



OPEN Predicting accrual success for better clinical trial resource allocation

Sisi Ma^{1,2✉}, Yinzhaoh Wang¹, John Wagner², Steve Johnson¹, Serguei Pakhomov³ & Constantin Aliferis^{1,2}

Accrual success is one key determining factor for the success of clinical trials. Global data analyses of all terminated trials reported that 55% of trials were terminated due to low accrual rates. Failure to meet accrual goals have a significant impact on costs for sponsors, academic institutions, investigators, and society at large. The ability to predict trial accrual success with high precision before the trial starts would be highly valuable, preventing the allocation of critical resources for trials unlikely to meet accrual goals. In the present study, we constructed a dataset for predicting clinical trial failure based on poor accrual using clinicaltrial.gov data containing information on 57,846 trials. Features of the dataset were informed by prior literature and constructed using data-driven natural language processing methods. We built predictive models for accrual failure using state-of-the-art supervised machine learning protocols and methods. Models resulted in good predictive performance that was stable over a 10-year time period, with predictive performance of cross-validation AUC = 0.744 (+/- 0.018) and prospective validation AUC = 0.737 (+/- 0.038). We also improved model calibration and examined model performance with the reject option. These modifications enable model translation into decision support tools for various real-world settings. To the best of our knowledge, this is the first study to develop models for predicting clinical trial failure due to accrual based on a large dataset with a comprehensive set of trial features.

Keywords Clinical trial, Clinical trial accrual, Prediction, Decision support

Accrual success is a dominant determining factor for the success of clinical trials (CT). Global data analysis of all terminated trials reported 55% of trials terminated due to low accrual rate¹. An Institute of Medicine report found 71% of phase III trials approved by the National Cancer Institute's (NCI's) Cancer Therapy Evaluation Program closed without meeting their accrual goals². Failed accrual dramatically affects costs for sponsors (Pharma, National Institutes of Health), academic institutions, investigators, and society.

The ability to predict CT accrual success with high precision before the trials start is highly valuable. It allows for early identification of trials that are unlikely to meet accrual goals before allocating critical resources. Moreover, identifying these trials enables more focused and targeted efforts to improve accrual.

While several methods for predicting accrual are available, limitations exist. For example, *Cohort identification tools*^{3–5} forecast “best case scenario” accrual by estimating the number of available subjects that meet inclusion/exclusion criteria using electronic health record (EHR) data. This method estimates the upper limit for accrual but fails to consider factors that limit accrual, such as the probability of consent, the complexity of the trial design, and the effectiveness of recruitment strategy. Existing *accrual forecasting tools* estimate accrual based on an estimated accrual rate (a parameter that is hard to determine before the start of the trial) or a limited number of associated factors^{6–13}. Some commercial vendors^{14–16} have recently offered such tools, but these products often lack sufficient evidence supporting their efficacy. Academic studies are typically much more rigorous, have detailed description of methods and clearly stated statistical assumptions. Many studies address complex clinical trial designs such as the multi-center trials, considers accrual as well as randomization, and allows for adaptive adjustment as the trial progresses. Many of the studies leverages the Bayesian approach for predicting the average and variation of accrual (see Anisimov, 2020¹⁷ for a comprehensive review). Earlier Bayesian approaches model the enrollment process as Poisson processes with accrual rate from a uniform distribution¹². A more sophisticated approach, the Poisson-gamma model, was introduced to account for variations among

¹Institute for Health Informatics, University of Minnesota, Minneapolis, MN 55455, USA. ²Medical School, University of Minnesota, Minneapolis, MN 55455, USA. ³College of Pharmacy, University of Minnesota, Minneapolis, MN 55455, USA. ✉email: sisima@umn.edu

accrual rate in different centers in multi-center trials, where the prior distribution of accrual rate is a gamma distribution^{7,13}. This was further extended to capture the situations where clinical centers can be initiated with random delays and also can be closed earlier⁶ and to model accrual in specific centers, regions, or countries¹⁸. For predicting accrual prior to the trial start, the main limitation of these and related approaches^{6–9} is that they rely on estimated accrual rate (or specifying a prior distribution), which can be difficult to obtain. Additionally, studies that investigate the relationship between accrual and trial-related factors often only examine a small number of trial-related factors. Also, these studies are often descriptive, without assessing the joint predictive capacity of these factors for accrual success^{10,11}. In addition, many studies focus on trials from a specific medical domain where recruitment is particularly challenging, or for specific recruitment methods and scenarios^{19–21}. A third category is the *manual academic administrative review*^{22–24}. This approach uses experienced scientists and research staff to subjectively predict accrual success (i.e., in the context of feasibility assessment). However, it frequently suffers from incomplete knowledge and is susceptible to undesirable variability due to varying composition of committees and other biases. In addition, this method is resource intensive and not scalable.

A large number of barriers and facilitators for CT accrual have already been identified^{1,22–25}, including trial design and complexity, number of eligible and available participants, experience of trial investigators, and characteristics of the institution hosting the trial (e.g. size, location, recruitment performance in prior trials). In principle, a model based on these factors can be constructed for predicting accrual failure. However, only a small fraction of the previously identified barriers and facilitators were examined simultaneously in prior quantitative studies^{6–11,26}, partially because many of these factors are hard to obtain from a large number of CTs. Also, these studies are commonly based on trials that are collected in a single organization, which can result in bias. Moreover, prior studies on accrual failure risk utilized relatively simple statistical methods (e.g. parametric models capturing additive relationships, such as generalized linear models) that may not fully capture the complex relationships among the predictors and the prediction target. Furthermore, these studies focus on identifying the factors associated with accrual, but did not evaluate the predictive capacity of the derived models formally. Therefore, the ability of these models to assess accrual failure risk in the real world remains unclear.

To address the above challenges, we leveraged information reported on clinicaltrials.gov website and external sources, utilized both rule-based and data-driven machine learning natural language processing (NLP) methods to create a dataset for predicting clinical trial accrual failure. Our sample covers CTs across a large variety of disease types conducted in different regions in the US from 1995 to 2022. Importantly, we manually reviewed failed trials reported on clinicaltrials.gov and identified trials that failed due to accrual. These efforts make modeling accrual failure possible. To assess the generalizability of the model over time, we employed a pseudo-prospective validation design, where models were built using historical data and validated on future data. We employed state-of-the-art predictive modeling techniques and protocols to capture complex, non-linear, interactive relationships among a large number of variables available before the trial starts, eliminating the need to estimate accrual rate, and to optimize predictive performance while avoiding overfitting. Further, we explored several methods to enable robust and cost-effective translation of the predictive model into the real world. We employed Markov boundary based feature selection methods which choose a minimum number of features with maximal predictive performance^{27,28}, enabling the cost-effective implementation of decision support tools (DST). We examined and improved model calibration such that the model predicted probability of success better aligns with the actual probability of accrual failure. We also implemented prediction with reject option to improve predictive performance in the predicted cases.

Several recent publications also used clinicaltrials.gov information to construct predictive models for related tasks such as, predicting accrual rate categories (low, median, vs. high)²⁹ and predicting trial termination^{30,31}. The individual accrual rate categories from the study by Bieganek et al.²⁹ were relatively broad, with low accrual rates defined as ≤ 25 participants/year, medium accrual rates defined as $25 < r \leq 100$ participants/year, and high accrual rates defined as > 100 participants/year. This limits the models' applicability in practical settings. The models reported by Elkin et al. and Kavalci et al.^{30,31} predict trial termination due to any reason, not specifically because of accrual failure. To the best of our knowledge, the current work is the first to focus on building a model specifically for predicting the binary outcome of trial failure due to accrual failure, which is defined as accrual that was too slow to meet accrual goals within a defined period. Accurately gauging the risk of trial failure due to accrual failures enables interventions to improve recruitment specifically for the trials that need them the most, ultimately enhancing the probability of trial success.

The accrual models developed here have resulted in good predictive performance that is stable over time, with predictive performance of cross-validation AUC = 0.744+/-0.018 and prospective validation AUC = 0.737+/-0.038. Furthermore, these models are also enhanced by improving calibration and implementing prediction with reject option, which enables model translation into DST suitable for a variety of real-world settings. These models, which use less than fifty easy-to-construct features, demonstrated good predictive performance and calibration.

Method

Goal

The primary goal of our study was to construct and evaluate models for clinical trial accrual failure risk assessment using information available before the start of the trial. A high-level summary of the analytical processes used to achieve this goal is visualized in supplemental Fig. S1.

Data

A total number of 111,494 clinical trials in the format of XML are collected from ClinicalTrial.gov, downloaded on 9/12/2022. We selected trials with four criteria: (1) U.S.-based; (2) interventional clinical studies; (3) the trial was initiated between 1995 and 2022; and (4) The trial recruitment status is either "Completed" or "Terminated",

this removes trials that are on-going, suspended, or withdrawn. The application of these criteria resulted in 57,846 trials in total.

Target of interest

The primary target of interest for this study was whether a trial is terminated due to accrual issues. To construct the target of interest, we first examined the recruitment status field from ClinicalTrials.gov. We started with all trials that have recruitment status of “Completed” and “Terminated”. The trials marked as “Completed” are the trials that ended normally, and participants are no longer being examined or treated (that is, the last participant’s last visit has occurred). None of these trials were terminated due to accrual issues. The trials that were marked as “Terminated” are the trials that have stopped early and will not start again. For the terminated trials, a text field provides unstructured data on reasons for termination provided by the responsible party. We manually reviewed the reason for termination for the remaining 7,965 terminated trials to determine if they were terminated due to accrual issues. One reviewer reviewed and labeled all 7,965 terminated trials, and a second reviewer randomly reviewed 1,000. The agreement between the two reviewers is 100%. The annotated outcomes of the 7,965 trials are described in supplemental Table S1.

We examined additional data fields on ClinicalTrials.gov to further assess if the outcome categories we assigned to the trials were valid. Specifically, we looked at the data fields “actual enrollment” and “estimated enrollment”. The estimated enrollment is the number of participants the trial planned to enroll. The actual enrollment is the actual number of participants enrolled, updated after the trial is completed or terminated. We computed for each trial the ratio of estimated enrollment minus actual enrollment over estimated enrollment (see Fig. 1 for distribution of the ratio for completed trials and terminated trials). For terminated trials, we expect this ratio to be positive. For completed trials, we expect this ratio to be fairly close to zero. There are 4,583 trials out of the 49,881 completed trials with the ratio greater than 0.5. We flagged these trials as abnormal since it is unlikely that a completed trial only recruited less than half of the participants originally planned.

We generated two datasets: one with the flagged abnormal trials included, one without the flagged abnormal trials. Our modeling procedure was conducted on both of the datasets as a sensitivity analysis. The model performance was very similar. We reported the results based on the dataset without the flagged abnormal trials in the main text. This dataset has 53,263 trials, consisting of 45,298 completed trials, 4,986 trials terminated due to non-accrual issues, and 2,979 trials terminated due to accrual issues. And the results on the other dataset in the supplemental file.

Feature construction

We are interested in predicting termination due to accrual failure before the trial starts. Therefore, we used information available before the initiation of the trials for feature construction. Some of the features constructed were informed by prior literature on related tasks^{30,31}. Novel features were also constructed for this study. We categorize the constructed features into two broad categories: (1) Design Features: a small set of features capturing trial characteristics using hand-crafted feature construction methods designed specifically for this study. (2) Text Features: a large set of features capturing information in the text description of the trials using text and natural language processing methods.

The 87 Design Features (Table 1) are constructed to capture factors that have been previously reported to correlate with clinical trial accrual success. All of them are derived from the ClinicalTrials.gov XML data, except for the study population and institutional score. The study population feature represents the size of the population from which participants can be recruited³². The institutional score³³ estimates the research capacity and output of the institution responsible for the clinical trial which was considered to be a facilitator of accrual success^{23,34–36}. Detailed descriptions of the construction methods for these features are presented in our prior study²⁹. Given the reported positive correlation between the complexity of eligibility criteria and risk for accrual failure^{23,25,37}, we included features to capture high-level characteristics of the eligibility criteria of trials, such as average number of words per eligibility criteria, number of eligibility criteria, etc. We adopted the method for constructing these features from a related study on predicting trial termination³⁰. All other design features were directly extracted from structured fields of ClinicalTrials.gov data, capturing information such as study design and study administration.

The 6,085 text features consist of 5,985 Medical Subject Headings terms (MeSH), and 100 word embedding vectors. We included the MeSH terms since they contain information regarding the research topic or target disease of the clinical trial, a factor reported to be related to accrual success in prior literature^{23,25,35,38}. The word embedding features are constructed to capture the information contained in the “Detailed Description” field of ClinicalTrials.gov. The party responsible for a trial is instructed to provide the following information in this field: “Extended description of the protocol, including more technical information, if desired. Do not include the entire protocol; do not duplicate information recorded in other data elements, such as Eligibility Criteria or outcome measures”. NLP methods are ideal for capturing information in this field since it contains unstructured textual data. We used BioWordVec³⁹, a pre-trained Doc2Vec model, to embed the “Detailed Description” of the trial into vectors of length 100. BioWordVec was derived from a biomedical corpus based on PubMed data contains 27,599,238 articles and MeSH terms. Benchmarking indicated that BioWordVec demonstrated superior performance to other embeddings for tasks in the biomedical domain.

Analytical strategy for predictive modeling

Overall design

To evaluate the performance of predicting future accrual failure based on historical data, we implemented a prospective cross-validation design, where clinical trial data from 1995 to 2022 were divided into discovery data and validation data. The discovery data consists of data from clinical trials that occurred earlier in time,

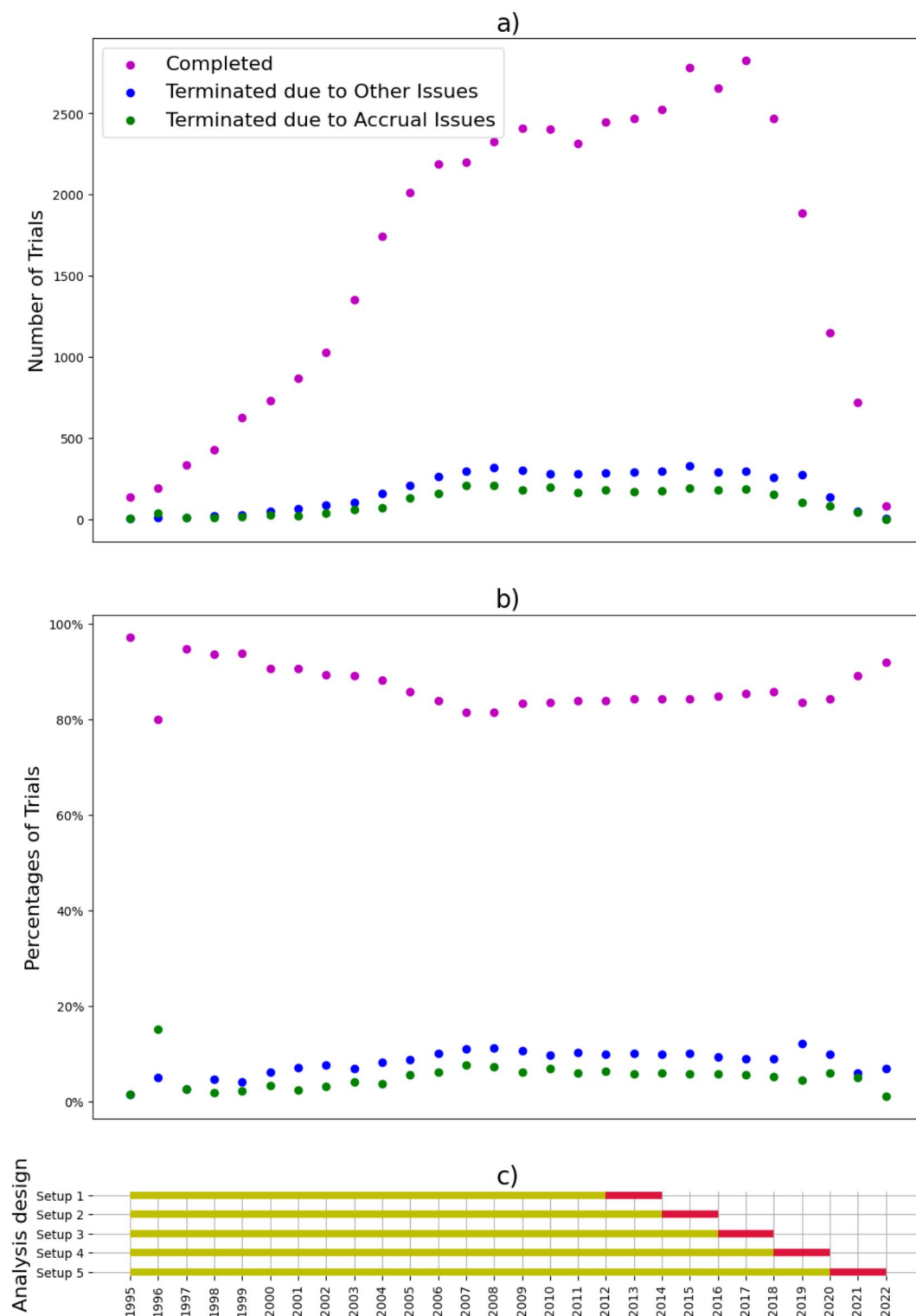


Fig. 1. The number (a) and percentages (b) of the three type of trials over time. To access the generalizability of our models over time, we used a prospective cross-validation design (c) where historical data were used as discovery data (yellow) for model training and selection, whereas future data were used as validation data (red). x-axis represents time.

whereas the validation data consists of data from clinical trials that occurred later in time. To evaluate the model performance over time, we evaluated five pairs of discovery and validation datasets. The five discovery datasets are all trials initiated during and before 2012, 2014, 2016, 2018, and 2020. Their corresponding validation datasets are all trials initiated in 2013–2014, 2015–2016, 2017–2018, 2018–2020, and 2021–2022 (Fig. 1).

	Trial type		
	Completed	Terminated due to other reason	Terminated due to accrual
Study design: randomized allocation			
N/A*	23.3%	35.1%	35.7%
Non-randomized	10.6%	15.1%	11.7%
Randomized*	60.0%	47.6%	51.7%
Missing	6.2%	2.2%	0.9%
Study design: intervention model			
Crossover assignment*	9.2%	4.9%	4.3%
Factorial assignment*	2.1%	1.2%	1.0%
Parallel assignment	49.2%	44.5%	48.5%
Sequential assignment	1.3%	2.5%	1.3%
Single group assignment*	31.6%	44.7%	43.7%
Missing*	6.7%	2.3%	1.2%
Study Design: intervention type			
Behavioral*	19.4%	5.2%	4.2%
Biological	7.2%	8.4%	6.7%
Combination product	0.2%	0.3%	0.2%
Device	9.3%	12.1%	10.9%
Diagnostic test	0.3%	0.4%	0.4%
Dietary supplement	3.2%	1.6%	2.5%
Drug*	45.6%	58.5%	60.0%
Genetic	0.2%	0.3%	0.2%
Other*	9.3%	6.8%	5.7%
Procedure*	4.8%	5.2%	7.1%
Radiation*	0.6%	1.2%	2.1%
Missing	0.01%	0.0%	0.0%
Study design: masking			
None (Open label)*	55.6%	66.6%	68.6%
Single*	0.7%	0.3%	0.1%
Single (Care provider)	0.3%	0.2%	0.1%
Single (Participant)*	4.5%	3.9%	2.7%
Single (Investigator)*	1.5%	0.9%	0.7%
Single (Outcomes assessor)*	5.8%	2.2%	2.8%
Double*	2.4%	1.8%	0.7%
Double (Care provider, Investigator)	0.1%	0.1%	0.1%
Double (Care provider, Outcomes assessor)	0.1%	0.0%	0.1%
Double (Investigator, Outcomes assessor)	0.8%	0.4%	0.5%
Double (Participant, Care provider)	0.7%	0.8%	0.8%
Double (Participant, Investigator)	5.9%	5.5%	5.8%
Double (Participant, Outcomes assessor)	1.7%	1.4%	1.4%
Triple (Care provider, Investigator, Outcomes assessor)	0.2%	0.0%	0.2%
Triple (Participant, Care provider, Investigator)*	2.8%	3.2%	4.4%
Triple (Participant, Care provider, Outcomes assessor)	0.6%	0.5%	0.5%
Triple (Participant, Investigator, Outcomes assessor)*	2.6%	1.9%	1.7%
Quadruple (Participant, Care provider, Investigator, outcomes assessor)	8.3%	8.5%	8.0%
Missing*	5.5%	1.9%	0.8%
Study design: arm group			
Active comparator*	14.6%	13.8%	18.5%
Experimental*	60.4%	68.0%	66.5%
No intervention	2.9%	2.1%	2.3%
Other	4.8%	4.0%	4.4%
Placebo comparator	3.2%	3.0%	3.0%
Sham comparator	0.3%	0.4%	0.3%
Missing	13.8%	9.7%	4.9%
Continued			

	Trial type		
	Completed	Terminated due to other reason	Terminated due to accrual
Study design: devices			
Studies a US FDA-regulated device product	3.8%	5.3%	4.9%
Studies a US FDA-regulated drug product*	10.4%	17.3%	19.0%
If has DMC*	33.2%	42.7%	49.1%
If the Clinical Trial has expanded access	0.2%	0.1%	0.1%
If the trial has DMC or FDA regulation*	42.3%	54.0%	60.8%
If the trial has FDA drug or FDA device regulation, *	13.9%	22.1%	23.3%
Eligibility: Gender			
All	85.9%	87.2%	84.8%
Female	10.0%	9.0%	10.8%
Male	4.1%	3.8%	4.4%
Missing	0.01%	0.0%	0.0%
Eligibility: healthy volunteer			
Accepts healthy volunteer*	30.8%	16.4%	10.5%
No*	69.1%	83.5%	89.4%
Missing	0.2%	0.1%	0.1%
Numerical eligibility features (continuous)			
Average words per eligibility criteria	11.9 (9.5)	12.9 (10.1)	12.6 (9.6)
Average words per inclusion criteria	10.3 (11.3)	12 (12.4)	12 (12.7)
Average words per exclusion criteria	11.4 (10.1)	12.2 (10.6)	11.7 (9.8)
Number of eligibility criteria*	13 (15)	18 (20)	17 (19)
Number of numbers in eligibility*	10 (20)	15 (20)	15 (27)
Number of words in eligibility*	159 (246)	226 (355)	214 (315)
Number of inclusion criteria*	4 (6)	5 (8)	5 (8)
Number of numbers in inclusion criteria*	3 (8)	5 (13)	5 (13)
Number of words in inclusion criteria*	43 (87)	60 (143)	62 (136)
Number of exclusion criteria*	8 (11)	10 (14)	10 (13)
Number of numbers in exclusion criteria*	10 (20)	15 (29)	15 (27)
Number of words in exclusion criteria*	96 (180)	130.5 (230)	120 (200)
Minimum eligibility age	18 (0)	18 (0)	18 (0)
Maximum eligibility age	85 (44)	99 (44)	99 (44)
Study administration: site and investigators			
Number of participated facilities	1 (1)	1 (1)	1 (0)
Population	4,729,484 (9170123)	4,875,390 (11562753)	4,875,390 (7769983)
Institution score	0 (42.88)	0 (23.59)	0 (41.63)
Number of groups	2 (1)	2 (1)	2 (1)
Number of collaborators*	0 (1)	0 (1)	0 (1)
Number of officials	2(1)	2(1)	2(1)
Study administration: responsible party			
Principal investigator*	29.4%	25.1%	32.8%
Sponsor	51.9%	60.1%	55.4%
Sponsor and principle investigator*	3.8%	5.8%	6.8%
Text features: top mesh terms			
Continued			

	Trial type		
	Completed	Terminated due to other reason	Terminated due to accrual
Breast neoplasms*	2.70%	3.60%	4.40%
Leukemia*	2.50%	5.00%	4.80%
Lymphoma*	2.20%	3.90%	4.30%
Syndrome	2.20%	2.90%	2.90%
Depression*	2.40%	1.50%	1.50%

Table 1. Distribution of design features and the top 5 most frequent MeSH terms for completed trials, trials that terminated due to other reasons, and trials that terminated due to accrual reasons. We categorized the hand-crafted design features into 3 categories: study design, eligibility, and study administration. *Indicates a feature is statistically significantly associated with termination due to accrual vs. other trial completion/termination types. For continuous variables, the t-test was used. For categorical variables, the chi-squared test was used.

Classification

We considered the following classification methods: multinomial logistic regression, random forest (mtry = sqrt(number of variables), number of trees = 500), and Adaptive Boosting, i.e. Adaboost (number of trees = 500). We used the multi-class version of these algorithms to categorize the outcome into three categories: completed (class 0), terminated for reasons unrelated to accrual (class 1), and terminated due to accrual failure (class 2). We chose the multi-class setup since it gives better performance compared to directly predicting the binary outcome of if the trial terminated due to accrual in our initial exploration (Supplemental Table S2).

We provide brief descriptions of the three classification method below, and also supplied reference to more detailed discussions of them.

Multinomial logistic regression⁴⁰ is a classification method that generalizes logistic regression to multiclass problems, where the outcome of the prediction is a categorical with more than two classes. The Multinomial logistic regression describe the probability of outcome Y being in class c with a generalized linear combination of the predictors X . $P(Y = c) = \frac{e^{\beta_c X}}{\sum_i e^{\beta_i X}}, i \leq C$, where C is the total number of outcome class.

The random forest⁴¹ is an ensemble learning method. It constructs an ensemble of decision trees and the prediction is made by voting using the output of all the decision trees. Specifically, the random forest is a ensemble $e(X)$ of K decision Trees T_1 to T_K : $e(X) = (T_1(X), T_2(X), \dots, T_K(X))$. The estimated probability of an observation being associated with class c is determined by the proportion of the trees that returns class c as output, i.e., $P(Y = c) = \frac{\sum_i (T_i(X)=c)}{K}$. The random in random forest comes from the fact that it uses bootstrap samples to build each decision tree and randomly select a subset of features when considering a candidate split.

AdaBoost⁴² is also an ensemble learning method where the predictions of many weak learners are combined into a weighted sum that represents the final output of the boosted classifier, i.e. $e(X) = \alpha_1 T_1(X) + \alpha_2 T_2(X) + \alpha_3 T_3(X) \dots + \alpha_K T_K(X)$, where T represent each individual weak learner and α represents their weights. We used decision tree as the weak learner (hence the notation T), since our outcome is categorical. AdaBoost is adaptive in the sense that subsequent weak learners are constructed to focus on those instances misclassified by previous classifiers.

Feature selection

For the feature selection, we use all features, fisher’s test⁴³, and generalized local learning (GLL). The fisher’s test assesses the univariate correlation between the outcome and individual candidate predictors, whereas the GLL assesses the conditional dependence among the outcome and a candidate predictor conditioned on combinations of other candidate predictors. Under broad assumptions, GLL guarantees the selection of the most compact (i.e., minimal) set of variables that contain the maximal information regarding the prediction target. GLL in addition to being theoretically optimal, has also been shown to be highly successful in real world benchmarks and applications, and finally possesses causal interpretability under well-defined conditions^{27,28}. Specifically, we used the GLL variant GLL-PC (K=1, 2, 3).

Model selection and performance estimation

The models were developed on the discovery data. The performance of the models was validated with cross-validation in the discovery data to estimate the model performance on data with a similar distribution as the discovery data. To test the generalization performance of these models to future data, the models were applied to validation data. To select the model that results in the best predictive performances among several model families and tune parameters for each, and obtain unbiased performance estimation on the discovery datasets, we used a five-fold nested-cross-validation procedure (NCV). The inner loop of the NCV is used to select the best classification, feature selection, and their hyperparameter combinations, and the outer loop of the NCV evaluates the performance of the selected models. The nested-cross validation procedures were repeated four times (we refer to each of them as a NCV repeat) to reduce the variation related to random cross-validation splits (see the “Imbalance” subsection below) and splitting the data into five folds. We conducted model selection and

performance estimation based on the model's ability to determine if the trial terminated due to accrual (e.g. AUC for distinguishing class 2 vs. the rest), since this is our outcome of interest. For detailed description of the NCV protocol can be found in here⁴⁴.

Missing values

Treatment for missing values was incorporated into the modeling pipeline so that imputation on the validation data is done according to the distribution of the discovery data. This prevents information leakage (i.e., to ensure that the error estimates are not biased). Median imputation was used for the continuous variables, and missing indicator columns were added to retain the missingness information. For the categorical variables with missing data, we added a “missing” level to the categories for that variable to represent the missingness information. Mesh and embedding variables are free of missingness. Out of 87 design variables the percentage of missingness ranges from 2 to 33%, among which 80 variables do not have any missing value. More than 91% of the variables have missingness less than 2%. The missingness are due to the corresponding values not being reported in clinicaltrials.gov.

Information content analysis

To examine the predictive performance of the two different types of features in the dataset (design features and text features), we trained classifiers on them individually and compared the predictive performances to the model trained with all features.

Imbalance

The proportion of trials that terminated due to accrual failure in our dataset is 2,979, constituting 5.59% of the total number of trials. An imbalance in the proportions of different outcome classes often results in suboptimal performance^{45,46}. Therefore, we explore if subsampling, a common technique to handle imbalanced data, improves performance. We explored four subsampling settings when training the models: (1) C1TO1TA1: sampling equal number of trials that were completed (C), terminated due to other reasons (TO), and terminated due to accrual issues (TA). (2) C2TO1TA1: sampling twice as many completed trials, compared to terminated due to other reasons, and terminated due to accrual issues (3) C5TO1TA1: sampling five times as many completed trials, compared to terminated due to other reasons, and terminated due to accrual issues (4) No subsampling. For the first three subsampling settings, the number of TA trials is the smallest of the three categories, so all of them were always sampled. For C trials and TO trial, a random subset of observations were sampled in each NCV repeat. For the last subsampling setting, i.e. no subsampling, all observations are used. The subsampling was only conducted on the discovery data when training the model. The validation data were not subsampled.

Performance metrics

We used area under the receiver operating characteristic curve (AUC), sensitivity, specificity, precision, and negative predictive value to evaluate the predictive performance of the models. We used the Brier score to evaluate model calibration. The perfect prediction will have AUC of 1, Brier score of 0. Random guess will result in AUC of 0.5 and Brier score of 0.25.

Improving predictive models for translation into decision support tools

Model calibration

The close correspondence between model predicted risk vs. observed risk is important if the model is to be deployed in a real world setting⁴⁷. The deviation of model predicted risk from the actual risk can result in model misinterpretation and misuse. Therefore, we evaluated the calibration of our models with the Brier Score⁴⁸. To improve model calibration, we applied isotonic regression, Platt scaling and spline calibration⁴⁹ to recalibrate models prediction on the discovery data, and evaluated model calibration on the validation data with Brier Score.

Prediction with reject option

Prediction with reject option is a framework aiming to prevent misclassification by not making a prediction for a subset of observations^{50–52}. When the cost of misclassification exceeds that of withholding a decision, prediction with reject option is preferred. In our application, the costs of misclassification include the cost of starting a trial when it would fail accrual and the cost of not starting a trial when it would succeed in accrual. The cost of withholding decision is the additional cost associated with deciding if the trial is to be started, such as, the cost of manual review by a group of experts that will lead to a decision outside the scope of the model. To explore if prediction with reject option improves predictive performance, we implemented a method termed “double threshold”. The intuition underlying this method is that, the model predicted score relates to the confidence of model prediction. And the confidence of model prediction are lower for observations with predicted scores that fall in the midrange of the predicted values. We empirically evaluate this by withholding prediction for the observations with predicted scores in the midrange and examine if the predictive performance improves for the rest of the observations. To achieve this, we introduced two thresholds on the predicted score, such that the trials with scores between these thresholds are classified as undecided. The trials with scores under the lower threshold are predicted to succeed in accrual. And the trials with scores higher than the upper threshold are predicted to fail in accrual. The selection of the thresholds is done by a grid search on the discovery data. The performance of the double threshold is evaluated on the validation data.

Results

Characteristics of the data

Our dataset contains 45,298 completed trials, 4,986 trials terminated due to non-accrual issues, and 2,979 trials terminated due to accrual issues. Figure 1 shows the number and the distribution of the three trial categories: completed (C), terminated due to other reasons (TO), and terminated due to accrual issues (TA). It is notable that the percentages of trials in the three categories changed over time. This could be due to a combination of the following factors: (1) actual percentages of trials in different categories changed over time, (2) changes in the reporting requirements for clinical trials over time^{53,54}, and (3) many of the trials that started in the more recent years are still ongoing, resulting in a bias in the estimated percentages of trials in different categories. For example, out of all the trials that started in 2021, we observed them for less than 2 years (up to the point of our data download on Sep, 2022). Supplemental Table S3 shows the percentages of trials that are completed or terminated till our data collection time out of the total number of trials started in a particular year. Table 1 shows descriptive statistics for key characteristics of the trials in our study.

Predictive modeling results

In this section we present predictive modeling results for predicting trial termination due to accrual. We first present sensitivity analysis for different level of potential noise in the outcome and for different subsampling. We then focused on reporting results for the dataset with low noise level in the outcome without subsampling. We presented results from models using all features, and compared that the models using design features and text feature respectively to assess the information content in different feature types.

Influence of noise in outcome category and subsampling on model performance

As stated in the method section, we faced two choices when deciding what trials to include when training our models. First, whether to include data from trials that potentially contain errors related to the outcome category. Second, whether to use subsampling to address class imbalance.

To assess the influence of potential noise in outcome category on predictive performances, we ran models based on data excluding and including the potentially problematic trials. The average validation AUC over all subsampling schema for the selected models determined by the model selection procedure on the discovery dataset was 0.714+/-0.033 and 0.706+/-0.022 respectively, for models built on data excluding or including the potentially problematic trials.

To assess the influence of subsampling on predictive performances, we applied different ratios of subsampling on the discovery datasets and evaluated the performance on the validation data where the proportion of different outcome categories were unaltered. We observed that when we train our models without subsampling (i.e. preserving the original proportion of outcome where 5.59% of trials terminated due to accrual), the performance on the validation data is on average nominally higher compared to when the three subsampling procedures were applied. Specifically, the average AUC for the model determined by the model selection procedure applied to validation datasets without subsampling is 0.732+/-0.028. Whereas, the average AUC for C1TO1TA1, C2TO1TA1, C5TO1TA1 are 0.693+/-0.032, 0.701+/-0.03, 0.721+/-0.015, respectively.

Given that excluding potentially problematic trials and without subsampling achieved nominally the best results, we focus on reporting results based on these models. The results on the dataset including potentially problematic trials, with subsampling C1TO1TA1, C2TO1TA1 and C5TO1TA1 can be found in supplemental Table S4.

Predictive performance of models built with all features

We first assess what is the best predictive performance can be achieved using all features. Using all 6,172 features, the model selected by the model selection procedure achieved good cross-validation predictive performance in the discovery datasets. The average cross-validation AUC over all discovery datasets are 0.733+/-0.03. The cross-validation performance is stable over the five discovery data sets, indicating consistent performance over time (Fig. 2c). Applying the models derived from the discovery dataset to the validation data resulted in similar performance. The average AUC for the prospective validation datasets are 0.732+/-0.028. The performance on the validation datasets increases over time (Fig. 2c). We hypothesize that this is largely due to the change of trial composition over time. Specifically, the more recent validation data contains more trials with shorter duration. When we applied our model to subsets of trials that completed or terminated within 2,4,6, and 8 years, the predictive performances decreased as the timespan increased, with average AUCs of 0.747, 0.728, 0.718, and 0.704, respectively (details can be found in supplemental Table S5). This result is consistent with our hypothesis.

Information content in models with different feature types

To assess the information contained in the design features and the text features regarding accrual failure, we built models using features from these domains respectively. The performance of the models using the text features resulted in cross-validation AUC = 0.682+/-0.029 and prospective validation AUC = 0.681+/-0.029 over all the discovery datasets. It is significantly worse compared to the models using all features, with cross-validation AUC = 0.733+/-0.029 and prospective validation AUC = 0.732+/-0.028 (cross-validation AUCs: $t = -11.2$, $p < 0.01$; prospective validation AUCs: $t = -2.87$, $p = 0.02$). The performance of the models using design features resulted in cross-validation AUC = 0.744+/-0.018 and prospective validation AUC = 0.737+/-0.038, it is not statistically significantly different from the models using all features (cross-validation AUCs: $t = 0.33$, $p = 0.756$; validation AUCs: $t = -0.19$, $p = 0.851$). As shown in Fig. 2, the cross-validation performance of models using different feature types are also stable over time. These results indicate that the hand-crafted design features representing factors previously reported in the literature contain more information compared to the text features (mesh terms and

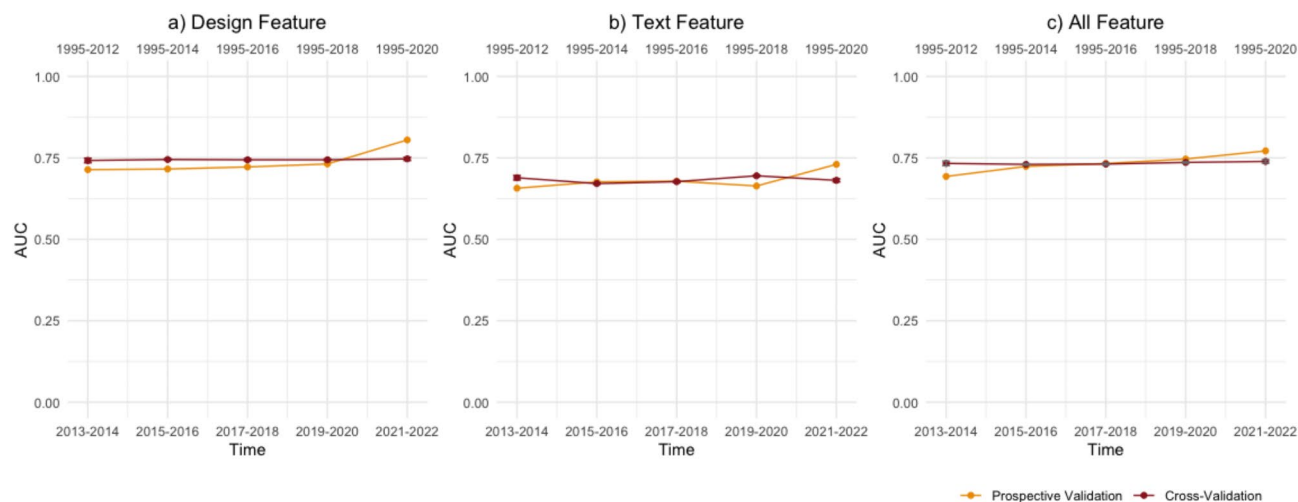


Fig. 2. Predictive performance of Models Built with Different Feature Types: (a) Design Features, (b) Text Features, and (c) All Features (Design + Text Features). The AUC of the model selected by the models selection procedure were estimated with cross-validation on the discovery datasets (yellow lines) and prospectively on the validation datasets (red lines). X-axis tick label on top of the subplots indicate the timespan of the discovery datasets. X-axis tick label on the bottom of the subplots indicate the timespan of the validation datasets.

embedding vectors) we constructed. Further, models using both the design and text features do not result in better performance for accrual failure prediction compared to models using only the design features alone.

Improving models for decision support

In this section we describe several methods we employ to enhance various aspects of model translation, including reducing model size, refining model calibration, and improving predictive performance by introducing prediction with reject option.

Identifying models with a smaller set of input features

The models selected by our models selection procedure for all discovery datasets typically contain all 6,172 features. These models can be converted to a decision support tool. However, the decision support tool requires relatively high resource commitment. It will require all 6,172 features as input, extracted and computed from raw data, adequate computational resources to store and execute the prediction model, and expert monitoring and maintenance. Though missing imputation can be conducted at prediction time if not all features are available, but that may reduce model performance^{55,56}.

To improve the model with the goal of obtaining a cost-effective decision support tool, we examined if there are models with smaller numbers of features that achieve similar predictive performance. As mentioned in the previous section, using the design features achieved predictive performance that is not statistically significantly different from using all features. Therefore, using only design features is one solution to reducing the number of features while retaining model performance. In this section, we explore potential further reductions of the number of features in the model using GLL-PC feature selection. We chose the GLL-PC feature selection since in principle GLL-PC can identify the smallest feature set that preserves the maximal information regarding the target of interest^{27,28}. As shown in Fig. 3b and d, the GLL-PC models applied to all 6,127 features resulted in models with, on average, 718 features, resulting in average cross-validated AUC = 0.722 $\pm$ 0.003 (as compared to cross-validated AUC = 0.733 $\pm$ 0.029 from the model with all 6,127 features) and prospective validation AUC = 0.706 $\pm$ 0.029 (as compared to AUC = 0.732 $\pm$ 0.028 from the model with all 6,127 features). The predictive performance difference between the GLL-PC vs. the full model is not statistically different on both cross-validation set, with $t = -1.52$, $p = 0.138$ and prospective validation set, with $t = 1.9306$, $p = 0.064$. As shown in Fig. 3a and c, The GLL-PC models applied to the design features resulted in models with on average 42 features, resulting in average cross-validated AUC = 0.724 $\pm$ 0.003, (statistically significantly different as compared to cross-validated AUC = 0.744 $\pm$ 0.018 from the model with all 87 design features, $t = -11.476$, $p < 0.01$) and prospective validation AUC = 0.705 $\pm$ 0.029 (not statistically significantly different compared to prospective validation AUC = 0.737 $\pm$ 0.0383 from the model with all 87 design features $t = 1.4972$, $p = 0.1753$). Moreover, the GLL-PC models also demonstrated stable performances over time. Our results suggests that, reduction of the number of features can be achieved by using the GLL-PC feature selection impacting model performance marginally or not at all. Selected AUCROC (area under the receiving operating curve) plots are shown in Fig. 3e and f. The features selected in the most compact models (design features selected by GLL-PC) and feature importance are presented in supplemental Table S6a and b.

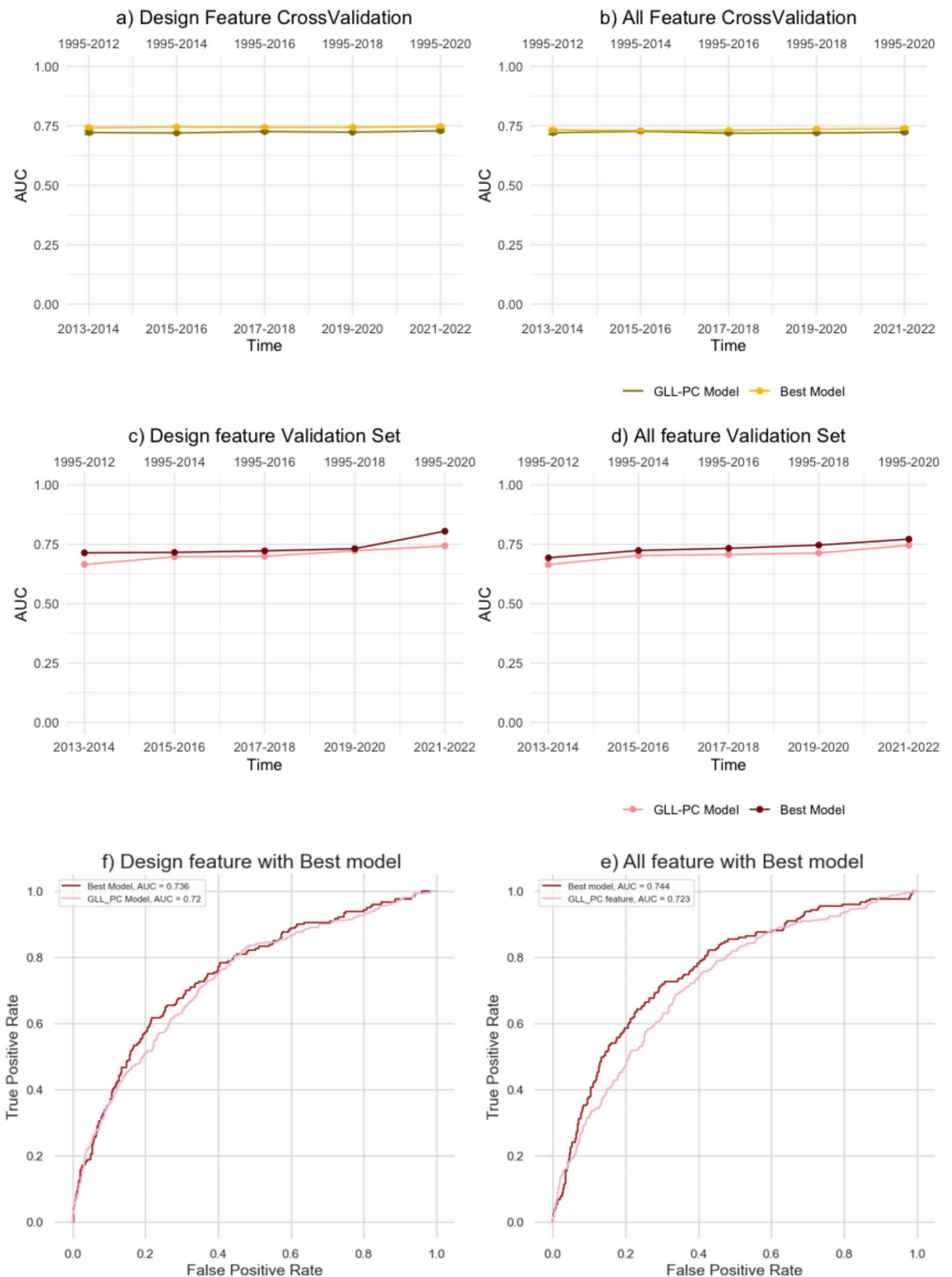


Fig. 3. Models Constructed with the GLL Feature Selection Showed Similar Predictive Performances Using a Smaller Number of Features. This is observed both in the cross-validated performance estimation in the discovery datasets (**a**, **b**) and the prospective validation performances (**c**, **d**). For models using design features (**a**, **c**) and all features (**b**, **d**). For (**a**–**d**), X-axis tick label on the top of the subplots indicate the timespan of the discovery datasets. X-axis tick label on the bottom of the subplots indicate the timespan of the validation datasets. Panel (**e**), (**f**) shows ROC curves for model performance on the validation sets with design features and all features respectively. For legibility, we only show ROC curves for year 2017–2018, ROC curves on the other validation datasets looks similar and are included in supplemental Fig. S2.

Model calibration

Another important consideration for a decision support tool is model calibration, which is how closely the model predicted probability of failure due to accrual aligns with the actual probability⁴⁷. To assess the model calibration, we computed the Brier Score. The average Brier score for the models selected by the model selection procedure is 0.274 ± 0.005 and for GLL-PC is 0.274 ± 0.005 on the validation data. To improve the calibration we applied the isotonic regression method, this resulted in significant improvement in calibration ($p < 0.01$). After model calibration, the average Brier score for the models is 0.068 ± 0.017 and for GLL-PC is 0.071 ± 0.021 on the validation data. We also applied two other calibration methods Platt scaling and spline calibration. The three methods work similarly well, we report isotonic regression in the main result section and the results of other methods in the supplementary Table S7.

Prediction with reject options

To further improve model performance and applicability in real-world decision support settings, we investigated the model performance under learning with reject option (LRO). Specifically, we consider three potential decision-support recommendations given the prediction output of the predictive model: (1) model prediction has low reliability, recommend expert review; (2) model predicts with high confidence for accrual success, recommend proceeding with current accrual plan; (3) model predicts with high confidence for accrual failure, recommend delay the initiation of accrual and explore additional resources to improve accrual. We examined one simple method to categorize model predictions into the above three categories, i.e., implementing two threshold values on model prediction. The predictions that are lower than the lower threshold are considered to be in category (2), the predictions that are higher than the higher threshold are considered to be in category (3), and the predictions that are between the two thresholds are considered to be in category (1). Different values of the thresholds would result in different predictive performances for the trials in (2) and (3), and will also affect the number of trials needing manual review, increasing institutional burden. Therefore, the optimal threshold for different institutions might be different application settings, depending on the expectation of model performance and available resources, i.e. the trade-off between misclassification cost and rejection cost.

We illustrate this method by applying different thresholds to the model using all features, at 20%, 30%, and 40% rejection rate of the total number of trials. In general, we found that as the percentage of rejects increases the predictive performance of the model also increases. The rejection rate of 20%, 30%, and 40% are AUC = 0.730 ± 0.018 , 0.747 ± 0.019 , and 0.759 ± 0.022 , respectively, averaged over the five prospective validation dataset. Among three rejection rates, $40\% \pm 2\%$ showing significant improvement over that without rejection (AUC 0.732 ± 0.028 , $p = 0.0498$). These results indicate that model performance can be further improved by withholding decisions on trials with low prediction reliability. Supplemental Table S8 shows the AUC, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for all prospective validation data and different rejection rates. Performances are stable over time except for year 2021–2022. This is likely related to this validation set is small and has different trial proportions compared to the other validation sets (Fig. 1a and b).

Discussion

The key contributions of this study are threefold. First, we constructed a dataset for predicting clinical trial failure due to accrual based on the *clinicaltrial.gov* data with information for 57,846 trials. We manually annotated the reasons for failure for 7,965 failed trials to construct our outcome of interest. We also extracted and constructed features informed by prior literature and using data-driven NLP methods. This dataset can benefit future studies with similar goals. Secondly, we successfully constructed models for predicting clinical trial accrual failure with good performance that generalizes well to prospective data through a 10-year span. To the best of our knowledge, this is the first study to develop models for predicting clinical trial failure due to accrual based on a large dataset with a comprehensive set of trial features. Thirdly, we demonstrated that enhancements can be made to the models to further improve their performance and applicability in real-world decision support settings.

Several directions for future work can address the limitations of the current study and may result in improved prediction performances. The first direction for future work is to evaluate the models built in the current study in a dataset with percentages of trials failed due to accrual that better approximate that of the real world. Our dataset, extracted from *clinicaltrial.gov*, has an accrual failure rate of 5.59%, which may not reflect the accrual failure rate in the real world due to bias in reporting^{53,54}. We expect the model performance to hold if the trials failed due to accrual reported on *clinicaltrial.gov* was representative of all trials that failed due to accrual. Otherwise, our models are biased due to the bias in the data and may result in reduced model performance when applied in a real-world setting. Further, building models de-novo in a dataset with percentages of failed trials that better approximate the real-world can result in improved predictive performances. Secondly, constructing additional features regarding the trials can potentially improve the predictive performance. Many barriers and facilitators of accrual identified in prior literature were not captured in *clinicaltrial.gov*. Examples include patient compensation^{25,35}, patient burden^{25,35,57–59}, the effectiveness of communication to patients^{23,35,57,58,60–62} and among the trial team^{35,36,57,58,61,63}, concurrent trials competing for participants and the trial team^{25,35}. These data are available and can be constructed from enterprise-level databases such as the clinical trial management systems. We are not aware of an existing dataset that is representative of the real-world accrual failure rate, contains a large variety of trials covering many diseases and geographical areas, and has a comprehensive set of trial characteristics. Constructing such a dataset can greatly enhance the ability to predict accrual failure. Thirdly, our models only flag the trials that are more likely to fail due to accrual, but do not point to interventions that can potentially lead to accrual improvement. The identification of intervention requires the knowledge of causal factors impacting accrual. In general, models and risk factors derived solely for predictive purpose are associative, and are not guaranteed to be causally relevant due to the potential presence of observed and

hidden confounding. Applying computational causal modeling techniques^{64,65} to a dataset that contains a large number of potential remediable causal factors for accrual can reveal trial-specific interventions for improving accrual. Lastly, our study identified models with good predictive performance and a small set of parameters that are cost effective to implement and maintain. In addition, adding the prediction with reject option further enhances the models performance. Our study provides a set of models that can be implemented in the real-world setting, however, the specific model of choice (e.g. percent of reject) depends on several aspects in the application setting that might be intertwined, including what decisions are to be made given the model output, the expected model performance, and the available resource. For example, an institution that have ample existing resource for improving accrual may choose a larger percent of reject value such that more trials will go through an expert review process for potential improvement of accrual. In general, institution-specific information about the cost of the following items can be leveraged to formally guide the choice of models⁴⁴: false positive (model judge the trial to be able meet accrual goal but in fact the trial would not), false negative (model judge the trial to not be able to meet accrual goal but in fact the trial would), expert review for the reject trials, and the institution's goals and budget for clinical trials. The financial implications of model implementation in specific institutions should be evaluated in a case by case manner.

Conclusion

The current study produced predictive models for accrual failure with good predictive performance that is stable over a ten year period. We also identified models that are better suited for translation into a real-world decision support tool, characterized by great calibration, cost-effectiveness for implementation and maintenance, and an option to withhold prediction. This study demonstrated a first step towards a decision support tool for clinical trial resource allocation.

Data availability

The data used in this study can be downloaded from the following urls: <https://clinicaltrials.gov/ct2/resources/download>, https://www2.census.gov/programs-surveys/popest/tables/2010-2018/state/totals/PEP_2018_PEPAN_NRES.zip, <https://www.nature.com/nature-index/institution-outputs/generate/all/global/all>. Derived models in Matlab format will be made available upon request for research purposes.

Received: 21 June 2024; Accepted: 28 January 2025

Published online: 31 January 2025

References

- Desai, M. Recruitment and retention of participants in clinical studies: Critical issues and challenges. *Perspect. Clin. Res.* **11**, 51–53. https://doi.org/10.4103/picr.PICR_6_20 (2020).
- Cheng, S. K., Dietrich, M. S. & Diltz, D. M. A sense of urgency: Evaluating the link between clinical trial development time and the accrual performance of cancer therapy evaluation program (NCI-CTEP) sponsored studies. *Clin. Cancer Res.* **16**, 5557–5563. <https://doi.org/10.1158/1078-0432.CCR-10-0133> (2010).
- Thadani, S. R., Weng, C., Bigger, J. T., Ennever, J. F. & Wajngurt, D. Electronic screening improves efficiency in clinical trial recruitment. *J. Am. Med. Inform. Assoc.* **16**, 869–873. <https://doi.org/10.1197/jamia.M3119> (2009).
- Embi, P. J., Jain, A., Clark, J. & Harris, C. M. Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. In *AMIA Annu Symp Proc*, 2005 231–235 (2005).
- Lai, Y. S. & Afseth, J. D. A review of the impact of utilising electronic medical records for clinical research recruitment. *Clin. Trails* **16**, 194–203. <https://doi.org/10.1177/1740774519829709> (2019).
- Anisimov, V. V. Statistical modeling of clinical trials (recruitment and randomization). *Commun. Stat. Theory Methods* **40**, 3684–3699. <https://doi.org/10.1080/03610926.2011.581189> (2011).
- Anisimov, V. V. & Fedorov, V. V. Modelling, prediction and adaptive adjustment of recruitment in multicentre trials. *Stat. Med.* **26**, 4958–4975. <https://doi.org/10.1002/sim.2956> (2007).
- Barnard, K. D., Dent, L. & Cook, A. A systematic review of models to predict recruitment to multicentre clinical trials. *BMC Med. Res. Methodol.* **10**, 63. <https://doi.org/10.1186/1471-2288-10-63> (2010).
- Gajewski, B. J., Simon, S. D. & Carlson, S. E. Predicting accrual in clinical trials with bayesian posterior predictive distributions. *Stat. Med.* **27**, 2328–2340 (2008).
- Carlisle, B., Kimmelman, J., Ramsay, T. & MacKinnon, N. Unsuccessful trial accrual and human subjects protections: An empirical analysis of recently closed trials. *Clin. Trials* **12**, 77–83. <https://doi.org/10.1177/1740774514558307> (2015).
- Treweek, S. et al. Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis. *BMJ Open* **3**, e002360. <https://doi.org/10.1136/bmjopen-2012-002360> (2013).
- Carter, R. E., Sonne, S. C. & Brady, K. T. Practical considerations for estimating clinical trial accrual periods: Application to a multi-center effectiveness study. *BMC Med. Res. Methodol.* **5**, 11. <https://doi.org/10.1186/1471-2288-5-11> (2005).
- Anisimov, V. V. & Fedorov, V. V. Design of multicentre clinical trials with random enrolment. In *Advances in Statistical Methods for the Health Sciences: Applications to Cancer and AIDS Studies, Genome Sequence Analysis, and Survival Analysis* (eds Auger, J.-L., Balakrishnan, N., Mesbah, M. & Molenberghs, G.) 387–400 (Birkhäuser, 2007). https://doi.org/10.1007/978-0-8176-4542-7_25
- Advarra Case Studies in Accrual Prediction (Advarra Whitepaper, 2021).
- Epic New Life Sciences Program Will Unify Clinical Research with Care Delivery (Epic Website, 2022).
- Cytel Forecast Enrollment Reliably. Cytel Website.
- Anisimov, V. V. Modern analytic techniques for predictive modeling of clinical trial operations. In *Quantitative Methods in Pharmaceutical Research and Development: Concepts and Applications* (eds Marchenko, O. V. & Katenka, N. V.) 361–408 (Springer International Publishing, 2020). https://doi.org/10.1007/978-3-030-48555-9_8.
- Anisimov, V. & Austin, M. Centralized statistical monitoring of clinical trial enrollment performance. *Commun. Stat. Case Stud. Data Anal. Appl.* **6**, 392–410. <https://doi.org/10.1080/23737484.2020.1758240> (2020).
- Unger, J. M., Xiao, H., LeBlanc, M., Hershman, D. L. & Blanke, C. D. Cancer clinical trial participation at the 1-year anniversary of the outbreak of the COVID-19 pandemic. *JAMA Netw. Open.* **4**, e2118433 (2021).
- Watson, N. L., Mull, K. E., Heffner, J. L., McClure, J. B. & Bricker, J. B. Participant recruitment and retention in remote eHealth intervention trials: Methods and lessons learned from a large randomized controlled trial of two web-based smoking interventions. *J. Med. Internet. Res.* **20**, e10351 (2018).

21. Kim, E., Yang, J., Park, S. & Shin, K. Factors affecting success of New Drug Clinical trials. *Ther. Innov. Regul. Sci.* 1–14. <https://doi.org/10.1007/s43441-023-00509-1> (2023).
22. Kost, R. G. et al. Accrual and Recruitment practices at Clinical and Translational Science Award (CTSA) institutions: A call for expectations, expertise, and evaluation. *Acad. Med.* **89**, 1180. <https://doi.org/10.1097/ACM.000000000000308> (2014).
23. Huang, G. D. et al. Clinical trials recruitment planning: A proposed framework from the clinical trials Transformation Initiative. *Contemp. Clin. Trials* **66**, 74–79. <https://doi.org/10.1016/j.cct.2018.01.003> (2018).
24. Brown, R. F. et al. Enhancing decision making about participation in cancer clinical trials: Development of a question prompt list. *Support Care Cancer* **19**, 1227–1238. <https://doi.org/10.1007/s00520-010-0942-6> (2011).
25. Fogel, D. B. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemp. Clin. Trials Commun.* **11**, 156–164. <https://doi.org/10.1016/j.conctc.2018.08.001> (2018).
26. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 (2019).
27. Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S. & Koutsoukos, X. D. Local causal and Markov Blanket Induction for Causal Discovery and feature selection for classification part I: Algorithms and empirical evaluation. *J. Mach. Learn. Res.* **11**, 171–234 (2010).
28. Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S. & Koutsoukos, X. D. Local causal and Markov Blanket Induction for Causal Discovery and feature selection for classification part II: Analysis and extensions. *J. Mach. Learn. Res.* **11**, 235–284 (2010).
29. Bieganek, C., Aliferis, C. & Ma, S. Prediction of clinical trial enrollment rates. *PloS One*. **17**, e0263193 (2022).
30. Elkin, M. E. & Zhu, X. Predictive modeling of clinical trial terminations using feature engineering and embedding learning. *Sci. Rep.* **11**, 3446. <https://doi.org/10.1038/s41598-021-82840-x> (2021).
31. Kavalci, E. & Hartshorn, A. Improving clinical trial design using interpretable machine learning based prediction of early trial termination. *Sci. Rep.* **13**, 121. <https://doi.org/10.1038/s41598-023-27416-7> (2023).
32. U.S. Census Bureau. Metropolitan and micropolitan statistical areas population totals and components of change: 20102019. <https://www.census.gov/data/datasets/time-series/demo/popest/2010s-total-metro-and-micro-statistical-areas.html>.
33. Nature index. <https://www.natureindex.com/faq>.
34. McNair, A. G. et al. Maximising recruitment into randomised controlled trials: The role of multidisciplinary cancer teams. *Eur. J. Cancer* **44**, 2623–2626 (2008).
35. Kaur, G., Smyth, R. L. & Williamson, P. Developing a survey of barriers and facilitators to recruitment in randomized controlled trials. *Trials* **13**, 1–12. <https://doi.org/10.1186/1745-6215-13-218> (2012).
36. Fletcher, G. F. et al. Exercise standards for testing and training: A scientific statement from the American Heart Association. *Circulation* **128**, 873–934. <https://doi.org/10.1161/CIR.0b013e31829b5b44> (2013).
37. Peterson, J. S. et al. Growth in eligibility criteria content and failure to accrue among National Cancer Institute (NCI)-affiliated clinical trials. *Cancer Med.* <https://doi.org/10.1002/cam4.5276> (2022).
38. Tang, C. et al. Clinical trial characteristics and barriers to participant accrual: The MD Anderson Cancer Center experience over 30 years, a Historical Foundation for Trial Improvement. *Clin. Cancer Res.* **23**, 1414–1421. <https://doi.org/10.1158/1078-0432.CCR-16-2439> (2017).
39. Zhang, Y., Chen, Q., Yang, Z., Lin, H. & Lu, Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci. Data*. **6**, 52. <https://doi.org/10.1038/s41597-019-0055-0> (2019).
40. Friedman, J., Hastie, T. & Tibshirani, R. *The Elements of Statistical Learning*. Springer Series in Statistics (2001).
41. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
42. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
43. Fisher, R. *Statistical Methods for Research Workers* (Oliver and Boyd, 1925).
44. Simon, G. J. & Aliferis, C. (eds) *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls* (Springer International Publishing, 2024). <https://doi.org/10.1007/978-3-031-39355-6>
45. Branco, P., Torgo, L. & Ribeiro, R. P. A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv. CSUR* **49**, 1–50 (2016).
46. Hasanin, T., Khoshgoftar, T. M., Leevy, J. L. & Seliya, N. Examining characteristics of predictive models with imbalanced big data. *J. Big Data* **6**, 1–21 (2019).
47. Van Calster, B. et al. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **17**, 230. <https://doi.org/10.1186/s12916-019-1466-7> (2019).
48. Brier, G. W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950).
49. Huang, Y., Li, W., Macheret, F., Gabriel, R. A. & Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J. Am. Med. Inform. Assoc.* **27**, 621–633. <https://doi.org/10.1093/jamia/ocz228> (2020).
50. Bartlett, P. L. & Wegkamp, M. H. Classification with a reject option using a Hinge loss. *J. Mach. Learn. Res.* **9**, 1823–1840 (2008).
51. Saria, S. & Subbaswamy, A. Tutorial: Safe and reliable machine learning. arXiv preprint <https://doi.org/10.48550/arXiv.1904.07204> (2019).
52. Chow, C. On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory* **16**, 41–46 (1970).
53. Tse, T., Fain, K. M. & Zarin, D. A. How to avoid common problems when using ClinicalTrials.gov in research: 10 issues to consider. *BMJ* **361**, k1452. <https://doi.org/10.1136/bmj.k1452> (2018).
54. Mitra-Majumdar, M. & Kesselheim, A. S. Reporting bias in clinical trials: Progress toward transparency and next steps. *PLoS Med.* **19**, e1003894. <https://doi.org/10.1371/journal.pmed.1003894> (2022).
55. Li, J., Wang, M., Steinbach, M. S., Kumar, V. & Simon, G. J. Don't do imputation: Dealing with informative missing values in EHR data analysis. In *2018 IEEE International Conference on Big Knowledge (ICBK)* 415–422. <https://doi.org/10.1109/icbk.2018.00062> (2018).
56. Saar-Tsechansky, M. & Provost, F. Handling missing values when applying classification models. *J. Mach. Learn. Res.* **8**, 1623–1657 (2007).
57. Ross, S. A., Tildesley, H. D. & Ashkenas, J. Barriers to effective insulin treatment: The persistence of poor glycemic control in type 2 diabetes. *Curr. Med. Res. Opin.* **27**(Suppl 3), 13–20. <https://doi.org/10.1185/03007995.2011.621416> (2011).
58. Nipp, R. D., Hong, K. & Paskett, E. D. Overcoming barriers to clinical trial enrollment. *Am. Soc. Clin. Oncol. Educ. Book* 105–114. https://doi.org/10.1200/EDBK_243729 (2019).
59. Subbiah, V. The next generation of evidence-based medicine. *Nat. Med.* **29**, 49–58. <https://doi.org/10.1038/s41591-022-02160-z> (2023).
60. Siembida, E. J. et al. Systematic review of barriers and facilitators to clinical trial enrollment among adolescents and young adults with cancer: Identifying opportunities for intervention. *Cancer* **126**, 949–957. <https://doi.org/10.1002/cncr.32675> (2020).
61. Unger, J. M., Vaidya, R., Hershtman, D. L., Minasian, L. M. & Fleury, M. E. Systematic review and meta-analysis of the magnitude of structural, clinical, and physician and patient barriers to cancer clinical trial participation. *J. Natl. Cancer Inst.* **111**, 245–255. <https://doi.org/10.1093/jnci/djy221> (2019).
62. Pinto, H. A., Mccaskill-Stevens, W., Wolfe, P. & Marcus, A. C. Physician perspectives on increasing minorities in Cancer clinical trials: An Eastern Cooperative Oncology Group (ECOG) Initiative. *Ann. Epidemiol.* **10**, S78–S84. [https://doi.org/10.1016/S1047-2797\(00\)00191-5](https://doi.org/10.1016/S1047-2797(00)00191-5) (2000).

63. Townsley, C. A., Selby, R. & Siu, L. L. Systematic review of barriers to the recruitment of older patients with cancer onto clinical trials. *J. Clin. Oncol.* **23**, 3112–3124. <https://doi.org/10.1200/JCO.2005.00.141> (2005).
64. Pearl, J. Causal inference in statistics: An overview. *Stat. Surv.* **3**, 96–146. <https://doi.org/10.1214/09-SS057> (2009).
65. Kummerfeld, E., Andrews, B. & Ma, S. Foundations of causal ML. In *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls* 197–228 (Springer, 2024).

Acknowledgements

This work is partially supported by Grant UL1TR002494.

Author contributions

Conception of the work: S.M., C.A., J.W., S.J., S.P. Design of the analytical experiments: S.M., Y.W., C.A. Data acquisition, processing, execution of the experiments: Y.W. Manuscript preparation and review: S.M., Y.W., J.W., S.J., S.P., C.A.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-88400-x>.

Correspondence and requests for materials should be addressed to S.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025