The AAPS Journal (2022) 24: 57
https://doi.org/10.1208/s12248-022-00703-3

RESEARCH ARTICLE

Theme: Applications of Machine Learning and AI to Drug Discovery, Development, and Regulations

Machine Learning Prediction of Clinical Trial Operational Efficiency

Kevin Wu[1] · Eric Wu[2] · Michael DAndrea[3] · Nandini Chitale[3] · Melody Lim[3] · Marek Dabrowski[4] · Klaudia Kantor[4] · Hanoor Rangi[5] · Ruishan Liu[2] · Marius Garmhausen[6] · Navdeep Pal[3] · Chris Harbron[7] · Shemra Rizzo[3] · Ryan Copping[3] · James Zou[1,2]

Received: 21 January 2022 / Accepted: 31 March 2022 / Published online: 21 April 2022

Abstract
Clinical trials are the gatekeepers and bottlenecks of progress in medicine. In recent years, they have become increasingly complex and expensive, driven by a growing number of stakeholders requiring more endpoints, more diverse patient populations, and a stringent regulatory environment. Trial designers have historically relied on investigator expertise and legacy norms established within sponsor companies to improve operational efficiency while achieving study goals. As such, data-driven forecasts of operational metrics can be a useful resource for trial design and planning. We develop a machine learning model to predict clinical trial operational efficiency using a novel dataset from Roche containing over 2,000 clinical trials across 20 years and multiple disease areas. The data includes important operational metrics related to patient recruitment and trial duration, as well as a variety of trial features such as the number of procedures, eligibility criteria, and endpoints. Our results demonstrate that operational efficiency can be predicted robustly using trial features, which can provide useful insights to trial designers on the potential impact of their decisions on patient recruitment success and trial duration.

KEY WORDS clinical trials · macxhine learning · operational efficiency

INTRODUCTION

Clinical trials have become significantly more expensive due to their increased complexity (1), with trials now involving more endpoints, procedures, eligibility criteria, countries, sites, and patients than in the previous decade (2). The median cost of a phase III trial is $20 million, with an additional $670K for each month delayed (3). A study of pivotal trials across a decade reports that close to a quarter of studies fail due to cost-related issues (4) Low operational efficiency has contributed to decreasing approval rates for drugs, with only 1–3% of oncology trials reaching the approval stage (5). Therefore, there exists a great need to improve the operational efficiency of trials in order to reduce costs and shorten the lag of improving patient access to novel and innovative treatments. However, the efforts by investigators to curb inefficiencies are hindered by several factors.

One of the main drivers of trial complexity is a growing number of stakeholders (6). For example, payers are increasingly requiring evidence of comparative effectiveness in trials, leading to additional endpoints and procedures (7). A rising number of targeted biomarker-based trials compete for overlapping patient populations, and efforts to increase patient diversity require more focused patient recruitment strategies (8, 9). The regulatory environment has also grown more stringent following

Guest Editors: Lawrence Yu, Hao Zhu and Qi Liu

✉ Kevin Wu
kevinywu@stanford.edu

1 Department of Biomedical Data Science, Stanford University, Stanford, California, USA

2 Department of Electrical Engineering, Stanford University, Stanford, California, USA

3 Genentech, South San Francisco, San Francisco, California, USA

4 Roche Pharmaceuticals, Warsaw, Poland

5 Roche Pharmaceuticals, Mississauga, Canada

6 Roche Pharmaceuticals, Basel, Switzerland

7 Roche Pharmaceuticals, Welwyn Garden City, UK

market withdrawals of high-profile pharmaceutical products (7). Furthermore, there has been an increased need for globalized drug development in order to expand patient access, mitigate against competition, and reach wider patient pools. This has led to multi-region clinical trials that need to satisfy multiple different regulatory requirements and synchronize operations across different regions (10, 11). Therefore, trial designers face the challenge of reducing operational inefficiencies while satisfying multiple competing interests.

Efforts to improve operational efficiency are often based on investigator experience and conventional wisdom (12). The expertise of the trial designer has been shown to be a significant factor in determining trial success (13). Currently, sponsors and investigators rely on subject matter experts to forecast the operational burden of each individual trial. These are mostly bespoke processes that consider largely *ad hoc* and qualitative information about the specific disease, patients, and history of sites for a particular trial. However, as efficiency metrics from clinical trials are not collected and made publicly available in large-scale datasets (14–17), it is difficult to produce systematic forecasts that can be calibrated across a large number of trials.

In this study, we develop machine learning models on data from over 5000 clinical trials by a large, multinational pharmaceutical company. Our models predict multiple efficiency metrics that are typically not systematically collected and made available to the public. While the features gathered in this analysis do not encompass all decisions made by clinical trial teams, our work is a validation of systematically gathering and organizing study features to predict operational efficiency.

MEASURING OPERATIONAL EFFICIENCY

Complex trials often include extensive patient recruitment requirements and protocol-related delays, leading to significant operational inefficiencies. Previous studies have shown that 90% of all clinical trials worldwide have to extend their enrollment period, with average delays of 6 weeks (18, 19). These delays can be very expensive, with case studies reporting costs of nearly $90,000 per patient and screening failures as a principal cost driver (20). Additionally, the rise in protocol procedures and amendments greatly increases site work burden and operational delays. High costs are a major reason why clinical trials fail to move to the next phase (12) and are a barrier to patients receiving the potential benefits of new drugs. In this study, we assess a trial's operational efficiency through various metrics associated with *patient recruitment* and *duration*.

Patient Recruitment

Patient recruitment success is crucial for accessing the appropriate patient population for a study, and insufficient enrollment is a common reason for trials to be suspended or terminated (12, 21). There can be several driving forces behind recruitment failure, including unnecessarily strict eligibility criteria, burdensome visit requirements, and additional investigations beyond endpoint and safety evaluations (1). Additionally, factors such as whether patients are provided with stipends, transportation, and options for remote/local visits which can impact the enrollment duration and dropout rate. Historically, patient recruitment success has been hard to model and understand (21).

We use *screen failure ratio* and *dropout ratio* to measure patient recruitment success (Fig. I). Screen failure ratio is the fraction of screened patients that do not end up enrolled in a trial and is commonly used to measure patient recruitment (15–17). A high screen failure ratio means a trial requires more money and time to acquire its patients. Dropout ratio is the fraction of enrolled patients that do not complete the trial and is an important metric to estimate in the study design phase (22). A patient can be withdrawn from a trial for a variety of reasons, such as adverse events, noncompliance, protocol deviations, and safety. Excessive dropout can lead to costly protocol amendments or underpowered studies (12), affecting the quality of data that can be used to improve patient outcomes.

Trial Duration

Operational delays are significant roadblocks to the overall success of clinical trials (1). Trial length is a key determinant of the financial risk and reward of drug development projects, where overextended trials are at risk of being suspended or terminated (5). Studies have also shown that site activation and regulatory approval alone can take nearly 5 months (23), and complex trials are likely to be delayed (6). Additionally, increasing regulatory burdens impact greater resources and can cause significant delays (24).

We aim to capture operational delays using three metrics: *pre-enrollment duration*, *enrollment duration*, and *study duration* (Fig. I). Pre-enrollment duration is the median number of days per site between site selection and first patient enrolled. This measures the time required to complete organizational prerequisites (e.g., contract negotiation and site training). A lengthy pre-enrollment period can imply a high regulatory and organizational burden. Enrollment duration is the median number of days per site between enrolling the first patient and the last patient,

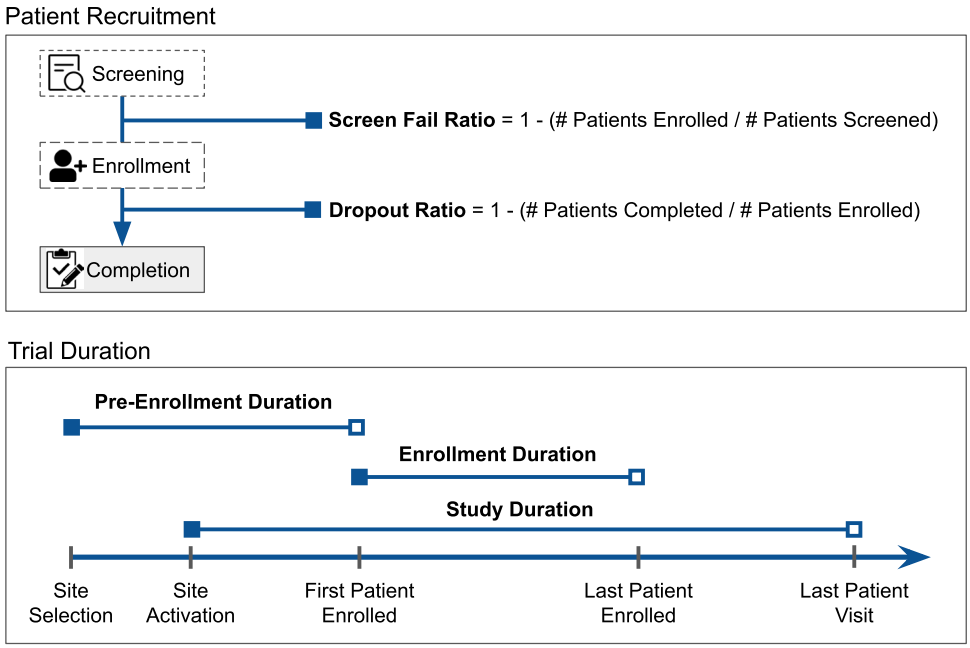

Fig. 1 Patient recruitment metrics displayed across the patient funnel from screening to completion and trial duration metrics across an abridged timeline of clinical trials. The timeline presented applies to a single site, and these events can be asynchronous between sites

across sites. Extended site enrollment delays can be due to unnecessarily stringent eligibility criteria and screening protocols, rare patient populations, and competing clinical trials. Study duration refers to the median days per site between site activation and last patient visit, and captures the end-to-end time required for a study to complete, across sites.

RELATED WORKS

There exists extensive literature studying the growing complexity of clinical trials (25–29), including patient recruitment and trial duration (1, 6, 12, 15–18, 21, 22, 24, 30–33). (12) provides a systematic overview of how operational inefficiencies impact the likelihood of overall trial success.

(34) shows that more complex trials with more procedures perform worse in patient recruitment and retention compared to low complexity trials. (35) finds that complex trials discourage trial participants, while (36) finds that trials with more eligibility criteria tend to be more prone to delays. Additionally, (37) finds that patient dropout is much higher in more complex trials. Our work aims to unify these analyses by modeling a variety of trial features on patient recruitment and trial durations. In doing so, we can also estimate the collective impact of multiple trial features on trial efficiency.

Prior work in applying machine learning methods to improving clinical trial efficiency includes natural language processing (NLP) methods for patient recruitment as well as extracting structured data from eligibility criteria (38, 39). Machine learning has been applied to clinical trial data for purposes of predicting the overall likelihood of approvals (5, 40, 41). Our work differs from prior work in that we focus on specific efficiency metrics, a level of granularity that is of particular interest to trial designers.

DATA

In total, the dataset contains 2051 completed clinical trials conducted by Roche with starting dates from 2009 to 2020. Due to differing levels of missingness across efficiency metrics, there are $N = 1922$, 1395, 932, 526, and 361 trials included in our analyses for enrollment duration, screen fail ratio, pre-enrollment duration, study duration, and withdrawal ratio. The varying levels of missingness are a result of differences in the data collection pipeline. For example, enrollment duration can be more easily estimated with start and end times, while withdrawal ratios require follow-up reports on each patient. We include a total of 23 operational features. The features and their detailed descriptions can be found in Supp. Table I. Among the full list of features included are the study phase, therapeutic area, experimental design, number of endpoints, number of eligibility criteria, and details about the planned procedures involved. In the data, there are 288 unique drugs and 219 unique indications represented across all trials and an average of 11.4 inclusion criteria, 15.3 exclusion criteria, and 3.9 countries per trial.

Table I Model Performance Results

	Overall	Therapeutic area (C-index)				Study phase (C-index)			
Efficiency metric	C-index	I2O	Neuroscience	Oncology	Other	I	II	III	IV
Screen failure ratio	0.801	0.795	0.765	0.789	0.808	0.622	0.788	0.802	0.771
Dropout ratio	0.791	0.750	0.651	0.715	1.000	0.784	0.801	0.804	0.771
Pre-enrollment duration	0.705	0.724	0.635	0.611	0.687	0.675	0.565	0.587	0.597
Enrollment duration	0.706	0.680	0.709	0.683	0.672	0.764	0.692	0.647	0.609
Trial duration	0.728	0.644	0.766	0.624	0.756	0.808	0.656	0.610	0.666
Average	0.746	0.719	0.705	0.684	0.784	0.731	0.700	0.690	0.683

Main results reporting c-index across all trials and stratified by therapeutic area and study phases. The therapeutic area "I2O" is an abbreviation for immunology, infectious diseases, and ophthalmology. The average across metrics is reported at the bottom of each column

METHODS

Data Preprocessing

Categorical features such as drug names and indications are encoded using a one-hot encoding scheme. To handle the heavy tail-end of features, we group features that appear in less than 1% of the data into a single category labeled "other." Missing values are imputed using the mean. Due to noise in data collection, trials with response variables outside two standard deviations are considered outliers and not included in the dataset.

Model Design

Given the different degrees of missingness across all 5 response variables, we train a separate model for each one. We use LightGBM (42), an efficient implementation of gradient boosted tree algorithms, due to its relative robustness of feature processing and strong performance on tabular data.

In the context of providing decision support for study design, we also seek to quantify the uncertainty around our model predictions. Not only are upper and lower bounds useful to understand the confidence of our estimates, but they can also be useful for decision-making. For example, an upper bound or "worst-case" prediction of trial duration can be more useful than a point estimate for resource planning purposes. To produce these predictive intervals, we use a quantile loss function for our gradient boosted trees (Equation 1), trained at quantiles 0.05 and 0.95 to achieve a 90% predictive interval. For point estimates, we use the quantile 0.5, which is equivalent to the median. Compared to mean squared error, which computes the conditional mean of the response variable, a model trained on quantile loss at 0.5 computes the conditional median. Similarly, a model trained at the quantile 0.05 should, in expectation, have residuals that are positive 95% of the time and negative 5% of the time.

$$\rho_\tau(y, \hat{y}) = (\hat{y} - y)(\mathbb{1}\{y \leq \hat{y}\} - \tau) \quad (1)$$

Our dataset is divided into training, test, and validation (hyperparameter tuning) sets using a random 60/30/10% split. Results are reported from our test set. We perform a grid search over a set of hyperparameters (number of leaves, minimum data in each leaf, maximum depth, maximum bins, and learning rate) and monitor performance on the validation set before setting the model for each response variable. Additionally, we perform a search over algorithm selection and hyperparameter tuning using AutoSklearn (43), a Python package that utilizes Bayesian optimization, meta-learning, and ensemble construction to optimize model performance. We test this method on predicting screen failure ratio and find that the top-performing model found does not outperform LightGBM after running AutoSklearn. As such, we find evidence that our model approach is well-optimized across a range of algorithms and hyperparameters.

Evaluation

To provide a uniform evaluation metric across different regression tasks, we use the c-index (44), which is a generalization of the area under curve (AUC) to continuous response variables. The c-index is defined as the proportion of concordant pairs among all evaluation pairs in the test set. A c-index value of 1 means that for any pair of trials, the model predicts a higher response variable for the trial with the actual higher response variable. Conversely, a c-index of 0.5 implies that the model does not perform better than chance at correctly assigning correctly ordered values. For example, when interpreting the c-index of screen failure ratio, two trials with actual values of 0.75 and 0.90 could have predicted values of 0.60 and 0.80 and be counted as a concordant pair. In practical terms, one can interpret the c-index as a measure of confidence that the model will correctly predict the direction of change in operational efficiency based on a set of trial features. In addition to the c-index, we report the *R*-squared score and mean absolute error (MAE) in Table IV.

RESULTS

Main Results

We report the performance of our models on each efficiency metric over all trials in the test set, as well as subsets corresponding to therapeutic area and study phase (Table I and Supp. Table II). Trial features can explain a substantial proportion of the variance in efficiency in total, while study duration is harder to predict on average compared to patient recruitment (Table IV). The models had excellent performance in predicting patient recruitment, with c-index values around 0.80 for both metrics in this category. Models that predicted trial duration response metrics also performed relatively well with c-index values of around 0.70 in all three metrics in this category. In addition to the point predictions, our model uses quantile regression to provide the 90% predictive interval for each prediction, which is empirically well-calibrated (Supp. Table III, Supp. Fig. I). The performance results are consistent across therapeutic areas and study phases.

Validation on Unseen Drugs

Clinical trials that study the same drug can potentially be very similar in their levels of operational efficiency, which may lead our models to overfit when trials on the same drug appear in both the training and test set. We seek to measure the effect of this by splitting the training and test sets based on randomly selected sets of drugs. Out of 288 total unique drugs in our full dataset, we train the model on 209 unique drugs and test the model on 79 different unique drugs which are chosen at random. Within the training drug set, we perform an additional training/test/validation split and report the results on the test split. The results are reported in Table II and show that our model performs worse slightly without knowledge of the drug being evaluated, revealing that knowledge of operation efficiency of a prior drug trial can help in predicting the efficiency of future trials involving the same drug. At the same time, the difference is not large, meaning that even without this knowledge, our model still performs reasonably well.

Table II Validation on Unseen Drugs Across Five Efficiency Metrics

Validation on unseen Roche drugs (C-index)	Training drug set (N = 339)	Testing drug set (N = 359)
Screen failure ratio	0.781	0.712
Dropout ratio	0.757	0.738
Pre-enrollment delay	0.674	0.634
Enrollment duration	0.673	0.665
Trial duration	0.699	0.679
Average across metrics	0.717	0.686

Table III Validation of Trials from Two Time Periods Across Five Efficiency Metrics

Validation across time (C-index)	Trials completed 2009–2012 (N = 439)	Trials completed 2012–2020 (N = 376)
Screen failure ratio	0.742	0.726
Dropout ratio	0.630	0.682
Pre-enrollment delay	0.673	0.680
Enrollment duration	0.711	0.669
Study duration	0.704	0.717
Average	0.692	0.695

Validation on Temporally Separate Trials

In deployment, these models would be run on trials occurring after the trials used in the training data. In order to evaluate time-specific biases, we divide our data into two time periods such that the sample sizes from each time period are roughly equal. We train our models on trials from the first period (2009–2012) and evaluate on trials beyond this period (2012 onward) (Table III). For the trials from 2009 to 2012, we split our training data into training/test/validation sets and report the results on the test set. Similarly, while we observe lower overall performance due to a smaller training set size, we do not observe large differences in performance when our models are evaluated on trials occurring after the trials they were trained on.

Correlation Between Actionable Trial Features and Operational Efficiency

We perform additional analyses to quantify how actionable features of trial design correlate with the five response metrics of operational efficiency. In particular, we only report the effects of features that can be changed during trial design

Table IV R-Squared Values and Mean Absolute Error from Our Model, Across 5 Efficiency Metrics

Efficiency metric	R-squared	Mean absolute error
Screen failure ratio	0.463	0.097
Dropout ratio	0.513	0.179
Pre-enrollment duration	0.319	60.0
Enrollment duration	0.26	245
Study duration	0.32	405
Average	0.375	-

(i.e., the number of eligibility criteria, endpoints, countries, and procedures), as opposed to fixed features such as a trial's phase, drug, and therapeutic area. Because it is challenging to interpret the nonlinear relationships captured by our GBM, we fit a separate multivariate regression model to all trial features and directly assess the association between these features and the trial's operational outcomes. Supp. Table IV reports the total list of actionable features, their respective coefficients, and the p values.

First, we find that the number of countries is associated with efficiency metrics in several ways. A larger number of planned countries are connected with longer pre-enrollment duration ($p = 0.01$) and longer study duration ($p = 0.005$), possibly reflecting the underlying difficulties of satisfying multiple regulatory requirements inherent in multi-regional clinical trials. Additionally, a larger number of countries are connected to higher screening failure ratios ($p <$ 1e-3), which may reflect the tendency for studies with more complex patient recruitment requirements to extend across multiple countries. On the other hand, as the number of countries increases, we observe shorter site-specific enrollment durations ($p <$ 1e-3) at the cost of longer total durations across sites. Specifically, for each country added, the linear model expects a shortened enrollment duration by almost 8 days while simultaneously making overall study duration over 10 days longer.

Second, having more primary and secondary endpoints, which could suggest that the trial is more complex, is associated with higher screen fail ratio ($p = 0.006$ and $p <$ 1e-3 for primary and secondary endpoints). Third, a higher number of planned patient visits are correlated with increased dropout ratio ($p <$ 1e-3) and longer study duration ($p <$ 1e-3), possibly due to higher burdens on patients and site investigators. Each additional examination procedure is associated with more than one percentage point increase in dropout ratio.

Finally, as expected, having higher planned patient enrollment is correlated with longer enrollment duration ($p = 0.003$), with an average of 1 extra day per 60 patients added. Additionally, in Fig. II, we use accumulated local effects (ALE) to directly visualize the effects of a subset of features on our GBM predictions (45). ALE is a method of understanding the average influence certain features have on a model's predictions that is more robust given highly correlated covariates. We find the interpretations from ALE are consistent with our interpretation of the linear coefficients. Furthermore, we provide the importance of a subset of actionable features used in our GBM, defined as the information gained from using the feature in the model (Fig. III). Interestingly, we find that planned patient enrollment and the number of eligibility criteria are among the most important features, even though they are not significant in our linear model in most metrics. This suggests a strong nonlinear relationship, affecting operational efficiency through their interactions with other features. We also report that the number of planned visits has a disproportionate effect on dropout ratio as compared to other metrics, reflecting the impact of the undue patient burden on patient retention.

DISCUSSION

Our results suggest that trial features can be relatively robust predictors of a trial's operational efficiency. We support our results through analyses by drug names, time periods, and sponsor companies. The results are also consistent across therapeutic areas and study phases. Correlations found in the data generally support findings in previous studies that complex trials perform worse in patient recruitment and trial durations (34). However, this large-scale analysis of trials shows that this relationship is not straightforward, as some types of trial complexity are correlated with improvements in operational efficiency.

We find models that predict patient recruitment metrics using trial features have higher performance than models that predict trial duration metrics. Additionally, more fine-grained time metrics like pre-enrollment duration can be subject to regulatory burdens that are out of the scope of trial design. As a whole, operational efficiency is multifaceted and must be evaluated in the context of the interaction of multiple trial features. In practice, investigators and sponsors face tradeoffs between operational and scientific efficiency (e.g., smaller studies may be shorter but have less statistical power). Additionally, trial features do not have strictly linear relationships with efficiency — adding or subtracting certain features does not lead to efficiency improvements across the board. Furthermore, our count-based features for eligibility criteria and procedures do not take into account the specific content but instead capture a relative scale of complexity. Nonetheless, our analysis underscores the importance of a data-driven approach to modeling and understanding clinical trial design, as well as the large-scale curation and collection of efficiency metrics for clinical trials.

Our model provides a proof-of-concept for using machine learning to forecast trial operational metrics such as study duration and screen fail ratio. Such predictions can provide trial designers and companies with additional information while planning and designing trials. In improving the operational efficiency of a trial, care should be taken not to impact the scientific utility of the trial or affect patient safety. A priority of the trial design is to deliver robust answers to the scientific questions it is addressing and to generate important data for key stakeholders. Any actions taken to increase the operational efficiency should be taken in the context of ensuring the scientific objectives of the trial can still be met.

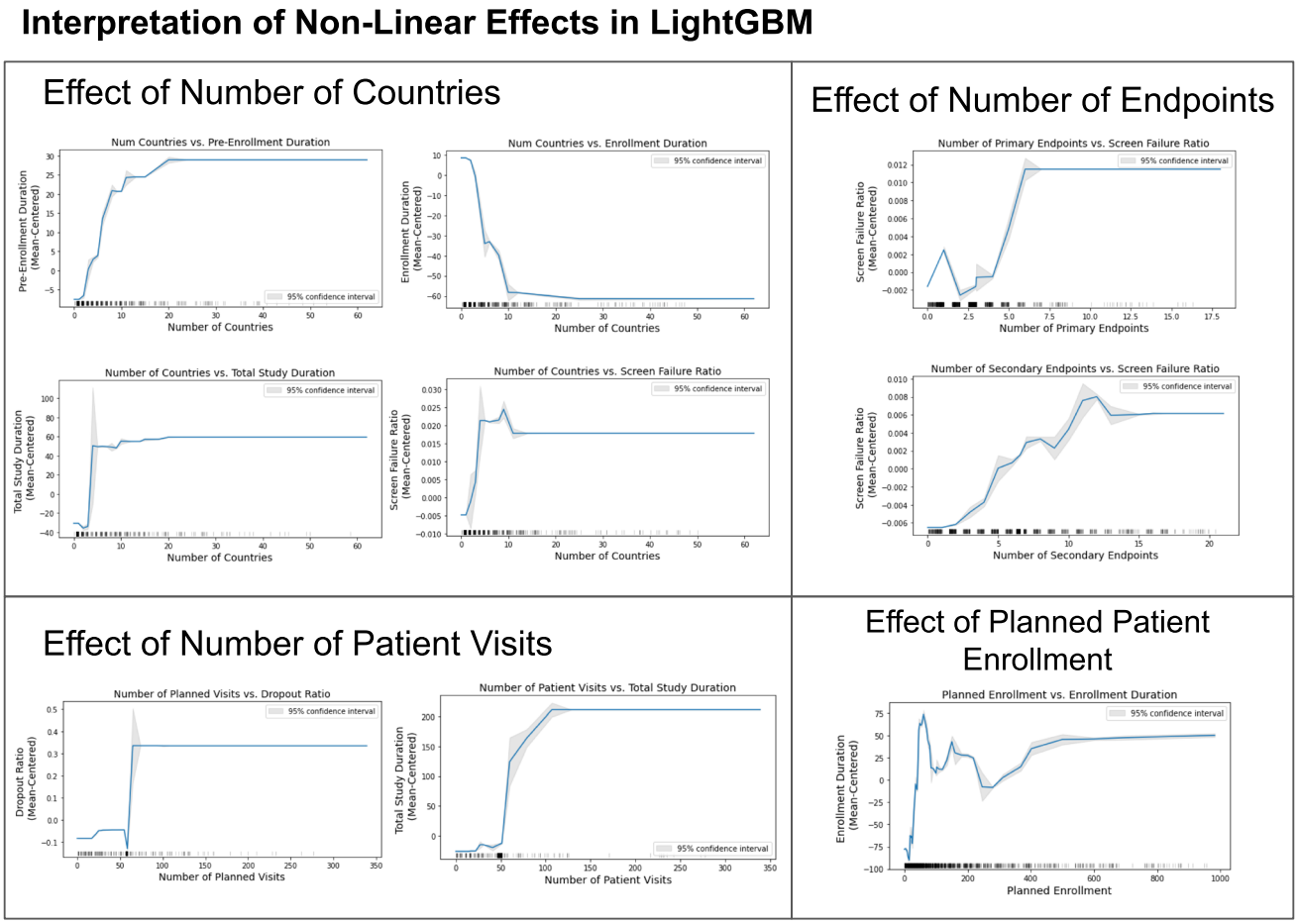

Fig. II We directly analyze the effects of a subset of actionable features in our LightGBM model using accumulated local effect (ALE) plots, which show the average effect of features on the prediction of a machine learning model. The *x*-axis represents the observed values of a feature, while the *y*-axis represents the effect of that feature compared to an average prediction. For example, in the top left plot, an ALE estimate of 20 days when the number of countries is 10 means that the model predicts the pre-enrollment duration to be 20 days more than the average prediction when there are 10 countries. The 95% confidence intervals are provided as the gray area around the blue lines. The black lines on the *x*-axis represent a rug plot that is denser around values that are more represented in the dataset. These directions of the influence found in the plots are in concordance with our simpler linear model

Computational predictions and models like ours should not be used in isolation but rather in combination with interactions with patient advocacy groups, investigators, payers, and health authorities.

As with most machine learning and forecasting models, the predictions are made based on correlations rather than causal relationships. For example, our model learns that trials with a larger number of primary endpoints tend to have longer durations, a relationship that is not necessarily causal. Moreover, some of the trial features that our model takes as input (e.g., therapeutic area, phase) are not design decisions that can be modified by the study team. Finally, the current trial features do not reflect all characteristics of a clinical trial, specifically those that are subjective like specific regulatory or ethical characteristics. Despite these limitations, model forecasts can still be useful for operational planning and fill an existing gap in the design process, which is to improve trial design decisions without full dependence on conventional wisdom or legacy norms.

CONCLUSION

The operational efficiency of clinical trials is inexplicably tied to the rate of progress of medicine. In this study, we show that data-driven forecasts of efficiency metrics are feasible using commonly collected study design features. Additionally, we use our model to examine relationships between trial features and operational outcomes, which help in understanding the impact of clinical trial design on efficiency. We hope that these results can reinforce the systematic collection and modeling of clinical trial data at a large-scale. We

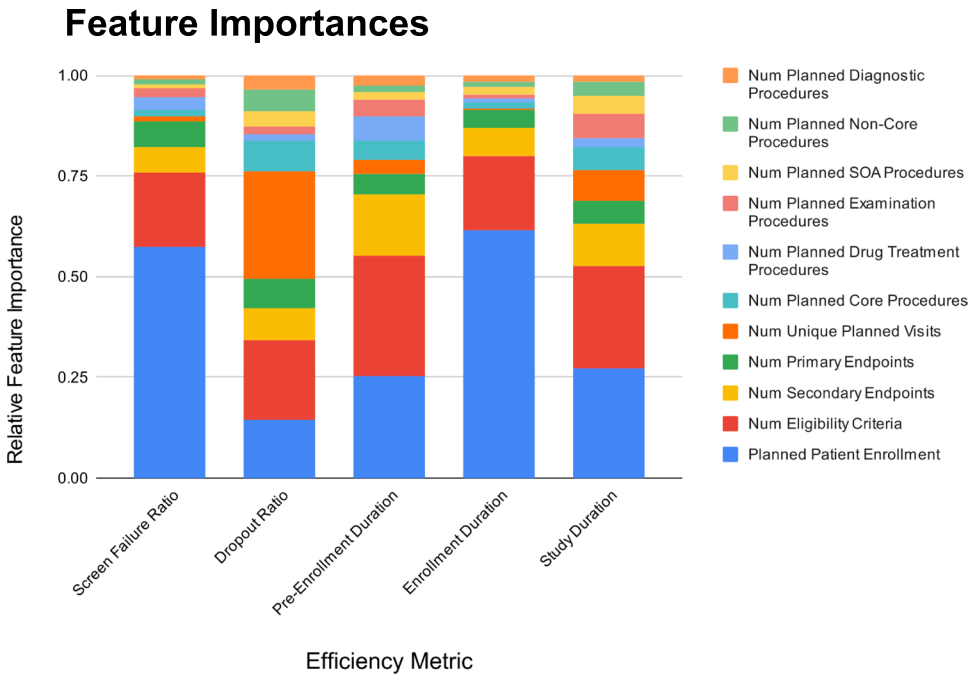

Fig. III We report the importance of features used in our LightGBM model, defined as the information gained from each feature with respect to the loss function. For visual clarity, we normalize the importance scores to sum to one for each metric. Additionally, for interpretability, we only report the importance of a subset of actionable features, rather than the whole set of features used by the model

anticipate the predictive power and reliability of such models to improve as more rich data is collected over time.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1208/s12248-022-00703-3.

Acknowledgements We thank Ernesto Guarin, Nicole Kim, Katelyn Bechler, James Harper, Jacek Basta, and Jennifer Copping for their helpful feedback.

Author Contribution KW and EW were involved in the design, analysis, and interpretation of data. JZ was involved in the design and interpretation of the data. MD, NC, ML, MD, and KK were involved in feedback on content and interpretation. HR, RL, MG, NP, CH, SR, and RC were involved in the review and feedback on content.

Funding This work was funded by Genentech, part of the Roche Group.

Code Availability The trained machine learning model is available at https://github.com/kevinwu23/clinical-trial-efficiency.

Declarations

Competing Interest MDA, NC, ML, MD, KK, HR, MG, NP, CH, SR, and RC are employees of Roche.

REFERENCES

1. Kelly D, Spreafico A, Siu LL. Increasing operational and scientific efficiency in clinical trials. Br J Cancer. 2020;123(8):1207–8.
2. Rosenblatt M. The large pharmaceutical company perspective. N Engl J Med. 2017;376(1):52–60.
3. Martin L, Hutchens M, Hawkins C, Radnov A. How much do clinical trials cost. Nat Rev Drug Discov. 2017;16(6):381–2.
4. Hwang TJ, Carpenter D, Lauffenburger JC, Wang B, Franklin JM, Kesselheim AS. Failure of investigational drugs in late-stage clinical development and publication of trial results. JAMA Intern Med. 2016;176(12):1826–33.
5. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics. 2019;20(2):273–86.
6. Getz K. Improving protocol design feasibility to drive drug development economics and performance. Int J Environ Res Public Health. 2014;11(5):5069–80.
7. Kaitin KI. Deconstructing the drug development process: the new face of innovation. Clin Pharmacol Ther. 2010;87(3):356–61.
8. Ledford H. Translational research: 4 ways to fix the clinical trial. Nat News. 2011;477(7366):526–8.
9. Liu R, Rizzo S, Whipple S, Pal N, Pineda AL, Lu M, *et al.* Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature. 2021;592(7855):629–33.
10. Shenoy P. Multi-regional clinical trials and global drug development. Perspect Clin Res. 2016;7(2):62.
11. Song SY, Chee D, Kim E. Strategic inclusion of regions in multiregional clinical trials. Clin Trials. 2019;16(1):98–105.
12. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp Clin Trials Commun. 2018;11:156–64.
13. Allison M. Reinventing clinical trials. Nat Biotechnol. 2012 Jan 9;30(1):41–9.
14. Cavallo C, Labib MA, Honea N, Nakaji P. Enrollment-to-screening ratio: an undervalued data in randomized clinical trials. Neurosurgery. 2018.
15. Craven BC, Balioussis C, Hitzig SL, Moore C, Verrier MC, Giangregorio LM, *et al.* Use of screening to recruitment ratios as a tool for planning and implementing spinal cord injury rehabilitation research. Spinal Cord. 2014;52(10):764–8.
16. Blanton S, Morris DM, Prettyman MG, McCulloch K, Redmond S, Light KE, *et al.* Lessons learned in participant

recruitment and retention: the EXCITE trial. Phys Ther. 2006;86(11):1520–33.
17. Harris-Brown TM, Paterson DL. Reporting of pre-enrolment screening with randomized clinical trials: a small item that could impact a big difference. Perspect Clin Res. 2015;6(3):139.
18. Giffin RB, Lebovitz Y, English RA. Transforming clinical research in the United States: challenges and opportunities: workshop summary: National Academies Press; 2010.
19. Lamberti MJ, Smith Z, Henry R, Howe D, Goodwin M, Williams A, *et al.* Benchmarking patient recruitment and retention practices. Ther Innov Regul Sci. 2021;55(1):19–32.
20. Stergiopoulos S, Calvert SB, Brown CA, Awatin J, Tenaerts P, Holland TL, *et al.* Cost drivers of a hospital-acquired bacterial pneumonia and ventilator-associated bacterial pneumonia phase 3 clinical trial. Clin Infect Dis. 2018;66(1):72–80.
21. McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, *et al.* What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials. 2006;7(1):1–8.
22. Keith SJ. Evaluating characteristics of patient selection and drop-out rates. J Clin Psychiatry. 2001;62:11–6.
23. Dicesare J. Improve the clinical trial startup process with just-in-time site activation. Applied Clinical Trials [Internet]. 2014 Aug 19; Available from: https://www.appliedclinicaltrialsonline.com/view/improve-clinical-trial-startup-process-just-time-site-activation
24. Getz KA. Characterizing the real cost of site regulatory compliance. Appl Clin Trials. 2015;24(6/7):18.
25. Eichler H-G, Sweeney F. The evolution of clinical trials: can we address the challenges of the future? Clin Trials. 2018;15(1_suppl):27–32.
26. Smuck B, Bettello P, Berghout K, Hanna T, Kowaleski B, Phippard L, *et al.* Ontario protocol assessment level: clinical trial complexity rating tool for workload planning in oncology clinical trials. J Oncol Pract. 2011;7(2):80–4.
27. Cunanan KM, Gonen M, Shen R, Hyman DM, Riely GJ, Begg CB, *et al.* Basket trials in oncology: a trade-off between complexity and efficiency. J Clin Oncol. 2017;35(3):271.
28. Malik L, Lu D. Increasing complexity in oncology phase I clinical trials. Invest New Drugs. 2019;37(3):519–23.
29. Yuan G, Wang L, Li J, Feng H, Ji J, Gu W, *et al.* Complexity in clinical trials: blind spots, misleading criteria, winners and losers. Clin Cancer Drugs. 2020;7(1):3–15.
30. Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D. Electronic screening improves efficiency in clinical trial recruitment. J Am Med Inform Assoc. 2009;16(6):869–73.
31. Frank G. Current challenges in clinical trial patient recruitment and enrollment. SoCRA Source. 2004;2(February):30–8.
32. Kadam RA, Borde SU, Madas SA, Salvi SS, Limaye SS. Challenges in recruitment and retention of clinical trial subjects. Perspect Clin Res. 2016;7(3):137.
33. Dilts DM, Sandler AB, Baker M, Cheng SK, George SL, Karas KS, *et al.* Processes to activate phase III clinical trials in a Cooperative Oncology Group: the Case of Cancer and Leukemia Group B. J Clin Oncol. 2006;24(28):4553–7.
34. Getz KA, Wenger J, Campo RA, Seguine ES, Kaitin KI. Assessing the impact of protocol design changes on clinical trial performance. Am J Ther. 2008;15(5):450–7.
35. Ross S, Grant A, Counsell C, Gillespie W, Russell I, Prescott R. Barriers to participation in randomised controlled trials: a systematic review. J Clin Epidemiol. 1999;52(12):1143–56.
36. Boericke K, Gwinn B. Planned to perfection. Int Clin Trials. 2010;17(8):26–30.
37. Andersen JW, Fass R, van der Horst C. Factors associated with early study discontinuation in AACTG studies, DACS 200. Contemp Clin Trials. 2007;28(5):583–92.
38. Tseo Y, Salkola MI, Mohamed A, Kumar A, Abnousi F. Information extraction of clinical trial eligibility criteria. ArXiv Prepr ArXiv200607296. 2020;
39. Liu H, Chi Y, Butler A, Sun Y, Weng C. A knowledge base of clinical trial eligibility criteria. J Biomed Inform. 2021;117:103771.
40. Munos B, Niederreiter J, Riccaboni M. Improving the prediction of clinical success using machine learning. 2020;
41. Feijoo F, Palopoli M, Bernstein J, Siddiqui S, Albright TE. Key indicators of phase transition for clinical trials through machine learning. Drug Discov Today. 2020;25(2):414–21.
42. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al.* Lightgbm: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30:3146–54.
43. Feurer M, Klein A, Eggensperger K, Springenberg JT, Blum M, Hutter F. Auto-sklearn: efficient and robust automated machine learning. In: Automated Machine Learning. Springer, Cham; 2019. p. 113–34.
44. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982;247(18):2543–6.
45. Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. J R Stat Soc Ser B Stat Methodol. 2020;82(4):1059–86.