinical trials based on funding sources.
  - Identify the most common sponsors and analyze the distribution of their studies based on study type and intervention type.
- **Apply machine learning techniques to identify trends and patterns in the dataset.**
  - Use clustering algorithms to identify patterns in trial design, patient demographics or trial documentation.
  - Apply PCA to analyze high-dimensional data, such as study type, study duration, intervention type, documentation, to facilitate data visualization and interpretation.

1.3 Sustainability, diversity, and ethical/social challenges

While the main aim of the project is to advance breast cancer treatment and care, it is strongly connected to sustainability, diversity, and ethical/social challenges.

From a sustainability perspective, the project aligns with SDG 9 by applying innovation in the data analysis and machine learning techniques to explore breast cancer clinical trials. This project promotes the use of advanced technologies for healthcare research, thus having a positive impact on sustainability.

One of the goals of the project is to understand how funding sources impact breast cancer clinical trials and this is directly aligned with the ethical behaviour and social responsibility. Once the study is concluded we will have a better understanding on the impact of funding on the clinical trials themselves and also the gains that sponsors obtain by participating in clinical trials.

Where we can identify an even more direct impact is on diversity, gender and human rights as the project aims to review the impact of the research based on the demographic characteristics of the enrolled patients as well as the geographic distribution of clinical trials. Although breast cancer affects predominantly women, it can also impact men and this project will allow us to promote gender equality in healthcare and research participation. We expect this project to have a positive impact on diversity, gender and human rights by outlining the areas where research efforts need to be focused to reduce inequalities in healthcare.

1.4 Approach and methodology

The methodology that will be employed for this data analysis project will be based on the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework. For the current project, the below phases are defined:

- **Business understanding**: The initial stage where the goals and methodology of the project are defined, and the project plan is produced.
- **Data understanding**: This stage, described in chapter 3.1, involves the data collection and initial assessment. It also includes the review of the state of the art in regard to breast cancer clinical trials to gain a deeper understanding of the data.
- **Data preparation**: Through data processing techniques the dataset will be cleaned and formatted adequately for the analysis. This phase, described in chapter 3.2, includes normalization, handling of missing values and transforming variables if necessary.
- **Modelling**: This phase, described in chapter 3.3, consists of the application of data analysis and machine learning techniques to the processed dataset. This will allow the identification of trends, patterns, and relationships within the dataset.
- **Evaluation**: In this phase, included in chapter 3.3, we will evaluate the results obtained and choose the most adequate models. We will also determine next steps, whether the project should move to deployment, or we need to iterate further.
- **Deployment**: In this phase, we will review the project and produce the final report.

Figure 1.2. CRISP-SM Diagram. Source Wikipedia

The data processing and analysis will be performed in Python by using Pandas, NumPy, Scikit-Learn libraries, among others, as well the NLTK (Natural Language Toolkit) for the processing and analysis of text data. Python libraries

such as Matplotlib and Seaborn will be used for data visualization. The project planning and tracking will be performed in Asana.

1.5 Schedule

The planning of the project has considered the CRISP-DM framework phases and has also integrated each continuous assessment test defined by the didactic plan of the master's thesis. Figure 1.2 shows the detailed schedule that includes the phases described in the previous section as well as all the deliveries planned by the teaching team.

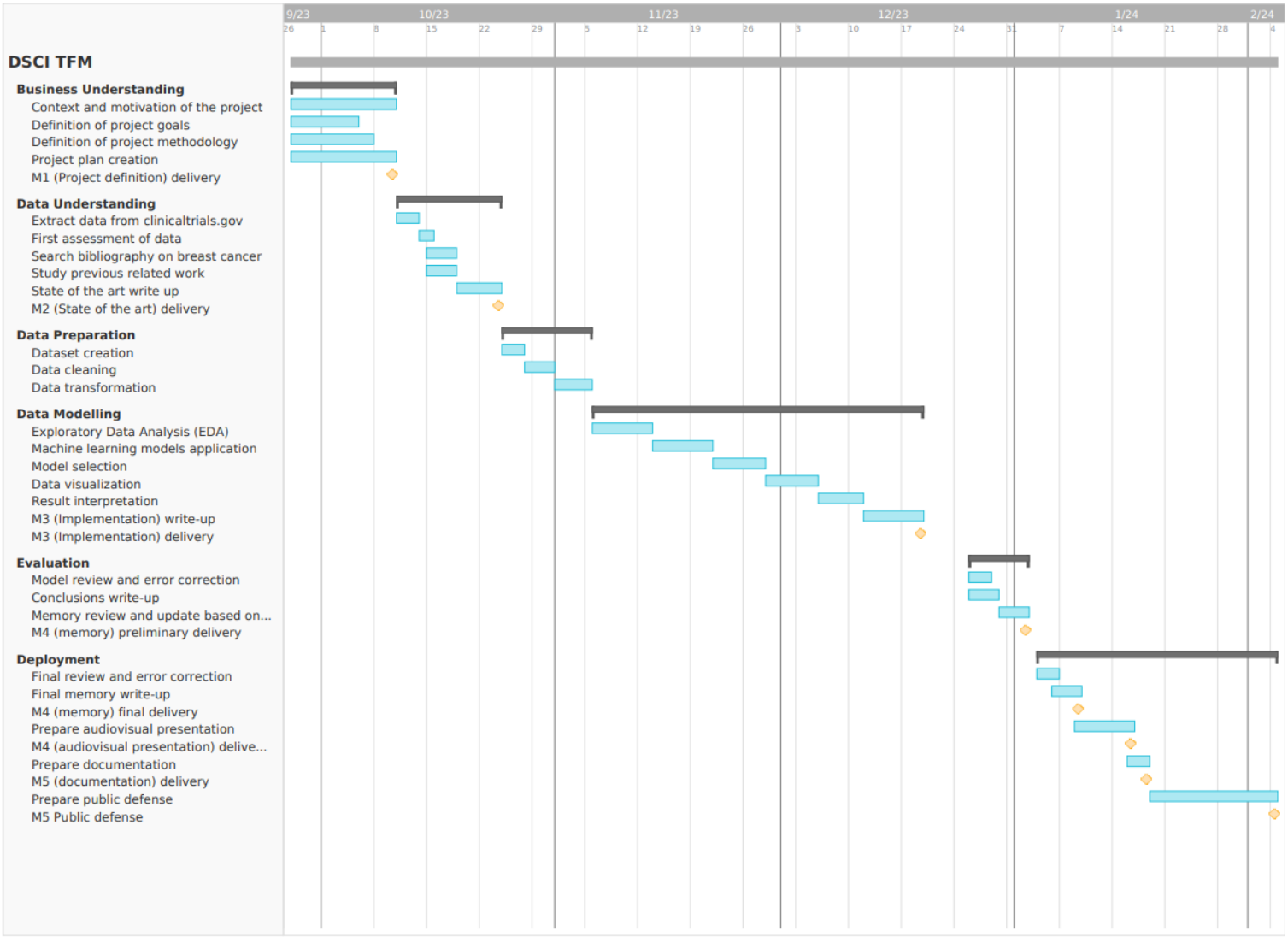


Figure 1.3. Project Gantt Chart

2. State of the art

In this chapter, we will conduct a review of the literature related to breast cancer clinical trials, including its history, the evolution, and challenges over time. We will also explore the utilization of machine learning techniques on clinical trials to contextualize the work developed in the current thesis.

2.1 Brief history of breast cancer clinical trials

According to the American Cancer Society, cancer is the leading cause of death worldwide with nearly one in every six deaths being caused by cancer in 2020 (1). The evolution of cancer is strongly connected with the diagnosis techniques and changes in the medical practice(5). Figure 2.1 illustrates long-term trends in overall cancer incidence in men and women, reflecting these changes such as the use of screening tests. For instance, the spike in incidence for men during the 1990s reflects a surge in prostate-specific antigen (PSA) testing among previously unscreened men (6). For women the rate was more stable until the 1980s and then the increase was slower than for men. Nowadays, the gender gap is narrowing with similar incidence rate ratios for men and women.

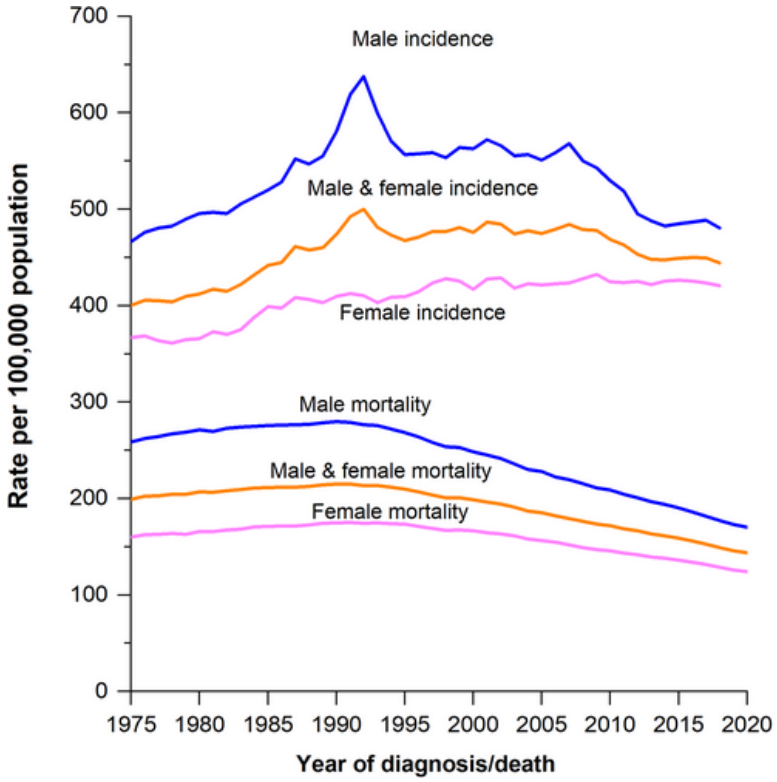


Figure 2.1. Trends in cancer incidence (1975–2019) and mortality (1975–2020) rates by sex, United States. Rates are age adjusted to the 2000 US standard population. Incidence rates are also adjusted for delays in reporting. Source (1)

The most common cancer types for men in the United States are prostate cancer with 29% of all incident cases, while for women the most prevalent is breast cancer with 31% of all female cancers(1). At a more global level, according to the latest statistics published by the World Health Organization (WHO) in 2020, breast cancer represents 47% of all female cancers worldwide, while the most

common cancer type in men is represented by lung cancer with a 31% of all men cancer, followed by prostate cancer accounting for a 30% of all cancer cases in men (7).

As mentioned previously, the spike illustrated in Figure 2.2 around 1990s in prostate cancer for men was due to the surge in screening, while the evolution in breast cancer detection for women was slower even if breast cancer screening programs were introduced at a similar time (5) This showed that cancers are a heterogeneous group of diseases and not all precancerous lesions lead to invasive cancers, meaning that generalized screening did not necessarily have a positive impact causing overdiagnosis and overtreatment.

To address this challenge, new risk-based models were introduced to identify individuals who had a higher risk of cancer than the general population as screening candidates. One of the first initiatives in this direction, was the creation of the Breast Cancer Risk Assessment Tool (8) that takes into account the personal history, family history, age of menarche, age of first live birth and number of previous biopsies, among others.

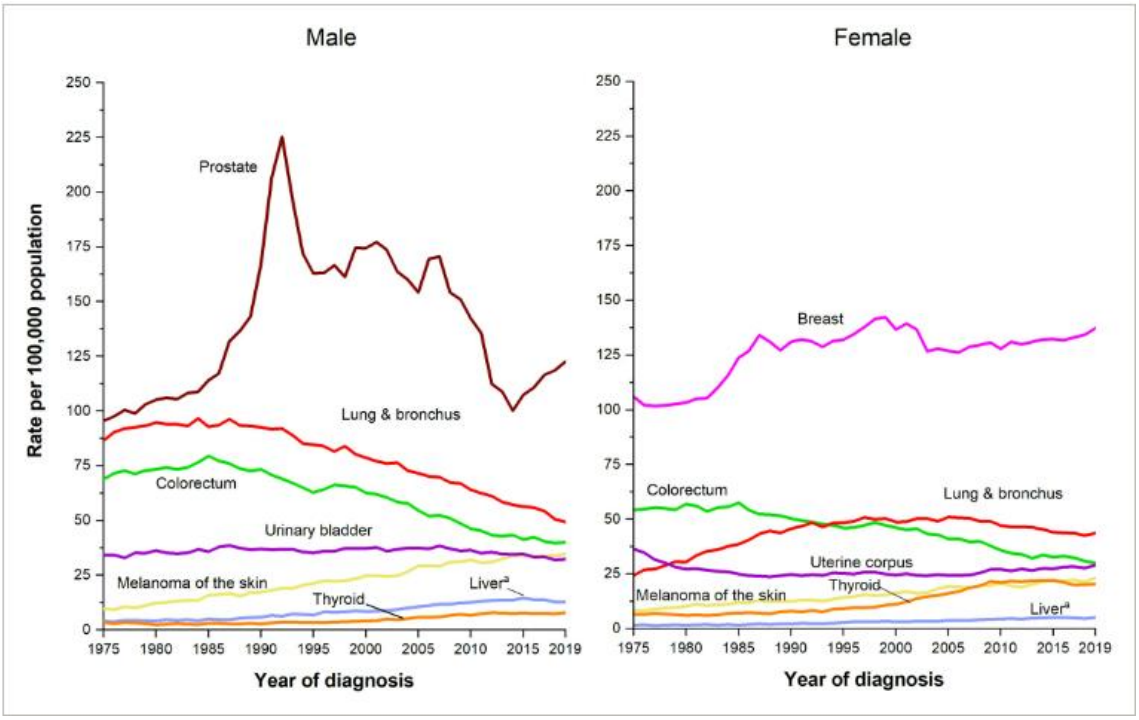


Figure 2.2. Trends in incidence rates for selected cancers by sex, United States, 1975–2019. Source (1)

Clinical trials have been fundamental in supporting with better screening for early diagnosis and cancer treatment. It is not an overstatement to say that contemporary medical oncology is built around the performance of clinical trials (9) and that they are the vector for the development of cancer treatments. The system of oncological clinical trials is a fairly recent innovation introduced after World War II. The first randomized cancer clinical trial was organized by the NCI (US National Cancer Institute) in 1954 in the context of cancer-drug screening

program for acute lymphocytic leukaemia (9). Since then, numerous research groups have been established in the United States to collaborate in cancer investigations under the umbrella of the NCI. Although initially these cancer clinical trials were focused on drug testing, in the mid-1960s they evolved to include testing of hypothesis concerning therapy and sought means for the prevention of cancer.

The earliest references to breast cancer date back to prehistory and the ancient world, with the first mention found in *The Edwin Smith Surgical Papyrus*, which is traced back to the pyramid age of Egypt (3000-2500 BC)(10). This paper refers to suturing wounds and cauterization with fire drills as a treatment for tumours in the breast, all in men and mostly due to wounds. As science, and mainly medicine, continued to develop the main treatment strategy for breast cancer was the mastectomy, first performed in the 17th century by the French surgeon Jean Louis Petit (10). Even in the early decades of the 20th century, most of the research was focused on “extended” mastectomies.

In terms of screening tools, the German surgeon Albert Salomon preformed studies with radiographs of breasts resected for carcinoma as early as 1913, but his work was interrupted by World War II. It was not until the 1960s, when Robert Egan developed the soft tissue techniques that allowed mammography to be used as a screening technique. Mammography is undoubtedly the most important advancement to date in the detection of breast cancer, and it allowed for many breast cancers to be detected when clinically occult.

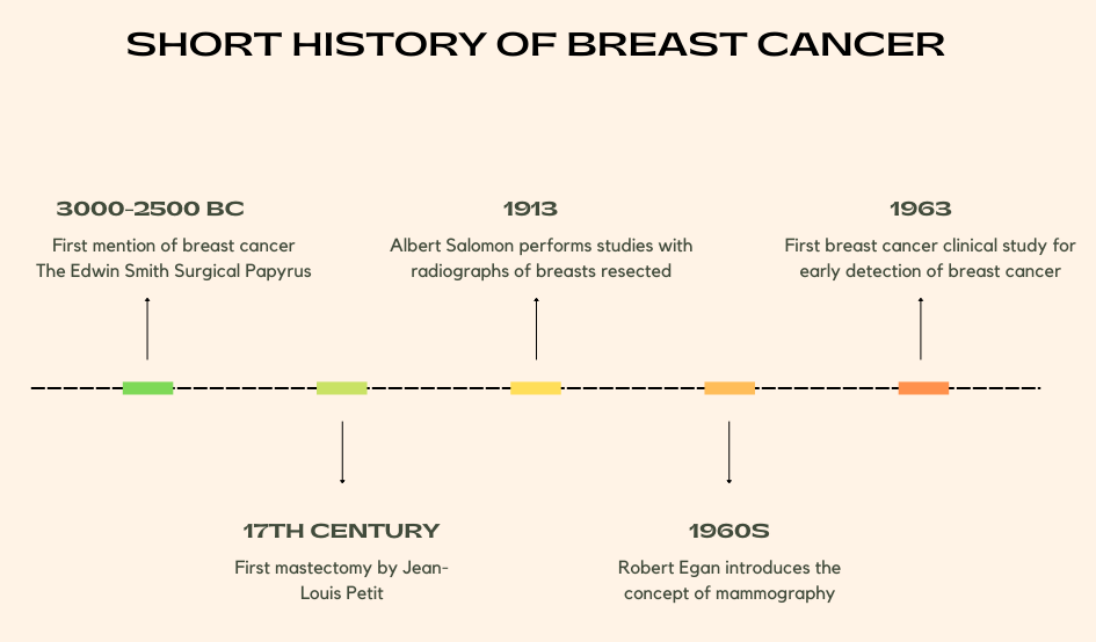


Figure 2.3. Short history of breast cancer. Source: Own work

One of the first randomized clinical trials focused on breast cancer was conducted in 1963 in New York by Sam Shapiro and Philip Strax who demonstrated that 30% of cancers could be detected by mammography alone, and deaths from cancers across screened women were reduced 30% compared to unscreened.

This led to a study performed in 1973 across 283,222 asymptomatic women who were screened showing that regular mammograms could detect up between 85% and 90% of asymptomatic breast cancers leading to a reduction of breast cancer (10).

This motivated the NCI and numerous other groups to support the introduction of the periodic mammography in asymptomatic women 40 years and older as means for detection of breast cancer.

Throughout the 20th and 21st century numerous clinical trials were performed to focus on better screening techniques and strategies, effective treatment, and improvement of life quality for cancer patients and their families. While until the 1940s the mastectomy and radiation therapy were the most common treatment approaches, thanks to multiple clinical trials it became evident that chemotherapy reduced breast cancer deaths in young women without the need for patients to lose the whole of their breasts (11).

In the last years, the focus has been on the development of sub-specialism within oncology, leading to patients benefitting from the combined expertise of a range of health professionals working together in a multi-disciplinary team. The numerous clinical trials performed on breast cancer have allowed for significant advances in the treatment of breast cancer and the ability to screen for the disease leading to it being one of the most curable types of cancer nowadays (12).

2.2 Key research themes and trends in breast cancer clinical trials

Currently, the breast cancer research is focused on precision treatment strategies based on molecular sub-typing of breast cancer. Being able to build targeted therapies for HER2, hormone receptors, and other molecular markers is fundamental for the advancement of breast cancer treatment (13). Figure 2.4 shows a brief summary and timeline of the advancements of breast cancer treatment in the last century. We can observe that immunotherapy continues to be an important focus in various cancer subtypes, both pre- and post-surgery.

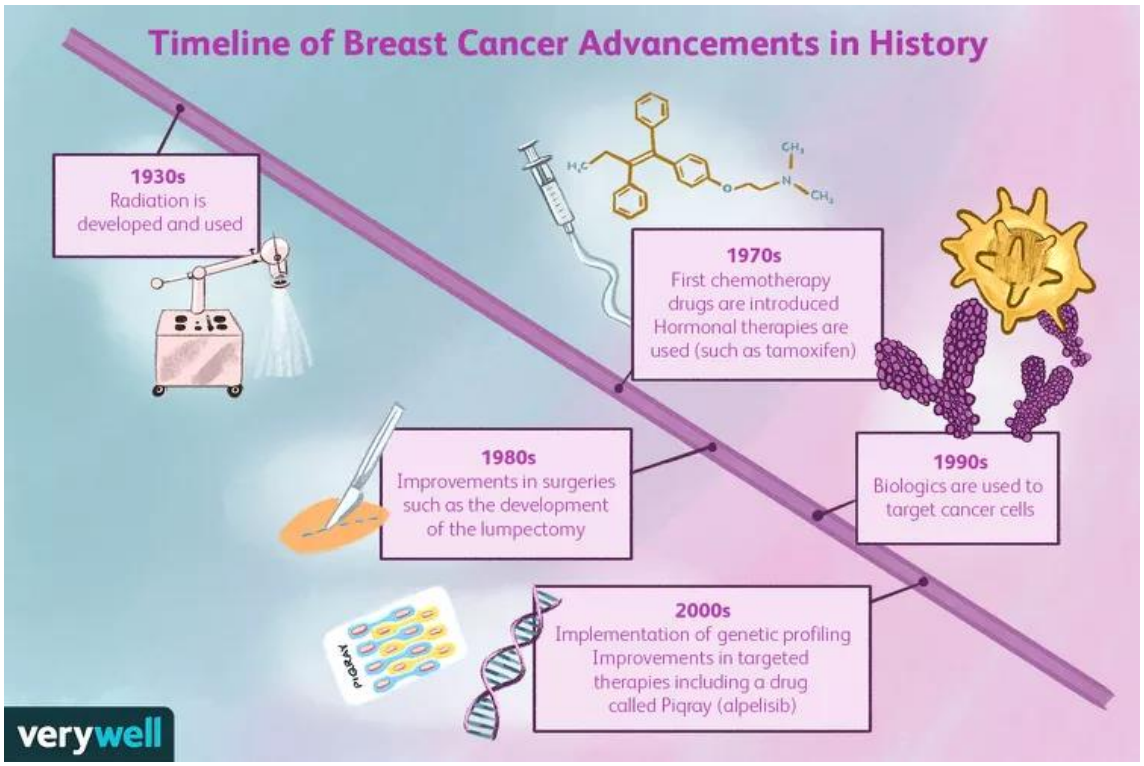


Figure 2.4. Timeline of breast cancer advancements in history. Source (14)

Metastatic breast cancer is still considered incurable, and many efforts are focused on understanding and treating metastatic cancer with novel treatments to extend survival and improve quality of life for patients.

Research in the recent times place a great emphasis on patient-reported outcomes, quality of life, and shared decision-making to enhance the patient's experience and well-being.

According to Hong et al. (13) there are two major questions that remain unanswered in the field of breast cancer research: whether breast surgery can be omitted in patients achieving a pathological complete response (pCR) after neoadjuvant therapy, and whether certain patients can avoid axillary surgery for both staging and treatment purposes.

2.3 Challenges faced in breast cancer clinical trials

Although clinical trials aim to include all the possible patients and explore the most important aspects of the illness they are focused on, sometimes it is challenging for patients to access clinical trials or be benefited by them. In this section, we will explore some of the most common challenges faced by patients with regards to clinical trials nowadays.

There are multiple barriers for patients to enter a clinical trial, as illustrated in Figure 2.5. According to a study conducted by Unger et al. (10) in 2019 analyzing 13 studies with a total of 8883 patients, identified that more than half of the patients (55,6%) were not able to join any clinical trial due to the unavailability of a trial for the patient's cancer type and stage.

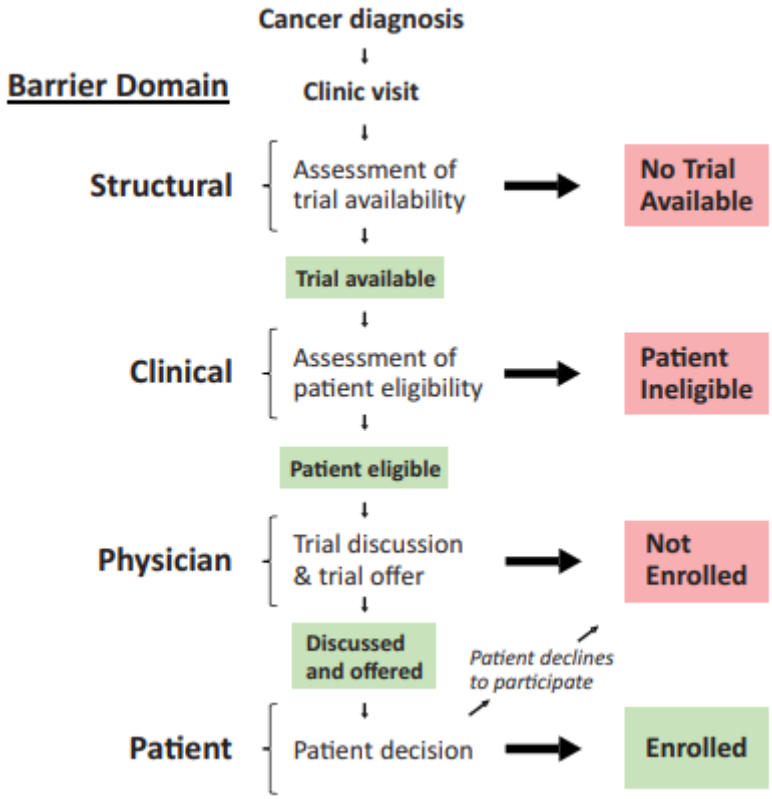


Figure 2.5. A framework for describing the clinical trial decision-making pathway. Source (15)

The same study (10) outlines that many clinical trials exclude patients due to the desire to establish a study cohort with similar patient profiles to assess more accurately the response of the patients to the different treatments or interventions. Nevertheless, many times this is unnecessary and can lead to the eligibility criteria being too narrow and affecting thus the results of the study due to the too narrow population.

In addition, there are physician and patient factors that can affect the participation in clinical trials. Physicians are the entry point of patients to clinical trials and sometimes they do not inform patients about existing clinical trials due to time or reimbursement constraints, treatment preference or other reasons. This takes away from the patients the opportunity to participate in a study. According to the multiples studies (10, 12), more than 50% patients accept to participate in a study if their physician recommends it and they are eligible. There is a small proportion of patients who refuse to participate, and it is usually due to their treatment preference, costs, or logistical barriers.

Another challenge that affects clinical trials is early stopping of the trial as described by Cuzick et al. (16) in his study of breast cancer clinical trials. The authors found that in some cases the trials are ended too early in favor of standard treatments which can delay full acceptance of new treatments as well as the credibility of the clinical trials themselves. In some cases, this early stopping is justified by the higher benefits obtained through standard treatment,

but in other cases when early indicators are positive, the author defends that there should be clearer stopping rules to avoid missing the whole benefit of the study.

As mentioned in section 1.3, diversity is an important aspect we would like to explore in this thesis. Currently, there are many challenges related to diversity in the participation in clinical trials. Recent findings by Bea et al. (17) show that African Americans have been under-represented in clinical trials despite carrying a disproportionately high breast cancer mortality burden. Some of the reasons for this absence from the clinical trials are the lack of financial support from the recruiting institutions, the language barrier, and the lack of patient education. A lack in reporting patient diversity when informing about a clinical study makes it difficult to quantify the impact of this misrepresentation.

Gender diversity is a key aspect as well as patients with a prior cancer are usually excluded from clinical trials, and men, for whom breast cancer is not a very common one, can be excluded from the study due to having another cancer previously. According to a study conducted by Rathod et al. (18) among 2317 men that were included in the study, almost a quarter (24,3%) had a different type of cancer previously and were excluded from clinical trials. Given the low prevalence of breast cancer in men, reconsidering this exclusion criteria might help with investigation advanced in breast cancer for men.

Another recent challenge faced by clinical trials in general was the impact of Covid-19 in many aspects of the healthcare system. Many of the research resources were directed towards Covid-19 and many randomized clinical trials were launched to support with different aspects of the pandemic (19). The social distancing regulations led to many ongoing studies having to reduce the number of follow-up imaging, general health measurements and history and physical collection for patients. In addition, patients were uneasy with leaving their homes and visiting medical centers. The NCI predicts that 10,000 excess cancer deaths will occur over the next decade because of missed screenings and delays in diagnosis during the Covid-19 pandemic. On a positive note, the challenge surfaced by Covid-19 helped clinical sites and research groups to consider aspects of oncology that can be modernized while maintaining research integrity, such as data collection process, electronic consent for enrolment, telemedicine visits, and mail order pharmacy.

2.4 Machine learning techniques applied to clinical trials research

Big data has significantly changed the way we generate, manage, analyze and leverage data. Clinical medicine generates and hold a large volume of data from patient records, wearable devices, and insurance companies (20). Taking into account this volume of data, there are many questions that can be addressed, especially with predictive analysis.

The data-driven techniques find applications throughout the entire spectrum of the disease course, spanning disease prevention, diagnosis, treatment, and prognosis. When it comes to prevention, predictive analysis can help identify risk

factors for certain types of cancer (for instance, smoking for lung cancer) allowing for informed public awareness campaigns to educate and thus mitigate the incidence of that particular cancer.

Once a patient has been diagnosed with a certain type of cancer, predictive analysis comes into play by conducting risk assessments based on genetic and clinic characteristics. This process helps healthcare professionals identify the treatment approach likely to produce the most favorable outcomes. Moreover, it empowers personalized medicine by tailoring interventions to individual patient profiles, ensuring that treatment plans are optimized for each patient.

Finally, data analysis extends its reach to the predictive of long-term outcomes and life expectancy. This information serves as a crucial compass for healthcare professionals, guiding them into offering patients comprehensive support measures to enhance their quality of life. Additionally, these predictions also offer valuable insights for patients and their families, assisting them in making informed decisions regarding treatment and care.

Looking at the specific field of clinical trials, there are many applications of Machine Learning (ML) to this field as illustrated in Figure 2.6.

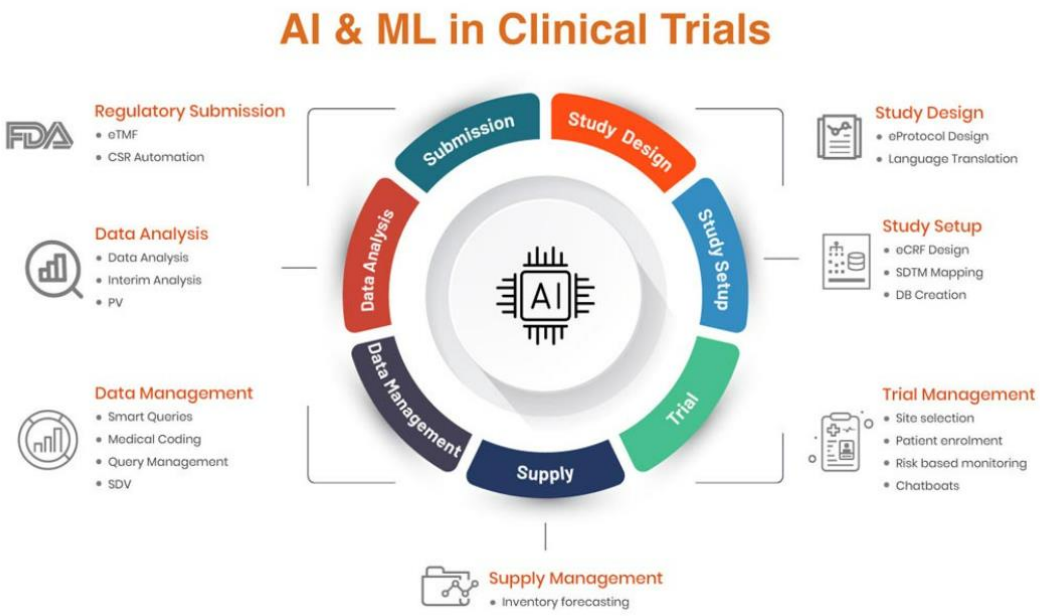


Figure 2.6. Use of Artificial Intelligence (AI) and Machine Learning (ML) in Clinical Trials. Source (21)

Many of these applications are meant to enhance progress in clinical trials and research, in their different phases. A study performed by Zippel et al. (22) based on the data registered on ClinicalTrials.gov observed an evolution of studies that employed Machine learning techniques in the last years.

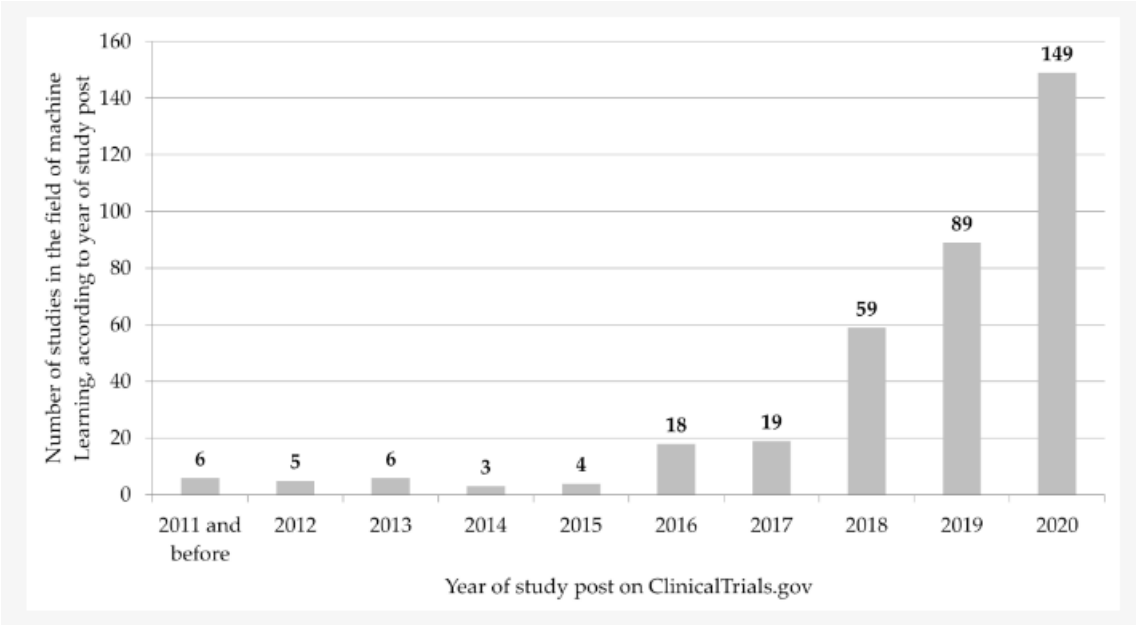


Figure 2.7. Number of clinical studies involving Machine Learning by year of publication on ClinicalTrials.gov. Source (22)

These studies were mostly related to the field of imaging (radiology, nuclear medicine, oncology), followed by cardiology, psychiatry, intensive care and neurology.

Examining the specific application of machine learning in analyzing data from clinical trials listed on ClinicalTrials.gov, it's essential to mention several noteworthy studies. These studies server as valuable references and will guide the development of the current research.

In a study conducted by de Glas et al. in 2014 (23), a comprehensive review of 463 clinical trials related to breast cancer treatment, as published on ClinicalTrials.gov, examined the approaches taken regarding older patients. The findings of this study were striking: only 2% of all ongoing clinical trials at the time were specifically designed for older patients. Furthermore, the study pointed out that the assessment of quality of life and preservation of functional capacity were not given high priority as endpoints in determining if patients could tolerate specific treatments. Ultimately, the study's conclusion emphasized that the ongoing clinical trials during that period were unlikely to yield substantial advancements in the treatment of older breast cancer patients.

Another study published in 2017 by Shepshelovich et al. (24) looked at the relationship between the clinical trials publication on ClinicalTrials.gov and their inclusion in journals. The authors reviewed 583 phase I adult cancer clinical trials out of which only 163 had entries in matching publications.

When reviewing these clinical trials, the authors found that many of them did not have complete reporting of all the clinical data on ClinicalTrials.gov. For instance, for 62% of reviewed trials, the primary outcome reported on ClinicalTrials.gov matched the one described in the journal publication, while this percentage was

much lower for secondary outcome with only 27% matching the journal publications.

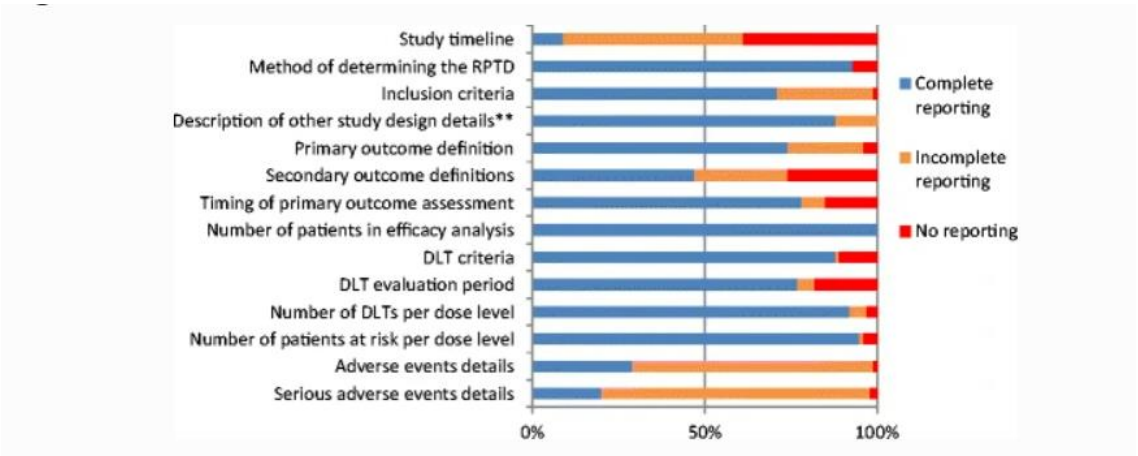


Figure 2.8. Completeness of reporting of assessed items in primary publications compared with ClinicalTrials.gov. Source (24)

The study found inconsistencies and shortcomings in the way results from early-phase clinical trials were reported in primary publications when compared to the corresponding listing on ClinicalTrials.gov. The authors recommended improving reporting and raising awareness of ClinicalTrials.gov as a valuable data repository for completed early-phase trials.

In 2013 Hirsch et al. (25) performed a review of oncological clinical trials published on ClinicalTrials.gov to perform a systematic analysis of the records. The authors extracted 40970 studies from ClinicalTrials.gov, out of which 21,8% represented oncological studies, followed by mental health (9%), infectious disease (8,3%), and cardiology (5,7%). Out of all oncology trials, 65,1% of the trials at the time included a North American study site, with only 34,9% of the studies conducted purely in other regions. In terms of sponsorship, 47,1% trials were funded primarily by the industry, and just 6,8% funded by the government.

This study shows that the focus of these clinical trials was mainly on finding new treatments and testing new drugs, as opposed to the better understanding and improvement of the existing treatments. According to the authors, by following this approach many of the research questions remained unanswered.

A recent study by Gresham et al. from 2022 (26) looked at all the clinical trials published on ClinicalTrials.gov between 2000-2022 and reviewed their design characteristics, eligibility criteria, interventions, conditions, and funders by year. This study showed the impact of different regulations on the completeness of data on ClinicalTrials.gov, where a direct correlation was found between new regulation being released and an increase in the completeness of data reporting on ClinicalTrials.gov. It also identified that over three quarters of the primary sponsors for registered trials were categorized as “other” making it difficult to identify relationships between other characteristics of the trials and the funding source. The authors suggest that future work should also include the results

database, including trial composition and demographics, primary outcome results, and safety data.

These studies collectively highlight important findings and areas for future work in clinical trials research. In 2013, Hirsch et al. (25) demonstrated a predominant focus on new treatment rather than improving existing ones, leaving many critical questions unanswered. In 2014, De Glas et al. (23) revealed that a mere 2% of clinical trials for breast cancer were designed for older patients, emphasizing the need for more inclusive trials and a focus on patient's qualify of life. In 2017, Shepshelovich et al. (24) uncovered inconsistencies in reporting between clinical trial publications and ClinicalTrials.gov, underscoring the importance of improving data reporting and raising awareness of ClinicalTrials.gov as a valuable repository. Lastly, in 2022, Gresham et al. (26) emphasized the impact of regulations on data completeness and the need for comprehensive data reporting, including trial composition and demographics, primary outcomes, and safety data, for future research.

3. Methodology and outcomes

In this chapter, we will detail the methodologies and resources employed in our comprehensive analysis of breast cancer clinical trials, while including the results obtained on each step. Our process involved (1) collecting data from ClinicalTrials.gov, followed by (2) thorough cleaning and transformation, and then (3) modelling with the conduction of an extensive Exploratory Data Analysis (EDA), and the application of clustering and Principal Component Analysis (PCA) to uncover patterns and trends in the evolution of these trials. For the processing and analysis of the data we have used Python, a programming language with open-source libraries such as Pandas, Numpy and Scikit-learn, among others.

3.1 Data understanding

3.1.1 Data collection

The data collection process involved extracting data from ClinicalTrials.gov and given that for our study we specifically targeted breast cancer clinical trials, we extracted only relevant studies by filtering for trials that focused on breast cancer, ensuring that we were capturing the most important data for our analysis. In this process we extracted a CSV file containing the variables related to these studies.

The dataset, as of December 12nd, comprises a total of 13,524 records related to breast cancer, each of them representing a unique clinical trial. This dataset provides 30 different variables that describe each clinical trial that we will use to build a complete landscape of the breast cancer clinical trials.

3.1.2 Description of data

The dataset includes variables that describe different aspects of clinical trials, from the participants' demographics to information about the study design, its sponsors, and other relevant information. Below we include a list of all variables, with their description and values that they take.

	Variable	Description
1	NCT Number	National Clinical Trial (NCT) Identification Number. The format is "NCT" followed by an 8-digit number.
2	Study title	A short title of the clinical study.
3	Study URL	Direct link to the clinicaltrials.gov study's page.
4	Acronym	An acronym or abbreviation used publicly to identify the clinical study, if any
5	Study Status	Represents the current status of the clinical research. Possible values: - UNKNOWN - COMPLETED - TERMINATED - RECRUITING - NOT_YET_RECRUITING - ACTIVE_NOT_RECRUITING - WITHDRAWN - ENROLLING_BY_INVITATION - SUSPENDED - APPROVED_FOR_MARKETING - AVAILABLE - NO_LONGER_AVAILABLE - TEMPORARILY_NOT_AVAILABLE
6	Brief Summary	A brief summary of the clinical trial.
7	Study Results	Indicates if the study has published results or not. Possible values: YES/NO
8	Conditions	Conditions targeted by the clinical trials. May include multiple conditions separated by a pipe symbol.
9	Interventions	Intervention/s associated with the clinical trial. It includes multiple interventions separated by a pipe symbol. Possible values: - BEHAVIORAL - BIOLOGICAL - COMBINATION_PRODUCT - DEVICE - DIAGNOSTIC_TEST - DIETARY_SUPPLEMENT - DRUG - GENETIC - PROCEDURE - RADIATION - OTHER
10	Primary Outcome Measures	The planned outcome measure for the clinical trials.
11	Secondary Outcome Measures	Secondary outcome for the clinical trial.
12	Other Outcome Measures	Any other outcome measure that is specified for the clinical trial.
13	Sponsor	Name of the organization that sponsors the clinical trial.
14	Collaborators	Name of other organizations that collaborate in the clinical trial. Can include multiple values separated by a pipe symbol.

15	Sex	Gender of the participants eligible to participate in the clinical study. Possible values: FEMALE, MALE, ALL.
16	Age	Age of the participants eligible to participate in the clinical study. Possible values: CHILD, ADULT, OLDER_ADULT. May include multiple comma-separated values.
17	Phases	Phase of the study as described in Chapter 1. Possible values: - EARLY_PHASE1 - PHASE1 - PHASE2 - PHASE3 - PHASE4 May include multiple pipe separated values.
18	Enrollment	Number of participants enrolled in the clinical trial.
19	Funder Type	The type of funder that supports the clinical trial. Possible values: - INDUSTRY - NETWORK - NIH - OTHER_GOV - FED - INDIV - UNKNOWN - OTHER
20	Study Type	The nature of the investigation or the investigational use for which the clinical study information is being submitted. Possible values: - INTERVENTIONAL - OBSERVATIONAL - EXPANDED_ACCESS
21	Study Design	Information about clinical trial allocation, interventional model, masking, and primary purpose.
22	Other IDs	Other identification numbers assigned by other organizations.
23	Start Date	The actual date when participants enrolled in the clinical study or estimate date when they should be able to enroll.
24	Primary Completion Date	The date when the final participant was examined or received an intervention for the purposes of final collection of data for the primary outcome.
25	Completion Date	The date when the final participant was examined or received an intervention for purposes of final collection of data for all outcome measures and adverse events (last participant's visit).
26	First Posted	The date when the record was first available on ClinicalTrials.gov.
27	Results First Posted	The date on which the sponsor or investigator first submits a study record with results.
28	Last Update Posted	The most recent date on which the study sponsor or investigator submitted changes to a study record on ClinicalTrials.gov.
29	Locations	Locations of the facilities participating in the study. It can include one or more addresses separated by a pipe symbol.
30	Study Documents	Name of the documents provided by the sponsor or investigator. It can include one or multiple documents separated by pipe symbol.

Table 3.1. Variable description

3.2 Data preparation

Through data processing techniques the dataset will be cleaned and formatted adequately for the analysis. This phase will include handling of missing values and transforming the necessary value to prepare the dataset for the Exploratory Data Analysis. This represents the initial processing of data, as in the course of our analysis the data will be modified again depending on the requirements of the Machine Learning algorithms that will be applied.

3.2.1 Data cleaning

As recommended by Chapman et al. (27) the data cleaning process should address the quality issues that are required for the selected analysis techniques. In this case, we will perform an initial selection of data followed by addressing the missing values from our dataset.

Upon the initial exploration of the data, we recognized that the variables 'Study URL' and 'Other IDs' do not add analytical value for our study, so we decided to drop these columns.

We confirmed that no duplicate entries were present in our dataset based on the 'NCT Number', so no action was necessary to address duplicate entries.

We then listed the missing values for each variable and the percentage they represented in relation to the totality of entries. We obtained the result displayed in the table below:

Variable	Missing values, n	Missing, %
NCT Number	0	0.00
Study Title	0	0.00
Acronym	9901	73.21
Study Status	0	0.00
Brief Summary	0	0.00
Study Results	0	0.00
Conditions	1	0.01
Interventions	1139	8.42
Primary Outcome Measures	586	4.33
Secondary Outcome Measures	3341	24.70
Other Outcome Measures	12373	91.49
Sponsor	0	0.00
Collaborators	8177	60.46
Sex	12	0.09

Age	0	0.00
Phases	6542	48.37
Enrollment	235	1.74
Funder Type	0	0.00
Study Type	0	0.00
Study Design	30	0.22
Start Date	83	0.61
Primary Completion Date	575	4.25
Completion Date	624	4.61
First Posted	0	0.00
Results First Posted	11656	86.19
Last Update Posted	0	0.00
Locations	1111	8.22
Study Documents	12618	93.30

Table 3.2. Missing values list with total count and percentage for each variable

To address the missingness, we performed the following actions:

1. **Drop variables**: We took this action for 'Acronym' variable as it showed over 73% missing values, and it was irrelevant for our study.
2. **Values replaced or imputed:**
   - For the 'Conditions' variable we replaced the one missing value with 'Breast Cancer' as it was representative for the study.
   - For 'Study Documents' as it had more than 93% missing values, we decided to fill them with 'NO' and process the rest of the values later during the Exploratory Data Analysis.
   - For 'Start Date' we replaced the 0,61% missing values with 'First Posted' for approximation. We understand that studies might have different Start and First Posted dates, but for such a small portion of the data, we decided to make the compromise and choose First Posted for replacement.
   - For 'Sex' variable we decided to impute the 0.09% of missing values with the most frequent value ('FEMALE').
   - For 'Study Design' missing values, we explored the 30 studies that has missing values and concluded they were all related to 'EXPANDED_ACCESS' type of studies. Given the circumstances surrounding these studies, as described in Chapter 1, we decided to replace the missing values with the 'EXPANDED_ACCESS' fixed value.
   - For 'Enrollment' variable we imputed the missing values with the median value for the variable.
3. **No action:**
   - We decided to take no action and leave missing values as NaN for the rest of variables with missing values for the Exploratory Data Analysis (EDA). We will address those missing values before the Machine Learning techniques application.

For the ‘Start Date’ variable we identified some outliers, for which we decided to take the following actions:

- Studies that started before 2000 – we decided to keep them as they had enough relevant information for our study.
- Studies set to start in the future – we decided to keep them as they include enough information to be included in the study.
- One study scheduled to start in 2100 – after analyzing the information regarding this clinical trial, we decided to update the value of ‘Start Date’ to ‘First Posted Date’ as the study is related to a review of an already closed study from 2020 (28).

This summarizes the data cleaning strategies followed to prepare the dataset for the next steps in our study.

3.2.2 Data transformation

Data transformation is a crucial step in preparing the data for modelling. It involved modifying data to make it more suitable for analysis. This process often includes the creation of new variables, modification of existing ones, and conversion of data types.

According to our analysis needs, we performed the following data transformation tasks:

1. **Text Field Transformation**: For text fields, new variables were derived to capture the required information in the appropriate format. This was relevant for the following variables: Conditions, Interventions, Collaborators, Age, Phases, Study Design, and Locations. For each of these variables, we split the string into a list or a list of tuples that better represent the data.
2. **Feature engineering**: Feature engineering is the process of creating new variables or features from existing data to provide deeper insights during analysis or to improve the performance of machine learning modules. For our study, we created the following variables:
    - **‘Countries’ and ‘Cities’** were extracted from ‘Locations’, with ‘cities’ encompassing regions or states, depending on the country.
    - **‘Duration’** was calculated based on ‘Start Date’ and ‘Completion Date’ to represent the length of each clinical trial in days.
    - **‘Study Has Documents’** variable derived from ‘Study Documents’. It categorizes the variable into ‘YES’ if any documentation is present and ‘NO’ otherwise. Before this transformation we extracted the types of documents and the frequency of their presence in the ‘Study Documents’ column:

Document Name	Count
Study Protocol and Statistical Analysis Plan	522
Study Protocol	273
Informed Consent Form	78
Statistical Analysis Plan	19

Table 3.3. Types of ‘Study Documents’ and frequency.

3. **Date and Time Extraction**: In the present dataset date-related variables were stored in string format and their conversion to date format was crucial for temporal analysis. We have converted to standard date format the following variables: Start Date, Primary Completion Date, Completion Date, First Posted, Results First Posted, and Last Update Posted

These initial tasks on data transformation prepared the dataset for the Exploratory Data Analysis (EDA), ensuring that each variable provided maximum value for our analysis. Upon completion of these transformation tasks, the dataset had the following structure:

#	Variable	Non-Null Count	Dtype
0	trial_id	13524 non-null	object
1	title	13524 non-null	object
2	status	13524 non-null	object
3	brief_summary	13524 non-null	object
4	study_has_results	13524 non-null	object
5	primary_outcome_measures	12938 non-null	object
6	secondary_outcome_measures	10183 non-null	object
7	other_outcome_measures	1151 non-null	object
8	sponsor	13524 non-null	object
9	sex	13524 non-null	object
10	phases	6982 non-null	object
11	enrollment	13524 non-null	float64
12	funder_type	13524 non-null	object
13	study_type	13524 non-null	object
14	start_date	13524 non-null	datetime64[ns]
15	primary_completion_date	12949 non-null	datetime64[ns]
16	completion_date	12900 non-null	datetime64[ns]
17	first_posted	13524 non-null	datetime64[ns]
18	results_first_posted	1868 non-null	datetime64[ns]
19	last_update_posted	13524 non-null	datetime64[ns]
20	study_has_documents	13524 non-null	object
21	conditions	13524 non-null	object
22	interventions	12385 non-null	object
23	collaborators	5347 non-null	object
24	age	13524 non-null	object
25	phases_split	6982 non-null	object
26	study_design	13524 non-null	object
27	countries	13524 non-null	object
28	cities	13524 non-null	object
29	duration	12900 non-null	float64

Table 3.4. Structure of the prepared dataset (results of data.info())

3.3 Data modelling

This phase consists of the application of methods for data analysis and machine learning techniques to the processed dataset. The aim of this process is the identification of trends, patterns, and relationships within the dataset.

3.3.1 Exploratory Data Analysis (EDA)

As a first step, we will perform an Exploratory Data Analysis (EDA) to better understand the data, uncover patterns, and draw insights that could be helpful for the next phases of our study.

We started by exploring the number of breast cancer clinical trials over time and plotted this evolution on the bar graph in Figure 3.1.

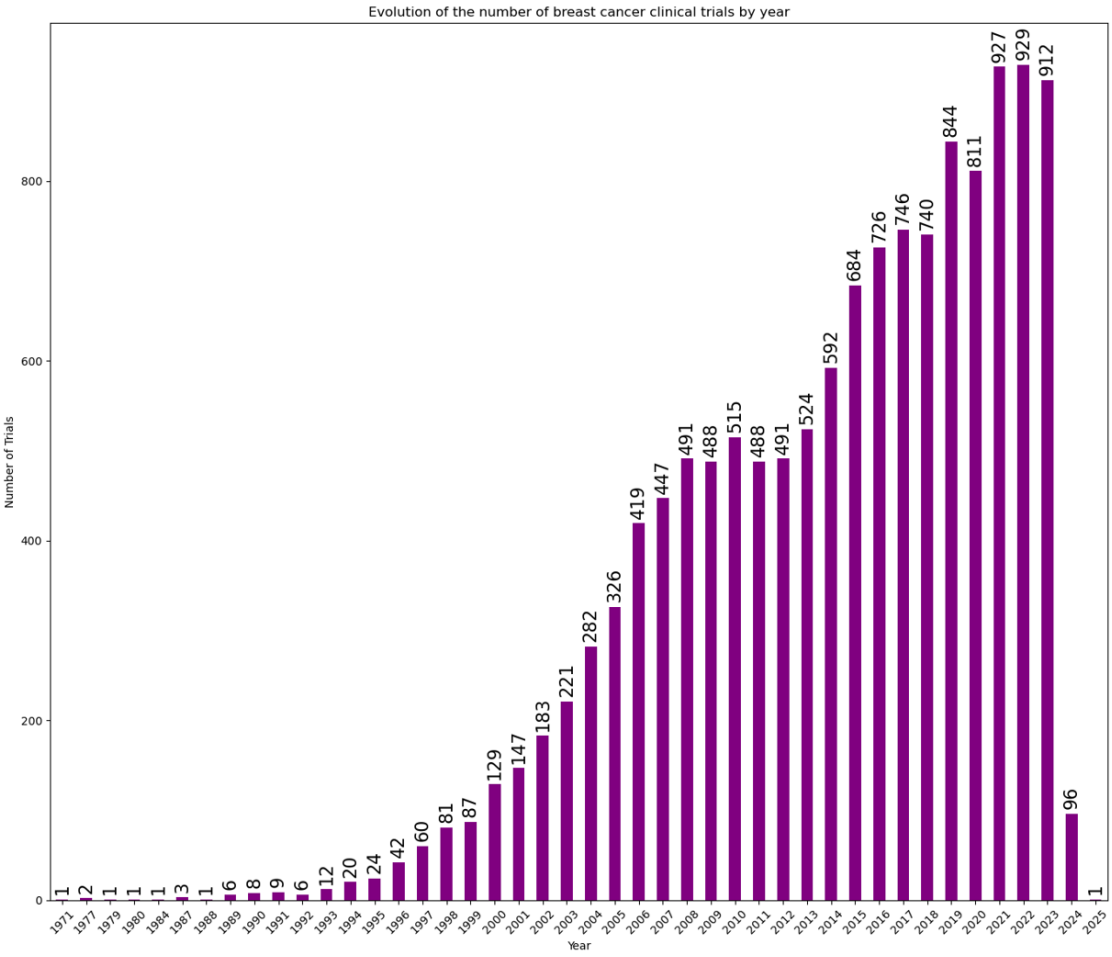


Figure 3.1. Evolution of the number of breast cancer clinical trials by year

As stated in section 3.2.1, ClinicalTrials.gov includes clinical trials that were started long before its launch in 2000, and studies that are scheduled to start in the next couple of years. After the launch of ClinicalTrials.gov in 2000 we can observe a steady increase in the number of reported clinical trials, with the exception of two periods that show a slight decrease in the number of studies. The first period was between 2011 and 2013 which is related to the global financial crisis that impacted the ability to fund research studies (29) and the

second one was in 2020, when the COVID-19 pandemic required extraordinary funding and resources to investigate the disease and produce vaccines and treatment, meaning that resources allocated to cancer research were shifted to COVID-19 research(19).

We noticed that approximately 20% of the studies in UNKOWN status involve a location from China, representing a 24% of the total clinical trials located in China. This indicates that the transparency of studies where China is one of the locations is not optimal, and that we are missing the benefits from a quarter of the total studies based in China.

When inspecting the top sponsor with most clinical trials in UNKNOWN status (Chinese Academy of Medical Sciences), we observed that 49% of the clinical trials sponsored by this institution are currently in UNKNOWN status. Although this is a small number in comparison with the totality of clinical trials, it is a missed opportunity of learning from the research conducted by this institution.

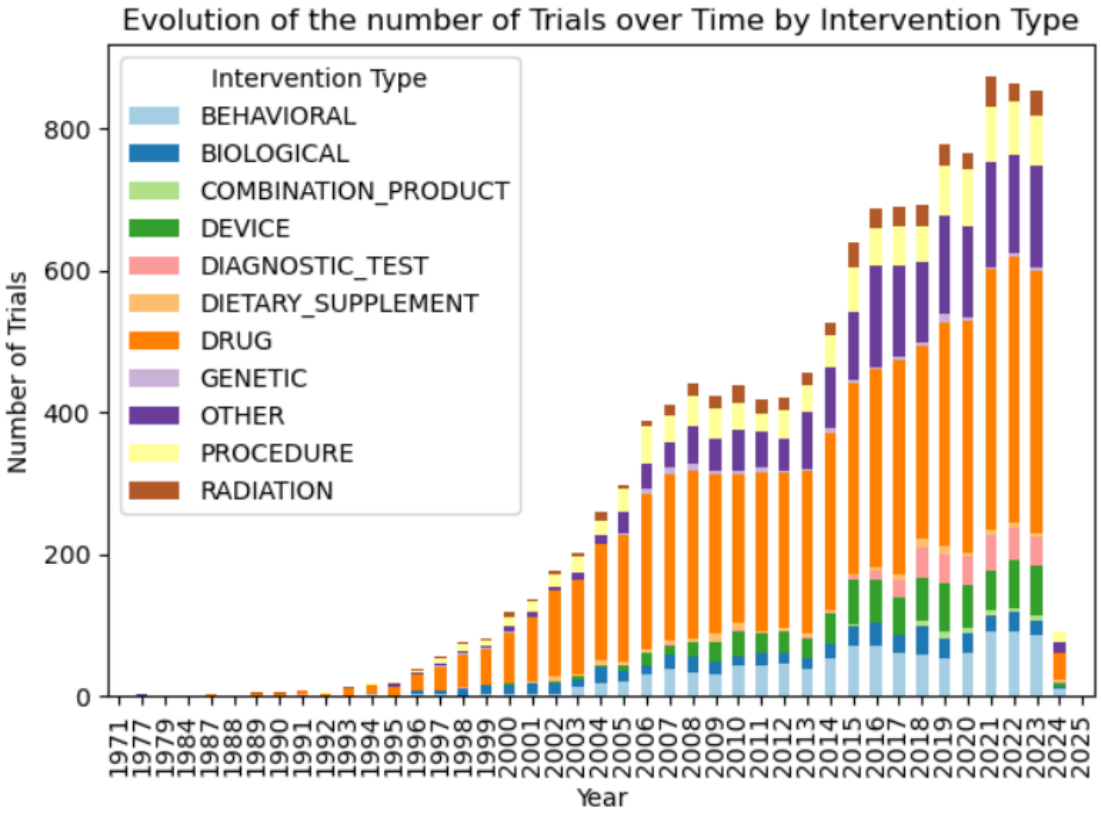


Figure 3.2. Evolution of the number of Trials over Time by Intervention Type

Figure 3.2 is a great summary of the evolution of breast cancer research, showing that before the 2000s reported clinical trials were mostly focused on drugs, and even though drug testing has continued to play an important role in breast cancer research, we are now observing clinical trials with focus on other intervention types. For example, starting from 2003 there is a steady increase in Behavioral intervention type and starting from 2017 there is a focus on diagnostic test for

breast cancer, as it was observed that early detection can significantly improve survival rates.

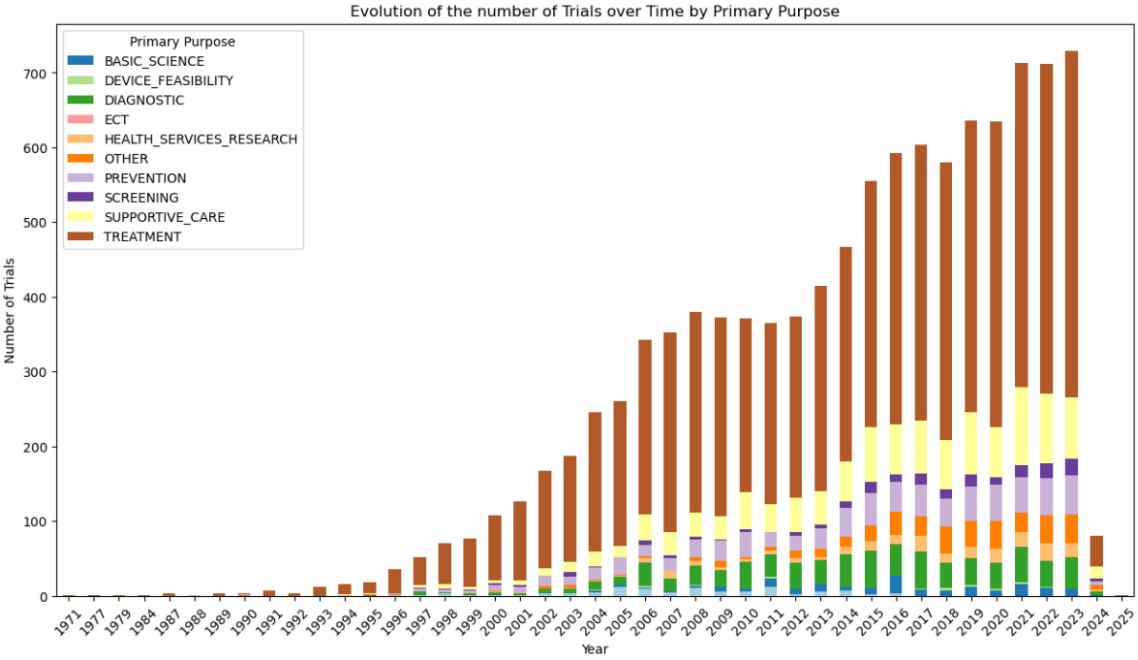


Figure 3.3. Evolution of the number of Trials over Time by Primary Purpose

Figure 3.3 further elaborates on the insights observed previously, offering a more detailed view on the research focus for breast cancer. We observe that treatment has always been one of the primary purposes, accounting for 45% of all clinical trials across all the reporting periods, and 63% of all interventional trials. This aligns closely with the predominance of drug interventions, as noted earlier. Analyzing the data year-over-year we can observe a significant diversification in the primary purpose of clinical trials. Particularly, there is a noticeable emphasis on supportive care, which has seen a significant increase since 2015.

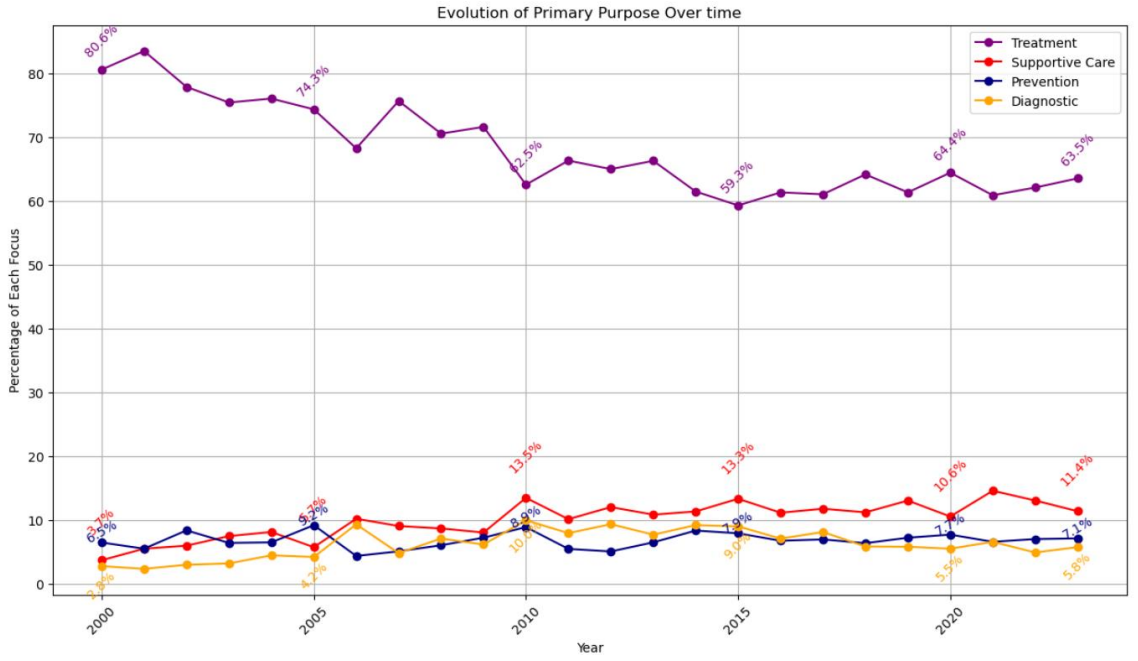


Figure 3.4. Evolution of Trial Primary Purpose over time

The figure above offers a detailed view of the evolution in the main primary purposes of clinical trials over the last 23 years. Notably, the emphasis on treatment has decreased from 80% to 63%, indicating a trending change in focus, by including other primary purposes in studies while treatment remains as the main focus. In contrast, supportive care has seen a significant increase from 3.7% to 11.4%. The attention towards prevention has remained relatively constant throughout this period. Lastly, the focus on diagnostic has nearly doubled, reflecting its growing importance in clinical trials as early detection has a direct impact on the patient's evolution.

After reviewing the data relating to intervention type and primary purpose, we wanted to dive deeper and learn more about the specific conditions that were targeted by the clinical trials. Figure 3.5 shows the ten most common breast cancer related conditions in current breast cancer research as part of the clinical trials reported on ClinicalTrials.gov.

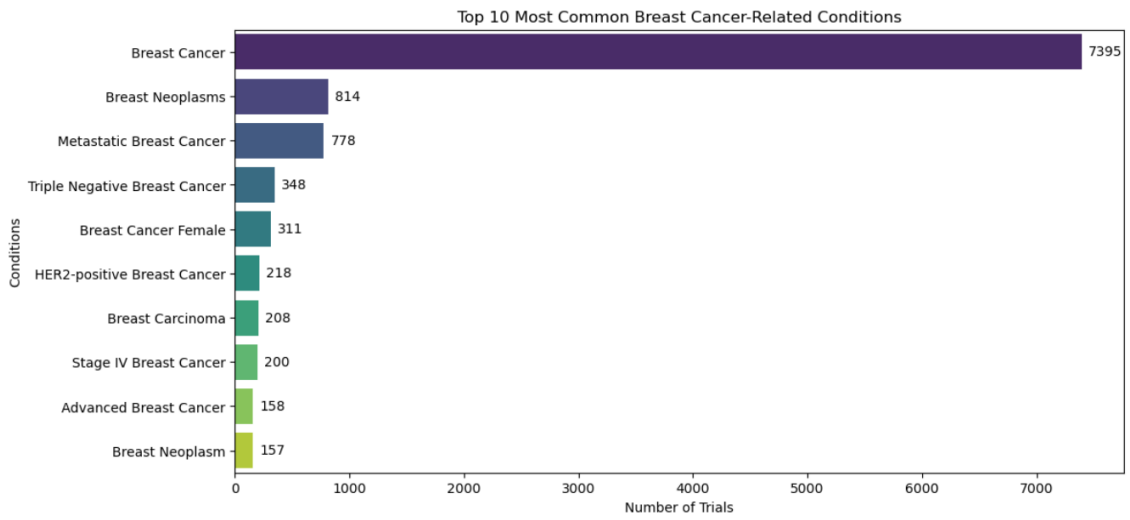


Figure 3.5. Top 10 Most Common Breast Cancer- Related Conditions

We can observe that some of the most commonly researched conditions are Metastatic (Stage IV) Breast cancer, HER-2 positive Breast Cancer and Triple-negative Breast Cancer (TNBC). In section 3 of this chapter, we will utilize this data to conduct a comparative analysis with real-world incidence rates of these conditions. This comparison aims to uncover potential disparities or gaps in research focus, providing insights into whether the most common conditions are receiving proportional research attention. Such an analysis is crucial for aligning clinical research efforts with the actual impact of breast cancer subtypes in the population.

To have a more general understanding of the conditions targeted by all studies related to breast cancer reported on ClinicalTrials.gov we have created a

wordcloud visualization. This shows an important focus on Stage III and Stage IV (Metastatic) breast cancer in research.

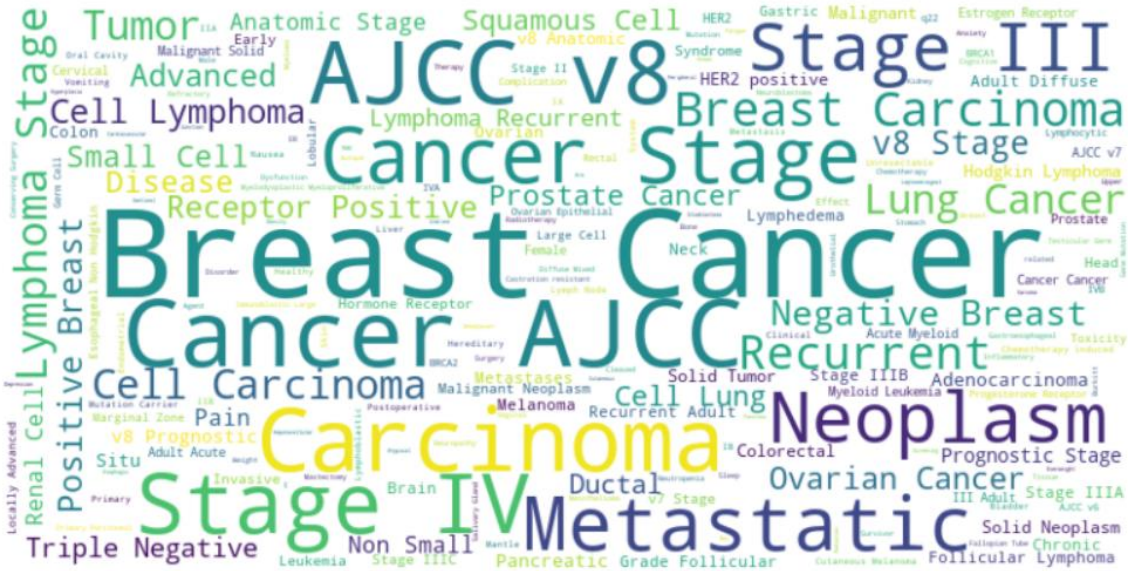


Figure 3.6. Wordcloud representation of the conditions targeted by breast cancer clinical trials

To continue our thorough analysis of the ClinicalTrials.gov dataset, we will review other characteristics of clinical trials, such as their status, results availability, or duration to gain more knowledge about their evolution.

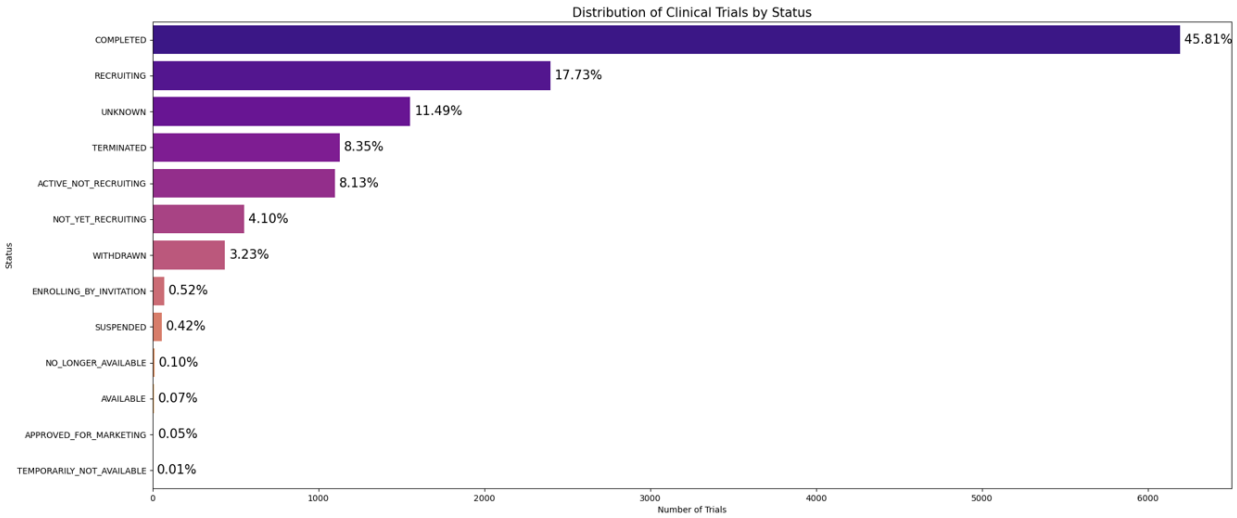


Figure 3.7. Distribution of Clinical Trials by Status

Upon examining the status of clinical trials, we can observe, Figure 3.7 shows a significant percentage of studies in ‘UNKNOWN’ status, representing 11% of the totality of studies. In this context, ‘UNKNOWN’ means that these trials have passed their completion date, and their status has not been verified within the past two years (30). This percentage raises concerns, as it implies that researchers and health care professionals lack data from one out of every ten studies.

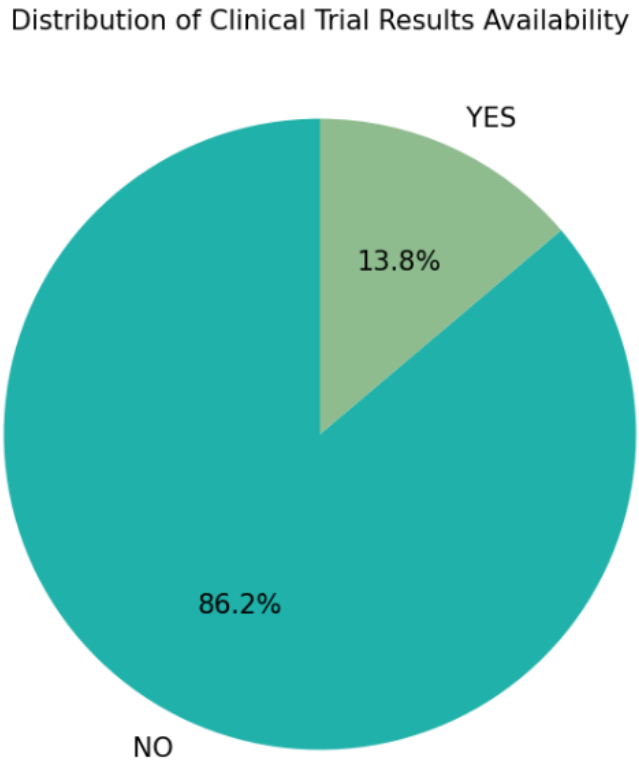


Figure 3.8. Distribution of Clinical Trial Results Availability

Figure 3.8 reveals a concerning statistic: less than 14% of the registered clinical trials have published results on ClinicalTrials.gov database. This lack of transparency about the study results slows down or hinders advances in the development of medical products and procedures.

In the United States, federal law mandates the submission of clinical trials results information within a year of the study's completion date. Despite the legal requirement, there is still a high percentage of clinical trials lacking reported results on ClinicalTrials.gov. For this reason, in April 2021, the U. S. Food and Drug Administration (FDA) released a statement regarding the enforcement of compliance measures for failure to submit required results. For these cases, the FDA will issue Notices of Noncompliance that provides sponsors with a number of days to submit the required information, and if they failed to do so, the FDA is authorized to seek money penalties from these companies.(31)

It is important to note that ClinicalTrials.gov serves an international community, and regulations might vary from country to country. In consequence, not all countries are required to submit results, which leaves concerning gaps in the database's information.

In terms of duration, we found that the average time for completing a clinical trial is of 1674 days. To understand what variables have an impact on clinical trial duration, we will look at the relationship between duration and some of the most important characteristics of clinical trials.

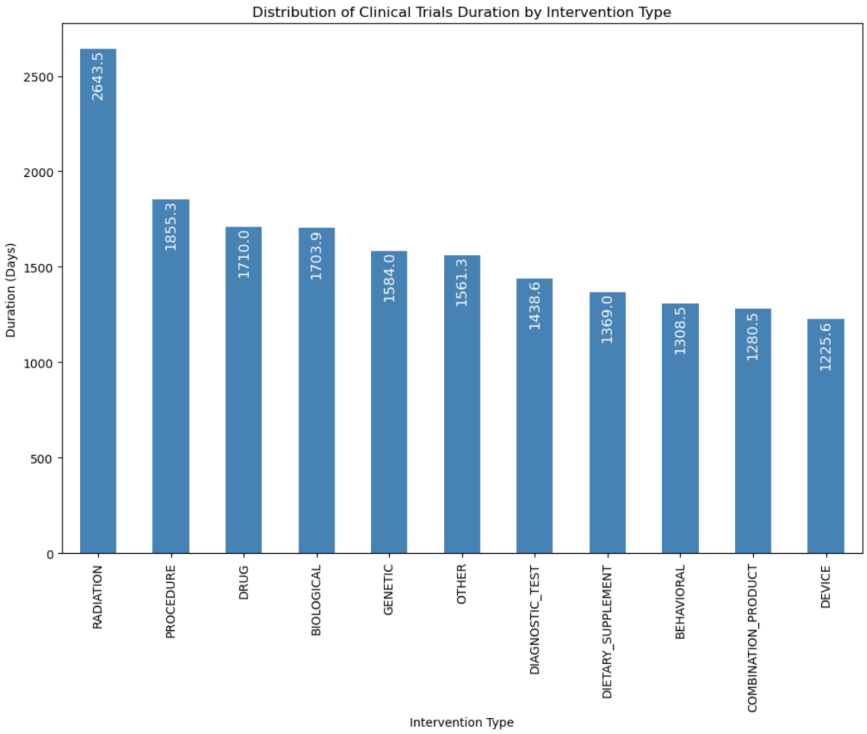


Figure 3.9. Distribution of Clinical Trials Duration by Intervention Type

In our analysis of clinical trial durations, we observed a notable trend where trials involving device intervention tend to have shorter duration. This could be attributed to the higher costs associated with maintaining these devices over prolonged periods of time. In contrast, trials that involve radiation interventions are typically the longest, likely due to the slower manifestation of radiation effects.

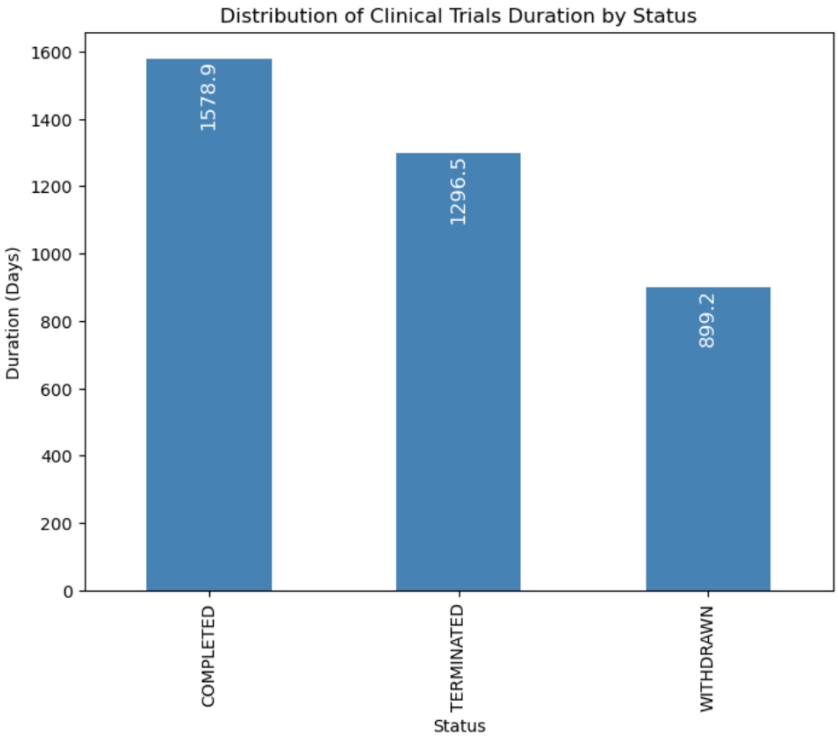


Figure 3.10. Distribution of Clinical Trials Duration by Completed Status

Analyzing the duration of closed trials, particularly those that are withdrawn or terminated, offers valuable insights into the research process. We found that, on average, withdrawn studies last over two years and terminated trials often exceed three years in duration. These durations suggest that researchers take substantial time to collect data and observe the results before reaching a decision to discontinue. However, this raises a yet unanswered critical question about resource allocation: is it possible to predict trial viability earlier in the process in order to promote a more efficient use of resources in clinical studies?

Lastly, in terms of clinical trial duration, we wanted to explore the relationship between duration and funder type:

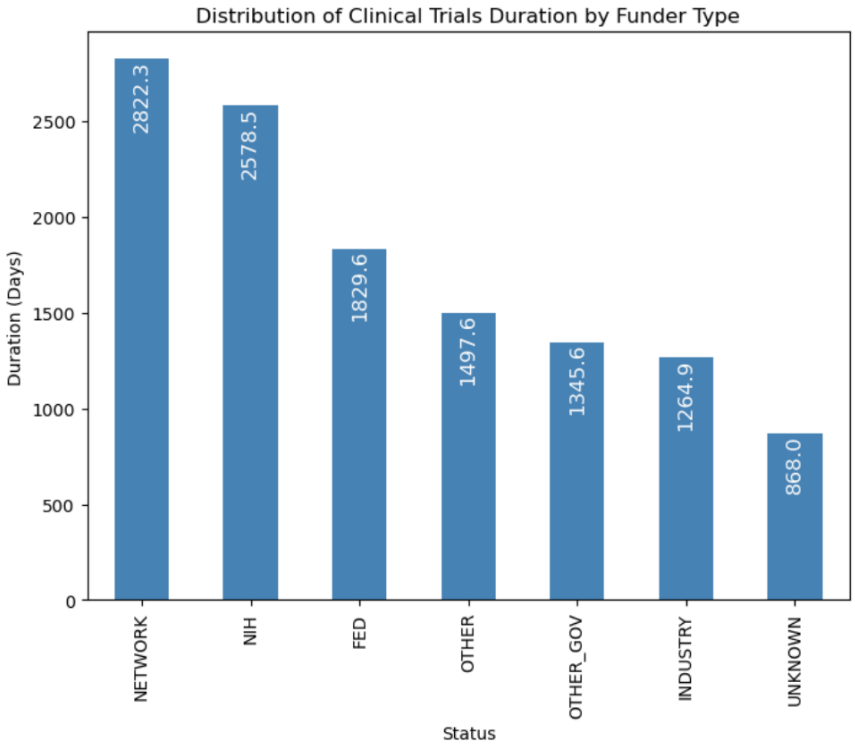


Figure 3.11. Distribution of Clinical Trials Average Duration by Funder Type

This graph suggests that publicly funded clinical trials have longer durations comparted to those funded by the industry. Industry-funded trials often adhere to strict objectives and timelines, driven by the need for cost efficiency and market-oriented results. In contrast, publicly funded trials might have more flexibility to focus on broader research objective over more extended periods.

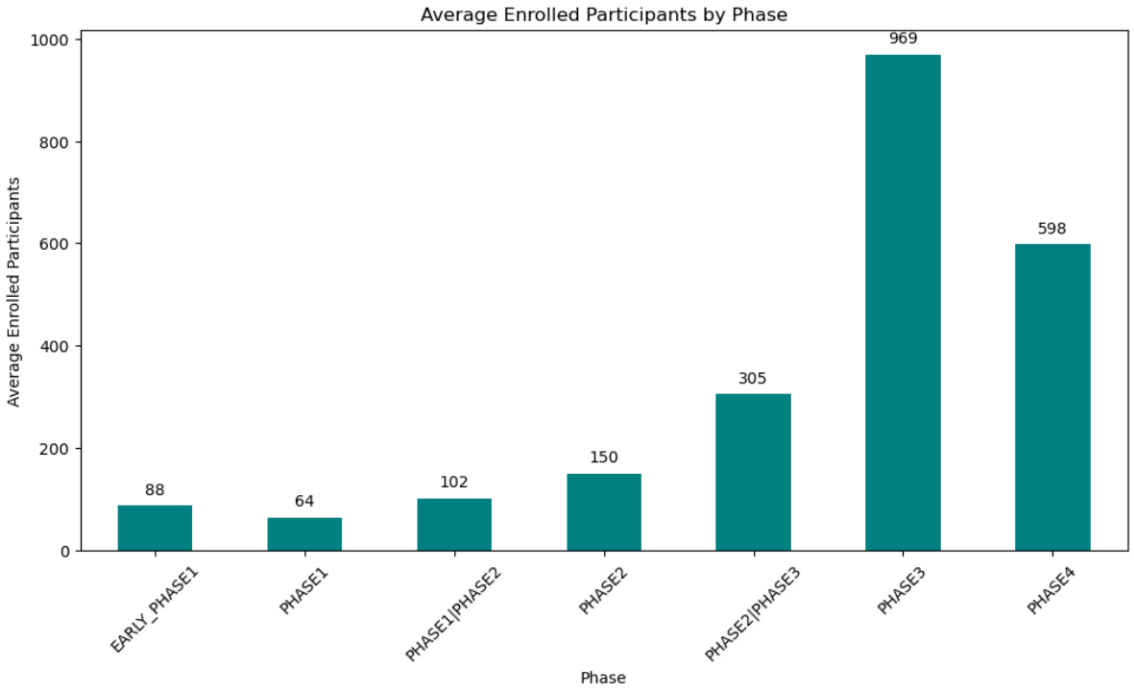


Figure 3.12. Average Enrolled Participants by Phase

According to U.S. Food and Drug Administration (FDA) (32) the size of clinical trials should follow these recommendations:

- Phase 1 - Between 20 and 100 healthy volunteers or people with the condition/disease
- Phase 2 - Up to several hundred people with the condition/disease
- Phase 3 - Between 300 and 3000 people with the condition/disease
- Phase 4 - Several thousand volunteers who have the condition/disease

Based on these specifications, upon analyzing the data from ClinicalTrials.gov we observe that sizes of Phase 1, Phase 2 and Phase 3 trials align with the expected values. Nevertheless, Phase 4 studies are noticeably smaller than the recommended size. This discrepancy raises concerns regarding the adequate monitoring of drug effectiveness and side effects once released to the general public. These concerns are also echoed in an article by Zhang et al. (33) which advocates for respecting an appropriate size for Phase 4 trials. This article emphasizes that larger sample sizes in this phase are crucial for the safe use of medications.

Another surprising observation is that the average size of Early Phase 1 trials is higher than that of the Phase 1 trials. Usually, Early Phase 1 trials are proof-of-concept rather than a full investigation to identify the correct dose to move to Phase 2 testing, and should include a small number of participants, typically between 10 and 15 (34). As Early Phase 1 clinical trials do not offer any therapeutic benefit, they are often prone to controversy due to their nature that can be seen as “experiments in people”, and sometimes even considered unethical (34). For this reason, their size should be better controlled to ensure there is a positive relationship between risks and benefits.

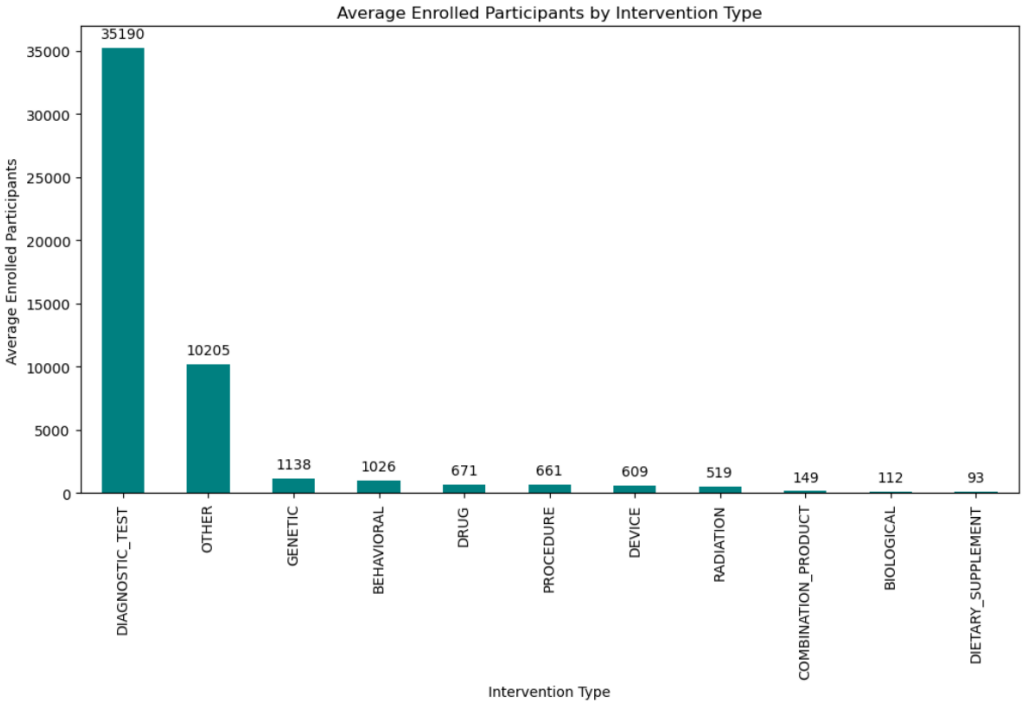


Figure 3.13. Average Enrolled Participants by Intervention Type

The largest average trial size is observed in diagnostic test interventions, which is anticipated given that screening processes typically require larger population samples comparted to other intervention types. Also, the available population is much higher for diagnostic tests, as it encompasses the whole population. In contrast, for other types of interventions, it is usually required to enroll only patients with the disease/condition which limits the scope of the clinical trials.

In order to shed some light on these numbers regarding study size, we looked at the participants' demographics to understand who participates in these clinical trials.

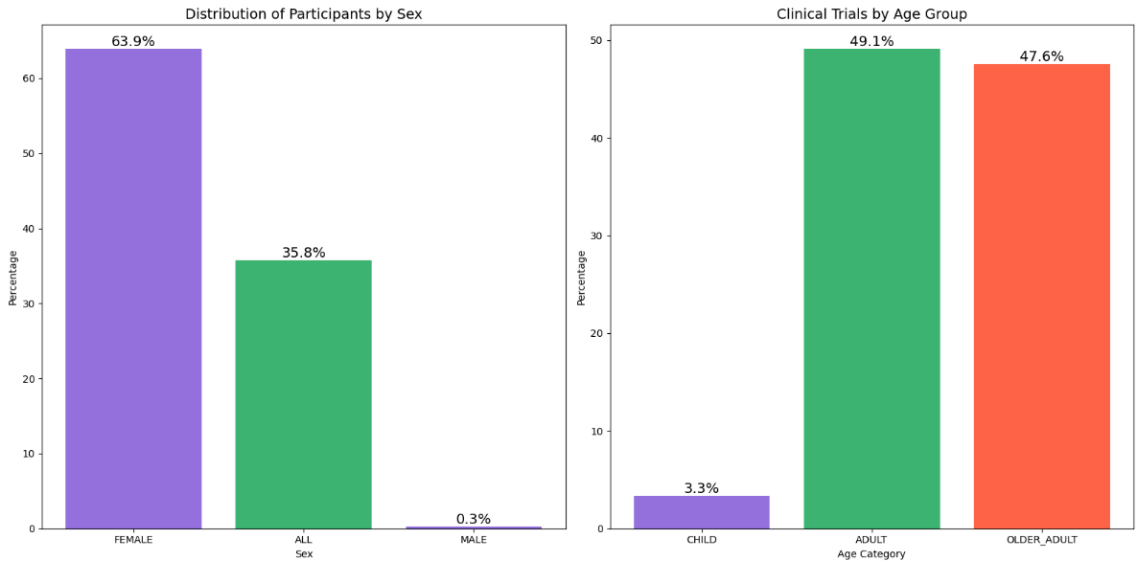


Figure 3.14. Participants' demographics (Gender and Age)

Our analysis reveals that a significant proportion of clinical trials are predominantly designed for female participants, accounting for 63,8% of the total participants. Meanwhile, about 36% of trials are structured to include both men and women, however, the actual percentage represented by each gender is unclear from the data that is available to us. Male participation in these trials represents only 0.3% of all participants, and 20% of these clinical trials are related to prostate cancer. This uncovers a gender inequality problem in breast cancer research, where there is opportunity for deeper research on breast cancer affecting men to come up with a better and more personalized approach for this type of cancer.

Regarding age demographics, 49% of trials include adults aged 18 to 64 years, while a 47% includes older adults, defined as those over 65 years of age. In contrast, only 3% of studies involve children.

This distribution will be further analyzed in comparison with the most recent breast cancer statistics in section 3 of this chapter, providing insights into potential gaps in research a cross age or gender groups.

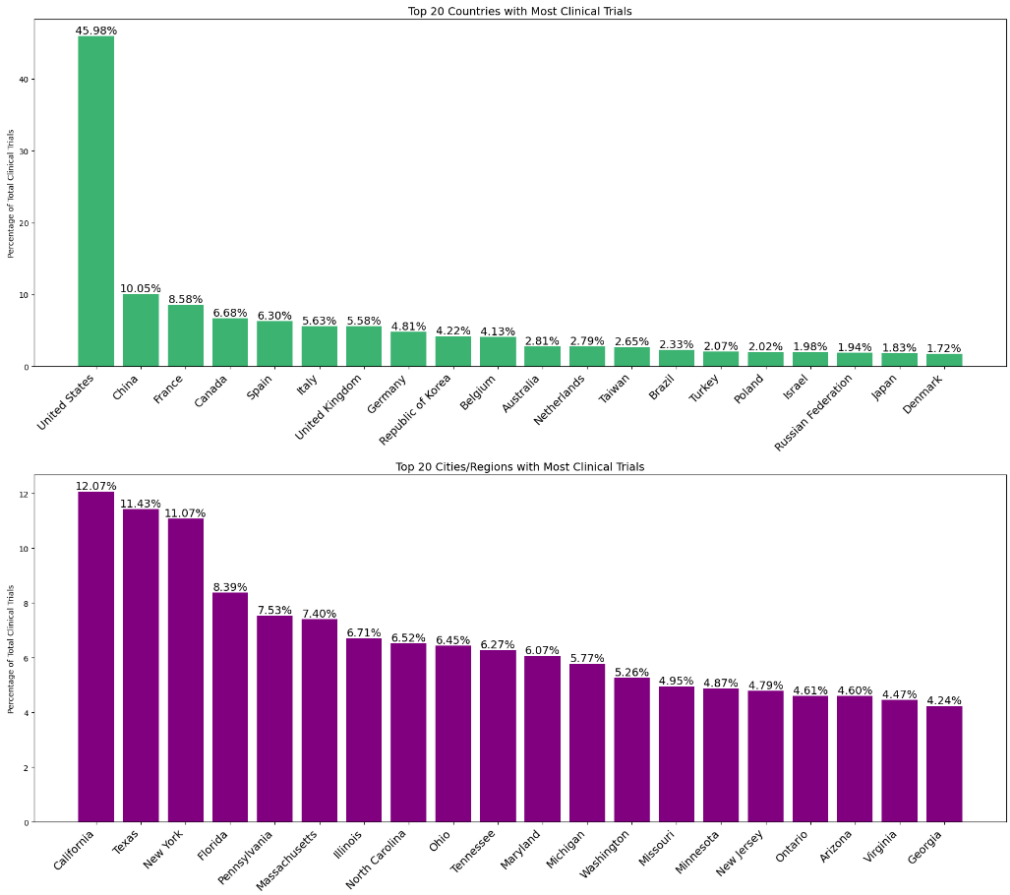


Figure 3.15. Top 20 countries and top 20 cities with most clinical trials

Our analysis reveals that almost half of the clinical trials are occurring in the United States (46%). Moreover, from the top 20 cities/regions where clinical trials are located, 19 are US cities or states. The countries appearing in the top 20 are located either in North America, Europe, or Asia, while other continents are notably absent from the top 20. This is especially concerning when observing that the 20th country from the top 20 represents 1.72% of all clinical trials.

Figure 3.16 shows the 10 cities outside of the United States with most clinical trials conducted, as we can observe that Spain and Canada have a very strong presence in these trials.

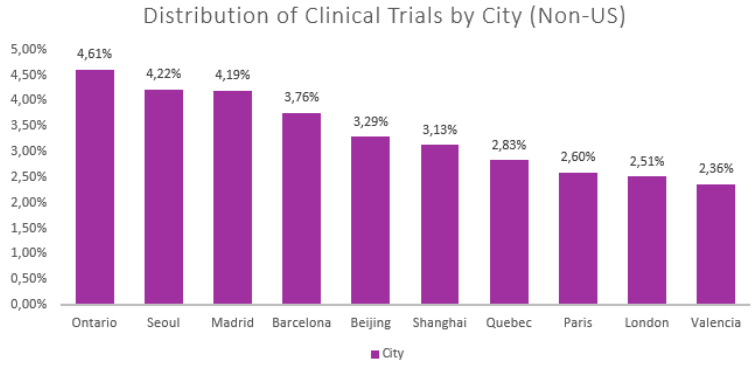


Figure 3.16. Top 10 cities with most trials (Non-US)

To gain a deeper understanding of this distribution, we compared the location of clinical trials with the population of each country. For this task, we imported population data from the World Bank, and we explored the relationship between each country's proportion of global clinical trials and its proportion of the world's population. This comparative analysis will help identify any gaps in breast cancer research based on location.

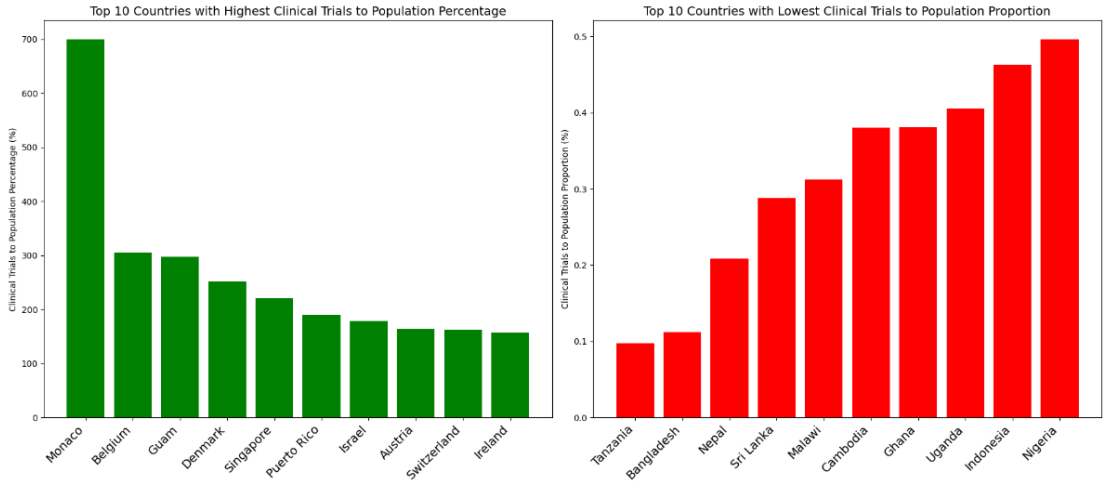


Figure 3.17. Top 10 Countries with Highest/Lowest Clinical Trials to Population Proportion

We can observe a list of countries with high gross domestic product (GDP) per capita on the left. In fact, Monaco is the country with the highest GDP per capita in the world and from the 10 countries on the left, 8 are in the top 20 countries with highest GDP per capita in the world. (35) If we observe the chart on the right, we find the other side of the coin, with the highest ranked country in the top GDP per capita being on the 112nd position. This reflects that there is a strong correlation between the resources that a country disposes of and the access to innovative treatments for their population. While it is true that the research that is realized in the rest of the world, has a positive impact on poorer countries as drugs can be released to the public in their countries, we could be missing a significant amount of data about breast cancer if we don't include a wider range of population in the clinical trials.

The choropleth map below is a clear visual representation of the data discussed above:

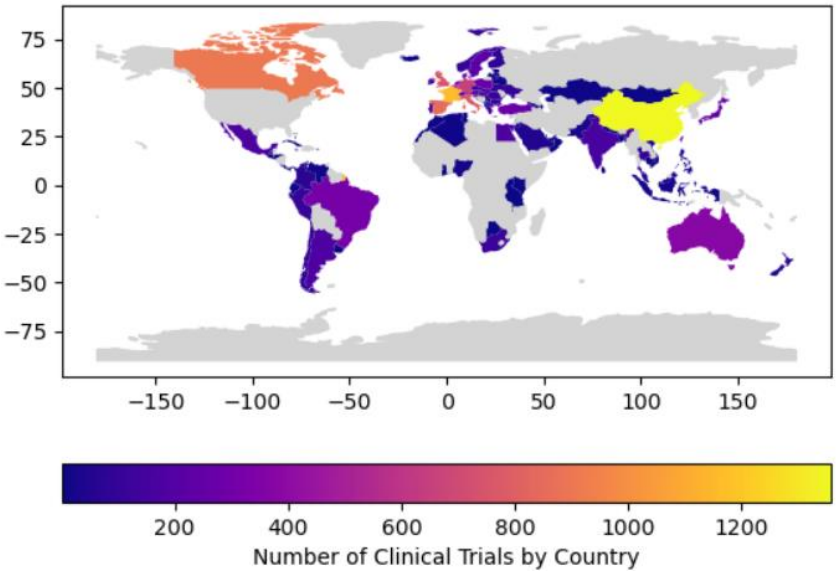


Figure 3.18. Cloropleth map representing number of clinical trials by country.

Having thoroughly analyzed various key aspects of clinical trials, including their evolution over time, types of interventions, primary purposes, conditions targeted, status, availability of results, duration, and size, as well as examined the demographics of participants (age and sex) and geographical locations, or focus now shifts to understanding the funding and collaborative dimensions of these trials. In the next phase of our analysis, we will concentrate on identifying the sponsors and collaborators involved in clinical trials to understand the broader context and networks collaborating on this essential research work.

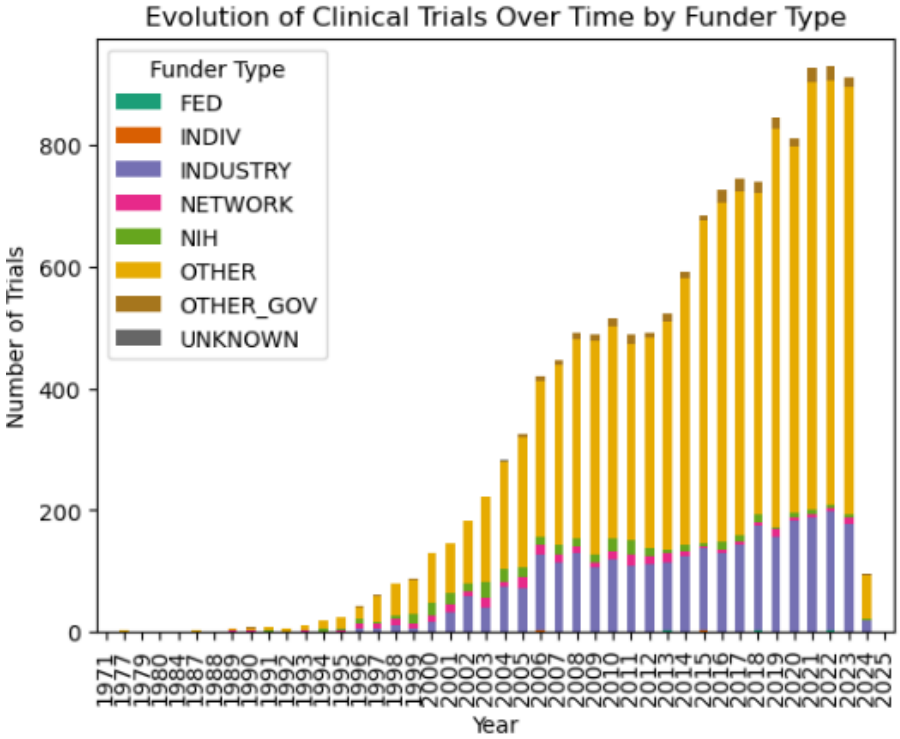


Figure 3.19. Evolution of Clinical Trials Over Time by Funder Type

The bar chart in Figure 3.18 illustrates the evolution of clinical trials over time by funder type. We observe that a significant percentage of clinical trials have a funder type defined as ‘OTHER’. In order to better understand what this category refers to we will look at the top 10 sponsors that are related to these clinical trials.

Sponsor	Clinical Trials, n
M.D. Anderson Cancer Center	251
Memorial Sloan Kettering Cancer Center	247
Fudan University	185
Mayo Clinic	150
Dana-Farber Cancer Institute	145
Alliance for Clinical Trials in Oncology	96
City of Hope Medical Center	84
Northwestern University	74
Washington University School of Medicine	74

Table 3.5. Top sponsors with funder type ‘OTHER’.

Most of these sponsors are universities and cancer research centers that usually obtain funding from different sources, such as government grants, private donations, foundations and other university founds. This variety of sources together with the lack of reporting, makes it hard to see transparency in the funding of clinical trials. It would be important to know exactly who is funding these studies to better understand their purposes and ensure that trials are not biased based on their funding sources.

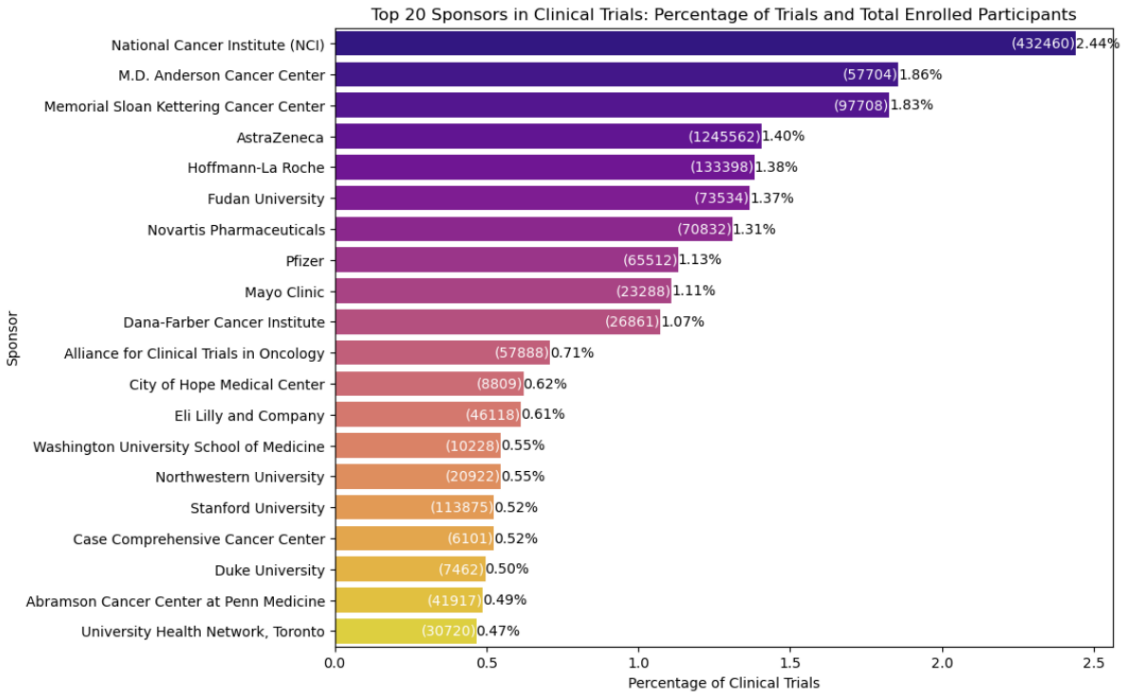


Figure 3.20. Top 20 Sponsors in Clinical Trials: Percentage of Trials and Total Enrolled Participants

The figure above presents the list of the top 20 sponsors who managed more breast cancer clinical trials, with the percentage of trials they were involved in, and the total number of participants enrolled in their studies. We observe that most these sponsors are located in the United States, which is expected as we learned earlier that almost half of the total clinical trials have United States as location. We also observed that there are 5 big pharmaceutical companies listed in this top 20 (AstraZeneca, La Roche, Novartis, Pfizer and Eli Lilly and Company).

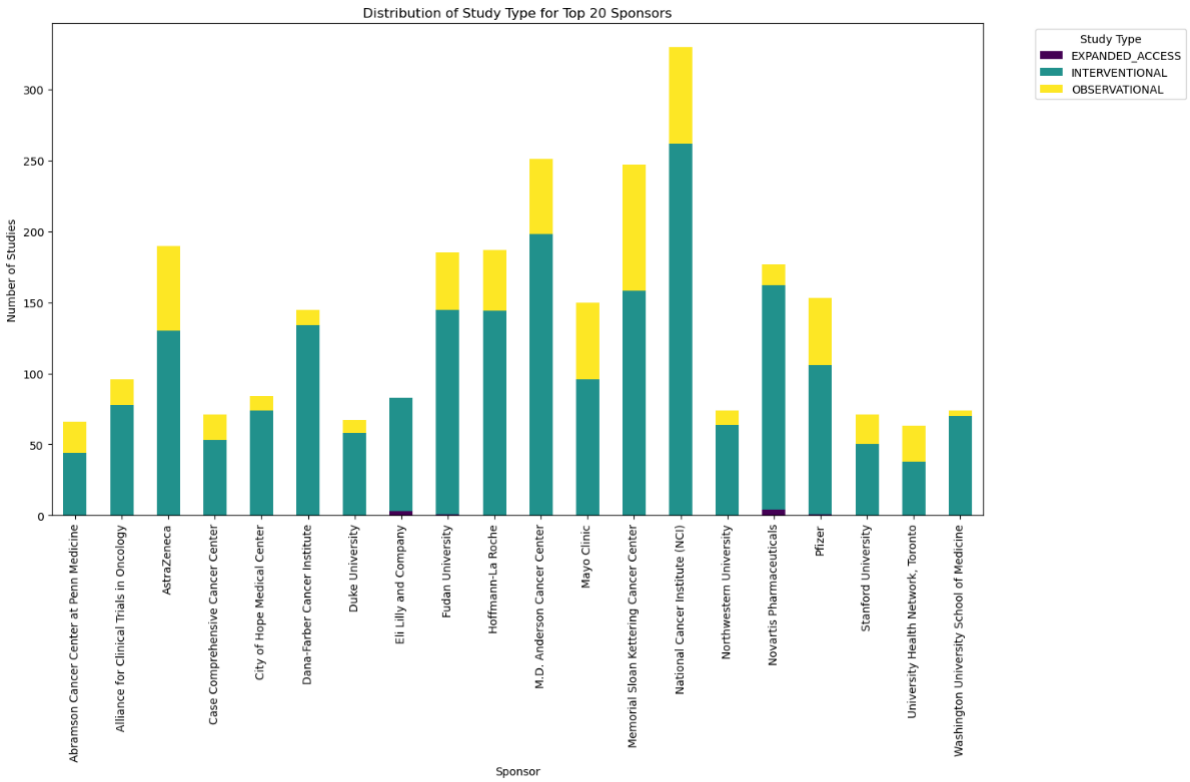


Figure 3.21. Distribution of Study Type for Top 20 Sponsors

Our analysis of the sponsors listed in the top 20 reveals that they manage both Interventional and Observational studies, while only “Eli Lilly and Company” and “Novartis Pharmaceuticals” perform a significant amount of expanded access clinical trials. This distinction is likely related to the company’s policies regarding compassionate use. For instance, Novartis has established a program called ‘Managed Access Programs’ for compassionate use which enables patients with serious or life-threatening medical conditions to access locally unlicensed medication.(36)

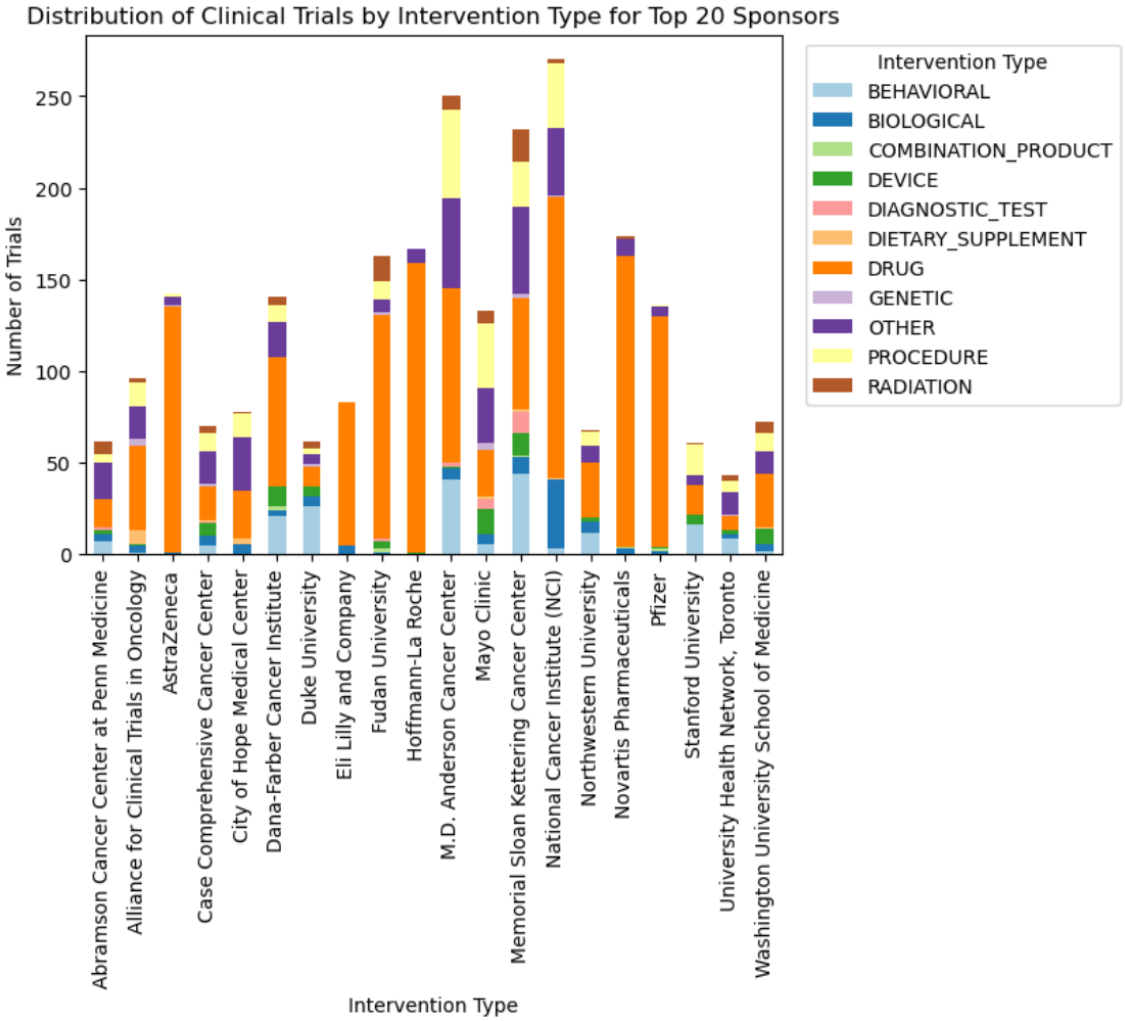


Figure 3.22. Distribution of Clinical Trials by Intervention Type for Top 20 Sponsors

Diving deeper into our analysis on sponsors, we observe distinct patterns emerge based on the type of sponsor. Predominantly, pharmaceutical companies are involved in trials related to drug development, which aligns to their commercial activity of creating new drugs for which they are expected to conduct the necessary clinical trials before commercialization. On the other hand, behavioral studies are more often associated with Cancer Centers and Universities. Additionally, diagnostic test studies are more often conducted by Cancer Centers and other clinics, most likely due to their direct access to diagnostic tools and patient populations.

Lastly, when considering the diversity of interventions covered by the clinical trials, we observe in Appendix B that “Hoffman-LaRoche” and “Eli and Lilly and Company” focus their studies on a very limited number of intervention types, device, drug and other for the first, and biological and drug for the latter. On the other hand, institutions like Memorial Sloan Kattering Cancer Center, covering all the intervention types and Mayo Clinic, covering 9 out of the 10 intervention types available offer a much more diverse approach to breast cancer research.

In order to visualize the collaboration between sponsors and collaborators, we created an undirected graph and used the Louvain method for community detection.

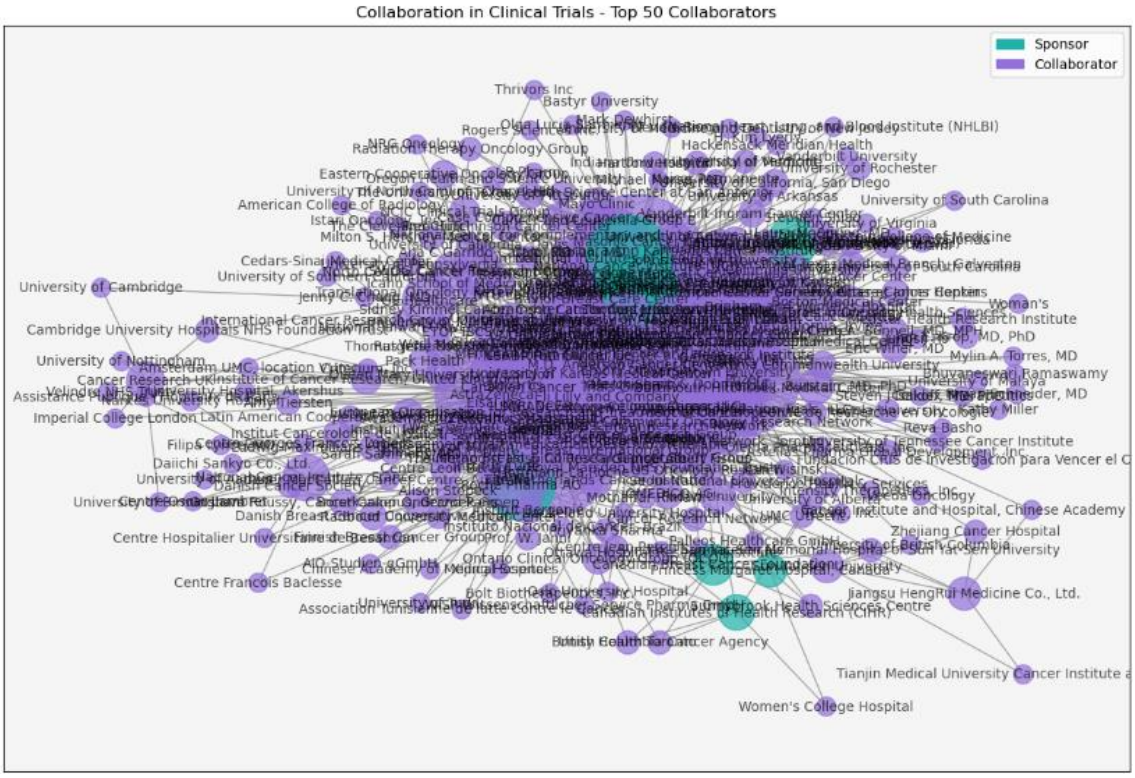


Figure 3.23. Collaboration Network with Community Structure

This graph shows the collaboration communities that were created between sponsors and collaborators. Upon examination, we observe that many of the collaboration networks are local, for example, the Cancer Research UK organization collaborates with different hospitals and universities in UK on breast cancer research. For this visualization, only the top 50 most active collaborators were considered, and leaves (nodes with only one connection) were excluded in order to reduce the complexity of the graph. A few interesting relationships within the network of collaboration have been included in Appendix C and D.

In conclusion, this comprehensive EDA of the trials network has unveiled some critical insights into the planning and development of breast cancer clinical trials. By examining the various aspects of the trials, such as intervention types, primary purpose, participant demographics, among others, we gained a deeper understanding of the trends and patterns in clinical research.

3.3.2 Machine learning techniques applied to clinical trial analysis

As we transition from the exploratory data analysis of our clinical trials dataset, our next step is to employ machine learning techniques to uncover deeper patterns and trends. The initial step in this process will be the application of clustering algorithms. Clustering will enable us to group similar trials based on various attributes such as study design, demographics, geographical distribution, or intervention types. By doing so, we aim to identify grouping criteria and

relationships between variables that might not have been apparent during the EDA.

Following the clustering, we will implement a Principal Component Analysis (PCA) to help us reduce the complexity of the dataset to its most informative components. This step will allow us to uncover correlations that might not have been discovered during the EDA and clustering phases.

3.3.2.1 Clustering

The algorithm we have chosen for this phase is K-means, as it is a powerful technique to group data into distinct clusters based on data similarities. This method can handle large datasets efficiently which makes it an ideal choice for our clinical trials dataset.

Before being able to apply K-means, we need to review the variables, address the missing values that were not addressed during the data cleaning phase and perform encoding for categorical variables.

For missing values, we decided to fill in the missing values for ‘phases’ and ‘most_common_intervention’ variables with the most frequent value, and the missing values for ‘duration’ with the median value.

As a next step, we tested two different approaches to encoding: one-hot encoding and label encoding. We found that one-hot encoding was a better option for K-means, which label encoding was more suitable for PCA.

Prior to applying the K-means algorithm, we used the following techniques to make the dataset adequate for clustering:
1. **One-hot encoding** for some variables ('status', 'study_type', 'funder_type', 'most_common_intervention', 'sex')
2. **Binary encoding** for ‘study_has_results’ and ‘study_has_documents’
3. **Variable reduction** - dropped some of the variables to avoid duplication as they were already encoded with one-hot encoding or because they did not have analytical relevance for the study.
    - Variables such as {'status', 'sex', 'phases', 'funder_type', 'study_type', 'start_date', 'interventions', 'collaborators', 'age', 'phases_split', 'study_design' and 'most_common_intervention',}, and
    - Variables such as {'trial_id', 'title', 'brief_summary', ‘sponsor’, 'primary_outcome_measures', 'secondary_outcome_measures', 'other_outcome_measures', 'primary_completion_date', 'completion_date', 'first_posted', 'results_first_posted', 'last_update_posted', 'conditions', 'countries', 'cities', 'combined_outcomes', 'cleaned_outcomes'}

After this processing phase, the clinical trials dataset contains 43 variables that will be employed for clustering. This higher number of variables is motivated by

the application of one-hot encoding, as it creates a new column for each unique category of the variables that it encodes.

	study_has_results	enrollment	study_has_documents	duration	year	status_ACTIVE_NOT_RECRUITING	status_APPROVED_FOR_MAR
0	0	100.0	0	1280.0	2014	False	
1	0	30.0	0	1603.0	2016	False	
2	1	55.0	0	3073.0	2010	False	
3	1	10.0	1	953.0	2017	False	
4	0	30.0	0	401.0	2021	False	

5 rows × 43 columns

Figure 3.24. Head of dataset prepared for clustering

Before applying the K-means algorithm, we explored the relationship between variables through a correlation matrix represented in the figure below.

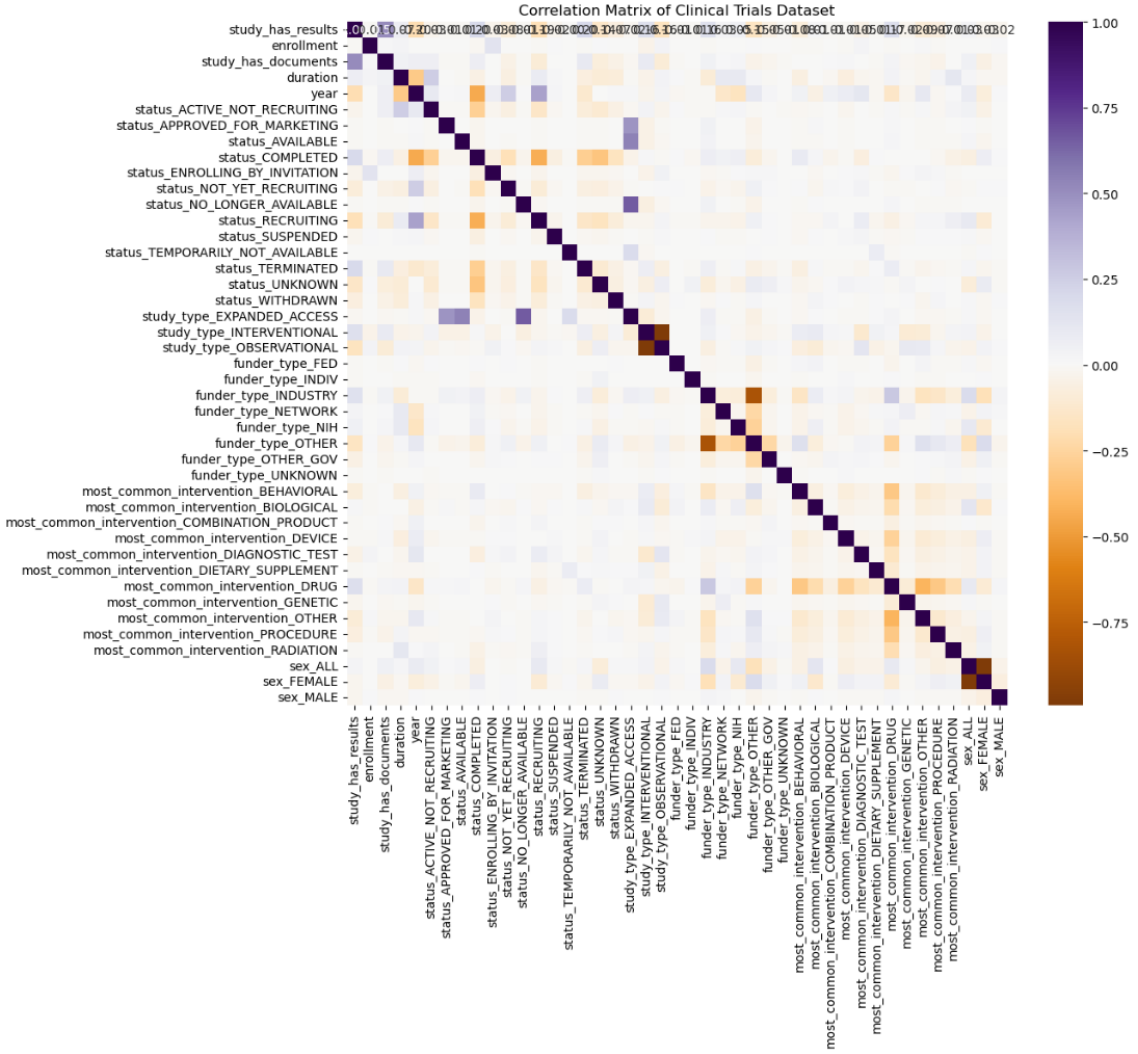


Figure 3.25. Correlation Matrix of Clinical Trials Dataset

The most notable correlations that we observed were between the variables ‘study_has_documents’ and ‘study_has_results’. When looking at the types of documents that are usually included in studies, those are either Study Protocol or Statistics Analysis Plan, or even both, in some cases. Publishing these

documents shows a compromise with transparency, and this is likely related to the publication of results upon the trial's completion.

There is also a reasonably strong correlation between EXPANDED_ACCESS study type and both AVAILABLE and NO_LONGER_AVAILABLE status values. Other interesting correlations that we observed during the Exploratory Data Analysis is between the funder type 'INDUSTRY' and the intervention type 'DRUG'.

Before applying the K-means algorithm, we used the elbow method to find the optimal value of k, which defines the number of clusters the data will be divided into.

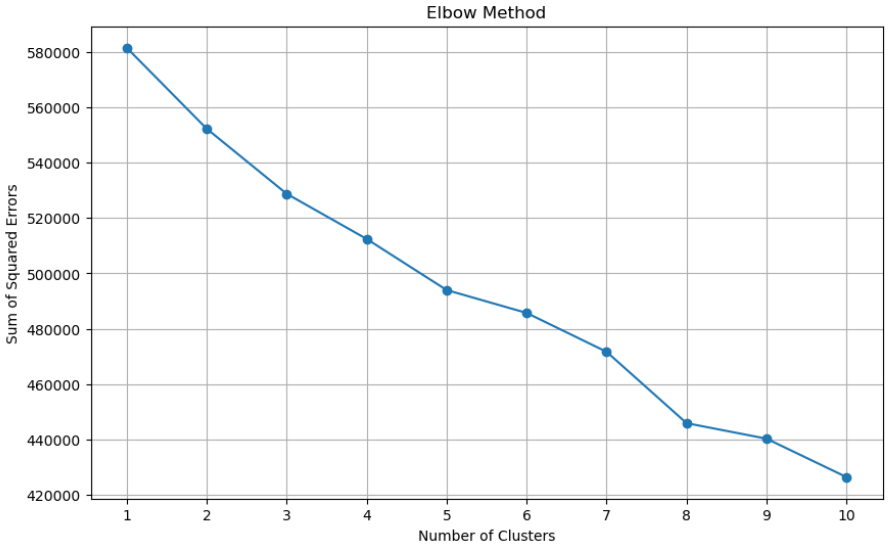


Figure 3.26. Elbow method visualization

Given that the Elbow method did not signal any strong elbow point in its visualization, we turned to the silhouette method as an alternative approach to find the optimal k.

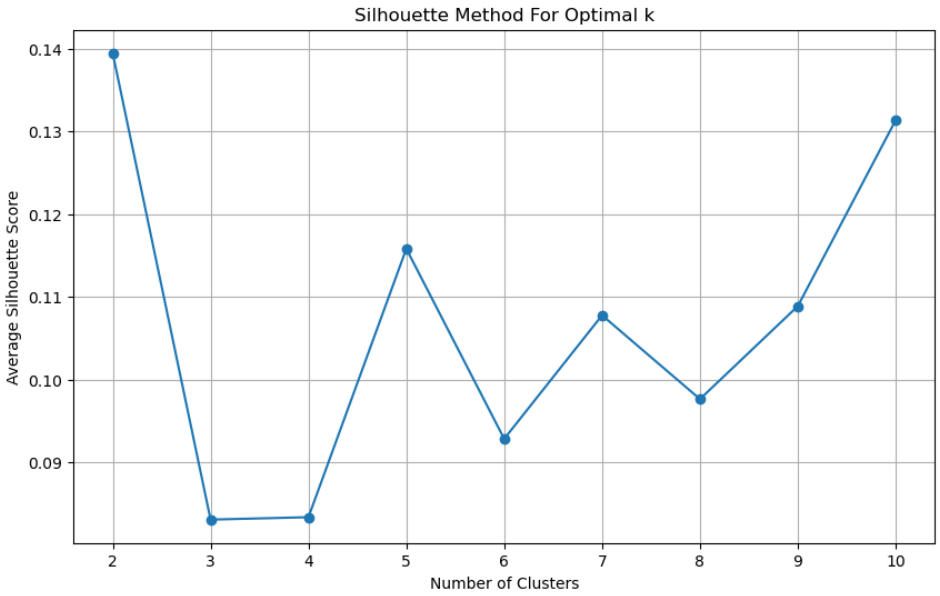


Figure 3.27. Silhouette method visualization

The Silhouette method has indicated 2 as the optimal number of clusters (k) for our K-means algorithm. Based on this data, we will proceed by applying the K-means algorithm to our scaled dataset, setting k to 2. This approach will enable us to segment the database into two clusters.

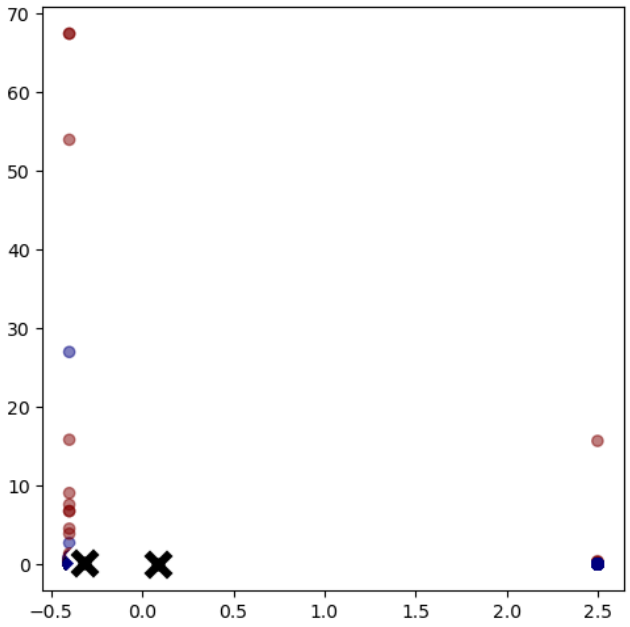


Figure 3.28. K-means generated clusters visualization

The scatter plot for the data points has not provided us with enough information to understand the grouping of data points. A helpful method of identifying common characteristics between clusters is by reviewing the centroids of the clusters to find defining patterns.

Variable	Centroid 0	Centroid 1
study_has_results	0.08	-0.31
enrollment	-0.02	0.08
study_has_documents	0.05	-0.18
duration	-0.02	0.08
year	-0.02	0.06
status_ACTIVE_NOT_RECRUITING	0.02	-0.07
status_APPROVED_FOR_MARKETING	-0.02	0.09
status_AVAILABLE	-0.03	0.10
status_COMPLETED	0.01	-0.03
status_ENROLLING_BY_INVITATION	-0.02	0.08
status_NOT_YET_RECRUITING	-0.00	0.01
status_NO_LONGER_AVAILABLE	-0.03	0.12
status_RECRUITING	-0.01	0.05
status_SUSPENDED	-0.01	0.02
status_TEMPORARILY_NOT_AVAILABLE	-0.01	0.03
status_TERMINATED	0.03	-0.13
status_UNKNOWN	-0.03	0.13
status_WITHDRAWN	0.01	-0.05
study_type_EXPANDED_ACCESS	-0.05	0.18
study_type_INTERVENTIONAL	0.52	-1.94
study_type_OBSERVATIONAL	-0.51	1.93
funder_type_FED	-0.00	0.00
funder_type_INDIV	-0.00	0.00
funder_type_INDUSTRY	0.03	-0.11
funder_type_NETWORK	0.00	-0.01
funder_type_NIH	-0.01	0.03
funder_type_OTHER	-0.02	0.08
funder_type_OTHER_GOV	-0.01	0.04
funder_type_UNKNOWN	-0.00	0.02
most_common_intervention_BEHAVIORAL	0.04	-0.15
most_common_intervention_BIOLOGICAL	0.05	-0.18
most_common_intervention_COMBINATION_PRODUCT	0.01	-0.05
most_common_intervention_DEVICE	0.01	-0.03
most_common_intervention_DIAGNOSTIC_TEST	-0.06	0.23
most_common_intervention_DIETARY_SUPPLEMENT	0.02	-0.09
most_common_intervention_DRUG	0.01	-0.04
most_common_intervention_GENETIC	-0.05	0.20
most_common_intervention_OTHER	-0.05	0.18
most_common_intervention_PROCEDURE	-0.01	0.04
most_common_intervention_RADIATION	0.02	-0.07
sex_ALL	0.02	-0.08
sex_FEMALE	-0.02	0.07
sex_MALE	-0.00	0.01

Table 3.5. K-means clusters centroids

Upon reviewing the centroids data, we identified some notable differences between the two clusters that we have captured in the table below:

	Cluster-1	Cluster-2
Studies with results	Higher presence	Lower presence
Timeline	Less recent trials	More recent trials
Enrollment size	Lower than average enrollment size (smaller trials)	Higher than average enrollment size (larger trials)
Interventional Studies	Higher presence of INTERVENTIONAL type of studies	Lower presence of INTERVENTIONAL type of studies
Observational Studies	Lower presence of OBSERVATIONAL type of studies	Higher presence of OBSERVATIONAL type of studies
Variety of studies	Higher variety of intervention types	More focus on DIAGNOSTIC TESTS
Gender diversity	More gender diversity	More focus on female gender

Table 3.6. Features of the k-means generated clusters

Applying k-means clustering to our breast cancer clinical trials dataset, with an optimal number of clusters of 2, has provided interesting insights into different types of trials. The centroids data revealed important distinctions between the two clusters. The first cluster seems to be focused on smaller studies on a variety of intervention types that are not very recent, while the second cluster includes more recent trials, with a more diverse focus on both interventional and observational studies and a focus on diagnostic tests, which explains the higher than average enrollment size.

3.3.2.2 Principal Component Analysis (PCA)

As we progress in our analysis of our breast cancer clinical trials dataset, the next step is to employ Principal Component Analysis (PCA) to reduce dimensionality of our dataset while preserving its essential characteristics. This will allow us to visualize and interpret data more easily. By identifying the principal components, we aim to uncover the most influential variables causing variation in our dataset.

Due to the fact that one-hot encoding is not the most appropriate encoding method for PCA, we will use label encoding on the processed dataset before applying the PCA algorithm.

We applied the PCA algorithm to the dataset with different number of components and observed the total explained variance:

Number of components	Total explained variance
2	29.4%
3	40.5%
4	50.2%

Table 3.7. PCA total explained variance by number of components

We observed that the new components do not contribute significantly to the total explained variance, thus we would be adding complexity without much added value. For this reason, we will stop at a 3 component PCA, understanding that we need to take a very careful approach in the analysis of the data. Due to the high dimensionality and complexity of the data, increasing the number of components will reduce the benefit that we could obtain from the PCA.

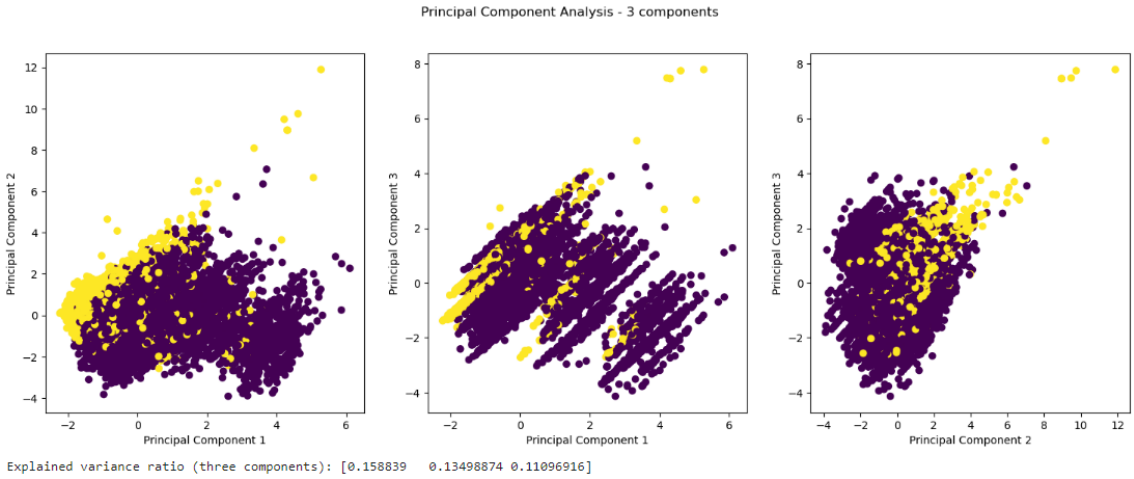


Figure 3.29. Principal Component Analysis – 3 components

In order to explore the three components, we will print out the loadings for the three components and review what is the contribution of the different variables to each component.

Variable	PC1	PC2	PC3
study_has_results	0.61	-0.08	-0.26
enrollment	0.00	0.06	0.03
study_has_documents	0.49	-0.19	-0.39
duration	0.26	0.46	0.34
year	-0.32	-0.41	-0.30
study_type_encoded	-0.23	0.24	0.10
funder_type_encoded	-0.25	0.32	-0.32
sex_encoded	-0.09	0.41	-0.44
phases_encoded	0.07	0.41	-0.41
status_encoded	-0.30	-0.13	-0.32
most_common_intervention_encoded	-0.03	0.26	0.01

Table 3.8. PCA Loadings for 3-component PCA

By reviewing the PCA loadings, we can observe the following characteristics of the components:

1. **Principal Component 1** – The highest positive loadings are represented by study_has_results and study_has_documents, and the highest negative loadings are year and status. This could indicate that Principal Component 1

is capturing the contrast between the presence of study documents and results in relation to the year of study or its status. We also observed a strong correlation between study_has_results and study_has_documents in previous phases of our analysis and this Principal Component 1 seems to leverage that correlation.

2. **Principal Component 2** – The highest positive loadings are represented by duration, sex and phases, while the highest negative loadings are represented by year. This would indicate that Principal Component 2 is capturing aspects related to the duration of the studies and specific demographic focus of the studies over different periods.
3. **Principal Component 3** – The highest positive loadings are represented by duration, and the highest negative loadings are sex and phases. Principal Component 3 might distinguish studies based on the duration versus their focus on different sexes and phases. We also observed in the K-means clustering that sex was a differentiator between the two clusters, where one was more diverse, and the other one more focused on female gender.

Overall, the PCA results suggest that the key differences in the breast cancer clinical trial dataset are related to the presence of study results and documentation, the trial's duration, the year and phase of the study and the focus on different genders. Although these components do not explain all the data, they give interesting insights into the factors that differentiate the clinical trials from the dataset.

3.3.3 Comparison between analysis results and latest cancer statistics

In the concluding section of this chapter, we will contrast our research findings with the most recent cancer statistics published in December 2020 by the World Health Organization (WHO) through the Global Cancer Observatory (7). This comparison is aimed at illustrating any disparities or unexplored areas in current research, thereby opening avenues for future research.

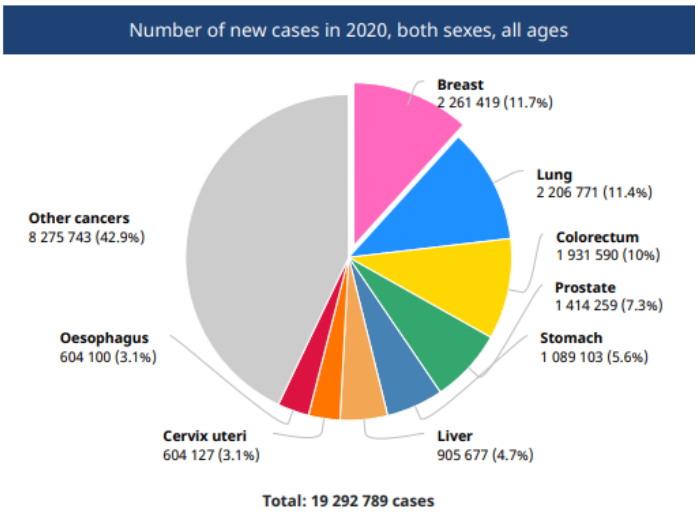
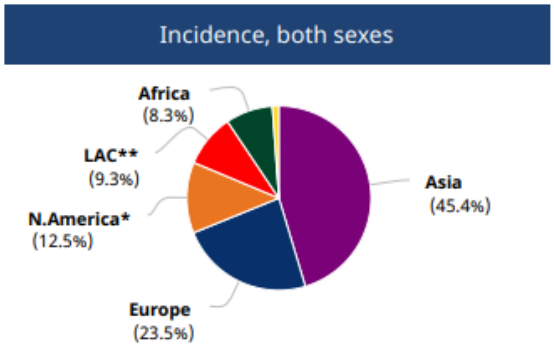


Figure 3.30. Number of new cases of Breast Cancer in 2020, for both sexes and all ages. Source: Globocan 2000 (7)

According to Globocan 2020 (7), breast cancer surpassed lung cancer in 2020 as the most commonly diagnosed cancer, accounting for 11.7% of all new cases detected in 2020. An analysis of the clinical trial data from ClinicalTrials.gov reveals that out of 101,922 cancer related clinical trials, 13524 are related to breast cancer (although they can include other cancer types in their design). This represents 13.26% of all cancer clinical trials, nevertheless these trials may also include other cancer types in their scope. This percentage is reasonably aligned with the proportion of cancer cases across all types of cancer, suggesting that the focus on breast cancer research is proportionate to its incidence worldwide.



	Population	Number
	Asia	1 026 171
	Europe	531 086
	*Northern America	281 591
	**Latin America and the Caribbean	210 100
	Africa	186 598
	Oceania	25 873
	Total	2 261 419

Figure 3.31. Incidence across continents, both sexes. Source: Globocan 2000 (7)

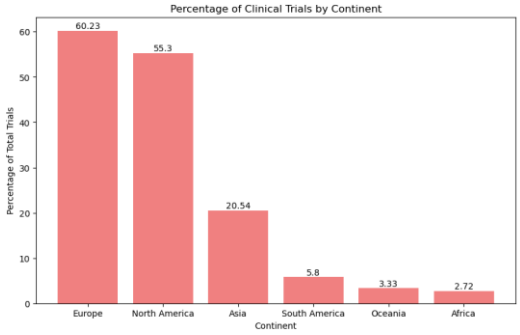


Figure 3.32. Distribution of Clinical Trials by Continent

In analyzing breast cancer incidence across continents, we find a notable disparity. Asia accounts for 45.4% of diagnosed cases worldwide, yet only 20.54% of clinical trials are conducted in Asian countries. On the other side, Europe hosts 60.23% of clinical trials, while the incidence of breast cancer on the continent is of 23.5%. Africa and Latin America have a smaller presence in the research priorities, with a 2.7% and 5.8% respectively, while the cancer incidence is significantly higher on these continents, 1.6 times higher in Latin America and three time higher in Africa. This data highlights a critical mismatch between research focus and disease incidence in various parts of the world.

These disparities are even more concerning if we consider the findings of the American Cancer Society(37) in regards to mortality from breast cancer: the breast cancer death rate in the U.S. is 40% higher in black women versus white women. This could be related to a combination of factors, including later diagnosis, higher rates of unfavorable tumor characteristics, or higher prevalence of other health conditions. Adopting a more inclusive research process could highly improve the life expectancy of black women who are diagnosed with breast cancer. This concern is even more acute when considering black women living in

the African continent, who may have limited access to screening and healthcare services, suggesting that the disparities might be even higher in a global context.

Although Globocan only offers breast cancer statistics pertaining to women, the American Cancer Society (37) indicates that men are not exempt from this disease, accounting for 1% of all breast cancer cases. However, when examining clinical trials data only 0.3% of studies specifically are designed for a male group. Of the totality of clinical trials, only 5 targeted Male Breast Cancer in particular. Men are often entirely absent from numerous statistical analyses in breast cancer research, underscoring a significant gap in representation and understanding of the disease in the male population.

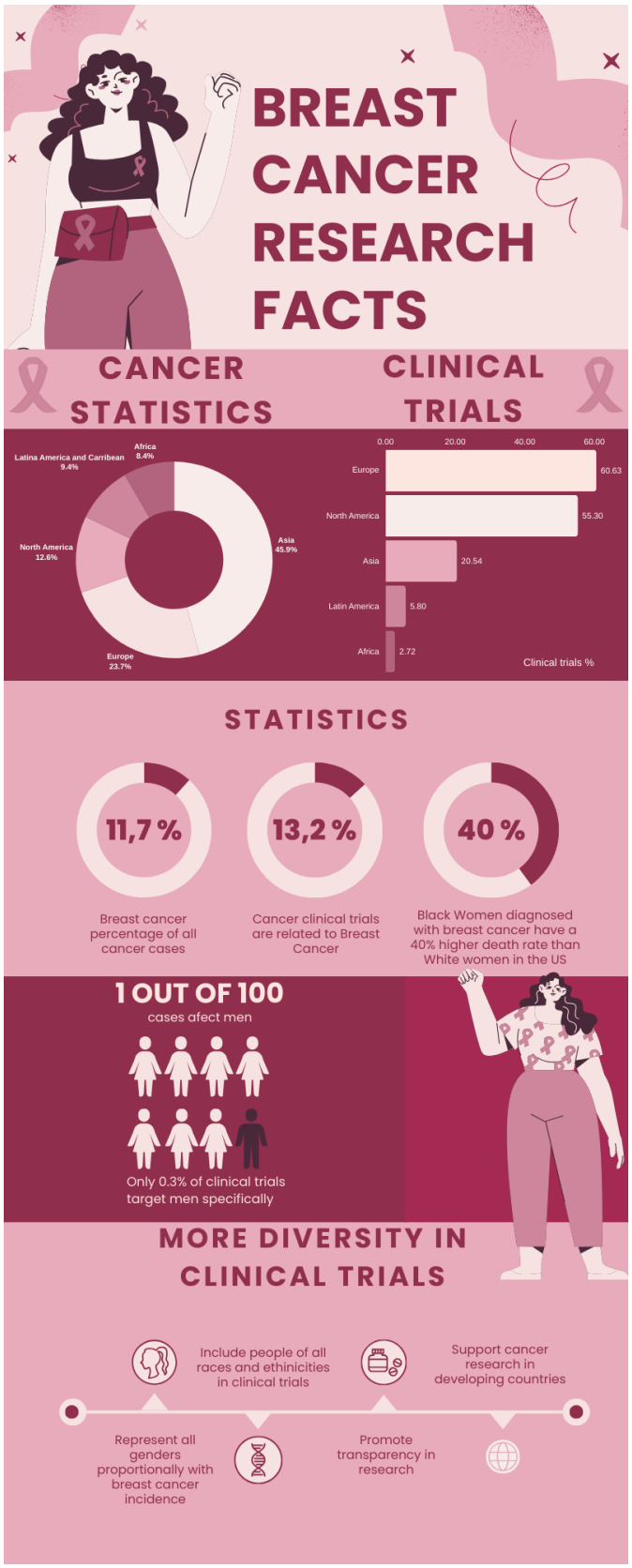


Figure 3.33. Breast Cancer Research Facts. Own work.

To sum up, as observed on the infographic above the highlighted disparities emphasize a broader issue of inequality in health research related to breast cancer. The geographical and gender imbalances in clinical trials not only reflect a gap in our current understanding but also signal a deeper issue rooted in the inequitable allocation of research resources. As anticipated in chapter 1.3, diversity of clinical research is closely related to fundamental human rights issues. **Every individual, regardless of gender or geographic location, deserves equal consideration and representation in health research.** As we move forward, the focus must shift towards a more inclusive research standard, one that actively addresses these disparities, thereby ensuring that the fruits of research are accessible and beneficial to the whole of the global population.

4. Conclusions and future work

This comprehensive analysis of breast cancer clinical trials, leveraging data from ClinicalTrials.gov, has uncovered several key insights that are vital for understanding the current landscape of breast cancer research. We consider that we have met the proposed goals for this work by performing a deep analysis on the data utilizing multiple data analysis techniques, such as Exploratory Data Analysis (EDA), clustering with K-means algorithm ad Principal Component Analysis (PCA).

We observed a **steady increase in the number of reported clinical trials**, particularly after 2000, with a couple of temporary dips influenced by major global events such as the financial crisis and COVID-19 pandemic. This growth demonstrates an ongoing commitment to advancing breast cancer research. In recent years, we have observed a significant **shift in focus of breast cancer research**. Initially dominated by drug intervention type, researchers are now embracing a more varied approach, **incorporating behavioral interventions and advanced screening techniques**.

Our study also uncovered a notable **gender imbalance** in breast cancer research, emphasizing that although 1% of all breast cancer cases are diagnosed in men, they are notably underrepresented in clinical trials, accounting for only 0,3% of the studies. This discrepancy underscores the **urgency for more inclusive research practices**. Aligning with the Sustainable Development Goal 5 - Gender Equality (38), it is crucial that all genders receive equitable attention in clinical research.

Regarding age demographics, our findings show that 3% of breast cancer clinical trials include children, which is consistent with the expectations as breast cancer is very rare in small children, with only 3% of all cases diagnosed in people under 30 years old.

Geographically, **the United States leads in breast cancer research**, hosting nearly half of all trials, followed by China with 10%. However, the analysis reveals significant geographic disparities in the distribution of clinical trials. Regions like Asia, Africa, and Latin America show an underrepresentation in research activities, despite having higher cancer incidence rates. This **geographic imbalance highlights an important gap in global research efforts**, emphasizing opportunities for a more equitable distribution of clinical trials across different regions.

Finally, our analysis revealed a **notable discrepancy in trial sizes**, particularly concerning Early Phase 1 and Phase 4 studies. We found that the sizes of Early **Phase 1 trials are larger than recommended, while Phase 4 studies are considerably smaller than the FDA recommended size**. Adhering to the

recommended trial sizes set by public health institutions is crucial to minimize risks for participants and ensure the broader safety of the population. Properly sized studies are essential for balancing the need for responsible research with the objective of protecting individual participants and public health at large.

4.1 Future work

As for **future work**, multiple observations from the present study could benefit from a deeper analysis to identify opportunities for advancement in breast cancer research.

A critical area for further exploration involves **racial disparities**. For instance, black women have a 40% higher chance of mortality from a breast cancer in the United States. Unfortunately, the data extracted from ClinicalTrials.gov did not include a split on race for the participants in clinical trials, preventing us from performing this analysis. However, it would be highly beneficial to analyze the clinical trials geographical distribution by race to identify gaps that need to be covered in order to reduce inequalities, as promoted by the Sustainable Development Goal 10 – Reduced Inequalities of the 2030 Agenda (38).

Additionally, the **impact of funding sources on the design and focus of clinical trials** requires further exploration. Our dataset included a high percentage of clinical trials categorized under the funder type ‘OTHER’ that did not allow us to obtain a good understanding of the origin of the investment in research. Increased transparency in funding sources could provide insights into potential biases in clinical trials and help ensure that research objectives are aligned with the patient needs rather than commercial or political interests.

Lastly, future research efforts can be greatly benefited by **higher transparency in study and result documentation**. By making detailed results readily available on ClinicalTrials.gov, researchers can more effectively correlate these outcomes with other dataset variables, improving predictions and understanding of study outcomes as a result. Such transparency not only benefits individual studies, but also enriches the global research community. Sharing findings broadly can contribute to the improvement of global health and provide foundational knowledge for further research. This approach also helps avoid duplication of efforts, ensuring that resources are allocated efficiently on relevant research topics.

5. Bibliography

1. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin. 2023 Jan 12;73(1):17–48.

2. Nardin S, Mora E, Varughese FM, D'Avanzo F, Vachanaram AR, Rossi V, et al. Breast Cancer Survivorship, Quality of Life, and Late Toxicities. Front Oncol. 2020 Jun 16;10.

3. U.S. Food & Drug Administration. FDA. 2023 [cited 2023 Dec 18]. Expanded Access. Available from: *https://www.fda.gov/news-events/public-health-focus/expanded-access*

4. National Institutes of Health. https://www.nih.gov/health-information/nih-clinical-research-trials-you/basics. 2023. NIH Clinical Research Trials and You. The Basics.

5. Loud JT, Murphy J. Cancer Screening and Early Detection in the 21 st Century. Semin Oncol Nurs. 2017 May;33(2):121–8.

6. Potosky AL. The Role of Increasing Detection in the Rising Incidence of Prostate Cancer. JAMA: The Journal of the American Medical Association. 1995 Feb 15;273(7):548.

7. International Agency for Research on Cancer. Global Cancer Observatory. 2020 [cited 2023 Dec 13]. Globocan 2020. Available from: *https://gco.iarc.fr/today/home*

8. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. JNCI Journal of the National Cancer Institute. 1989 Dec 20;81(24):1879–86.

9. KEATING P, CAMBROSIO A. Cancer Clinical Trials: The Emergence and Development of a New Style of Practice. Bull Hist Med [Internet]. 2007;81(1):197–223. Available from: *http://www.jstor.org/stable/44451744*

10. Donegan WL. History of breast cancer. In: Breast cancer. 2006. p. 1–14.

11. Yarnold J. NHS England. 2023. Game changers in breast cancer treatment.

12. Crago AM, Azu M, Tierney S, Morrow M. Randomized Clinical Trials in Breast Cancer. Surg Oncol Clin N Am. 2010 Jan;19(1):33–58.

13. Hong R, Xu B. Breast cancer: an up-to-date review and future perspectives. Cancer Commun. 2022 Oct 8;42(10):913–36.

14. Welsh J. VeryWell Health. 2021 [cited 2024 Jan 2]. What Is the History of Breast Cancer? Available from: *https://www.verywellhealth.com/history-of-breast-cancer-5207255*

15. Unger JM, Vaidya R, Hershman DL, Minasian LM, Fleury ME. Systematic Review and Meta-Analysis of the Magnitude of Structural, Clinical, and Physician and Patient Barriers to Cancer Clinical Trial Participation. JNCI: Journal of the National Cancer Institute. 2019 Mar 1;111(3):245–55.

16. Cuzick J, Howell A, Forbes J. Early stopping of clinical trials. Breast Cancer Research. 2005 Oct 21;7(5):181.

17. Bea VJ, Taiwo E, Balogun OD, Newman LA. Clinical Trials and Breast Cancer Disparities. Curr Breast Cancer Rep. 2021 Sep 30;13(3):186–96.

18. Rathod A, Murphy CC, Rahimi A, Pruitt SL. Revisiting Exclusion of Prior Cancer in Clinical Trials of Male Breast Cancer. J Cancer. 2023;14(5):737–40.

19. Ndumele A, Park KU. The Impact of COVID-19 on National Clinical Trials Network Breast Cancer Trials. Curr Breast Cancer Rep. 2021 Sep 12;13(3):103–9

20. Zhang Z. Predictive analytics in the era of big data: opportunities and challenges. Ann Transl Med. 2020 Feb;8(4):68–68.

21. Vangipurapu M. Clinion. 2022 [cited 2023 Nov 15]. AI and Automation in Clinical Trials. Available from: *https://www.clinion.com/insight/ai-and-automation-in-clinical-trials/*

22. Zippel C, Bohnet-Joschko S. Rise of Clinical Studies in the Field of Machine Learning: A Review of Data Registered in ClinicalTrials.gov. Int J Environ Res Public Health. 2021 May 11;18(10):5072.

23. de Glas NA, Hamaker ME, Kiderlen M, de Craen AJM, Mooijaart SP, van de Velde CJH, et al. Choosing relevant endpoints for older breast cancer patients in clinical trials: an overview of all current clinical trials on breast cancer treatment. Breast Cancer Res Treat. 2014 Aug 9;146(3):591–7.

24. Shepshelovich D, Goldvaser H, Wang L, Abdul Razak AR, Bedard PL. Comparison of reporting phase I trial results in ClinicalTrials.gov and matched publications. Invest New Drugs. 2017 Dec 14;35(6):827–33.

25. Hirsch BR, Califf RM, Cheng SK, Tasneem A, Horton J, Chiswell K, et al. Characteristics of Oncology Clinical Trials. JAMA Intern Med. 2013 Jun 10;173(11):972.

26. Gresham G, Meinert JL, Gresham AG, Piantadosi S, Meinert CL. Update on the clinical trial landscape: analysis of ClinicalTrials.gov registration data, 2000–2020. Trials [Internet]. 2022 Oct 6 [cited 2023 Dec 10];23(1):858. Available from: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9540299/*

27. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, et al. CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc. 2000;9(13):1–73.

28. Mai PL, Miller A, Gail MH, Skates S, Lu K, Sherman ME, et al. Risk-Reducing Salpingo-Oophorectomy and Breast Cancer Risk Reduction in the Gynecologic Oncology Group Protocol-0199 (GOG-0199). JNCI Cancer Spectr. 2020 Feb 1;4(1).

29. Jeong S, Sohn M, Kim JH, Ko M, Seo H won, Song YK, et al. Current globalization of drug interventional clinical trials: characteristics and associated factors, 2011–2013. Trials. 2017 Dec 21;18(1):288.

30. CTG labs - NCBI. ClinicalTrials.gov. 2023 [cited 2023 Dec 17]. About ClinicalTrials.gov. Available from: *https://clinicaltrials.gov/about-site/about-ctg*

31. Woodcock J. U.S. Food and Drug Administration. 2023. FDA takes action for failure to submit required clinical trial results information to Clinicaltrials.gov.

32. U.S. Food and Drug Administration. FDA. 2023 [cited 2023 Dec 13]. Step 3: Clinical Research. Available from: *https://www.fda.gov/patients/drug-development-process/step-3-clinical-research*

33. Zhang X, Zhang Y, Ye X, Guo X, Zhang T, He J. Overview of phase IV clinical trials for postmarket drug safety surveillance: a status report from the ClinicalTrials.gov registry. BMJ Open. 2016 Nov 23;6(11):e010643.

34. Kummar S, Rubinstein L, Kinders R, Parchment RE, Gutierrez ME, Murgo AJ, et al. Phase 0 Clinical Trials: Conceptions and Misconceptions. The Cancer Journal. 2008 May;14(3):133–7.

35. International Monetary Fund. IMF. 2023. GDP per capita, current prices.

36. Novartis. Novartis Website. 2022 [cited 2023 Dec 15]. Compassionate use: Providing access to much needed treatments. Available from: *https://www.novartis.com/stories/compassionate-use-providing-access-much-needed-treatments*

37. American Cancer Society. Breast Cancer Facts & Figures 2022-2024 [Internet]. Atlanta; 2022 [cited 2023 Dec 17]. Available from: *https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/2022-2024-breast-cancer-fact-figures-acs.pdf*

38. United Nations. United Nations Website. 2023 [cited 2024 Jan 2]. The 17 goals. Available from: *https://sdgs.un.org/goals*

6. Appendices

Appendix A: Python Code

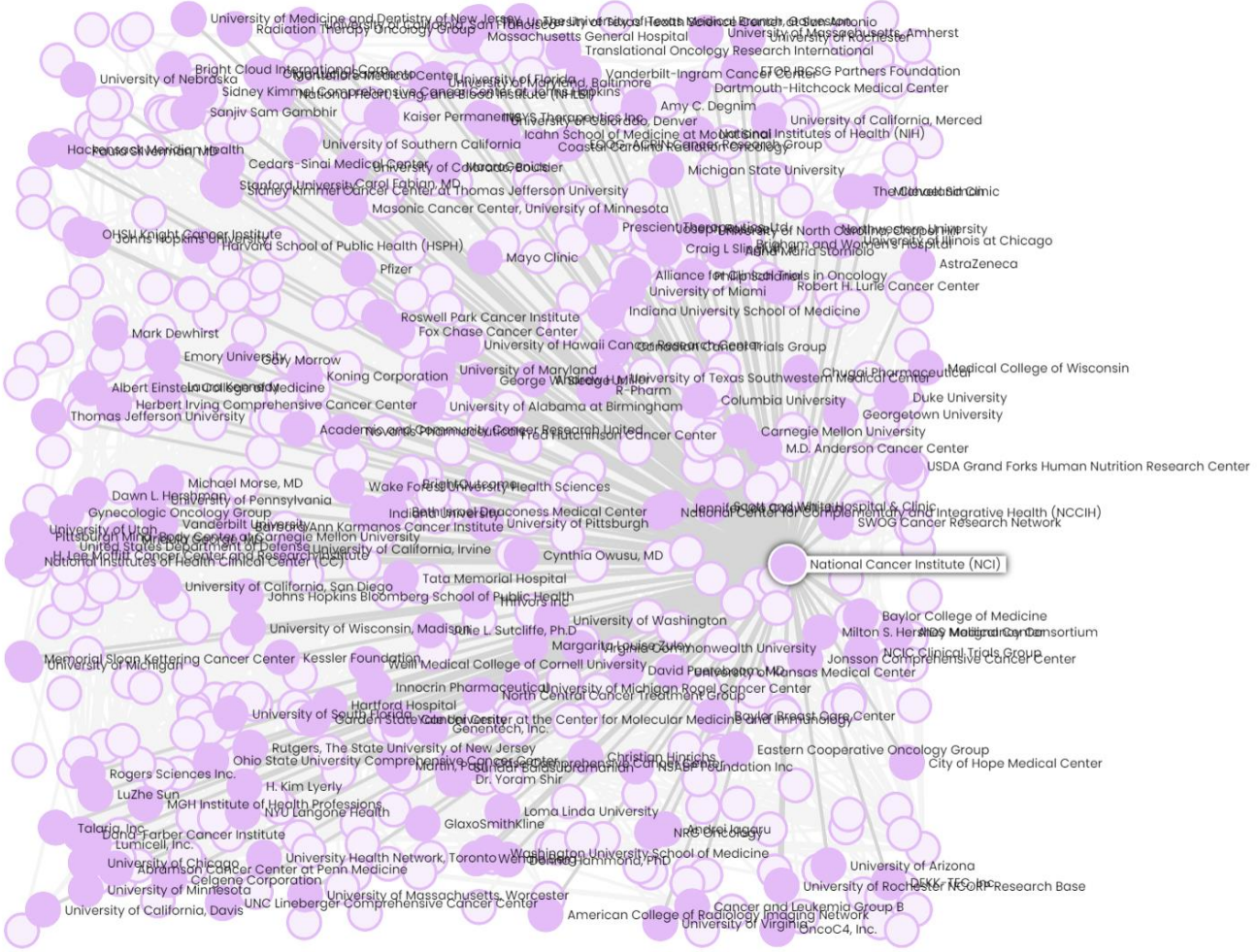
The Python code for this project is available in the following Google Drive location, with access permissions for all the UOC community:

https://drive.google.com/drive/folders/1ZnyAPXNWjWtMjCszDz7rsFmAZSvRKx8M?usp=sharing

Appendix B: Intervention type distribution chart for top 20 sponsors

INTERVENTION	BEHAVIORAL	BIOLOGICAL	COMBINATION_PRODUCT	DEVICE	DIAGNOSTIC_TEST	DIETARY_SUPPLEMENT	DRUG	GENETIC	OTHER	PROCEDURE	RADIATION
SPONSOR											
Abramson Cancer Center at Penn Medicine	11.29	6.45	0.00	3.23	3.23	0.00	24.19	0.00	32.26	8.06	11.29
Alliance for Clinical Trials in Oncology	1.04	4.17	0.00	1.04	0.00	7.29	47.92	4.17	18.75	13.54	2.08
AstraZeneca	0.00	0.70	0.00	0.00	0.00	0.00	94.37	0.70	3.52	0.70	0.00
Case Comprehensive Cancer Center	7.14	7.14	0.00	10.00	1.43	1.43	25.71	2.86	24.29	14.29	5.71
City of Hope Medical Center	0.00	7.69	0.00	0.00	0.00	3.85	33.33	0.00	37.18	16.67	1.28
Dana-Farber Cancer Institute	14.89	2.13	1.42	7.80	0.00	0.00	50.35	0.00	13.48	6.38	3.55
Duke University	41.94	9.68	0.00	8.06	0.00	0.00	17.74	1.61	9.68	4.84	6.45
Eli Lilly and Company	0.00	6.02	0.00	0.00	0.00	0.00	93.98	0.00	0.00	0.00	0.00
Fudan University	0.00	0.61	1.23	2.45	1.23	0.00	74.85	0.61	4.29	6.13	8.59
Hoffmann-La Roche	0.00	0.00	0.00	0.60	0.00	0.00	94.61	0.00	4.79	0.00	0.00
M.D. Anderson Cancer Center	16.40	2.40	0.00	0.40	0.80	0.00	38.00	0.00	19.60	19.60	2.80
Mayo Clinic	4.51	3.76	0.00	10.53	3.76	1.50	18.80	3.01	22.56	26.32	5.26
Memorial Sloan Kettering Cancer Center	18.97	3.88	0.43	5.17	5.17	0.43	26.29	0.86	20.69	10.34	7.76
National Cancer Institute (NCI)	1.11	14.07	0.00	0.00	0.00	0.37	56.67	0.37	13.70	12.96	0.74
Northwestern University	17.65	8.82	0.00	2.94	0.00	0.00	44.12	0.00	13.24	11.76	1.47
Novartis Pharmaceuticals	0.00	1.72	0.57	0.00	0.00	0.00	91.38	0.00	5.17	0.00	1.15
Pfizer	0.00	1.47	0.74	0.74	0.00	0.00	92.65	0.00	3.68	0.74	0.00
Stanford University	26.23	0.00	0.00	9.84	0.00	0.00	26.23	0.00	8.20	27.87	1.64
University Health Network, Toronto	20.93	4.65	0.00	4.65	0.00	0.00	18.60	2.33	27.91	13.95	6.98
Washington University School of Medicine	2.78	5.56	0.00	11.11	0.00	1.39	40.28	0.00	16.67	13.89	8.33

Appendix C – National Cancer Institute (NCI) Network Relationships



Appendix D – AstraZeneca Network Relationships