Visualizing and Predicting Trends in Clinical Trials: A Deployed Dashboard Using Clinical Trial Data

Final project write-up

Team: Mission Possible

Team member: Can Wang, Xiao Wu, Xindi Shan

Importance of the project

Clinical trials are a crucial component of drug development, playing a vital role in advancing public health. They provide essential evidence of drug effectiveness and safety, guiding regulatory approvals and informing clinical practice. According to ClinicalTrials.gov, an online database of clinical research studies, the number of registered studies increased from 2219 in 2000 to more than 500,000 by the end of 2024 (ClinicalTrials.gov, 2024). With the rapidly evolving medical innovations, efficiently analyzing clinical trial data is important for informed decision-making. In this project, we developed a dashboard using data from ClinicalTrials.gov, visualizing and predicting the trend in clinical trials.

Multiple projects have taken advantage of the extensive data available in ClinicalTrials.gov for analysis and visualization, such as visualizing the geographic distribution of EEG-related trials (Tibbs, 2017), summarising the trends of interventional trials (Chiswell, 2022), and providing comprehensive summary statistics (ClinicalTrials.gov, 2024). However, the existing tools often rely on outdated data and lack comprehensive, user-friendly features for dynamic trend exploration.

Our project aims to explore trends in the number of clinical trials conducted across various countries, disease types, and phases, as well as how these trends evolve over time. Using the most up-to-date data from ClinicalTrials.gov, we developed an interactive dashboard, providing dynamic visualizations based on disease type, trial phase and trial location, as well as making predictions. This dashboard would help to identify patterns in clinical trial data, supporting policy-making, research prioritization and funding allocation. It could also provide foresight for researchers, drug developers and healthcare providers.

Overview of technical challenges

The data is collected from the PostgreSQL AACT database, which contains all information about the registered trials in ClincialTrials.gov and is refreshed daily (Clinical

Trials Transformation Initiative, n.d.). The key variables we extracted include trial ID, start date of the trial, phase, disease and country. Additionally, we only included the trials that were conducted in the past 10 years in the top 10 most studied diseases and top 10 most popular countries. Trials with missing values in any of the key variables were excluded. The code for data collection is available in the data_collection folder on GitHub.

We built LASSO regression models to predict clinical trial trends across countries, diseases, and trial phases, using cross-validation to optimize model parameters. After training separate models for each category using parallel computing, we created a streamlined prediction function, tailored for seamless integration into a dashboard. All models performed well, and we saved the complete work in the prediction_modelling folder.

We need to combine our work into one dashboard, and make it user-friendly. So we need to figure out a reasonable layout, use some interactive designs like selection boxes and click buttons, and provide clear guides so that users can use it intuitively and conveniently. The code for making the dashboard is available in the “Dashboard” folder on Github.

Effectiveness of paradigm integration

Our project effectively integrates machine learning and parallel computing paradigms to address the challenges of modeling clinical trial data. The machine learning paradigm is exemplified by our LASSO regression models, which effectively handled the extreme imbalance in trial counts across countries by prioritizing features with stronger predictive power (Kavlakoglu, 2024). It ensures robust and interpretable results. The effectiveness of LASSO is also validated by our strong model performance metrics, with R^2 values above 0.97 for country predictions and reasonable mean absolute errors across all categories.

To improve computational efficiency, we applied parallel computing during model training and cross-validation. This reduced processing time significantly, enabling us to train multiple models for countries, diseases, and trial phases in a timely manner, even with large-scale data.

The successful integration of these paradigms is demonstrated in our streamlined prediction function, which combines LASSO's ability to handle complex datasets with parallel computing's speed to create an efficient, practical tool for analyzing clinical trial trends. This approach not only solved our immediate technical challenges but also created a reusable foundation for effective dashboard integration.

Summary of the data analytic product

Our final data analytic product is an interactive dashboard. This dashboard has three panels. The first part provides a descriptive function. Users can select by county/disease/phase, and it will display a line graph showing the trend of counts of clinical trials. In the second part, we provide a predicting function. Users can select by country/disease/phase, and choose a year to predict. It will show the predicted clinical trial counts as well as some information about the predicting model. In the third part, we provide a search function. Users can filter the clinical trial records, conditioned by country, disease and year. The dashboard works as intended, and it combines our work in data collecting, data processing and predicting model construction.

The dashboard is easy to use, with simple design, clear guides and clear titles and legends in the graphs. It is original and complex, as we use up-to-date data, and combine searching, descriptive and predicting functions into one product. This product avoids the limitations of previous works, and provides a comprehensive overview of global clinical trial trends.

The final product is also consistent with the original goal. We constructed the dashboard as planned and predicted the number of trials for future years. However, fewer variables were chosen than those in the proposal for better visualization and machine learning model building. Additionally, the LASSO model was chosen as the final model based on its small MSE.

Lessons from the project

Through this project, we learned how to efficiently filter and extract data from an SQL database and to use GitHub for collaborative projects through command-line coding.

We learned to combine our work into a data analytic product, which was an interactive dashboard, and make it user-friendly. The model prediction phase taught us that simpler approaches like LASSO regression can be highly effective, especially when focusing on interpretability and practical use.

Most importantly, we learned to balance technical sophistication with user experience in creating an interactive dashboard, while maintaining clean, modular code that can be easily maintained and updated.

References

Chiswell, K. (2022). GitHub - Ctti-Clinicaltrials/Aact: Improving public access to aggregate content of ClinicalTrials.gov. *GitHub*. Retrieved from https://github.com/ctti-clinicaltrials/aact

ClinicalTrials.gov. (2024). *ClinicalTrials.gov*. Retrieved from https://clinicaltrials.gov/about-site/trends-charts

Clinical Trials Transformation Initiative. (n.d.). *AACT database: Aggregate analysis of ClinicalTrials.gov*. Retrieved from https://aact.ctti-clinicaltrials.org/

Kavlakoglu, E. (2024). Apply lasso regression to automate feature selection. *IBM Developer*. Retrieved from https://developer.ibm.com/tutorials/awb-lasso-regression-automatic-feature-selection/

Tibbs, S. (2017). Where are EEG-related Clinical Trials Being Conducted? *CTTI-ClinicalTrials.org*. Retrieved from https://aact.ctti-clinicaltrials.org/use