**ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ**

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ  
«Βιοστατιστική και Επιστήμη Δεδομένων Υγείας»**

A close up of a sign

AI-generated content may be incorrect.

**Διπλωματική Εργασία**

**Πρόβλεψη πρόωρου τερματισμού κλινικών δοκιμών με μοντέλα μηχανικής μάθησης vs κλασσικής παλινδρόμησης,  
βάση ανάλυσης των χαρακτηριστικών σχεδιασμού τους**

**Diploma Thesis**

**Predicting early termination of clinical trials with machine learning vs. classical regression models, based on analysis of their design characteristics**

**Συγγραφέας**: Ανδρέου Ευγενία

**Επιβλέπων**: Σπύρος Δεναξάς

ΑΘΗΝΑ, 2025

**Contents**

[1. Summary 4](#_Toc208126361)

[2. Introduction 6](#_Toc208126362)

[Clinical Trials Design 6](#_Toc208126363)

[Advantages/Disadvantages 6](#_Toc208126364)

[Approvals and stages 6](#_Toc208126365)

[Phases 7](#_Toc208126366)

[Reasons and Consequences of Early Termination 10](#_Toc208126367)

[ClinicalTrials.gov Registry 10](#_Toc208126368)

[Clinical Trials.gov Data issues 10](#_Toc208126369)

[Importance of Accurate Registration of Clinical Trials 12](#_Toc208126370)

[Background 12](#_Toc208126371)

[Termination Reasons 12](#_Toc208126372)

[Background Research 13](#_Toc208126373)

[3. Statistical Theory 18](#_Toc208126374)

[1. Evaluation Metrics - Feature Selection 18](#_Toc208126375)

[Binomial Distribution 28](#_Toc208126376)

[1. Models 28](#_Toc208126377)

[1. Logistic regression 28](#_Toc208126378)

[Generative Models 34](#_Toc208126379)

[K-Nearest Neighbours - KNN 36](#_Toc208126380)

[SVC 41](#_Toc208126381)

[Classification Trees 44](#_Toc208126382)

[2. Performance Comparison 49](#_Toc208126383)

[4. Methods 57](#_Toc208126384)

[Tools and Technologies 57](#_Toc208126385)

[Environment 57](#_Toc208126386)

[Libraries 58](#_Toc208126387)

[Data Sources 58](#_Toc208126388)

[Data Engineering - Feature Formats 59](#_Toc208126389)

[General 59](#_Toc208126390)

[Missing values 59](#_Toc208126391)

[Data Engineering 60](#_Toc208126392)

[Features Dropped – Not used 63](#_Toc208126393)

[Visualization Pre-Processing 64](#_Toc208126394)

[Dummies 64](#_Toc208126395)

[Interactions 64](#_Toc208126396)

[Models 64](#_Toc208126397)

[5. Results 65](#_Toc208126398)

[Descriptive Statistics – Visualization 65](#_Toc208126399)

[Selected Model 65](#_Toc208126400)

[Features Contribution 65](#_Toc208126401)

[6. Discussion 66](#_Toc208126402)

[Comparison with Literature - Strengths and Limitations 66](#_Toc208126403)

[Recommendations for Practice 66](#_Toc208126404)

[7. Appendices 67](#_Toc208126405)

[Python Code 67](#_Toc208126406)

[Data Dictionaries 67](#_Toc208126407)

[8. References 67](#_Toc208126408)

# Summary

**Πρόβλημα/Υπόβαθρο:**

Ο πρόωρος τερματισμός των κλινικών δοκιμών έχει ως αποτέλεσμα τη σπατάλη πόρων, ανθρώπινου δυναμικού αλλά και χρηματοδοτήσεων. Το αντίκτυπο είναι μεγαλύτερο σε προχωρημένες φάσεις, όπως η 2, 3 και 4, δεδομένου του μεγαλύτερου μεγέθους των παραπάνω πόρων.

Οι λόγοι μπορεί να οφείλονται μεταξύ άλλων, στα ίδια τα χαρακτηριστικά σχεδιασμού της έρευνας. Συνεπώς, ο τερματισμός μιας τόσο σύνθετης ερευνητικής διαδικασίας είναι πολύ παραγοντικός αλλά και μοναδικός για την εκάστοτε κλινική δοκιμή. Το σύμπλεγμα αιτιών καθίσταται συχνά δύσκολα ανιχνεύσιμο από τα κλασσικά μοντέλα πρόβλεψης, όταν αντίθετα τα μοντέλα μηχανικής μάθησης εκπαιδεύονται συγκεκριμένα στα εκάστοτε δεδομένα.

**Σκοπός:**

Το περιεχόμενο της παρούσας εργασίας είναι η ανάλυση των χαρακτηριστικών των κλινικών δοκίμων μέσω μοντέλων πρόβλεψης της πιθανότητας πρόωρου τερματισμού τους. Η αναγνώριση των χαρακτηριστικών αυτών μπορεί να συμβάλλει στη βελτίωση του σχεδιασμού των κλινικών δοκιμών και κατ’ επέκταση στην επαρκέστερη διαχείριση πόρων.

**Μέθοδοι:**

Για το σκοπό αυτό αναλύθηκαν οι κλινικές δοκιμές διαστήματος 2011-2024 από το <https://www.clinicaltrials.gov/> και <https://aact.ctti-clinicaltrials.org/> . Εκπαιδεύτηκαν μοντέλα μηχανικής μάθησης και κλασσικής παλινδρόμησης σε δεδομένα σχεδιασμού - αποτελέσματος και συγκρίθηκε η μεταξύ τους απόδοσή.

**Αποτελέσματα:**

**Συζήτηση/Συμπεράσματα:**

**Problem/Background:**

The premature termination of clinical trials results in the waste of resources, such as human resources and funding. The impact is greater in late phases, such as 2, 3 and 4, given the greater size of the above resources.

The reasons may be due, among other things, to the design characteristics of the research itself. Consequently, the termination of such a complex research process is very factorial but also unique to each clinical trial. The complex of causes often becomes difficult to detect by classical prediction models, when on the contrary, machine learning models are specifically trained on each data.

**Purpose:**

The content of this work is the analysis of the characteristics of clinical trials through prediction models of the probability of their premature termination. The identification of these characteristics can contribute to improving the design of clinical trials and, by extension, to more efficient resource management.

**Method:**

For this purpose, clinical trials from the period 2011-2024 were analysed from <https://www.clinicaltrials.gov/> and <https://aact.ctti-clinicaltrials.org/> . Machine learning and classical regression models were trained on design-outcome data, and their performance was compared.

**Results:**

**Discussion/Conclusions:**

# Introduction

## Clinical Trials Design

Clinical trials are designed to observe outcomes of human subjects under interventional experiment conditions controlled by the scientist. Randomized controlled trials, where participants are allocated at random (by chance alone) to receive one of several clinical interventions, are the ultimate evaluation of a healthcare intervention.[1]

### Advantages/Disadvantages

RCTs main advantage is the randomization which solves selection bias problems. Selection bias could be formed due to unknown confounders.

In contrast some biases that may occur are ‘misclassification-information bias’ where outcomes or exposure are incorrectly recorded, co-interventions (e.g., additional interventions received), contamination (e.g., subjects receiving the intervention outside the trial etc.). Another problem ‘volunteer bias’ (sub-type of ‘selection bias’) i.e., when the sample is not representative of the population from which it was drawn. This happens mostly due to restricted inclusion or exclusion criteria. More specifically, possible participants may be excluded due to eligibility criteria such as comorbidities, or other attributes like distance from study facility etc. Restricted criteria are defined to study design to achieve homogeneity of groups, control of confounders and thus ‘efficacy’ (measure of success of intervention under experimental conditions). On contrast, this leads to loss of generalizability to real world population, where comorbidities or other parameters may occur. The result is that RCT although offering very good ‘efficacy’ measures, cannot promise intervention’s ‘effectiveness’ (real world measure of success of intervention).

A well-structured RCT needs a well-defined and representative sample from population. Sample size needs to be sufficient for the study to achieve the appropriate power needed to detect a statistically significant difference. Studies’ outcomes must be reliable and meaningful measures. Above all else, ensuring safety is the top priority.[1]

### Approvals and stages

For a promising intervention in pre-clinical stage, the sponsor or/and investigator submits an investigational new drug (IND) application to FDA, with all necessary drug information and results till this stage. If approved phases I-III follow. Similarly, if those phases results are promising regarding not only efficacy but safety too, New Drug Application (NDA). Except FDA an external committee is also involved with this application. After final approval, phase IV follows, in which safety and effectiveness for the indicated population is monitored.[1]

### Phases

As mentioned above, safety is above all matters in an RCT, so the first of 4 phases purpose to test the safety and maximum tolerated dose (MTD) of a drug, the pharmacokinetics, pharmacodynamics and drug–drug interactions. [1] Participants are allocated at random to receive clinical interventions. RCT are considered the ultimate evaluation of a healthcare intervention.[2]

Below are explained all stages of an RCT in detail.

#### Pre-Clinical

Pre-Clinical stage includes animal experiments and evaluation of drug production and purity. Main concepts studied are: 1) Drug safety per dose, approximated to human quantities. 2) Pharmacodynamics (e.g., mechanisms, dose-response), 3) Pharmacokinetics (e.g., absorption, metabolism, drug interactions). Those data are all included in IND approval submitted to FDA for further drug investigation in humans.

#### Phases I-II (‘dose-escalation’ or ‘human pharmacology’)

It is the first human experiment stage. In this stage are determined the maximum tolerated dose (MTD) and doses before toxicity. Subjects are followed closely for all dose levels for any toxic effects.

Although, the objective of this stage is dose and pharmacology testing, there is a misinterpretation from volunteers that it could be therapeutic. Some improvements regarding the consent form, could help solving more clearly this ‘therapeutic misconception’. Participants constitute a small number of healthy or/and diseased volunteers. This stage is open labeled.

#### Phase II (‘therapeutic exploratory’)

They mostly test safety, pharmacokinetics, pharmacodynamics etc. An especial role these trials have is that they are also a supportive stage to proceed to phase III, as they answer crucial questions regarding doses, dose frequencies, administration routes, endpoints etc.

They may also study preliminary evidence of efficacy by comparing the drug with other controls (e.g., drug from published trials, standard therapies etc.), randomize and examine different dose arms (e.g., control arm).

Despite the possibility of studying all the above and that they are larger that phase I trials, they are also conducted with a small number of participants but only with the disease of interest. In combination with the safety concerns of early stages, this phase cannot serve any efficacy establishment but only support the continuity and planning of phase ΙΙΙ. That is the reason FDA usually requests for additional phase IV less common adverse events to be identified.

With the completion of the above early phases, a meeting may be conducted between sponsors, investigators and FDA to review preliminary data, IND or even manufacturing concerns, to decide if phase III will be initialized.

#### Phase III (‘therapeutic confirmatory’, ‘comparative efficacy’ or ‘pivotal trial’)

They are much larger that the above with a more diverse population. Their main topic is to identify efficacy and adverse events.

However, it has no more than 300-3000 participants, which gives a statistical power to establish an adverse event no less than 1 in 100 participants [3].This means that less common adverse events are not probable to be identified in this stage, and this is the reason why occasionally FDA requests for more than one phase III clinical trials or additional phase IV trial after drug approval.

* Comparative Efficacy trials (‘superiority’ or ‘placebo-controlled trials’):

It is the most common type of phase III RCTs where an intervention is compared to a standard therapy or placebo. It is notable that many ‘placebo effect’ instances (health improvement by placebo administration). Some say it's due to better care under research conditions, while others say that participants with acute symptoms would either way improve as time passes even without intervention. However, this indicates that a promising indication from a study does not always establish efficacy, as placebo effect may occur to both interventions. Another problem in phase III is a debate regarding comparative design in cases like surgical placebo procedures.

* Equivalency trial (‘positive-control study’):

The objective of this type of study is to study whether the intervention of interest is similar or different from another comparator. In this type, a placebo is almost never used. The margins are prespecified by the investigator and although they are based on statistical evidence, clinical experience etc. guidance for margin specification remains poor.

* Non-inferiority trial:

This type is a variant of equivalency studies. Its objective is to exclude that drug of interest is less effective from comparator i.e., null hypothesis is there is difference.

Non-inferiority trials often appear major issues, such as biased results towards null hypothesis (no differences) or else false positive, due to incorrect design and analysis. It is notable that they are most prone to false positive than any other study. It is notable that a false positive for an equivalency study would be a false negative in a comparative efficacy study.

*Phase III analysis/balance methods*

* Randomization:

The massive positive of Phase III studies is randomization in treatment allocation. Its purpose is to eliminate confounders and systematic differences (bias) between groups, thus any difference is due to treatment and not confounders. A common method is to just randomly assign subjects per group like flipping a coin.

However, this could lead to imbalances in treatment assignments or covariates’ distribution. A solution is ‘block randomization’ in which the number of subjects per arm is equal and balanced after a specified block size. For example, in a trial with 2 arms, a block size of 4 subjects would have 2 positions in arm A and 2 positions in arm B.

Yet, this could also result in ‘unblinding’ issues. For example, in a block size of 2, investigator will understand that if the first subject belongs to treatment A the next belongs to treatment B.

* Stratification:

This method is usually used in combination with randomization. Stratification ensures that some prognostic factors of clinical importance are balanced within arm groups.

Though, in small sample sizes when block randomization and stratification had both been applied, the original intended balance may be lost. To avoid such problems, alternative randomization methods can be used such as minimization or dynamic allocation that reduce imbalances among multiple study arms.

* Blinding-Masking

Blinding aims to reduce ‘information bias’ of subjects’ outcomes. Blinding could be:

* ‘Single Blinding’: Subject unaware of the given treatments.
* ‘Double Blinding’: Subject and investigator unaware of the given treatment.
* ‘Triple Blinding’: Data analyst, subject, and investigator unaware of the given treatment.

Still, not all types of studies can be blinded. For example, drug delivery methods or expected adverse and toxicities could indicate which drug is provided.

Trial design can be customized to the above problems like ‘crossover’ where each subject is also its own control or ‘factorial’ where more than one treatment are simultaneously evaluated.

*Analysis methods*

* ‘Intention-to-treat’ analysis (or ‘analyzed as randomized’ rule):

Subjects are evaluated based on the arm they belong, no matter what treatment they receive. It is the primary analysis method and most used, as it has the benefits of randomization’s selection bias reduction.

* ‘As-treated’ or ‘per-protocol’ analysis:

Subjects are evaluated based on the treatment received, no matter in arm they belong to. It is used mainly as complementary analysis as it diminishes the randomization benefits and is prone to selection bias. Although in comparison to the above it seems not so useful method, it has much better results in the case that, there are ‘adherence’ or ‘contamination’ issues, where the above method fails.

#### Phase IV (‘therapeutic use’ or ‘post-marketing’)

FDA may request phase IV trial after a drug approval. Their purpose is to identify less common adverse events and even evaluate cost, drug effectiveness in populations and doses similar or even different from the previous study.

After post-marketing phase IV many drugs require new black box

warnings post-marketing than pre-marketing phases (e.g, phase III) or many drugs even withdraw because of safety reasons.[1]

## Reasons and Consequences of Early Termination

### ClinicalTrials.gov Registry

Clinical Trials.gov is a national registry of clinical trials, created from FDA in 2000 [4] due to a law of Food and Drug Administration Modernization Act of 1997 which required from National Library of Medicine (NLM) to create and manage a public database for investigational drugs [5][6]. Although today it includes clinical trials and observational studies, happening now, will happen, or completed/terminated, with corresponding information e.g., study names, sponsors, investigators, dates, conditions, intervention types, documents etc.

It includes studies from over 200 countries. Registry of studies should be done from Investigators and Sponsors [6]. Registration data are displayed as table containing sponsors, design features, sample eligibility criteria, study endpoints, study results, and other [5]. Especially, for FDA approved drugs results must be uploaded within one year of the study’s completion, but this does not apply for all studies. [4] [6]

Finally, Clinical Trials ‘expanded access’ provide the possibility to patients with serious illness, unable to participate in trials, to register for an unapproved treatment. [6]

### Clinical Trials.gov Data issues

Many issues about inaccurate data registration at Clinical Trials.gov have been noticed. Most common problems are missing or misleading data (e.g., Intervention types, RP names or roles, enrolment size, dates, sites etc.)[7] [8] [9]and non-reporting results on time. However, most of this information are being mandatory to mention (e.g., sponsor’s name).

First field regarding RPs/PI personal information. For example, there have been noticed trials not mentioning one or both of responsible Parties (PR), Principal Investigator (PI) names, others not matching PRs’ number with corresponding roles etc. Additionally, many RP names are typed with various formats, within different fields of registry, or are even presented with personal information of fake non-existent person. Registries with more than one PI per trial, despite normally only one should bear responsibility. Even for multi-location trials each location’s PI registry should be distinguished from the main.[7] Another rarer issue is multiple primary outcome registries. [9] Further, problem is the wrongly labelled studies e.g., RCTs labelled as observational and via versa or even more general as ‘NA’ etc.[7] [9]

Another issue concerns the delayed or belated publication of studies and results.

Trials should register prospectively and upload main primary outcomes, before enrolment starts. Also, they should upload results within one year of study completion. However, not all obligations are met, as there are still trials not registering [7] or data uploaded later than they should, affecting investigators’ analysis by missing information etc., Another, serious data misleading is about enrolment sizes. Many studies did not mention pre-designed or actual sample sizes or mentioned extreme outlier values which was obviously mistaken. Moreover, in some the size referred to cluster samples, population proportion sizes etc. rather than actual participant’s size. Several, discrepancies between sample estimates during register and actual accruals (usually less sample achieved actually), something noticeable if compare enrolment registries to their text reports. [9]

Another issue that noticed in the precent work and complicates analysis, is the great amount of unstructured text data in ClinicalTrials.gov. Explanation summaries and descriptions are conducted through brief texts without formatting restrictions, although they include precious information e.g, termination reasons. Finaly, the most serious problem, is the discrepancies found between registries and their publication [8] [5] [9]. However, here ClinicalTrials.gov usually seemed to be more accurate. Specifically, publications studies were found not to mention part or whole number of adverse events (usually ClinicalTrials.gov were accurate) or actual efficacy, which is harmful for the health community. [5][8]

It is notable that on ClinicalTrials.gov registry documentations, is referred that some fields were made mandatory only after specific dates, while others are not mandatory even today as seen in icons below.[10]

|  |  |
| --- | --- |
| **\*** | Required |
| **\*§** | Required if Study Start Date is on or after January 18, 2017 |
| **[\*]** | Conditionally required |

For example, Study Documents **\*§** column from registry, has more than 100.000 null values out of 180.000 and more that were downloaded in precent analysis.

Due to the directions and pressure of the FDA and ICMJE the situation has improved considerably over the years [7], although it is important to be noted that ClinicalTrials.gov does not review the accuracy of registrations [6].

### Importance of Accurate Registration of Clinical Trials

The importance of accurate registry of clinical trials has many benefits, while the opposite may even have harmful consequences.

Because great funding is applied to RCT trials the appropriate registration of their design characteristics (especially, those leading to early termination), could be evaluated from registries’ statistical analysis. The detection of those characteristics could prevent the risk of wasted resources. Early termination is most costly on later phases where human and fund resources are greater. [11]

Clinical trials involve experiments implemented on human samples. They aim to health science evolution, and their results may have implications to future patient’s care [7]. The impact of clinical trials extends to individual patients by providing alternative therapies, and the society by increasing health care value.[1]

Especially, Randomized controlled trials (RCTs), which are fundamental for health science evolution and decision-making, should be reported transparently, regarding not only the treatment efficacy but also adverse events or safety risks that patients may occur e.g., death risk, hospitalization or abnormalities. As such it is harmful and thus unethical not to be reported accurately [8][1].

Furthermore, lots of studies are possibly to be published at journals with great impact on the clinical care community [5].

The reduction of literature bias will benefit prescribers, inform patients for new treatments and researchers doing systematic reviews [4], [7][4], without leading them to biased results. All in all, such registries are of public accountability [7][12].

Finally, misleading records also cancels the participants' attempt to contribute to health knowledge, especially of those relying on the study results [12].

Moreover, public registration platforms, such as CLinicalTtials.gov, use the data for variety of analysis regarding the sample sizes, sponsor type (e.g., industry, on- industry), number of Principal Investigator (PI), sponsor or funder per trial, sources distribution for global studies, termination reasons, results, type of intervention, locations that trials are distributed globally, nations health burden, unmatched or missing registries within platforms for same trials etc. [7].

Due to the above, International Committee of Medical Journal Editors (ICMJE), decided in 2005 that trials should be registered in corresponding public platforms, such as [ClinicalTrials.gov](https://clinicaltrials.gov/) , before patients recruitment, if they need to qualify for publication [7][4]. Similar, requirements were decided from Food and Drug Administration FDA in 2007[4] which required all clinical trials, with one or more sites in the United States, to be registered to ClinicalTrials.gov.[2]

## Background

### Termination Reasons

Clinical trials depend on human volunteers and cost significant investments of human, physical, and financial resources. Given the above, an early terminated trial raises financial, ethical, and scientific concerns such as cost of resources

that could have supported other trial, ethical issues regarding volunteers whose participation may not contribute to meaningful scientific knowledge, scientific issues which lead to termination decision and finally safety concerns regarding efficacy and most important safety of provided drugs. [13]

RCTs initiation, involve extensive planning, cost, researchers’ and sponsors’ deliberation. Funding agencies (e.g., NIH) are trying to screen only quality studies before funding, by conducting many peer reviews, which are costly for enterprise and time consuming. Even though, still many trials achieve moving on but unfortunately, they end up early terminated [14][15]. Unfortunately, later phases have more complicated explanation for termination. [2]

Termination reasons seem to be difficult to obtain, even though registered outcomes.[14]

The most common reason is fully unsuccessful or low enrollment, especially in specific study fields(e.g., rarity of disease hardens the collection of sufficient sample size). [16] [17][13][2][15][18]. Larger sample size trials seem to have lower termination rates [16].

Field of study is also an indicator, with most common the cancer/neoplasms trials, followed by surgical, while pediatric trials were mentioned rarer [14][15][19][11]. Also, trials not accepting healthy volunteers were noticed to have greater rate of termination in one study [11] .Moreover, disease type was affecting the low accruals rates e.g., rarer diseases had lower enrollment. [2][19]

Poor accruals affect also the statistic power of a study, thus it is important to fulfil the pre-designed participant’s number. [18]. Lastly, number of sites impacts the enrolment outcome, as it is related to limitations of participants leaving nearby or not from sites etc.[11]

Another main explanation is drug efficacy, toxicity or more general the risk-benefit ratio [14] [13] [11][20][15].

Moreover, funding, type of sponsor etc. or even due to location of facilities [14].

Greater number of collaborators has been noticed to result in higher probability of termination[2]. Administrative explanations were also detected regarding protocols, investigators, drug supply or withdrawal etc. [13]

It is notable that ‘terminated’ trials refer to those who no longer treat participants, while ‘withdrawn’ refer to trials for which enrollment was never initiated, thus did not enroll any participant.[13]

### Background Research

While all institutions and registries try to overcome early terminations and false registries to save human and funding resources, this topic has become popular to researchers who study reasons for RCTs early termination. Many technics have been used apart from classic models such as ML algorithms or even NLP processes to track differences indicating an early termination as soon as possible. [14]

Some of the research that dealt with the topic is:

[Follett L, Geletta S, Laugerman M (2019)](#_References) who studied only ‘completed’ and ‘terminated’ prior to 2015 trials from ClinicalTrials.gov. They used a combination of structured data offered from ClinicalTrials.gov public format with unstructured text data they from ‘description’ column. The study included transformation of ‘description’ column to a structured format by creating binary columns based on the presence of most frequently detected words in each category of studies. After, data engineering by removing common or unwanted characters-‘tokenization’, separating words, creating ‘tidy data frame’ of words-trials, inverse document frequencies, binarization etc. Words were assigned to one of two outcome categories, if some pre-specified frequency threshold was exceeded to this category trials while rarely detected the other category trials. Many words were evenly detected in both completed and terminated categories.

Finally, all data fitted with random forest models. Study resulted in specific words as model predictors, with the top five (based on their predictive importance), are ‘treat’, ‘chemotherapy’, ‘cancer’, ‘patients’, and ‘tumor’. Study was conducted in all phases and not separate per phase. [14]

[Geletta S, Follett L, Laugerman M (2019)](#_References) used NLP and ML algorithms to predict patterns of early trials’ termination on trials started prion 2015. Specifically, Latent Dirichlet Allocation (LDA) was used on unstructured texts data from descriptions of trials (e.g., Brief Summary). LDA uses Bayesian methods to model each document as a mixture of topics and each topic as a mixture of words, which are after corresponded to probabilities. Thus, a document can be represented by a vector of topic probabilities while each topic can be represented by a vector of word probabilities. Number of topics choice is mostly arbitrary.

Pre-processing of LDA included ‘tokenization’ as above study (data transformed to ‘tidy data frame’ so that there is one line per word -’token’ per clinical trial). Common English words and characters were also excluded. Document-term matrix is created by calculating the frequencies per trials.

Then, LDA was applied to the document-term matrix to create topic-word probabilities (β matrix) and document-topic probabilities (γ matrix).

Form all topics only 25 were used based on their probabilities.

After, those 25 topics were fitted with random forest classifier with 6 other structured data, to assess importance of the LDA topic probabilities. Another Random forest was fitted also for the 6 structured data only. Then a logistic regression was used for interpretation reasons. It was proven that both type of predictors gave better results, as structured alone lost in rare diseases.

For each topic, top 10 words in terms of the term-topic probabilities were evaluated and resulted in a top 5 of diseases fields that seemed to be most valuable indicators of termination (e.g., ‘surgery’, ‘surgical’ , ‘pain’ underlying surgical procedures. First topics were the surgical procedures, then dermatological, heart conditions, pregnancy and last HIV. [14]

[Kavalci E, Hartshorn A (2023)](#_References) studied only ‘completed’, ‘terminated’ or ‘withdrawn’ (no enrolment), clinical trials from ClinicalTrials.gov register for year 2012-2022 (before 2011 missing data were greater). This study combined two different public datasets, ClinicalTrials.gov and CHIA dataset (CHIA contains thousands of inclusion and exclusion criteria based on phase 4 studies). The first was used as source for RCT data (CSV format) and disease categorization features, while the second dataset to generate new set of eligibility criteria. ‘Text mining’ from unstructured text data with already structured data were combined.

Double phase characterizes studies were assigned to both phases e.g., phase 2/3.

Withdrawn trials were also characterized as ‘terminated’ in order to create a binary outcome.

Some examples of features used from ClinicalTrials.gov, is the number of primary and secondary outcomes, number of sites, termination by disease (e.g., neoplasm studies were most likely to fail) etc. Other regarding eligibilities such as gender, age, healthy volunteers, number of inclusion/exclusion criteria etc.

Especially, CHIA dataset was used to generate more complex eligibility criteria which were also used as search terms in text data. Those, data were transformed into binary a binary dataset of 12,864 features.

Other, important data and statistic handling: Missing data for numerical features were handled via Multiple Imputation by Chained Equations (MICE). Phases were studied at separate datasets each. Train test split 70:30 ratio. Imbalance dataset towards completed status (85:15 ratio) was handled with random under sampling on the train set only, in order for the test set to be representative of true status distribution. Feature election was used because eligibilities created enlarged the dataset and created noise. Firstly, elbow point of features vs model error plot was used to identify k number of features. Then, ANOVA F scores determined which features to keep.

For model evaluation fivefold cross validation was also used, because of small dataset occurring after under sampling. Dataset was split into 5 folds and each time a different fold was used as the test set. Thus, model is trained 5 times in total. Performance of the model was the average of 5.

Then features were analysed with logistic but also ensemble models like Random Forest and XGBoost to predict probability of termination. XGBoost model evaluated as the best performing.

After, two different types of XGBoost models were trained for all phases together. One with only study characteristics features, and the other with eligibility criteria search features and disease categorisation features additionally. McNemar test for paired nominal data was used to test the null hypothesis is that models perform the same.

For model interpretation (SHapley Additive exPlanations) SHAP plots for local overview were used. [11]

[Elkin M, Zhu X (2021)](#_References) studied interventional and observational studies from 2000 and after. Like above, features were created from the registries structured and unstructured data.

Features were formed in three categories:

1) ‘statistic features’ (e.g., number of collaborators, industry/non-industry sponsor, study design like masking type, eligibility criteria like gender, USA/non-USA location, number of countries etc.).

2) ‘Keyword features’ resulting from more unstructured data from ‘description’ field. Analysed with TF-IDF (term frequency-inverse document frequency). TF is the frequency of words in a document while IDF is the frequency of words in several documents. Based on IDF a term that occurs in many documents (e.g., ‘the’), is not a good discriminator and is given less weight.

3) ‘Embedding features’ to overcome the problem of sparse data for keywords appearing in small number, using Doc2Vec to generate vector representations of words.

Furthermore, five feature selection methods were used, including ANOVA, ReliefF, Mutual Information (MI), CIFE (Conditional Informative Feature Extraction) and ICAP (Interaction Capping). Dowdall method was implemented to combine and select features from all five methods. Dowdall favours feature with many first preferences, thus, If a feature is accidentally ranked to the bottom by a method, it will have little impact the aggregation value.

Finally, data were fitted to four classifiers, Neural Networks, Random Forest, XGBoost, and Logistic Regression. Three types of features were fitted and evaluated separately and combined. Highest metrics resulted from combination of features (all three together). Ensemble models like random forest and XGBoost outperformed logistic. Similar to first study above, cancer words were top detected to relate with early termination but also some rarer diseases (‘Mycosis’, ‘Fungoides’, and ‘Sezary’) which may reflect the difficulty of finding sufficient sample.[2]

[Williams R, Tse T, […]](#_References) studied ‘primary outcomes’ field from ClinicalTrials.gov for RCT studies to detect reasons of early termination and pattern that primary outcomes were reported on registry or public literature.

Study defined 3 types of explanation for trial termination:

1. Scientific reasoning was most common e.g., risk-benefit
2. Other than scientific study, e.g, low accrual as main in this category, administrative regarding to protocol or investigators, funding, product withdrawal, insufficient drug supply, etc.
3. Unknown-not provided [13]

[Gayvert K, Madhukar N, Elemento O (2016)](#_References) studied RCT provided drugs from ClinicalTrials.gov based on their features to predict toxicity related early termination of trial. In addition, were compared drugs that although been approved from FDA but failed in toxicity trials (FFT drugs).

Features regarding drug characteristics but also tissue expression (eg., genes etc.) where considered with additional data from DrugBank register. Examples of drug features are Hydrogen Bond Acceptor Count, Molecular Weight, Polar Surface Area etc. and examples of target tissue features are Target Liver Expression, Target Network Degree, Target LOF Mutation Frequency etc.

This study developed a new approach for predicting the odds of clinical trial outcomes using random forest (PrOCTOR). PrOCTOR integrates the two types of features mentioned above in order to distinguish FDA-approved drugs from FTT drugs.

Principal components analysis (PCA) was used due to high correlation of target expression features.

Study results showed that frequent adverse events are acceptable for drugs treating serious diseases such as cancer, which was mentioned to all above studies as well.

This because many drugs may have partly failed as FFT but alternative others may have failed to all similar requirements. Analysis also indicated that both categories of features were important indicators, while combination of other although uncorrelated features provided greater discriminative power as well. [20]

[Chapman S, Shelton B, […] Bhangu A (2014)](#_References) studied phase III/IV clinical trials from ClinicalTrials.gov database for year 2008-2009 for surgical procedures specifically, using logistic regression. Result were that surgical procedure trials were more likely to early terminate and appear low accrual rates as well. [19]

[Carlisle B, Kimmelman J, […] MacKinnon N (2015)](#_References) studied trials ‘completed’ or ‘terminated’ in 2011 from ClinicalTrials.gov registry, which had unsuccessful or fulfilled less than 85% of pre-designed enrolment. Other factors which resulted in poor enrolment, were greater number of eligibility criteria, non-industry funding, early phases, smaller number of sites, use of placebo etc. Statistical methods used, where chi squared test, backward feature selection and logistic model.[18]

Some of commonly used methods in most of the studies are: Choosing location this with most facilities, as main location, for multi-location trials. Evaluate models on test sets or parameter tunning with validation set cross validation and k-fold cv accordingly. Evaluation metrics always include balanced accuracy (imbalance towards ‘completed’), sensitivity, specificity and ROC curves. Imbalance of ClinicalTrials.gov datasets towards completed was usually treated with random under sampling.

# Statistical Theory

## Evaluation Metrics - Feature Selection

#### Hypothesis Testing

Hypothesis testing explores how likely it is that the observed difference would be seen by chance alone if the null hypothesis were true.

The goal is not to accept or reject the null hypothesis but to rather to evaluate how likely it is for an observed difference to happen if the null hypothesis is true [21].

: Null Hypothesis assumes no difference between groups (e.g., in their mean or proportions). Any difference lies due to chance. [22]. It expresses a specified value of an unknown parameter of population , such as , where a known constant [23].

: Alternative Hypotheses assumes signiﬁcant diﬀerences between two groups or else that there is significant deviance of from .

The alternative hypothesis must be based on a logical explanation based on the problem background [22]. Alternative Hypothesis can be two-sided (or two tailed) such as , or one-sided (one tailed) such as depending on the problem (e.g., one tailed: assuming the true time results of a drug appear equals to 10, then ) [23].

Then a random sample of the evaluated populations is chosen. The acceptance or rejection of is based on the value of a given test statistic, which needs to have a defined distribution under the . Afterwards, values of statistic test are separated into two discrete spaces. In one region is accepted while in the other rejected with second space being called [23].

##### p value

is the probability that an observed effect is simply due to chance, if is true[24] or else is the probability of seeing the observed difference, or greater, just by chance if the null hypothesis is true.

As a probability it can take any value between 0 and 1. Values close to 0 indicate that the observed difference is very unlikely to be due to chance and in contrast values near 1 indicate that observed difference is very likely to occur due to chance and random variation.

As an example, assume p value = 0.08 for a problem studied with logistic regression with given Odds ratio = 0.33. p value indicates that there is an 8% chance or else (8 out of 100 times) to observe such a difference of 66% risk reduction (or extremer) if no effect occurs, just by chance. [21]

Finally, is a measure of the strength of an association but does not provide any measure of the effect size and so it must not be used as the only guide for evaluating a result.[25]. More specifically the ‘significant’ or ‘non-significant’ result based on p value is obsolete. This is because, for e.g., p = 0.05 values, researchers will wrongly accept 5% of the times (5% of times the result occurs due to randomness), meaning to fall into a , but wrongly rejecting the hypothesis can induce a too. The wrong evaluation is more probable in small sample sizes, where significance may be missed due to small sample. Furthermore, p value should not be used for clinical decisions, as they offer no evidence for effect size, adverse events, etc.[21]

##### Test Statistics on Logistic

For logistic regression statistic test is statistic that divides the difference in observed and proportions by the null standard error of sample’s . z statistic measures the number of null standard errors that the sample proportion falls from the null hypothesized proportion .

: sample size

: sample proportion

: mean of

: Standard error of (3.2)

: population proportion under (3.3)

: Null standard error of , that holds under

Note that as sample increases the standard error of samples proportion decreases and approximates populations proportion and distribution of p approximates the normal. [26]

##### Confidence Intervals CL

Confidence intervals determine the range of all plausible values that a parameter can be assigned [26] or else a range of values within which it is likely that the

true population value lies.[21]

: The estimated standard error of p.

Note that this formula substitutes unknown population parameter with the sample proportion in the population standard error

: The standard normal percentile having right-tail probability equal to .

: a 95% conﬁdence interval for i.e., 95% of the times values of belong in this range of values. Also, .

We can be 95% conﬁdent that the population proportion is between the two values that arise from equation (3.4). [26]

##### Type I, II Error

As the rejection of is based upon one value of the statistic test, two mistakes arise to make:

* : Is to reject while it is actually true. This error can occur when statistic test’s value belongs to the but should not.

A white rectangular box with black text

AI-generated content may be incorrect.

* : Is to accept it when it is actually untrue. This error can occur when statistic test’s value belongs outside the n but should not.[23]

[27]

It cannot be accurately specified what type of error occurs. However, the possibility of making can be computed through the . In contrast, possibility of Type II error cannot be computed as it is based on unknown true value of parameter. The unknown probability of is symbolized as its domain is 𝜃 values specified by the alternative hypothesis.

The probability of rejecting while actually true is called ‘power’ and defined as .

The critical region is chosen based on where the probabilities of Type I and II error could be minimum. The concurrent minimization of those two is practically not achievable. So values of , and which lead to small values of , should be chosen wisely in order to minimize those two errors [23].

#### Akaike information criterion (AIC)

AIC tests a model by how close the ﬁtted values are to the true values, as summarized by a certain distance between the two. The optimal model is that which tends to have the closest predicted and true outcome probabilities for a logistic regression. This model also minimizes the following:

AIC depends on loglikelihood and number of predictors. Increase in predictors increases AIC score and thus has a negative impact.

In case of same number of predictors are compared or even same predictors with different formatting (e.g., log, sqrt etc.), AIC part of equation that is compared is as the predictors part is the same for all compared models.[28]

#### Cross-validation (CV)

The more complex and higher dimensional the data becomes, the more the model’s fit reflect the noise values in the data set. A sign of overfitting is that the model fits well to trained data but poorly to test data from the same initial sample .

k-fold has common use in which the data split into k different subsets (folds). The model is trained on every subset except one held-out subset in which the model is tested. CV needs to assure a total isolation of the test data. No test data must be part of selective (e.g., variables, parameters etc.) or training procedures. If not, then ‘information leaking’ is going to happen from test data’s information to fitting procedure, which results to inflated accuracy scores. [29]. Also, variable selection must not be done prior to cv, however, many studies miss this and have biased results.[30]

##### k-fold cv

Firstly, dataset is randomly divided into k groups (folds), of approximately equal size.

Then the statistical method is fitted to folds and tested to the remaining held out fold (validation set). This process is repeated by having different fold as validation each time. From this process k number of error scores (number of misclassifications for binary outcome), that are averaged to the final error. [31][32]

The error rate calculated from each iteration of cv, is the probability of misclassification for classification problems. The error rate is a measure of predictive capacity of the classifier – model.

Under random sampling, it can be proved that it is cv is almost unbiased.[32]

Per fold error rate:

k- fold cv averaged error rate:

[32][31]

: true error (probability of misclassification) of a classifier designed on a sample of size n.

Bias is not too great as long as is small.[32]

##### Leave One Out Cross Validation (LOOCV)

The extreme version of the k-fold CV approach in which k equals the total number of individual data entries n of dataset (e.g., training dataset). Instead of splitting the dataset of interest into k-folds, here only one data point is left out as test point each time for each cross-validation step.[33] So, LOOCV can be referred as a special case of k-fold that here equals to all data points .

LOOCV cv averaged error rate:

[32][31]

##### Validation Set approach

Validation set s used for evaluation of a method, usually to find the performance of a model on unseen data. It includes simply dividing the initial dataset into two subsets, the ‘training set’ and ‘test set-validation set or hold-out set’. The model fitted on the training data and then predict values of unseen test data. Resulting test error rate is like above.

: test set data points

##### Uses

CV is commonly used to overcome overfitting by model selection and models’ best parameters (parameter tuning) through grid search cross validation. Is also used for model assessment.[30] Another use is for estimation of test error rate, through training set, as training error rate is usually, much lower than test’s. The estimation of test error rate is done by the method described above. The error curve resulting from this process from the training data is similar to the one resulting from true test data error. [31]

##### Drawbacks of cv

One pitfall cv may fall, is that when splitting data into folds, different subsets will be produced each time, so different tuning parameters may be chosen. For this problem ‘repeated grid search cv’ may be used, which repeats a specified times, and produces same amount of cv errors, so the minimal cv error parameter is chosen. [30]

Another one is the assumption of CV is that observations are independent. The existence of this assumption is critical to the validity of k-fold cross-validation e.g., it is violated for data from family members, where its members tend to have more similarities. [29][34]

##### Train-test Split – Validation set CV

#### Chi-Square Test

**Chi square Distribution**

Chi-square test was proposed from Karl Pearson back in 1900 and is used for comparing unpaired group data, for categorical/nominal variables [35] [23]. Chi square test uses frequencies (i.e., the number of observations per category). As a non-parametric test, it doesn’t make population characteristics assumptions e.g., about distributions. However, non-parametric tests are generally less powerful to identify significant differences and reject null hypothesis [35][36] i.e., greater Type II error rate. [36]

Chi square test is used to test if observed frequencies differ significantly from the expected ones. Firstly, to recognize deviations of the sample observed frequencies from the theoretically expected ones (one-way tables) and secondly relationships between categorical variables (contingency tables). The closer to accept Ho i.e., closer the observed to the expected frequency, the smaller χ2 value will be and via versa. Also, as the df increase the χ2 value increases, as more squared differences are summarizing and the critical value for rejection of Ho also increases. [36] [37].

Chi square distribution is skewed, with skewness decreasing as the df increase, so as n 🡪 ∞ it approximates Ν(0,1) normal distribution [37].

N should be large enough for the chi square achieve approximation i.e., greater than sample 5.[38]. The area of a Chi Square distribution below 4 is same as of standard normal distribution below 2. [37].

**Degrees of freedom**

A standard normal deviate is a random sample from the standard normal

Distribution. The Chi Square is the distribution of the sum of squared standard normal deviates [37], thus all values are positive.[36] The degrees of freedom (df) equal the number of independent normal deviates e.g., only 1 independent normal deviate means a χ2(1) distribution with 1 df. [37]

**Chi square applications**

There are always two hypothesis H0 and H1. To reject Ho, it must be χ2 > c, where

c = critical value, defined from literature, based on df and a sig. level. Beyond c threshold null hypothesis is being rejected. [38][36]

Most common applications of chi square test are the two below.

* 1. Testing Goodness of Fit (One-Way Design)

Tests if observed data (from sample) differ signiﬁcantly from an expected population frequency, that can be theoretically determined even rationally or empirically.

One-way table example:

|  |  |  |
| --- | --- | --- |
| **Outcome Category** | **𝐸𝑖** | **𝑂𝑖** |
| 1 | 25 | 18 |
| 2 | 10 | 22 |
| 3 | 13 | 30 |
| 4 | 5 | 21 |
| 5 | 12 | 17 |
| 6 | 20 | 12 |

X = Random variable from which observations occur with mutual independent outcome.

Oi = (i.e., Ο1,Ο2, …,Ο𝑘) = observed frequencies (number of observations) that belong to set Ii (i.e., I1, …, Ik).

𝐸𝑖 = 𝑛𝑝i = expected frequency of set Ii (or ith category) based on H0.

pi = probability that a random observation’s outcome is from set Ii (or ith category).

𝑖 = 1, … , 𝑘, categories of outcomes.

df = k-1.

Degrees of freedom are calculated based on the number of categories k - 1. [36] as one category’s frequency can be calculated from the sum of all other frequencies and total number of sample/experiments occurring.

So df = number of categories – 1 (i.e., (e.g., 6 category outcomes df = 6 - 1 = 5). [39]

Oi = each observed frequency of corresponding category

n = total number of independent observations or total number of independent random experiments. [37]

* 1. Independency test (Two- Way Design)  
     (Contingency - frequency - cross tabulation tables)

Tests whether the distribution of frequencies in one variable is related to the distribution of frequencies to the second variable. Where if true (H1 accepted) the two of them are related.

H0: 2 Variables A,B are independent

H1: 2 Variables A,B are related

Issue: distribution of frequency of one variable is or is not associated with the distribution of frequency of the other.[36]

Degrees of freedom are calculated based on rows and columns of the table, where df =(R − 1)(C− 1).

R = number of rows

C = number of columns. [37]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***B1*** | ***B2*** | ***…*** | ***Bc*** | Sum A |
| ***A1*** | O11 | O12 | … | O1c | **O1j** |
| ***A2*** | O21 | O22 | … | O2c | **O2j** |
| ***…*** | … | … | … | … | **…** |
| ***Ar*** | Or1 | Or2 | … | Orc | **Orj** |
| Sum B | **Oi1** | **Oi2** | **…** | **Oic** | **N** |

The expected frequency for a cell in ith row and jth column:

Oi = Total for the ith row,

Oj = total for the jth column,

N = total number of observations.

Each observation contributes to only one cell (unpaired sample). thus, the sum of all cell frequencies is the same as the number of observations (sample) in the experiment

Ai = Variables with categories: 𝑖 = 1, …, 𝑟,

Bj = Variables with categories: 𝑗 = 1, …, 𝑐.

k = categories of combination i,j of variables A,B.

Ο𝑖,𝑗 = number of observations belonging in categories of Ai and Bj variables.

𝑝𝑖𝑗 = probability of an observation to belong in cell (𝑖,𝑗)

𝑝𝑖 = probability of an observation to belong in a category of A variable.

𝑝𝑗 = probability of an observation to belong in a category of B variable.

Under 𝐻0 ∶ 𝑝𝑖𝑗 = 𝑝𝑖 i.e., probability of belonging to one of ith categories of A variable, does not depend on probability of belonging to jth category of B variable.

**2x2 Tables Chi Square test**

|  |  |  |
| --- | --- | --- |
| A | B | **A+B** |
| C | D | **C+D** |
| **A+C** | **B+D** | **N** |

A, B, C, D = observed frequencies, Oi, in each cell

AD = OA × OD

BC = OB × OC

N = Total number of observations [36]

**Cramer V - Effect size of Chi Square**As the sample sizegets greater the more Chi square test can detect smaller effects, as the statistical power increases. However, the effect size of this measure is calculated through Cramér’s V(or Cramér’s phi) value.

φ ≤ 0,1 are considered to have small effect size, while

φ ≥ 0.3 are considered to have large effect size. [36]

#### Bias – Variance Tradeoff

##### Quality of Fit

To evaluate the quality of fit, there needs to be measured how well predictions match the true data values. The interest lies in evaluating the accuracy of predictions on unseen test data, rather than training data.

The quality of fit is evaluated through the accuracy of estimation , i.e., how well it performs when used to predict data. As for classification:

* *Training error rate:* The proportion of mistakes if estimator is used to the training set.

: Predicted label of ith observation by using

: True label of ith observation

: Indicator variable. if and if .

* *Test error rate:* It is associated with a set of test observations on which has been applied:

: Predicted label by using classifier to the test observation.

A good classifier is one for which the test error is smallest.[31]

Overfitting: The model appears to be a good predictor on the training set while underperforming on unseen data of test set. This is due to its low bias and high variance in which the model may adapt too strongly to the data which includes noise.

Underfitting: The model has high bias and lower variance. This is due to potential interrelationships between data features may have been ignored.[33]

##### Bias – Variance Tradeoff

Bias variance trade-off refers to finding the best balance between bias and variance to have a more generalizable model. Increased bias translates into underfitting, while increased variance translates into overfitting [33].

* Variance:

Refers to the amount by which would change if another data set was used (e.g., different training data). Ideally, should not change much within different sets. High variance model means that small changes of dataset (e.g., change any data point) result in great changes in estimate. More flexible models tend to have greater variance.

* Bias:

Refers to the *error* introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model. Low bias translates into that the model fits the data more closely. For example, a linear relationship may not fit well in real-life problems, so there will be bias in the estimation of .

* Flexibility:

More flexible methods, have more variance and less bias. Flexibility is associated with degrees of freedom of the model, where less df mean more restricted and robust model.[31]

#### Normalization (Scaling)

## Binomial Distribution

Distribution results from number of Bernoulli trials which could have two possible outcomes referred as ‘success’ and ‘failure’, with ‘success’ usually labelled as the preferred outcome.

: Success probability for a given trial,

: The number of successes out of the n trials is denoted with. So, has the binomial distribution with index n, parameter π.

trials are assumed to be:

* Identical (same probability of success is same for all trials) and
* Independent random variables (the outcome of one trial is independent from other trials).

Binomial distribution’s mean and standard deviation:

The probability of outcome y for Y:

[26]

## Models

### Logistic regression

#### Maximum Likelihood Method

Maximum Likelihood is the method used to fit the logistic regression model. It searches for coefficients that maximize the likelihood function:

From the equation it can be concluded that maximizing likelihood means finding those coefficients, for which each data point is assigned with a predicted probability that gives as close as possible the actual observed value for each data point i.e., assign value close to 0 for Y = 0 and close to 1 for Y = 1, when predicted

replaces coefficients in the above equation.[31]

#### The Need for Classification models

A Linear form of the relationship of X with a binary outcome y to be interpreted: [31]

: Probability of ‘success’.

Outcome Y is encoded as 0,1 for convenience.

A comparison of a number of graphs

AI-generated content may be incorrect.This relation cannot be explained through linear regression, because negative probabilities or greater than 1 occur, as seen below.

#### Logistic Fit

Logistic regression uses this transformation to interpret the binary outcome.

From the equation is notable that p(X) cannot take values equal or less than zero and equal or greater than 1, for whatever X is used.

#### Odds

By transforming above equation, we receive the Odds quantity: , which takes positive values [0, +∞). A unit increase in X multiplies odds by:

* Odds: [31]
* Logarithm of Odds (Log Odds/Logit): [26]

Log Odds : is linear to X. Thus, a unit increase in X changes the log odds by β1.

It must be noted clearly that relation between p(X) and X is not straight line, as is in simple linear regression. (i.e., it is not ). This is understandable when plot X ~ p(X) (fig above), that logistic gives an shaped curve. Thus, a unit change in X does not correspond to change in p(X) by β1 (but the log Odds by β1 as notes above). The rate of change of p(X) per unit of X depends on the value of X.[31]

#### Interpreting logistic Output

Logit (log Odds) increases/decreases by β for every 1 unit change in X.

: Determines the rate of increase or decrease of change in logit with the corresponding change of X.

: Increase of logit. Ascending S curve.

: Decrease of logit. Descending S curve.

: Y is then independent of X. Horizontal straight line. Then, .

: The curve’s slope or else actual rate of change. For example, for

The slope approaches 0 as the probability approaches 1.0 or 0.

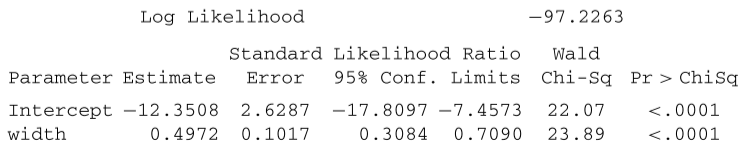
: median effective level, where 0 (outcome has a 50% chance).

has slope approached zero(straight line).

.Slope [26]

**Interpret Probability:**

For example, for x = 33.5 probability of success is 0.987 or 98.7% as seen below.



[26]

**Odds Ratio Interpretation**

The odds multiply by for every 1-unit increase in x. Thus, the odds at level x + 1 equal the odds at x multiplied by .

Thus, for 1 unit increase of x:

For each unit increase in x there is (1+0.64) = 64% increase in Odds.

##### Confidence Intervals

Wald conﬁdence interval for the parameter β coefficient and (confidence intervals for odds) accordingly:

##### Signiﬁcance Testing

Wald test statistic, following approximately standard normal distribution:

follows chi-squared distribution with df = 1.

The likelihood ratio test is more reliable. It compares the maximum of log-likelihood when β = 0 to the maximum of log-likelihood for β≠0. It follows chi-squared distribution with df=1.

#### Regularization

Regularization is when the coefficient estimates are regularized i.e., constrained or else shrink towards zero.

This way, models bias may increase slightly but the variance will be reduced. [40]

A case where regularization helps is when the number of predictors is large compared to the sample size. Another one is when covariates are correlated with each other. Overfitting and collinearity yield very unstable maximum likelihood estimates (MLEs) with some of the being even infinite. [40]

##### Ridge

Ridge regression has effects on coefficient estimators by shrinking them towards zero, but never equal to zero, which can be a problem for high dimensional data.

: Shrinkage Penalty.[31]

: Fraction of Ridge estimators towards initial ones.

or else : . [31][41]

: Probability , given the value :

Log likelihood :

: Tunning parameter. Controls the size of shrinkage

: no coefficients shrink, and model is fitted as original

λ = ∞ : coefficients will approach zero

As increases, shrinkage increases, and decrease. Its value is usually selected through cross validation

Τhe estimate of Ridge coefficients is expected to be on average closer to the real value of than the unrestricted , i.e. for regression.[40]

: depends on λ value and scaling of predictors e.g., a greater scale of predictors would estimate much different on the outcome as the coefficients are also shrunk. [31]

Ridge regression shrinks correlated predictors’ coefficients towards each other, allowing them to borrow strength from each other. In the extreme case of k identical predictors, they all get identical coefficients with 1/kth the size that any single one would get if fit alone. It is ideal if there are many predictors, and all have non-zero coefficients.[42]

**Lasso**

Lasso regularization works like Ridge, but here coefficients can get exact zero. Another difference of lasso is that is used. [31]

= : [43][41]

: Shrinkage Penalty

It is useful for high dimensional data where variable selection is needed and also it results in and easier interpretated model. [31]

In contrast to Ridge, lasso is indifferent to very correlated predictors and as it will pick one and ignore the rest. In the extreme case of k identical predictors however it fails. [42]

##### Elastic Net

Elastic Net combined both Ridge and Lasso.

It is useful in high dimensional data and when correlated predictors occur.

: Shrinkage Penalty

* : Ridge regularization
* : Lasso regularization
* : In between Ridge and lasso regularization [45][41]

Or format:

### Generative Models

Generative models use another method to fit the classification, based on Bayes’ theorem/Bayes’ classifier if K classes assumed:

The idea is that it categorizes each observation x to the class where is largest.

This method is useful for cases where: 1. there is a serious separation between classes, where logistics coefficients become very unstable, and 2. if X follows an approximate normal distribution in each class with small sample size.

= prior probability that an observation x belongs to kth class.

= density function of X for an observation x which belongs to kth class. is large if probability of X ≈ x (i.e., belong to kth class) and small otherwise.

= posterior probability of an observation X = x belongs to the kth class, given the predictor value for this observation.

The idea is that instead of computing , to just plugin and to the equation (1). Specifically, estimations:

= By simply taking the fraction of ktvh class from a training set from a random sample of population.

= it is more difficult to estimate, and each of the LDA and QDA models make corresponding assumption. [31]

#### LDA – Linear Discriminant Analysis

**Unique predictor, p = 1**

Assumptions of LDA to satisfy Bayes’ theorem:

* 1. = follows normal gaussian distribution
  2. = variance is equal among all classes.

Other assumptions:

* 1. = is unknown and simply assumed as the average of the training sample of kth class.
  2. = is unknown and simply assumed as a weighted average of the sample variances for each of the K
  3. classes.
  4. = for no other information, is simply assumed as the fraction of the number of observations in the kth category to the entire sample size.

Thus,

By replacing :

or the log, named Discriminant function which is linear to x :

which replacing for mean, variance assumptions is∶

is linear to x (linear discriminant analysis) and decision depends only on x. Decision boundaries are ‘lines’ themselves as well. In case i.e., equal prior probability to belong to each class, then classifies observation to class for which :

**Generalize for Predictors, p > 1**

Assumptions**:**

* 1. : predictors, follow multivariate gaussian (normal) distribution, with some correlation between them.
  2. : mean of X
  3. : is unknown p × p covariance matrix of X, common to all classes. Common covariance matrix makes LDA model linear to x.
  5. = are unknown means for each class.
  6. = are unknown.

Simple assumption to approximate those values happen also for p>1.

Decision boundaries number is based on pair of predictors. E.g., 3 predictors mean 3 boundaries or 3 separate spaces for each class k.

Error rates are a function of posterior probability threshold, where changes can differ sensitivity for specificity and via versa.

LDA in general is less flexible with its linear boundaries, with less variance. Moreover, assumption for common covariance is not actually happening to data, LDA will perform high bias. The fact that LDA depends on all data points, considering within class means and covariance matrices, makes it very vulnerable to data changes and thus less robust to outliers or noise. However, it is recommended when there is small training sample, as variance is reduced. [31]

#### QDA - Quadratic discriminant analysis

Generally, the concept of QDA is like LDA, using the Bayes’ theorem with plugged in values of , while differences lie to covariances and boundaries shape.

Assumptions:

* 1. : predictors, follow multivariate gaussian (normal) distribution, with some correlation between them.
  2. : mean of X
  3. : are unknown covariance matrices, different for each class k. Unequal covariance matrix makes QDA model quadratic to x.
  5. = are unknown means for each class.
  6. = are unknown.

QDA model has the drawback that many covariance matrices need to be computed. However, it is more flexible and better for great sample sizes, especially when variance of classifier is not of big concern. [31]

### K-Nearest Neighbours - KNN

KNN is a non-parametric supervised machine learning method which was developed by Evelyn Fix and Joseph Hodges in 1951 and later optimized by Thomas Cover.

KNN method has various applications, especially in registering closest points, recognizing patterns, fraud detection etc.[47]

**Method**

KNN model is instance based i.e., it does not use a function defined from training set, but it rather makes predictions by analysing the data in real-time when new data points appear, without necessarily having to go through a training phase.

KNN works based in neighbour data points. Similar data points usually cluster nearby at space. Thus, a prediction for a query data point happens based on the distance and thus similarity to the exciting training data points. [47].

**Parameters**

1. : Number of neighbour data points. Crucial tuning parameter. For large , computational demands increase. Choice is associated with dataset’s characteristics and impacts the accuracy.

* Small values (few neighbors) increase model flexibility and noise susceptibility i.e., overfitting.
* Large values (lots of neighbors) can underfit, while also being computationally more demanding. Boundaries are also smoother.

1. Distances : Usual metrics :

* Euclidean : Most used.
* Manhattan : Better for high dimension data, for different scaling or sensitivity to specific predictors.
* Minkowski : Similar uses as Manhattan.

1. Neighbors:   
   All data points are ordered based on their distance and the number of closest data points of neighbors are becoming the neighbors.
2. Response:   
   For classification response is simply the most common occurring class of neighborhood. For regression could be similarly the mean, median etc. Note that test points will always take the same value as response, based e.g., on mean of neighborhood. [47]

**Exact KNN**

Exact KNN methods tend to be more accurate in finding the actual neighbor data points (i.e., actual close ones) and perform better where precision is important in the outcome e.g., health data. Exact KNN has two main types of methos : KNN search and KNN join for selecting the nearest neighbor’s data points. [47]

1. KNN Search (Brute Force - BF or Exhaustive Search):   
   Finds nearest points of one point from the rest points. The process is repeated for all data points. However, due to calculating all distances for each data point, it is extremely computationally expensive in big datasets.   
   This method is better for problems which need speed and accuracy. In more detail:   
   : dataset  
   n: data points   
   : number of dimensions,   
   : a query (test) point,  
   : number of nearest neighbors  
   : a function that calculates the distance between point and query point .   
   : a set of k neighbor points.   
   KNN search finds a set of k neighbor points such that for each point p ∈ R , there is no other point where .
2. KNN Join : Finds nearest points for each data point, but the process does not repeat as above per data point but happens for all data points as the same time. Better when uncovering hidden relationships or patterns needed and also computationally more efficient. This method is applied to k-means clustering, Outlier detection, KNN classification, Missing value computation etc. More detailed:  
   : dataset  
   : dataset  
   n: data points in   
   : number of dimensions  
   : number of nearest neighbors  
    : a function that calculates the distance between point

and query point   
: a set of k neighbor points.   
KNN Join find pairs of points such that for each , there are exactly k points with minimal distance. Or else, for each point in , we find its closest points in .[47]

**Drawbacks**

Computational requirements: A data point’s distance must be calculated from all the other data points. This escalates when it comes to bigger sample datasets [47]. Roughly, the time needed to find nearest neighbours is about time, which can be a serious time passing by with large n samples. [48]

1. Noise, Outliers:   
   The fact that method is based on neighbors makes it susceptible to noise and outliers, in case a non-significant predictor or an extreme value in considered as neighbor.
2. Accuracy:   
   By finding the efficient number and neighbor points precisely. Especially, in high dimensions, where distances lose their meaning, it is more difficult to find the actual near data points.
3. ‘Curse of Dimensionality’ :   
   Model’s performance can decrease in high dimensions because distances become less meaningful.[47]

**Alternative methods and KNN models**

Other alternative methods help in overcoming the drawbacks that exact ΚΝΝ cannot by itself.

* Computational efficacy:   
  Alternative KNN models like ‘Approximate Nearest Neighbors’ (ANN) for more complex data, are faster but with a cost of accurately choosing the neighbors, i.e., neighbors chosen might not be the actual nearest ones.
* Noise, Outlier:   
  Advanced distance metrics, weights (i.e., far neighbor points are more influential than close ones etc.). For example, variations of model K such as Adaptive KNN, Weight adjusted KNN and Fuzzy KNN to improve precision.
* Other:   
  e.g., k‑distance join, reverse KNN join, dynamic KNN join etc.[47]

Below methods are useful for computing the actual neighbor points accurately but also as fast as possible. Especially in high dimension B or R-tree indexing methods, perform purely.

1. Parallelization: divides the concurrent process of distance computing into different units. For example, this method is regarding device’s CPU way of procedure with command (e.g., = 1 work only on 1 core which is computationally slow, = 4 work on 4 cores, = -1 work on all available cores).
2. Dimensionality reduction (DR): Simply, decreases dimensionality e.g., through ‘principal components analysis’ - PCA.
3. Partitioning: Data is segmented through tree structures (e.g., R-tree, R\*-tree, Ball tree, KD-tree etc.) that perform pruning on part of data.

* Space based
* Data based [47]

**Tree Structures of Exact KNN**

Tree structures are a partitioning method, with the most common to be . Tree’s root is the whole data space which is the initial cell. Then data split at their median along the direction of a chosen coordinate. The process is repeated until split leaf cells are left with a threshold of data points. Final tree has a depth of about . Time that is consumed for this process is about

When a query point appears two methods of prediction may be chosen:

* Defeatist search: the query point is assigned down to the appropriate leaf and takes the value of the closest to it leaf point. Time consumption is more efficient at about or for constant . The great negative of this is that the ‘nearest point’ may be giving very wrong prediction i.e., may not be actual as similar as needed and also another leaf’s point may be the actual closest one to the query point. Unfortunately, failure rates may be unacceptable.
* Comprehensive search: Checks which other leaves may be needed to be taken into consideration for NN doners and returns the true closest neighbor point. Computationally it may take max time.

A big drawback of tree KNN is that their performance decreases a lot in high dimensions. For this problem, extra methods for dimension reduction are applied such as PCA as noted above. [48]

**A table of informational text

AI-generated content may be incorrect.A white sheet with black text

AI-generated content may be incorrect.Summary of different KNN models and methods** [49]

### SVC

#### Maximal Margin Classifier

**Separating Hyperplane**

A Hyperplane is a dimension flat subspace, which separates classes, in a p dimensional space or else a linear decision boundary. For example, for , it is a simple line separating two classes. For more than two classes multiple hyperplanes are used.

Generalizing for p the equation of a hyperplane is simply:

(7.1)

Any that satisfies this equation, is a point of the hyperplane and defines the hyperplane.

For training set of n training observations and

X lies on the hyperplane and defines it.

and X lies to one side of hyperplane

and X lies to the other side of hyperplane. Or more combined, ensuring all points are on the correct side:

Thus, for a test data point, the categorization is achieved based on the sign (positive or negative) the equation is assigned for a data point.

Thus for:

The farthest from zero is, the surer for the classification, as it belongs far from the boundaries space.

**Maximal Margin Classifier – Hyperplane Margins**

If data are perfectly separable, then there may be infinite hyperplanes, slightly rotated, that can separate them too. In order to stick with one hyperplane, the ‘maximal margin hyperplane’ is used among all the other hyperplane choices. More specific, two margins are defined in both sides of the hyperplane. These margins are defined by calculating the maximum distance M from each side of the hyperplane, from the closest to the hyperplane data points. So, we are looking for coefficients to satisfy the maximized value of M. The distance between margin and hyperplane is constant all along the hyperplane. Those closest to hyperplane data points that define where margins and thus hyperplane are oriented, are practically lying on the margin boundaries and are called ‘support vectors’. All in all, maximal margin hyperplane depends only on the close neighbour data points, that can ‘touch’ the margin boundaries, rather than the whole dataset.

Thus, any addition, removal or other change of support vectors or data points capable of being support vectors, will make the hyperplane orientation to change, when in contrast changes in the rest of the dataset, will not make any difference. This, however, implies overfitting issues may occur.

Finally, another problem arises when it comes to non-perfectly separable datasets, where the violation of margins and/or hyperplane is inevitable from some data points. This problem is solved by ‘soft margin classifiers’ or else ‘support vector classifiers’.

**Support Vector – Soft margin Classifier**

Support Vector classifier solves the above problems, by allowing misclassification and boundary violation to some data points close to margins and hyperplane. Those violators can be in the wrong side not only of margins but also hyperplane itself. This makes the model more robust and overcomes the problem of overfitting.

Similarly, with above M is the margin width and needs to be as large as it can be. Support vectors are now the data points, lying on margins or being inside them i.e. all data points which violated.

are ‘slap’ variables, which allow observations to violate boundaries.

For the training set:

Classification of a test data point:

* : , test point on correct side.
* : , test point is inside the margin.
* : , point on the wrong side of hyperplane.

**Tuning parameter C:**

C parameter defines the number of violations allowed from the classifier and thus the bias-variance trade-off.

* , each , no violations allowed.
* , maximum number hyperplane violations allowed are equal to C.

The higher the C value, the wider the margins and the more the support vectors.

**Support vector Machines – Non-Linear Boundaries**

Support vector machines are a support vector classifier, which use kernels to enlarge the feature space, and produce non-linear boundaries.

The mathematical solution to above functions includes the inner products of training observations, e.g., the inner product of two observations is

.

For observations and parameters, which are the trained coefficients per training observation and are only for support vector data points, the above function is formed as follows:

S is the collection of indices of support vectors.

and are estimated using number of n data point pairs.

The classification of test point x needs to compute all the inner products of that point with all the training points.

General form of function, where K is a kernel function, which quantifies the similarity of two observations. Te following most common kernels:

* Linear Kernel:  
  , where classifier is linear to x. It uses Pearson (standard) correlation, to quantify 2 observations similarity.
* Polynomial Kernel:  
  , . The greater the d the more flexible the boundaries.
* Radial Kernel:  
  . Works very local with close to boundaries points.

The advantage of using a kernel rather than simply enlarging the feature space as we would do in a linear regression model (with polynomials, interactions etc.), is computational efficacy, as using kernels needs only to compute for all [31]

### Classification Trees

#### Recursive Binary Splitting

For predictors, Trees are separating this predictor’s space into distinct regions. Regions could have any shape, but usually, ‘boxes’ are chosen for interpretation facilitate. A split can result in two or a single region or leaf.

The process initiates from the top split, by considering all as first the first candidate for splitting, to finally choose the one that will give less Classification error rate (or RSS for regression) in both resulting regions. The first feature is the most important predictor.

The process continues similarly, from top to bottom. However, it is a greedy process as splits continue to be decided based on the rest of predictors and the already formed regions i.e., next regions are divisions of the previous ones. The process ends when a cut point is reached, e.g., a specific number of observations in leaves.

* 1. Internal nodes:   
     The node where a predictors space splits.

e.g., Left: , Right: , the response measure for regression or classification accordingly.

* 1. Leaves (Terminal nodes) :   
     The last outcomes of the last internal nodes, at the bottom of the tree.

Practically, each split leads to a Region , where the outcome values are always the same for all data points belonging in . For example, for classification this value is the most ‘common occurring class’ in this region, while for regression it usually is the mean value of the response for points in the region. In classification the interest lies not only to the most common class but also to the proportion of this class, among others. Finally, this process runs faster for less predictors (less dimensions p).

**Separating criteria**

Classification:

* 1. Classification error rate:

Is the fraction of observation not belonging to kth class in a particular region and needs to be minimum. However, it is not used in practice.

: is the proportion of training observations from the kth class in the mth region.

* 1. Gini Index:   
     A measure of variance across classes and node purity as well. For close to zero or one. i.e., great purity in node. For very pure nodes (i.e., most of observations belong to kth class).   
     Gini Index has a small value for pure nodes..
  2. Entropy:  
     For near zero or one (pure node) entropy takes value near zero, i.e., entropy takes small value for pure nodes.

Regression:

= mean response or common label for all points in

= mean response or common label for all points in

Note that regression criteria is oriented to both nodes, while in classification per node. [31]

#### Pruned Trees

With simple trees there is a risk of overfitting, as more the splits and leaves, i.e., think that the bias is minimized (overfitting) on training data but does not mean the same will happen for test observations.

For this reason, pruned trees create less nodes and leaves in order to decrease variance with the trade of bias increase.

**Cost complexity pruning**

Initially, a very large unpruned tree is made. Then a series of other pruned subtrees are defined by pruning the initial . In order not to compute all the possible combination of , a parameter is defined, based on which a series of subtrees are created i.e., for each value of corresponds a subtree .

: number of terminal nodes (leaves)

: is a tunning parameter, that considers the bias-variance trade-off of complexity vs fitting to training set. It is usually selected through cross validation.

* : subtree equals the initial unpruned .
* : As increases less leaves remain i.e., more pruning happens.

Finally, after cross validation in the training set, the chosen subtree is the one for which the below obtains minimum value.

Reason why many subtrees are calculated, rather than one pruned based on a defined threshold (e.g., number of leaves to stop) is because tree structure is a greedy process as mentioned above. So, the pruning may have ended to a not so important split, while a more important may have followed, if continued.

**Advantages and Drawbacks of Trees**

Firstly, trees are simple and easy to interpret, especially because of their Graphical representation not only regarding their ‘tree’ structure but also the region spaces in Cartesian plane. They will also outperform classic liner models for non-linear or complex relationships.

However, in general they are less accurate than best supervised methods.

Most important, they are very prone to data changes, i.e., they lack robustness or else suffer from high variance. For example, 2 different datasets for the same classification problem and predictors or even 2 different random splits from the same dataset will give different trees.

For the accuracy optimization, the below alternative tree methods have been defined, but with the trade of decreased interpretability. [31]

#### Ensemble Method Trees

Ensemble method combines a number of simpler models to final predictor/model in order to achieve better performance. Such as, bagging trees, random forests, boosting and xgboost trees, that are discussed below, aggregate several decision trees into a single predictor. [50]

##### Bagging Trees (Bootstrap)

Because of high variance and lower accuracy scores in simple trees, there have been defined other methods to solve those problems.

Random forests use the bagging method to define more than one trees, while also using the property of reducing variance by averaging this set of trees.

More specifically, bootstrapping is taking repeatedly random samples from the training set, separated random training samples are generated from the initial one. Thus, it is and for each of the training sets. [31]

A test point prediction in a classification problem, is decided by considering the class that each one of bagging trees gave and choose the most commonly occurring. [31]

Finally, the number of trees in not so crucial. As far as an efficient number of bagging trees is chosen, a greater one will not increase the accuracy or even overfit further. Thus, one positive aspect of bagging is the lower risk of overfitting.

Moreover, bagging trees offer the capability of summarizing predictor’s importances. For example, the total amount of Gini index decreased averaged over all B trees, by splits over a given predictor, can indicate and important predictor if this value is large enough. Similarly, happens with RSS for regression problems.

In contrast, due to more than one trees, they lose their interpretability efficacy compared to a single tree. [31]

##### Random Forests

Random Forests are an optimized form of bagging trees regarding the ‘tree decorrelation’. The concept of bootstrap defined trees is the same. The difference lies on the splitting methods. Here, only a random sample of the predictors (usually ) are candidates for each split and from those, only one is chosen. This way, bagging trees defined are not the same with one another and if for example, a very strong predictor occurs among others, he will not be present on average to (p − m)/p of the splits. This predictor will not prevail, and trees will not give similar results as it would happen in bagging trees. Averaging, uncorrelated trees lead to variance reduction, while with correlated ones this does not happen. Additionally, using a part of predictors for fitting, is extremely useful when it comes to high dimensional datasets, or correlated predictors [31]. Random Forests are proven an extremely good performing model for classification, which requires less pruning than other algorithms and is also straightforward as it is simply a form of ensemble learning involving multiple classification trees.[14]

##### Gradient Boosted Decision Tree (GBDT) [51]

GBM is another ensemble method, i.e., it combines a number several decision trees into a single predictor, i.e., many trees are combined as well. [31] [50]. Boosting differs from the above ensemble methods, in that trees grow sequentially, i.e., each tree grows based on the previous and also that trees are fitted from the whole initial dataset, rather than bagging samples of it. [31]

It is notable that GMB works based not on parameters, but on functions’ space and also that not but residuals are used.[31][50]

Each new decision tree is trained based on compensating the prediction errors

of the previous iteration’s tree [50].

More specifically, GMB works by approximating the function , of the relation of , by computing of in an additive manner. Firstly, given the initial model, a decision tree is fitted to the residuals, which equal to true only for this iteration. After the first iteration, tree, each new tree is trained to minimize the following: [50]

: Residuals of previous iteration .

: loss function. Measure of prediction’s error.

: Prediction of current mth tree [50].

: matrix of features

: output vector

: Number of trees.

: , .

: Number of splits for terminal nodes.

: Initializing residuals for all in the training set.

: Learning rate. Constant regularization parameter, that limits the influence of trees ensembled to avoid overfitting. It takes small values (e.g., 0.001, 0.01) Boosting trees ‘learn slowly’, so , slows the process, by determining how much trees’ change during the iterations.

Small values usually need greater . Generally, the slower the learning rate the better the performance .

: Initial function (e.g., that is going to be compensated.

: Tree fitted in the iteration.

: True function of relationship between

: Approximately computed function of

Then, the tree defined is added to the initial model to update the residuals by adding a shrunken version of the new tree, and this is repeated similarly.

Update previous tree [31][50][52]:

Update residuals [31]:

Final, model [50]:

Trees do not need to have many splits. Even one or two splits can be enough, as fitting small trees results in optimizing to areas that does not perform well. Although, a great number of trees can overfit. parameter is selected, as usual, through cross validation. [31]

#### Extreme Gradient Boosting - XGBoost

XGBoost is a regularized alternative of GMB. The difference lies in the computation of residuals per iteration , that is no more a loss function. Formula used for residuals is now:

: Regularization hyperparameters [50]

: Number of leaves in the tree

: L2 norm of tree’s leaf weights [50][52]

XGBoost learns simpler trees with smoother weights, which leads to better generalization. Additionally, XGBoost employs Newton descent instead of gradient descent to optimize its trees, which leads to faster convergence. Finally, XGBoost also introduced a new feature split finding algorithm to speed up training. [50]

XGBoost gives very good performance in many problems, outperforming most of the times the commonly used models. This happens mainly, because it a scalable method to all scenarios, while also runs notable faster than them offering a computational relief even if ran on a common desktop. [52]

## Performance Comparison

#### SHAP Plots

Machine learning models steadily gain large applications on drug development, however their low interpretability of results, limits their implementation. SHAP plots is a feature-based plot method that enhanced ML models to better interpretability and thereby their trustworthy.

SHAP analysis is based on Shapley values, a concept from ‘collaborative game theory’. SHAP provides a fair distribution of a result payout among collaborative factors (e.g., features of a drug) where factors cooperate for a common purpose, without necessarily contributing equally.

For example, instead of ‘players’, let’s assume Feature A, Feature B and Feature C. Further, let’s assume a model’s outcome (e.g, outcome in percentage) instead of ‘game’, for which features contribute in order to be predicted.

Total response rate:

Feature A, Feature B, Feature C: 90%.

Each drug’s response rates:

Feature A: 40%

Feature B: 50%

Feature C: 60%

Paired drugs’ response rates:

Feature A, Feature B: 70%

Feature A, Feature C: 65%

Feature B, Feature C: 80%

Assumptions of SHAP:

• Efficiency: The sum of each drugs’ contributions equals the total response (e.g., 90%)

• Symmetry: If two Feature contribute always the same whichever the subset they belong (eg., pair or individually), they should receive an equal result payout.

• Additivity: If an intervention has multiple sub-samples and each has a separate payout result, then the contribution of each feature to the outcome is equal to the sum of contributions to each sub-sample. (e.g., two sub-populations of patients, the feature A contribution to the total population equals the sum of its contribution to each of the sub-populations).

• Null player: If a feature doesn't contribute to any subset (e., pair or individually), its share of the payout result is 0.

: Result payout

: ‘Player’ (e.g., Drug in this example)

SHAP value. One is assigned to each drug (‘player’ ) and corresponds to its contribution to the total outcome. SHAP values can be described as the weighted average of a drugs contributions across all possible subsets.

SHAP values satisfy above assumptions.

: Set of all drugs

: size of subset (eg., S = {Drug A, Drug B}) of size

: quantifies the contribution of drug j to subset S

: Weight of contribution.

: Weighted contribution of subset of size

sums all possible subset contributions without

In practice, other approximate methods are used to calculate SHAP values, as this formula would be calculating expensive.

Especially, in Pyhton SHAP package is used with e.g.,DeepExplainer, GradientExplainer etc.

: sample

: Model prediction for sample

: Average model prediction

: Sum of all SHAP values for sample i

Thus, for sample i, the SHAP values assigned to each feature describe how that feature contributed to the difference between the individual prediction of this sample from the average model prediction.

##### SHAP interpretation

Reported SHAP plots refer to test dataset. Different SHAP plots can illustrate explanation of feature contributions, locally i.e., on one data point or globally i.e., whole dataset.

* Bar Plot

Displays the mean absolute SHAP value for each feature, considering all data points i.e., global overview. Thus, it is a measure of feature importance but also ranks features top to bottom based on their mean absolute SHAP value or else impact on predictions.

Advantages:

Provides same units as model’s predictions, provides good interpretability.

A bar graph with numbers and symbols

AI-generated content may be incorrect.Disadvantages:

Lack of direction (eg., positive, negative) and monotonicity of impact.

* **Beeswarm Plot**

Displays SHAP value for each feature, considering all data points i.e., global overview, ordered by the mean absolute SHAP value. Thus, it is a measure of feature importance, which ranks features top to bottom based on their mean absolute SHAP value or else impact on predictions.

The dots are the SHAP values of each data point. Color range indicates feature values. High feature values are displayed as red, while lower as blue. Missing values, if any, are displayed as grey. Each dot is placed based on its SHAP value on x-axis.

Advantages:

The combination of dots’ color and place on x-axis indicates the direction of feature-prediction relationship, monotonicity, possible alike spreading or even outlier SHAP values.

Disadvantages:

A diagram of a graph

AI-generated content may be incorrect.Colour scale makes it difficult to surely explain the characteristics of relationship due to overlapping colored points, more complex relationships etc.

* **Scatter Plot**

Displays the relationship between one feature’s values (x-axis) and SHAP values (y-axis), considering all data points i.e., global overview but not for all features. On x-axis it may be displayed a histogram regarding feature value distribution.

Advantages:   
Relationship of feature-SHAP values trend indication (eg., linear/non-linear)

Interactions with other features if a vertical spread of SHAP values, or through coloring of points. The addition of trend lines makes patterns more obvious.

Can be used for k-fold validation, displaying all folds’ samples, and indicating different patterns within different samples of dataset. Each fold is assigned with a specific data point coloring.

A graph of value and values

AI-generated content may be incorrect.Disadvantages:  
Only one feature per plot.

* **Waterfall plot**

Displays SHAP values of a single data point but for all features i.e., it is a local overview. Displays features (y-axis) and SHAP values (x-axis). The value of that feature for that specific point is denoted in gray. SHAP values are displayed on the ‘arrow’ of each feature. Arrows are colored, with red color meaning that SHAP value increases the prediction , and blue if decreases it.

Advantages:

Outliers, SHAP value outliers, evaluate specific features.

A graph with numbers and a graph

AI-generated content may be incorrect.Disadvantages:  
Display only one feature.

##### SHAP for Classification

A close-up of a graph

AI-generated content may be incorrect.In classification models, SHAP plots may interpret: 1) probabilities, or 2) log-Odds. When probabilities are displayed, it is easier to interpret predictions, however, log-odds offer easier relationship interpretation between features (e.g., interactions etc.). Best practice is to evaluate both formats.

Example left probabilities are displayed on x-axis, while on the right log odds. [53]

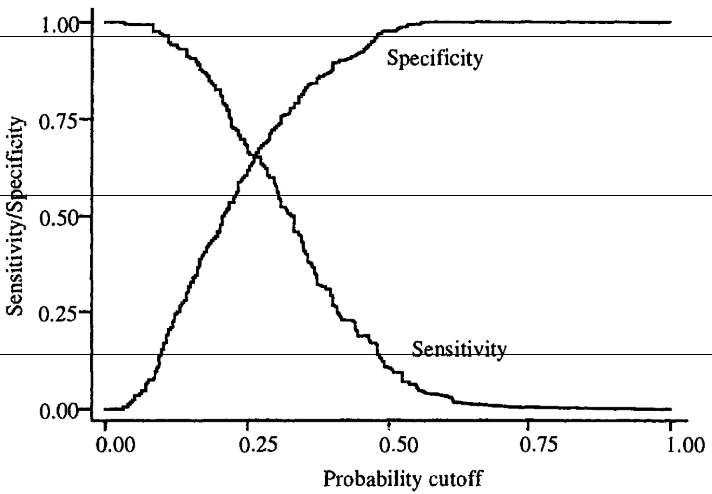
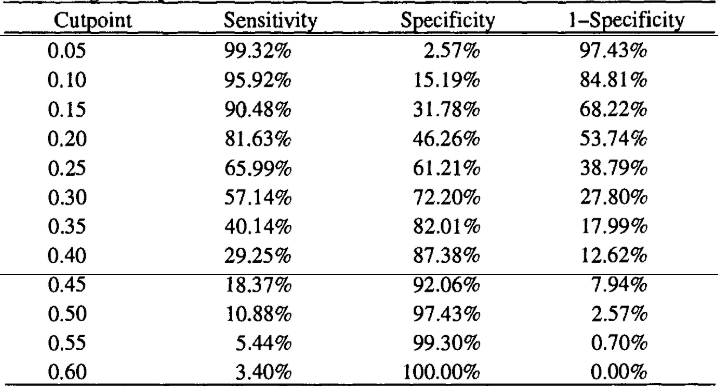
#### Roc Curve – ROC

##### Roc Curve

Receiver Operating Characteristic Curve (ROC) plots the probability of true signal detection (sensitivity) vs false signal detection (1-specificity) for a range of possible cut points. ROC curve is a completed description of a model’s accuracy.

Area Under the Curve (AUC) is a measure of the ability of the model to discriminate among the individuals who would achieve the outcome (Y=1) vs those who would not (Y=0). AUC values lie from 0 to.

As far as it concerns cut points, the usually used in classification is and accordingly. The use of other cut points induces a trade-off between sensitivity and specificity and thus 1-specificity as the example seen below. [54]

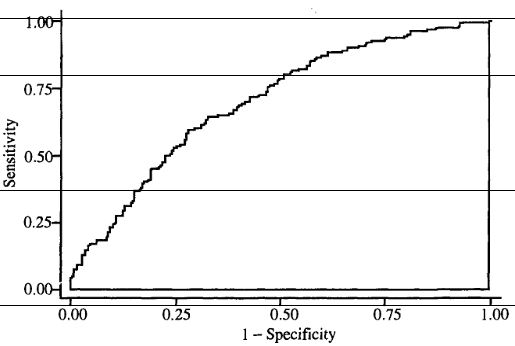
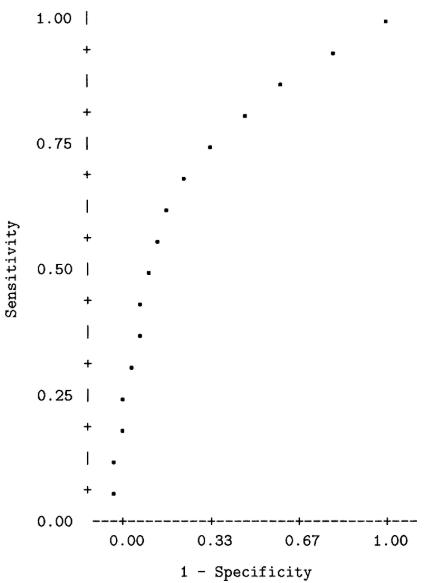


If cut point is near 0, almost all predictions are , sensitivity is near 1 and

speciﬁcity is near 0, and the point for (1 – speciﬁcity, sensitivity) has coordinates near (1, 1). If cut point is near 1, almost all predictions are , sensitivity is near

0 and speciﬁcity is near 1, and the point for (1 – speciﬁcity, sensitivity) has coordinates near (0, 0).

One could choose another cut point which maximizes both sensitivity and specificity, however there are statistical benefits from the usual 0.5 value. By taking the sensitivity and 1- specificity for each cut point, a dot arises, and connecting those dots, result to the ROC curve [54]. ROC curve has a concave shape connecting the points (0, 0) and (1, 1) [28].



[28][54]

##### Area Under the Curve AUC

The AUC is the probability that a subject achieving Y=1 outcome will have higher probability [54]. It is a measure of predictive power called ‘concordance index’. More specifically, let’s assume:

: all data points for which

: all data points for which

: pairs of each is paired with each

So, AUC is the proportion of those pairs for which (larger y) subject indeed had the higher the two probabilities (larger ). [28]

* : Good. is generally, considered acceptable.
* : Fair
* : Poor
* : Fail
* : like random guessing [55]
* : Excellent. In practice it is rarely observed and if yes it may be due to perfect separation problems in data (sparse data). [54]

#### Accuracy Score – Sensitivity - Specificity

##### A diagram of positive and negative AI-generated content may be incorrect.Sensitivity vs Specificity

* True Positives (TP): Samples for which true and predicted labels are both 1.
* True Negatives (TN): Samples for which true and predicted labels are both 0.
* False Positives (FP): Samples for which true label is 0 and predicted label is 1.
* False Negatives (FN): Samples for which true label is 1 and predicted label is 0.

: True state of a subject

= 1 = positive,

Y = 0 = negative

* Sensitivity (recall): The probability the diagnostic test is positive, given that subject is truly positive. Fraction of correct predicted positive to all truly positive.
* Speciﬁcity : The probability the diagnostic test is negative, given that subject truly does not have the outcome. Fraction of correct predicted negative to all truly negative.
* Positive predictive value (PPV) (precision): Fraction of correct predicted positive to all predicted positive.
* Negative predictive value (NPV): Fraction of correct predicted negative to all predicted negative.
* Accuracy: Fraction of correct predicted outcomes to all predictions.
* Balanced accuracy (BA): Accuracy metric that accounts for unbalanced samples.
* F1 score: Harmonic mean of PPV (precision) and sensitivity (recall).

Usually, the scores checked are sensitivity and speciﬁcity, for which the higher they are the better the diagnostic test.

[28][55][56]

# Methods

## Tools and Technologies

### Environment

Python version 3.12.10 through Visual Studio Code

pgAdmin4: version 9.0

All processes were managed with Python through Visual Studio Code and organized to jupyter notebook files, except part of processes regarding database source data, as explained below.

Jupyter files were organised based on the order of data engineering and analysis processes:

1. SQL

2. Data engineering

3. Visualization

4. Dummies-

5. Interactions

6. Models

7. Results

### Libraries

Python: Pandas, numpy, sklearn, statsmodels, sqlalchemy, pickle, matplotlib, seaborn, collections, geonamescache, scipy, itertools, patsy etc.

## Data Sources

* + - 1. [AACT Database](https://aact.ctti-clinicaltrials.org/downloads) of ClinicalTrials.gov:

AACT is a publicly available relational database that contains all information about every study registered in ClinicalTrials.gov. Content is downloaded from ClinicalTrials.gov daily and loaded into AACT. The Clinical Trials Transformation Initiative (CTTI) enhanced AACT in October 2016. Data are directly accessible in the cloud and can also be downloaded as a static copy. [57]

For this study, a static copy was downloaded in postgres.dmp format. The .dmp file provides SQL code for each table. Tables could be loaded by executing the corresponding SQL code. Specific tables were chosen to be studied, based on the features needed extraction. The use of database was the main source of data, for coherence and consistency of process, even if provided in structured format through CSV. Thus, database with combination to CSV, enhanced the creation of features, especially when one source provided a more unstructured format of a feature that the other.

For all data extracted, only ‘INTERVENTIONAL’ trials with start data >= ‘2011-01’ and study status 'COMPLETED', 'WITHDRAWN' or 'TERMINATED' were filtered (table columns: studies.study\_type, studies.study\_type,studies.overall\_status accordingly).

Tables from database was executed through on pgAdmin4 environment. After, Visual Studio Code was connected to database, and any data extracted or new tables created, were managed through this environment.[57]

* + - 1. [Clinical Trials.gov](https://clinicaltrials.gov/search) : Data downloaded in CSV format [6]. CSV was mainly used for comparison with database’s data, extraction of structured data that may not be provided in database source and also as the main source of trials according to their unique NCT numbers. It was not preferred as main source, because many columns provided data in unstructured text format, while in database source same data were structured.

**Filters applied:**

Study Status

* No longer looking for participants: Completed, Terminated
* Other: Withdrawn

Study Type

* Study Type: Interventional

Date Range

* Study starts: From: 0/01/2011
* Primary completion: To: 12/12/2024
  + - 1. [MeSH Terminologies Tree View](https://meshb.nlm.nih.gov/treeView): Disease, Conditions and phenomena categories. Terminologies and their codes were used in combination with corresponding tables of database. [58]

## Data Engineering - Feature Formats

### General

Three main data types/formats of features were analysed, with their columns renamed based on that:

* **Binary** : Suffix ‘\_Bin’ for easier recognition of data type/format.
* **Categorical**: Suffixes such as ‘\_Categ’ or ‘\_List’ for easier recognition of data type/format.   
  For, ‘\_Categ’ features, they all included two levels, thus no processing needed.   
  Especially for ‘\_List’ data that were initially unstructured text data that referred categorical features (e.g., ['CHILD', 'ADULT', 'OLDER\_ADULT'] for age feature). They were transformed into list element-data for convenience. This way two things were achieved, firstly, one trial record could belong to more than one levels of the same category, while data frame’s number of rows remained the same, without inflating the sample number. List element features were expanded and displayed as dummy variables for model and visual processes.
* **Continues**: Suffixes such as ‘\_counts’, ‘\_log’, ‘\_sqrt’ etc were added to naming for easier recognition of data type/format. Continues features were evaluated in different scales e.g., initial numeric, logarithmic, root etc. Continuous features were also examined as categorical with 2 levels (see ‘\_Categ’ above). The best format was chosen based on tests like AIC.

### Missing values

Missing values regarding numerical features, where dropped, as only a small amount was noticed compared to dataset (eg., only 54 rows for enrolment).

For categorical features, text mining was used to extract words from unstructured text data. For example, ‘Brief Summary’ column of csv source was widely used for text mining and null values filling. Similarly, happened for database text mining through different columns of tables.

Many missing values contained in columns: regarding Study Documents (145000 out of almost 180000 records), Locations, Outcomes etc. Rows remaining null which could not be filled through text mining, were dropped as their number was small comparing the dataset. Method of missing values filling is not referring to each feature/column of ‘Features Pre-Processing’ field for convenience, as it is similar for all columns.

### Data Engineering

#### AACT database

Specific tables were loaded by executing corresponding code from .dmp file. The choice of used tables was based on features needed extraction, that CSV format may not completely provide. From these tables others were created with the prefix ‘my\_’.

nct\_id column from studies table of AACT database, was used as the unique identifier of trials. It was joined with any table extracted or created from AACT and for any merge of AACT tables with csv tables etc. Thus, studies.nct\_id is selected for all below processes. It is not mentioned, for convenience, except the first table’s creation. Specifically, tables created from initial database tables and columns:

**my\_studies**: Tables: studies (Columns: nct\_id, brief\_title, official\_title, source, phase, number\_of\_arms, enrolment)  
For comparison to csv, number of arm groups, titles etc.

**my\_terminations**: Tables: studies (Columns: why\_stopped, overall\_status)  
Detection of termination related to accrual, funding, efficacy, toxicity etc. However, this table was not used.

**my\_conditions**:  Tables: mesh\_headings (Columns: qualifier, heading), conditions (Columns: name)

Conditions’ categories from [MeSH tree view](https://meshb.nlm.nih.gov/treeView) [58] was also used for this table. Detection of possible relation between disease/conditions’ category with probability of termination. Two levels of conditions from MeSH tree view were candidates for analysis, the general (e.g., C = Diseases) and the more detailed (e.g., C10 = Nervous System Diseases).

**my\_covid**: Tables: browse\_conditions (Columns: mesh\_term)  
Separated trials to covid and non-covid. Evaluation of termination relation to covid. Defined a binary feature.

**my\_placebo**: Tables: design\_groups (Columns: group\_type)   
Separated trials to containing placebo and not. Evaluation of termination relation to placebo intervention. Defined a binary feature .

**my\_interventions**: Tables: browse\_interventions (Columns: mesh\_term)   
Table of detailed intervention (e.g., Rifampin). However, it was not used for analysis.

**my\_interventions\_types:** Tables: interventions (Columns: intervention\_type)Evaluation of type of intervention relation with termination eg., ['DRUG']

**my\_intervention\_methods**: Tables: interventions (Columns: intervention\_type, description)

Several words were searched through the columns of this table, in order to categorise the intervention method to ['Oral', 'Topical', 'Injection', 'Surgical'].

**my\_soc**: Tables: interventions (Columns: description), design\_groups (Columns: description, group\_type)

Separates trials to including standard care (SOC) treatment and not. Evaluation of termination to standard care. A binary feature was produced.

**my\_adverse**: Tables: reported\_event\_totals (Columns: event\_type, subjects\_affected)

Produced table with features regarding adverse level of severity e.g., ['Death', 'Other', 'Serious'] and number of adverse events.

**my\_adverse\_system**: Tables: reported\_events (Columns: organ\_system)

Produced table regarding organs affected from adverse events e.g., ['General', 'Nervous System', 'Skin']

**my\_designs**: Tables: designs (Columns: allocation, intervention\_model, primary\_purpose, masking, subject\_masked, caregiver\_masked, investigator\_masked, outcomes\_assessor\_masked)

Produced table regarding study design eg., allocation (randomized/non-randomized), masking (single, double etc.), intervention model (parallel, crossover etc.), primary purpose (treatment etc.). These data were also available in csv in same format.

**my\_eligibilities**: Tables: eligibilities (Columns: gender, healthy\_volunteers)

Produced table containing features of gender eg., [‘MALE’, ‘FEMALE’, ‘ALL’] and trials accepting healthy volunteers or not (binary feature).

**my\_outcomes**: Tables: outcomes (Columns: outcome\_type)

Produce table of outcomes eg., ['PRIMARY', 'SECONDARY', ‘OTHER’]

**my\_locations**: Tables: countries (Columns: name), facilities (Columns: country, city)

Produced table of facilities sites locations and number of them. Geonamescache library was used for this table too.

**my\_documents**: Tables: provided\_documents (Columns: has\_protocol, has\_icf, has\_sap)

Produced table of documents uploaded in clinical trials.gov e.g., ['Analysis Plan', 'Protocol', 'Consent Form']. Unfortunately, not all trials fill this field properly as there are many missing values, thus this feature was not used (145.000 plus missing values out of almost 180.000 rows in dataset). Field seems to be optional to fill, as those documents are necessary to a study design.

#### Features Pre-Processing

All above features resulted from combination of csv with tables of database.

**Age\_List**: CHILD, ADULT, OLDER\_ADULT

**Sex\_List**: ALL, FEMALE, MALE. Was transformed into list. Missing values were filled with text mining through ‘Brief Summary’ column.

**Funder\_Type\_List**: Used directly from csv. OTHER, NETWORK, INDIV, UNKNOWN, AMBIG, INDUSTRY, OTHER\_GOV, FED, NIH were the initial levels of categorical feature. Some of those levels had too little sample and produced sparse data errors in visualization models and analysis models. So, these levels were merged to greater ones as follows: 'OTHER':'FUNDER\_OTHER', 'NETWORK':'FUNDER\_OTHER', 'INDIV':'FUNDER\_OTHER', 'UNKNOWN':'FUNDER\_OTHER', 'AMBIG':'FUNDER\_OTHER', 'OTHER\_GOV':'GOVERM', 'FED':'GOVERM', 'NIH':'GOVERM', 'INDUSTRY': 'INDUSTRY'.

Also, text mining was applied on ‘Sponsors’ and ‘Collaborators’ columns to create another level for educational/university and hospital facilities, that replaced some rows of ‘'FUNDER\_OTHER’ and ‘GOVERM’ levels. Final, levels were defined as follows: EDU\_UNIV, INDUSTRY, HEALTH, FUNDER\_OTHER, GOVERM.

‘Sponsors’ and ‘Collaborators’ columns were not used further in the analysis.

**Datetime columns:**

Datetimes of dataset appeared into two formats: YYYY-MM-DD or YYYY-MM. All datetime data were converted to format : YYYY-MM. Were used for below feature creation.

**Completion\_Gap\_Counts**: Feature was created from Completion Date - Start Date columns in month counts. Also, evaluated as **Completion\_Gap\_Categ** format.

**Intervention\_Type\_List**: 'BEHAVIORAL', 'BIOLOGICAL', 'COMBINATION\_PRODUCT', 'DEVICE', 'DIAGNOSTIC\_TEST', 'DIETARY\_SUPPLEMENT', 'DRUG', 'GENETIC', 'INTERV\_OTHER', 'PROCEDURE', 'RADIATION'. Feature used as provided from dataset.

**Intervention\_Type\_Counts**: Number of ‘Intervention\_Type\_List’ column. Also evaluated as **Intervention\_Type\_Categ** format.

**Intervention\_Method\_List**: ['Injection', 'Oral', 'Surgical', 'Topical']. Created by merging database table with csv (see AACT database tables above field). Missing values were filled with text mining applied on ‘Brief Summary’ and ‘Interventions’ columns.

**Intervention\_Method\_Counts**: Number of Intervention\_Method\_List. Also evaluated as **Intervention\_Method\_Categ** format.

**Placebo\_Bin**, **Standard\_Care\_Bin**, **Healthy\_Bin**, **Covid\_19\_Bin**: Binary data for placebo intervention, standard care intervention, healthy volunteers’ acceptance, covid trials accordingly. Created by merging of database tables (my\_placebo, my\_soc, my\_eligibilities, my\_covid accordingly) and csv (see AACT database tables above field).

**Conditions\_List, Conditions\_Detail\_List**: Created by merging database table (my\_conditions) with csv (see AACT database tables above field). Two levels of information depth (see ‘AACT database’ field above): Conditions\_List is general (e.g., C = Diseases) Conditions\_Detail\_List is moredetailed (e.g., C10 = Nervous System Diseases).

**Adverse\_List**: Severity level of adverse events. Created by merging database table (my\_adverse) with csv (see AACT database tables field). Missing data were considered as no adverse events occurring.

**Adverse\_Detail\_List:** Organ system that adverse was located.Created by merging database table (my\_adverse\_system) with csv (see ‘AACT Database’ field). Missing data were considered as no adverse events occurring.

**Adverse\_Counts**: Number of Adverse\_List. Created by merging database table (my\_adverse) with csv (see ‘AACT Database’ field). Was evaluated in different formats : **Adverse\_Counts\_Log, Adverse\_Counts\_Sqrt, Adverse\_Categ, Adverse\_Bin.**

**Allocation\_List**: RANDOMIZED, NON\_RANDOMIZED, NA\_RANDOMIZED (For one arm trials). Created by merging database table (my\_designs) with csv (see ‘AACT Database’ field).

**Intervention\_Model\_List**: PARALLEL, SINGLE\_GROUP, CROSSOVER, SEQUENCIAL, FACTORIAL. Created by merging database table (my\_designs) with csv (see ‘AACT Database’ field).

**Masking\_List**: MASK\_NONE, SINGLE, DOUBLE, TRIPLE, QUADRUPLE. Created by merging database table (my\_designs) with csv (see ‘AACT Database’ field).

**Masking\_Detail\_List**: ['CARE\_PROVIDER', 'INVESTIGATOR', 'PARTICIPANT', 'OUTCOMES\_ASSESSOR']. Created by merging database table (my\_designs) with csv (see ‘AACT Database’ field).

**Arm\_Counts**: Created by merging database table (my\_studies) with csv (see ‘AACT Database’ field). Was evaluated also as **Arm\_Counts Log, Arm\_Counts\_Sqrt, Arm\_Categ**

**Country\_Counts**, **Continents\_List, Continent\_Counts, City\_Counts:** Created by merging database table (my\_locations) with csv (see ‘AACT Database’ field). Geonamescache library was used too. Also evaluated as **City\_Categ, Country\_Categ, Continent\_Categ** formats.

**Enrollment\_Counts**: regarding number of participants. Used directly from csv. Also, evaluated as **Enrollment\_Log**, **Enrollment\_Sqrt, Enrollment\_Categ**

**Phases\_List**: Five data frames df0, df1, df2, df3, df4, df5 for were produced based on phases, and analysed separately from this study. The number each data frame name ends is based on the phase (e.g., df1 for phase 1) except df5 referring to NA phase type which was named like this for convenience.

#### Features Dropped – Not used

* Unique identifiers of trials, not study status associated:

**Study Title, official\_title, Other IDs, Acronym, Study URL, Brief Summary, nct\_id**

* Dates were not used for analysis, as they are not casual factor, but events happened during these periods are e.g., for year 2019 when covid studies started, covid trials were analysed.

**Start Date, Completion Date, Primary Completion Date,** **First Posted, Results First Posted, Last Update Posted**

* Too many missing data:

**Study Documents, Outcomes** (['OTHER\_PRE\_SPECIFIED', 'POST\_HOC',  
'PRIMARY', 'SECONDARY'])

* Not pre - Designed characteristics of trials, but results.

**Study Results, Primary Outcome Measures, Secondary Outcome Measures, Other Outcome** **Measures**

* Used through other columns/features:  
  **Sponsors, Collaborators**

Initial dataset for all phases with data frame name df had shape: df: (177586 rows)

## Visualization Pre-Processing

In this section plots and descriptive statistics were displayed per df of phase. Thus, pivot tables, chi-squared tests, count-plots, boxplots, kde-plots were displayed. Each feature was analysed through univariable logistic regression in relation to binary completion/termination outcome.

Additionally, in this section the best performed format of features were also chosen e.g., initial numeric, logarithmic etc. scale, through tests such as AIC. Tables for p-AIC an also p-values, confidence intervals from logistic models’ outputs were also displayed here.

All procedures were implemented through the creation of functions.

For achieving analysis per phase of trial, an input cell was placed on top of this step’s jupyter file, so one can input the data frame corresponding to trial want to analyse.

## Dummies

‘\_List’ data: List element data, were encoded through dummy creation. The problem was that list-element data cannot be used from models. Firstly, list-element rows were exploded to one row per element of list. However, this inflates sample size, as rows increase and duplicate for same nct\_id if more than one element occur in a list-row. For this reason, these data were then grouped by nct\_id, so number of rows remained the same as initial datasets (df0, df1, df2, df3, df4, df5).

‘\_Categ’ data: Were encoded as simple categorical data types with 2 levels. None of them had more than 2 levels, and no record could belong to more than one level.

‘\_Counts’, ’\_Log’, ’\_Sqrt’: Continues data were used as is and scaled to next level of analysis as mentioned below.

In this section, suffixes helped to choose which columns to binarize (0,1), make dummies when more than 2 levels or keep as in case of continuous data.

## Interactions

Interaction terms for Funder Type (e.g., INDUSTRY) and Intervention Type (e.g., DRUG) were created. As mentioned in literature, funder type is associated with trial termination, especially if industry (see Background field). However, this analysis purpose was to evaluate if funder type like industry (which most likely is pharmaceutical industry), considering intervention like drug (i.e., pharmaceutical product) results to greater probability of termination.

## Models

In this section, each trials phase was analyzed for early termination probability based on its characteristics. The data frames for each phase with shapes:   
df0: (2236, 130), df1: (26045, 155), df2: (28409, 155), df3: (16093, 155)

df4: (12813, 130), df5: (95731, 155), df: (173541, 160) containing all phases data frames.

Initially, each dataset was split into train and test set on with different proportion for each df. The train-test set proportion was decided based on background similar study that proposed 70:30 [11]. However, under-sampling used after train-test split (explained below), minimizes train set so that in some phases test set is greater.

Then train set only was Under sampled, due to great imbalance towards ‘Completed’ outcome. Test set should be left with the true distribution. Random Under sampling was chosen as the simplest method. In contrast, random over-sampling includes resampling methods to create greater ‘termination’ category (e.g., resampling minority class with replacement). This decision was also made from literature background [11]. In under-sampling methods, train set is minimized, as many rows are dropped from majority category. Combining the great sample sizes of data frames and the appropriate train-test split analogy per data frame, train samples remain efficient, with predictors by far less in size .

Scaling, used on train and test set. Scaling by using mean produced negative values in continues features, which is not justified from dataset values, so it was chosen to scale only by using standard deviation.

Finaly, models selected for performance comparison were:

* Classic Statistical Models:

Logistic Regression with Lasso (L1 penalty), Elastic Net Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA)

* Machine Learning Models:

K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Classification Pruned Tree, Random Forest, Gradient Boosting, Extreme Gradient Boosting (XGBoost).

Parameter tuning for models, was managed through k-fold cross validation with k=5 folds. Evaluation of models was done on test set, and metrics of accuracy, precision, recall, f1-score and AUC curves.

Furthermore, predictors contribution was evaluated through SHAP plots.

For achieving analysis per phase of trial, an input cell was placed on top of this step’s jupyter file, so one can input the data frame corresponding to trial want to analyse.

# Results

After pre-processing data engineering each phase was analysed separately through corresponding data frame.

## Descriptive Statistics – Visualization

In this section distribution of sample.

## Selected Model

XGBoost model outperformed at almost all phases (phase 1-phase 4 and non-applicable study phase), with exception of early phase 1 (named as phase 0 in this study) where QDA model was evaluated ss the best classifier. Model performance is in perfect alignment with background studies’ selected models. Models’ performance was evaluated with accuracy, precision, recall, f1-score and AUC curves.

## Features Contribution

Top feature contributor was for all phases the Enrolment for all phases. Another feature is Neoplasms disease RCTS, in mostly in early phases. Especially for phases 0 to 2, neoplasms undertake the second position in SHAP plots, while for phase 3 they much lower. Another, feature regards the Adverse event counts, which are present at the third position for phases 2-3 and lower for phase 4.

Feature contribution seems to be related with study phase. For example, adverse events are top contributor to corresponding phases 2-3, which main purpose is efficacy and drug safety, or later phases couldn’t initiate. Facilities and locations are another feature arising in different formats (e.g., city\_categ, country\_counts, continents etc.). Additionally, industry funder type was not a top contributor for this analysis, while mentioned by literature.  
Result of this study is aligning with background results of previous similar studies in most of the outcomes, as far as it concerns feature contribution.

Differences are observed mostly to disease categories or type of interventions. This study did not conclude any surgical or cardiovascular field disease as high contributors, as mentioned by literature, while intervention method of injection was in top 20 features in many phases and not surgical as mentioned from background studies.

# Discussion

## Comparison with Literature - Strengths and Limitations

Although, most of the results are aligned with background similar studies, some topics differ from this study either on models, results and data engineering methods. Features contribution and models’ performance are in line with background studies. Methods used were also similar. In general, this study analysed more features and evaluated more models, for all trial phases. So it is a more general approach not limited to specific topics or stages of n RCT.

### Strengths

More specifically, previous studies did not consider interactions such as funder type and intervention type (eg., drug-industry sponsor). Regarding adverse events, this study considered the type and the number of adverse events, but also the organ system located (e.g., ear), while in literature disease category is mostly referred and not adverse counts. The type of intervention method was analysed e.g., oral, topical, injection, surgical.

This analysis checked for different formats of a feature (e.g., enrolment, enrolment\_log etc.). Additionally, more models were evaluated than in literature but also ML to non-ML models were compared. In background studies mostly Random Forests and secondly XGBoost models are evaluated. Precent study had double purpose, first to compare classic statistical models with Machine Learning models and secondly to evaluate which model outperformed to prediction of termination probability.

Feature of Funder Type was analysed not only based provided categories (e.g., industry) but also two additional categories that were created for educational and hospital facilities, but also the interaction with intervention type (e.g., INDUSTRY x DRUG).

Finaly, all phases and even ‘Not Applicable’ type of phases were analysed separately, while in literature, most of the times, not all stages are evaluated.

### Limitations

The presence of missing values decreased sample size for many features and to features defined from others e.g., country\_counts as defined Countries\_Listetc.

Missing data were filled, where possible, while background studies usually dropped them based on a pre-specified field needed to have no nulls. However, current study did not use NLP text mining methods, which is mentioned in literature studies. Text mining methods used here, were based on words mining from unstructured text data (e.g., str.contains() command etc.), with which it is not possible to search for all words leading to the desired result.

What need to be done additionally, is the analysis of enrollment feature as interaction to disease field, location etc. to conclude why low enrolment occurs as other studies have done.

## Recommendations for Practice

Literature seems to provide common outcomes, regarding the early termination reasons of RCTs. What needs to be done is further investigation those specific reasons, including interactions with other features, in order to evaluate them better. For example, low accrual is a common top contributor for trial termination, but the factors which lead to low accruals rates are not deeply evaluated. Similarly, for the second top contributor, which is neoplasm diseases, where one could expect to have greater completion rates.

All these need further investigation to detect deeper reasons leading to early termination, thus improving trials’ design or even communicating consent forms in a more efficient way to attract more volunteers, especially for crucial disease fields.

# Appendices

## Python Code

## Data Dictionaries

# References

[1] C. A. Umscheid, D. J. Margolis, and C. E. Grossman, “Key concepts of clinical trials: A narrative review,” Sep. 2011. doi: 10.3810/pgm.2011.09.2475.

[2] M. E. Elkin and X. Zhu, “Predictive modeling of clinical trial terminations using feature engineering and embedding learning,” *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-82840-x.

[3] E. C. Eypasch RolfLefering K Kum Hans Troidl, E. Eypasch, C. K. Kum, and H. Troidl, “Probability ofadverse events that have not yet occurred: a statistical reminder.”

[4] A. P. Prayle, M. N. Hurley, and A. R. Smyth, “Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: Cross sectional study,” *BMJ (Online)*, vol. 344, no. 7838, Jan. 2012, doi: 10.1136/bmj.d7373.

[5] R. Talebi, R. F. Redberg, and J. S. Ross, “Consistency of trial reporting between ClinicalTrials.gov and corresponding publications: One decade after FDAAA,” *Trials*, vol. 21, no. 1, Jul. 2020, doi: 10.1186/s13063-020-04603-9.

[6] “ClinicalTrials.gov Study Search.” Accessed: Sep. 02, 2025. [Online]. Available: https://clinicaltrials.gov/search

[7] N. Chaturvedi, B. Mehrotra, S. Kumari, S. Gupta, H. S. Subramanya, and G. Saberwal, “Some data quality issues at ClinicalTrials.gov,” *Trials*, vol. 20, no. 1, Jun. 2019, doi: 10.1186/s13063-019-3408-2.

[8] E. Tang, P. Ravaud, C. Riveros, E. Perrodeau, and A. Dechartres, “Comparison of serious adverse events posted at ClinicalTrials.gov and published in corresponding journal articles,” *BMC Med*, vol. 13, no. 1, Aug. 2015, doi: 10.1186/s12916-015-0430-4.

[9] S. G. Nicholls *et al.*, “A review of pragmatic trials found a high degree of diversity in design and scope, deficiencies in reporting and trial registry data, and poor indexing,” *J Clin Epidemiol*, vol. 137, pp. 45–57, Sep. 2021, doi: 10.1016/j.jclinepi.2021.03.021.

[10] “ClinicalTrials.gov Protocol-Definitions.” Accessed: Sep. 03, 2025. [Online]. Available: https://clinicaltrials.gov/policy/protocol-definitions

[11] E. Kavalci and A. Hartshorn, “Improving clinical trial design using interpretable machine learning based prediction of early trial termination.,” *Sci Rep*, vol. 13, no. 1, p. 121, Jan. 2023, doi: 10.1038/s41598-023-27416-7.

[12] G. Antes and I. Chalmers, “Under-reporting of clinical trials is unethical,” Mar. 22, 2003, *Elsevier B.V.* doi: 10.1016/S0140-6736(03)12838-3.

[13] R. J. Williams, T. Tse, K. DiPiazza, and D. A. Zarin, “Terminated trials in the clinicaltrials.gov results database: Evaluation of availability of primary outcome data and reasons for termination,” *PLoS One*, vol. 10, no. 5, May 2015, doi: 10.1371/journal.pone.0127242.

[14] L. Follett, S. Geletta, and M. Laugerman, “Quantifying risk associated with clinical trial termination: A text mining approach,” *Inf Process Manag*, vol. 56, no. 3, pp. 516–525, May 2019, doi: 10.1016/j.ipm.2018.11.009.

[15] S. Geletta, L. Follett, and M. Laugerman, “Latent Dirichlet Allocation in predicting clinical trial terminations,” *BMC Med Inform Decis Mak*, vol. 19, no. 1, Nov. 2019, doi: 10.1186/s12911-019-0973-y.

[16] B. Kasenda *et al.*, “Prevalence, characteristics, and publication of discontinued randomized trials,” *JAMA*, vol. 311, no. 10, pp. 1045–1051, Mar. 2014, doi: 10.1001/jama.2014.1361.

[17] I. Baldi, C. Lanera, P. Berchialla, and D. Gregori, “Early termination of cardiovascular trials as a consequence of poor accrual: Analysis of ClinicalTrials.gov 2006-2015,” *BMJ Open*, vol. 7, no. 6, Jun. 2017, doi: 10.1136/bmjopen-2016-013482.

[18] B. Carlisle, J. Kimmelman, T. Ramsay, and N. MacKinnon, “Unsuccessful trial accrual and human subjects protections: An empirical analysis of recently closed trials,” *Clinical Trials*, vol. 12, no. 1, pp. 77–83, Feb. 2015, doi: 10.1177/1740774514558307.

[19] S. J. Chapman, B. Shelton, H. Mahmood, J. E. Fitzgerald, E. M. Harrison, and A. Bhangu, “Discontinuation and non-publication of surgical randomised controlled trials: Observational study,” *BMJ (Online)*, vol. 349, Dec. 2014, doi: 10.1136/bmj.g6870.

[20] K. M. Gayvert, N. S. Madhukar, and O. Elemento, “A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials,” *Cell Chem Biol*, vol. 23, no. 10, pp. 1294–1301, Oct. 2016, doi: 10.1016/j.chembiol.2016.07.023.

[21] “Statistics review 3: Hypothesis testing and P\_values,” 2002. [Online]. Available: http://ccforum.com/content/6/3/222

[22] K. F. . Weaver, *An introduction to statistical analysis in research : with applications in the biological and life sciences*. John Wiley & Sons, Inc., 2018.

[23] Σ. Μαλεφάκη, Α. Μπατσίδης, and Π. Οικονόμου, “Στατιστική Ανάλυση Δεδομένων,” 2023. Accessed: Aug. 14, 2025. [Online]. Available: http://dx.doi.org/10.57713/kallipos-321

[24] E. Whitley and J. Ball, “Erratum: Statistics review 3: Hypothesis testing and P values (Critical Care (2003) vol. 7 (15)),” Feb. 2003. doi: 10.1186/cc1868.

[25] E. Whitley and J. Ball, “Erratum: Statistics review 3: Hypothesis testing and P values (Critical Care (2003) vol. 7 (15)),” Feb. 2003. doi: 10.1186/cc1868.

[26] A. Agresti, “An Introduction to Categorical Data Analysis Second Edition.”

[27] “P  VALUE, A TRUE TEST OF STATISTICAL SIGNIFICANCE A CAUTIONARY NOTE”.

[28] A. Agresti, “An Introduction to Categorical Data Analysis Second Edition.”

[29] R. A. Poldrack, G. Huckins, and G. Varoquaux, “Establishment of Best Practices for Evidence for Prediction: A Review,” May 01, 2020, *American Medical Association*. doi: 10.1001/jamapsychiatry.2019.3671.

[30] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, “Cross-validation pitfalls when selecting and assessing regression and classification models,” *J Cheminform*, vol. 6, no. 1, Mar. 2014, doi: 10.1186/1758-2946-6-10.

[31] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, “An Introduction to Statistical Learning,” 2023.

[32] U. M. Braga-Neto, A. Zollanvari, and E. R. Dougherty, “Cross-validation under separate sampling: Strong bias and how to correct it,” *Bioinformatics*, vol. 30, no. 23, pp. 3349–3355, Dec. 2014, doi: 10.1093/bioinformatics/btu527.

[33] H. H. Rashidi, N. K. Tran, E. V. Betts, L. P. Howell, and R. Green, “Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods,” 2019, *SAGE Publications Ltd*. doi: 10.1177/2374289519873088.

[34] J. White and S. D. Power, “k-Fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation,” *Sensors*, vol. 23, no. 13, Jul. 2023, doi: 10.3390/s23136077.

[35] P. Ranganathan, “An introduction to statistics: Choosing the correct statistical test,” *Indian Journal of Critical Care Medicine*, vol. 25, no. S2, pp. S184–S186, 2021, doi: 10.5005/JP-JOURNALS-10071-23815.

[36] L. G. . Grimm and K. Paul. Nesselroade, *Statistical applications for the behavioral and social sciences*. Wiley, 2019.

[37] D. M. Lane, D. Scott, M. Hebl, R. Guerra, D. Osherson, and H. Zimmer, “Introduction to Statistics.”

[38] “Introduction to Mathematical Statistics.”

[39] M. E. Solari, “The Distribution of the Chi Square Test of Fit Statistic,” *The Statistician*, vol. 13, no. 4, p. 263, 1963, doi: 10.2307/2987306.

[40] S. Le Cessie and J. C. Van Houwelingen, “Ridge Estimators in Logistic Regression,” 1992.

[41] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” 2005.

[42] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization Paths for Generalized Linear Models via Coordinate Descent.”

[43] T. Hastie, R. Tibshirani, and J. Friedman, “The Elements of Statistical Learning Data Mining, Inference, and Prediction.”

[44] J. Taylor and R. Tibshirani, “Post-selection inference for ℓ1-penalized likelihood models,” *Canadian Journal of Statistics*, vol. 46, no. 1, pp. 41–61, Mar. 2018, doi: 10.1002/cjs.11313.

[45] J. K. Tay, B. Narasimhan, and T. Hastie, “Elastic Net Regularization Paths for All Generalized Linear Models,” *J Stat Softw*, vol. 106, 2023, doi: 10.18637/jss.v106.i01.

[46] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent,” 2011. [Online]. Available: http://www.jstatsoft.org/

[47] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, “Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications,” *J Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00973-y.

[48] S. Dasgupta and S. Kpotufe, “Nearest-Neighbor Classification and Search.”

[49] N. Bhatia, “Survey of Nearest Neighbor Techniques,” 2010. [Online]. Available: http://sites.google.com/site/ijcsis/

[50] D. Boldini, F. Grisoni, D. Kuhn, L. Friedrich, and S. A. Sieber, “Practical guidelines for the use of gradient boosting for molecular property prediction,” *J Cheminform*, vol. 15, no. 1, Dec. 2023, doi: 10.1186/s13321-023-00743-7.

[51] xgboost developers. © Copyright 2022, “DMLC XGBoost Documentation.”

[52] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[53] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, “Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development,” *Clin Transl Sci*, vol. 17, no. 11, Nov. 2024, doi: 10.1111/cts.70056.

[54] “Applied Logistic regression”.

[55] F. S. Nahm, “Receiver operating characteristic curve: overview and practical use for clinicians,” *Korean J Anesthesiol*, vol. 75, no. 1, pp. 25–36, Feb. 2022, doi: 10.4097/kja.21209.

[56] G. Varoquaux and O. Colliot, “Evaluating Machine Learning Models and Their Diagnostic Value,” in *Neuromethods*, vol. 197, Humana Press Inc., 2023, pp. 601–630. doi: 10.1007/978-1-0716-3195-9\_20.

[57] “AACT.” Accessed: Sep. 06, 2025. [Online]. Available: https://aact.ctti-clinicaltrials.org/

[58] “MeSH Tree View”, Accessed: Sep. 06, 2025. [Online]. Available: https://meshb.nlm.nih.gov/treeView