Clinical Trials Exploratory Data Analysis

This is the official data report of the exploratory data analysis done on heart disease clinical trials. It is a partial fulfilment of the requirement by DTE Consultancy as part of the conclusion of the DTE DATATHON programme.

By:
Getrude Obwoge - Team Lead
Teofilo Ligawa - Member

Technologies used include:
- Python
- Shell
- Amazon S3

Background of the Study

Clinical research is medical research that examines humans to better comprehend health and illness. Clinical research helps enhance the way doctors treat and prevent illnesses. Through clinical research, scientists discover:

- How the organism functions
- How disease progresses in humans, including how diseases improve or deteriorate over time
- How the body reacts to a potential treatment
- Which behaviours promote health and illness prevention, and which behaviours increase the likelihood of becoming ill?

- The objective is to use science to enhance the health care and health of individuals over time. Clinical research studies may or may not provide participants with personal benefits.

Clinical trials are research studies where participants are randomly assigned to experimental interventions, often without physicians' prescription. These trials are also known as interventional investigations, allowing participants to evaluate their experiences without choosing the intervention.

Researchers seek individuals who meet a certain set of eligibility criteria. These criteria specify who can and cannot participate in a study and may include:

- Individuals of a particular age or gender
- Those with or without a particular malady, disease, or health condition
- Individuals with or without a particular health history, such as a prior treatment, will be evaluated.
- People who are subjected to or are in contact with something that affects their health

Researchers use eligibility criteria to ensure the safety of participants and enroll the appropriate individuals in order to collect the data necessary to answer the research question. There are numerous varieties of research studies, each with distinct eligibility requirements.

Feel free to read more about clinical research here.

Problem Statement

Heart conditions, such as coronary artery disease, heart failure, and atrial fibrillation, are ubiquitous global health concerns. While advances in medical research and treatment options have improved outcomes, the availability and accessibility of clinical trials for heart-related maladies continue to be a challenge, especially for underprivileged individuals in regions with limited resources.

This research endeavour seeks to address two crucial concerns:

1. Identifying Sponsors for Researchers Numerous cardiology researchers seek funding and support for their clinical trials. Identifying appropriate sponsors who share a commitment to advancing heart disease research is frequently a difficult endeavour. The goal of this

project is to give researchers a complete list of possible sponsors based on past interactions with them. This will make it easier to write research proposals and get sponsorship from good institutions for important cardiovascular studies.

2. Access to clinical trials is often limited for individuals from underprivileged communities, especially in countries with limited healthcare resources. Patients in these regions face significant barriers when seeking treatment options for heart conditions. This project intends to recommend to patients healthcare providers that participate in diverse regions to discover ongoing clinical trials related to heart diseases.

Objective:

The objective is to determine which are the better sponsors in cardiovascular disease clinical trials.

Research Questions:

- Which type of study takes the most days? Which type of study has the most primary outcome measures?
- What is the mean number of countries that a clinical trial is involved in?
- Which sponsor takes part in the highest number of clinical trials, has the longest duration, and has the most conditions?
- Which sponsor considers the highest number of primary outcome measures?
- What is the relationship between study status and duration days, the number of primary outcome measures, and number of countries?
- What is the relationship between age and number of primary outcome measures?
- Which study type, sponsor, and age enrolled the most?
- What is the relationship between number of conditions and duration?

Data Understanding

The data was loaded from the S3 bucket to the Jupyter notebook on Google Colab. The study then performed a preliminary data inspection. The findings from the preliminary data inspection were as follows:

- There are a total of 30 columns in this dataset

- 29 columns out of the 30 columns in this dataset are string (object in pandas).
- the only numeric column is `enrollment`.
- The column names in the data need to be made uniform, e.g., there are columns like `Study URL` and `Study Type` - the naming system used here is not uniform, thus the need to make it uniform.
- There are columns like `Start Date`, `Primary Completion Date`, `Completion Date`, `First Posted`, `Results First Posted`, `Last Update Posted` that should be datetime objects yet are categorized as Pandas objects (strings). The study also feels that some of these date elements are redundant, and as such, some of these columns shall be dropped. The fact that some of these columns contain missing values could also be used to inform the decision making on the columns that are to be dropped/ maintained.
- The data contains missing values in a number of the columns, and the missing values shall be dealt with based on the type of data and the number of missing values per column. The default mechanism will be to drop the missing values, knowing that in the real world, data from one trial cannot be filled using data from another trial.
- Some columns like the `Study URL` might not really be helpful in terms of adding value to this study.
- The variable `conditions` would be considered a categorical column, but the fact that some conditions are repeated but with different phrasing puts the study in a difficult position to consider it as such.

Data Preparation

The study took note of the observations from the data understanding and used them to preprocess the data.
The following steps were taken:

- A threshold of 40% was set to aid in removal of missing values; essentially, this means that if a column has at least 40% of its rows as missing values, then the column would be dropped altogether. Columns that were removed like this include the acronym, other outcome measures, collaborators, phases, results first posted, and study documents.
- Columns with less than 40% missing values were not dropped, but just the rows with missing values from those columns were dropped.
- The columns NCT number, study url, other IDs, primary completion date, first posted, and last update posted were dropped because they would not be useful during the course of the study.

- Uniform naming of the columns was done by first lowering the case and filling whitespaces between the words in the column names with the underscore ‘_’. Forexample, the Study Title was now study_title.
- Columns that were categorical in nature were converted to type categorical. For example, funder type and study status

Feature Engineering and Text Analytics

Upon granular inspection of the data, the study engineered new columns from already existing columns. The first column to be engineered was countries. The locations column provided information on the locations where the trials were conducted. It was a single string, but we realized that each location in that string was separated by the pipe operator '|' . The pipe operator was then used as the delimiter that separated the various elements in the list. The country of each location was extracted and we cross-checked the country names extracted using the python package geonamescache for correctness. Names that were not found include ’Reunion’, ‘Iran, Islamic Republic of’, ‘Russia Federation', and ‘Korea, Republic of’. The study corrected these to France (Reunion is an island under the control of France), Iran, Russia and South Korea.
The study then found out the number of countries that a study was involved in by counting the number of elements that were in the resulting countries column of each row.

The study used the above mentioned delimiter to find the number of conditions, primary outcome measures and secondary outcome measures that each study had. The last feature that was engineered was the duration days which was basically the difference between the completion date and the start date of a study.

In summary, 5 new columns were generated, they include:

- Countries
- Num countries - number of countries
- Num conditions
- Num primary measures
- Num secondary measures

The summary of the numeric columns was as follows:

- The mean number of enrollment in the cases was 4835
- The mean number of countries involved in a study was 3.
- The mean number of primary measures was 2, while that of secondary measures being at around 8.
- The mean number of conditions is 2 while the mean duration days for the studies was 1204.
- There is at least one study that did not even last a single day.

The text column brief summary was preprocessed using the following steps using a spacy nlp model:

- Contractions like can't were fixed to cannot. The words were filtered and those that had less than 3 letters were removed.
- Noise was removed, noise was regarded as special characters or phrases that did not make sense like n/n/*.
- Digits like 2 were converted to words like two.
- The words were then joined to form one single text.
- The resulting text was then lemmatized, where words were transformed into their canonical base or lemma. For example, the lemma of "running" is "run," and the lemma of "better" is "good." Stopwords like 'is' and 'the' were removed. The text was then joined once more so that it could be returned as a single text.

The above steps were used to create the corpus for the study. A corpus is a collection of text.

The study then performed some basic text analytics and found that the top 5 unigrams (words) in the data are:

- *Patient*
- *Study*
- *Heart*
- *Cardiac*
- *coronary*

The top 5 bigrams were:

- *Heart failure*
- *Purpose study*
- *Coronary artery*
- *Atrial fibrillation*

- *Study evaluate*

The top 3 trigrams were:

- *Coronary artery disease*
- *Percurteneous coronary intervention*
- *Acute coronary syndrome*

It is worth noting that '*coronary*' is in the top 5 of unigrams, the top 3 ofbigrams, and the most common trigram. It also worth pointing out that in the top 10 trigrams, it appears 5 times.

From the above, clearly we can deduce that in all these trials, the coronary factor was major.

The study then created a word cloud where some of the dominating words were:

- *Heart failure*
- *Atrial fibrillation*
- *Patient*
- *Coronary artery*
- *Purpose study*

Univariate Data Analysis

Countries

Univariate data analysis was performed where columns were considered in singular fashion.
A list of countries from the list of countries in each of the studies was created. The list was then flattened and the number of times that a country appeared was counted to obtain the countries that participated in most trials. The top 10 was as follows:

- United States: 5035 studies

- France: 2681 studies
- Germany: 1668 studies
- Canada: 1613 studies
- China: 1479 studies
- United Kingdom: 1351 studies
- Italy: 1279 studies
- Netherlands: 1016 studies
- South Korea: 1016 studies
- Russia: 1016 studies

It was not surprising to see that the countries in the top 10 were all first world countries.
The findings from the other columns include:

- *interventional* study type dominated with 80% of the data while *observational* study type formed 20% of the studies.
- 12.4 % of the study results had been posted, 87.6 % of the trials did not post the study results.
- 96.2 % of the trials included all sexes, 2.4 % females and the remaining 1.4 % males.
- From the study status distribution, most of the studies have been completed - about 8000 and about 3500 studies are recruiting. Less than 400 studies have been suspended. This shows that most studies follow the regulations.
- The most common sponsors include *Abbott Medical Devices*, *Assistance Publique - Hôpitaux de Paris*, *Mayo Clinic* and *Bayer*.
- The most common conditions include *Coronary Artery Disease*, *Heart Failure* and *Atrial Fibrillation* - this confirms the findings from the text analytics which was based on the columns brief summary.

Distribution of Numeric Columns

The study went ahead and checked the distribution of the numeric columns, we realized that none of them was normally distributed.
The numeric columns in the study include:

- Enrollment
- Num countries
- Num conditions
- Num primary measures

- Num secondary measures

The mean number of countries that clinical trials were involved in was 1.77. This could be rounded off to 2 countries.

Bivariate Data Analysis

The study began the bivariate analysis by investigating the association between the categorical variables using Cramer's V.
Here is how to interpret the result from a Cramer's V:
Cramér's V is an effect size measurement for the chi-square test of independence. It measures how strongly two categorical fields are associated.
$ES \leq 0.2$ The result is weak. Although the result is statistically significant, the fields are only weakly associated.
$0.2 < ES \leq 0.6$ The result is moderate. The fields are moderately associated.
$ES > 0.6$ The result is strong. The fields are strongly associated.
For example there is a strong association of 0.88 between the funder type and the sponsors of a study.

Categorical columns were compared against numeric columns. This was done to answer the research questions that have been mentioned above and it was done by grouping the data by the categorical columns and then using the mean as the aggregate measure with which to compare the various categories for a particular column. Vizualizations were then made to show the comparisons.

Study type

The first was study type against the numeric columns and the observations were as follows:

- Observational trials took more days to complete than interventional trials. At 1200 vs 1100.
- Observational trials had more primary outcome measures on average than interverntional trials. At about 1.75 vs 1.60
- Observational trials, on average had more conditions that interventional trials. At 2.2about vs 1.8
- On average, interventional trials involved more countries than observational trials. At about 1.8 vs 1.55
- On average, observational trials have more enrollment at about 15000 than interventional trials at about 3000.

Age

The observations that were made when age was compared to numeric variables include:

- Trials that include all ages enroll more people than the others, these are the *CHILD, ADULT, OLDER_ADULT* with an average of over 1400 and those of adults only (excluding older adults) enroll the fewest participants with about 800.
- Trials for adults only have the highest number of primary outcome measures at an average of just under 2.5 while those of older adults having the fewest at an average of about 1.6.
- Trials involving adults only have the fewest conditions while those involving children and adults(excluding older adults) have the most conditions.
- Most clinical trials enroll all age groups(*CHILD,ADULT,OLDER_ADULT*)

Sponsor

The case for this column was special because there was a realization that there are extreme values. For example, when the dataframe was grouped by this column and then aggregated using the mean to find the top 10 sponsors based on the number of enrollments they have, there came up a sponsor by the name *Bentong Yu* who had enrolled about 20 million participants in a study. Upon inspection, it was realized that it only registered one trial and this was not fair for the other sponsors who had registered several trials. We then decided to use the top 1% sponsors based on the number of trials they have registered, the precise figure was 52 studies for the least value among the top 1%, thus the study adopted a threshold of 50 studies as the minimum number of studies for a sponsor to be considered in finding the best sponsors.

The findings were as follows:

- *Norvatis Pharmaceuticals* leads the way in terms of the number of countries of operations, having about 9 countries, followed by *GlaxoSmithKline* with about 6, *Bayer* and *AstraZeneca* with about 5 each.
- *Edwards Lifesciences* takes the longest days with about 3000 days on average. It is followed by *Seung-Jung Park* at around 2300 and *Rigshopitalet, Denmark* at just under 2000.
- *Bayer* leads the way by enrolling about 55000 participants, followed by *Stanford University* at around 40000 participants.
- *VA Office of Research and Development* enrolls just about 25000 participants.
- The sponsor that considers the most number of conditions - at about 3 is *Brigham and Women's Hospital*.

- *AstraZeneca* leads by an average of 4 primary measures in the studies that it is involved in followed closely by *GlaxoSmithKline* (3.7) and *Bayer* (3.2).

Correlation of Numeric Columns

The correlation of numeric columns was investigated. The pairings were:

- Number of conditions vs duration - the correlation coefficient stood at 0.046 which is a very weak but positive linear association.
- Number of conditions vs number of primary outcome measures - the correlation coefficient stood at 0.0488 which is weak as well.

Conclusion

The conclusion from the analysis was as follows:

1. 80% of the trials are interventional with the remaining 20% being observational.
2. 12.4 % of the study results had been posted, 87.6 % of the trials did not post the study results.
3. 96.2 % of the trials included all sexes, 2.4 % females and the remaining 1.4 % males.
4. The most common sponsors according to the number of appearances include Abbott Medical Devices, Assistance Publique - Hôpitaux de Paris, Mayo Clinic and Bayer
5. The most common conditions include Coronary Artery Disease, Heart Failure and Atrial Fibrillation.
6. The most common word in the trial summaries is patient - this can serve to show that patient care is at the top of most clinical trials' priorities.
7. The word patient is followed by study, heart and cardiac. This serves as a testament that these clinical trials are focused on the heart.
8. The most common two-word sequence is heart failure, purpose study and coronary artery. It could be argued out that 'purpose study' was initially 'purpose of the study' but due to the fact that 'of' and 'the' are stopwords, they have been removed.
9. The most common three-word sequence is coronary artery disease, followed by percutaneous coronary intervention, purpose study determine and acute coronary syndrome.
10. In most of the studies, the coronary factor was major problem.

11. On average, observational trials have more enrollments (15000 vs 3000), conditions, primary outcome measures and take longer than interventional trials (At 1200 vs 1100 days.).
12. Interventional trials take part in more countries than observational trials (1.8 vs 1.55).
13. Trials that include all ages enroll more people than the trials of specified age ranges.
14. Trials for adults only have the highest number of primary outcome measures.
15. Trials involving adults only have the fewest conditions while those involving children and adults(excluding older adults) have the most conditions.
16. Most clinical trials enroll all age groups(CHILD,ADULT,OLDER_ADULT)
17. Trials that are active and not recruiting have generally enrolled more people than the other trials in that category, they also take longer durations.
18. Funder type OTHER enrolls more people in their trials than any other funder type.
19. Industry funders fund trials in more countries than other funder types, with slightly over an average of 3.5.
20. NIH funds studies that take the longest duration - about 2000 days on average.
21. Funder type of unknown funds studies with the most number of conditions, averaging at about 4 conditions.
22. Industry funds trials with the highest number of primary outcome measures at an average of slightly above 1.75.

Among the 1% of the top sponsor by appearances

The sponsors that make at least 3 out of 5 appearances in the 5 metrics include:

- *Norvatis Pharmaceuticals*
- *Bayer*
- *Yonsei University*
- *VA office of Research and Development*
- *Abbot Medical Devices*

The sponsors that make at least 2 out of 5 appearances in the 5 metrics include:

- *GlaxoSmithKline*
- *AstraZeneca*
- *Boston Science Corporation*

- *Samsung Medical Center*
- *Mayo Clinic*
- *Seung-Jeung Park*
- *VA office of Research and Develpment*
- *Rigshospitalet, Denmark*
- *Brigham and Women's Hospital*
- *University of Sao Paulo General Hospital*

Recommendations

The following recommendations can be drawn from the study:

The study has found that some of the best sponsors in reasearching heart diseases include:

- Norvatis Pharmaceuticals.
- Bayer.
- Yonsei University.
- VA office of Research and Development.
- Abbot Medical Devices.
- GlaxoSmithKline.
- AstraZeneca.
- Mayo Clinic.

The above sponsors would be recommended to a researcher who has drafted a proposal to perform a clinical trial in cardiovascular diseases. The above sponsors are also the sponsors that would be recommended to a patient that wants to sign up for treatment of a certain condition.

Increase Transparency in Posting Study Results: Since a significant percentage of clinical trials did not post their study results, it's important to encourage sponsors to improve transparency by posting the results of their trials. This can enhance the overall quality and trustworthiness of clinical research.
Given that certain conditions such as Coronary Artery Disease, Heart Failure, and Atrial Fibrillation are common, sponsors should continue to focus on researching and developing treatments for these prevalent health issues.
Patient-Centered Approach: The prevalence of the word "patient" in trial summaries underscores the importance of a patient-centered approach.
Efficiency in Trial Design: Observational trials tend to have longer durations, higher enrollments, and more primary outcome measures than interventional trials. Sponsors should consider the trade-offs between observational and interventional trial designs to ensure efficiency while maintaining scientific rigor.
Regular Evaluation: Continuously evaluate and assess the performance of sponsors, both in terms of quantity and quality of trials, and recognize and reward sponsors who consistently contribute to the advancement of clinical research.

Future Work

Having the financials, such as the budget used in the study and the revenue that would be generated with the adoption of the approaches established in the study or even the lives saved by using the approaches from the study, would be nice to have to further the study.
Further exploration of variables such as the number of countries, conditions, primary outcome variables, days and enrolment is needed to determine whether more is good or bad or how the tradeoff happens. Knowing whether a study was successful or not would be helpful in this.