A Machine-Learning-Based Enrollment Rate Forecasting System

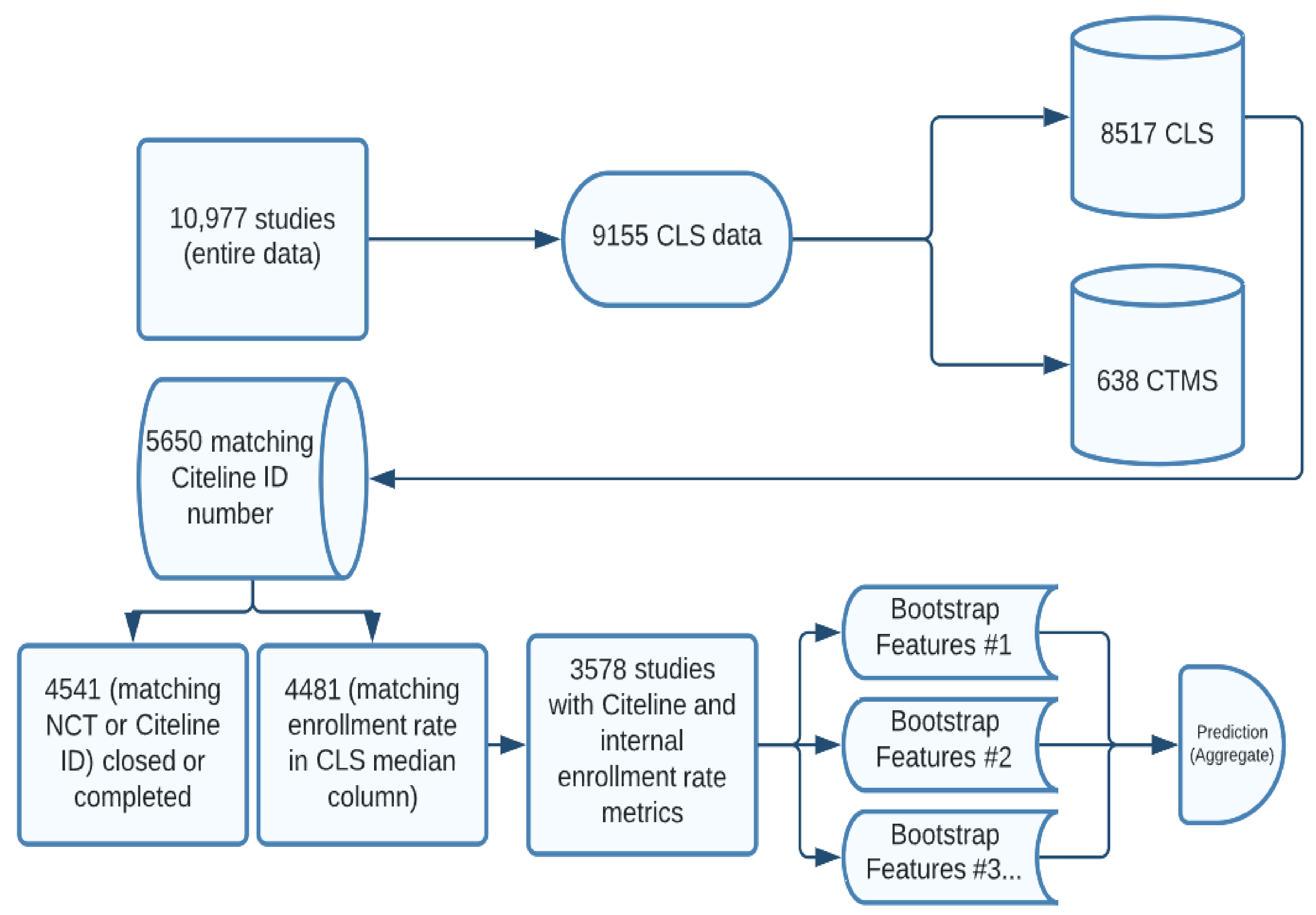
Yiqiao Yin, Sr. Data Scientist, Labcorp Drug Development Inc., Princeton, NJ

Introduction

- To bring forth unrivaled real-time patient and indication data to develop drug trial studies, important practices include working with contract research organizations (CROs) as partners to collaborate on trial studies.
- There have been significant achievements 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 trial design stage is required to reduce the risk of trial recruitment delays and failures (Matthias Briel, 2016; Healy P, 2018).
- Research has 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 that 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.

Methods

- In this study, we propose a statistically enhanced, data-driven, machine-learning-based approach to evaluate and compute the enrollment rate.
- The contributing factors include historical enrollment rates, investigator enrollment estimates (site outreach) and additional considerations such as competitive trial and therapeutic landscape. To assist practitioners in distinguishing the difference between operational and labs experience, the enrollment rate is computed for each drug study and an overall recommendation score is computed.



10,977 studies (entire data)

9155 CLS data

8517 CLS

638 CTMS

5650 matching Citeline ID number

4541 (matching NCT or Citeline ID) closed or completed

4481 (matching enrollment rate in CLS median column)

3578 studies with Citeline and internal enrollment rate metrics

Bootstrap Features #1

Bootstrap Features #2

Bootstrap Features #3...

Prediction (Aggregate)

Figure 1. Diagram of the proposed machine-learning-driven approach. After data is received and processed, there are 3,578 studies with Citeline ID and internal enrollment rate metrics. Different subsets of features are extracted to build bootstrap features. These features are used to build the final prediction of the enrollment rate.

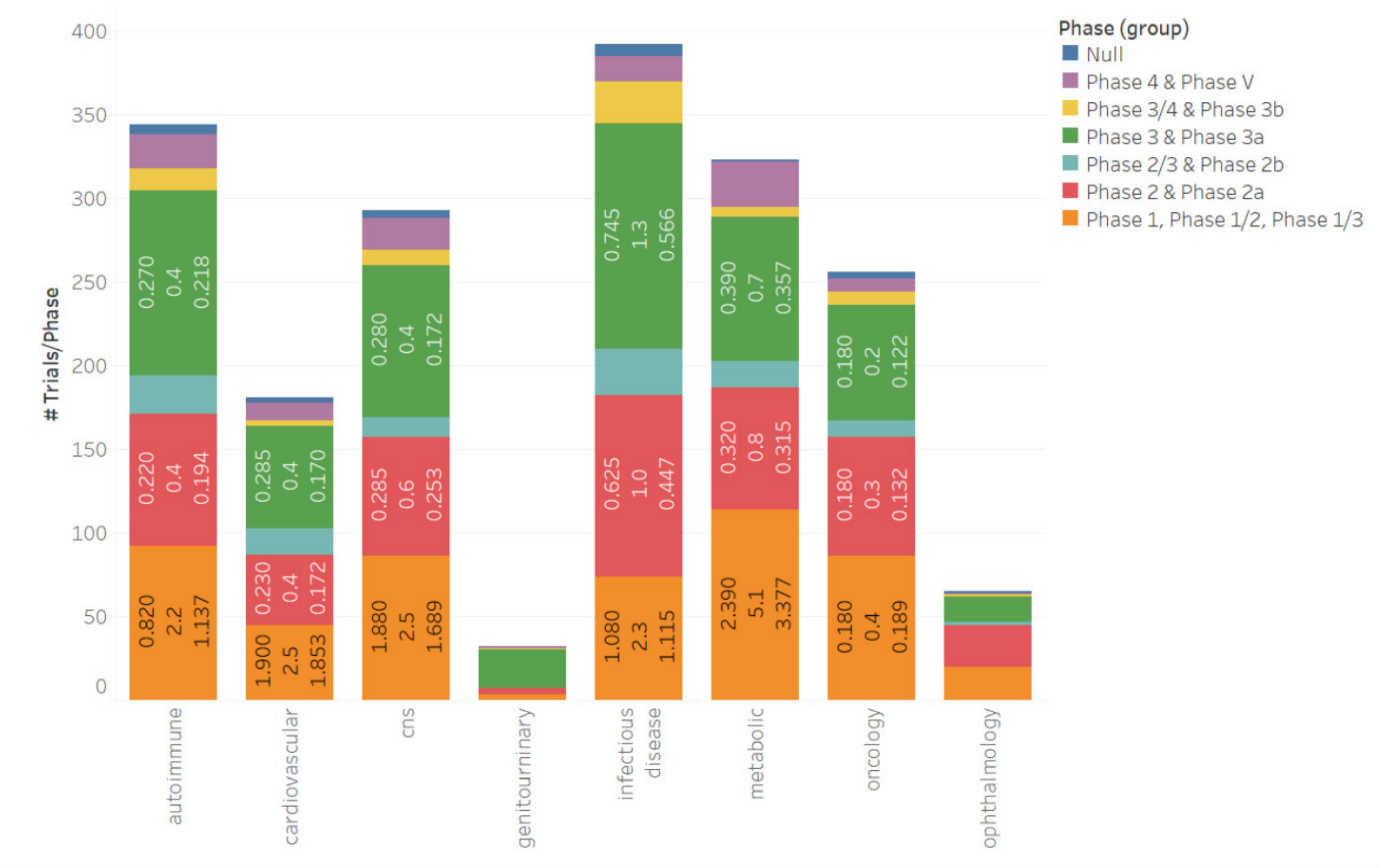


Figure 2. Bar chart of different levels of enrollment rate from the Citeline, internal metrics and machine-learning approach, color coded by different phases. Therapeutic areas are on the horizontal axis and trials/phase are on the vertical axis.

Table 1. Prediction Performance of Out-of-Sample or Test Set Studies

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

Results are presented according to different therapeutic area with sample size, and measured using Pearson correlation for the stakeholders.

Results and Discussion

- In Table 1, a summary of different correlations is presented in different therapeutic areas. For 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 2).
- Due to heteroscedasticity, a log-scale correlation adjustment is used in the training process.
- The machine-learning algorithms used are Bagging, Linear Regression, Random Forest and Multi-Layer Perceptron (MLP).

Presented at OHDSI Symposium 2022

labcorp Drug Development