

Predicting Success of Phase III Clinical Trials

Colin Scherer

scherc3@rpi.edu

Rensselaer Polytechnic Institute

Troy, New York, USA

Abstract

Clinical trials often take years and millions of dollars to perform, usually yield disappointing results, and end up with a treatment that is ultimately infeasible for public use. In this paper, I explore the inclusion of machine learning techniques in optimizing clinical trial results by predicting whether a given trial's primary endpoint will be statistically significant. I expand upon previous research by considering mainly trial design parameters that are changeable, allowing for an alteration within the design to increase the chance of a significant result. When considering these parameters, I obtain a promising accuracy of 0.720 for random forest classification, and can accurately predict the p-values of a given clinical trial.

1 Introduction

In this paper, I compare four potential ML-based strategies for predicting whether the null hypothesis of a Phase III clinical trial will be rejected, based **purely on trial design parameters**. I also aim to predict the exact p-values of the primary endpoints.

Due to the monetary and temporal costs of clinical trials, it is imperative that the trial design maximizes the chance of meaningful results, and ML models can be used to guide changes in parameters such as target enrollment, minimum / maximum age, and number of sites.

I hope to improve on the current usage of ML in clinical trials by creating an effective model that can take as input current trial parameters and then predict whether the p-value will be statistically significant. With this prediction, principal investigators (PIs) and statisticians can modify certain trial parameters to increase the likelihood that the results will be statistically significant, or avoid conducting the trial entirely, saving money and time. I only consider Phase III trials here due to their importance in the drug approval process – while Phase I trials tend to focus on the pharmacokinetics of the drug and Phase II on proof of concept and dose finding, Phase III trials are used for regulatory approval and primarily test one dose of the drug versus an approved treatment and/or placebo. The robustness of the parameters in a Phase III trial is therefore of the utmost importance in the probability of drug approval.

I will be comparing four different models to develop an effective strategy to accomplish this task. Although there is some research within this area, I construct a dataset from scratch with different features in hopes to represent a more realistic scenario.

2 Related Work

The influence of certain characteristics on the success of clinical trials has already been investigated using ML methods such as neural networks, support vector machines (SVM), or random forest (RF)

[3][5]. In this paper, I expand on this research by omitting certain characteristics previously used, such as date and duration. With these alterations, I consider a more realistic setting: the removal of some variables ensures that all input features are known before the beginning of the trial, and *can be changed*.

Machine learning methods have already been applied to the task of predicting the likelihood of success of a clinical trial using neural networks along with drug information ([6], [8]). Compared to this research, I tackle a more difficult task by considering only trial design parameters, omitting any drug information besides previous (Phase II) performance.

3 Dataset Creation

As there are no existing datasets for this task to my knowledge, all trial information had to be compiled and processed manually using the API from clinicaltrials.gov. With this API, 17 features were extracted for each of the 527,941 clinical trials listed. Since results are required for ML training, the data was first filtered by trials with results reported, and afterwards filtered to only include Phase II or Phase III trials. Since I would be predicting whether the trial was statistically significant, I also required a reported p-value for the primary endpoint. Other restrictions included: interventional study type, information about arms reported, and information about trial design reported.

After compiling data from the API, further preprocessing is performed to obtain the final features listed in Table 1. All categorical features are one-hot encoded and, in the case of categorical features in a list, multiple ones are recorded per entry. In the case of the "Derived Terms" feature, terms with 'high' relevance are assigned a 1, and terms with 'low' relevance are assigned a 0.25. Continuous features are z-normalized (setting $\mu = 0$ and $\sigma = 1$). Because reported clinical p-values tend to follow a highly skewed-right distribution, results are normalized by first taking the logarithm of the p-values and then z-normalizing the resulting distribution as described above.

3.1 Phase II Result Integration

One key feature that could be highly indicative of the success of Phase III is Phase II results. Because the source API does not directly list the Phase II success of a particular treatment, these data must be manually extracted. For each Phase II trial in the dataset, experimental treatment, derived terms, and p-value are gathered. Afterwards, a list of potential relevant Phase II results for the experimental treatment in a given Phase III trial is gathered. Since a given treatment could be tested for various indications and dosages, there are many potential candidate Phase II trials for each Phase III trial.

The best candidate from these trials is determined using a nearest-neighbor algorithm – the derived terms for each candidate trial are

Feature	Description	Type	Unique Values
Conditions	A list of conditions the trial is meant to treat	Categorical, multi-label	1996
Collaborators	A list of agencies with a vested interest in the trial	Categorical, multi-label	510
Organization	The type of organization overseeing the trial	Categorical	8
Allocation	The method of allocation used for assigning treatments	Categorical	2
Interventional Model	The method of assigning treatments	Categorical	6
Primary Purpose	Purpose of treatment	Categorical	9
Blinding	The number of groups blinded	Categorical	4
Enrollment	The <i>actual</i> enrollment number for the trial	Continuous	N/A
Healthy Volunteers	Whether or not the trial includes healthy volunteers	Boolean	2
Sex	The sexes included in the trial	Categorical	3
Minimum Age	Minimum age included	Continuous	N/A
Maximum Age	Maximum age included	Continuous	N/A
Number of locations	The number of centers conducting the trial	Continuous	N/A
Derived terms	Terms related to the trial and the relevance of each term	Categorical, multi-label	3713
Interventions	Type of interventional arms in trial	Categorical, multi-label	6
Phase 2 Results	p-value of primary endpoint of most similar Phase II trial	Continuous	N/A
Results	p-value of primary endpoint in trial of interest	Continuous	N/A

Table 1: List of features used in model.

compared to the derived terms for the Phase III trial in question, and an affinity score for each candidate is calculated depending on the overlap between the derived terms. Then, the p-value and affinity score of the Phase II candidate with the highest affinity score are z-normalized and used as features for the Phase III trial in question.

Statistics and selected distributions of features are provided in Appendix A.

4 Methodology

4.1 Classification Task

For the task of classifying whether the results of a Phase III clinical trial would be statistically significant, I compare 4 models: support vector machines (SVM), random forest (RF), XGBoost [2], and multi-layer perceptron (MLP).

SVM was chosen due to its ability to handle high-dimensional data well due to built-in regularization. Random forest also tends to handle high-dimensional data well, and is more scalable than SVM. XGBoost tends to perform better than random forest in most settings, so it is also a good candidate for this task.

The MLP contains L hidden layers – each hidden layer i has l_i hidden neurons and is followed by Batch Normalization [4], ReLU activation, and a dropout layer with dropout probability p_d . Therefore, the output of each layer is given by:

$$z^{(i+1)} = \text{Dropout}(\text{ReLU}(\text{BatchNorm}(\mathbf{W}^{(i)}z^{(i)} + b^{(i)}))) \quad (1)$$

where $\mathbf{W}^{(i)} \in \mathbb{R}^{l_{i-1} \times l_i}$ and $b^{(i)} \in \mathbb{R}^{l_i}$ are weights and biases learned during training.

For the output layer, the BatchNorm and dropout layers are removed, and the ReLU activation is replaced by a sigmoid activation, which reduces the possible output to the range [0,1].

The loss function for the MLP is binary cross-entropy, which is defined by

$$\frac{1}{B} \sum_{n=1}^B y_n \log(\hat{y}_n) + (1 - y_n) \log(1 - \hat{y}_n) \quad (2)$$

where B is the batch size, y_n and \hat{y}_n are the true and predicted labels for a given sample.

4.2 Regression Task

In addition to classification, I also consider a regression task – predicting the exact p-value of a clinical trial, given its parameters. For this task, I compare three different models: RF, XGBoost, and MLP.

The MLP architecture for regression is exactly the same as the described classification architecture, with the exception of the output layer, which does not have an activation function. The loss function used for optimization was L1 (mean absolute error) loss.

5 Experimental Setup and Results

Both MLPs were implemented using PyTorch 2.3.0 and optimized using the AdamW optimizer. The other models were implemented using scikit-learn. Hyperparameter tuning was performed for each model, which is described in Appendix C.

After filtering the clinicaltrials.gov data, my dataset contains 4228 trials with 6265 dimensions for each trial. Because the data is high-dimensional, PCA is done to decrease the number of features per trial to 256. Model performance on the reduced dataset is compared with performance when dimensionality reduction is not applied in order to analyze the impacts of this reduction.

The dataset is randomly divided into training, validation, and test sets in a 3:1:1 ratio. Due to the unbalanced nature of the classes, upsampling is performed on the training set by randomly duplicating ‘unsuccessful’ trials with replacement until the class balance is even, i.e. there are as many unsuccessful as successful trials in the training set. To ensure accurate results and reduce random seed

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
SVM	0.680±0.019	0.725±0.028	0.761±0.026	0.742±0.025	0.658±0.013
RF	<u>0.685±0.021</u>	<u>0.705±0.026</u>	<u>0.827±0.019</u>	<u>0.761±0.021</u>	<u>0.647±0.019</u>
XGB	0.675±0.013	0.702±0.023	0.805±0.020	0.750±0.016	0.640±0.012
MLP	0.697±0.029	0.697±0.035	0.884±0.023	0.779±0.026	<u>0.647±0.019</u>

Table 2: Results for the classification task with PCA

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
SVM	0.708±0.013	0.739±0.014	0.816±0.013	0.776±0.012	0.673±0.017
RF	0.721±0.008	0.742±0.016	0.842±0.013	0.789±0.008	0.682±0.014
XGB	0.696±0.012	0.731±0.018	0.807±0.016	0.767±0.009	0.661±0.016
MLP	0.706±0.004	0.731±0.010	0.838±0.012	0.781±0.002	0.661±0.012

Table 3: Results for the classification task without PCA

Model	MAE	Median negative p-value	Median positive p-value
RF	0.961±0.017	0.02532±0.0065	0.00438±0.0007
XGB	0.958±0.014	0.0217±0.0048	0.0055±0.0008
MLP	0.717±0.018	0.0274±0.0048	0.0037±0.0009

Table 4: Results for the regression task

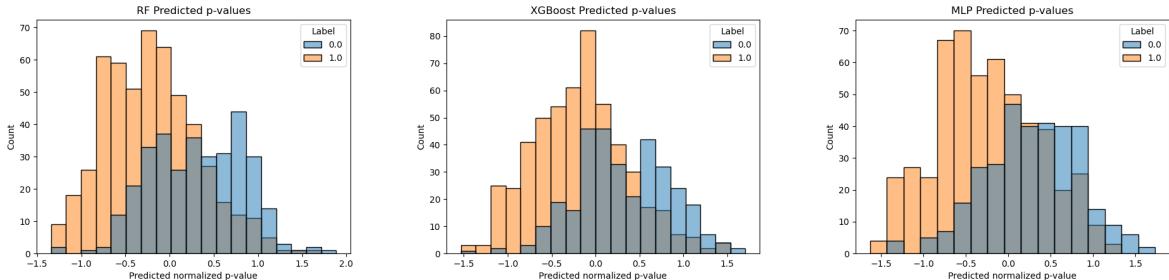


Figure 1: Histograms of predicted p-values for each regression model. Blue bars correspond to ground truth failed trials, while orange bars correspond to ground truth successful trials. The gray bars are overlaps between the distributions. Predicted p-values are in the transformed space as described in Section 3 (For reference, 0.05 ≈ 0.48 in the transformed space).

dependence, results are averaged over five separate training and evaluation runs.

5.1 Classification Results

Accuracy, precision, recall, F1, and ROC-AUC scores are reported for each classification model.

Table 2 summarizes the results (mean±SD) obtained with PCA, while table 3 summarizes the results without PCA.

Random forest clearly shows the best results in the classification task, outperforming every other model for the five reported metrics when dimensionality reduction is omitted. Somewhat surprisingly, transforming the data with PCA reduces overall performance, which implies that dimensionality reduction somehow removes some meaning from the dataset. However, it tends to significantly decrease runtime as well, which could be an important factor for larger datasets.

Due to the inferior results for classification after applying PCA, I omit PCA analysis for the regression task.

5.2 Regression Results

Mean absolute error of the transformed p-value is reported for each regression model. In addition, I report the median predicted p-values (untransformed) for each ground truth class.

Table 4 summarizes the results for the regression task. In general, the results follow expected trends: the median predicted p-values for unsuccessful trials are about 5 times greater than those for successful trials – the models tend to correctly predict successful trials as having lower p-values. However, the median predicted p-values for unsuccessful trials are still below the significance level of 0.05, which means that most of the unsuccessful trials would still be predicted to be statistically significant, which isn't an ideal result. For this task, MLP seems to perform the best, with the lowest MAE,

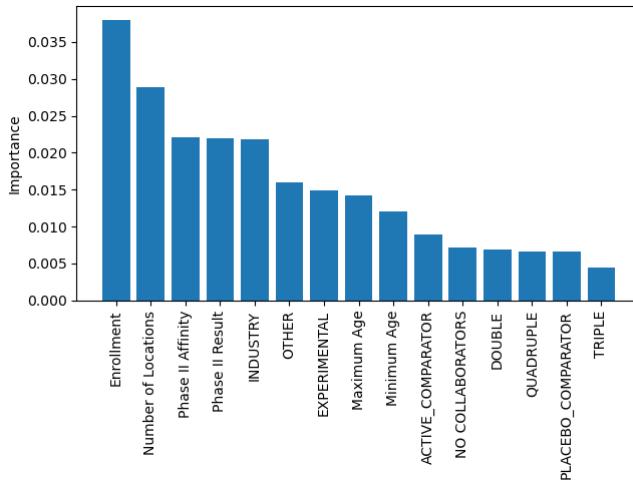


Figure 2: Relative importances for selected features in determining RF classifier.

highest predicted negative p-value, and lowest predicted positive p-value.

Along with these metrics, I also provide a visualization of the p-value distributions generated by each model in figure 1. Although MLP obtains the best metrics, RF and XGBoost seem to have a higher separation in the visual distributions, implying their potential in certain scenarios. Additional scatter plots of predicted vs. actual p-values can be found in appendix B.

5.3 Feature Importance

One useful application of these models is in determining the importance of various parameters in the design of clinical trials. Because RF tracks the importance of each feature in determining the final classifier, we can visualize how each factor impacts the final decision.

Feature importances are shown in figure 2. As expected based on the results obtained by [3], enrollment seems to have the highest importance, with the number of locations being the second-most important. The importance of inclusion of Phase II results in my model is also apparent, since both the affinity and results of a previous Phase II trial make up 2% of the RF classifier. Also included among the important features are the other continuous variables and some categorical variables, such as the type of organization and the blinding scheme used.

6 Discussion

Limitations. There are a few limitations to these results. First, the omission of entries with missing data can create heavily biased datasets, since missing data is much more common for failed trials rather than successful trials [6]. Second, up to now, I have implied that a p-value of less than 0.05 means the trial is successful and therefore the drug will get approved for public use. However, this is not necessarily true and can depend heavily on the drug, indication, and secondary endpoints. For example, there have been instances in which drugs have been approved without meeting the primary

endpoint criteria ([7], [1]). Additionally, there are instances where a trial's results *are* statistically significant but the drug does not get approved [9]. Therefore, my results cannot be interpreted as predicting whether a given treatment will be approved, but simply whether a trial will have statistically significant results. Finally, my results depend on the assumption that target enrollment was met or the models are being run after actual enrollment is known. In reality, target enrollment numbers may be missed, leading to errors in one of the driving factors for classification.

Impacts and Applications. Generally, these results are promising in the context of clinical trial design. High recall scores for the classification task imply that the tested models rarely misclassify trials as unsuccessful. If random forest were used in this context, and predicted a given trial as unsuccessful, there is a high probability that the final results will indeed be statistically insignificant. Optimizing with models like these in mind can potentially improve trial design by catching infeasible parameters before they occur.

Additionally, if changes to the parameters are made according to the RF prediction and the classification is still negative, there is a chance the trial is infeasible to begin with, which can halt unsuccessful trials before they ever occur.

7 Conclusion

By considering a scenario where most input parameters can be modified before trial genesis, I explore the inclusion of machine learning in clinical trial design. I consider a more challenging and realistic scenario than previous research by restricting input features to those that can be modified before the trial starts and removing any temporal information.

I also investigate the importance of certain parameters in the prediction of clinical trial success, which can be used to help guide trial designers in the future.

I obtain promising results with four different models, showing the feasibility of a task like this. I also strengthen the arguments for usage of random forests for future ML integration into clinical trial design.

8 Future Work

Although these results are promising, there are still many shortcomings. In the future, I will increase the size and diversity of my dataset while decreasing bias by implementing statistical imputation techniques such as described in [6]. I also wish to integrate inclusion and exclusion criteria into my models, which would mean generating embeddings using models such as transformers. Because p-values are not necessarily representative of FDA approval, I integrate secondary endpoints and other treatments for the same indication into my predicted values; these values are often also taken into account during regulatory review.

References

- [1] BROCKMANN R, NIXON J, L. B., AND I, Y. Impacts of fda approval and medicare restriction on antiamyloid therapies for alzheimer's disease: patient outcomes, healthcare costs, and drug development., Mar. 2023.
- [2] CHEN, T., AND GUESTRIN, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug. 2016), KDD '16, ACM, p. 785–794.
- [3] CHOI, J. W. *Analysis of Clinical Trial Design and Prediction of Success*. PhD thesis, UCL (University College London), 2024.
- [4] IOFFE, S., AND SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.
- [5] KANGAS, A. Analysis of the most important variables impacting clinical trial success rates, 2024.
- [6] LO, A. W., SIAH, K. W., AND WONG, C. H. *Machine learning with statistical imputation for predicting drug approvals*, vol. 60. SSRN, 2019.
- [7] PFLAUM, C. Fda expands approval of gene therapy for patients with duchenne muscular dystrophy [press release], June 2024.
- [8] SEO, S., KIM, Y., HAN, H.-J., SON, W. C., HONG, Z.-Y., SOHN, I., SHIM, J., AND HWANG, C. Predicting successes and failures of clinical trials with outer product-based convolutional neural network. *Frontiers in Pharmacology Volume 12 - 2021* (2021).
- [9] SUN D, GAO W, H. H., AND S, Z. Why 90% of clinical drug development fails and how to improve it?, Feb. 2022.

A Selected Feature Distributions

Shown below are unnormalized selected distributions for features mentioned in table 1.

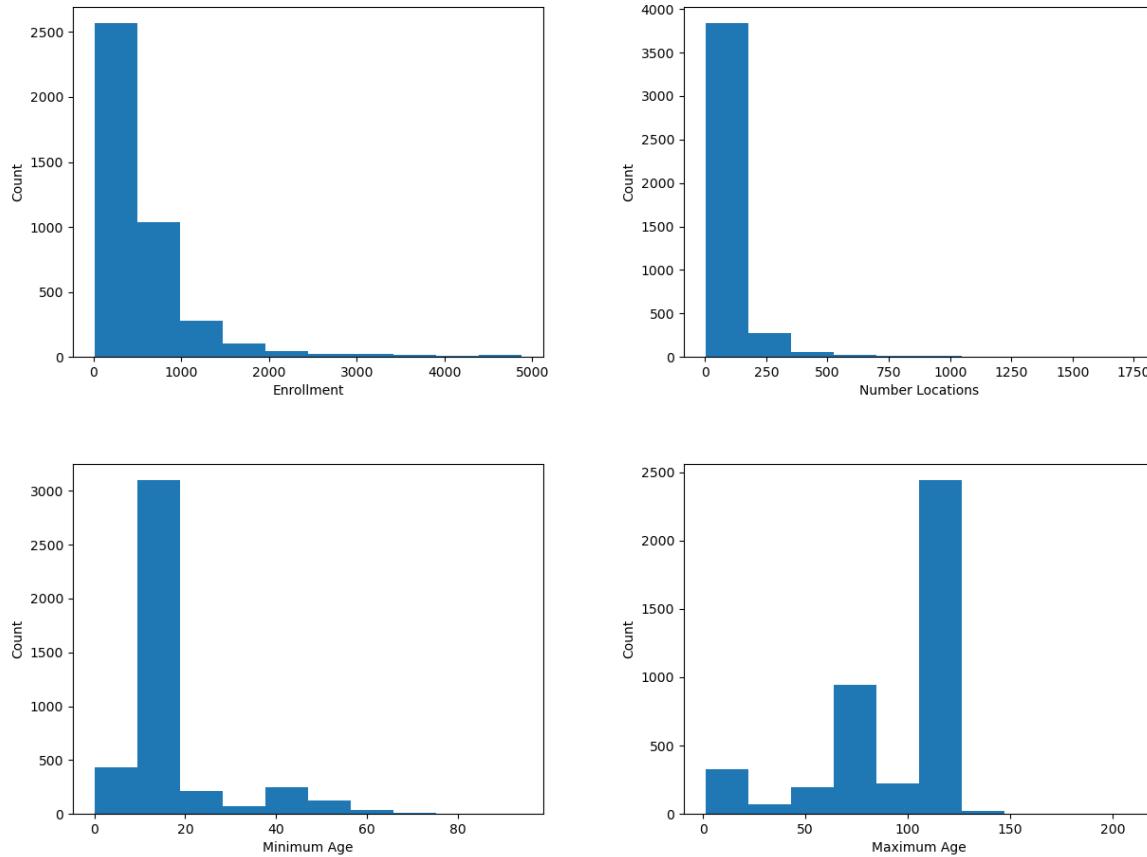


Figure 3: Unnormalized distributions of continuous features.

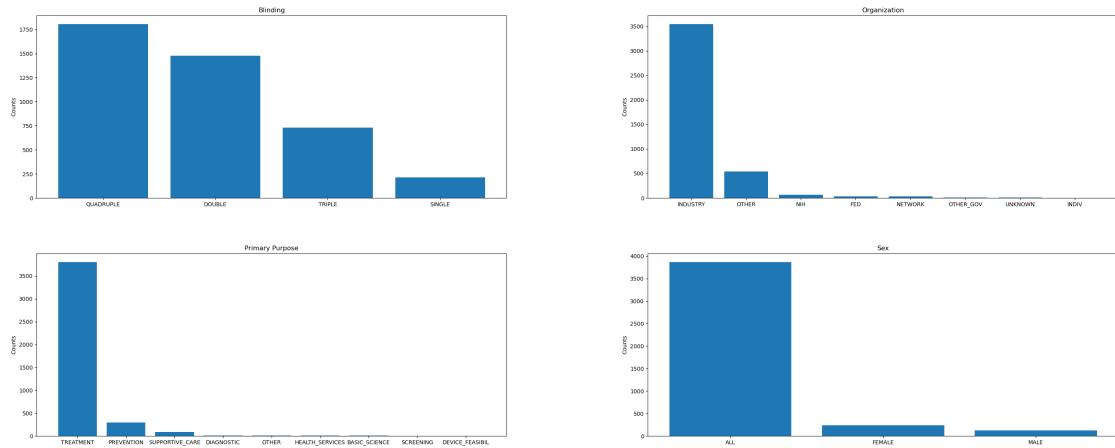


Figure 4: Distributions of categorical features.

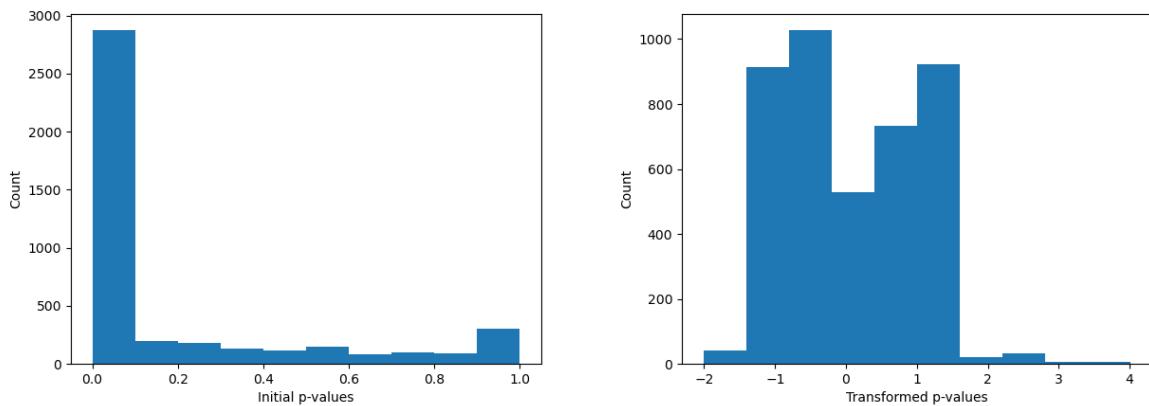


Figure 5: Initial and transformed p-values.

B Actual vs. Predicted p-values

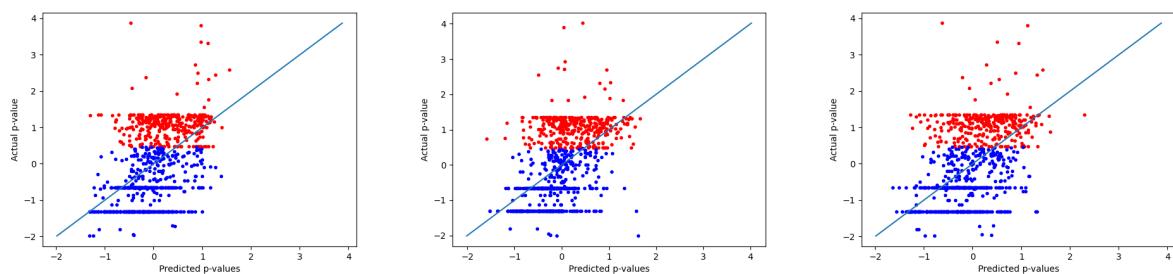


Figure 6: Actual vs. predicted p-values of RF, XGB, and MLP models, respectively, along with the 'perfect fit' line.

C Hyperparameter Tuning

I trained the MLP for classification for 50 epochs with PCA beforehand and 100 epochs without PCA, performing validation every epoch. I selected the model based on the lowest validation loss. I tuned the MLPs with $p_d \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, $d_L \in \{64, 128, 256, 512\}$, and $L \in \{2, 4, 6, 8\}$.

For XGBoost, hyperparameter tuning was performed with $\text{max_depth} \in \{3, 5\}$, $\text{lr} \in \{0.001, 0.01, 0.1\}$, $\alpha \in \{0, 3, 5\}$, and $\lambda \in \{5, 10\}$. Early stopping was used for model selection, with the number of rounds without validation improvement equal to 5.

For random forest, hyperparameters were tuned with $\text{n_estimators} \in \{50, 75, 100, 150\}$ and $\text{max_depth} \in \{3, 5, 7, 9\}$.

D Source Code

All code for data extraction, model implementation, and training/evaluation can be found at <https://github.com/cscl7/Clinical-Trial-Prediction>.