# Analysis of Clinical Trials Data: Investigating Study Areas, Collaborators Availability, Interventions, Enrollment, Funder Types, and Masking Types

**Team 5**

Kelvin Kiplagat

Thomas Muriuki

**Mentor**

Julius Owuonda

October 20, 2023

DTE Consultancy Datathon Challenge

DTE Team 5 Phase II.ipynb

# TABLE OF CONTENTS

# Introduction

Clinical trials represent a cornerstone of medical research, serving as the crucible in which new treatments, interventions, and therapies are rigorously evaluated. According to Seyhan and Carini (2019), these trials offer a critical pathway for transforming groundbreaking scientific discoveries into practical, patient-centric healthcare solutions. The data presented in this report, drawn from six distinct datasets, each corresponding to a specific study area, sheds light on the broad yet indefinite landscape of clinical research and its profound implications for medical science and public health.

*ClinicalTrials.gov*, the source of the data underpinning this analysis, serves as a repository for an ever-expanding field of clinical trials worldwide. Researchers, clinicians, and policy-makers depend on this vast archive of clinical studies to get insights, make informed decisions, and direct resources effectively. Each dataset represents a different study area – Cancer, HIV, pneumonia, malaria, heart, and the unprecedented challenge of COVID-19 (*ClinicalTrials.gov*, 2021). To make use of them, we delve into the comprehensive analysis of clinical investigations.

## Problem Statement

### *The Broader Significance of Clinical Trials*

Beyond the specific data points examined in this analysis, clinical trials embody a universal force in modern healthcare where studies encompass a plethora of therapeutic and diagnostic interventions and bear consequences far-reaching and enduring. They constitute the bedrock of evidence-based medicine, enabling healthcare practitioners to refine their practices and make well-informed decisions. Clinical trials also drive medical innovation by fostering interdisciplinary collaborations since researchers from diverse backgrounds and institutions unite to challenge medical frontiers, expanding the bounds of scientific knowledge. It is in the sense of clinical trials that diseases once considered insurmountable became amenable to treatment and management, and novel therapeutics emerged (Novitzke, 2008).

### *Unveiling the Richness of Clinical Research*

In this report, we comprehensively examine clinical trials, traversing their geographic, temporal (time), and thematic diversity. We address questions that illuminate facets often concealed within the complexities of clinical research. Our inquiry extends to topics such as the impact of masking types in different study areas, the allocation of collaborators, the prevalence of interventions, the influence of demographics, the role of funding sources, and the diversity of interventions employed. As we navigate the intricate web of clinical trials, we endeavor to draw insights that transcend individual datasets and we aspire to contribute not only to the refinement of therapeutic approaches but also to the body of knowledge guiding the design, implementation, and evaluation of future clinical studies.

## Objective of the Study

The main objective of this study is to analyze clinical trials data and investigate various aspects related to study areas, collaborators, interventions, enrollment, funder types, masking types, and the outcomes of clinical trials. This analysis aims to answer specific research questions utilizing appropriate statistical methods.

***Essentially, we aim to comprehensively explore and gain insights into the diverse landscape of clinical trials across different study areas and their associated attributes, including collaborators, interventions, enrollment, demographics, funder types, and masking, with the ultimate goal of informing evidence-based decision-making, healthcare policies, and future research priorities.***

The research questions pertinent to the study are:

1. Are there masking types common in different study areas than others?

2. a) Are there differences in the number of collaborators different study areas get? b) Does the availability of collaborators affect study status in completed, terminated, and suspended studies?
3. Does the number of interventions differ in different study areas?
4. Are there differences in enrollment across sex and age?
5. Does the funder type affect the study status?
6. Are there differences in the number of interventions used in masking type and intervention model categories?

***The Journey Ahead***

Our exploration into clinical trials begins with a meticulous examination of the data associated with each study area, followed by a comparative analysis that elucidates patterns, relationships, and potential opportunities for enhancement. The findings presented in this report hold the potential to inform research priorities, healthcare policy, and clinical practice in ways that may lead to improved patient outcomes and a more efficient and responsive healthcare system. Taking a leap into these realms, we embark on a journey through the evolving landscape of medical research, demonstrating the enduring commitment of the scientific community to the pursuit of knowledge and the betterment of public health. ***DTE Consultancy***, having granted us a chance to analyze the clinical trial datasets, has stretched the capacity of human knowledge in this invaluable field by a mile.

# Methods

## Data and Development

The data for this study was sourced from ClinicalTrials.gov, a comprehensive and publicly accessible repository of clinical trials information. Six distinct datasets were utilized, each corresponding to a specific study area: Cancer, HIV, pneumonia, malaria, heart, and COVID-19. These datasets were systematically compiled for analysis. Since there was a massive disparity in sample sizes between different datasets, some inferential statistical tests required resampling (majorly undersampling) the datasets to bring a balance in categories.

Development was done over Google Colaboratory Environment, Python Version 3.10.12. This came in handy given the computational intensity required by the data thus easing the analysis a bit.

### Data Preparation and Preprocessing

Prior to analysis, the datasets underwent a series of preprocessing steps. The datasets were merged into a single dataset and data cleaning commenced. Handling missing values, and outliers, and ensuring data consistency was carried out. Some features were merged as one. Other features were created by extracting information from other columns.

It is worth noting that data we considered as outliers were not dropped but rather analyzed as a group to gain some insights into them. A broad instance would be the enrollment numbers of COVID-19 which were extremely high, yet so crucial in understanding why such a scenario happened.

### Statistical Analysis

Aside from the descriptive statistics and visualizations encompassing descriptive analytics, the research also sought to answer some pertinent questions listed below using inferential statistical measures encompassing; Chi-Square Tests, One-Way and Two-Way ANOVA as well as post hoc tests done using Tukey HSD (Honesty Significance Difference) to underscore the significance of the results which would be tested at $\alpha = 0.05$.

Developed by Karl Pearson, the Chi-Square statistical test examines the differences between categorical variables to establish whether it is by chance or a relationship between them (Biswal, 2023). Analysis of Variance (ANOVA), whose focus is on differences of group means, is popular in medical research since its need stems from "the error of alpha level inflation, which increases Type 1 error probability (false positive) and is caused by multiple comparisons" (Kim, 2017)

### Topic Modeling

Topic modeling was employed to ascertain the thematic focus of clinical trials based on primary and secondary outcome measures. Latent Dirichlet Allocation (LDA) was applied to extract key topics from the text data. Blei et al. (2003) defined Latent Dirichlet Allocation as a generative probabilistic model for collections of discrete data such as text corpora with a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topic probabilities.

### Visualization

Various visualization techniques were utilized to present findings effectively. Python libraries such as Matplotlib and Seaborn were employed to create visualizations, including bar charts, box plots, pie charts, contingency tables, and interactive plots. The application of rigorous statistical methods, coupled with data visualization, enabled a comprehensive analysis of clinical trials data, resulting in important insights into the multifaceted landscape of clinical research across different study areas. The combination of these methodologies provides a robust foundation for the interpretation and communication of study outcomes.

In summary, the analysis will employ a combination of statistical methods, including Chi-Square Tests, One-Way ANOVA, Two-Way ANOVA, Post-Hoc tests such as Tukey HSD

(Honesty Significance Difference), and topic modeling, to address the research questions and fulfill the main objective of this comprehensive analysis.

## Results and Discussion

This section sheds light on the results of the study. Before the main analysis was done, we investigated the data and noted a few anomalies. Some trials had no enrollment (enrollment was 0) while there were serious outliers that needed attention as established below.

### Anomalies & Outliers

Figure 1 shows the frequency of observations(trials) with no enrollment per study status. Table 1 investigates the completed trials that had no enrollment data. The common funder type was the NIH. The pneumonia study area had only 1 trial with no enrollment while malaria and HIV were the most common. The absence of these figures in completed studies could have been caused by the erroneous data ingestion since their frequency as a ratio of all completed studies is minuscule.

Figure 2 shows the average of outliers in enrollment numbers of clinical trials and time taken by the trials in days to completion for funder type, study area, and study status. COVID-19 trials had the highest enrollment figures treated as outliers on average while Malaria had the least. Outlier cancer trials took the longest time to complete. Trials funded by the INDUSTRY had the highest enrollment figures yet took the least time to complete. NIH on the other hand took the most number of days in outlier terms.

*Figure 1: Frequency of Trials with no Enrollments per Study Status*

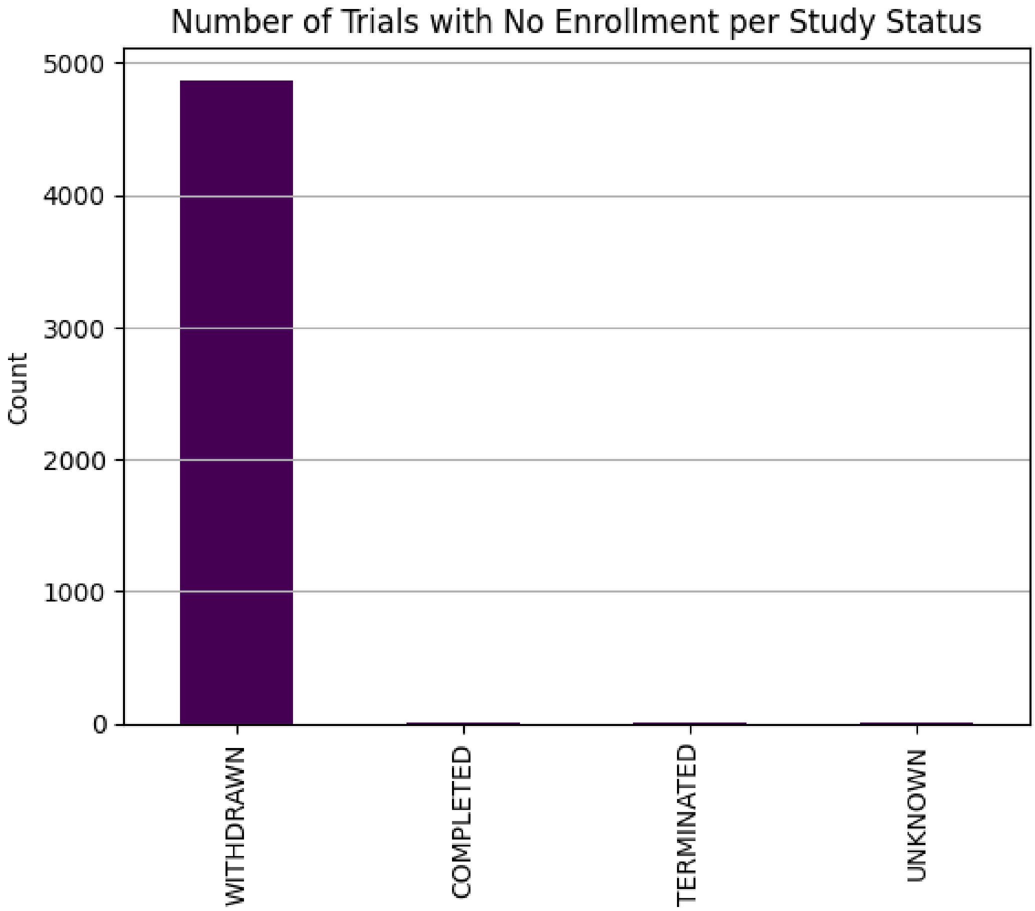
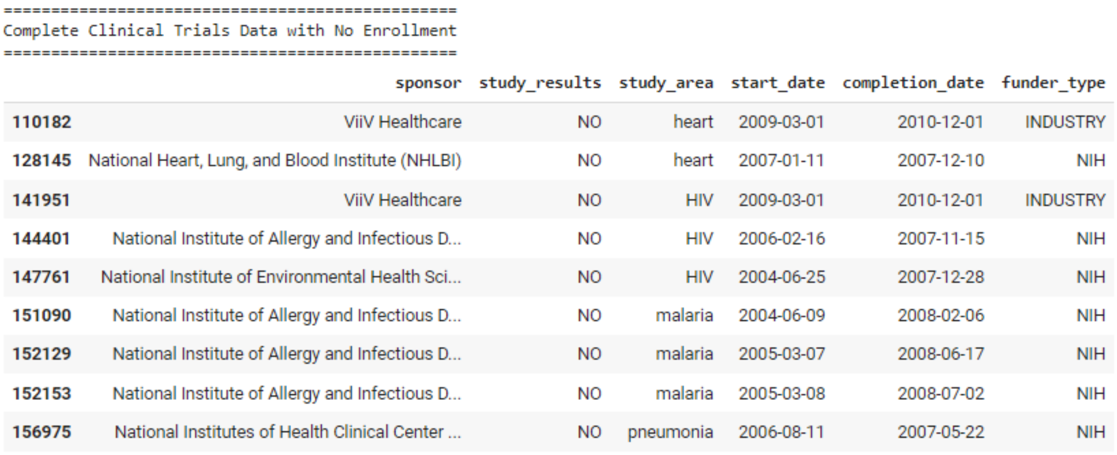


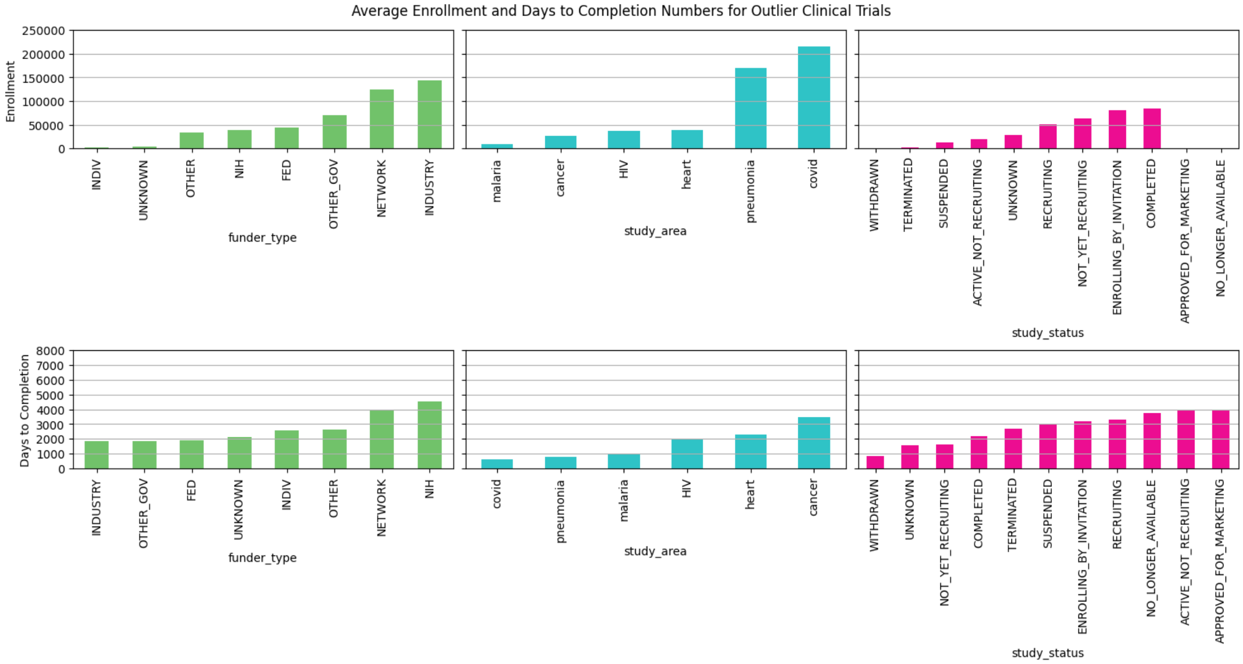
*Table 1: Completed Trials with no Enrollment Data*



```
================================================
Complete Clinical Trials Data with No Enrollment
================================================
```

| | sponsor | study_results | study_area | start_date | completion_date | funder_type |
|---|---|---|---|---|---|---|
| **110182** | ViiV Healthcare | NO | heart | 2009-03-01 | 2010-12-01 | INDUSTRY |
| **128145** | National Heart, Lung, and Blood Institute (NHLBI) | NO | heart | 2007-01-11 | 2007-12-10 | NIH |
| **141951** | ViiV Healthcare | NO | HIV | 2009-03-01 | 2010-12-01 | INDUSTRY |
| **144401** | National Institute of Allergy and Infectious D... | NO | HIV | 2006-02-16 | 2007-11-15 | NIH |
| **147761** | National Institute of Environmental Health Sci... | NO | HIV | 2004-06-25 | 2007-12-28 | NIH |
| **151090** | National Institute of Allergy and Infectious D... | NO | malaria | 2004-06-09 | 2008-02-06 | NIH |
| **152129** | National Institute of Allergy and Infectious D... | NO | malaria | 2005-03-07 | 2008-06-17 | NIH |
| **152153** | National Institute of Allergy and Infectious D... | NO | malaria | 2005-03-08 | 2008-07-02 | NIH |
| **156975** | National Institutes of Health Clinical Center ... | NO | pneumonia | 2006-08-11 | 2007-05-22 | NIH |

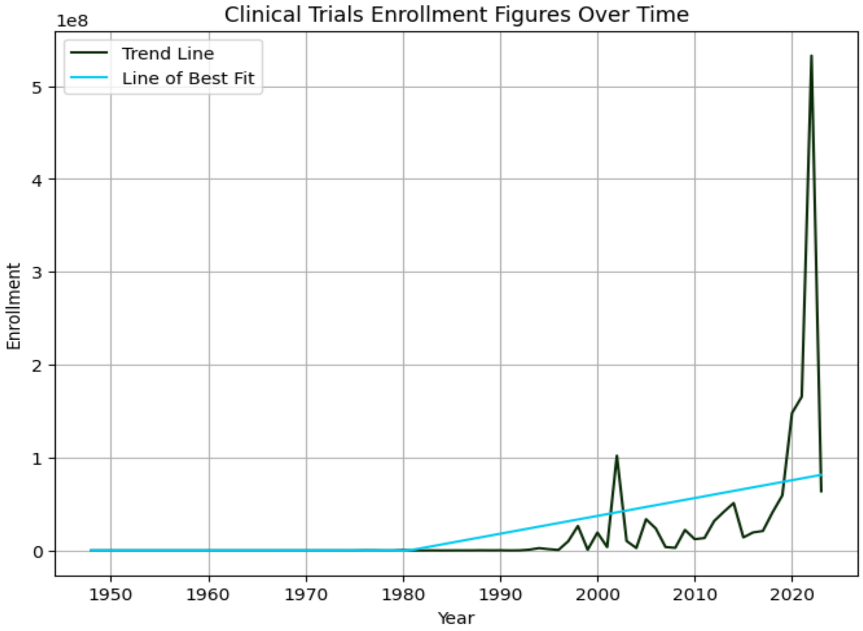
*Figure 2: Average Enrollment and Days to Completion Numbers for Outlier Clinical Trials*



## Descriptive Analytics

### *Clinical Trials Enrollment Over Time*

*Figure 3: Enrollment Figures over Time*

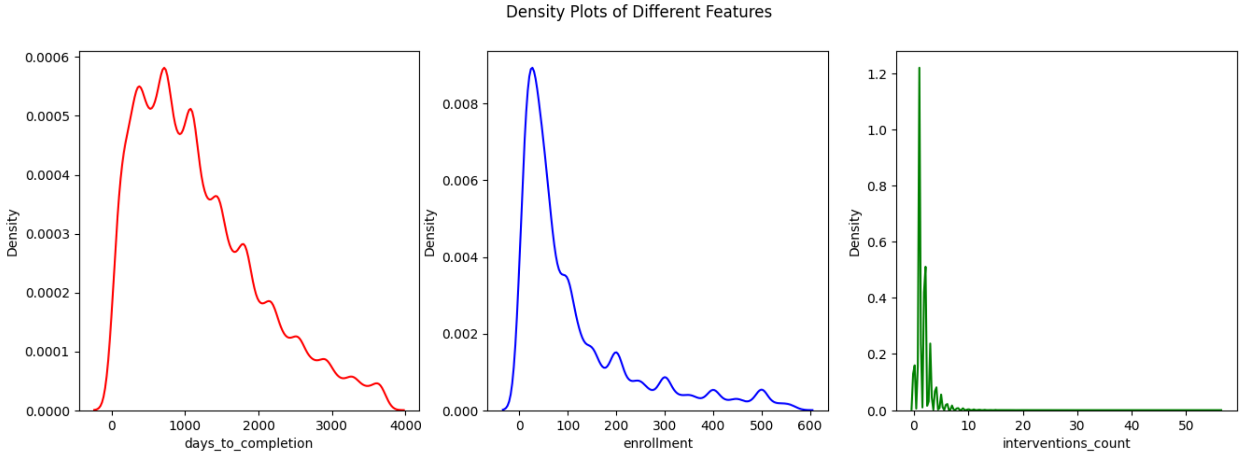


This section focuses on the clinical trial enrollments as a whole over time regardless of outliers but capped at 2023. Enrollments that begin in 2024 going forward are not included. Figure 3 shows that there has been an overall upward trajectory over time of clinical trial enrollments. However, 2002, 2019, 2020, and 2021 show rather uneven peaks with the latter three years having been contributed by the COVID-19 Pandemic. 2002 may have been a result of the SARS-COV2 virus among other trials.

### *Distribution of Days to Completion, Enrollment, and Number of Interventions per Trial*

Figure 4 shows the distribution of various numerical features after outliers had been eliminated for both days to completion and enrollment. We can underscore that all the features are right-skewed with the intervention count being heavily right-skewed. In right-skewed distributions, the mode is lesser than the median which is often lesser than the mean because the long tail pulls the average value towards the right (Bensken et al., 2021). All the plots are leptokurtic given their very high peak in the mode values, however, the time taken to completion is nearly multimodal given the three close peaks. Ideally, most trials took between 700 and 1400 days to complete. The majority of trials had an enrollment figure of approximately 70. The majority of trials took approximately between 2 and 5 interventions.

*Figure 4: Density Plots*



### *Average Days to Completion by Study Area, Sex*

*Figure 5: Average Days to Completion by Study Area, Sex*

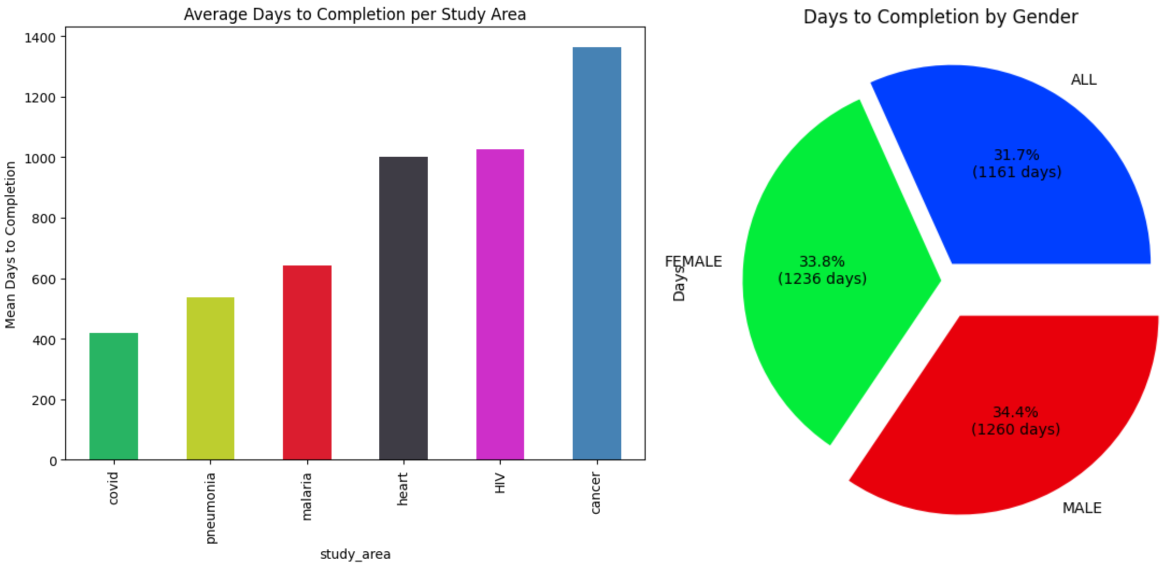


Figure 5 shows that Cancer trials took the highest number of days to complete, while COVID took the least number of days. Heart and HIV-related trials are nearly similar coming fourth and fifth respectively. In terms of classification by sex, all trials have nearly similar values percent-wise, however, we noted that trials that encompassed all sex categories took the least number of days (1161 days) while trials that focused only on Males took the highest number of days (1260) to complete. Trials whose focus was on Females took an average of 1236 days.

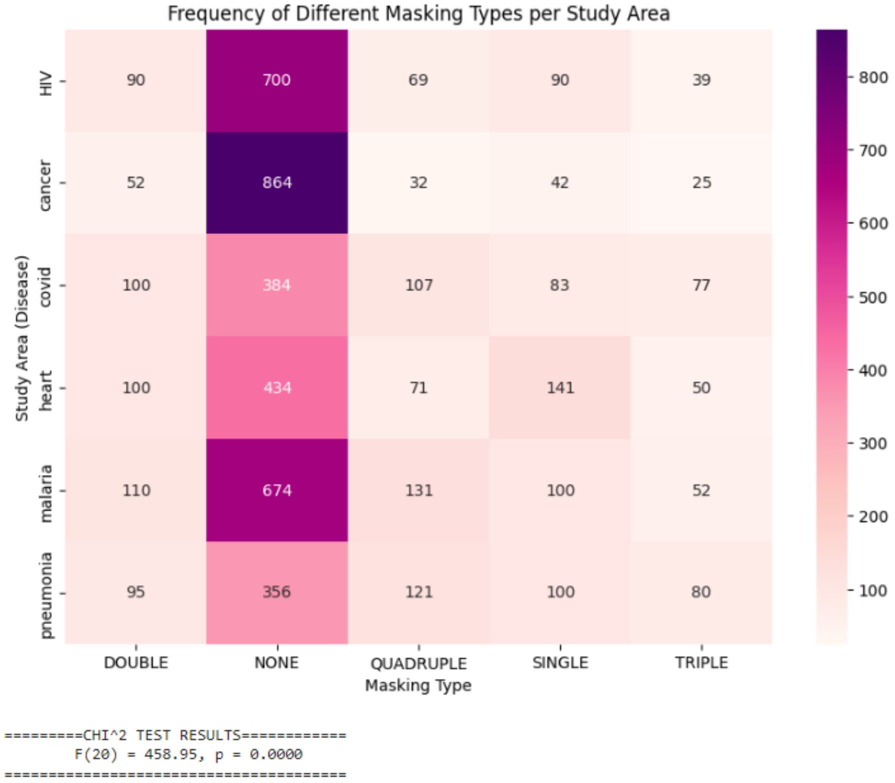
## Inferential Analytics

This section seeks to answer the questions earlier established.

### *1) Are there masking types more common in different study areas than others?*

Figure 6 contains the contingency table and the chi-square results at the bottom. The Chi^2 test results (F(20) = 458.95, p = 0.0000) underscores that there are differences in masking types used in different study areas. To underscore the significance of the results we check the contingency table. For instance, aside from NONE, which means no masking type was used, common in all areas, we can see that SINGLE masking type is common in HIV and heart study areas. QUADRUPLE masking type is common in COVID and Malaria studies. Double masking type is common in Cancer studies. TRIPLE masking type is barely preferred since it doesn't rank at least third in all the study areas.

*Figure 6: Contingency Table and Chi^2 Results*

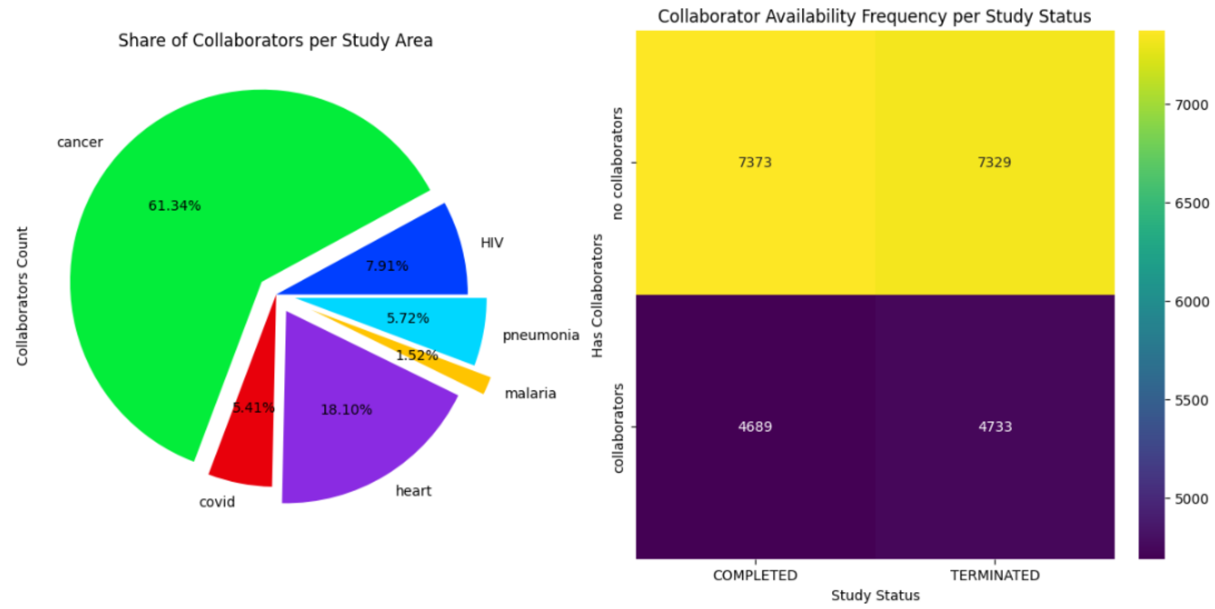


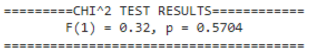
```
=========CHI^2 TEST RESULTS============
       F(20) = 458.95, p = 0.0000
=======================================
```

2) ***a) Are there differences in the number of collaborators different study areas get? b) Does the availability of collaborators affect study status in completed, terminated, and suspended studies?***

The pie chart in Figure 7 answers (a). We can evidently see that different study areas received different numbers of collaborators with Cancer having the highest number of collaborators and the least number of collaborators in Malaria. These disparities may not be representative because of the differences in samples per study area. That’s why for (b) we investigated if collaborator availability affects study status. We limited the study status to only completed and terminated trials and undersampled the terminated studies to match the completed studies. *The findings (F(1) = 0.32, p = 0.5704) show that collaborator availability does not affect the study status.*

*Figure 7: Collaborators Availability*



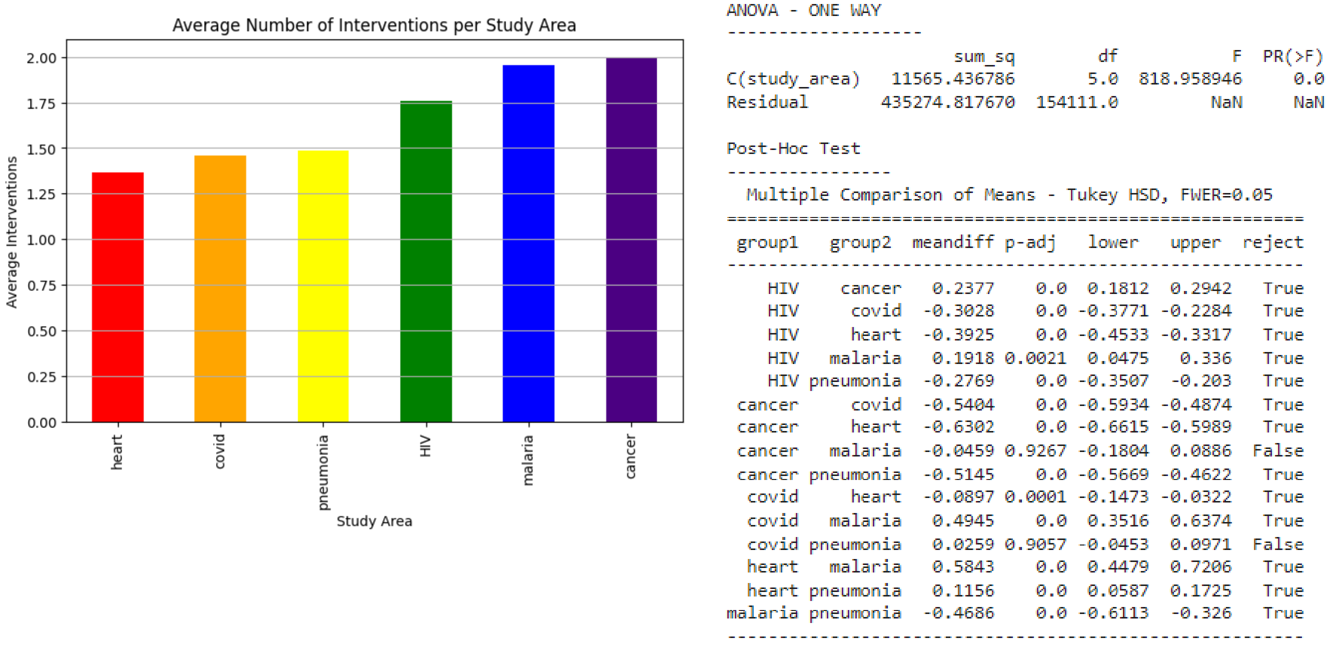


```
=========CHI^2 TEST RESULTS============
        F(1) = 0.32, p = 0.5704
=======================================
```

### *3) Does the number of interventions differ in different study areas?*

Figure 8 shows a bar chart displaying the average number of interventions per study area and One-way ANOVA analysis to compare differences in means. The findings from one-way ANOVA (F = 818.96, p = 0.0000) show that the number of interventions used in different study areas differs. The significance of the findings supports the bar chart insights with heart studies receiving the least number of interventions on average while Cancer received the most number of interventions. The post hoc tests show that only cancer and malaria, and covid and pneumonia didn't have a difference in means.

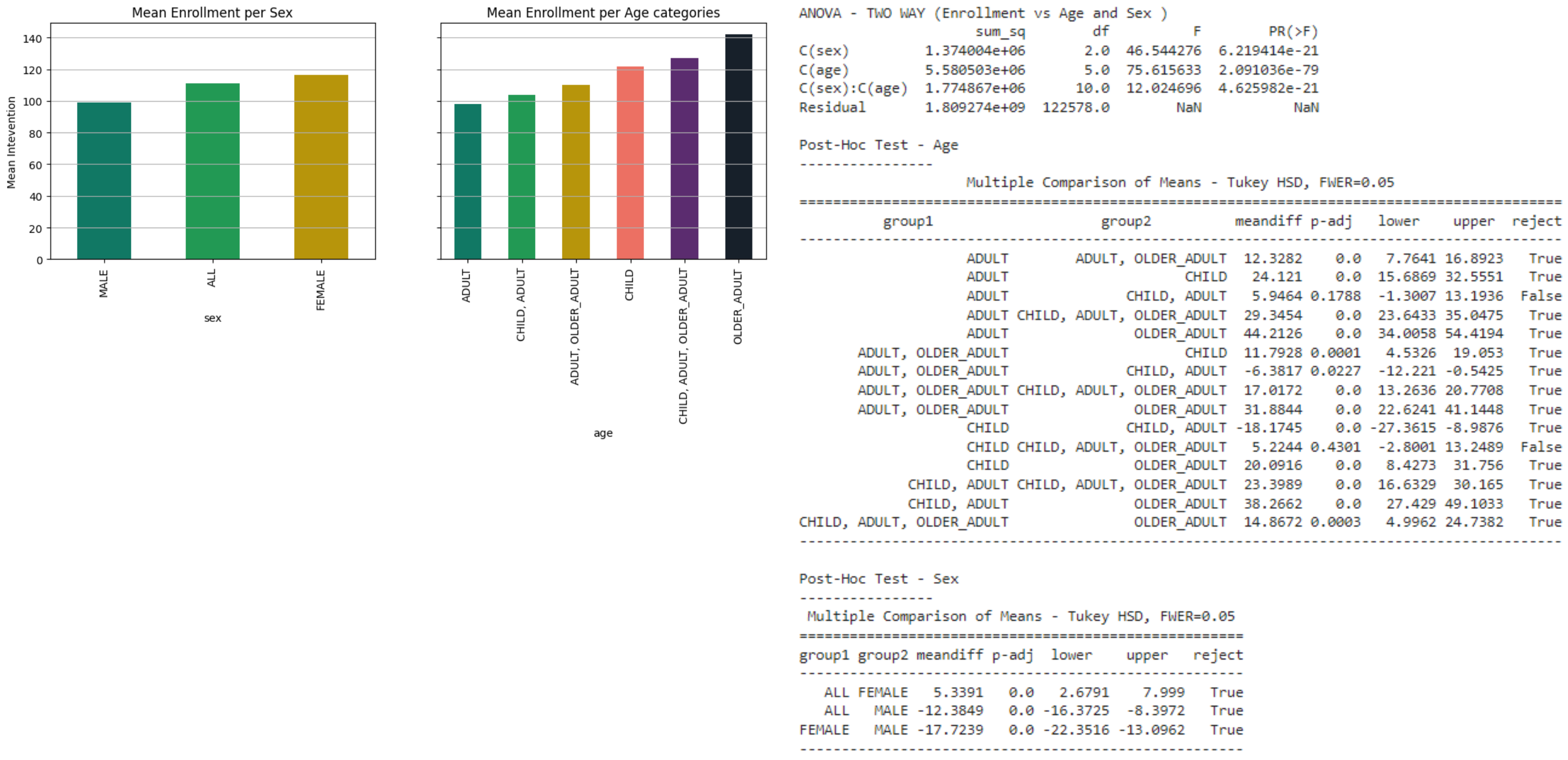
*Figure 8: ANOVA - Interventions per Study Area*



```
ANOVA - ONE WAY
-------------------
                      sum_sq        df           F  PR(>F)
C(study_area)   11565.436786       5.0  818.958946     0.0
Residual       435274.817670  154111.0         NaN     NaN

Post-Hoc Test
----------------
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
========================================================
 group1   group2  meandiff p-adj  lower   upper  reject
--------------------------------------------------------
    HIV    cancer   0.2377    0.0  0.1812  0.2942   True
    HIV     covid  -0.3028    0.0 -0.3771 -0.2284   True
    HIV     heart  -0.3925    0.0 -0.4533 -0.3317   True
    HIV   malaria   0.1918 0.0021  0.0475   0.336   True
    HIV pneumonia  -0.2769    0.0 -0.3507  -0.203   True
 cancer     covid  -0.5404    0.0 -0.5934 -0.4874   True
 cancer     heart  -0.6302    0.0 -0.6615 -0.5989   True
 cancer   malaria  -0.0459 0.9267 -0.1804  0.0886  False
 cancer pneumonia  -0.5145    0.0 -0.5669 -0.4622   True
  covid     heart  -0.0897 0.0001 -0.1473 -0.0322   True
  covid   malaria   0.4945    0.0  0.3516  0.6374   True
  covid pneumonia   0.0259 0.9057 -0.0453  0.0971  False
  heart   malaria   0.5843    0.0  0.4479  0.7206   True
  heart pneumonia   0.1156    0.0  0.0587  0.1725   True
malaria pneumonia  -0.4686    0.0 -0.6113  -0.326   True
--------------------------------------------------------
```

### 4) *Are there differences in enrollment across sex and age?*

*Figure 9: Enrollment per Sex and Age Categories*



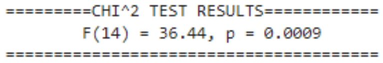
```
ANOVA - TWO WAY (Enrollment vs Age and Sex )
                     sum_sq        df          F        PR(>F)
C(sex)         1.374004e+06       2.0  46.544276  6.219414e-21
C(age)         5.580503e+06       5.0  75.615633  2.091036e-79
C(sex):C(age)  1.774867e+06      10.0  12.024696  4.625982e-21
Residual       1.809274e+09  122578.0        NaN           NaN

Post-Hoc Test - Age
----------------
                  Multiple Comparison of Means - Tukey HSD, FWER=0.05
=======================================================================================
         group1                  group2          meandiff p-adj   lower    upper  reject
---------------------------------------------------------------------------------------
                   ADULT       ADULT, OLDER_ADULT  12.3282    0.0   7.7641 16.8923   True
                   ADULT                    CHILD   24.121    0.0  15.6869 32.5551   True
                   ADULT             CHILD, ADULT   5.9464 0.1788  -1.3007 13.1936  False
                   ADULT CHILD, ADULT, OLDER_ADULT  29.3454    0.0  23.6433 35.0475   True
                   ADULT              OLDER_ADULT  44.2126    0.0  34.0058 54.4194   True
      ADULT, OLDER_ADULT                    CHILD  11.7928 0.0001   4.5326  19.053   True
      ADULT, OLDER_ADULT             CHILD, ADULT  -6.3817 0.0227  -12.221 -0.5425   True
      ADULT, OLDER_ADULT CHILD, ADULT, OLDER_ADULT  17.0172    0.0  13.2636 20.7708   True
      ADULT, OLDER_ADULT              OLDER_ADULT  31.8844    0.0  22.6241 41.1448   True
                   CHILD             CHILD, ADULT -18.1745    0.0 -27.3615 -8.9876   True
                   CHILD CHILD, ADULT, OLDER_ADULT   5.2244 0.4301  -2.8001 13.2489  False
                   CHILD              OLDER_ADULT  20.0916    0.0   8.4273  31.756   True
            CHILD, ADULT CHILD, ADULT, OLDER_ADULT  23.3989    0.0  16.6329  30.165   True
            CHILD, ADULT              OLDER_ADULT  38.2662    0.0   27.429 49.1033   True
CHILD, ADULT, OLDER_ADULT              OLDER_ADULT  14.8672 0.0003   4.9962 24.7382   True
---------------------------------------------------------------------------------------

Post-Hoc Test - Sex
----------------
 Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj  lower    upper   reject
-----------------------------------------------------
   ALL FEMALE   5.3391   0.0   2.6791    7.999   True
   ALL   MALE -12.3849   0.0 -16.3725  -8.3972   True
FEMALE   MALE -17.7239   0.0 -22.3516 -13.0962   True
-----------------------------------------------------
```

Figure 9 shows bar charts and ANOVA tests conducted to show if there are differences in means between the interaction of Age and sex in enrollment. The findings show that Age (p = 6.26e-21), Sex (p = 2.09e-79), and Interaction between Sex and Age (p = 4.63e-21) have differences in means.

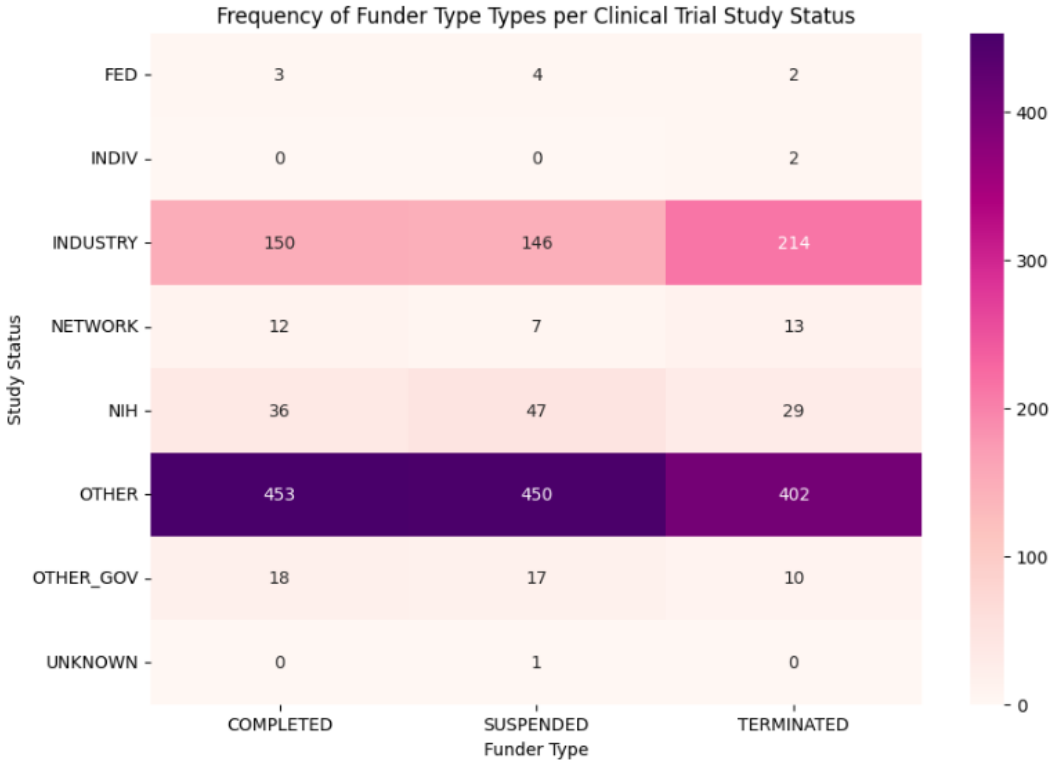
To check the significance of the results, the post hoc tests show that all group pairs show differences in means except for categories in the age variable; there were no differences in the means of CHILD and ADULT enrollment and CHILD and CHILD, ADULT, OLDER_ADULT group.

***5) Does the funder type affect the study status?***

*Figure 10: Chi^2 and Contingency Table (Funder type and Clinical Trial Status)*



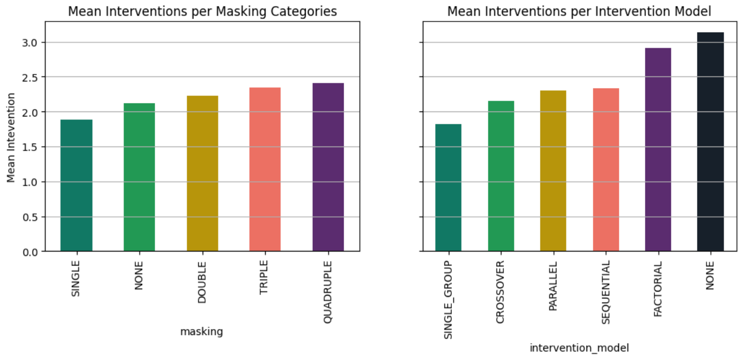
```
=========CHI^2 TEST RESULTS============
        F(14) = 36.44, p = 0.0009
=======================================
```

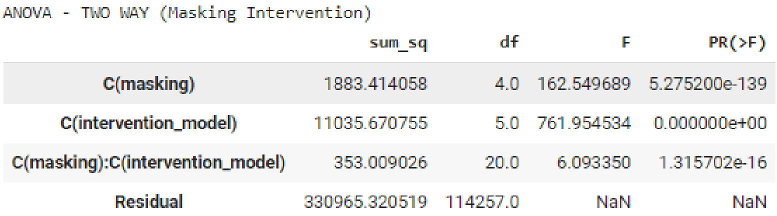


The study status was resampled to ensure even frequency in all categories. Chi-square findings from the analysis (F(14) = 36.44, p = 0.0009) imply that funder type actually affects study status given the significance of the results. From the contingency table, the majority of the trials that were terminated received had the largest number of funders in the Industry category while completed trials had the largest number of funders in the “other” category. The NIH category had the most funders in the Suspended category.

6) ***Are there differences in the number of interventions used in masking type and intervention model categories?***

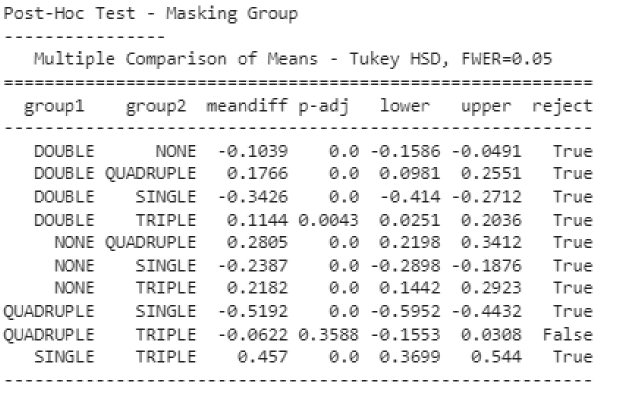
*Figure 11: Masking Type and Intervention Model - ANOVA*



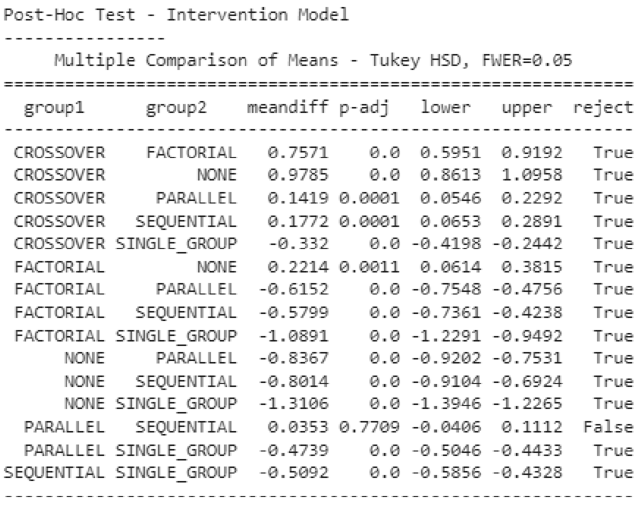


ANOVA - TWO WAY (Masking Intervention)

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| C(masking) | 1883.414058 | 4.0 | 162.549689 | 5.275200e-139 |
| C(intervention_model) | 11035.670755 | 5.0 | 761.954534 | 0.000000e+00 |
| C(masking):C(intervention_model) | 353.009026 | 20.0 | 6.093350 | 1.315702e-16 |
| Residual | 330965.320519 | 114257.0 | NaN | NaN |



```
Post-Hoc Test - Masking Group
---------------
   Multiple Comparison of Means - Tukey HSD, FWER=0.05
==========================================================
  group1    group2  meandiff p-adj   lower   upper  reject
----------------------------------------------------------
   DOUBLE      NONE  -0.1039    0.0 -0.1586 -0.0491   True
   DOUBLE QUADRUPLE   0.1766    0.0  0.0981  0.2551   True
   DOUBLE    SINGLE  -0.3426    0.0  -0.414 -0.2712   True
   DOUBLE    TRIPLE   0.1144 0.0043  0.0251  0.2036   True
     NONE QUADRUPLE   0.2805    0.0  0.2198  0.3412   True
     NONE    SINGLE  -0.2387    0.0 -0.2898 -0.1876   True
     NONE    TRIPLE   0.2182    0.0  0.1442  0.2923   True
QUADRUPLE    SINGLE  -0.5192    0.0 -0.5952 -0.4432   True
QUADRUPLE    TRIPLE  -0.0622 0.3588 -0.1553  0.0308  False
   SINGLE    TRIPLE    0.457    0.0  0.3699   0.544   True
----------------------------------------------------------
```



```
Post-Hoc Test - Intervention Model
---------------
     Multiple Comparison of Means - Tukey HSD, FWER=0.05
==============================================================
  group1      group2    meandiff p-adj   lower   upper  reject
--------------------------------------------------------------
 CROSSOVER    FACTORIAL   0.7571    0.0  0.5951  0.9192   True
 CROSSOVER         NONE   0.9785    0.0  0.8613  1.0958   True
 CROSSOVER     PARALLEL   0.1419 0.0001  0.0546  0.2292   True
 CROSSOVER   SEQUENTIAL   0.1772 0.0001  0.0653  0.2891   True
 CROSSOVER SINGLE_GROUP   -0.332    0.0 -0.4198 -0.2442   True
 FACTORIAL         NONE   0.2214 0.0011  0.0614  0.3815   True
 FACTORIAL     PARALLEL  -0.6152    0.0 -0.7548 -0.4756   True
 FACTORIAL   SEQUENTIAL  -0.5799    0.0 -0.7361 -0.4238   True
 FACTORIAL SINGLE_GROUP  -1.0891    0.0 -1.2291 -0.9492   True
      NONE     PARALLEL  -0.8367    0.0 -0.9202 -0.7531   True
      NONE   SEQUENTIAL  -0.8014    0.0 -0.9104 -0.6924   True
      NONE SINGLE_GROUP  -1.3106    0.0 -1.3946 -1.2265   True
  PARALLEL   SEQUENTIAL   0.0353 0.7709 -0.0406  0.1112  False
  PARALLEL SINGLE_GROUP  -0.4739    0.0 -0.5046 -0.4433   True
SEQUENTIAL SINGLE_GROUP  -0.5092    0.0 -0.5856 -0.4328   True
--------------------------------------------------------------
```

The findings from the table above Masking type (p = 5.28e-139), intervention model (p < 0.005), masking:intervention model (p = 1.32e-16) are significant showing that there are differences in mean interventions used between the different groups. To determine the significance of the results, all pairs were true to the findings except the

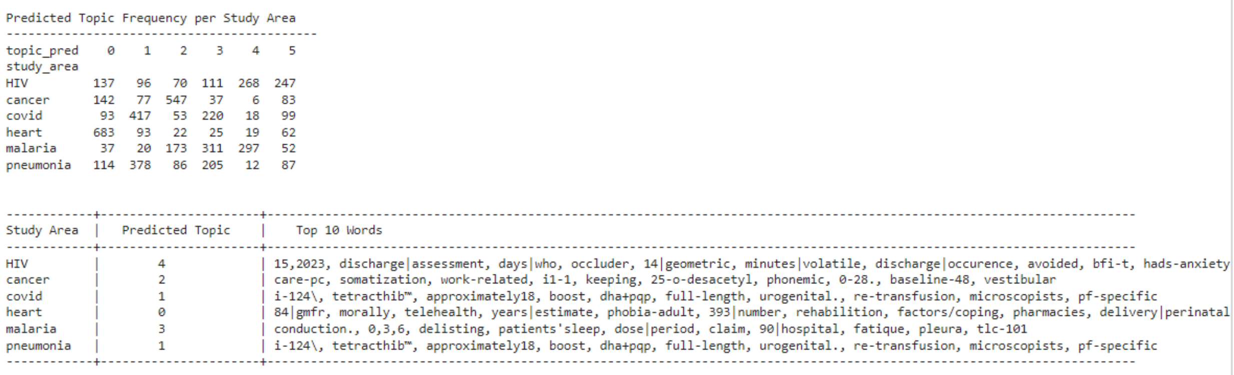
QUADRUPLE and TRIPLE pair in masking type and PARALLEL and SEQUENTIAL pair in intervention models

**Topic Modelling**

Topic modeling was done using LDA. Figure 12 shows 2 different tables. The first one shows the frequency of the predicted topic per study area while the second one shows the 10 most weighted words per topic and the study area that had the highest frequency in the first crosstab.

Selecting the most frequent topic per study area, we underscored that Pneumonia and COVID shared a single topic (1) while HIV, Cancer, Heart, and Malaria assumed topics 4, 2, 0, and 3 respectively.

*Figure 12: Topic frequency per study area and most frequent words per topic*



```
Predicted Topic Frequency per Study Area
-------------------------------------------
topic_pred    0    1    2    3    4    5
study_area
HIV         137   96   70  111  268  247
cancer      142   77  547   37    6   83
covid        93  417   53  220   18   99
heart       683   93   22   25   19   62
malaria      37   20  173  311  297   52
pneumonia   114  378   86  205   12   87


------------+----------------------+------------------------------------------------------------------------------------------------
Study Area  |    Predicted Topic   |    Top 10 Words
------------+----------------------+------------------------------------------------------------------------------------------------
HIV         |         4            | 15,2023, discharge|assessment, days|who, occluder, 14|geometric, minutes|volatile, discharge|occurence, avoided, bfi-t, hads-anxiety
cancer      |         2            | care-pc, somatization, work-related, i1-1, keeping, 25-o-desacetyl, phonemic, 0-28., baseline-48, vestibular
covid       |         1            | i-124\, tetracthib™, approximately18, boost, dha+pqp, full-length, urogenital., re-transfusion, microscopists, pf-specific
heart       |         0            | 84|gmfr, morally, telehealth, years|estimate, phobia-adult, 393|number, rehabilition, factors/coping, pharmacies, delivery|perinatal
malaria     |         3            | conduction., 0,3,6, delisting, patients'sleep, dose|period, claim, 90|hospital, fatique, pleura, tlc-101
pneumonia   |         1            | i-124\, tetracthib™, approximately18, boost, dha+pqp, full-length, urogenital., re-transfusion, microscopists, pf-specific
------------+----------------------+------------------------------------------------------------------------------------------------
```

To help characterize primary themes, we investigated the 10 most frequent words per topic. In the HIV study area, the top words for predicted topic 4 include "discharge assessment," "days," "who," "occluder," and "geometric." These words suggest that documents related to this topic likely focus on these themes.

In the cancer study area, predicted topic 2 is characterized by words like "care-pc," "somatization," "work-related," and "phonemic," indicating potential areas of interest in cancer-related trials.

In the COVID and Pneumonia study area, predicted topic 1 includes words such as "tetracthib™," "boost," "dha+pqp," and "microscopists," which might suggest a focus on specific interventions and measurements.

The top words for the heart study area's predicted topic 0 include "84|gmfr," "telehealth," "years," and "estimate," suggesting themes related to telehealth, estimates, and time-related factors.

## Conclusion and Recommendations

### Conclusion

In this comprehensive analysis of clinical trials data, several critical findings have been highlighted. The study began by addressing anomalies and outliers in the data. Notably, some trials had no enrollment, and the presence of certain outliers required attention. However, these anomalies were often related to specific study areas and could be attributed to data ingestion errors.

Descriptive analytics revealed important trends in clinical trial enrollments over time, showcasing an overall upward trajectory. Notable peaks were observed in 2002, 2019, 2020, and 2021, with the latter three years being heavily influenced by the COVID-19 pandemic. Furthermore, the distribution of days to completion, enrollment, and the number of interventions per trial were examined. The findings indicated that these features were right-skewed, with the intervention count being heavily skewed. Time to completion exhibited multimodality, with most trials taking between 700 and 1400 days to finish.

Inferential analytics addressed specific research questions. Actionable results included differences in masking types across study areas, with specific masking types being prevalent in different areas. Collaborator availability was also investigated, but it was found not to significantly affect study status. Differences in the number of interventions across study areas were identified, with heart studies receiving the fewest interventions and cancer studies the most. Furthermore, the analysis revealed differences in enrollment based on age, sex, and their interaction. Age, sex, and the interaction between age and sex were all found to have significant differences in means.

Additionally, funder type was also found to affect study status, with the majority of terminated trials having the most funders in the Industry category, while completed trials had more funders in the "other" category. Differences in the number of interventions used in masking type and intervention model categories were identified, with significant variations in means between different groups.

Finally, topic modeling using Latent Dirichlet Allocation (LDA) identified the primary themes within each study area, shedding light on the content and focus of clinical trials in each domain.

**Recommendations**

➔ **Data Quality Assurance:** Given anomalies and outliers observed in the data, it is essential to implement robust data quality assurance processes during data collection and ingestion. This can help minimize errors and ensure data integrity.

➔ **Investigate Enrollment Trends:** Continue to monitor trends in clinical trial enrollments, especially in response to significant events like pandemics. Understanding these trends can inform resource allocation and research priorities.

- **Research Focus:** The thematic analysis of clinical trials revealed varying research focuses in different study areas. Thus, researchers should consider these themes when designing and prioritizing future clinical trials.
- **Collaboration:** Collaborator availability was found to differ across study areas. Encouraging collaboration among research groups may lead to more balanced research efforts and improved resource allocation.
- **Funder Type and Study Status:** The influence of funder type on study status should be further investigated to understand the underlying factors and implications. This can help in optimizing funding strategies.
- **Masking Types and Intervention Models:** The differences in interventions and masking types in clinical trials may indicate varying approaches across study areas. Understanding the reasons behind these variations can inform best practices.
- **Topic Modeling Insights:** The identified topics can guide researchers and policymakers in prioritizing areas of healthcare research and development based on current trends and needs.

## References

Bensken, W. P., Pieracci, F. M., & Ho, V. P. (2021). Basic Introduction to Statistics in Medicine, Part 1: Describing Data. *Surgical Infections*, *22*(6), 590–596. https://doi.org/10.1089/sur.2020.429

Biswal, A. (2023). What is a Chi-Square Test? Formula, Examples & Application. *Simplilearn.com*. https://www.simplilearn.com/tutorials/statistics-tutorial/chi-square-test

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

*ClinicalTrials.gov*. (2021, December 8). Learn About Studies. Retrieved October 20, 2023, from https://clinicaltrials.gov/study-basics/learn-about-studies

Kim, T. K. (2017). Understanding one-way ANOVA using conceptual figures. *Korean Journal of Anesthesiology*, *70*(1), 22. https://doi.org/10.4097/kjae.2017.70.1.22

Novitzke, J. M. (2008, January 1). *The significance of clinical trials*. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3317309/

Seyhan, A. A., & Carini, C. (2019). Are innovation and new technologies in precision medicine paving a new era in patients centric care? *Journal of Translational Medicine*, *17*(1). https://doi.org/10.1186/s12967-019-1864-9