Clinical Trials Data Analysis Report

1. Introduction

Clinical trials play a crucial role in advancing medical knowledge, testing new drugs, and improving healthcare outcomes. This project analyzes a dataset of registered clinical trials with the goal of uncovering patterns in trial phases, sponsor categories, enrollment sizes, and timelines.

Objective: To perform exploratory data analysis (EDA) and generate insights about the structure and trends in clinical research.

2. Methodology

The analysis followed these steps:

Data Cleaning

- Handled missing values in enrollment and phases.
- Standardized categorical variables (e.g., sponsor categories).
- Converted start_date and completion_date to datetime format.

Exploratory Data Analysis (EDA)

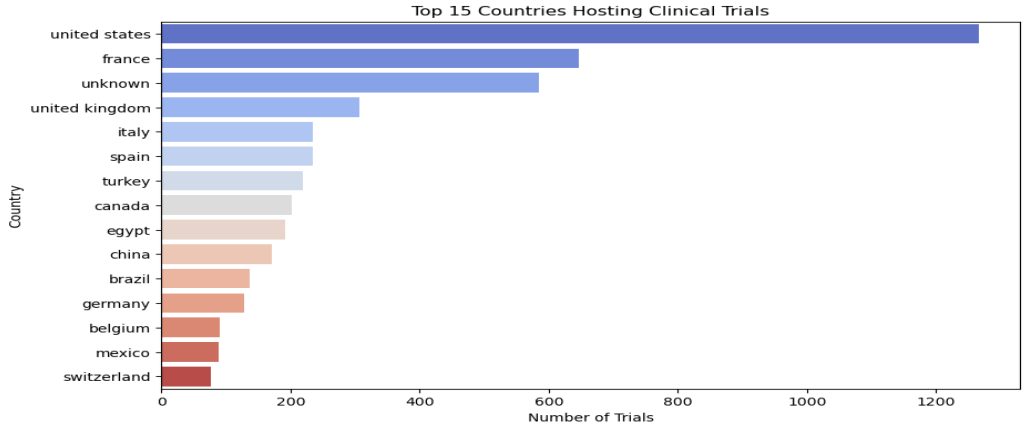
- Summarized distributions of trial phases, sponsors, and enrollment.
- Visualized trends across time.

Visualization

- Bar charts, histograms, and line plots were used to highlight patterns.
- Interpretation
- Each visual was followed by insights and possible implications.

3. Analysis & Key Insight

3.1 Top 15 Countries Hosting Clinical Trials



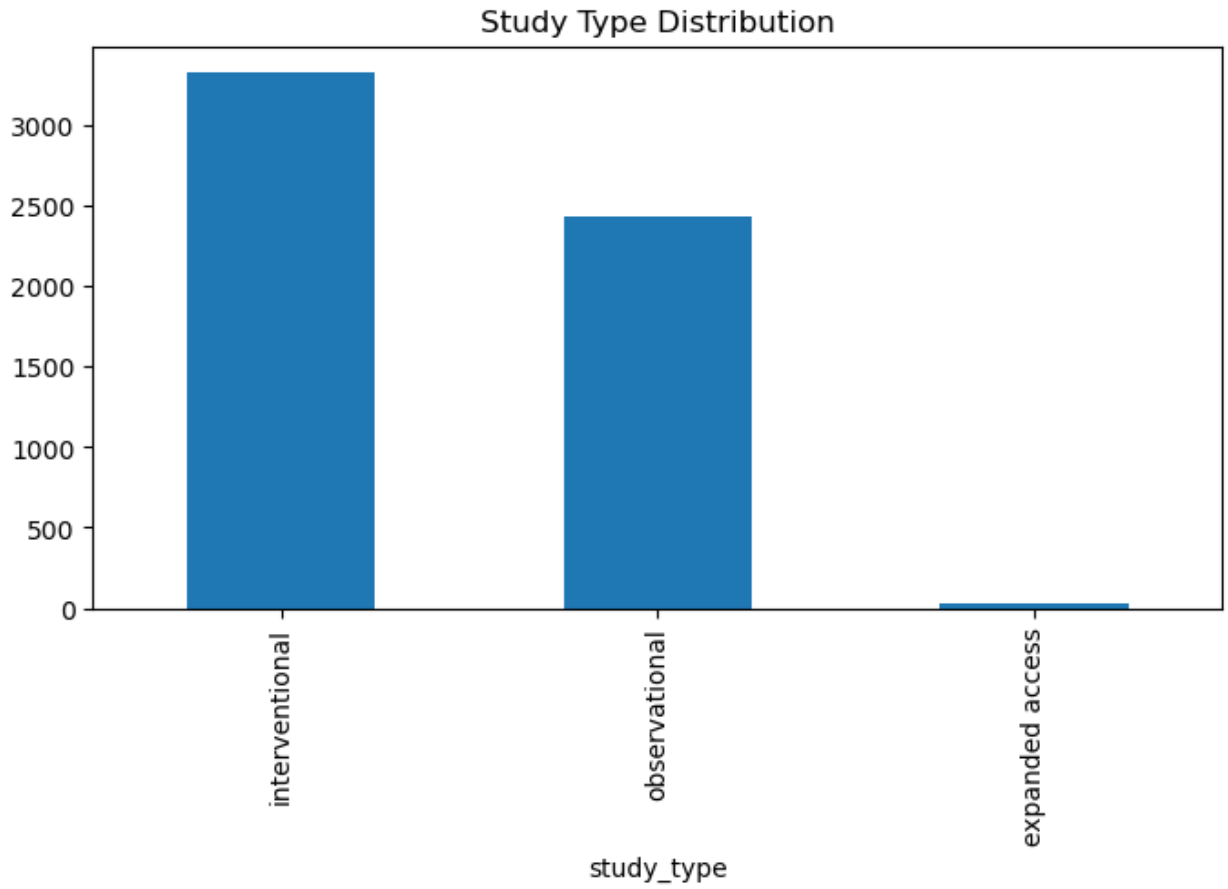
Observation:

A bar chart of the top 15 countries shows that the United States and France clearly dominate, hosting the largest share of clinical trials.

Inference:

These two countries are global leaders in clinical research infrastructure. The U.S. dominance reflects its pharmaceutical industry strength, while France's presence highlights Europe's investment in biomedical studies. Other countries have significantly smaller shares, underlining the geographic concentration of trials.

3.2 Study Type Distribution



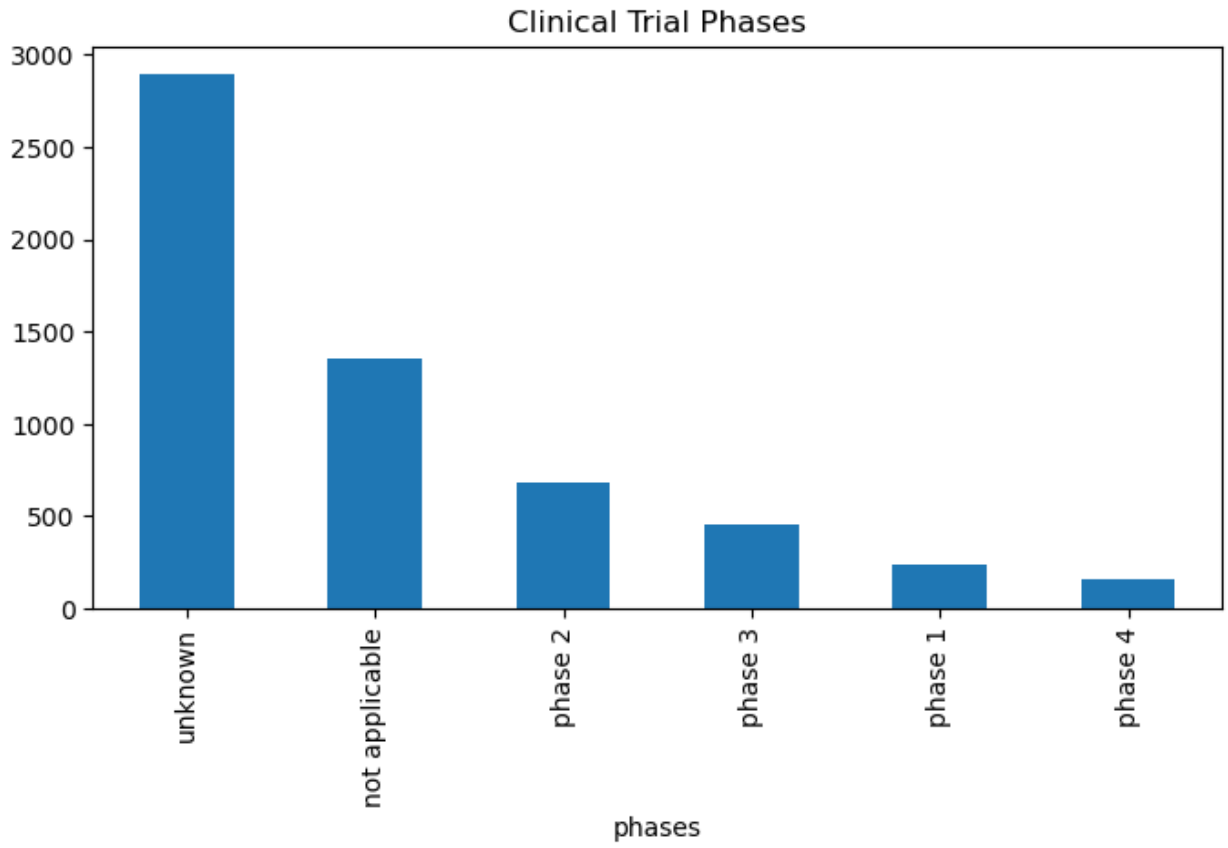
Observation:

The majority of studies are Interventional trials, followed by Observational studies. Expanded Access studies are nearly zero. Within COVID-19 research specifically, Interventional trials dominate.

Inference:

This makes sense, as Interventional designs are the primary way to test treatments in patients. Observational studies are important for secondary data collection, but play a smaller role. Expanded Access is rare because it applies to only very specific, life-threatening situations.

3.3 Trial Phases



Observation:

A large portion of trials fall under “Unknown” or “Not Applicable” categories. Among the well-defined phases, the order of prevalence is:

Phase II (most frequent)

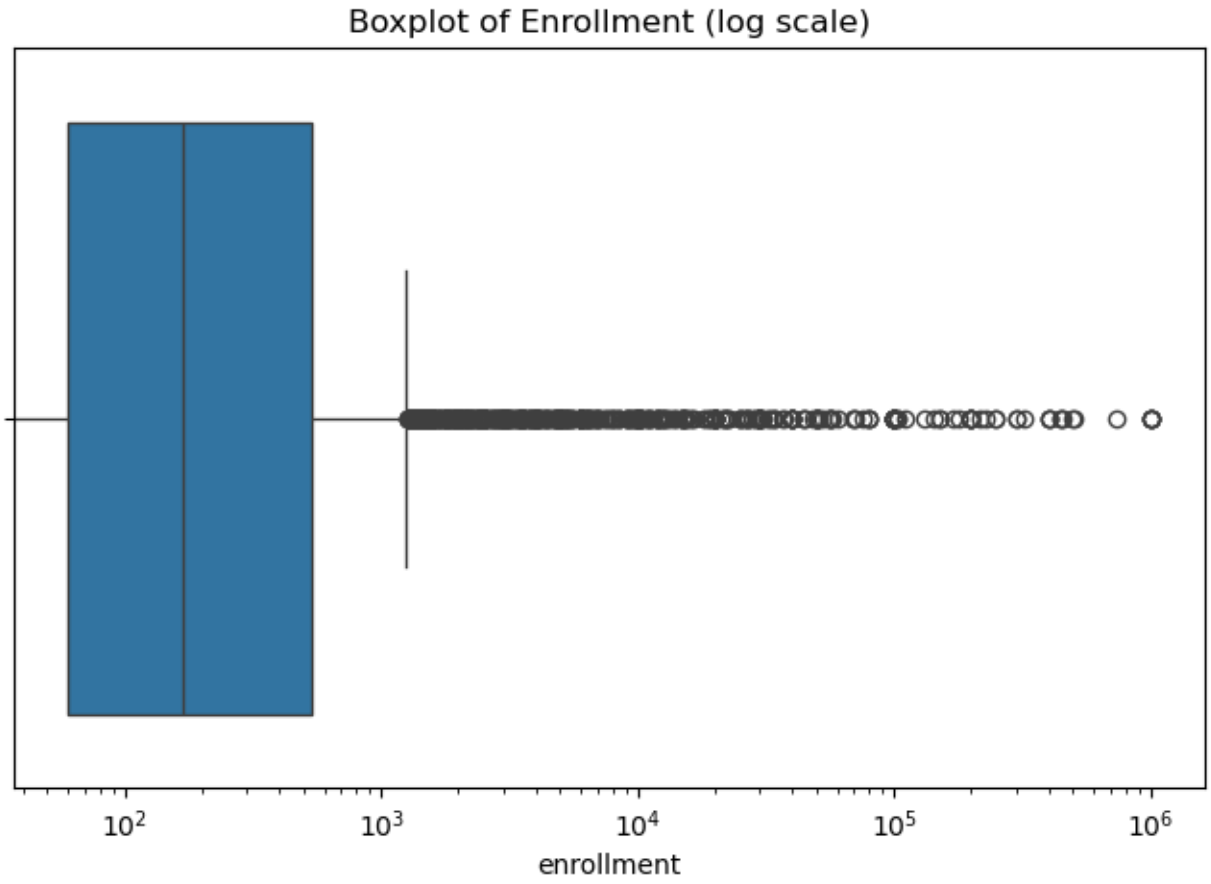
Phase IV (least frequent)

Inference:

The “Unknown/Not Applicable” values suggest incomplete reporting or non-traditional studies. Excluding those, the distribution aligns with the clinical research pipeline: most studies aim to

test efficacy and safety (Phase II & III), while relatively fewer reach long-term monitoring (Phase IV).

3.4 Enrollment Distribution



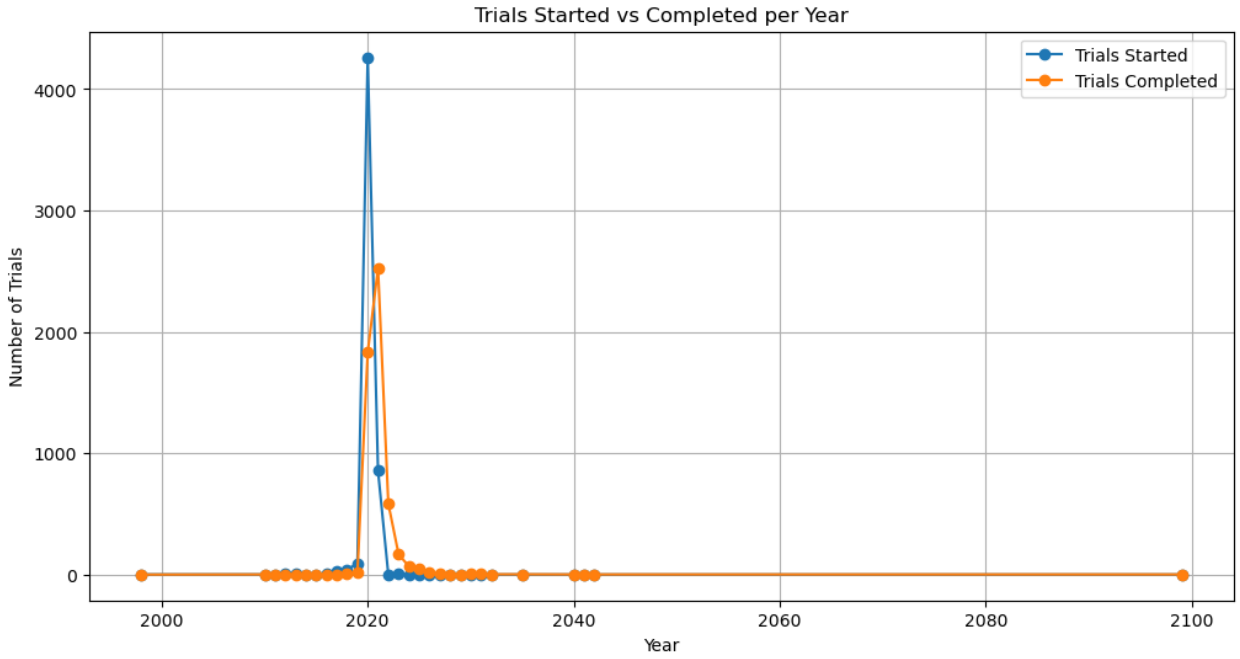
Observation:

Enrollment sizes vary widely, with most trials involving small participant groups and a few extreme outliers reaching into the thousands.

Inference:

The distribution is highly skewed, indicating that while small-scale studies are common, large trials are rare but impactful. Log-scaled plots help reveal this pattern more clearly, supporting better decisions around modeling, resource planning, and feasibility assessment.

3.5 Trials Started vs Completed per Year

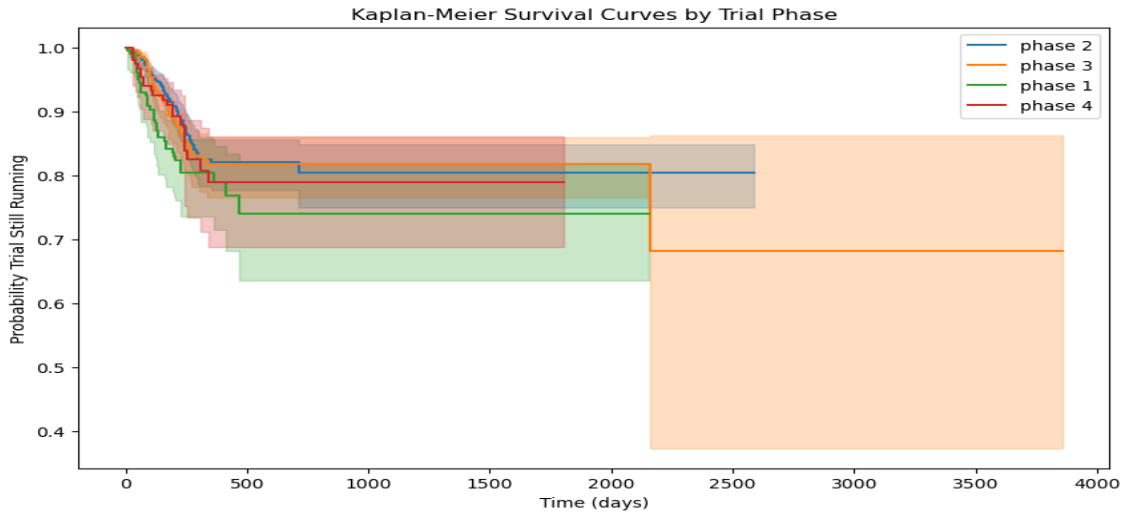


Observation:

A sharp increase in trial initiations occurs in a specific year, while the number of completed trials remains consistently low.

Inference: This disparity suggests operational bottlenecks—such as recruitment challenges, resource constraints, or regulatory delays—that prevent many trials from reaching completion despite increased launches.

3.6 Kaplan-Meier Survival Curves by Trial Phase



Observation:

Phase 3 trials show the longest duration with high survival probability—remaining above 0.4 even beyond 4000 days. Phase 2 and Phase 1 follow, with gradually shorter durations and lower probabilities. Phase 4 drops off earlier, indicating shorter trial lifespans.

Inference:

Trials in Phase 3 are more stable and tend to run longer, likely due to stronger funding, clearer protocols, and regulatory momentum. Earlier phases (1 and 2) show moderate longevity, while Phase 4 trials may be more targeted or post-approval, leading to shorter durations. This pattern reflects how trial phase impacts operational lifespan.

3.7 Hypothesis Test : Trials Started vs. Trials Completed

Research Question

Do the number of trials started in a given year significantly differ from the number of trials completed, and are they correlated over time?

Methodology

Paired t-test was used to compare yearly counts of trials started vs. completed.

Normality of differences was checked using Shapiro–Wilk test.

As the normality assumption was violated, a non-parametric Wilcoxon signed-rank test was also performed.

Pearson correlation was used to measure association between the two trends.

Results

t-test: $t = 0.0$, $p = 1.0$ → Fail to reject H_o.

Wilcoxon test: $p = 0.293$ → Fail to reject H_o.

Shapiro–Wilk test: $p \ll 0.05$ → Differences not normally distributed.

Correlation: $r = 0.70$, $p < 0.001$ → Strong positive correlation.

Inference

There is no significant difference between the number of trials started and completed per year. On average, they balance out.

However, there is a strong positive correlation: when more trials are initiated in a year, more are also completed, indicating synchronized trends in trial activity.

3.8 Hypothesis Test :Enrollment Sizes Across Trial Statuses

Research Question

Do enrollment sizes differ significantly across clinical trial statuses?

Methodology

Null hypothesis (H_o): Enrollment distributions are the same across all trial statuses.

Alternative hypothesis (H_1): At least one trial status has a significantly different enrollment distribution.

Because enrollment data were highly skewed with extreme outliers, ANOVA assumptions (normality, homogeneity of variance) were violated.

A non-parametric Kruskal–Wallis H-test was used instead.

Results

Kruskal–Wallis test: $H = 139.91$, $p \approx 1.05 \times 10^{-27}$.

Since $p \ll 0.05$, H_o is rejected.

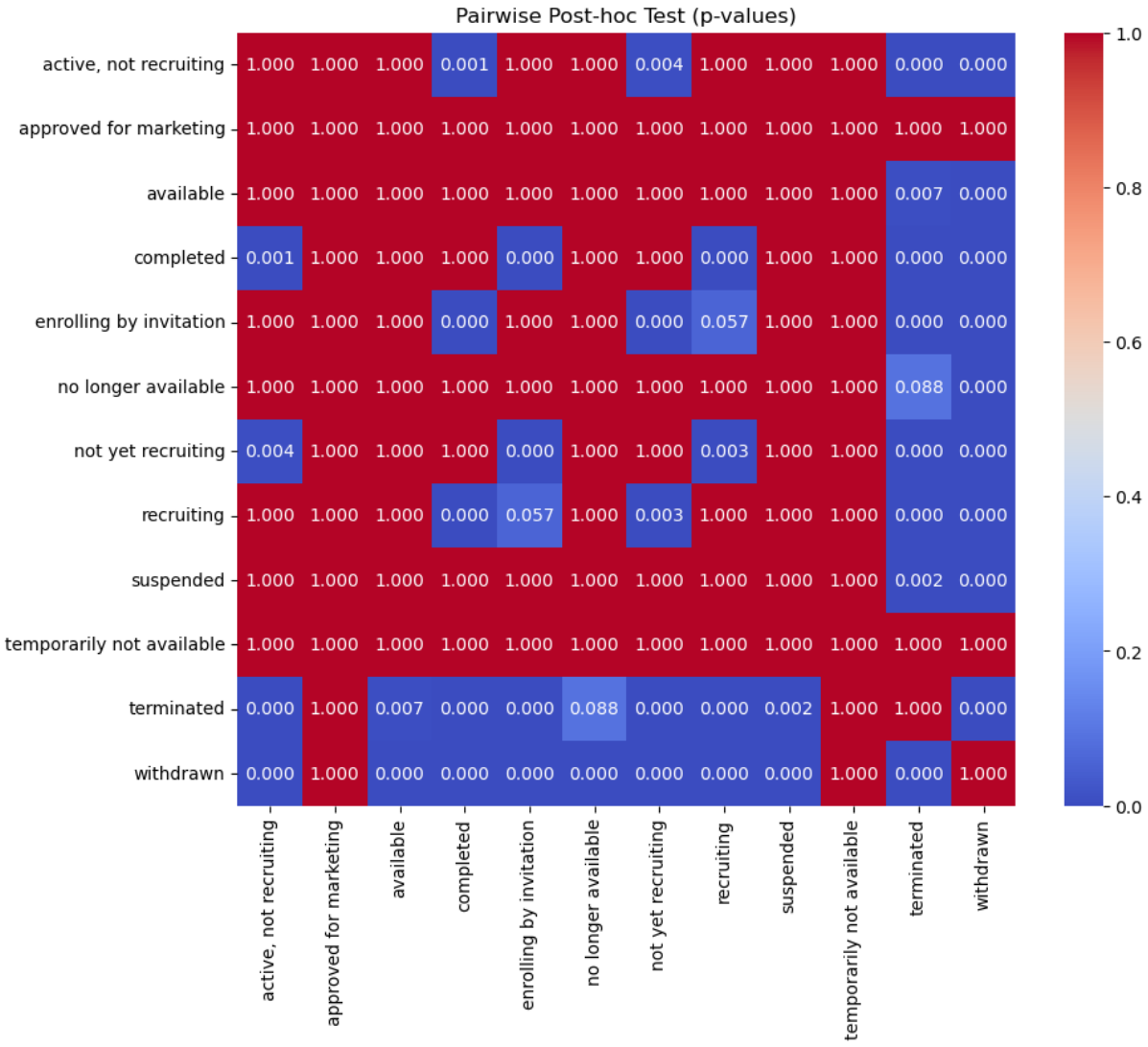
Inference

Enrollment sizes do vary significantly depending on clinical trial status.

This suggests that certain statuses (e.g., "Recruiting" vs. "Completed") tend to involve systematically different participant counts.

A post-hoc pairwise comparison (e.g., Dunn’s test with multiple-comparison correction) would be necessary to identify which statuses differ from which.

3.9 Pairwise Post-hoc Comparison of Enrollment by Trial Status



Observation

The post-hoc heatmap shows that enrollment distributions differ across multiple trial statuses. Strong contrasts are visible particularly between Completed, Recruiting, Terminated, and Withdrawn, while statuses like Temporarily Not Available remain uniform with no significant differences.

Inference

Enrollment size is not evenly distributed across trial statuses. Trials that end early (Withdrawn, Terminated) tend to have systematically different enrollments compared to those that progress further (Completed, Active). This confirms that trial status is a key factor influencing participant counts.

3.10 Hypothesis Test: Enrollment Across Trial Phases

Research Question

Do later-phase trials involve larger participant enrollments compared to early-phase trials ?

Methodology

Null hypothesis (H_o): Enrollment distributions are the same across all trial phases.

Alternative hypothesis (H_1): At least one phase differs in enrollment.

Since enrollment data are skewed with outliers, the non-parametric Kruskal–Wallis test was used instead of ANOVA.

A post-hoc pairwise comparison was conducted to identify specific phase-to-phase differences.

Observation

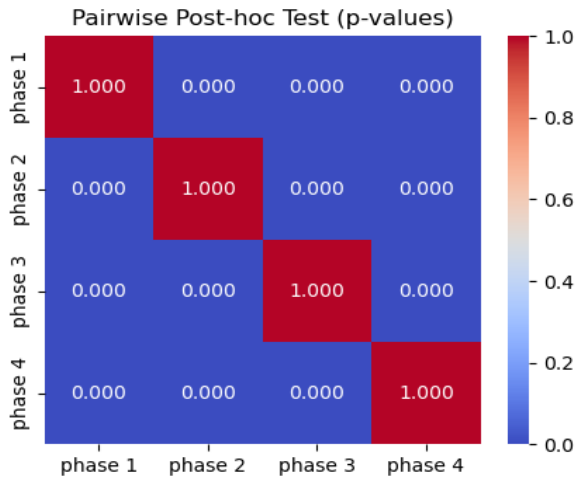
The Kruskal–Wallis test yielded H = 366.15, $p \approx 4.7 \times 10^{-79}$, strongly rejecting H_o.

Post-hoc results show that later phases (Phase 3, Phase 4) consistently have significantly higher enrollments than early phases (Phase 1, Phase 2).

Inference

Enrollment size increases with trial progression: larger participant groups are recruited in later phases to validate safety and efficacy, while early phases remain limited to smaller samples for safety and dosage testing. This aligns with the intended structure of clinical trials, where evidence strength and statistical power grow as trials advance.

3.11 Pairwise Post-hoc Comparison of Enrollment by Phases



Observation:

The heatmap shows statistically significant differences in enrollment between all trial phases. Every off-diagonal comparison yields a p-value of 0.000, indicating strong evidence that enrollment distributions differ across phases.

Inference:

Enrollment size varies meaningfully between trial phases. This suggests that each phase has distinct operational demands—Phase 3 and 4 trials likely require larger sample sizes for efficacy and safety validation, while Phase 1 and 2 trials operate on smaller, more exploratory scales. These differences are not random; they reflect structured design choices in clinical research.

3.12 Hypothesis Test 4 Enrollment Across Sponsor Types

Research Question

Do enrollment sizes differ significantly depending on the type of trial sponsor?

Methodology

Null hypothesis (H_o): Enrollment distributions are the same across all sponsor types.

Alternative hypothesis (H_1): At least one sponsor type has a different enrollment distribution.

Because enrollment data are skewed with outliers, a non-parametric Kruskal–Wallis test was applied.

Observation

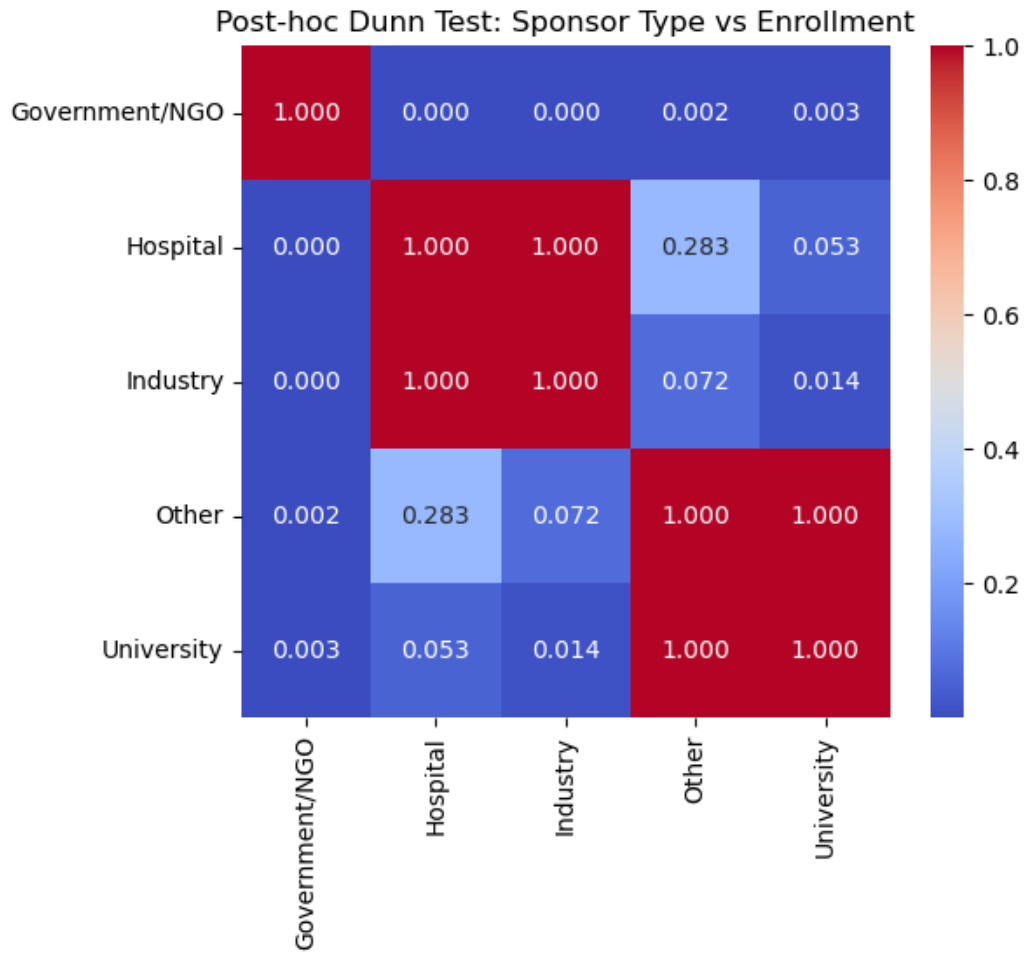
The Kruskal–Wallis test produced $H = 34.12$, $p \approx 1.87 \times 10^{-7}$, leading to a strong rejection of H_o.

This indicates that sponsor type does influence enrollment sizes, with some sponsor categories systematically conducting trials with larger or smaller participant groups.

Inference

Enrollment levels are not uniform across sponsor types. Certain sponsors (e.g., large pharmaceutical companies or government-funded studies) are likely to manage trials with larger participant counts, while smaller sponsors or academic institutions may run smaller-scale studies. This suggests that sponsor resources and trial objectives directly shape enrollment capacity.

3.13 Pairwise Post-hoc Comparison of Enrollment by Sponsor Type



Observation:

The Dunn test heatmap reveals statistically significant differences in enrollment between several sponsor types. Comparisons involving University, Industry, and Hospital show low p-values, indicating strong contrasts. Diagonal values remain at 1.000, as expected.

Inference:

Enrollment sizes vary meaningfully across sponsor categories. Trials sponsored by universities or hospitals may have broader recruitment networks or academic incentives, while industry-sponsored trials might be more targeted or resource-driven. These differences reflect how institutional context influences trial scale and design.

4. Conclusion

This study analyzed clinical trial data to uncover key insights into trial distributions, phases, sponsor involvement, and enrollment patterns. The results highlighted that:

The number of trials started and completed each year are strongly correlated, with no significant difference in averages.

Enrollment sizes vary significantly across trial statuses, phases, and sponsor types.

Later-phase trials (Phases 3 & 4) typically have higher enrollments than early-phase trials (Phases 1 & 2).

Withdrawn and terminated trials show distinct enrollment patterns compared to other statuses.

Sponsor type plays a crucial role in determining enrollment distributions.

These findings provide a clear picture of how clinical trials progress globally and the factors that influence their scale and outcomes

5. Future Work

I am planning to extend this study further using the same dataset. Potential directions include:

Predictive Modeling – Building machine learning models to forecast trial outcomes & enrollment sizes.

Survival Analysis – Estimating trial duration and completion probabilities over time.

NLP-Based Insights – Extracting trends from trial descriptions and intervention details.

Interactive Dashboards – Developing Power BI or Tableau dashboards for real-time trial exploration.