UNIVERSITÉ PARIS CITÉ

FACULTE DE PHARMACIE DE PARIS

Année 2023 - 2024 **N°**

THESE

pour l'obtention du Diplôme d'Etat de

DOCTEUR EN PHARMACIE

présentée et soutenue publiquement

par

SCHAEFER Yves

le 16 décembre 2024

What are the organizational and protocol variables increasing the risk of early termination in Phase 2 Rare Diseases Clinical Trials?

JURY :

Mme. GUIHENNEUC Chantal, Responsable et Président

Mme. ARNOUX Armelle, Directeur

Mme. KATSAHIAN Sandrine,

M. MARTELLI Nicolas,

M. FISMAN Paul

ABSTRACT

This study aims to identify the most impactful factors contributing to early termination in Phase 2 rare disease clinical trials, compared to Phase 2 common disease trials. We specifically focus on organizational and protocol-related factors, as opposed to the more commonly analyzed scientific and pharmaceutical drivers of termination.

We trained our model without imposing any prior assumptions or weights favoring specific variables over others. The model was allowed to learn patterns from the data autonomously. Notably, the model identified and highlighted variables that, based on existing literature and prior knowledge, are known to significantly influence early trial termination. This alignment between the model's findings and established research reinforces confidence in the model's predictive capacity and its ability to uncover meaningful insights from the dataset.

This conveys the neutrality of the model during training and how its results are aligned with prior knowledge, which validates its performance.

Methods

Using data from ClinicalTrials.gov, we built Logistic Regression (LR) and Extreme Gradient Boosting (XGBoost) models to compare the prediction of trial outcomes (completed vs. terminated early) for Phase 2 Rare Diseases trials versus Common Diseases trials conducted between 2019 and 2022. The dataset incorporated a range of variables, including trial protocol design variables, organizational variables and other variables found within the ClinicalTrials.gov database. Models were evaluated using precision, F1-score, and PR-AUC to account for the imbalanced nature of the dataset. To mitigate the impact of COVID-19, terminated trials linked to pandemic-related reasons were excluded via an NLP model. Then, SHAP (SHapley Additive exPlanations) values were employed to interpret the significance of each factor.

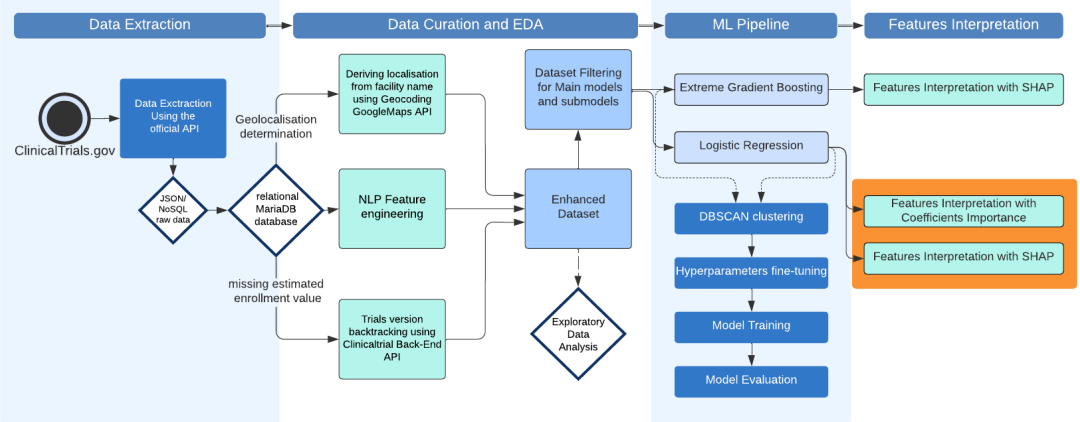
Results

Logistic Regression outperformed XGBoost in all the datasets, giving us some confidence for the feature interpretations. Apart from the moderate deviation (set as real enrollment count < 75% theoretical enrollment count) from the original enrollment target that significantly increased the risk of early termination, we found more subtle impactful variables including complex regulatory requirements for FDA-regulated drugs along with the use of traditional trial designs which may not always be a fit for Rare Diseases trials. Protocol stability, less restrictive eligibility requirements,

Academic sponsor entities and “Diseases and Abnormalities at or Before Birth” MeSH, were associated with a higher likelihood of trial completion.

Conclusion

This study highlights key protocol and organizational factors that increase the risk of early termination in rare disease trials, such as FDA regulatory complexity and patient recruitment challenges. Traditional clinical trial designs, such as parallel treatment models, are less effective in rare disease contexts, where accruing patients is difficult. Collaborations with multiple centers nationally and academic institutions, however, improve trial success rates by expanding access to specialized resources and participants. Understanding these context-specific challenges is crucial for designing more resilient Phase 2 trials in rare diseases.



Graphical Abstract of the study. This thesis will focus on the Features Importance with SHAP section. Informations on the others parts can be found in the Additional Information Section

INTRODUCTION

What is the current state of the clinical trial field?

In 2022, the global clinical trial market was valued at approximately $49.8 billion, and it is projected to reach $78.3 billion by 2030[1].

Despite their importance, clinical trials are often lengthy, costly, and have a low success rate. Numerous factors, including adverse side effects and insufficient efficacy, can lead to trial failures. Additionally, organizational and project management issues can significantly impact the ability to meet primary endpoints.

The field of clinical research is continually expanding, with an exponential increase in the number of trials and evolving regulatory requirements, such as new non-FDA phases or more mandatory fields in order to increase transparency and allow public access to a wider audience.

Several public data sources are available for analysis, the largest being the ClinicalTrials.gov database, which currently hosts over 502 990 registered trials worldwide as of July 2024. ClinicalTrials.gov, managed by the National Library of Medicine, is an essential resource for healthcare professionals, researchers, patients, and the public. It provides access to clinical trial registration data, enabling the analysis of trends in trial composition, size, design, and funding.

Phase II Clinical Trials

Phase II clinical trials are conducted to evaluate whether a treatment shows sufficient activity or other clinical benefits to justify further investigation in a larger, more definitive phase III trial. While the primary focus is often on the treatment's activity, secondary endpoints such as toxicity, quality of life (QOL), biomarkers, and disease-specific outcomes are also possible.

The sample size in phase II trials is typically moderate, ranging from tens to a few hundred participants. In single-arm trials, disease response is the preferred primary endpoint, particularly when testing cytotoxic agents. For randomized phase II trials, progression-free survival (PFS) and disease-free survival (DFS) are commonly used to measure disease progression and patient outcomes, with overall survival (OS) being more suited for diseases with poor prognosis.

Randomized phase II trials with a control arm are generally preferred, as they offer a more robust comparison. However, particularly in rare diseases, single-arm or adaptative designs may be appropriate when we have a limited pool of patients, or in settings where there is a lack of standard treatments or reliable historical controls.[2]

How are clinical trials on rare diseases different?

In the United States, the Food and Drug Administration (FDA) defines a rare disease as any disease that affects fewer than 200,000 Americans. In Europe, a disease is defined

---

[1] Research and Markets. Clinical Trials Market Size, Share & Trends Analysis Report, 2021 - 2028. Research and Markets; 2022.https://www.researchandmarkets.com/reports/4396385/clinical-trials-market-size-share-and-trends

[2] Torres-Saavedra PA, Winter KA. An Overview of Phase 2 Clinical Trial Designs. Int J Radiat Oncol Biol Phys. 2022;112(1):22-29. doi:10.1016/j.ijrobp.2021.07.1700

as rare when it affects less than 1 in 2,000 people.Worldwide, it is estimated that 350 million people suffer from rare diseases worldwide.[3]
According to the National Organization for Rare Disorders, more than 95% of Rare Diseases have no treatment. Clinical Trials are essential for identifying therapeutic options.

The number of clinical trials in rare diseases has grown significantly from the last 5 years (2018-2022) close to 16,000 globally. More than 500 orphan designated drugs are commercially available worldwide for various rare diseases, out of which oncology has the most drugs, followed by hematological and other disorders. [4]

Why can trials stop early ?

Reasons for early discontinuation of a trial include: evidence of benefit, evidence of harm, and evidence of futility. More than 1 of these elements will often be present:[5]

Reduction in sample size

The major cause is slow or lack of accrual. This can be due to the low incidence of the disease, an heterogeneous population but also potentially overly restrictive eligibility criterias. This lack of accrual can also result from lack of resources in order to be able to reach a wide enough audience or interest for the study.

Early evidence of benefit or harm

If the new treatment shows a significant beneficial/negative effect at an interim analysis it would be unethical not to stop the trial so that the new treatment may be made available or not.

Futility and other reasons

Interim analysis may reveal a neutral effect or even a trend toward a negative effect of the new treatment, and it may appear increasingly unlikely that a positive effect will be demonstrated by the end of the trial. It is often useful to determine the conditional power of the observed difference, calculated under the assumption that the current trend in the data will continue until the end of the trial[6].

However**,** not all trials stop early due to the efficacy of the experimental product itself. The major cause is the inadequate accrual of eligible patients, but other factors can include poor planning or a misunderstanding of key biological or drug development principles.[7]

---

[3] Rare Diseases Clinical Research Network. What are rare diseases? Rare Diseases Clinical Research Network. https://www.rarediseasesnetwork.org/about/what-are-rare-diseases.
[4] *https://novotech-cro.com/faq/rare-disease-clinical-trials-unveiling-insights-and-charting-progress*. Novotech. Published September 13th 2023
[5] Thom EA, Klebanoff M., Issues in clinical trial design: Stopping a trial early and the large and simple trial, Research Methods: State of the ScienceVolume 193, Issue 3, P 619-625,September 2005, DOI: 10.1016/j.ajog.2005.05.061
[6] *Lan KK, Wittes J. The B-value: a tool for monitoring data. Biometrics. 1988;44(2):579-585.*
[7] *https://www.allucent.com/resources/blog/why-do-clinical-trials-fail*, Allucent

Additionally, project and data management often pose greater challenges, especially in multicenter trials that require regular sponsor monitoring visits, data audits, trial infrastructure issues like eCRFs, sample shipping, along with all the additional time needed to process the additional paperwork.[8]

Additionally, the cause of the termination can be unrelated to the study itself: it could result from a purely financial loss or depletion of funds, a business decision, workforce issues or logistic issues such as the manufacturer not producing the experimental product before the end of the trial.

Due to these reasons, we will focus on Phase 2 trials, which are more influenced by protocol and organizational variables than ADME or eligibility factors that are more preponderant for the other phases.
With the same model pipeline we will compare side by side the most relevant variable effects for Rare Diseases trials compared to Common Diseases trials.

We will first build a model that includes enrollment-related variables, which are well-established as high-risk factors for early termination. This will allow us to assess the model's performance with these influential variables in place. Subsequently, we will rebuild the model excluding the enrollment variables, to investigate the impact of other, more subtle factors. This approach will enable us to better understand the contribution of less obvious variables, which may otherwise be overshadowed by the strong influence of enrollment data.

CONTRIBUTIONS

Previous analysis of ClinicalTrials.gov registration data have focused on specific fields, such as a single condition[9], specific phases, countries[10], or particular registration elements[11] within ClinicalTrials.gov. However, to our knowledge, no study has exploited the free text field "why_stopped" which provides insights into the reasons for early termination of studies. By using Natural Language Processing (NLP), we aim to curate new information from this field and incorporate it to filter and classify different types of study terminations.

---

[8] Booth C, Parexel, Advancing rare diseases drug development Report, Section 2 Effective regulatory strategies,Part 7 https://www.parexel.com/insights/new-medicines-novel-insights/advancing-rare-disease-drug-development/study-design-and-execution-rare-diseases/how-sites-manage-pediatric-gene-therapy-trials

[9] Chen D, Parsa R, Chauhan K, et al. Review of brachytherapy clinical trials: a cross-sectional analysis of ClinicalTrials.gov. Radiat Oncol. 2024;19(1):22. Published 2024 Feb 13. doi:10.1186/s13014-024-02415-8

[10] Di Tonno D, Perlin C, Loiacono AC, et al. Trends of Phase I Clinical Trials in the Latest Ten Years across Five European Countries. Int J Environ Res Public Health. 2022;19(21):14023. Published 2022 Oct 28. doi:10.3390/ijerph192114023

[11] Song SY, Koo DH, Jung SY, Kang W, Kim EY. The significance of the trial outcome was associated with publication rate and time to publication. J Clin Epidemiol. 2017;84:78-84. doi:10.1016/j.jclinepi.2017.02.009

In addition to this, we will try to see the impact of different specific features such as the countries, regions, organization and sponsor classes and how they compare between each other.

While the overall model pipeline isn't really different to the ones of different articles on the same topic, we introduce extra steps in order to curate more accurate data of higher quality. This may result in lower scores but depict a more accurate representation of the reality of a trial.

A more detailed walkthrough will be described in the subsequent Data curation section. The most pertinent novel features being :

- The exact estimated enrollment count at the start of the study was scrapped using a non official API allowing access to all the versions of a specific study. As accrual is a major issue for termination, using the actual enrollment count rather than the estimated one in the theoretical protocol would be a significant data leakage in our prediction models.
- An NLP categorization of study termination in order to distinguish the termination reason of one study (along with a categorization of individual patient dropout reason not used in our models)
- An accurate geolocation of the study : unlike other studies we won't define the first center of one study as its location but we will rather derive it from the name of the responsible party when possible.

RELATED WORKS

The inspiration for the insight of the exploitation of the eligibility criteria and the use of the SHAP framework to interpret feature importance for tree based models is based on previous work from E. Kavalci and A. Hartshorn, 2023[12].

MATERIALS AND METHODS

Data sources

ClinicalTrials.gov official API

Several public data sources are available for analysis, the largest being the ClinicalTrials.gov database, which currently hosts over 502 990 registered trials worldwide as of July 2024. ClinicalTrials.gov, managed by the National Library of Medicine, is an essential resource for healthcare professionals, researchers, patients, and the public. It provides access to clinical trial registration data, enabling the analysis of trends in trial composition, size, design, and funding.

[12] Kavalci, E., Hartshorn, A. Improving clinical trial design using interpretable machine learning based prediction of early trial termination. Sci Rep 13, 121 (2023). https://doi.org/10.1038/s41598-023-27416-7

We will first query the ClinicalTrials.gov official public API and retrieve the data from all the Phase 2 trials from the period “startDateStruct” 2019-01-01 to 2022-12-31 with the “overallStatus” COMPLETED,TERMINATED or WITHDRAWN. Any Phase 2 trials past 2022 will be kept as an external dataset.

A trial will be put to the set of Rare Diseases trials if it contains within the derived section the MeSH Term ‘Rare Diseases’, defined by the id Rare or D000035583 in the raw data (or as the RDF Unique Identifier D035583).

ClinicalTrials.gov unofficial Backend API

However, we will point out two flaws coming from solely using the raw data from the official ClinicalTrials.gov Official API :

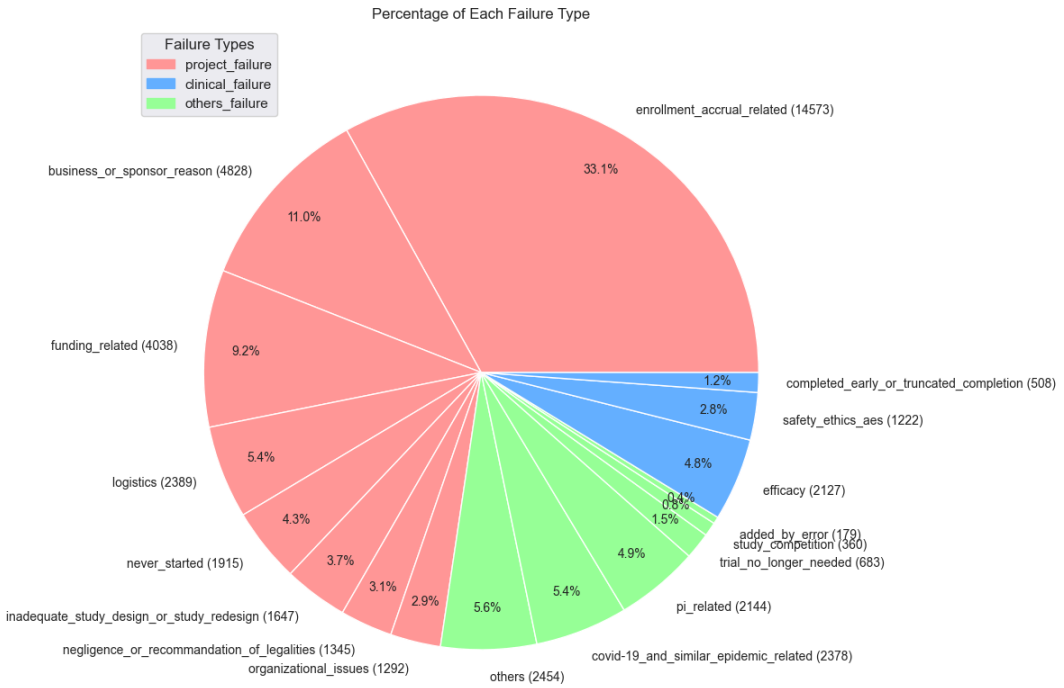
- For multicenter trials, the location can often be incorrect, as the first contact location infos isn’t necessarily the administrative location of the responsible party.
- The enrollment count is the final actual number, rather than the theoretical number defined during the elaboration of the protocol. This may result in a data leakage, as trials with a number of patients significantly inferior to the norm for a trial of said phase would very likely be a trial that terminated early.

Concerning the last two points, using the clinicaltrials.gov public API or the Automatized Aggregated ClinicalTrials database AACT would not solve the issue.
We made a script that parsed for each trial each version to get the latest estimated enrollment count and for the multi-center studies we derived the location from the facility name using GoogleMaps Geocoding API.
Additionally we will fetch data that is only available via ClinicalTrials.gov Backend API, which will provide us additional information across all record versions of each trial in order to retrieve the original planned enrollment count and determine the experimental drug used.

Long Short Term Memory NLP model for classification of early termination cause categories

In a previous study, we trained and built an NLP model from all the trials past 2017 parsing through all the terminated studies and their field stating the reason for the early stop. This model took the input of each study and categorized it amongst one of the 17 causes of early termination. We will use the results of this model to differentiate

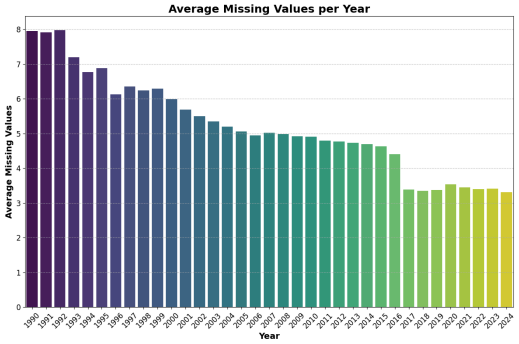
terminated studies caused by toxicity or futility and termination due to project failure such as low accrual, funding, logistics etc.



Pie chart of the distribution in percentage of all type of early termination causes for all studies past 2017

Dataset Preparation

We will only select trials that started past January 2017, as the trials past this date have on average less missing values fields and thus of higher quality.



Average amount of missing field value per trial per year.

However we will only keep the trials from 2019 to 2022, specifically the ones

happening during the covid period, in order to keep the dataset homogeneous. The trials from 2023 to 2024 will be kept as an external validation set.

As the studied period is happening during covid pandemic, many trials which have been otherwise potentially completed without issues were terminated due to the in-person covid restriction, resources prioritization and tainting the pool of eligible patients. We will use the results of our NLP model to filter out of our dataset all the terminated trials caused by the covid restrictions.

This will lead to a higher quality dataset, without the bias of turning the covid into a confounding factor for the outcome prediction.
The benefit of using a NLP model rather than a simple regex filtering out the keywords “covid” or “pandemic” will be the ability to filter out restrictions related termination without removing others covid reasons (i.e : terminations due to the low accrual caused by the decrease of incidence of covid cases won’t be filtered out).

However, we will have to keep in mind during the interpretations that the covid may still have an impact on others studies completed within our dataset.

Selection of variables used to build our prediction models

ClinicalTrials.gov data

We first processed the NoSQL JSON data into several relational tables :

- A main table with all the main variables
- Two tables of primary keys to link all countries and MeSH terms to their respective foreign keys across the others tables
- A table with each MeSH terms and tree/ancestor level, and all the trials associated with these terms
- A table containing all the unique center locations, their geolocation, and the trial nct_id associated with each center
- A table containing all the chemicals used in a trial, along with their categorization as experimental, placebo or active comparator, along with their DrugBank chemical ID.

Curated data

- A table with all the terminated studies and the termination cause determined by our NLP model
- A table with the location of the responsible party of each trials, along with their geolocation derived from the facility name with an algorithm using GoogleMaps Geocoding API

Variables used

From these tables, we extracted the following variables we will be using to train our model :

- the **study organization and lead sponsor names** and their **entity classes**
- whether or not the study has a **Data Monitoring Committee (DMC)**, has an **FDA regulated drug**, is **randomized** or allows **healthy volunteers**
- if it is a multi-phase study **Phase 1 - 2**, **Phase 2 - 3** or standalone **Phase 2**
- **Intervention Model** : single-group, parallel, sequential etc
- **Primary Purpose** : treatment, prevention etc
- **Mask type** and who amongst the Participant, Caregiver, Outcome assessor and Investigator is masked
- The **theoretical enrollment count needed**, as described in the protocol before the start of accrual
- total number of **arm groups**
- **Sex and Age** eligibility
- Administrative **country** and **region** location of a trial
- **Responsible Party**
- count of **Primary and Secondary Outcomes**
- Count of **total centers**, and count of distinct centers per city, country and region
- **Pregnancy** eligibility
- Binary variable stating if the count of outcomes, organization, responsible party and lead sponsor remained the same across all versions of the trial
- Experimental type : experimental treatment only, against placebo or against an active comparator
- total number of **collaborators** and the proportion of industrial class type collaborators
- **MeSH Terms** associated with the trial

Due to the high cardinality of some of the variables and their lack of ordinality, we will turn all these columns into individual binary variables.
For each column we will use the most prevalent value as the dummy variable and drop it.
Lastly, after separating the Rare Diseases and Common Diseases trials into their respective dataset, we drop all the columns where all the trials have the same value.
This gives us a Rare Diseases trials dataframe of 2742 trials and 511 variables and a Common Diseases dataframe of 4007 trials and 542 variables.

Further detailing entity classes using regex

Currently, ClinicalTrials.gov only defines the entity classes as INDUSTRY, NIH and OTHER. We further detailed these classes by searching with regex keywords contained in the name of these organizations and expanded the OTHER category into ACADEMIC, MEDICAL and NETWORK.

Additionally, we parsed the eligibility criterias free text field and created the variables pregnancy inclusion or exclusion.

Details of the regex will be provided in the Additional Information section.

We used a bool variable to indicate if current enrollment count is <75% of the original enrollment count. However the variable alone has too much impact and overshadows the importance of other variables, thus we create interactions of these variables with all the others variables in order to keep the information but lower its impact.

Models building

Definition of the model's outcomes

We will define the outcome as completed versus terminated early. This will not detail whether or not the study outcome was positive or negative as it'd require manually checking the results of each individual primary and secondary outcomes. Moreover, 86% of studies do not end up with results or do not make them publicly available.

Score metrics

We will be using the precision, F1 score and PR-AUC, as our datasets are imbalanced and these metrics are sensitive to class imbalance.
The hyperparameters for each model were calculated previously using GridSearchCV.
The scores will be calculated on a tenfold cross-validation.

Each model will be trained twice on each Rare Diseases and Common Diseases datasets. First time, we will give the information whether a trial succeeds to reach at least 75% of the theoretical accrual goal and the second time we will not share this information and only give the theoretical accrual goal. We do this in order to observe the difference of that one single variable on the score's results.
However we will only use the second variable to calculate the importance of each feature, as keeping this variable will overshadow other more subtle variables.

SHAP Feature Importance

We will then try to calculate the weight of each variable in the model's final outcome prediction.
The lack of accrual is the most well known cause for early termination, we will omit these variables and the categorical slices of numerical values in order to visualize the impact of the other variables by only keeping the value of the theoretical needed enrollment count.

First, in order to see if the list of the most impactful variables is specific to Rare Diseases, we will filter only the top values specific to the Rare Diseases dataset and train again our models and compare the results with our previous scores.

To validate that this set of variables is relevant to Rare Diseases trials, we expect not to see much change in the Rare Diseases model's scores and a decrease in Common Diseases model's scores.

Next, we will overview each model for each variable category. Then, we will summarize all the most impactful variables for the two models into a single plot.

Global interpretability: explaining the factors that influence the overall model output

The goal of global interpretation methods is to describe the expected behavior of a machine learning model with respect to the whole distribution of values for its input variables. With SHAP, this is achieved by aggregating the SHAP values for individual instances across the entire population.
This will be illustrated by the Importance plot of SHAP values, displaying as a bar plot the mean absolute impact magnitude of each variable.

Local interpretability: explaining individual predictions

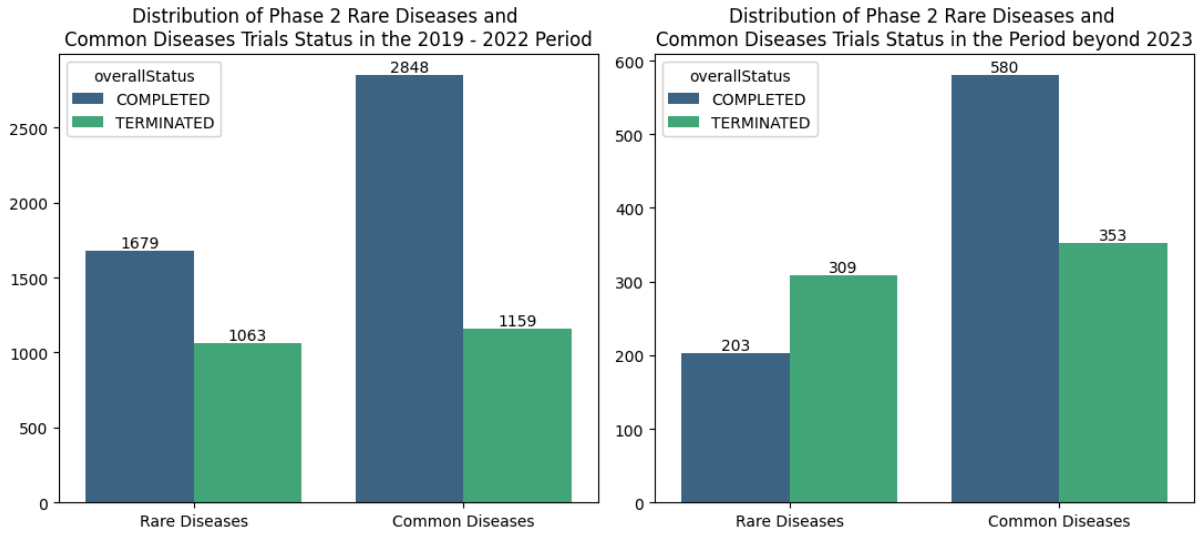
Explaining predictions for individual instances of the data is referred to as local interpretability. SHAP explains how individual predictions are arrived at in terms of contributions from each of the model's input variables. This is a highly intuitive approach that produces simple but informative outputs.
This will be represented by the plot of summary of SHAP values, as a beeswarm, where each dot represents the impact of the variable for a singular instance of a trial.

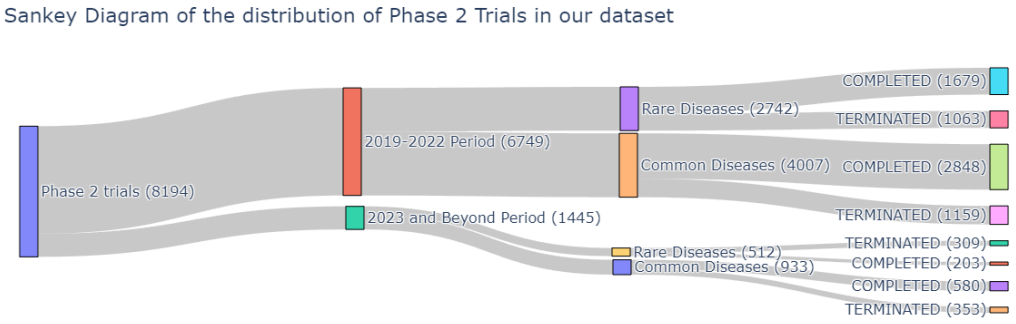
Due to the randomization and that the beeswarms are calculated only on the test set in order to avoid overfitting from the information from the training set, beeswarms generated from one seed to another could potentially illustrate opposite interpretations for variables with low magnitude importance. All effects describe the behavior of the model and are not necessarily causal in the real world.
Beeswarms : The beeswarms is a figure where each dot represent an individual trial, and the impact this specific variable had on this individual trial. The lowest or binary False values are represented in blue while the higher or binary True variables are represented in red. As the outcome of interest TERMINATED is set as the positive class, positive SHAP values represent variables that increase the likelihood of early termination. [13]

[13]Cooper A, Explaining Machine Learning Models: A Non-Technical Guide to Interpreting SHAP Analyses, https://www.aidancooper.co.uk/a-non-technical-guide-to-interpreting-shap-analyses/, November 1st 2021

Therefore we will refer to the mean average impact on model output magnitude and use the beeswarms to illustrate their effects over the different individual test set instances of trials.

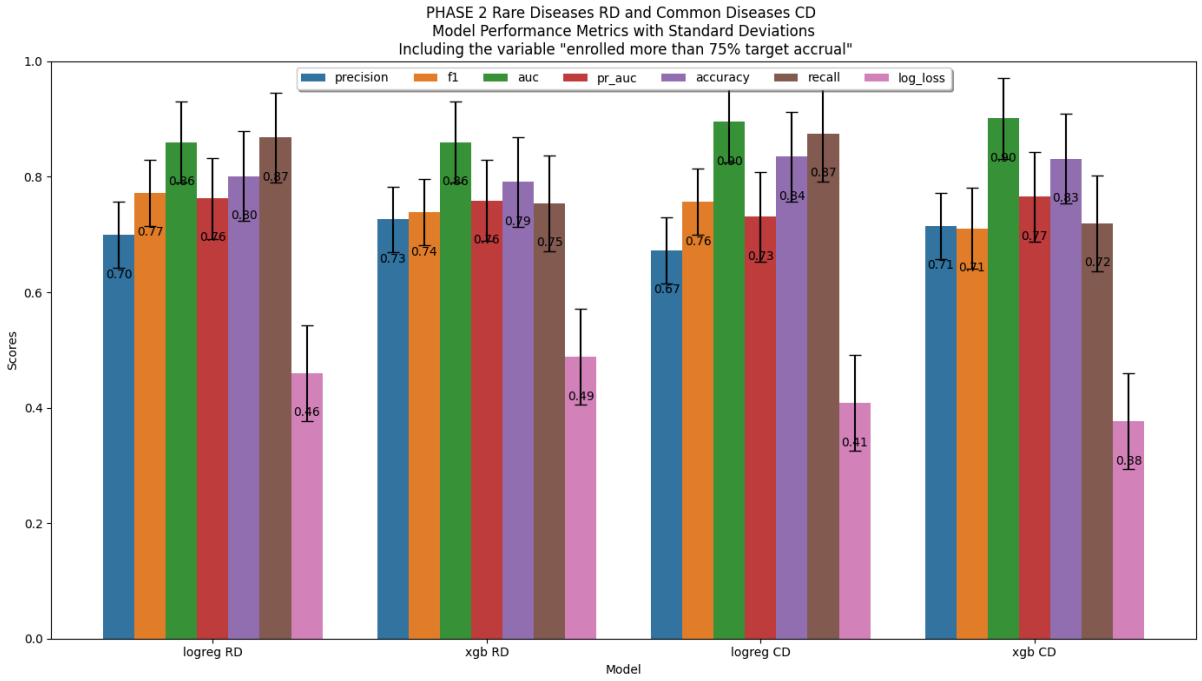
RESULTS





Distribution of the trial's status for the Rare Diseases and Common Diseases main and external set.

Phase 2 Models Results including the variable “enrolled more than 75% target accrual”



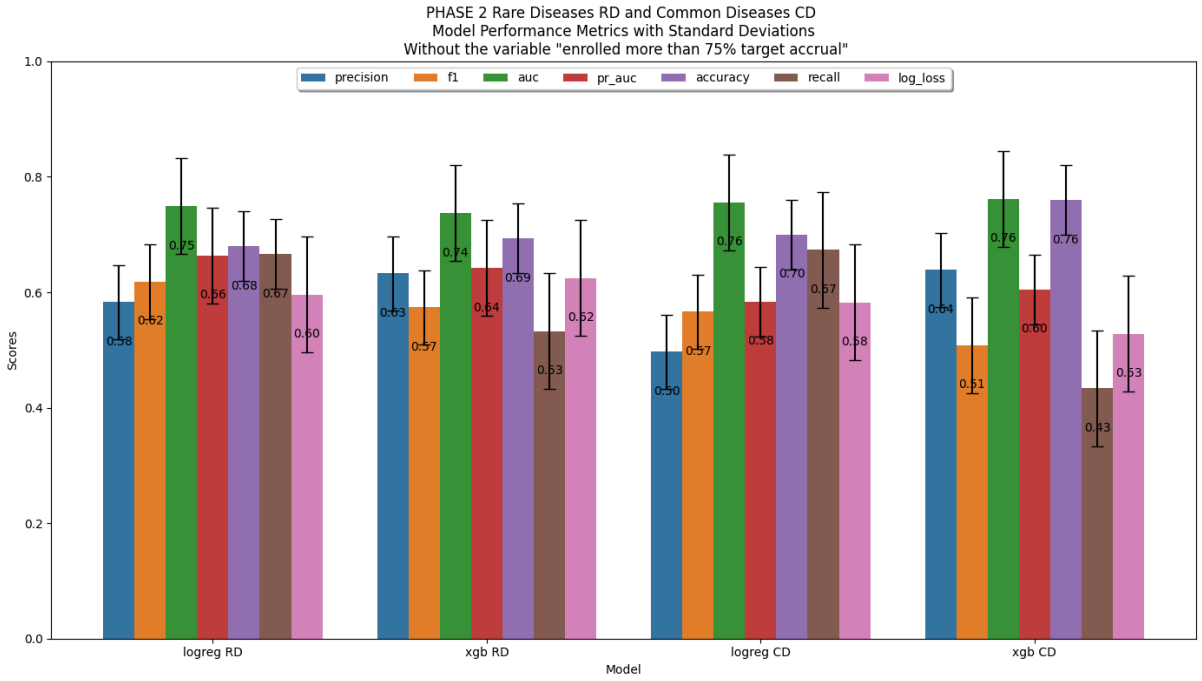
	Precision	F1	ROC-AUC	PR-AUC
non_Rare LR	0.672 ± 0.08	0.757 ± 0.04	0.896 ± 0.03	0.731 ± 0.08
non_Rare XGB	0.715 ± 0.08	0.711 ± 0.05	0.901 ± 0.02	0.765 ± 0.06
Rare LR	0.699 ± 0.06	0.773 ± 0.03	0.860 ± 0.03	0.763 ± 0.06
Rare XGB	0.726 ± 0.07	0.738 ± 0.04	0.860 ± 0.04	0.759 ± 0.07

Models Results on the external test dataset 2023 - 2024

These scores will tell us if the models’ performances are generalizable on trials outside of the training set period.

	Precision	F1	ROC-AUC	PR-AUC
non_Rare LR	0.913 ± 0.09	0.933 ± 0.07	0.957 ± 0.06	0.952 ± 0.07
non_Rare XGB	0.917 ± 0.08	0.913 ± 0.06	0.954 ± 0.06	0.934 ± 0.09
Rare LR	0.972 ± 0.05	0.958 ± 0.04	0.937 ± 0.09	0.970 ± 0.05
Rare XGB	0.972 ± 0.05	0.964 ± 0.03	0.946 ± 0.07	0.974 ± 0.04

Phase 2 Models Results without the variable “enrolled more than 75% target accrual”



	Precision	F1	ROC-AUC	PR-AUC
non_Rare LR	0.497 ± 0.06	0.566 ± 0.03	0.755 ± 0.04	0.584 ± 0.06
non_Rare XGB	0.639 ± 0.10	0.508 ± 0.07	0.762 ± 0.05	0.605 ± 0.06
Rare LR	0.583 ± 0.06	0.618 ± 0.04	0.749 ± 0.06	0.664 ± 0.07
Rare XGB	0.633 ± 0.08	0.574 ± 0.04	0.737 ± 0.05	0.642 ± 0.07

Models Results on the external test dataset 2023 - 2024

These scores will tell us if the models’ performances are generalizable on trials outside of the training set period.

	Precision	F1	ROC-AUC	PR-AUC
non_Rare LR	0.714 ± 0.16	0.689 ± 0.13	0.779 ± 0.12	0.766 ± 0.14
non_Rare XGB	0.732 ± 0.13	0.684 ± 0.12	0.821 ± 0.14	0.836 ± 0.13
Rare LR	0.872 ± 0.06	0.758 ± 0.13	0.744 ± 0.10	0.900 ± 0.05
Rare XGB	0.830 ± 0.05	0.833 ± 0.07	0.752 ± 0.10	0.905 ± 0.05

Phase 2 Models Results with original datasets filtered on most relevant Rare Diseases variables

	Precision	F1	ROC-AUC	PR-AUC
LR Rare Top Features for RD	0.592 ± 0.07	0.626 ± 0.05	0.743 ± 0.06	0.651 ± 0.07
LR Rare original dataset	0.583 ± 0.06	0.618 ± 0.04	0.749 ± 0.06	0.664 ± 0.07
LR non_Rare Top Features for RD	0.485 ± 0.06	0.550 ± 0.03	0.738 ± 0.04	0.558 ± 0.06
LR non_Rare original dataset	0.497 ± 0.06	0.566 ± 0.03	0.755 ± 0.04	0.584 ± 0.06
XGB Rare Top Features for RD	0.625 ± 0.09	0.564 ± 0.05	0.724 ± 0.05	0.632 ± 0.07
XGB Rare original dataset	0.633 ± 0.08	0.574 ± 0.04	0.737 ± 0.05	0.642 ± 0.07
XGB non_Rare Top Features for RD	0.609 ± 0.08	0.494 ± 0.06	0.752 ± 0.04	0.597 ± 0.06
XGB non_Rare original dataset	0.639 ± 0.10	0.508 ± 0.07	0.762 ± 0.05	0.605 ± 0.06

Shap Features Importance and Summary plots

The impact value of each variable is calculated in comparison with all the other variables of the dataset. For ease of interpretation, we will look at the filtered figures containing only the variables for one category. And at the end we will display a summary of all the most impactful variables of each category together to see the range of their impact in a more global picture.

Note that these results only explain how the models make their prediction and aren't necessarily a full picture of real-world scenarios. The main objective is to find which variables have the most influence within our models.

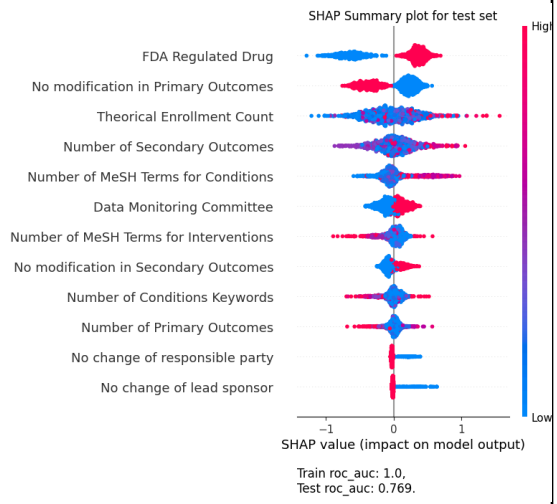
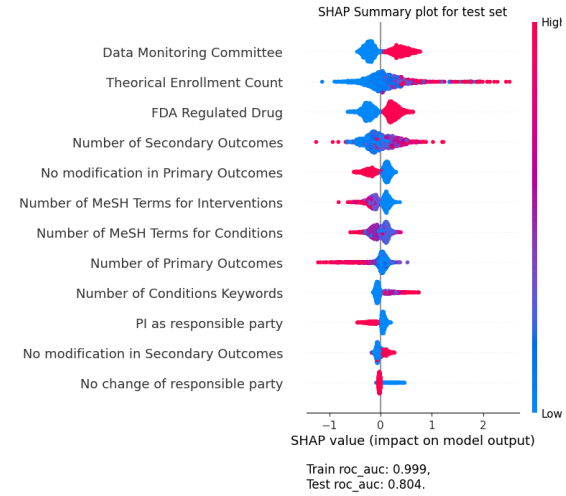
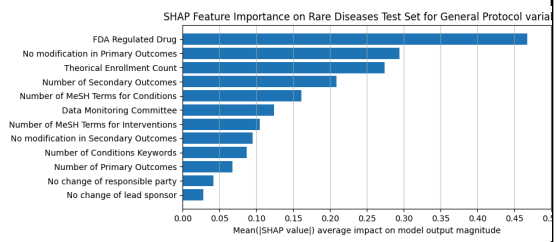
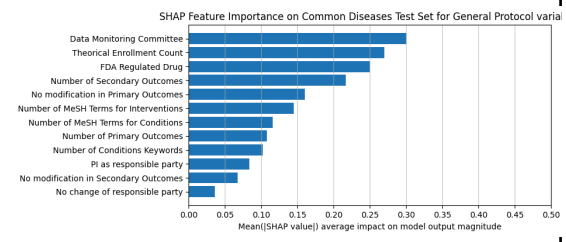
Features Importance barplots

This barplot represents the absolute mean value of each variable. This allows us to see which variables have the most impact on the final model outcome decision.
We will only consider the variables with an absolute mean SHAP value higher than a 0.05 threshold, as these values remain stable over different seeds of the model.

Beeswarms plots

The beeswarms is a figure where each dot represent an individual trial, and the impact this specific variable had on this individual trial. The lowest or binary False values are represented in blue while the higher or binary True variables are represented in red. As the outcome of interest TERMINATED is set as the positive class, positive SHAP values represent variables that increase the likelihood of early termination.

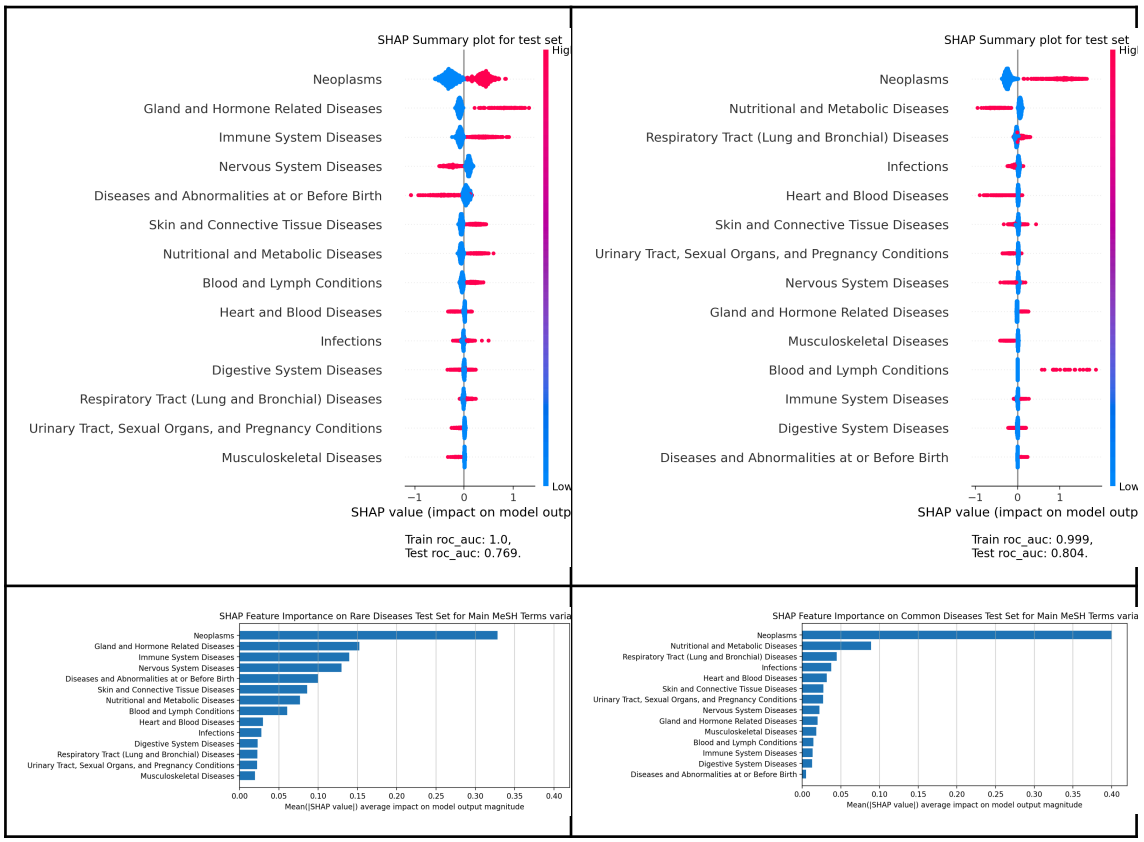
General Protocol variables

Rare diseases	Common diseases
 Train roc_auc: 1.0, Test roc_auc: 0.769.	 Train roc_auc: 0.999, Test roc_auc: 0.804.
	

This category contains some of the main variables found within the record history of a trial. Additionally, rather than using the final accrual number, it instead uses the theoretical accrual number planned in the protocol. Additionally, it provides some information concerning the evolution of changes in the trial such as the change of the responsible parties or count of outcomes.

Main BrowseBranch MeSH Terms variables

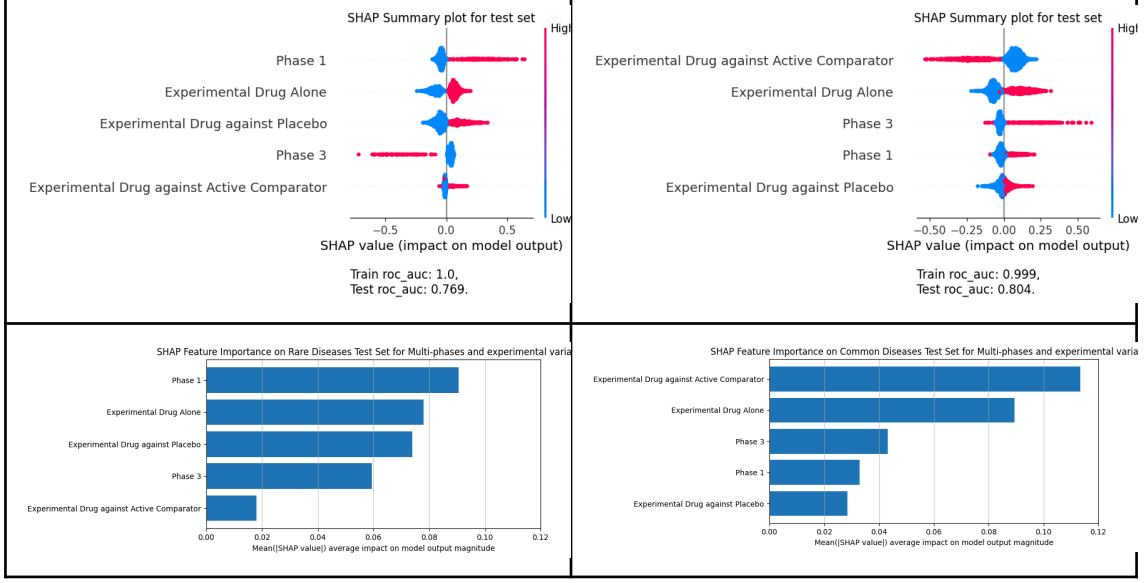
Rare diseases	Common diseases



The current dataset contains to this day over 8500 unique MeSH Terms. In order to limit the dataset sparsity, we will focus on interpreting the 77 unique values designated browsebranches MeSH Terms which are on a level above and encompass the general scope, or physiological systems the MeSH Terms of the trial belong to.

Multi-phases and experimental type variables

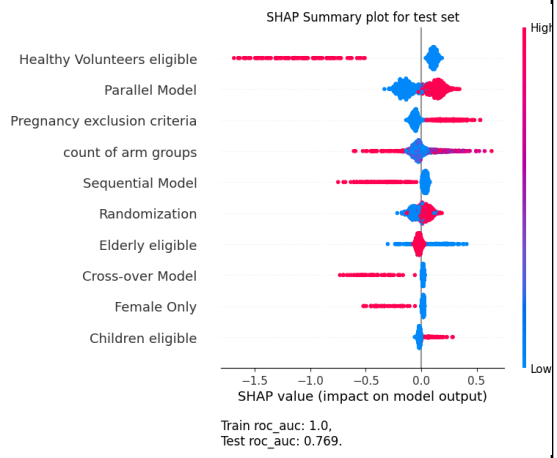
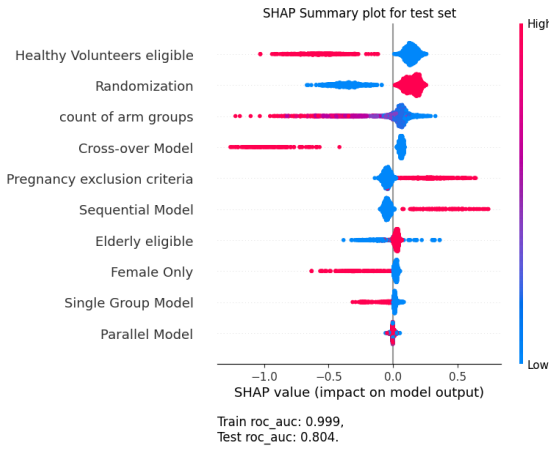
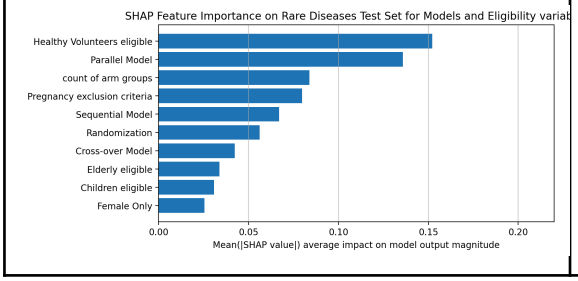
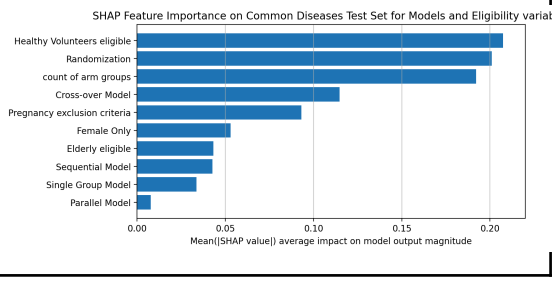
Rare diseases	Common diseases



Some of the Phase 2 trials in the dataset are Phase 1-2 or Phase 2-3 trials. Multi-phases trials can accelerate the research in order to make available faster treatments especially

for pathologies that don't have any standard treatment to date. However, this requires more ahead planning and paperworks than stand alone phases trials.
Additionally, we build an algorithm that retrieves all the drugs used during a trial and matches them to their respective arm group, which allows us to determine which drugs are experimental, active comparator or placebo.

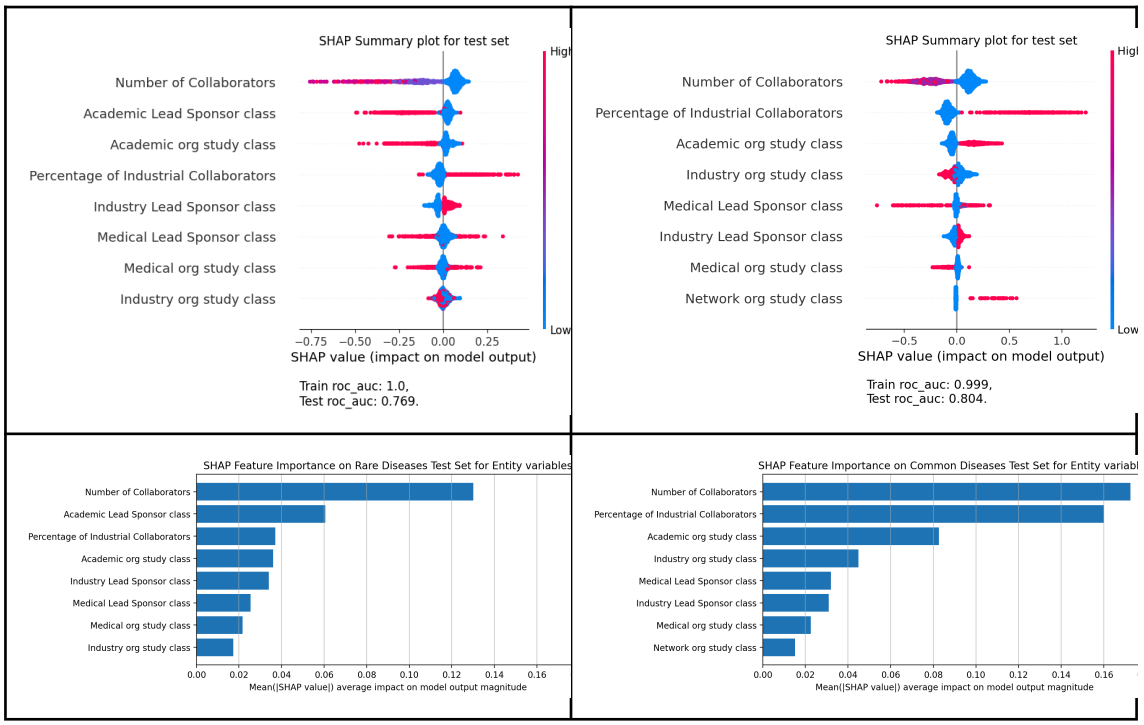
Intervention model and eligibility variables

Rare diseases	Common diseases
	
	

We will compare the different types of intervention models and see if traditional designs are as effective for rare diseases.
Eligibility being one of the main attrition factors, we will see which ones have the highest impact.

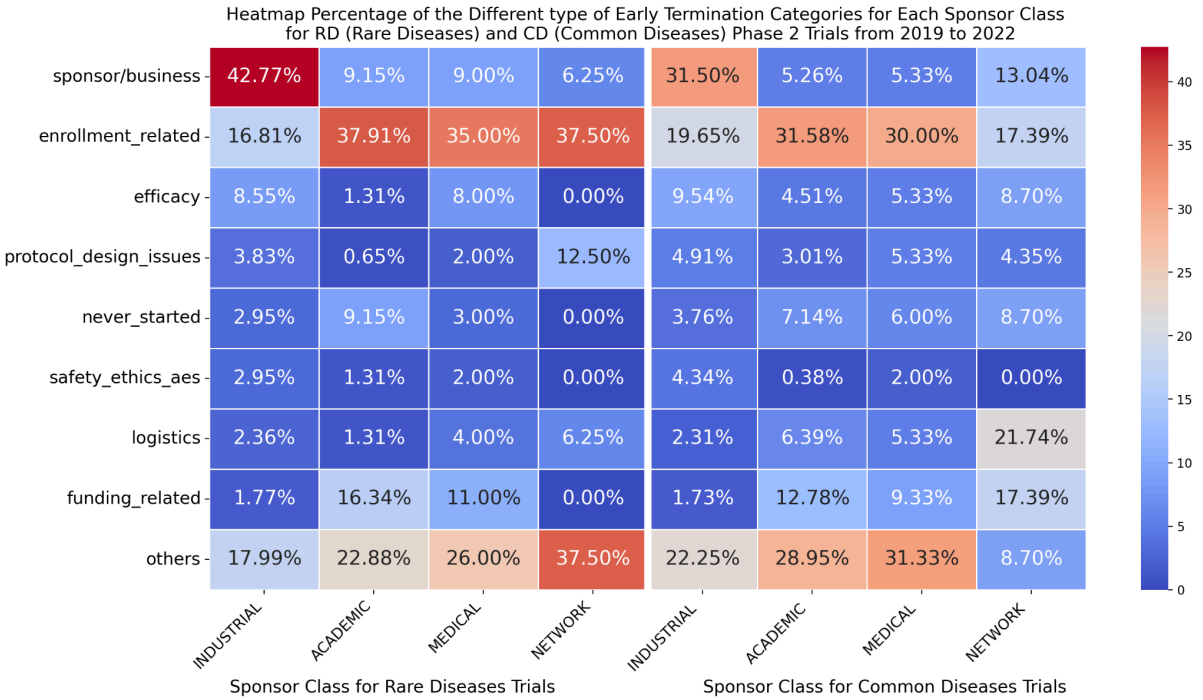
Entity and entity classes variables

Rare diseases	Common diseases



To this date, no study has been studying in detail the impact of the study or lead sponsor class. We will try to see if we find any variables that stand out.

Additionally, we will use our LSTM NLP model to illustrate in a heatmap the distribution of the main causes of early termination for each sponsor class for Rare and Common Diseases.

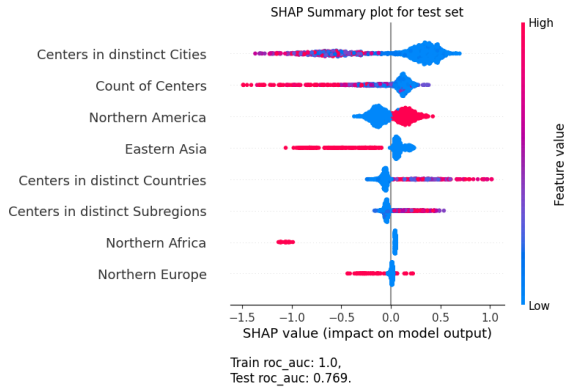
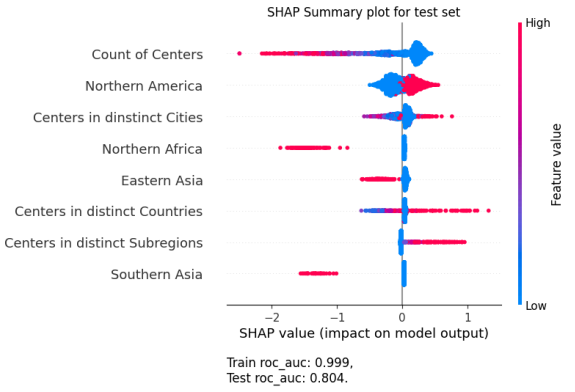
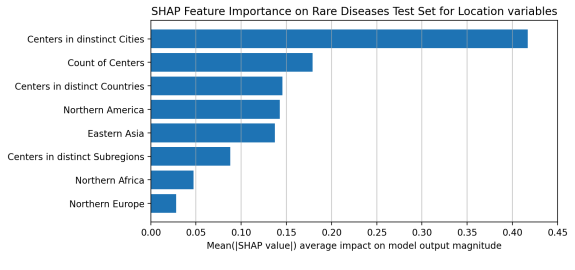
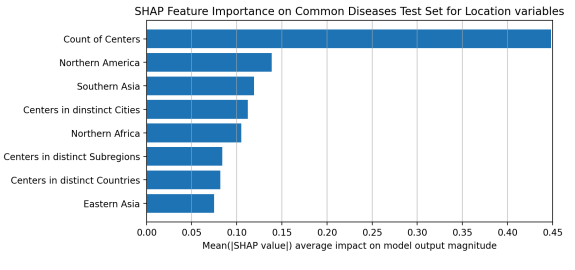


Distribution of the early termination causes for each sponsor classes for Rare and Common Diseases for Phase 2 studies within the 2019 - 2022 period. Vertically, each percentage represents the amount of terminated trials led by the sponsor of each class by termination cause.

Dataset	Rare Diseases Trials / Common Diseases Trials			
Cause	INDUSTRIAL	ACADEMIC	MEDICAL	NETWORK
sponsor/business	145 / 109	14 / 14	9 / 8	1 / 3
enrollment related	57 / 68	58 / 84	35 / 45	6 / 4
efficacy	29 / 33	2 / 12	8 / 8	0 /2
protocol design issues	13 / 17	1 / 8	2 / 8	2 / 1
never started	10 / 13	14 / 19	3 / 9	0 / 2
safety_ethics_aes	10 / 15	2 / 1	2 / 3	0 / 0
logistics	8 / 8	2 / 17	4 / 8	1 /5
funding related	6 / 6	25 / 34	11 / 14	0 / 4
others	61 / 77	35 / 77	26 / 47	6 / 2

Count of terminated trials by trial type, termination cause and lead sponsor class for Phase 2 studies within the 2019 - 2022 period

Locations variables

Rare diseases	Common diseases
	
	

The location of the study was derived from the address of the location of the responsible party of the trial. The sum of all the centers was counted as “Count of Centers”. Specifically, we also counted the number of different cities,countries and regions where these centers were located.

DISCUSSIONS

Phase 2 Models Results

Although being more simple than an extreme gradient boosting model, we can see that the Logistic Regression obtains higher F1 scores for both datasets.
This suggests that the model was able to extract more useful information out of the binary categorical slices of numerical values of the enrollment count rather than the numerical value of the enrollment count. Categorizing into slices with the last slice including all the trials with over a thousand enrolled patients is a potential way to deal with the outliers without filtering them out of the dataset.

When we compare the results with and without the variable enrolled more than 75% target accrual, the scores of the models goes up to 20% higher. This shows that even a moderate deviation from the original enrollment count greatly puts at risk the odds of success of a study.
As the goal of this study is to interpret other variables that might be overshadowed by this lack of accrual, rather than build a predictive model, we will use the model that doesn't use this variable.

Models Results on the external test dataset 2023 - 2024

Despite the model's results being excellent, we lack enough perspective to conclude anything as of now. Due to the set containing an insufficient number of trials, and the set being biased toward short studies that started and completed or terminated within two years. The external test should be done in a few years once we have more available trial data.

Phase 2 Models Results with original datasets filtered on most relevant Rare Diseases variables
By filtering the original 511-column dataset down to 45 top features for rare diseases , we effectively reduced the sparsity of the data while retaining key information. These 45 variables capture nearly as much predictive power as the full set of 511, leading to improvements in logistic regression for Rare Diseases predictions. For example, precision increased from 0.583 ± 0.06 to 0.592 ± 0.07, and the F1 score rose from 0.618 ± 0.04 to 0.626 ± 0.05.

However, when applying these Rare Diseases specific top features to the Common Diseases set, performance dropped, with LR precision decreasing from 0.497 ± 0.06 to 0.485 ± 0.06 and the F1 score from 0.566 ± 0.03 to 0.550 ± 0.03.
The slight decrease in other metrics reflects that the selected features are more aligned with Rare Diseases trials characteristics, leaving some nuances unexplained.

Despite the improvement of the F1 and the precision, other variables decrease slightly, potentially due to the loss of minor but potentially useful information and reduced variability.

Doing the same but on the top features for Common Diseases will yield the same opposite results trend.
This further validates that our following variable interpretations will be more specific for Rare Diseases trials.

General Protocol variables

- Common diseases trials are more highly impacted by the Data Monitoring Committee (DMC) for several reasons:
    - DMCs monitor participant safety and efficacy throughout the trial. If the interim data reveal that a treatment is causing unexpected adverse effects, or if the risk-to-benefit ratio becomes unfavorable, the DMC may recommend stopping the trial early to protect participants.
    - Ethical Considerations: DMCs are tasked with ensuring the trial's ethical integrity, and their conservative approach may increase the likelihood of early termination when there are concerns about continuing without significant benefit[14].
- The impact of a DMC on rare disease trials may be less pronounced compared to regular trials for several reasons:
    - Rare disease trials often have significantly smaller patient populations, making statistical assessments from interim data less reliable. In regular trials, DMCs review interim results to detect trends in safety and efficacy. However, with limited participants in rare disease trials, the data might not be robust enough for DMCs to make early decisions. This can reduce the frequency of early terminations based on interim analyses[15].
    - For rare diseases, there is often little or no established standard of care, which makes it difficult for DMCs to compare new treatments against well-defined benchmarks.
    - Since rare diseases typically lack many treatment options, DMCs may be more hesitant to terminate a trial early, even if interim data suggests modest benefits. The absence of alternative therapies gives more weight to continuing the trial to gather as much evidence as possible.

- FDA regulated drug is the most impactful factor for rare diseases. This is due to stricter regulatory requirements, with demands for more rigorous safety and

[14] Tharmanathan, P., Calvert, M., Hampton, J. et al. The use of interim data and Data Monitoring Committee recommendations in randomized controlled trial reports: frequency, implications and potential sources of bias. BMC Med Res Methodol 8, 12 (2008). https://doi.org/10.1186/1471-2288-8-12
[15] Bierer, B.E., Li, R., Seltzer, J. et al. Responsibilities of Data Monitoring Committees: Consensus Recommendations. Ther Innov Regul Sci 50, 648–659 (2016). https://doi.org/10.1177/2168479016646812

efficacy data, along with well-defined clinical endpoints and larger sample sizes to prove efficacy, which is harder to gather for rare diseases due to limited patient populations and available data[16].
Gathering high-quality Real-World Evidence to satisfy those regulations prove to be harder for rare diseases where the available data is scarce.
Finally, FDA-regulated trials involve more funds and are more time-consuming due to additional approval phases, making them financially riskier and more likely to be terminated prematurely if issues arise[17].

- In addition to the first variables we've described above, original enrollment count has a big impact on both models, as a higher count of needed patients is harder to reach.

- When looking at the number of primary and secondary outcomes, for common diseases, the lower the amount of secondary outcomes the better. However, this seems to be the opposite for the number of primary outcomes. Previous studies[18]stated that a low amount of both primary and secondary outcomes was to be preferred as to not have the trial scope becoming too broad.
  It is to note that primary outcomes count has half as much impact on the overall model than the secondary outcomes count. Therefore, we could potentially think that too many secondary outcomes has a negative impact with more certainty than we could think that more primary outcomes is better.

- For rare diseases trials, although the magnitude of impact of primary and secondary outcomes is in the same order as common diseases trials, the beeswarm seems to be roughly symmetrical. This could indicate that the overall number of outcomes is not a major factor for rare diseases trials. The number of outcomes has less impact on rare diseases because of the small patient population in rare diseases. Phase 2 trials may be smaller and rely on surrogate endpoints (biomarkers) rather than clinical outcomes, since the disease progression is often hard to track over short periods.[19]

Main BrowseBranch MeSH Terms variables

- Neoplasm is the main factor for both types of trials. This is due to the complexity of the oncological field, along with the rise of personalized medicine, making the pool of eligible patients much narrower than for other disease fields.
  The difference of magnitude between the two datasets is due to the prevalence

[16] Chow SC, Pong A, Chow SS. Novel Design and Analysis for Rare Disease Drug Development. Mathematics. 2024; 12(5):631. https://doi.org/10.3390/math12050631
[17] Miller E, DrugWatch,https://www.drugwatch.com/fda/approval-process/, Last modified September 5th 2023
[18] Vetter TR, Mascha EJ. Defining the Primary Outcomes and Justifying Secondary Outcomes of a Study: Usually, the Fewer, the Better. Anesth Analg. 2017;125(2):678-681. doi:10.1213/ANE.0000000000002224
[19] Pizzamiglio C, Vernon HJ, Hanna MG, Pitceathly RDS. Designing clinical trials for rare diseases: unique challenges and opportunities. Nat Rev Methods Primers. 2022;2(1):s43586-022-00100-2. Published 2022 Mar 10. doi:10.1038/s43586-022-00100-2

of the condition: 15% of the trials from the common diseases set contains the MeSH term Neoplasms whereas this number rises to 50% for the rare diseases set.

- The positive impact on common diseases for heart and blood and nutritional and metabolic MeSH terms can be explained by the quantity of trials on these topics. They fall under the category of the top 10 leading causes of premature death in the US within the general population in 2015-2020[20].
  Thus plenty of trials, ressources and documentations are available in order to plan a protocol design less likely to fail.
- Chronic lower respiratory diseases and infections are also significant from already being in the top 10 leading causes of death, the covid situation might also have a potential factor increasing the importance of the features Respiratory Tract (Lung and Bronchial) Diseases and Infections Diseases MeSH terms.
- In contrast, for rare diseases, Immunes or Hormones related trials are more challenging. This is potentially due to the focus on the most prevalent rare diseases of these MeSH Terms that currently do not have a definite cure, such Burkitt lymphoma or autoimmune type 1 Diabetes.
- Diseases and abnormalities at or before birth MeSH might be more feasible as they focus more on the prevention or diagnosis of the unborn individual, rather than focusing on treating the disease. Similarly to Neoplasms, the difference of magnitude importance is related to only 3.6% of the Common Diseases trials containing this MeSH Term compared to 14.5% in Rare Diseases trials. Rare diseases, particularly those with genetic origins, often have well-defined molecular targets, making therapeutic approaches more precise. However, many rare diseases have poorly defined clinical endpoints since they have not been previously treated.When damage occurs before birth, early intervention may be impossible. Despite valid targets, treatments may be ineffective due to irreversible damage. Nonetheless, rare disease trials may benefit from focused patient populations and specific regulatory pathways that allow for more tailored approaches, improving their chances of success compared to the broader and more variable nature of common disease trials.[21]

Multi-phases and experimental type variables

Before the interpretation, we need to remember that in some cases, trials that completed earlier than planned or partially completed (i.e : Phase 1 completed but the results deem a Phase 2 futile) are still classified as TERMINATED in the ClinicalTrials.gov database.

[20] Ahmad FB, Anderson RN. The Leading Causes of Death in the US for 2020. JAMA. 2021;325(18):1829-1830. doi:10.1001/jama.2021.5469

[21] Logviss K, Krievins D, Purvina S (2018) Characteristics of clinical trials in rare vs. common diseases: A register-based Latvian study. PLoS ONE 13(4): e0194494. https://doi.org/10.1371/journal.pone.0194494

- Phase 1 is a crucial step to estimate the toxicity of a drug, thus a Phase 2 is unlikely to happen if there are any toxicity concerns during Phase 1, regardless of other factors. This would explain why the factor Phase 1 - Phase 2 trials are at more risk, especially in rare diseases trials where the use of cytotoxic experimental drugs is prevalent.
- Phase 2-3 are potentially less risky in rare diseases trials, as the average smaller cohort sizes are closer than the required sizes for common diseases trials.

When comparing the different type of experimental type, we need to note the low statistical power and small number of trials in Rare Diseases experimenting against active comparator (due to the low number of Rare Diseases that already have a treatment) and experimenting against a placebo (due to the potential ethical issues to not give a patient a potentially beneficial treatment).

- In Common Diseases trials, comparing against an active comparator has the highest beneficial impact. This is usually the gold standard, allowing us to compare the efficacy of the experimental treatment against an already established one, ruling out the placebo effect.
  We note that, due to the trials' status format in ClinicalTrials.gov, a study has the status COMPLETED once it ends as planned, whether it had positive or negative results, therefore a study can be COMPLETED even if they failed to show superiority of the experimental treatment against the active comparator.
- Trials using only the experimental drug alone have more impact on rare diseases trials as the majority of these diseases do not have any treatment to this date. This also implies testing a novel drug is more risk prone and thus more at risk of early failure. The low number of Rare Diseases trials being able to compare against an active comparator also explains the lack of statistical power (mean SHAP < 0.05) of this value in the Rare Diseases dataset.
  The case of comparing against a placebo may raise ethical concerns, offering a placebo or withholding treatment could raise ethical concerns, as these patients may urgently need intervention. A single group assignment ensures all patients receive potentially beneficial treatment, which is ethically more justifiable in such cases where no other active comparators are available.

Intervention model and eligibility variables

- For both types of trials, having eligible volunteers is a major beneficial factor. In the case of Rare Diseases, especially Neoplasms and cytotoxic treatments, this is rarely feasible, explaining the lower magnitude impact of this variable on the Rare Diseases set compared to the Common Diseases set.
  However, these trials cannot use potentially harmful treatments to the healthy participants and this variable correlates with low risk trials and therefore lower risk of scientific failure. However, the increased amount of accrual could still

have an impact on the complexity of the organizational management, which could potentially increase termination risks.

- Parallel Model is a traditional design type, in order to differentiate the treatment and the placebo effects. The wide usage explains the low impact on the Common Diseases set. However, this model may not always be appropriate for Rare Diseases. Due to the difficulty in recruiting enough patients to fill all arms and the ethics concerning giving a placebo to one arm instead of a potentially beneficial treatment explains why single group models would be more appropriate. This also explains the different impact of the variables Randomization and count of arm groups in the two sets.
- This also explains why sequential and cross-over designs are more adapted for Rare Diseases trials, as these Adaptive Designs include less patients and allows the adaptation of the protocol at pre-specified endpoints.
- Many trials testing novel drugs exclude pregnant women until more is known on its metabolism. However, despite many Rare Diseases concerning infants or elderly, the inclusion criterias doesn't have a significant (>0.05) impact on the model outcome prediction compared to the other variables.

Entity and entity classes variables

Firstly, we see that an increased number of collaborators is beneficial, as it allows more resources to conduct the trial.
Apart from the variables Number of Collaborators and Academic Lead Sponsor class, all the other entity variables have a mean SHAP impact inferior to 0.05. Therefore we can only hypothesize their interpretations.

The industry ratio, which represents the ratio of industrial collaborators out of all collaborators for one study, may not be always beneficial; they potentially may have more resources and funding but their collaboration can be potentially influenced by ressources prioritization amongst all the concurring trials they are sponsoring.
This has less of an impact for rare diseases trials, potentially due to their prioritization on diseases without any current treatment and programs like Orphan Drug Designation (FDA/EMA) that offer regulatory flexibility, tax incentives, and longer exclusivity periods.

Seemingly counter-intuitive, industrial sponsor class is linked to increased early termination risk. This can be again due to resources and funding prioritization.

Universities as lead sponsors for rare diseases have the most impact, potentially due to their focus on academics and less influence on commercial impact. With half as much impact but still the most important entity class factor for normal trials is the organization study being a medical entity, as they have direct access and are more self-reliant for resources needed for the trial and decreased risk of manufacturing or logistics penury issues.

Network entities seem to be the class struggling the most. However the low mean SHAP value and the low count of network entities in our dataset is insufficient to give enough insight to this interpretation.

To give more weight to our hypothesis, we use the LSTM NLP model previously created that processes the free text field filled by trials that ended prematurely explaining the cause of the termination. This model classifies into several labels the potential cause of termination.

From these percentages, we see the main cause of termination for industrial entities is due to business or sponsor reasons rather than scientific or enrollment reasons.
In the same inverted trend, industrial entities have the lowest amount of termination due to funding or enrollment issues.
This indicates the large potential amount of resources that industrial entities have facilitated reaching more potential participants. But paradoxically might be less stable due to the potential business incentives reasons and resources allocation amongst the many trials they are sponsoring.

Locations variables

A higher number of centers is beneficial for both types of trials, reducing the likelihood of complete termination if one of the centers stops contributing.
For common diseases Phase 2 trials, the total number of centers, regardless of them being within the same cities or not, has the highest impact. This is likely due to the need for these trials to accrue a larger number of patients, and the interventions being able to be done in most centers without the need of highly specialized personnel or equipment. In comparison, for a given rare disease, a single city is unlikely to have a large number of highly specialized institutions, therefore the emphasis on the need to spread through several large cities.

Having centers abroad might actually be a negative complicating factor, as having centers across different countries with different healthcare system structures may further complicate the synchronized management of the study globally.

An additional point to take in account is despite the benefit of increased resources, more centers implies an increased administrative paperwork load, sample delivery, regular audits etc .. which may cause a strain on the workforce and overall trial. [22]

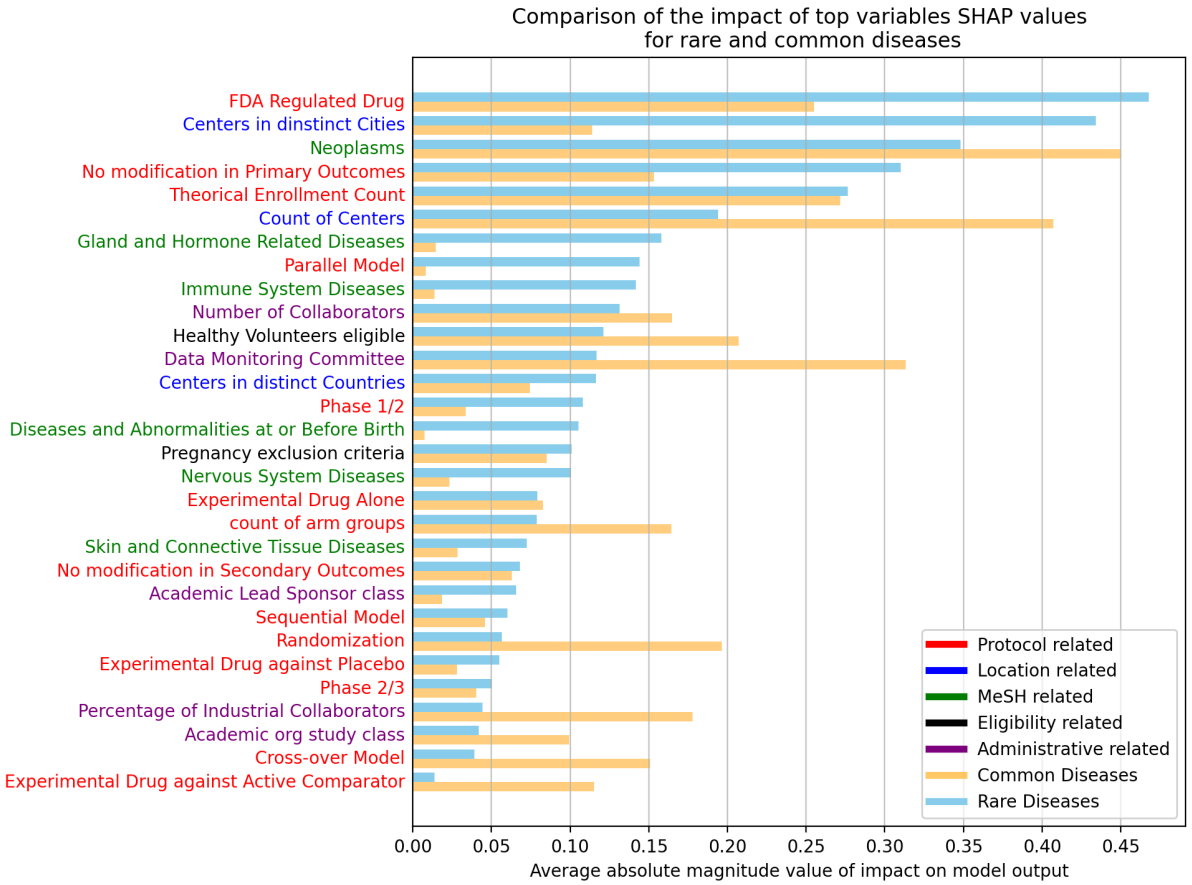
[22]Booth C, Parexel, Advancing rare diseases drug development Report, Section 2 Effective regulatory strategies,Part 7 https://www.parexel.com/insights/new-medicines-novel-insights/advancing-rare-disease-drug-development/study-design-and-execution-rare-diseases/how-sites-manage-pediatric-gene-therapy-trials

A noteworthy point is that the factors Northern Europe, Northern America and Eastern Asia are shown here as significant factors due to the fact that France, China and the United States are amongst the top three countries where the trials are started within our dataset.
However, Northern Africa is solely carried by Egypt and especially Turkey, which have a significant rate of trial success due to the benefits made available for clinical studies in these countries. This impact is more prevalent for common diseases as the dataset doesn't contain as many rare diseases trials done within those two countries.

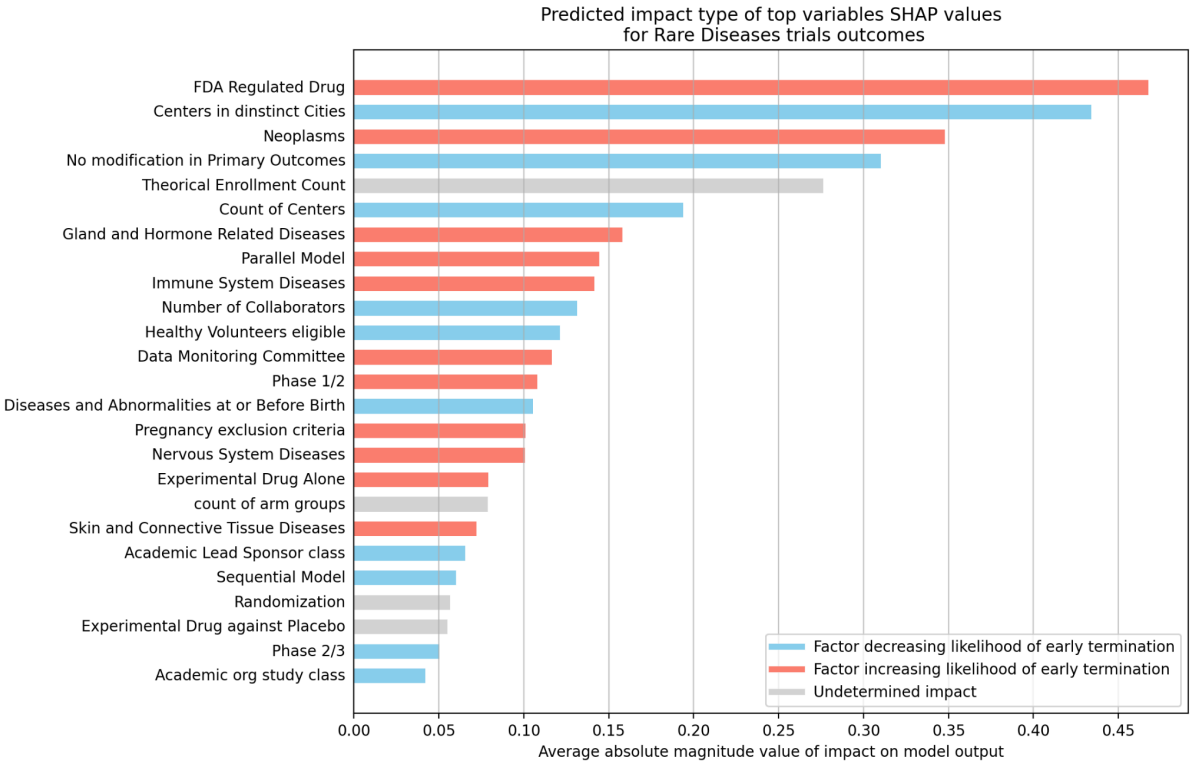
Turkey, in particular, stands out as an underrated but highly effective location for conducting clinical trials. The country has built a robust infrastructure, making it an increasingly attractive option for trial sponsors. Mainly thanks to the high percentage of treatment-naive patients who are more accessible and interested in participating in trials compared to those in Western Europe and the USA. Additionally, Turkey's clinical trial environment is characterized by full compliance with Good Clinical Practice (GCP) and quality standards, moderate clinical research costs, and a concentration of trial sites in major cities. The regulatory environment is also favorable, with no requirement for an EU QP statement or an Investigational Medicinal Product Dossier (IMPD) to initiate trials. These factors contribute to Turkey's efficiency and cost-effectiveness in clinical trial execution, making it a strategic choice that is often overlooked by sponsors focused on more traditional markets[23]. This may explain why Turkey consistently outperforms other regions in terms of trial success.

[23] https://smoothdd.com/en/company/geography/smooth-drug-development-in-turkey/

Summary of the most impactful variables in the final model output



Summary Plot of the most influential variables predicting early termination in Phase 2 Rare Diseases trials and their impact magnitude for Rare Diseases and Common Diseases trials.
Some variables have been omitted for clarity or relevance (i.e : continents, count of outcomes etc)



Summary Plot of the most influential variables predicting early termination in Phase 2 Rare Diseases trials and the observed type of impact they have on the model's outcome prediction

To summarize, we took the most impactful variables of each category and compared them against each other. The following figure represents all those features, sorted by magnitude impact for rare diseases trials.

We can see that the main factors for rare diseases trial termination are :

- using FDA regulated drugs, which impose several reglementations that may increase the complexity of a trial whereas the Data Monitoring Committee does not have as much of an impact as it has with Common Diseases trials.
- Having centers in multiple cities which indicates the benefit of having more access to resources and patients. This also highlights the need to have access to specialized infrastructures.
- No modification in Primary outcomes across study records, which indicate trials that didn't have to adjust their design mid-way are more likely to be successful, rather than the impact of the number of outcomes.
- Commonly used models such as parallel treatment or comparing the experimental product against an active comparator and placebo doesn't have the same impact for these Phase 2 trials compared to common diseases trials due to the unique challenges of these diseases.

- After Neoplasms, Endocrinal, Immune and Tissues Mesh terms appear due to the higher amount of terminated trials than completed trials in our dataset. Further analysis would be required to understand the actual causes.

Most of the variables are protocol related, implying the importance of the careful choice of these parameters in order to maximize likelihood of success.

The MeSH term Neoplasms has the most impact over all the other MeSH terms, as it differentiates itself due to the low prevalence of eligible patients, and the heterogeneity of the population and the lack of historical knowledge.

From the table comparing the proportions of different early termination causes, we saw that due to their resources, industrial entities are less likely to terminate early due to lack of accrual or funding issues but are highly susceptible to terminate due to business or sponsor reasons. Even without as much available resources, Universities and medical entities are seen as a more reliable sponsor source.

LIMITS OF THE CURRENT SCOPE OF THE STUDY

ClinicalTrials.gov as the sole source of data

This whole study is based on information gathered from the database ClinicalTrials.gov. In order to further validate our findings, in future works, these models might need to be done using additional data from several different databases.
Furthermore, the interface may not be user-friendly. In the case of consulting a single trial record it is clear enough. However, for the need to do more global studies requiring the comparison of hundreds of trials, this may complicate the search of more intricate query for someone not used to work with APIs and the parsing of JSON NoSQL raw data.

Additionally, we had to code additional scripts to retrieve the theoretical enrollment count and determine the study's experimental drug and correct administrative location of trials. Information that is not directly available in the raw data.

Phase 2 trials

This study focused on Phase 2 trials, as this is the phase more sensitive to organizational and protocol factors whereas the Phase 1 and 3 is highly influenced by other factors such as the ADME factors of the investigational product or the eligibility criterias.

Although the same methodology can be applied to the other phases, a dataset containing more information on these factors would be required in order to have a more accurate interpretation.
In future works, we could query the DrugBank.ca database using the extracted experimental drug names and include several new factors such as potential drug interaction or toxicity related in order to strengthen the accuracy results of the Phase 1 predictions.

For the Phase 3 trials, the eligibility criterias play a major role defining the eligible population able to be accrued. ClinicalTrials.gov contains a free text field detailing the inclusion and exclusion criteria that could potentially be exploited using a Natural Language Processing model to create more variables related to Phase 3 trials.

Lastly, several other potential factors unrelated to protocol design may have an impact on Phase 2 trials as well that weren't included in our model's dataset.

The distribution of trials in the Rare Diseases and Common Diseases set

The lack of statistical power in some variables like the MeSH terms is due to the sparsity of some of them within the sets. Complementary analysis focusing more specifically on one or a few of them would be required in order to have a more detailed picture.

Impact of keeping terminated due to low accrual trials in the set

Keeping trials that have been terminated due to low accrual without actually giving the information (the final low accrual number compared to the theoretical enrollment count needed) can yield some issues. The model could end up predicting the trial as completed if the theoretical accrual was reached or terminated by attributing the cause to other factors.
However, we decided to keep these trials, as despite the accrual being the official termination reason, it will also involve some other factors that increased this termination. For instance, a low number of centers, not enough collaboration or too stringent eligibility terms would also be factors contributing to the termination. And it would also help the model identify beneficial factors leading to the completion of a study.

In future works, we could potentially clusterize trials depending whether or not they reached more than 75% of the theoretical enrollment count and compare the difference in variable impacts.

Impact of covid trials in the set

The pandemic had a significant impact on the clinical trials landscape, including many restrictions, resources prioritization and more resources to find a vaccine. Although we

were able to filter out the early terminations caused by these restrictions, the remaining trials may have been influenced in a way or another. Along with the limitations of the generalization of our models, this would require the models to be trained and applied again in the next few years when the impact of the pandemic will be less prominent and cross validate our current findings.

The nature of SHAP values

SHAP feature importance is an alternative to permutation feature importance. There is a big difference between both importance measures: Permutation feature importance is based on the decrease in model performance. SHAP is based on magnitude of feature attributions.

The feature importance plot is useful, but contains no information beyond the importances. For a more informative plot, we use the summary plot displayed as a beeswarm plot.
However, it is to note that this plot only describes the behavior of the model. Which explains why its trends may slightly vary over different seeds.

NLP Limits

The current NLP model plateau around 60% accuracy on the validation dataset. Given the relatively large quantity of labels, obtaining higher accuracy would require a more complex model or a different approach.
Another limitation is the quality of the training dataset. This was done by a single annotator over the course of several days and the labels have been subject to multiple changes, which may lead to a subjective bias. One improvement would involve the use of several annotators in order to bring the methodology closer to the gold standard of annotating.

Generalization of the models' scores

One significant limitation in the external validation is the potential bias introduced by using data from 2023 and 2024 for external validation. These years encompass a specific subset of short-duration studies, which both started and completed within the two-year span, potentially skewing the results. Furthermore, 2024 represents the year with the fewest trials in the dataset, and is likely biased by a higher proportion of terminated studies, reducing the statistical power of the results for this period.

CONCLUSION

In conclusion, we have seen a large panel of different protocol and organizational variables. Especially rare disease trials, focusing on complex fields such as oncology or working with FDA regulated drugs that involve a lot of complicating restrictions, tend to experience higher trial failure rates, reflecting the inherent challenges of developing treatments in these areas. Standard intervention models methods such as Parallel cohorts or high theoretical enrollment count does not work as well for rare diseases trials that already struggle to accrue enough patients.
However they do benefit from having a high number of centers, spread through several cities, which helps reaching more eligible participants and having multiple collaborators, notably academic entities, in order to have access to more specialized resources in order to carry the study to term.

Additionally, indicators such as unchanged end outcomes across the record history of a trial, signifying not any major change in the protocol design during the trial is also a potential good factor to predict the final outcome of a study.
Lastly, our model was able to capture the influence of the regions where trials were more at chance to complete. It also highlighted the potential benefits of conducting trials in Northern Africa, specifically Egypt and Turkey, where their robust infrastructure and high percentage of treatment-naive patients make them good options when deciding where to plan a clinical trial.

The interpretations of specific feature categories, especially when comparing rare diseases trials against general phase 2 trials, underscores the importance of context in interpreting these findings. It highlights the need to carefully consider each factor within its broader context, avoiding a narrow focus that could lead to misinterpretation.

Additional Information Section

Regex

Defining new class variables for organization study and sponsor class

Sub-class for OTHER	**Keywords**
UNIVERSITY	universit%, college, school, academ%, faculty, campus, polytech% etc
HOSPITAL	hospit%, medic%, clinic%, klinik%, hôpit%,hopit%, krankenhaus Ospedale etc
NETWORK	group, network, society,foundation etc

Pregnancy Inclusion and Exclusion algorithm

$$\begin{aligned}\text{Pregnancy Inclusion} = \{C \mid C \subset \text{“inclusion\_criteria”}\} \cap \{C \mid C \text{ contains “pregnancy\_related”}\} \\ \cap \{C \mid C \not\supset \text{“negation\_words” followed by “pregnancy\_related”}\} \\ \cap \{C \mid C \not\supset \text{“contraception\_words”}\} \\ \cap \{C \mid C \not\supset \text{“negative\_test\_pattern”}\}\end{aligned}$$

$$\begin{aligned}\text{Pregnancy exclusion} = (\{C \mid C \subset \text{“inclusion\_criteria”}\} \cap \{C \mid C \supset \text{“negation\_words” followed by “pregnancy\_related”}\}) \\ \cup (\{C \mid C \subset \text{“inclusion\_criteria”}\} \cap \{C \mid C \supset \text{“pregnancy\_related”}\})\end{aligned}$$

$$\begin{aligned}((\{C \mid C \supset \text{“negative\_test\_pattern”}\} \cup \{C \mid C \supset \text{“contraception\_words”}\})) \\ \cup \{C \mid C \subset \text{“exclusion\_criteria”} \cap C \supset \text{“pregnancy\_related”}\}\end{aligned}$$

***Regex Patterns**:*

- *pregnancy_related: Matches words related to pregnancy, lactation, procreation, or childbearing.*
- *negation_words: Matches negation words like "no", "non", "not", etc.*
- *contraception_words: Matches terms related to contraception, effectiveness, or any other terms related to preventing pregnancy.*
- *negative_test_pattern: Specifically matches "negative" followed by "test".*

NLP Model Pipeline

We define termination reasons under three major failure types:

1. Project failure, e.g. the budget has been overspent, project targets haven't been achieved and deadlines haven't been met.
2. Research failure, e.g. not being able to reach statistical significance in a research area and so failed to prove the efficacy of a drug or obtain controversial results.
3. Others, events that could not been predicted such as personal health issues or the pandemic

The "Principal Investigator related" reason will not be placed in the "project failure" category for subjective reasons : a significant amount of termination reasons mentioning the principal investigator is due to personal reasons (health issues, death, end of contract) which aren't preventable like the other reasons.
In addition to that, a majority of reasons mention the principal investigator terminating the study due to another project failure reason, therefore the real reason wouldn't be because of the investigator themselves.
Studies mentioning covid are classified as trial no longer needed if the issue is that the goal of the study is no longer relevant with the rise of covid vaccines or decreased infection rate.

Dataset extraction

First we extract the values of nct_id with the why_stopped field as a csv. As the reasons are written in english and do not involve medical specific vocabulary we will use a model adapted for common vocabulary.

Spellcheck

First step of text cleaning is to try to catch any misspelling that would result in loss of valuable information. We will run these misspelled words through a spell-checker. In order to do this we will use the transformer model spelling-correction-english-base from oliverguhr[15]. Unlike regular python spell-check packages, the model will look at each word in the context of the sentence in order to make proper decisions.
NB : For this reason we will beforehand replace all the "pi" or "pts" contractions as "principal investigator" or "patients" respectively.
For computational reasons considering the size of our corpus, we will first create a list of the value count of each individual word instance and only keep the instances that appear less than two times in the whole dataset. This might contain highly domain specific terms but also potential misspelled common words as it is unlikely for the same typing error to appear more than twice. We cannot run the transformer on the word on

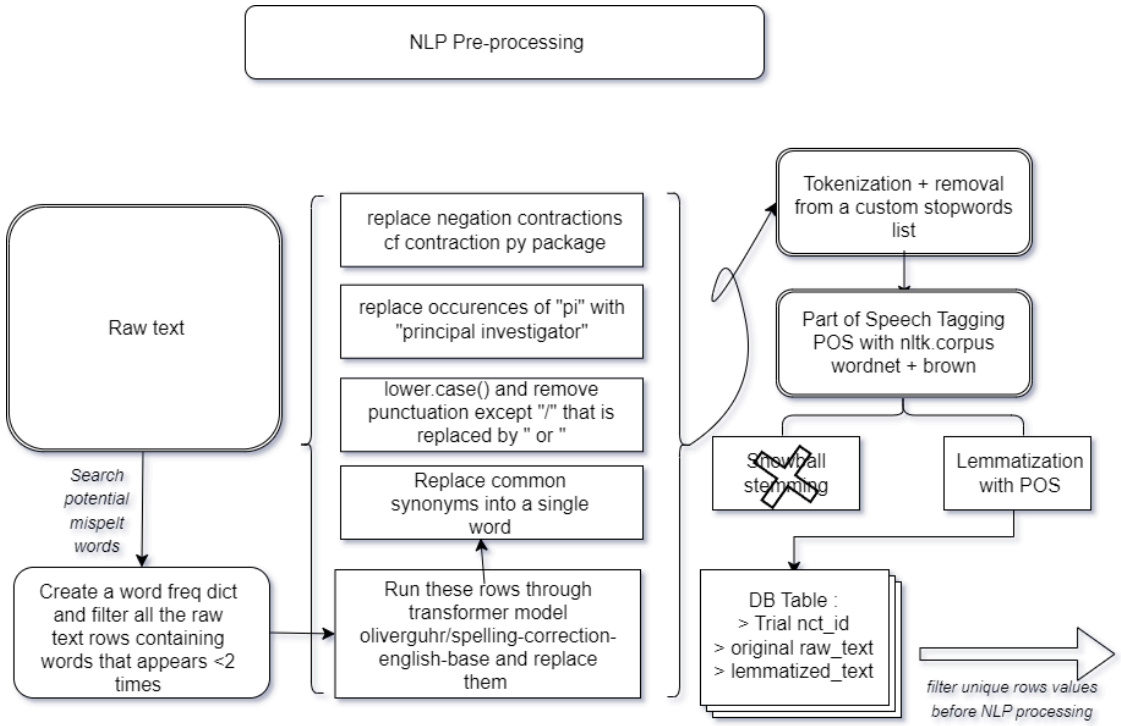
its own as we need it in the original sentence to give it context. From the filtered word count list we will only select the rows in our dataset containing those rare words and run these rows through the transformer model that will process the sub-selection in roughly an hour (versus seven hours for the whole corpus).

Text preprocessing

- Using the python package “contractions” we will replace negative contractions.
  (i.e : mustn't've => must not have)
- Set everything to lowercase and remove punctuation with the exception of dashes that will be replaced by the stopword “ or “.
  (i.e : lack/refusal of patient => lack or refusal instead of lackrefusal)
- To simplify the corpus, we convert the most common words to their plural synonyms.
  (i.e : volunteer, patient, individuals, subject => patients)
- The tokenization after filtering out common stopwords with a custom stopword list that does not remove negations and temporality.

Lemmatization with Part Of Speech

Using nltk.corpus wordnet + brown we will add a Part Of Speech (POS) to each word in each sentence that will allow us to perform a lemmatization of the text. Unlike stemming such as Snowball stemming that reduces the words to their apparent root, lemmatization will reduce words to their etiologic roots depending on their POS tagging. *(i.e “happening” will be reduced to the token “happen” or “happening” depending of their use as a verb or as an adjective)*
(See the text preprocessing pipeline in Fig 7.1 in the Figures Section)



NLP model building

The set is first manually annotated up to 40%. Then we will convert the sentences back into tokens. We will set the max_features to the number of unique words in our corpus which is 9273 here and the max_len to the size of the longest sentences which is 41 words. All the other sentences will be padded to 41 tokens in order to have all sentences with the same length.
We will then use the Pre-Trained Global Vector for Word Representation GloVe 200d to embed our training set into a high-dimensional space embedded matrix.

We will use a Long Short Term Memory LSTM layer. LSTMs are a type of recurrent neural network (RNN) that are effective at learning from sequential data over time.

Preliminary tests show that the model is very prone to overfitting on our training set. Therefore, the final model was fine-tuned to learn without relying too much on previous iterations' results.

First we will perform the embedding of our text data to add a third dimension in order to fit the dimensions of the LSTM layer. Each token is represented as a vector in a continuous vector space.

We pass it to a Dropout layer which indiscriminately "disable" some nodes so that the nodes in the next layer are forced to handle the representation of the missing data and the whole network could result in better generalization.

We set the dropout layer to drop out 50 % of the nodes, which helps prevent overfitting by ensuring the model does not become overly reliant on specific features.

After a dropout layer, we connect the output to a densely connected layer to produce an output dimension of 50 and then pass through a RELU function, which will either return 1 or nothing. Then applies a L2 regulation penalty in order to penalize large weights more heavily and increase resistance to overfitting.
The downside of an increased l2 penalty is a slower convergence speed, hence why we perform a large number of iterations.

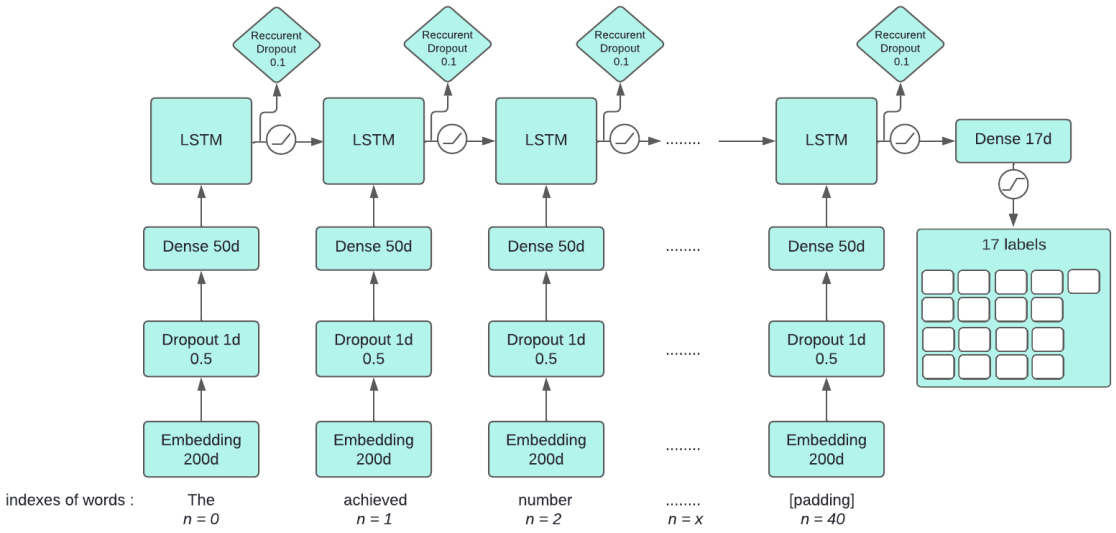
Next is a Long Short-Term Memory (LSTM) layer with 64 units. The dropout=0.1 parameter applies dropout to the input of the LSTM layer, and recurrent_dropout=0.1 applies dropout to the recurrent state, both of which help prevent overfitting.

Finally the output layer returns a dimension with 17 units, each of them corresponding to one of the labels to be predicted. The sigmoid function converts the result to a value between 0 and 1 which is the prediction score for each label.

The learning rate is set to 0.002 using the Adam optimizer which combines the benefits of Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp).

We set the model to train on 501 epochs with a batch_size of 264. We will create a custom class and functions to let us save the model every 25 epochs past the 300th epochs.

For each saved epoch, we will calculate the predictions on the test set and calculate the “error rate” by dividing the sum of rows with a prediction score lower than our threshold by the total count of rows.



Discussions of the model's results

After the iterations of the latest epochs, we select the 375th epoch having 9366 predictions above the 50% threshold out of the 15008 rows or 38.0% error rate.
The respective training and validation scores are 0.7296 and 0.6832 for the accuracy and 0.0737 and 0.0758 for the loss.
We will extract all the predictions with a score higher than 50% and we will manually annotate the remaining ones with the assumption that they contain words with not enough frequency in the corpus to allow the model to properly categorize them.
We can then use those labels to filter out noise data such as terminated trials labeled as "added by error" and filter out overfitting data such as trials labeled as "covid-19 related termination" out of the training set for our predictive models.

Aside from the random spikes due to the imbalance of the 17 labels, we can see from the loss that our current model does not under or overfit. However past 200 epochs we see that our model plateau and more training won't improve its accuracy.
Testing different extreme parameters, we see that the training accuracy will always stabilize around 60-65% even when overfitting. Considering the quality of the data and 17 labels being a large number of labels in which some of them are grouped smaller subsets of reasons, the imbalances of the labels and the use of embedding, it is unlikely that we could further improve this score without a more complex model.
Another caveat is the annotation subjectivity, ambiguous reasons and lack of multiple annotators could lead to a decrease of the training data quality.
The lack of proper filling of the why_stopped field in some studies can also be another factor regarding the data quality.
These elements suggest that the ClinicalTrials.gov trial registry form could implement along the free text field a more formatted category where trials are required to specify the cause of the trial termination along with the free text field if they deem more details to be necessary.

THANKS

I’d like to thank my director ARNOUX A. and KATSAHIAN S. for their inputs on the construction of this thesis.
FISMAN P. and MARTELLI N. for their interest in the topic and GUIHENNEUC C. for supervising the conformity of the paper.

REFERENCES

[1] Research and Markets. Clinical Trials Market Size, Share & Trends Analysis Report, 2021 - 2028. Research and Markets; 2022.https://www.researchandmarkets.com/reports/4396385/clinical-trials-market-size-share-and-trends

[2] Torres-Saavedra PA, Winter KA. An Overview of Phase 2 Clinical Trial Designs. Int J Radiat Oncol Biol Phys. 2022;112(1):22-29. doi:10.1016/j.ijrobp.2021.07.1700

[3] Rare Diseases Clinical Research Network. What are rare diseases? Rare Diseases Clinical Research Network. https://www.rarediseasesnetwork.org/about/what-are-rare-diseases.

[4] Novotech. Published September 13th 2023, *https://novotech-cro.com/faq/rare-disease-clinical-trials-unveiling-insights-and-charting-progress*.

[5] Thom EA, Klebanoff M., Issues in clinical trial design: Stopping a trial early and the large and simple trial, Research Methods: State of the ScienceVolume 193, Issue 3, P 619-625,September 2005, DOI: 10.1016/j.ajog.2005.05.061

[6] *Lan KK, Wittes J. The B-value: a tool for monitoring data. Biometrics. 1988;44(2):579-585.*

[7] Allucent, *https://www.allucent.com/resources/blog/why-do-clinical-trials-fail*

[8] & [22] Booth C, Parexel, Advancing rare diseases drug development Report, Section 2 Effective regulatory strategies,Part 7 https://www.parexel.com/insights/new-medicines-novel-insights/advancing-rare-disease-drug-development/study-design-and-execution-rare-diseases/how-sites-manage-pediatric-gene-therapy-trials

[9] Chen D, Parsa R, Chauhan K, et al. Review of brachytherapy clinical trials: a cross-sectional analysis of ClinicalTrials.gov. Radiat Oncol. 2024;19(1):22. Published 2024 Feb 13. doi:10.1186/s13014-024-02415-8

[10] Di Tonno D, Perlin C, Loiacono AC, et al. Trends of Phase I Clinical Trials in the Latest Ten Years across Five European Countries. Int J Environ Res Public Health. 2022;19(21):14023. Published 2022 Oct 28. doi:10.3390/ijerph192114023

[11] Song SY, Koo DH, Jung SY, Kang W, Kim EY. The significance of the trial outcome was associated with publication rate and time to publication. J Clin Epidemiol. 2017;84:78-84. doi:10.1016/j.jclinepi.2017.02.009

[12] Kavalci, E., Hartshorn, A. Improving clinical trial design using interpretable machine learning based prediction of early trial termination. Sci Rep 13, 121 (2023). https://doi.org/10.1038/s41598-023-27416-7

[13] Cooper A, Explaining Machine Learning Models:
A Non-Technical Guide to Interpreting SHAP Analyses,November 1st 2021
https://www.aidancooper.co.uk/a-non-technical-guide-to-interpreting-shap-analyses/,

[14] Tharmanathan, P., Calvert, M., Hampton, J. et al. The use of interim data and Data Monitoring Committee recommendations in randomized controlled trial reports: frequency, implications and potential sources of bias. BMC Med Res Methodol 8, 12 (2008). https://doi.org/10.1186/1471-2288-8-12

[15] Bierer, B.E., Li, R., Seltzer, J. et al. Responsibilities of Data Monitoring Committees: Consensus Recommendations. Ther Innov Regul Sci 50, 648–659 (2016). https://doi.org/10.1177/2168479016646812

[16] Chow SC, Pong A, Chow SS. Novel Design and Analysis for Rare Disease Drug Development. Mathematics. 2024; 12(5):631. https://doi.org/10.3390/math12050631

[17] Miller E, DrugWatch,https://www.drugwatch.com/fda/approval-process/, Last modified September 5th 2023

[18] Vetter TR, Mascha EJ. Defining the Primary Outcomes and Justifying Secondary Outcomes of a Study: Usually, the Fewer, the Better. Anesth Analg. 2017;125(2):678-681. doi:10.1213/ANE.0000000000002224

[19] Pizzamiglio C, Vernon HJ, Hanna MG, Pitceathly RDS. Designing clinical trials for rare diseases: unique challenges and opportunities. Nat Rev Methods Primers. 2022;2(1):s43586-022-00100-2. Published 2022 Mar 10.

[20] Ahmad FB, Anderson RN. The Leading Causes of Death in the US for 2020. JAMA. 2021;325(18):1829-1830. doi:10.1001/jama.2021.5469

[21] Logviss K, Krievins D, Purvina S (2018) Characteristics of clinical trials in rare vs. common diseases: A register-based Latvian study. PLoS ONE 13(4): e0194494. https://doi.org/10.1371/journal.pone.0194494

[23]
https://smoothdd.com/en/company/geography/smooth-drug-development-in-turkey/

SUMMARY :

This study examines the factors contributing to the early termination of Phase 2 clinical trials for rare diseases, comparing them with common disease trials. Using data from ClinicalTrials.gov, the focus is on organizational and protocol design elements rather than scientific factors. Logistic Regression and XGBoost models were applied to predict trial outcomes, revealing that complex regulatory requirements, particularly for FDA-regulated trials, and traditional clinical designs are major causes of early termination. The challenges of rare disease trials, such as small patient populations and lack of historical data, often result in high failure rates. However, these trials benefit from having multiple centers across the country and academic collaborators, which increase access to specialized resources. The study emphasizes the importance of contextual analysis when interpreting factors influencing trial outcomes.

FIELD :

Public Health, Data Sciences, Data Mining, Machine Learning, Natural Language Processing

---

KEYWORDS :

“Clinical Trials as Topic"[MeSH Terms] , ClinicalTrials.gov, NLP, Long Short Term Memory, Logistic Regression, Extreme Gradient Boosting, SHAP, Data mining, clinical trial outcome prediction, early termination, Databases

---

AUTHOR’S CONTACT INFORMATIONS :

SCHAEFER Yves
+33 6 64 97 28 87
yves.schaefer97@gmail.com