The current state of clinical trials amongst major diseases

The Worldwide Landscape of Clinical Trials for Major Diseases

The Group details.

The second group participating in Challenge 2 is known as "Group 2." This team is led by John Kamau, and its members include Lynn Ajema and Cosmas Kibett. The team consists of colleagues from L-IFT, a non-governmental organization based in the Netherlands, specializing in the study of low-income individuals and their decision-making processes. Pascal Amisi serves as the team's mentor, providing guidance on report structure, expectations, and any additional information they require.

The group's contributions resulted in the creation of this report, along with a straightforward dashboard developed using Python Dash. The code for the dashboard has been added to a GitHub repository for reference and sharing.

Table 1: Group, members, mentor, GitHub repository and dashboard link.

Data	Details
Group	2
Members	1.John Kamau 2.Lynn Ajema 3. Cosmas Kibet
Mentor	Pascal Amisi
GitHub	https://github.com/sophicist/DTE-Datathon
Dashboard	https://dashboarddte-42fd5a84086f.herokuapp.com/
Dashboard Repository	https://github.com/liftransform/DTE-dashboard

Introduction

This analysis delves into a clinical trials dataset accessible via the provided link. The overarching theme of this research is to foster a data-driven, innovative, and collaborative environment that encourages creative problem-solving for societal issues, ultimately yielding actionable results.

The clinical trials dataset, available at https://clinicaltrials.gov/, offers comprehensive information regarding ongoing, upcoming, and completed clinical research studies. These studies encompass a global scope, spanning all 50 states and more than 200 countries. We specifically retrieved data related to clinical studies involving Cancer, malaria, Covid-19, HIV, Heart Conditions, and pneumonia, as these health conditions represent some of the foremost causes of mortality in Kenya.

This analysis poses several key questions:

a. How can the data be most effectively analyzed to extract meaningful insights?
b. Are there discernible trends within the realm of clinical studies that merit identification and exploration?
c. What conclusions and recommendations can be drawn from the findings generated through this analysis?

Clinical Trials and Observational studies

A clinical study is a form of research involving human volunteers, often referred to as participants, with the aim of contributing to medical knowledge. There are two primary categories of clinical studies: clinical trials, also known as interventional studies, and observational studies. ClinicalTrials.gov encompasses both interventional and observational studies. In a clinical trial, participants undergo specific interventions as outlined in a research plan or protocol developed by the investigators. These interventions may encompass medical products like drugs or devices, medical procedures, or modifications to participants' behavior, such as dietary changes.

Conversely, in an observational study, researchers evaluate the health outcomes of participant groups based on a predetermined research plan or protocol. While participants in observational studies may receive interventions like medical products or procedures as part of their regular medical care, these interventions are not assigned to specific individuals by the investigator, as is the case in a clinical trial.

The data

The dataset encompasses six distinct studies, each featuring an equal number of attributes but varying in the quantity of data rows. Notably, the Cancer studies comprise the most substantial portion of the dataset, followed by the Heart studies, with Covid studies and HIV studies closely following. Pneumonia studies are represented with a somewhat smaller volume of data, while Malaria studies have the fewest entries.

To facilitate comprehensive analysis in this paper, all these datasets were harmoniously merged into a single, consolidated dataset. As part of this integration, an additional column was introduced to classify and track the specific type of study within this unified dataset.

Table 2: Consolidated dataset summary for each disease studied.

Dataset	Rows	Columns
Cancer - studies	100262	30
Heart - studies	32176	30
Pneumonia- studies	9640	30
Covid 19 - studies	9342	30
HIV - studies	9113	30
Malaria- studies	1330	30
Final Dataset	161863	31

The visual representation of the distribution of studies is depicted in the graph presented below:

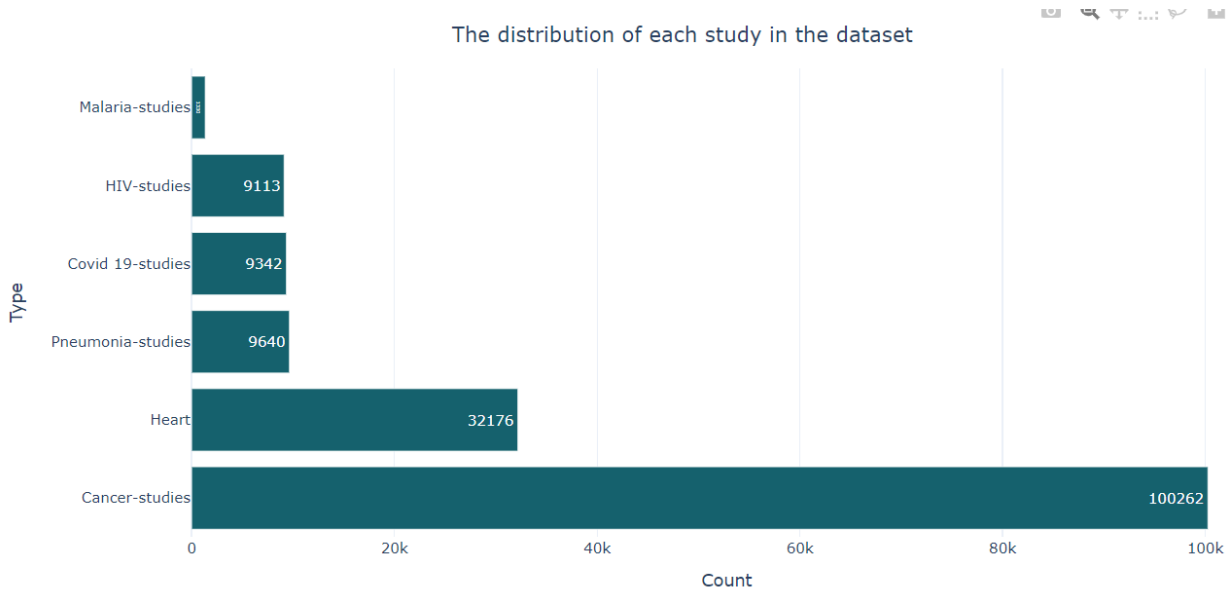

Figure 1: Disease type and the associated count.

Data cleaning and transformations

We employed distinct data cleansing techniques tailored to the specific column types. Textual data underwent processes such as tokenization, stop word removal, bigram extraction, and conversion to lowercase. Numeric data was subjected to outlier removal to enhance the clarity of visualizations, particularly for boxplots. The IQR method was applied for this purpose. Additionally, any missing data was eliminated when applicable, although this was performed selectively at the column level to support ongoing analysis rather than across the entire dataset.

The Analysis of the data

For the analysis we had several research questions that we used the graphs and other types of analytics to answer. The general theme was - what is the current state of clinical trials world wide and we used the graphs ton answer these questions as well as other analysis included in this study.

Demographics

The studies predominantly encompass both male and female participants. In the case of HIV studies, approximately 20% of the studies are specific to either men or women, a similar distribution is observed in Cancer studies with about 15% concentrating solely on females and approximately 6% exclusively on males. Notably, Malaria studies exhibit an interesting pattern, with around 9% emphasizing women and only about 3% focusing on men.

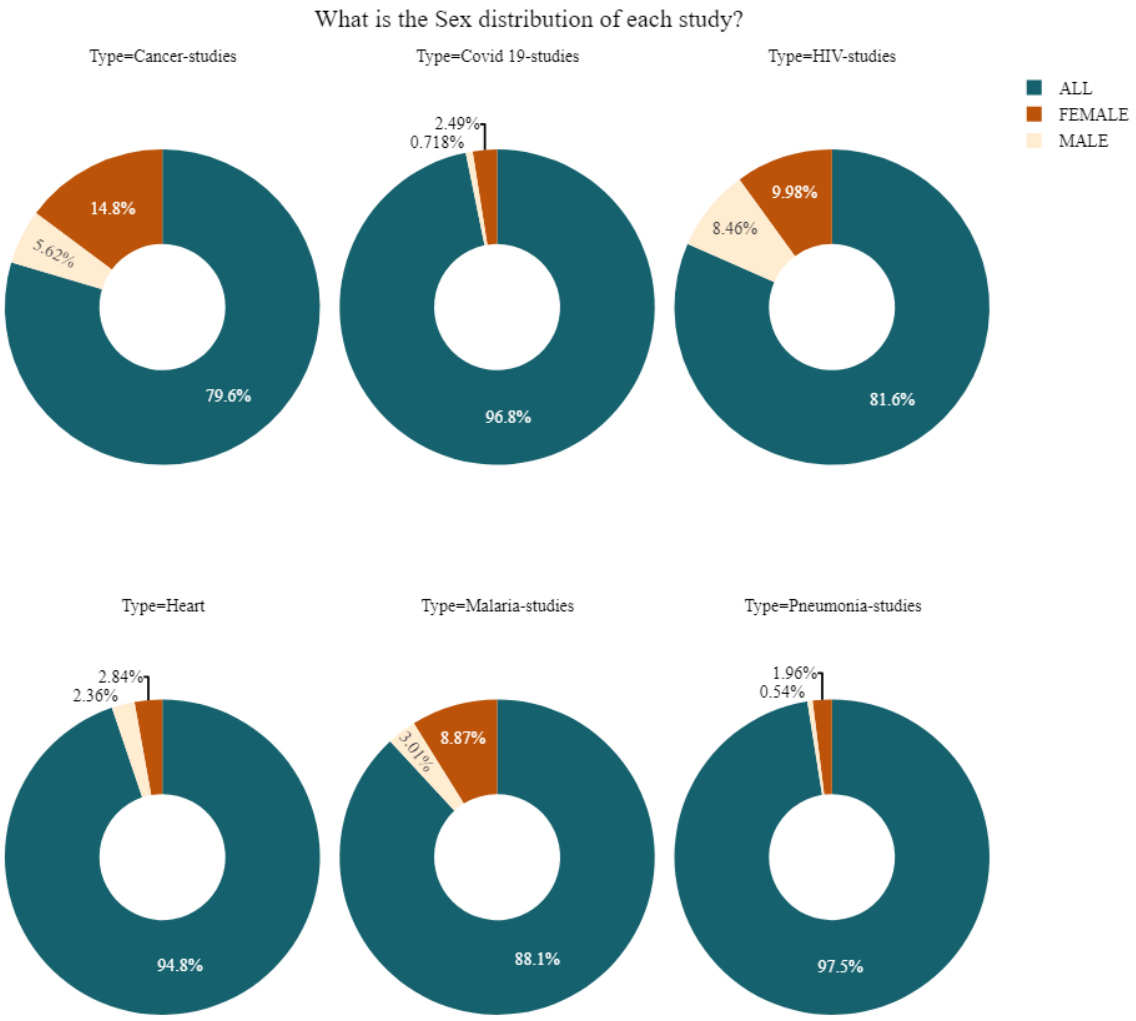

Figure 2: Demographic summary for each disease type.

Regarding age groups, the majority of studies, except for Malaria studies, include over 60% adult participants. Malaria studies, in contrast, allocate about one-third of their focus to child participants, another one-third to adults, and the remaining one-third to various age groups. HIV studies also allocate a significant portion to children, accounting for approximately 17% of the total.

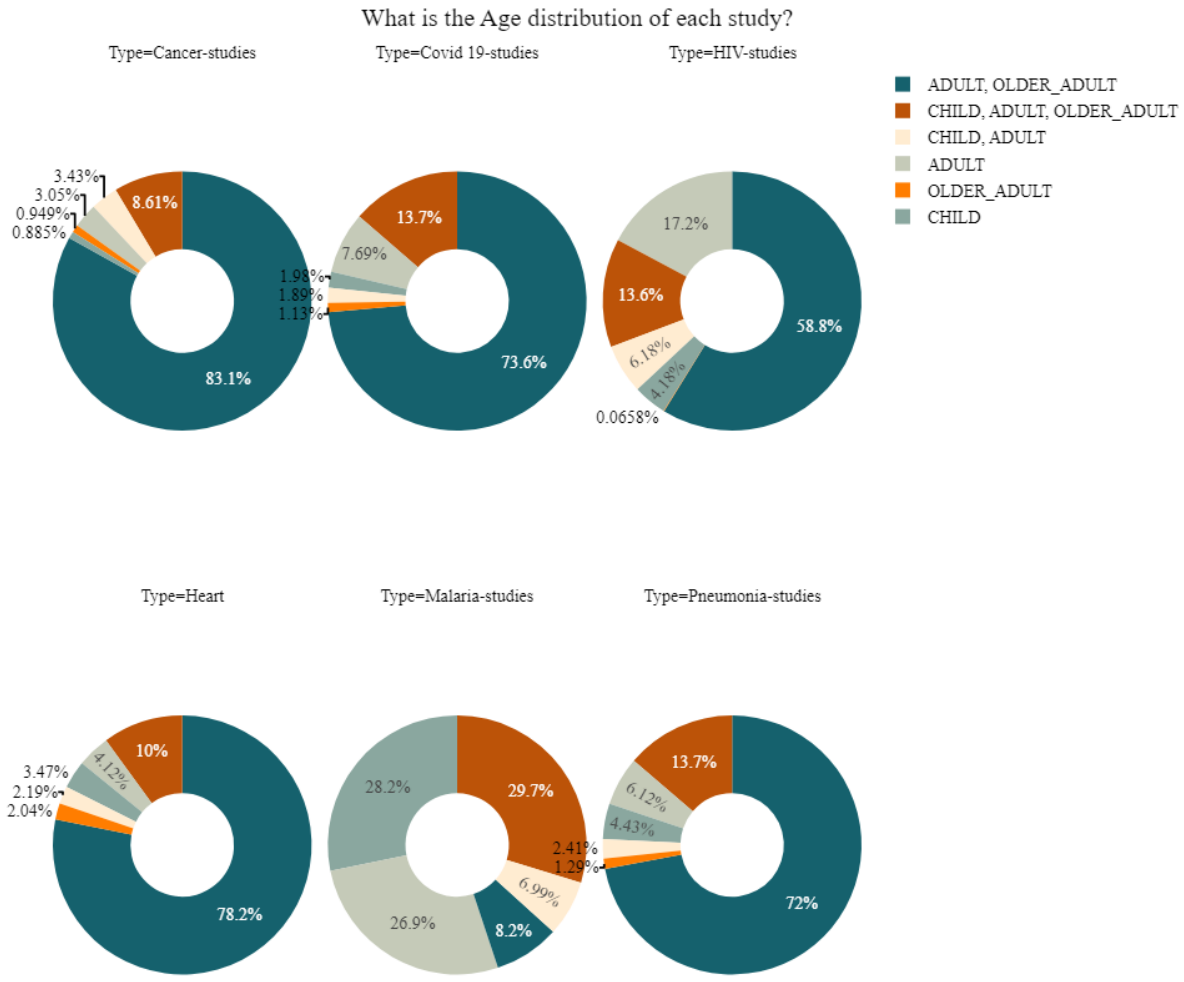

Figure 3: Age distribution for each disease.

In terms of study phases, most studies span from phase 1 to phase 4, with a relatively balanced distribution, particularly evident in HIV and Covid-19 studies. Cancer studies primarily encompass three phases, with phase 4 being the most prevalent. There are a few studies that incorporate a mix of phases, although these constitute a smaller portion of the overall study landscape.

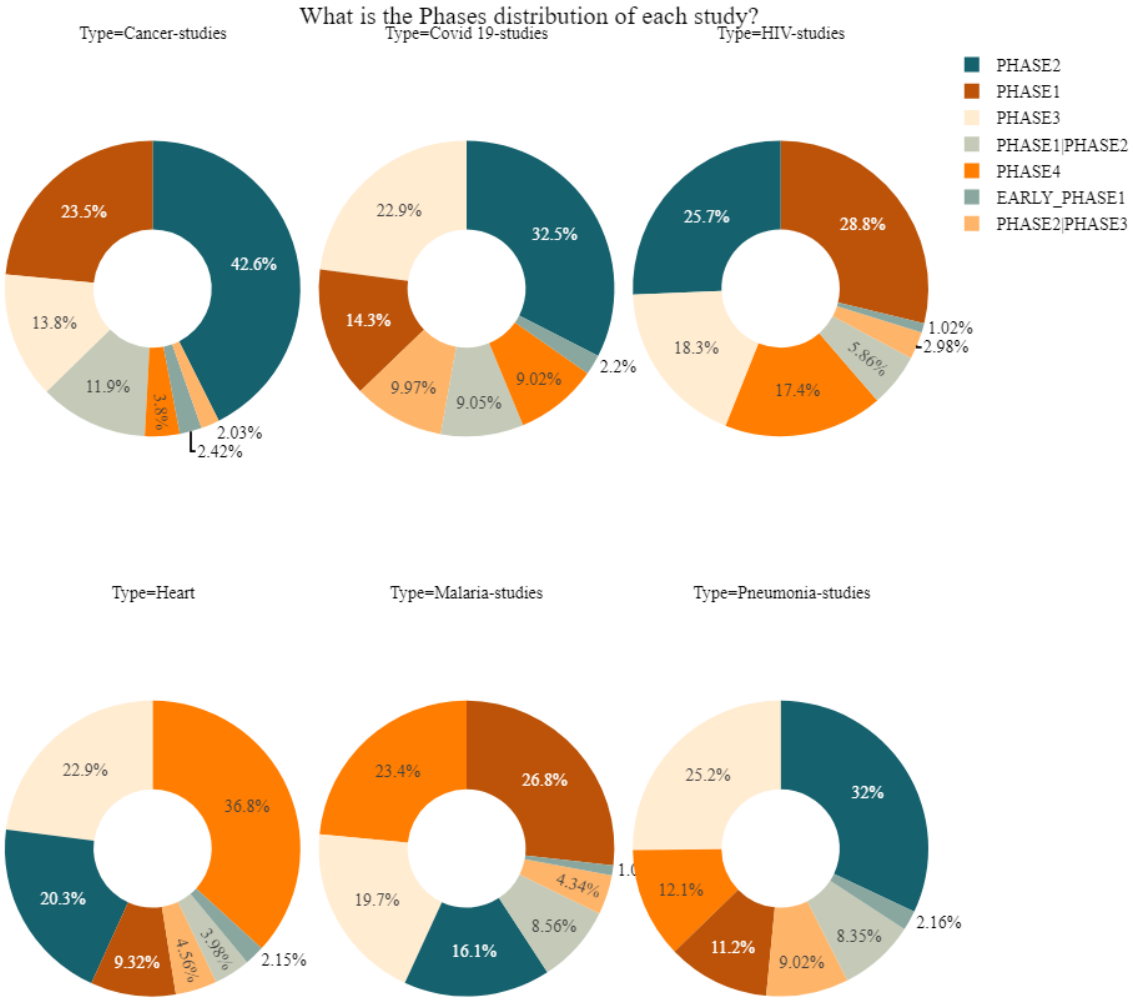

Figure 4: Disease study phases.

Two primary study types characterize the nature of clinical research: interventional studies, also known as clinical trials, and observational studies, which may encompass patient registries and expanded access programs.

Across the various domains, interventional studies dominated, constituting a significant portion of the research landscape. Specifically, Malaria studies featured 83.7% interventional studies, HIV studies had 81% interventional studies, and Cancer studies comprised 80.2% interventional studies. In comparison, Covid-19 studies demonstrated a lower proportion at 57%, while Heart and Pneumonia studies had 63% and 59.9% interventional studies, respectively.

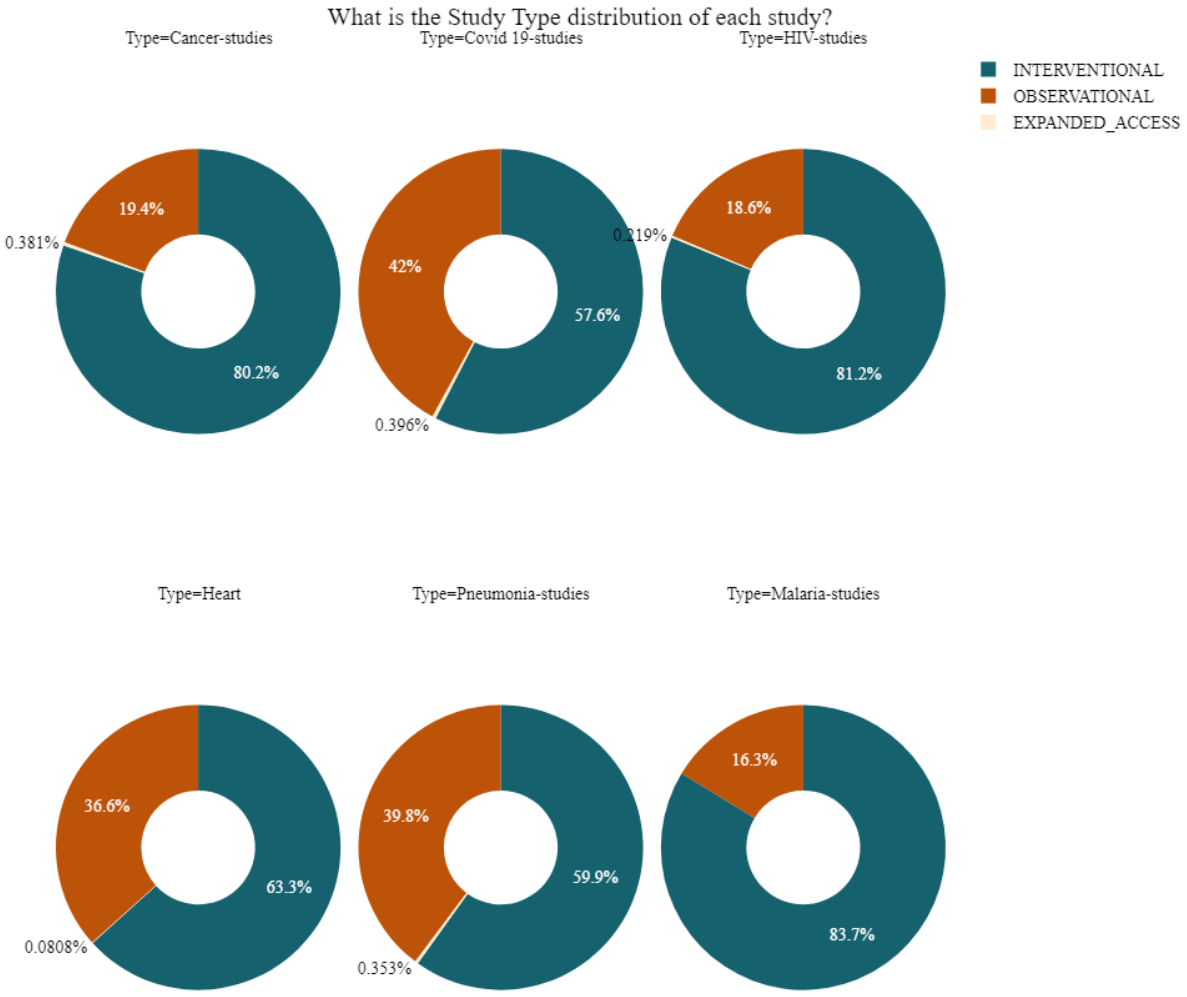

Figure 5: Research study designs.

Conversely, the largest share of observational studies was observed in the context of Covid-19 and Pneumonia research.

Does the study have results and what's its status?

A small proportion of the studies yielded results, with the highest percentages observed in Cancer studies and HIV studies at around 14-15%, followed by Malaria studies at 11.4%. Most other studies had results in less than 10% of cases.

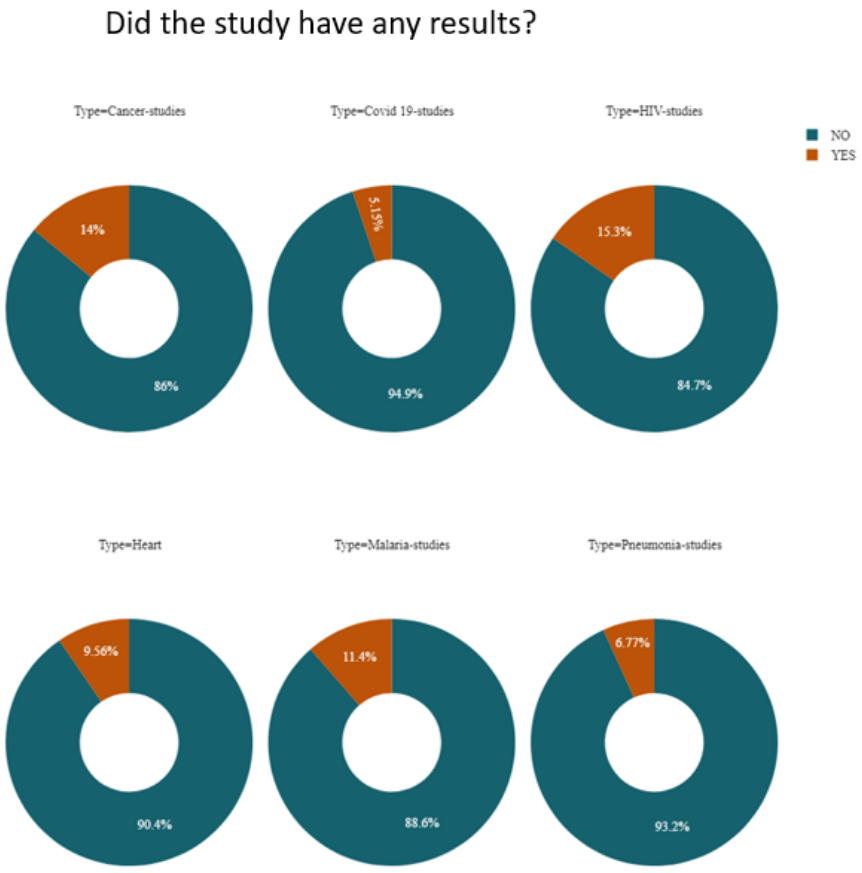

Figure 6: Study results and their status.

Although the results status was very low in comparison. Most of the studies across the studies are already completed so the reason of not having any results is not because they are ongoing.

Overall, between 43% to 71% of the studies are completed across the studies with about 7% - 18% having unknown status. About 15% of the studies are currently recruiting.

In terms of duration, Covid 19 took on average the lowest number of months on average to complete maybe due to international pressure to complete these studies as this was pandemic at 10.87 months to complete the studies on average. Pneumonia studies that were closely related came close at only about a year on average to complete the studies. HIV and Cancer had the highest number of months to complete at well over 3 years on average to complete the studies with Heart and Mararia taking 1.5 – 2 years on average to complete.

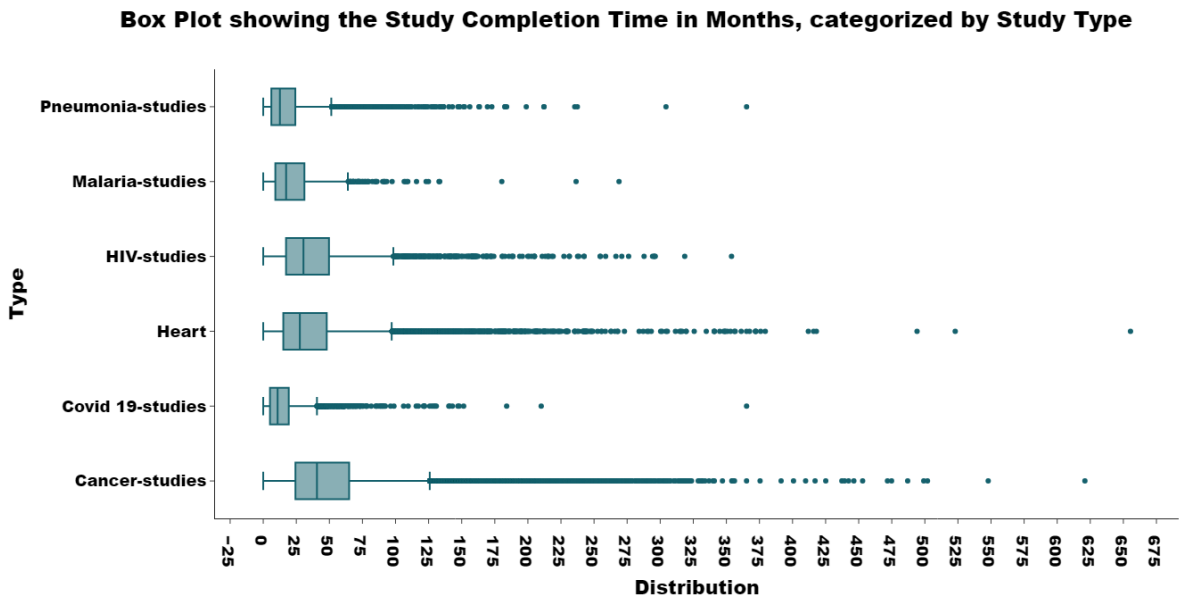

Figure 7: Box plot showing the relationship between disease type and their distribution.

What conditions are being studied?

In this analysis, we examined the top 10 medical conditions currently under study. The bar graph displayed below illustrates that a significant portion of these studies are related to various aspects of Cancer. Among these conditions, Breast cancer stands out as the most extensively researched, with nearly double the number of studies compared to the second most common cancer, Prostate cancer, which is primarily prevalent in men. Additionally, Lung cancer and Colorectal cancer are also subject to extensive research.

Beyond the realm of Cancer research, we find studies dedicated to other critical health issues such as HIV infection and COVID-19. In the field of cardiac health, Heart failure and Coronary Artery disease are prominent subjects of study.

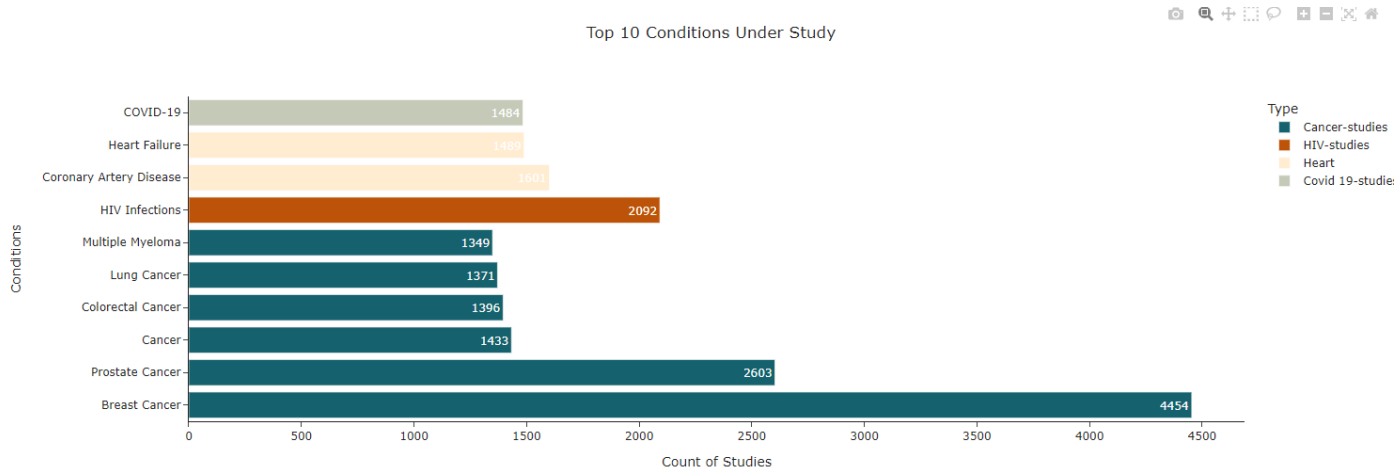

Figure 8: The count of top ten conditions for each disease.

What are the primary Outcome measures of these studies?

The dataset provides concise explanations of the main research findings. In our analysis, we utilized bigrams to identify the most frequently mentioned words associated with the primary research outcomes.

In the case of Cancer studies, the primary outcome measurement primarily revolves around the investigation of adverse effects, response rates, and complete responses. For Covid-19 studies, while adverse effects are a primary consideration, there are also references to specific timeframes such as "28 days," "day 1," and "seven days." In the context of Heart studies, the primary focus is on heart failure as the outcome measure, with mentions of Myocardial infarction, along with time intervals like "6 months," "12 months," and "30 days." In HIV studies, adverse effects remain significant, but viral load is also a primary measure, with references to various time intervals. Similarly, Pneumonia studies highlight adverse effects as a primary outcome measure.

It's noteworthy that all studies emphasize the assessment of adverse effects, but each study also places distinct emphasis on its own specific primary outcome measurement.

Figure 9: The bigram for the response time for each condition.

Who are the study sponsors?

In general, each category of study exhibited its own unique set of sponsors, although there were some sponsors that were common across different study domains. This analysis primarily emphasizes the significant sponsors, disregarding those with minimal involvement. For Cancer studies, the primary sponsors were the National Cancer Institute and the M.D. Anderson Cancer Center. In HIV studies, Healthcare and the National Institute of Allergy and Infectious Diseases

(NIAID) played prominent sponsorship roles. In the context of Heart studies, the leading sponsors included the National Heart, Lung, and Blood Institute (NHLBI) and Abott Medical Devices. Notably, many sponsors for Pneumonia studies were also associated with Covid-19 studies.

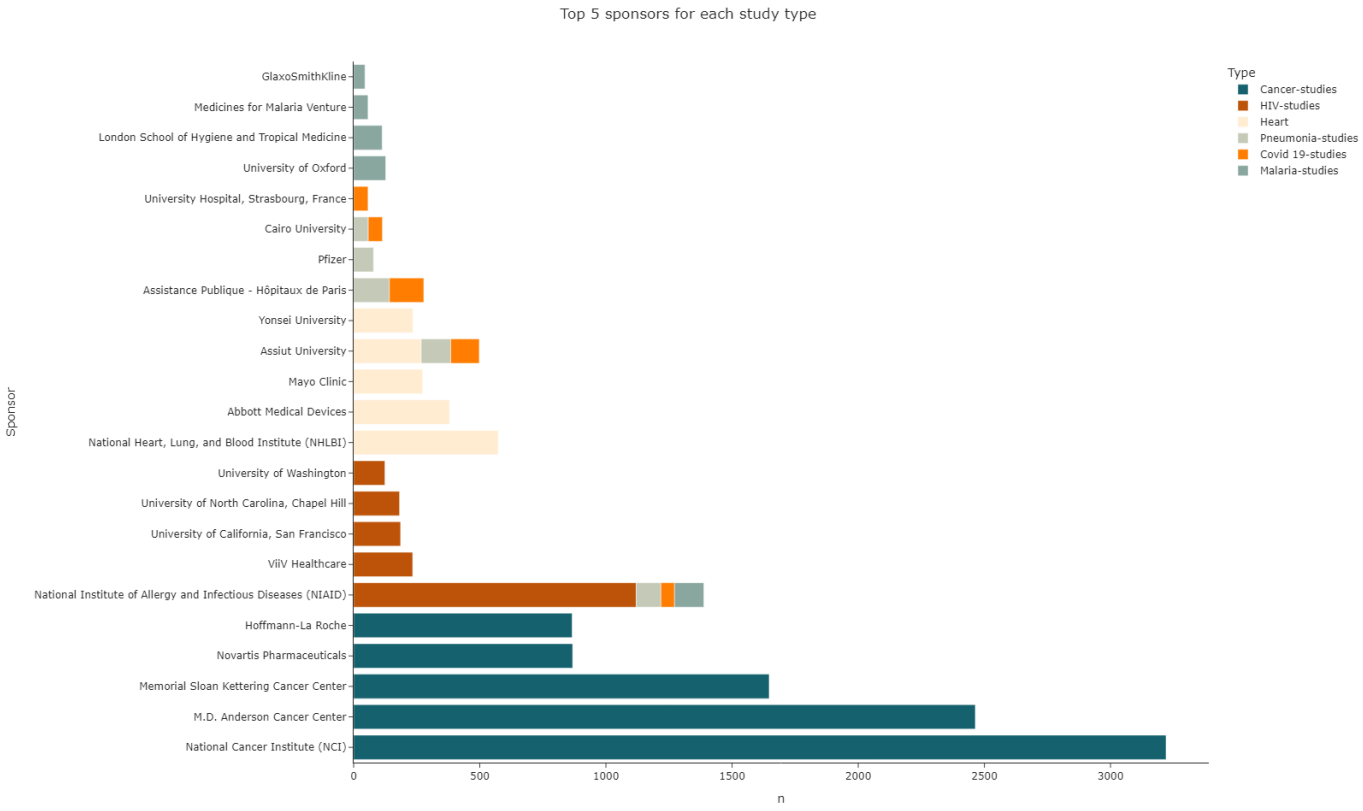

Figure 10: Top ten sponsors for the conditions.

Malaria studies received relatively limited sponsorship, primarily from the University of Oxford and the London School of Hygiene and Tropical Medicine. The NIAID emerged as the most prominent sponsor across various study areas, given its focus on infectious diseases. The only exceptions were heart diseases and non-infectious Cancer studies, which did not fall under NIAID sponsorship.

Who are the main collaborators?

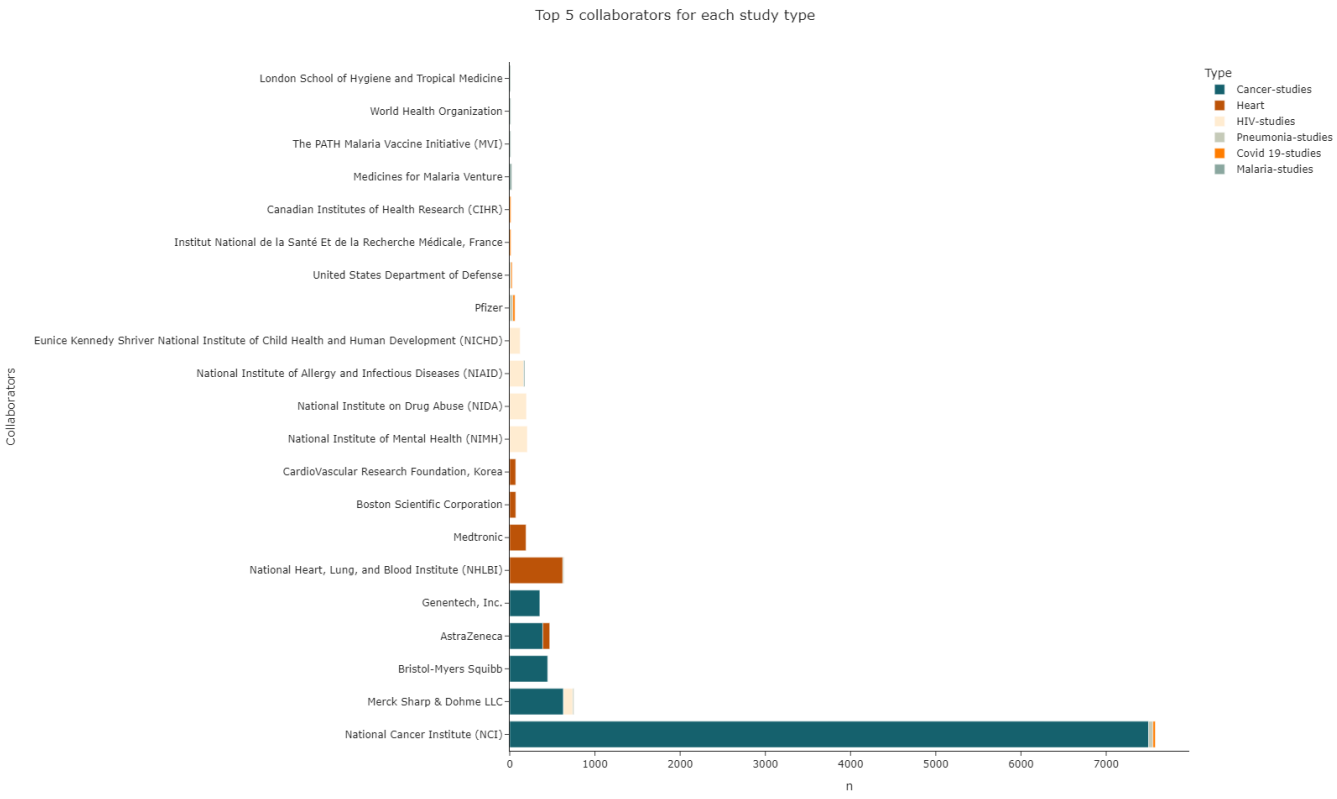

Figure 11: Collaborators for the named condition.

In line with our examination of sponsors, we also depicted the roles of collaborators in these studies. A collaborator is an organization, distinct from the sponsor, that contributes to the support of a clinical study. This support can encompass various aspects such as funding, study design, implementation, data analysis, or reporting.

A noteworthy example of dual roles can be seen in Cancer studies, where the National Cancer Institute (NCI) assumed both sponsorship and collaboration responsibilities. Within Heart studies, the primary collaborator emerged as the National Heart, Lung, and Blood Institute (NHLBI), closely followed by Medtronic. For HIV studies, frequent collaborators included the National Institute of Mental Health and the National Institute of Drug Abuse, with their involvement evident in a significant number of HIV-related studies. In contrast, Pfizer participated in collaborations, primarily in a limited number of studies, particularly notable in the fields of Covid-19 and Pneumonia research.

What was the enrollment?

Enrollment refers to the count of participants in a clinical study, and the "estimated" enrollment denotes the intended number of participants required for the research. In our analysis, the enrollment data exhibited outliers that could potentially skew the data's interpretation. To address this issue, we applied the Interquartile Range (IQR) method, removing any data falling beyond the lower and upper bounds defined by the IQR.

The median number of participants enrolled in the studies typically fell within the range of 49 to 168. The study with the lowest enrollment was Cancer, registering 49 participants. For Covid, Heart, and HIV studies, the enrollment figures were 100, 80, and 80, respectively. In contrast, Malaria and Pneumonia studies had higher enrollments, surpassing 100 participants, with 168 and 102 participants, respectively.

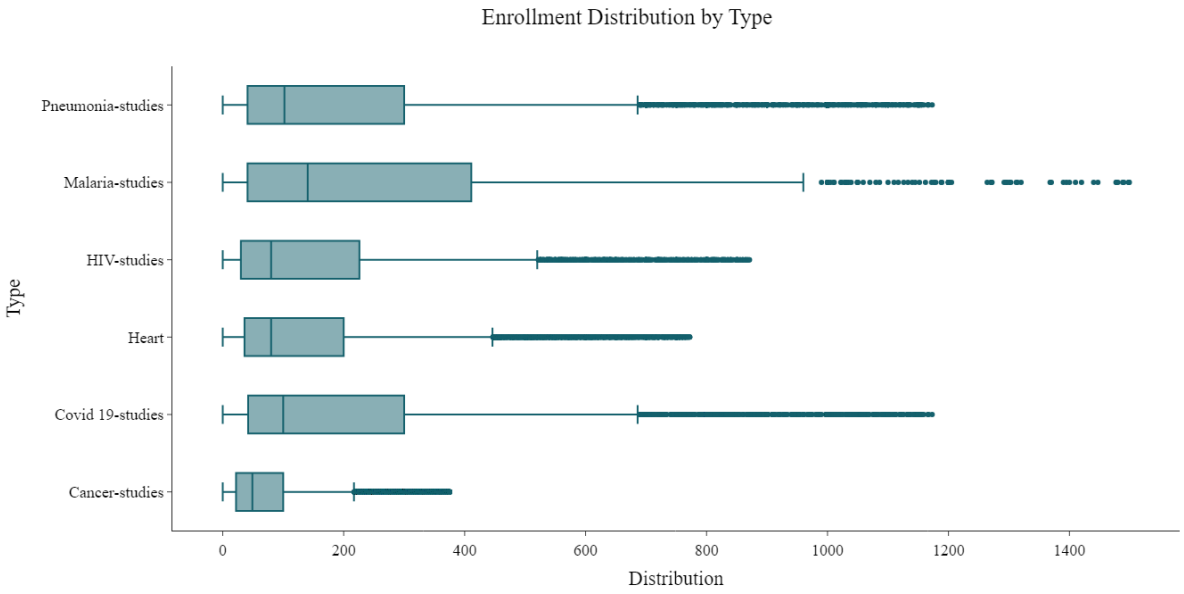

Figure 12: Boxplot showing the enrolment distribution for each disease type.

Notably, Malaria studies presented numerous outliers, with some studies having participant counts reaching as high as 2500. These extreme outliers were excluded from the analysis.

At what point in time did the studies commence

Around 1995, there was a notable surge in the number of studies, with the peak occurring in the 2010s, particularly in the case of cancer studies. Heart studies displayed a similar but less pronounced trend compared to cancer studies. HIV studies have maintained a relatively stable level over an extended period, especially during the 2000s. Covid-19 studies experienced a peak in 2020, along with pneumonia studies. Cancer studies, on the other hand, have consistently maintained a high level of research activity throughout the entire timeframe. In contrast, malaria studies exhibited the smallest number of studies, which have consistently remained at relatively lower levels over the given time.

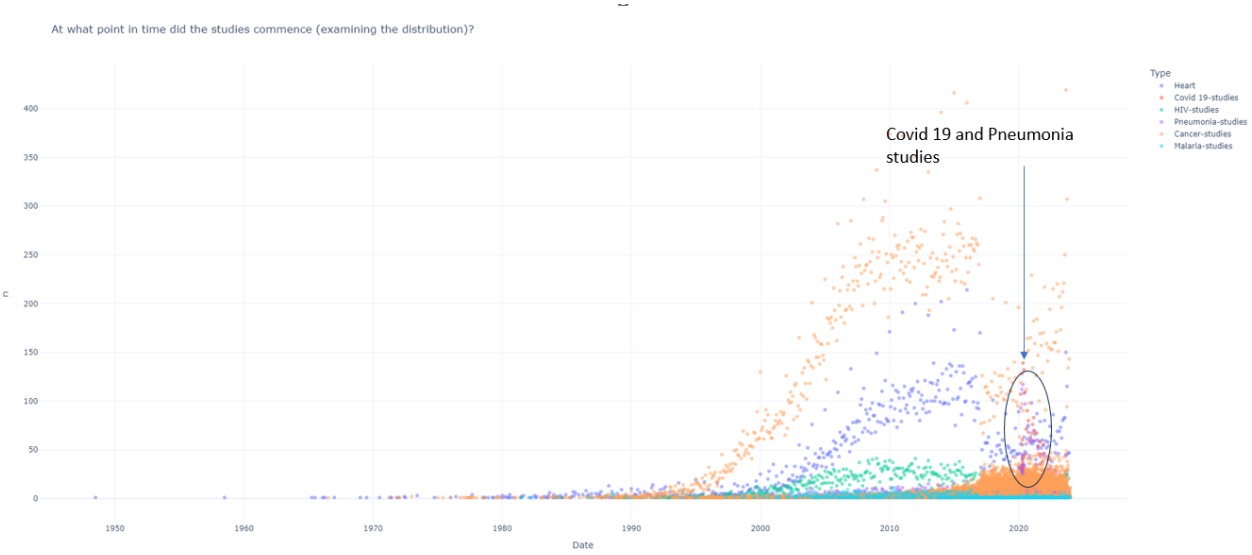

Figure 13: Year of commencement of the disease.

Where are these studies taking place?

The majority of cancer studies primarily take place in the United States, with substantial research efforts also happening in Europe. For Covid-19 studies, research is widespread, covering multiple regions around the world, including superpowers, although there are smaller-scale studies in Africa. In the context of HIV research, East and Central Africa play a significant role, making substantial contributions to the field.

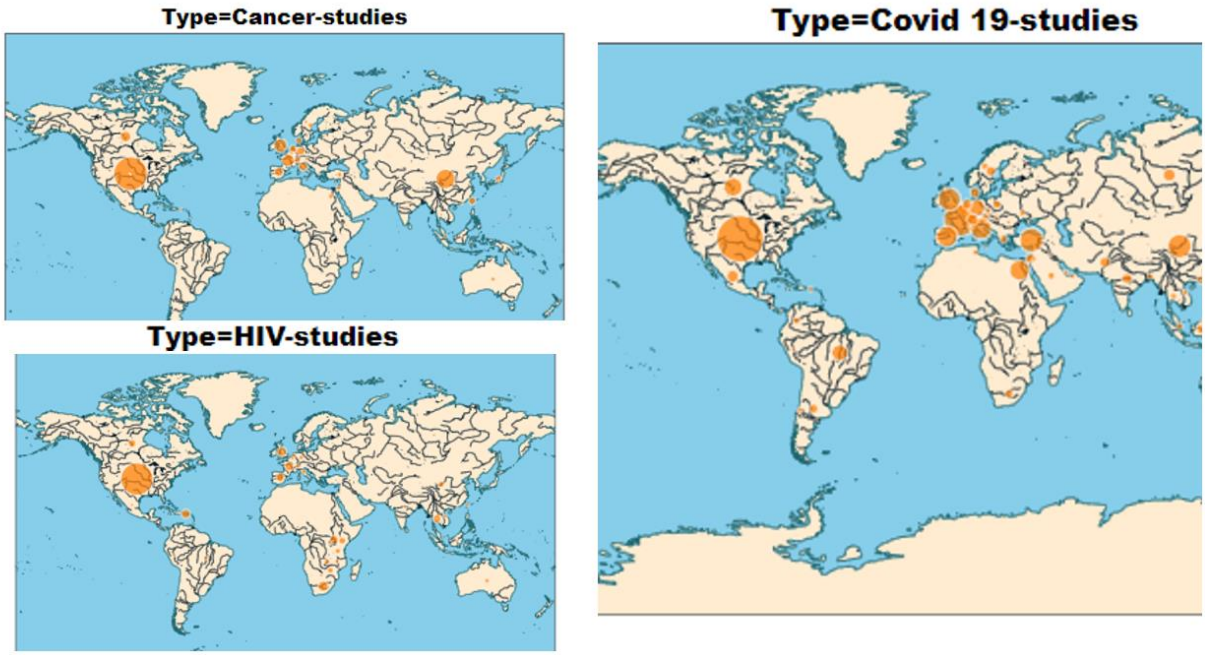

Figure 14: Cancer and HIV continental distribution.

Africa has a particularly notable presence in the field of malaria research, which is crucial given the significant number of people affected by the disease on the continent. This emphasis is particularly evident in East and West Africa, alongside research activities in Europe. The geographic focus coincides with the tropical regions where malaria is most prevalent. On the other hand, pneumonia and heart studies tend to concentrate primarily in the United States and

Europe, as well as China.

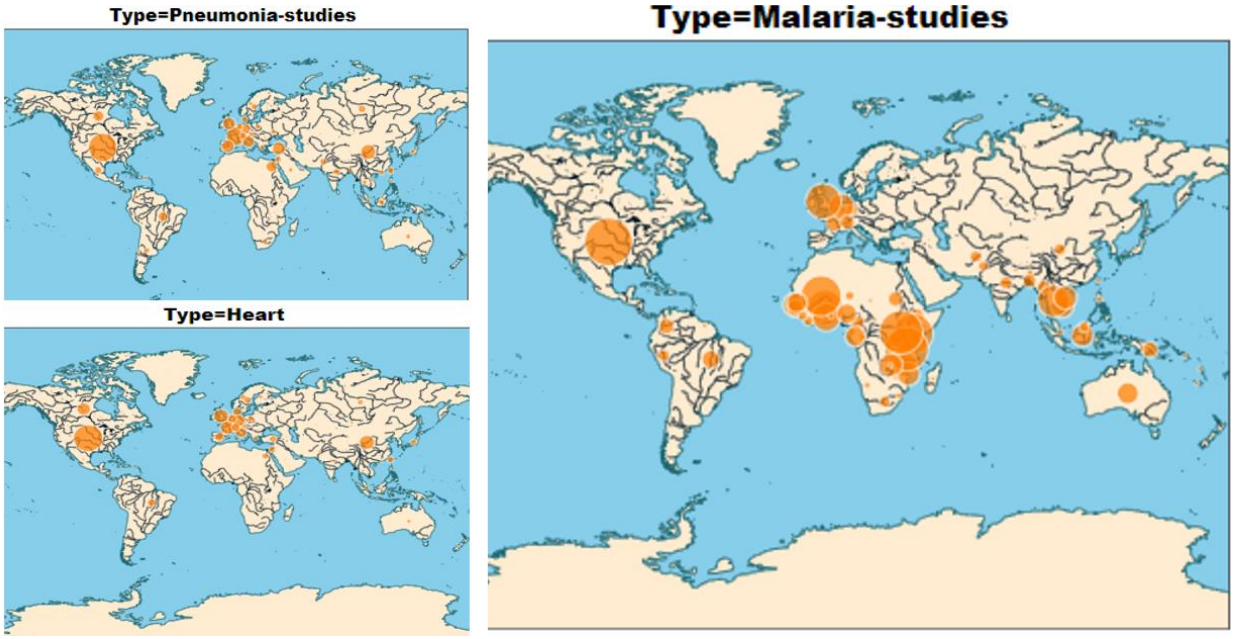

Figure 15: Malaria, TB and Heart disease distribution across the continents.

Conclusions and findings

It's intriguing to note that research efforts are thriving across the globe, with major sponsors predominantly located in Europe and North America, which are considered first-world regions. These are also the primary locations for the studies. Africa plays a significant role in research, particularly in the context of HIV and malaria studies, with Kenya being a notable hub of activity.

In terms of outcomes, a relatively small proportion of studies yield conclusive results, and more than half of the mentioned studies are already completed. The year 2000 marked a turning point when research efforts notably increased, especially in the field of cancer studies, a trend that has persisted over time. Covid and pneumonia studies experienced significant growth around 2020.

Most sponsors tend to focus on a specific type of study, while some organizations, like the Center for Infectious Diseases, sponsor a wide variety of studies. The majority of studies typically require 2 to 3 years for completion, but under high global pressure, such as in the case of Covid-19 studies, they can be concluded in less than a year.

These studies typically enroll around 100 adult participants, with a mix of males and females, although some studies, like those in cancer, may focus specifically on one gender due to differing effects. Some studies target children, while most are directed at adults or a combination of both. The primary study objectives often revolve around assessing adverse effects and response rates, with some specialized studies, such as in HIV, emphasizing measurements like viral load. These studies typically progress through four distinct phases.