A Machine Learning based Enrollment Rate Forecasting System

Yiqiao Yin

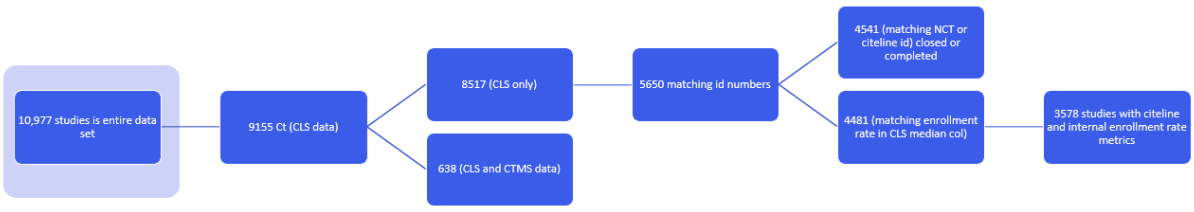
Background

To bring forth unrivaled real-time patient and indication data to develop drug trial studies, important practices include working with contract research organization (CRO) as partners to collaborate on trial studies. There is significant achievement made in recent years to optimize patient recruitment rate for clinical trials. However, due to limited volume of data access, it is suggested that more precise and granular prediction of patient recruitment in the trail design stage is required to reduce the risk of trial recruitment delays and failures (Matthias Briel, 2016; Healy P, 2018). Research have suggested that multiple statistical models can be used to project the enrollment recruitment numbers (Gkioni E, 2019; Bakhshi A, 2013; Gajewski BJ, 2008). One important assumption these past studies have been making is that the enrollment rates are constant over time or even across trial sites which is difficult to hold in practice (Liu J, 2021). For many years, both rule-based systems and machine learning approaches have been proposed (Kang T, 2017; Weng C, 2010). This enrollment rate is crucial for feasibility. In this study, we propose a statistically enhanced, data-driven, machine learning based approach to evaluate and compute the enrollment rate.

Data and Motivation

The study investigates the data from Clinical Trials website and each study has a unique National Clinical Trial (NCT) identified which provides information of different drug trials. In the sample in-house, there are 3,578 unique NCT identifies. The contributing factors include historical enrollment rates, investigator enrollment estimates (site outreach), and additional considerations such as competitive trial and therapeutic landscape, and so on. To assist practitioners to distinguish the difference between operational and labs experience, the enrollment rate is computed for each drug study and an overall recommendation score is computed. The clients observe publications from Clinical Trials, which would produce enrollment rates higher than internal data and cause discrepancy between enrollment rates calculated from LabCorp and that from Citeline data. This paper investigates this discrepancy using machine learning approaches and show that machine learning methods such Random Forests can mitigate this undesirable discrepancy.

Figure 1. In-house data and processing pipeline.



Methods

This section introduces the methods proposed in this paper. The new enrollment rate model we propose takes the following form

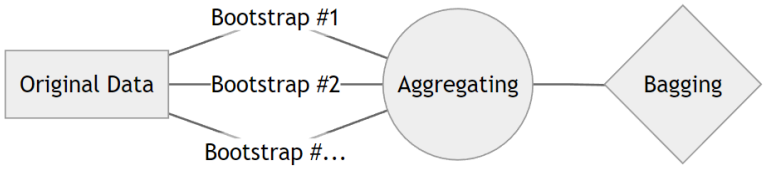
$$Enrollment\ Rate = Patients \frac{1}{Sites} \frac{1}{Months}$$

The formula takes the number of patients and divided by the number of Sites and the number of Months

of enrollment period. In this case, the enrollment period is computed using the date for the Last Patient First Visit from Citeline subtracting the First Patient First Visit from IPST data. The modification considers of the information Citeline has which is what clients observe directly.

This paper examines the following machine learning algorithms: Bagging, Linear Regression, Random Forest, and Multi-layer Perceptron (MLP). Bagging (also known as Bootstrap aggregating) is a type of machine learning algorithms that use ensemble learning to combine the weak learners to form aggregated estimates. The algorithm is presented in Figure 1.

Figure 2. Executive Diagram of Bagging (Bootstrap Aggregating).



Linear Regression is a common technique is statistical machine learning and it provides the fundamental setup of regression problem. The enrollment rate from IPST is considered y while the linear additive model produces an estimate $\hat{y}$ and linear regression, using the OLS assumption, uses a square error to measure the mistakes the estimate makes. The square error takes the form of $\sum(y - \hat{y})^2$ and this loss function can be used as a target to minimize. The optimal parameters can be sought using maximum likelihood estimates or gradient descent.

The next algorithm is Random Forest. Like Bagging, Random Forest also starts the algorithm with sampling of the attributes in the dataset. Every subset of attributes is randomly selected, and a decision tree can be created. Random Forest consists of many decision trees and the final estimate is created using ensemble learning.

The last algorithm selected for the task is a deep neural network. The model is also known as a multi-layer perceptron (MLP). A multi-layer perceptron starts with an input feature set. The input features are generic features from the dataset. The many different layers are artificially designed to create representations of the intrinsic knowledge from the data. The weights (or also known as the parameters) of the neural network or MLP are trained using backward propagation which is a gradient descent like algorithm propagating backward layer-wise.

Results

The correlations are shown in Table 1. In Table 1, a summary of different correlations is presented in different therapeutic areas. Each therapeutic area the new enrollment rate is calculated, and a benchmark correlation is computed using the new enrollment rate and the IPST enrollment rate (see Figure 3). Due to heteroscedasticity, a log-scale correlation adjustment is presented in the column named “Benchmark (log scale)”. The machine learning algorithms used are Bagging, Linear Regression, Random Forest, and Multi-layer Perceptron (MLP). Due to concerns of interpretation from clientele, correlation is chosen as the final evaluation metric though the Pearson correlation only measures linear association between two variables. However, the purpose of the report is to deliver insight to clients and hence the final measure is required to be correlation to suit for the final deliverable.

Figure 3. Bar plots of median enrollment rate calculation using Citeline data, CLS Data, and new enrollment rate.

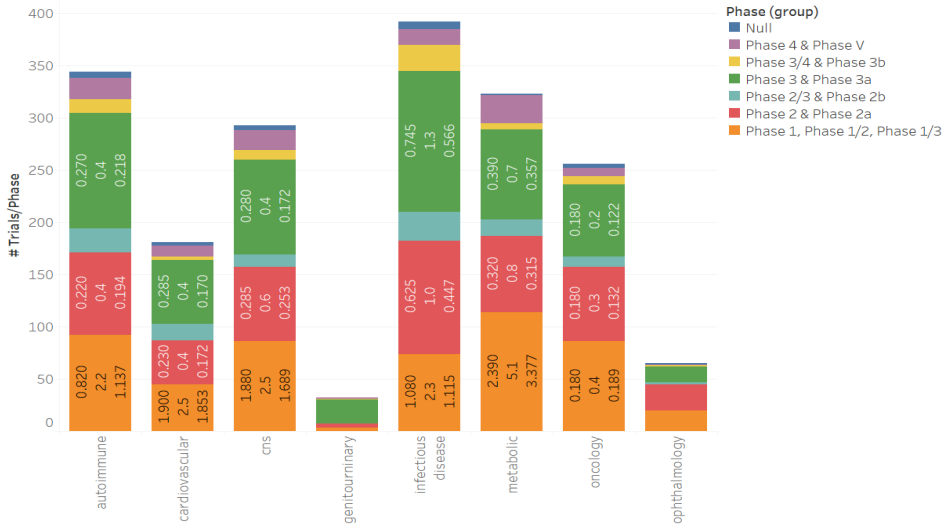


Table 1. Summary of the correlations amongst different methods are presented according to different therapeutic area.

Therapeutic Area	Sample Size	Benchmark	Benchmark (log scale)	Bagging	Linear Regression	Random Forest	MLP
Autoimmune	650	0.213	0.784	0.978	0.928	0.962	0.88
Cardiovascular	229	0.616	0.866	0.874	0.895	0.762	0.894
CNS	484	0.652	0.84	0.667	-0.51	0.712	0.896
Genitourinary	33	0.856	0.878	0.03	-0.675	0.984	0.318
Infectious Disease	507	0.721	0.711	0.951	0.917	0.833	0.659
Metabolic & Endocrinology	441	0.665	0.88	0.877	0.705	0.849	0.871
Oncology	1058	0.214	0.597	0.521	0.615	0.691	0.693
Ophthalmology	67	0.43	0.782	0.982	0.987	0.981	0.938
Average	433.625	0.546	0.792	0.735	0.483	0.847	0.769
SD	312.535	0.221	0.092	0.307	0.633	0.112	0.195

Conclusion

As a summary, this report investigates the discrepancy between the enrollment rates between IPST data and the Citeline data. The report developed machine learning based approach that produced an average of 84% correlation can be achieved by using machine learning algorithm Random Forest to produce and forecast the IPST enrollment rate using modified models from Citeline data, which implies that the Citeline data has information that can assist the prediction task of IPST enrollment rate. This work also calls for further investigation on enrollment rate forecasting on drug level and individual study level to analyze the richness of the IPST data.

References

Bakhshi A, S. S. (2013). Some issues in predicting patient recruitment in multi-centre clinical trials. *Statistics in medicine*, 5458-5468.

Gajewski BJ, S. S. (2008). Predicting accrual in clinical trials with Bayesian posterior predictive distributions. *Statistics in medicine*, 2328-2340.

Gkioni E, R. R. (2019). A systematic review describes models for recruitment prediction at the design stage of a clinical trial. *Journal of clinical epidemiology*, 141-149.

Healy P, G. S. (2018). Identifying trial recruitment uncertainties using a James Lind Alliance priority setting partnership–the PRioRiTy (Prioritising recruitment in randomised trials) study. *Trials*, 1-2.

Kang T, Z. S. (2017). EliIE: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, 1062-71.

Liu J, A. P. (2021). A Machine Learning Approach for Recruitment Prediction in Clinical Trial Design. *arXiv preprint arXiv:2111.07407*.

Matthias Briel, K. K. (2016). A systematic review of discontinued trials suggested that most reasons for recruitment failure were preventable,. *Journal of Clinical Epidemiology*, 8-15.

Weng C, T. S. (2010). Formal representation of eligibility criteria: a literature review. *Journal of biomedical informatics*, 451-467.