Clinical Trials: Phase Completion Prediction

Erdös Institute Bootcamp

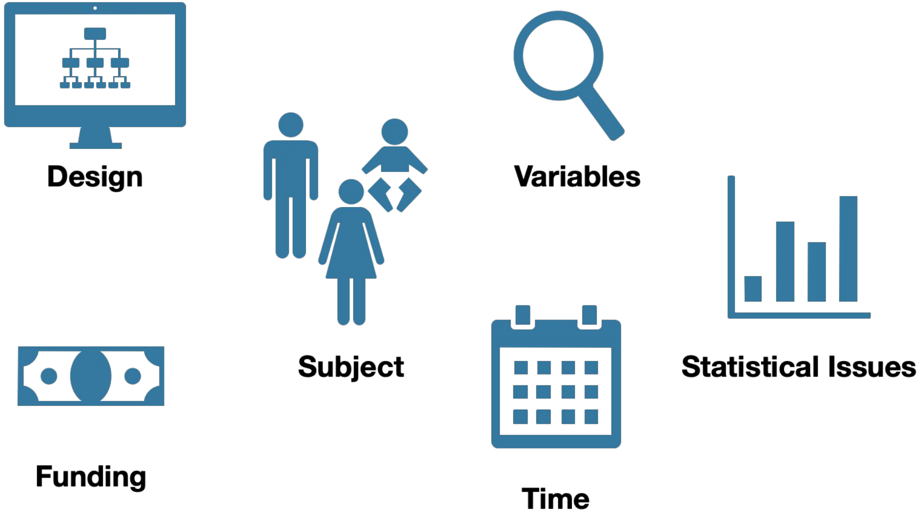
Devashi Gulati - University of Georgia
Adriana Morales Miranda- University of Illinois at Urbana-Champaign
Meghan Peltier- The Florida State University

June 4, 2022

Significance

- Clinical trials are vital for the discovery of new medicines and diagnostic methods
- Clinical trials are also resource intensive

"There are multiple reasons which can cause failure of a trail - a lack of efficacy, issues with safety, or a lack of funding to complete a trial, as well as other factors…" [1]



Design

Subject

Variables

Statistical Issues

Funding

Time

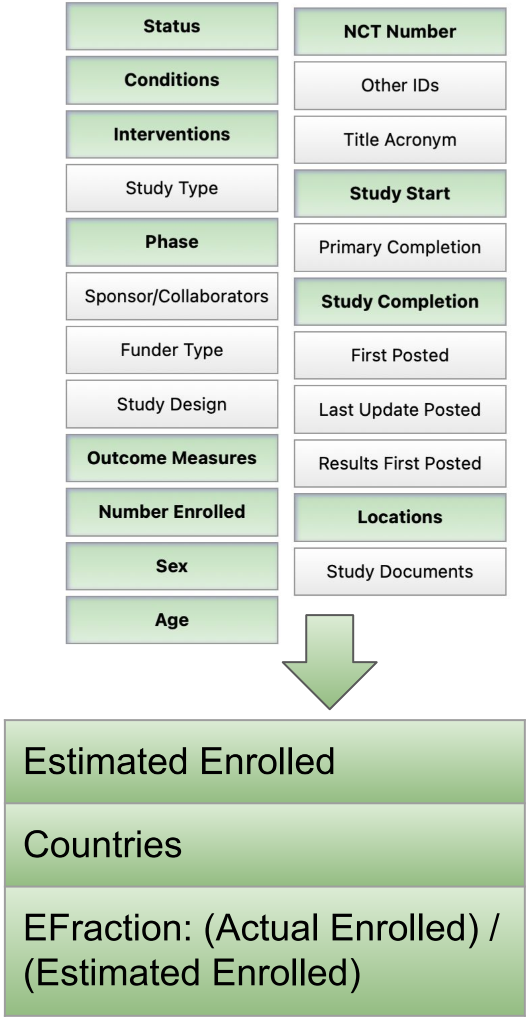
Predicting the successful completion of clinical trials can increase efficiency and development of clinical research

Data Collection

- NIH U.S National Library of Medicine's website [2]
- 10,000 trial for each:
  - ◆ Cancer
  - ◆ Cardiovascular diseases
  - ◆ Respiratory diseases

NIH
U.S. National Library of Medicine

shutterstock.com · 2088809149



Status
Conditions
Interventions
Study Type
Phase
Sponsor/Collaborators
Funder Type
Study Design
Outcome Measures
Number Enrolled
Sex
Age

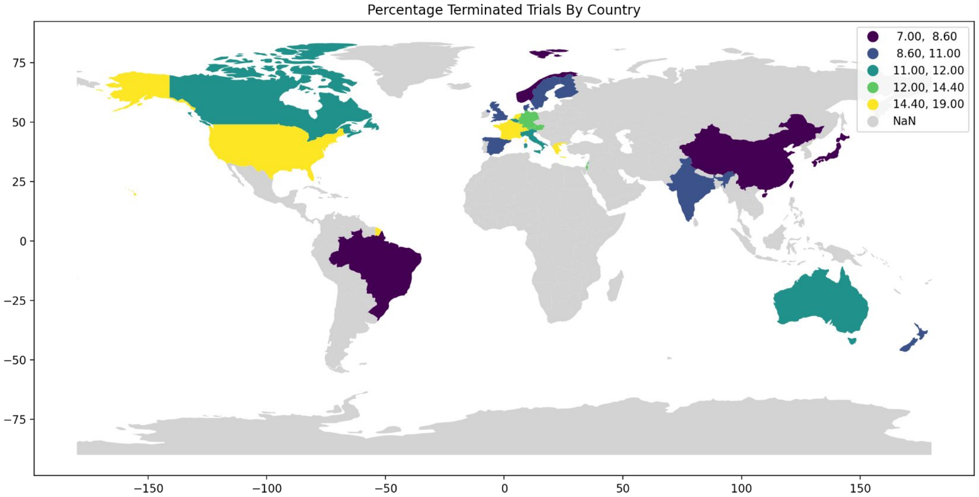
NCT Number
Other IDs
Title Acronym
Study Start
Primary Completion
Study Completion
First Posted
Last Update Posted
Results First Posted
Locations
Study Documents

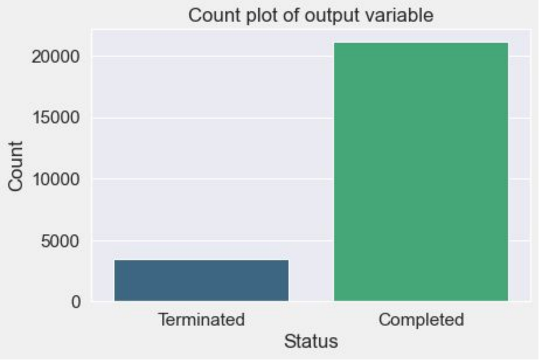
Estimated Enrolled

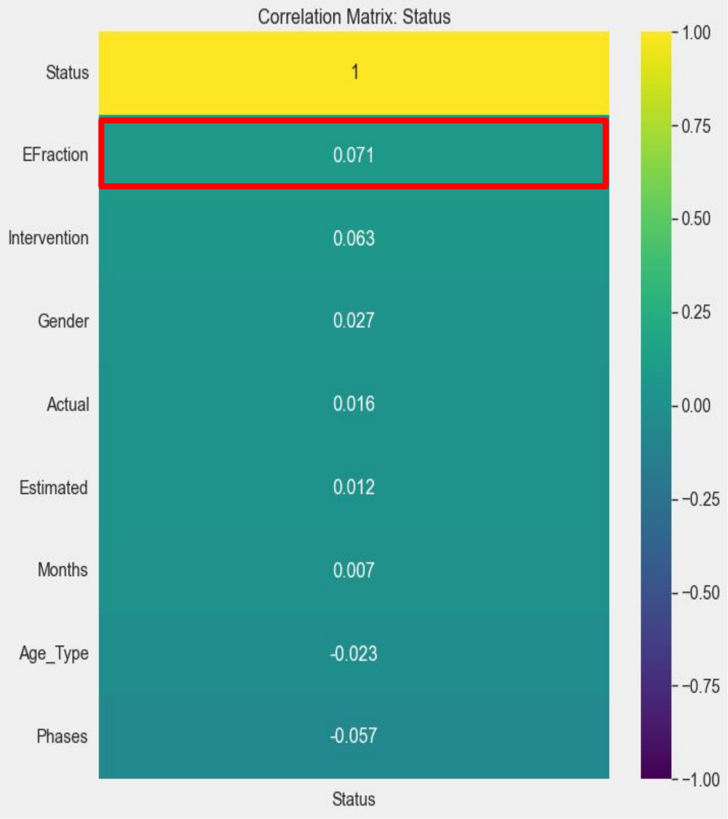
Countries

EFraction: (Actual Enrolled) / (Estimated Enrolled)

Exploratory Data Analysis



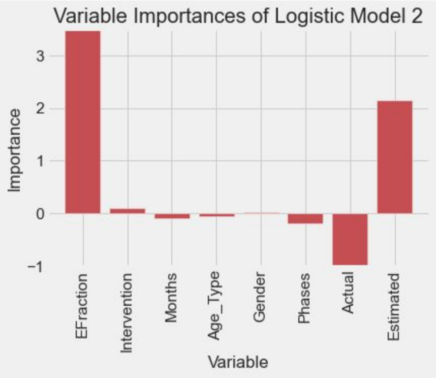


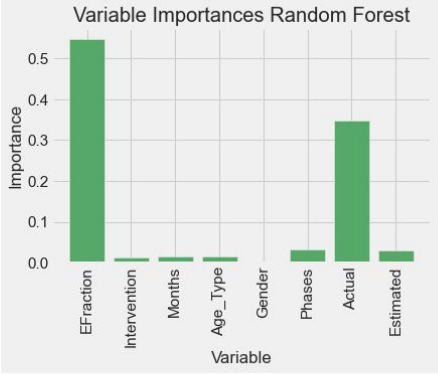


Modeling

- Classification models : Logistic Regression, Random Forest, Neural Network

Important Features





Logistic Regression

```
              precision    recall  f1-score

           0       0.43      0.76      0.55
           1       0.96      0.84      0.89

    accuracy                           0.83
   macro avg       0.69      0.80      0.72
weighted avg       0.88      0.83      0.84
```

Random Forest

```
              precision    recall  f1-score

           0       0.79      0.37      0.51
           1       0.91      0.98      0.94

    accuracy                           0.90
   macro avg       0.85      0.68      0.73
weighted avg       0.89      0.90      0.88
```

Neural Network

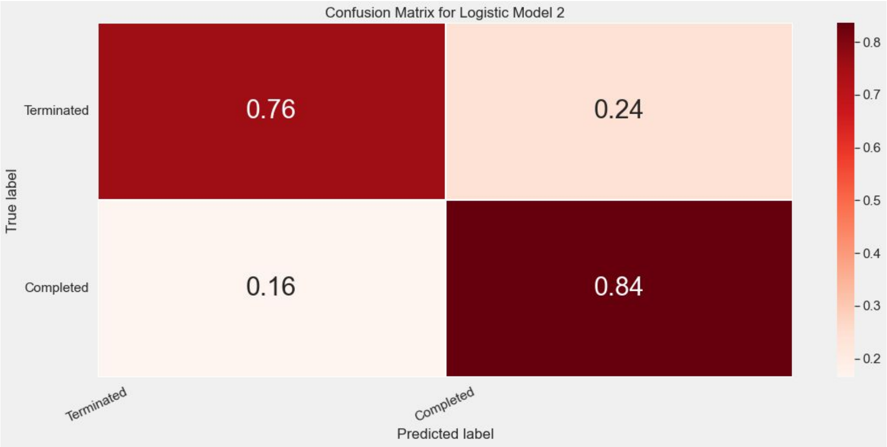
```
              precision    recall  f1-score

           0       0.70      0.47      0.56
           1       0.92      0.97      0.94

    accuracy                           0.90
   macro avg       0.81      0.72      0.75
weighted avg       0.89      0.90      0.89
```

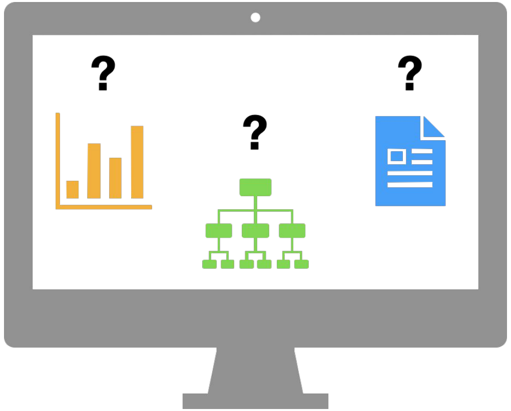
Model Summary

While the accuracy for all the models was good, the imbalance of our data leaves room for a lot of Type 1 Errors (False Positives)

There is room for improvement:

➔ balancing data
➔ examine larger data set
➔ improve model through feature selection
➔ add more features





Inference

Recommendations for research organizations :

- Prioritize clinical trials that have achieved maximum enrollment
- Factors like age, months or type of intervention should not play a significant role in allocation of resources to clinical trials
- Using our logistic prediction model, NIH could save **approx. 24%** of the 30 billion spent annually on clinical trials i.e approx. **7.2 billion dollars** of taxpayer money per year. This demonstrates the impact of clinical trial phase completion prediction.

Future directions:

- Analyze importance of population density of the location on trial completion
- Improve our predictive model
- Further data collection and analysis on how to increase enrollment
- Make an app to predict trial completion

Thank you!

Special thanks to Matthew Graham, our mentor.

References

[1] Fogel, David B. “Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review.” *Contemporary clinical trials communications* vol. 11 156-164. 7 Aug. 2018, doi:10.1016/j.conctc.2018.08.001

[2] https://clinicaltrials.gov/

[Image 1] Curry, Rowan. “Simplified Logistic Regression: Classification With Categorical Variables in Python.” *Medium*, 4 Jan. 2022, medium.com/@curryrowan/simplified-logistic-regression-classification-with-categorical-variables-in-python-1ce50c4b137.

[Image 2] “What Is a Random Forest?” *TIBCO Software*, www.tibco.com/reference-center/what-is-a-random-forest. Accessed 4 June 2022.

[Image 3] Gupta, Vikas. “Understanding Feedforward Neural Networks | LearnOpenCV.” *LearnOpenCV – OpenCV, PyTorch, Keras, Tensorflow Examples and Tutorials*, 20 Apr. 2021, learnopencv.com/understanding-feedforward-neural-networks.