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

Presented By: TEAM 5

1. Kelvin Kiplagat
2. Thomas Muriuki

Colab Code (link) : DTE Team 5 Phase II.ipynb

Google Colaboratory: Python Version 3.10.12

Powered by DTE & Mentor Owuonda Julius

Observations | Conclusion

Our Observations

- Outliers and anomalies were present. Some trials had no enrollment figures yet others were completed others had more than a million.
- Trend-wise, enrollment peaks were observed in 2002, 2019, 2020, and 2021 - COVID-19 and SARS-COV2
- Time to completion for clinical trials took between 700 and 1400 days to complete.
- We discovered that different masking types were common in different study areas.
- We discovered Collaborator availability was found not to affect study status..
- We found that the Number of interventions used per study area were different. However few category pairs illustrated no difference. For example Covid and Pneumonia had no differences in the number of interventions used because they are nearly similar.
- Age, sex, and the interaction between age and sex were all found to have significant differences in mean number of interventions.
- Masking type, Intervention model, and the interaction between masking type, and intervention model were all found to have significant differences in the mean number of interventions used.
- Funder type was found to affect study status. We found that clinical trials funded by the industry were more likely to be terminated.
- We came up with a model that identified the primary themes within each study area (diseases), shedding light on the content and focus of clinical trials.

Conclusion

- Given anomalies and outliers observed in the data, it is essential to implement robust data quality assurance processes during data collection to help maintain data integrity.
- There should be Continuous monitoring of trends in clinical trial enrollments, especially in response to significant events like pandemics.
- Encouraging collaboration among research groups will lead to more balanced research efforts and improved resource allocation to mitigate differences in collaborator availability.
- The funder type Influences the success of a research status. This means that the amount of resources a clinical trial receives affects the general outcome of the research.
- Differences in interventions and masking types in clinical trials may indicate varying approaches across study areas. This means different research methodologies should be used to avoid bias.

Introduction

Problem Statement

- 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.
- 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 well as thematic/topic modeling analysis

Research Objective

- 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.

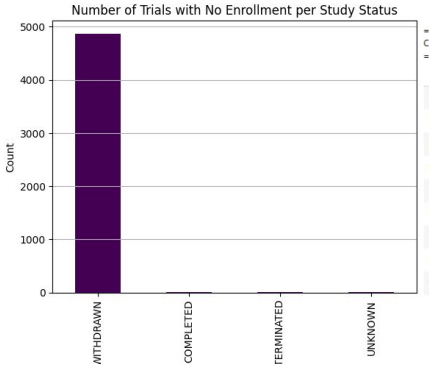
Research Questions

- Are there masking types common in different study areas than others?
- 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?
- Does the number of interventions differ in different study areas?
- Are there differences in enrollment across sex and age?
- Does the funder type affect the study status?
- Are there differences in the number of interventions used in masking type and intervention model categories?

Chi-Square Tests, Analysis of Variance (ANOVA) together with Post-Hoc tests (Tukey HSD) and topic modelling using Latent Dirichlet Allocation were used to gain the relevant answers to these questions.

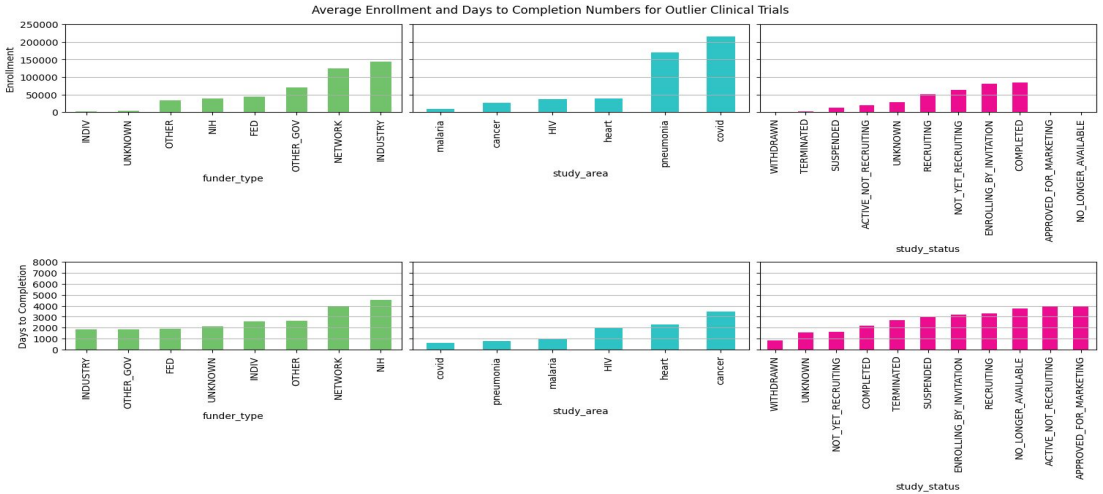
Data Exploration|Cleaning

- Originally 161863 observations (clinical trials) from all the six study areas (datasets). 159008 actionable trials remained after removing trials with no start date and those that were to begin in 2024 going forward.
- Study areas were; HIV, Malaria, Heart, Cancer, COVID, Pneumonia.
- Some trials had no enrollment, however majority of them had been **withdrawn** as seen on the right. although, there were **completed** trials with no enrollment. Data ingestion errors? Probably.
- NIH was the common funder type for completed trials with no enrollment.
- There were heavily pronounced outliers. Outliers were derived from time taken to completion (days) and enrollment features.
- COVID-19 trials had the highest enrollment figures treated as outliers on average
- Other outliers exploration can be seen on the visualizations to the right.

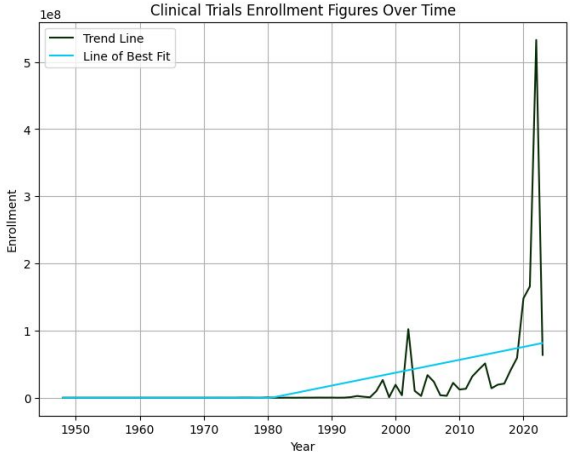


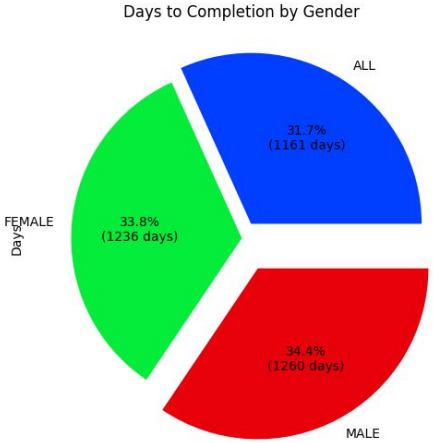
Complete Clinical Trials Data with No Enrollment

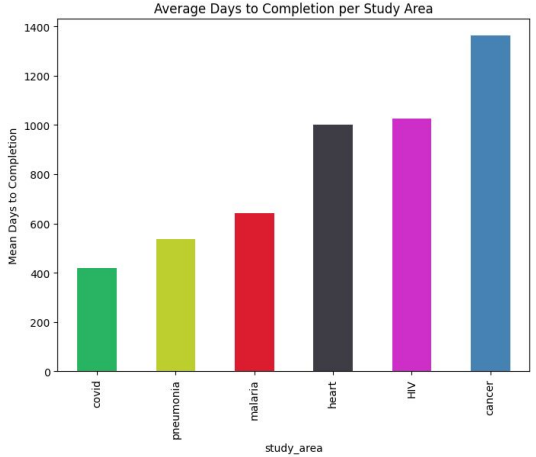
	sponsor	study_results	study_area	start_date	completion_date	funder_type
110182	ViiV Healthcare	NO	heart	2009-03-01	2010-12-01	INDUSTRY
128145	National Heart, Lung, and Blood Institute (NHLBI)	NO	heart	2007-01-11	2007-12-10	NIH
141951	ViiV Healthcare	NO	HIV	2009-03-01	2010-12-01	INDUSTRY
144401	National 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

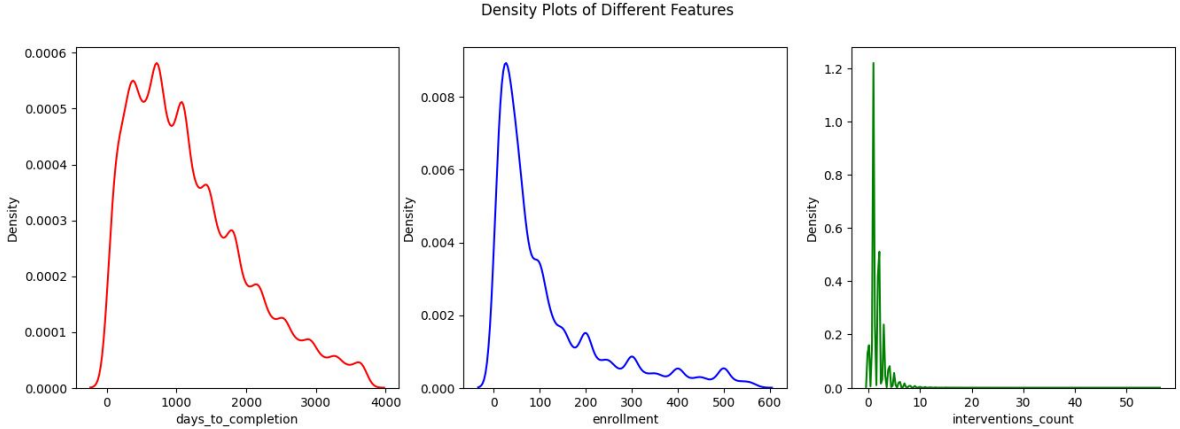


Descriptive Visualizations







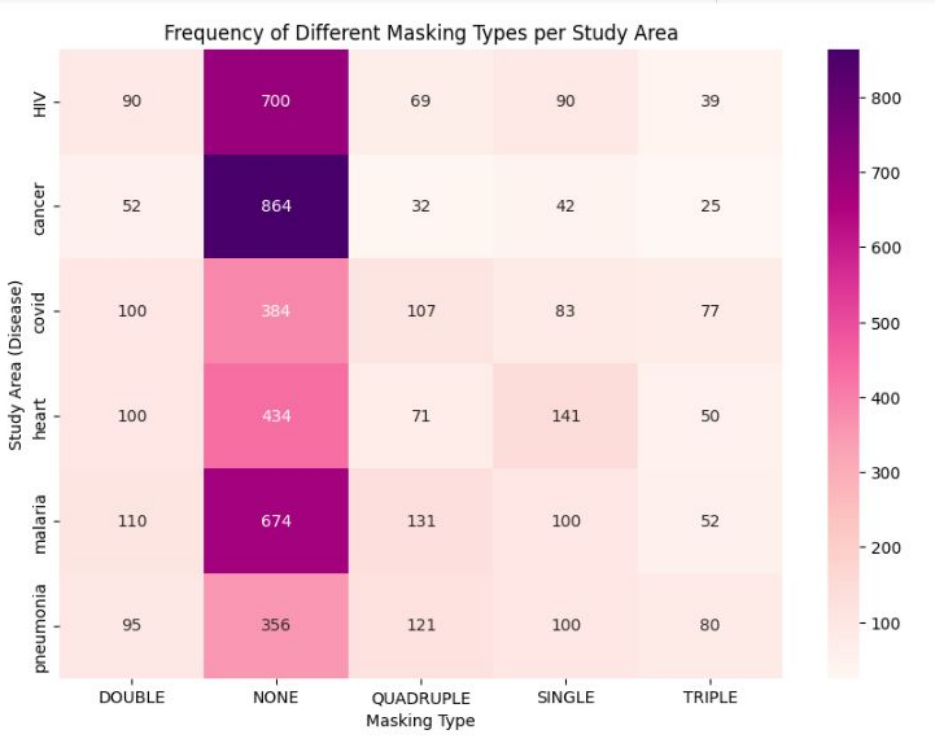


- ➔ Clinical trial numbers have been increasing over time.
- ➔ 2002, 2019, 2020, and 2021 saw concerning increases.
- ➔ Trial time to completion, enrollments, number of interventions are skewed to the right.
- ➔ Distribution of days to completion exhibits multimodality.
- ➔ Male trials take the most number of days to completion. All sex trials take least days.
- ➔ However, differences sex-wise are miniscule.
- ➔ Cancer takes the most number of days by far.
- ➔ COVID took the least number of days.

Inferential Statistics

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

- The Chi-square findings below the Contingency table are significant.
- 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




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

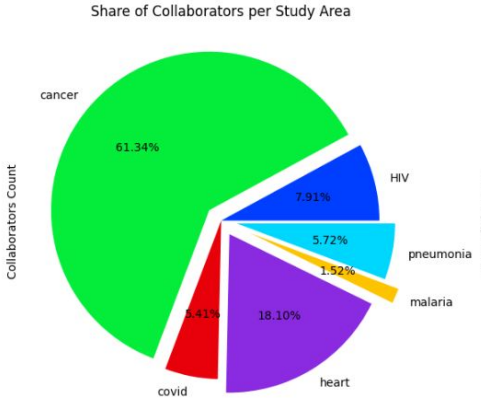
Inferential Statistics

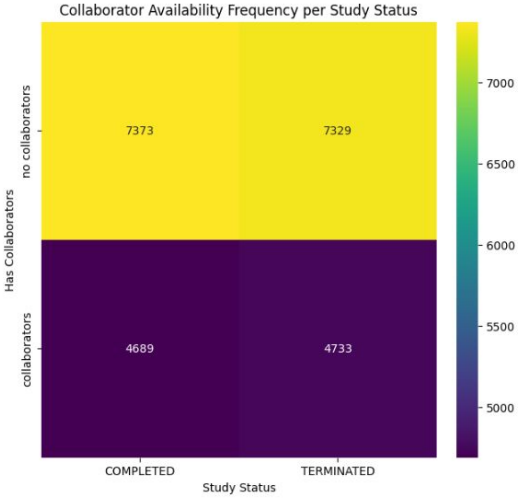
2 a) Are there differences in the number of collaborators different study areas get?

- 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.
- However, these disparities may not be representative because of the differences in samples per study area.

2 b) Does the availability of collaborators affect study status in completed, terminated, and suspended studies?

- The Chi-square findings are not significant at $\alpha = 0.05$.
- 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.






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

Inferential Statistics

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

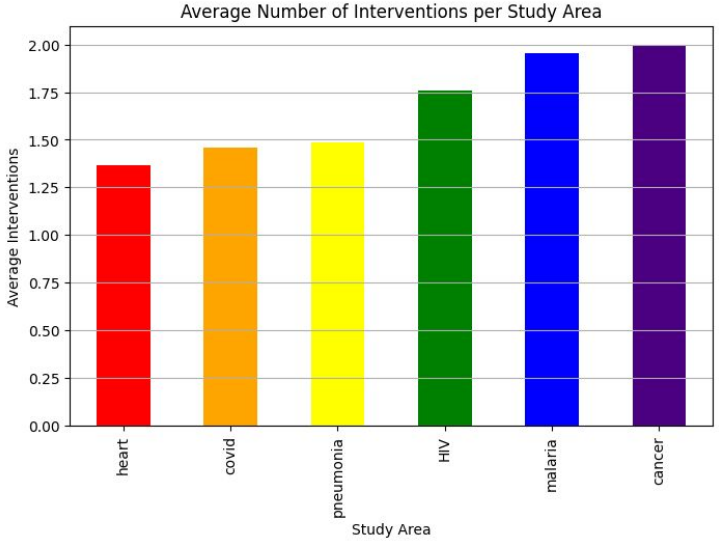
```
ANOVA - ONE WAY
-------------------
                   sum_sq        df           F  PR(>F)
C(study_area)   11565.436786       5.0  818.958946     0.0
Residual       435274.817670  154111.0         NaN     NaN

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
-----------------------------------------------------
```

- The findings from the one-way ANOVA ($F = 818.96$, $p = 0.0000$) show that the number of interventions used in different study areas differs in means.

- The significance of the findings and the post-hoc tests supports majority of the differences seen in the bar chart with heart studies receiving the least number of interventions on average while Cancer received the most number of interventions.

- However, the post-hoc tests show that only cancer and malaria, and covid and pneumonia didn't have a difference in means in the number of interventions used in the trials



Inferential Statistics

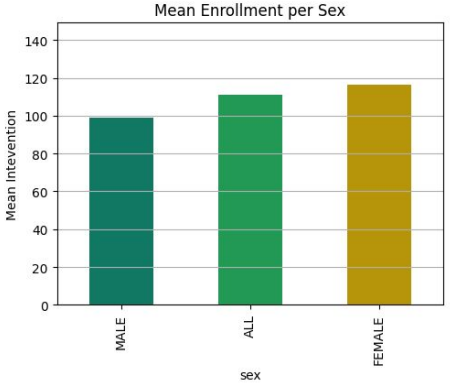
4) Are there differences in enrollment across sex and age?

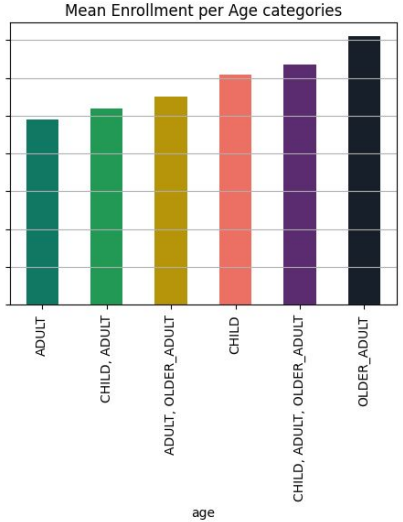
- The findings: Age (p = 6.26e-21), Sex (p = 2.09e-79), and Interaction between Sex and Age (p = 4.63e-21) are significant.
- Hence there are differences in means between majority of the categories in all the features.

```
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
----------------------------------------------------
```





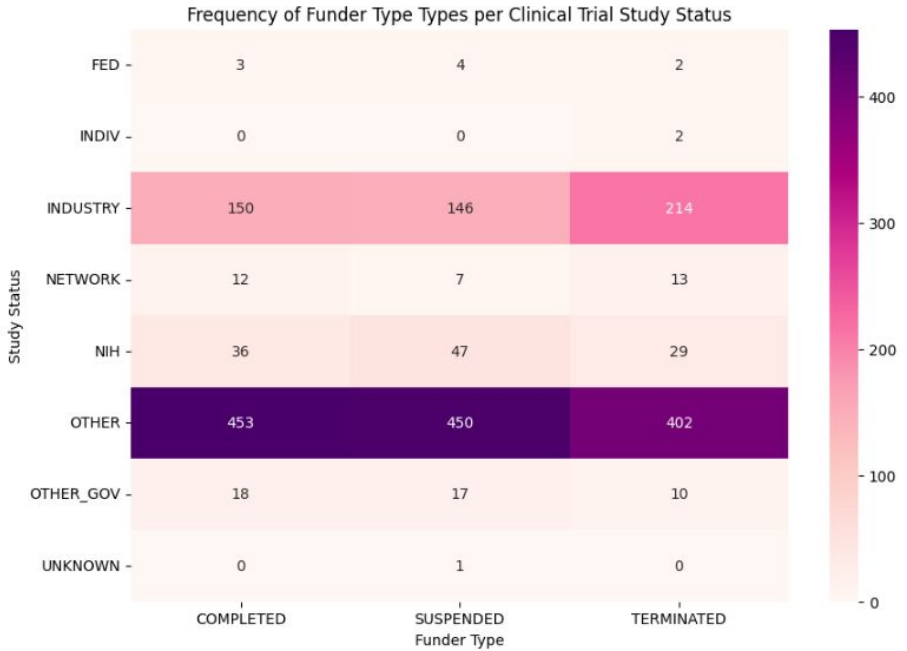
- To check the significance of the results, the post-hoc tests show that all group pairs show differences in means except for categories in the age variable; there were no differences in the means of CHILD and ADULT enrollment and CHILD and CHILD, ADULT, OLDER_ADULT group.

Inferential Statistics

5) Does the funder type affect the study status?

- Chi-square findings from the analysis ($F(14) = 36.44$, $p = 0.0009$) are significant at $\alpha = 0.05$.
- Thus funder type actually influences 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.

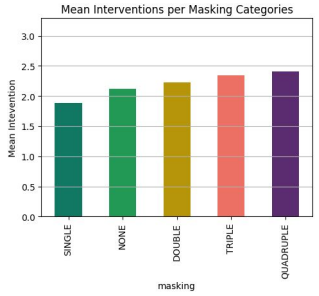
```
========CHI^2 TEST RESULTS============
        F(14) = 36.44, p = 0.0009
======================================
```

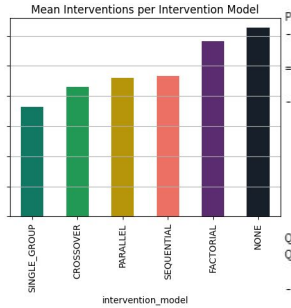


Inferential Statistics

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

- 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 category pair-wise, all pairs were true to the findings except the QUADRUPLE and TRIPLE pair in masking type and PARALLEL and SEQUENTIAL pair in intervention models





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

Topic Modelling - Thematic Analysis

Latent Dirichlet Allocation (LDA)

- 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.
- Ideally the Table to the right shows the frequency of the predicted topic per study area.
- The table below shows the top 10 words per topic and the study area that had the highest frequency in the first crosstab (table to the right).

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

Colab Code Link, Report and Data Sources

Data Source(s)		Link
	ClinicalTrials.gov	https://classic.clinicaltrials.gov/ct2/resources/download#UseURL
Colab		
	DTE Team 5 Phase II	https://colab.research.google.com/drive/1fXLmIrRFpxeAzdAH0mmEbLNNrDk0OpY5?usp=sharing
Report		
	DTE Datathon TEAM 5 Report	https://docs.google.com/document/d/1vc8SieUmyII2XHHwoVsOru4InUNnYNRvsEnGZEb0n8s/edit?usp=sharing

THAN DTE Consultancy

The place for Data, Technology and Expertise

e entirely grateful for the entire DTE Consultan *)wuonda for bein*
with us at the very step of the W

python™

ClinicalTrials.gov

Colab