Predicting Clinical Trial Success

**Introduction - C**linical trials are integral to the advancement of medical treatments, representing a cornerstone in the quest for new drugs and therapies. This vital undertaking extends beyond pharmaceutical companies, involving active participation from institutions and government-funded projects, collectively propelling healthcare innovation. However, the journey from initiating a trial to achieving success is fraught with challenges, with a mere 7.9% of trials reaching fruition. The stakes are high, marked by substantial costs, approximately $2.4 billion, invested by biopharmaceutical companies. Despite these formidable challenges, the prospect of creating successful treatments not only mitigates risks but also yields significant benefits, impacting both pharmaceutical companies and projects supported by institutions or the government. This collaborative effort contributes synergistically to the collective pursuit of advancements in medical discoveries.

In this study, we concentrate on three key factors: the quality of the clinical trial, the duration of the study, and demographic considerations. Drawing from a comprehensive dataset obtained from the World Health Organization's registry for Covid-19 trials, which incorporates clinical trial information from around the globe, we navigate the complexities of data accessibility. Accessing clinical trial data posed challenges, as many registries require sponsorship or are limited to investigators. Our dataset, comprising approximately 21,444 -entries covering different trial phases, serves as the foundation for our exploration. Our overarching goal is to delve into these trials, identify patterns, and unravel the factors contributing to trial success. Given the relatively low success rate, understanding these critical factors becomes paramount for pharmaceutical companies, institutions, and government-funded projects, collectively striving to progress medical discoveries.

**Goal Description:** The primary objective of this research is to identify patterns or trends within clinical trial data that may serve as predictive indicators of a trial's success. By scrutinizing historical data, we aim to unravel the intricate interplay of variables that contribute to the approval of a drug. This exploration is not only an academic pursuit but also holds practical significance for the biopharmaceutical industry, providing insights that could inform strategic decisions in future clinical trials.

**Learning Type:** This study will primarily adopt a supervised learning approach. Leveraging historical data on clinical trials, we intend to train our model to recognize patterns associated with successful outcomes. However, the possibility of incorporating unsupervised learning methods will be explored to identify hidden patterns or groupings within the data that might elude human observation. It was decided to used perceptron model as the background model and we will be training and testing our model with logistical regression.

**Personal Interest:** This project holds personal significance for me on two fronts. As a director of a pharmaceutical company, gaining a nuanced understanding of the clinical trial process and the associated data is crucial for informed decision-making in the industry. Additionally, the chosen dataset focuses on COVID-19, a virus that has left an indelible mark on my personal life. My father's passing during the pandemic due to COVID-19 creates a unique and profound connection to this research. It is not merely an academic pursuit; it is a personal exploration into a subject matter that transcends the realms of professional interest, infusing the study with a sense of purpose and personal connection.

**Approach:** A project is only good as how it is planned. The cornerstone of a successful project is how well it is framed. Which is why it is essential to address problem framing, data framing and objective framing. We will approach these by posing relevant questions and trying to answer them.

**Problem framing:**

What is the problem and how does it relate to the data? What is the objective of this study?

In framing the problem for our clinical trials analysis, we confront the stark reality of a mere 7.9% success rate in drug development. The core issue revolves around deciphering the intricacies that contribute to this low success rate. By analyzing the data we have, including attributes like TrialID, Recruitment Status, and Study design, we aim to identify patterns that can help predict whether a trial is likely to be successful. Ultimately, we want to create a model, based on historical data, that can forecast the outcomes of future trials. The success of our study will be measured by how well our model can predict trial results based on the information in the dataset.

**Data framing:**

What are the features that were selected and organized?

From the data set, there were many attributes that were present. Based on study, there are certain factors that are very important for a successful clinical trial. The table below will describe the factors that will be used in the study.

|  |  |  |  |
| --- | --- | --- | --- |
| **Factors consider for clinical trial success** | | | |
| **Factors** | **Details** | **Descriptions** | **Related Studies** |
| Speed  *H1* | Duration of the Trial | Time taken to conduct the clinical trial | 2 |
| Quality  *H2* | Clinical Trial Experience | The sponsors previous experience in the field | 6 |
| Demographic  *H3* | Gender Usage | Study using one or both gender | 3 |

**Objective framing:**

What are the parameters this study will be testing for and what are its performance metrics?

The objective of this study is to predict the success of clinical trials. The table below describes the objective framing based on each hypothesis mentioned in methodology.

|  |  |  |  |
| --- | --- | --- | --- |
| **Objective Framing** | | | |
| **Objective** | **Target Variable** | **Performance Metric** | **Evaluation** |
| Predicting the Success of Clinical Trials | Dependent: Status of Clinical Trials | Accuracy | Assess the model's accuracy on a separate dataset to ensure its generalization to new data. |
| Factors Influencing Clinical Trial Success | Dependent: Status of Clinical Trials  Independent: Success Ration  Independent: Duration  Independent: Gender | Correlation Coefficients | Assess the correlation between each independent variable and the dependent variable. Identify variables with the strongest influence on clinical trial success. |
| Predictive Model for Clinical Trial Success | Dependent: Status of Clinical Trials  Independent: Success Ration  Independent: Duration  Independent: Gender | Impact of each independent variable | Refine the model iteratively, adjusting parameters to improve accuracy while considering the influence of each independent variable. |

**(c)Methodology**

The baseline model will be to predict the success or failure for binary classification by using perceptron model.

*H1* Duration of the clinical trials will possess a negative impact on the clinical trials.

*H2* Sponsor's experience will possess a positive impact on the clinical trials.

*H3* Using both genders will have a negative effect in the studies.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Methodology** | | | | | |
| **Hypothesis** | **Evaluation Metrics** | **Experimental Design** | **Data Acquisition and Cleaning** | **Model** | **Related Studies** |
| *H1* | Obs  Mean  Std. Dev  Min  Max  F1 score  VIF | Choice of Algorithm: Logistic Regression  Data Splitting: 80/20  Cross-Validation: Splitting data into subsets.  Randomization:  Among Features | Source: WHO International Clinical Trials Registry Platform (ICTRP)  Covid-19 trials  Mentioned in data cleaning process | Logistic Regression | 2 |
| *H2* | 6 |
| *H3* | 3 |

|  |  |  |
| --- | --- | --- |
| **Data Cleaning Process** | | |
| **Stage** | **Process** | **# of Rows** |
| Stage 1 | Clinical trial data gathering:  Source: WHO International Clinical Trials Registry Platform (ICTRP)  Covid-19 trials: | 21,444 |
| Stage 2 | Study type: intervention  Status: completed or not  Duration: 01.2014 - 12.2023  Phase: 1 - 4 | 6,453 |
| Stage 3 | Drugs: Matched the word drug in the intervention | 3,475 |

2. **Data Summary**

The original dataset contains 42 attributes. The cleaned and deduced down to 4 factors and answer the hypothetical question.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Definition and measurement variables** | | | | | |
| **Variables** | **Factor** | **Variables** | **Definition** | **Measurement** | **Related Studies** |
| Dependent Variable | Status | Status | Success of clinical trials | If completion date = 1  Otherwise = 0 | 4 |
| Independent Variables | Quality | Success Ratio | Cumulative success ratio of clinical trial | Number of successful clinical trial / total number of clinical trials | 5 |
|  | Speed | Duration | Days taken to complete clinical trials | Number of days from the beginning to complete of clinical trials | 2 |
|  | Demographic | Gender | Gender used in the study | If male or female = 1  Otherwise = 0 | 3 |

3. **Model Application in Practice**

|  |  |
| --- | --- |
| **Model Application in Practice** | |
| Decision-Making Context | Pharmaceutical/Institutional Clinical Trial Optimization |
| Causal Question and Hypothesis  Does optimizing clinical trial parameters based on our predictive model lead to increased success rates? | We hypothesize that implementing the recommendations from our model will result in a higher success rate for clinical trials. |
| Outcome | Primary: Increase success rate of clinical trials  Secondary: Reduction in duration, better resource allocation and increased efficiency in drug development |
| Target Population | The model aims to assist pharmaceutical and institutional entities in optimizing their clinical trial strategies. |
| Experimental Conditions | All clinical trials in the dataset involve interventional studies, consisting of both experimental and control groups. |
| Unit of randomization | Randomizing at the level of individual clinical trials ensures that the impact of the model is assessed across a diverse range of drug development projects. |

**Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **Duration** | **Inclusion Gender** | **Date Completed Label** | **Inclusion Gender Label** |
| **Mean** | 357.47 | 0 | 1 | 0.012987 |
| **Standard Deviation** | 363.07 | 0 | 0 | 0.113402 |
| **Minimum** | -61.0 | 0 | 1 | 0 |
| **Maximum** | 3196.0 | 0 | 1 | 1 |
| **VIF** | 1.005441 | NaN | 2.007799 | 1.005441 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Training Accuracy** | **Test Accuracy** | **F1 Score** |
| **Logistic Regression** | 0.9878 | 0.9839 | 0.9759 |
| **Perceptron** | 0.9797 | 0.9677 | 0.9677 |
| **SVM** | 0.9879 | 0.9839 | 09759 |

Meanings

Mean: The mean duration of clinical trials is 357.47 days, and on average, trials did not explicitly include gender (mean inclusion gender label is 0), while most trials have completion dates (mean date completed label is 1).

Standard Deviation: The standard deviation of the duration of clinical trials is 363.07 days, indicating considerable variability. Inclusion gender and date completed label show no variability (standard deviation is 0).

Minimum and Maximum: The minimum duration recorded is -61 days, potentially an anomaly. The maximum duration is 3196 days, indicating a wide range. Inclusion gender and inclusion gender label are consistently 0, and date completed label is consistently 1.

VIF (Variance Inflation Factor): The VIF for duration is 1, suggesting minimal multicollinearity. However, inclusion gender and inclusion gender label have undefined VIF values, indicating potential collinearity issues. The VIF for date completed label is 1.97, suggesting moderate correlation with other variables.

**Methodology:**

Baseline Model: We utilized logistic regression as the baseline model for binary classification, predicting the success or failure of clinical trials.

Hypotheses:

H1 Duration Impact: Our hypothesis regarding the negative impact of trial duration was supported by the data, revealing a negative correlation between duration and trial success.

H2 Sponsor's Experience: The analysis affirmed our hypothesis that trials sponsored by experienced entities have a higher likelihood of success.

H3 Gender Usage: Contrary to our initial hypothesis, the data did not show a significant negative effect of using both genders in studies on trial success.

Data Cleaning: We performed a multi-stage data cleaning process, including gathering clinical trial data, filtering by study type, status, duration, and drug-related keywords.

**Discussion:**

The results reveal a dataset characterized by diverse trial durations, prevalent single-gender inclusion, and a consistent reporting of completion dates. VIF analysis indicates minimal collinearity among features, affirming the independence of predictors in the logistic regression model. Notably, the models, particularly Logistic Regression, exhibit high accuracy and F1 scores, reflecting successful training and robust predictive capabilities. The moderate VIF for the date completed label prompts consideration in future model refinement to enhance predictive accuracy further. In summary, the employed methodology, encompassing data summary, VIF analysis, and model evaluation, offers a comprehensive understanding of the dataset and underscores the successful performance of the models.

**Conclusion**:

The employed methodology offers a comprehensive understanding of the dataset, emphasizing successful model performance. Insights gained can inform strategic decisions in clinical trial optimization, providing a valuable contribution to the collective pursuit of medical advancements.

**Citations**

1. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemp Clin Tr Commun.* 2018;11:156–164.
2. Kim E, Yang J, Park S, Shin K. Factors Affecting Success of New Drug Clinical Trials. Ther Innov Regul Sci. 2023 Jul;57(4):737-750. doi: 10.1007/s43441-023-00509-1. Epub 2023 May 11. PMID: 37166743; PMCID: PMC10173933.
3. Lee H, Pak YK, Yeo EJ, Kim YS, Paik HY, Lee SK. It is time to integrate sex as a variable in preclinical and clinical studies. Exp Mol Med. 2018 Jul 23;50(7):1-2. doi: 10.1038/s12276-018-0122-1. PMID: 30038313; PMCID: PMC6056479.
4. Logan JK, Tang C, Liao Z, Lee JJ, Heymach JV, Swisher SG, et al. Analysis of factors affecting successful clinical trial enrollment in the context of three prospective, randomized, controlled trials. *Int J Radiat Oncol Biol Phys.*2017;97(4):770–7. doi: 10.1016/j.ijrobp.2016.11.035.
5. Thunecke M. Predicting Success of Clinical Trials. Journal of clinical trials. 2021.
6. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics.* 2019;20(2):273–286.