The current state of clinical trials amongst 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.

Data	Details
Group	2
Members	John Kamau, Lynn Ajema and Cosmas Kibett
Mentor	Pascal Amisi
GitHub	
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?

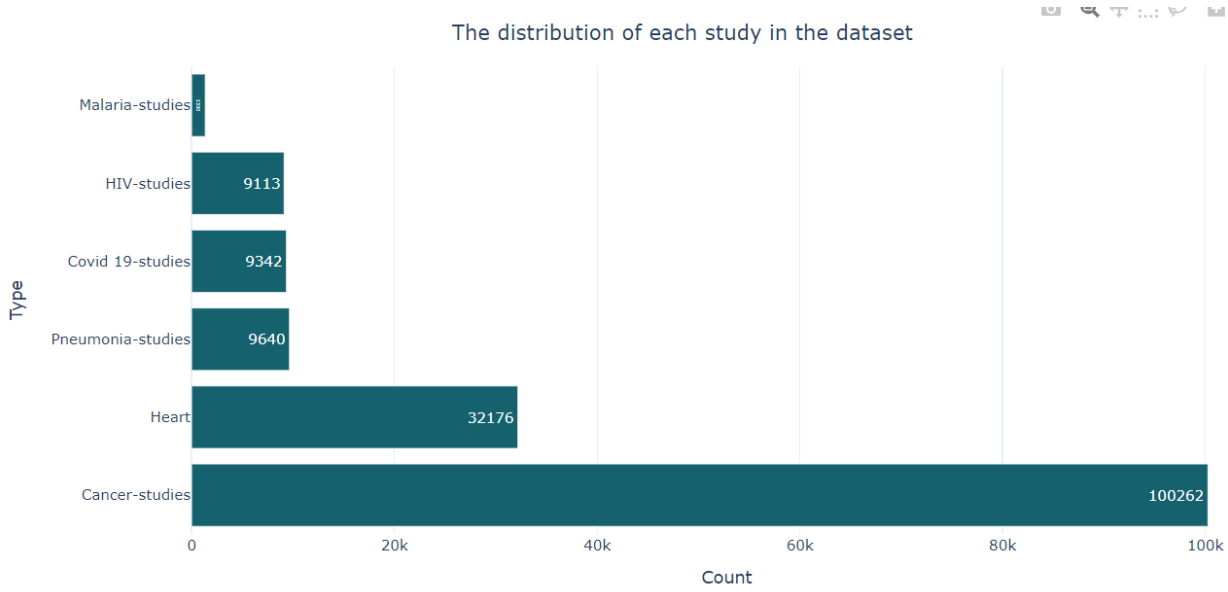
The data

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.

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:



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.

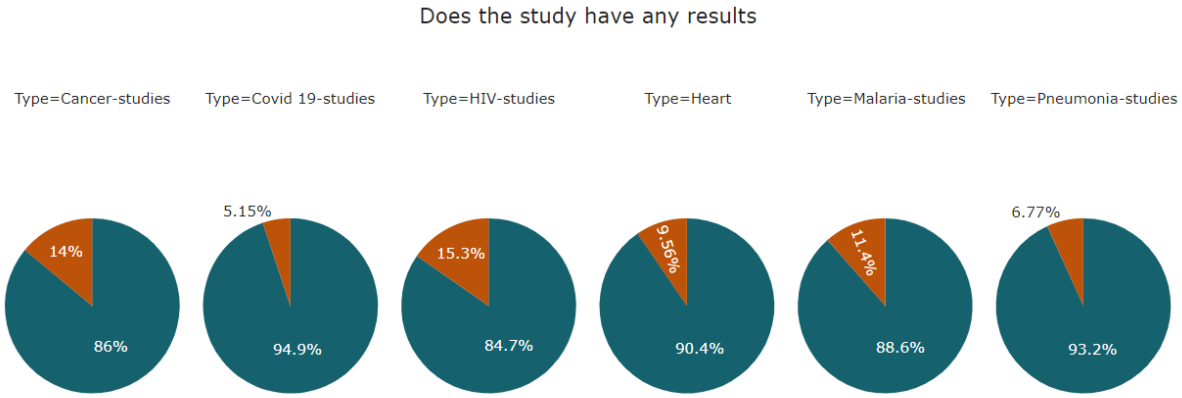
The Analysis of the data

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.

Does the study have results?

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.



What conditions are being studied?

Conclusions and findings