Predicting Drug Approvals: the Novartis Data Science and Artificial Intelligence Challenge

Kien Wei Siah[1,2], Nicholas Kelley[3], Steffen Ballerstedt[3], Björn Holzhauer[3], Tianmeng Lyu[4], David Mettler[3], Sophie Sun[4], Simon Wandel[3], Yang Zhong[5], Bin Zhou[5], Shifeng Pan[5], Yingyao Zhou[5], Andrew W. Lo[1,2,6,*]

[1]Laboratory for Financial Engineering, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America; [2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America; [3]Novartis, Basel, Switzerland; [4]Novartis, East Hanover, New Jersey, United States of America; [5]Genomics Institute of the Novartis Research Foundation, San Diego, California, United States of America; [6]Sante Fe Institute, Santa Fe, New Mexico, United States of America

*Corresponding author: Andrew W. Lo, MIT Sloan School of Management, 100 Main Street, E62–618, Cambridge, MA 02142. (617) 253–0920 (tel), (781) 891–9783 (fax), alo-admin@mit.edu (email).

This version: February 17, 2021

Word count: 4,164 words

Abstract

We describe a novel collaboration between academia and industry, an in-house data science and artificial intelligence challenge held by Novartis to develop machine learning models for predicting drug development outcomes, building upon research at MIT using data from Informa® as the starting point. Over 50 cross-functional teams from 25 Novartis offices around the world participated in the challenge. The domain expertise of these Novartis researchers was leveraged to create predictive models with greater sophistication, two teams developed models that outperformed the baseline MIT model through state-of-the-art machine learning algorithms and the use of newly incorporated features and data. In addition to validating the variables shown to be associated with drug approval in the earlier MIT study, the challenge also provided new insights into the drivers of drug development success and failure.

Table of Contents

1 Introduction

The rising cost of clinical trials and a shift to utilizing more complex biological pathways with greater therapeutic potential—but also greater chances of failure—have in the past decade caused drug development to become an increasingly lengthy, costly, and risky endeavor. The average drug now requires at least a decade of translational research involving multiple iterations of lead optimization and several phases of clinical studies costing hundreds of millions of dollars before it can be approved by drug regulatory authorities, such as the U.S. Food and Drug Administration (FDA).

Due to the capital-intensive nature of the drug development process, biotech and pharma companies are only able to afford to invest in a limited number of projects. When managing their portfolios of investigational drugs, these developers typically use historical estimates of regulatory approval rates, based on the therapeutic class and phase of development of the drug, combined with subjective adjustments, determined through unstructured discussions of project-specific risk factors, to make their investment decisions. Recently, however, there has been an increased interest in combining machine learning predictions with human judgments on project specific information in a more structured manner.[1]

In a recent large-scale study involving over 6,000 unique drugs and close to 20,000 clinical trials, Lo et al.[2] proposed using a range of drug and clinical trial features in machine-learning techniques to more accurately estimate the probabilities of success of pipeline candidates. Using two proprietary pharmaceutical pipeline database snapshots (2015Q4) provided by Informa® (*Pharmaprojects* and *Trialtrove*), Lo et al.[2] developed models that achieved promising predictive accuracy, measured at 0.78 and 0.81 AUC for predicting transitions from phase 2 to regulatory approval and phase 3 to regulatory approval, respectively. (The AUC, also known as the area under the receiver operating characteristic curve, is the estimated probability that a classifier will rank a positive outcome higher than a negative outcome.) Its models also identified the most useful features for predicting drug development outcomes: trial outcome, trial status, trial accrual, trial duration, prior approval for another indication, and sponsor track record.

With more accurate forecasts of the likelihood of clinical trial success and a better understanding of the drivers of drug approval, biopharma companies and investors should be better able to assess the risks of different drug development projects, and thus allocate their capital more efficiently.

As an extension of the previous study, the MIT team collaborated with Novartis, one of the largest multinational pharmaceutical companies in the world, to implement an in-house Data Science and Artificial Intelligence (DSAI) challenge based on updated snapshots (2019Q1) of the same Informa® databases. This challenge was designed to leverage the domain expertise of Novartis data scientists, statisticians, portfolio managers, and researchers to develop

more powerful models for predicting the probability of success of pipeline drug candidates and uncover deeper insights into the drivers of drug approval. Success in this context was defined as regulatory approval. Over 50 teams participated in the challenge, consisting of more than 300 individuals from 25 Novartis offices around the world, submitting approximately 3,000 models for evaluation in a head-to-head competition. In addition to their predictive performance, the teams were evaluated on the innovativeness and robustness of their models, and the potential business value of their findings.

In this paper, we summarize the findings of the top-performing teams. By examining their models, we validate the variables previously found to be associated with drug approval, and identify new features that contain useful signals about drug development outcomes.

2 Methods

a Data

For the DSAI challenge, we use two pharmaceutical pipeline databases from the commercial data vendor Informa® for the core dataset: *Pharmaprojects*, which specializes in drug information, and *Trialtrove*, which specializes in clinical trial intelligence.[3] These two databases aggregate drug and trial information from over 40,000 data sources in the public domain, including company press releases, government drug and trial databases (e.g., Drugs@FDA and Clinicaltrials.gov), and scientific conferences and publications. The database snapshots used in this paper are updated versions of that used in Lo et al.[2] (2019Q1 versus 2015Q4).

As in Lo et al.[2], we construct a dataset of drug-indication pairs, focused on phase 2 trial data that have known outcomes ("P2APP"), either successful registration or program termination. We extract a range of drug compound attributes and clinical trial characteristics as potential features for prediction, including three binary features, one date, seven numerical features, two multi-class features, sixteen multi-label features, and five unstructured free texts. These are summarized in Table 1. For the purpose of our analysis, we define the development status of suspension, termination, and lack of development as "failures," and registration and launch in at least one country as "successes" or approvals. (See Supplementary Materials A for further details.)

This dataset consists of 6,901 drug-indication pairs and 12,680 unique phase 2 clinical trials, with end dates spanning 1999 to early 2019, containing about two decades of data (Table 2). In our dataset, 796 drug-indication pairs (11.5%) were successes, and 6,105 drug-indication pairs (88.5%) ended in failure. The data covers fifteen indication groups: alimentary, anti-infective, anti-parasitic, blood and clotting, cardiovascular, dermatological, genitourinary, hormonal, immunological, musculoskeletal, neurological, anti-cancer, rare diseases,

respiratory, and sensory products. Drugs for cancer, rare diseases, and neurological diseases make up the largest subgroups. As expected, the majority of the trials in the dataset are sponsored by industry, rather than investigator-initiated academic trials.

Table 1. Features extracted from *Pharmaprojects* and *Trialtrove*. See Supplementary Materials A for examples of each feature.

	Description
Drug-indication Pair	
Biological target	Protein on which the drug acts.
Country	Country in which the drug is being developed.
Drug-indication development status	Current phase of development of the drug-indication pair.
Indication	Indication for which the drug is under development.
Mechanism of action	Mechanism through which the drug produces its pharmacological effect.
Medium	Physical composition of the material in which the drug is contained.
Name	Name of the drug.
Origin	Origin of the active ingredient in the drug.
Prior approval of drug for another indication	Approval of the drug for another indication prior to the indication under consideration.
Route	Route by which the drug is administered.
Therapeutic class	Therapy area for which the drug is in development.
Trial	
Attribute	Distinguishing attribute or feature of the trial, e.g., registration trials, biomarkers, immuno-oncology.
Actual accrual	Number of patients enrolled in the trial.
Disease type	Disease, disorder, or syndrome studied in the trial.
Duration	Duration of the trial.
Exclusion criteria	Criteria for excluding a patient from trial consideration.
Gender	Gender of the enrolled patients.
Investigator experience	Primary investigator's success in developing other drugs prior to the drug-indication pair under consideration.
Location	Country in which the trial is conducted.
Number of identified sites	Number of sites where the trial is conducted.
Outcome	Outcome of the trial.
Patient age	Minimum and maximum age of the enrolled patients.
Patient population	General information about the disease condition of the enrolled patients.
Patient segment	Disease segmentation by patient subtypes, therapeutic objectives, or disease progression/staging.
Phase 2 end date	Year phase 2 ended (end date of the last observed phase 2 trial).
Primary endpoint	Detailed description of primary objective, endpoint or outcome of the trial. Endpoints are classified into four main groups: efficacy, safety/toxicity, health economics and outcomes research, and pharmacokinetics/pharmacodynamics.
Sponsor	Financial sponsor of the trial.
Sponsor track record	Sponsor's success in developing other drugs prior to the drug-indication pair under consideration.
Sponsor type	Sponsor grouped by type.
Status	Recruitment status of the trial.
Design	Investigative methods used in the trial.
Design keywords	Keywords relating to investigative methods used in the trial.
Target accrual	Number of patients sought for the trial.
Therapeutic area	Therapeutic area of the disease studied in the trial.

Table 2. Sample sizes of the P2APP dataset, and the training and testing data used for the challenge.

	Drug-indication Pairs	Clinical Trials	Unique Drugs	Unique Indications	Unique Clinical Trials
P2APP					
Success	796	2,435	614	182	2,209
Failure	6,105	13,203	3,313	283	10,722
Total	**6,901**	**15,638**	**3,726**	**291**	**12,680**
Training Data					
Success	610	1,852	468	169	1,666
Failure	4,293	6,839	2,537	264	5,845
Total	**4,903**	**8,691**	**2,872**	**272**	**7,451**
Testing Data					
Success	186	583	160	93	557
Failure	1,812	6,364	1,096	218	5,065
Total	**1,998**	**6,947**	**1,229**	**229**	**5,561**

b Challenge Setup

The DSAI challenge was hosted on an Aridhia Digital Research Environment (Aridhia DRE), a cloud-based platform designed for collaborative data analytics on healthcare data.[4] Each team was provided a remote workspace for accessing the data, computing resources for developing their models, and a Git repository hosted by AIcrowd[5] for managing their source code. AIcrowd was also used to host a leaderboard and discussion forum for teams to interact and answer questions. See Figure 1 for an illustration of the setup.

For the leaderboard challenge, teams were required to predict the probability of regulatory approval (i.e., the drug-indication development status) given phase 2 trial data and drug compound characteristics (see Table 1). This corresponds to a real world decision-making scenario whereby a pharmaceutical company must decide whether to invest in a phase 3 program based on phase 2 results. We split the P2APP dataset chronologically, with drug-indication pairs that failed or succeeded before 2016 provided to the participants as training data, while those pairs that failed or succeeded in 2016 or later were held out as testing data for leaderboard evaluation. Table 2 shows the sample sizes of the training data and the testing data. Teams were encouraged to create new features in the core dataset in addition to those provided by linking new datasets (e.g., compound data) and through feature engineering.

The challenge spanned five months, from October 2019 to March 2020: one month for team registration and onboarding, two months for model development and submission, and two months for final evaluation. During the model development segment, teams built their models using the training data, and were able to receive real-time feedback on the performance of their models on a subset of the testing data (50%) and how this compared with other teams ("open-testing round"). This happened via a public leaderboard, which was updated with every submission. This gave participants the opportunity to refine and

calibrate their algorithms. Additionally, all submissions made by each team were evaluated on the complete testing set (100%) in the final evaluation round. This information was not shown to participants during the competition, defining the private leaderboard to assess performance. We used the binary cross entropy log loss function as the primary scoring metric.

We also trained a baseline model based on the algorithm described in Lo et al.[2] using the same training data provided to the participants. To obtain the confidence interval of the performance of each model, we bootstrapped the testing set 1,000 times and evaluated the models on the same bootstrapped datasets.

As part of the final evaluation process, teams were required to upload the code used to train their models, and a write-up describing their methods and results. A committee screened and ranked teams by the technical, data science, and business aspects of their submissions. Along the technical dimension, each team's source code repository was examined to ensure that the results reported were robust and reproducible. The submission history of the top-performing teams was also reviewed to ensure that they did not gain an unfair advantage by making frequent submissions. For data science, the novelty of the solutions was evaluated in terms of its data wrangling and adopted methodology. Finally, since the potential business value of the findings would be to inform portfolio and risk management decisions, the focus for the business evaluation was on the interpretability of the models, i.e., the ease of insight regarding the risk factors and key drivers of approval. Teams were ranked based on their leaderboard performance and the three dimensions above.

Subsequent to this evaluation, the two top-performing teams were selected to present their findings to a final committee consisting of Novartis leaders from its portfolio strategy and biostatistics divisions and its Digital Office, and MIT researchers Andrew W. Lo and Kien Wei Siah. Other teams with innovative approaches were also invited as part of a panel discussion to share their experience with the broader Novartis community.

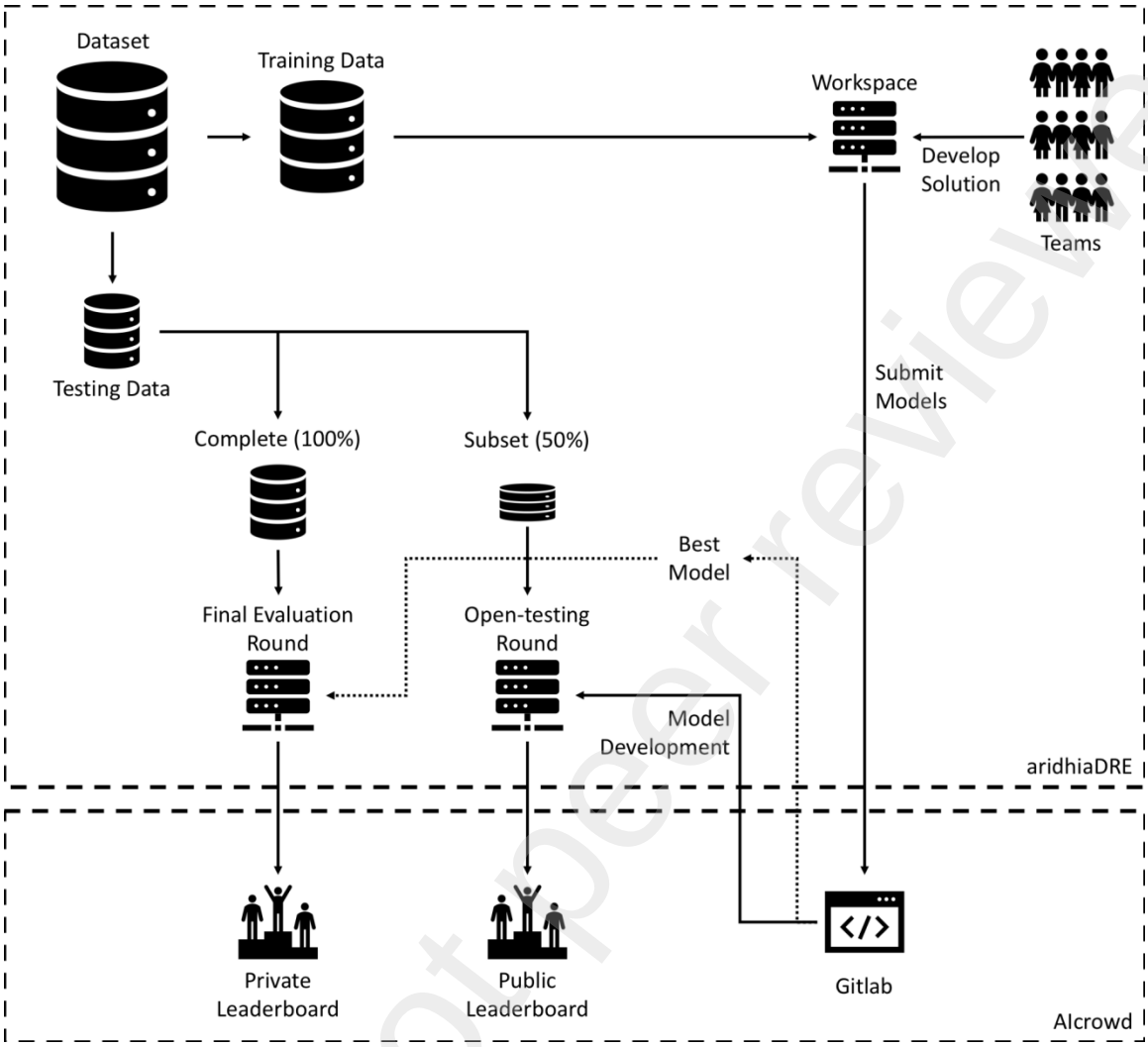

Figure 1. DSAI challenge setup. The challenge was hosted on Aridhia DRE and AIcrowd. It consists of an open-testing round for teams to refine and calibrate their models, and a final evaluation round.

3 Results

We received approximately 3,000 model submissions in the open-testing round of the leaderboard challenge. The teams explored a wide range of machine-learning models, ranging from traditional logistic regression, support vector machines, decision trees, and neural networks to ensemble methods such as random forests[6], gradient boosting machines, XGBoost[7], and combinations of multiple types of models. Figure 2 and Figure 3 show the public leaderboard scores of participating teams over time and their corresponding rankings, respectively.

Recognizing the dangers of overfitting that arise from the reuse of testing set data[8], we created a scatterplot of public and private leaderboard scores to assess the extent of adaptive

overfitting (see Figure 4). The public scores were evaluated on a subset of the testing set provided to the participants during the open-testing round, while the private scores were evaluated on the complete testing set in the final evaluation round. In the ideal case, the points would lay close to the diagonal, since the public and private performance of the models would be almost identical. In contrast, deviations from the diagonal suggest possible overfitting. We observe that our scores approximated the ideal case in Figure 4, indicating that there was little evidence of competitors overfitting to the public leaderboard score in the DSAI challenge.

In Figure 5, we compare the performance of the top ten ranking teams to the baseline model described in Lo et al.[2], using the private leaderboard log loss and the AUC as our metrics. While the baseline model had a worse log loss compared to the top ten best performing teams, its AUC (0.78 with 95% CI [0.75, 0.82]) was only lower than the top two teams in the challenge. This may be in part because the teams in the competition attempted to optimize log-loss.

We focus on the approaches of the two teams that outperformed the baseline model on all metrics. These teams had different strategies and backgrounds of expertise, but were aligned in the way they harnessed human insight into their model predictions:

- The team with the top-ranked model was primarily composed of biostatisticians with significant domain expertise in clinical trial data analysis. It relied on handcrafted features that incorporated their insights into drug development timelines and which data entries should be discarded. A team member with portfolio management experience also provided a different perspective.
- The runner-up team was primarily composed of data scientists with domain expertise in bioinformatics and cheminformatics. It relied on extensive data exploration and feature engineering, in particular developing a novel method to understand the interaction of these features, but also augmented them with clinical trial knowledge.

a Approach of the Top-Performing Team

The top-performing model was developed by a collaborative team (team "Insight-Out") from Novartis offices in the U.S. and Switzerland whose members had backgrounds in biostatistics, data science and portfolio management. Its model achieved an AUC of 0.88 (95% CI [0.85, 0.90]), corresponding to an improvement of approximately 0.10 over the baseline model. In addition to using the core features provided in the dataset, the team created several new features to capture information about orphan drug indications, to improve the granularity of therapeutic areas, to compare the relative size of phase 2 trials to the average by therapeutic area and disease, to classify the drug candidate as a novel compound, a Lifecycle

Management (LCM) project, or a generic, and to determine whether an international nonproprietary name (INN) has been registered for the drug.

The final model of the top-performing team was an ensemble consisting of two XGBoost models and one Bayesian logistic regression[9,10] (BLR) model. The XGBoost models, known to be highly effective for tabular data, were trained using 263 raw and derived features, using time-series cross validation with different levels of hyperparameter tuning (i.e., using simple heuristics and a more sophisticated approach involving differential evolution optimization[11]). Subsequently, logistic regression with a ridge penalty was used to combine the trial-level predictions of the XGBoost models into predictions at the drug-indication level.

The BLR model was trained using case weights based on covariate balancing propensity scores[12], with greater weights given to cases that had a greater propensity of appearing in the test set. The BLR model allowed the team to incorporate its judgment on the likely effects of a smaller set of features. These included granular therapeutic areas as a random effect, novelty (e.g., that a drug was non-generic, and not an insulin or a flu vaccine), the relative phase 2 accrual versus the disease average, the success rates of drugs with the same mechanism of action, INN assignment, and trial outcomes, as well as interactions between these features. Its parameters were estimated via Markov chain Monte Carlo sampling.

Ensembles of diverse models can generally outperform any individual model.[13] The ensemble predictions were obtained as a weighted average of the predictions from the XGBoost and the BLR models. Afterwards, the predictions were post-processed using heuristics derived from the team's domain expertise. For example, the predictions for trials after 2018 were rescaled between 0.001 and 0.1 because the team believed that obtaining approval within two years of completing phase 2 was unlikely. These limits were determined based on prior elicitation using the roulette method.[14] In addition, the team introduced upper and lower bounds for their predictions to reduce the impact of overconfident and over-pessimistic predictions on the log loss, since extreme predictions that are incorrect are heavily penalized under the log loss metric.

The team found that the phase 2 accrual relative to the disease average was one of the strongest predictors of approval. The likelihood of success increased for programs with above average accrual compared with other programs for the same disease. In contrast, programs with below average accrual were more likely to fail. The team also found prior approvals for any indication (e.g., LCM programs), past approvals of other drugs for similar indications, and well-established modes of action improved the odds of approval, suggesting that repositioning an approved drug for a new indication is less challenging than developing a first-in-class new chemical entity. On the other hand, it found that drugs that targeted difficult-to-treat diseases, such as cancer or Alzheimer's disease, were more likely to fail. Trial termination (whether due to lack of efficacy, safety issues, or pipeline reprioritization),

poor patient enrollment versus planned accrual, and the absence of an INN were also strong indicators of failure.

b Approach of the Second Place Team

The second place model was developed by a team of data scientists and researchers from the Genomics Institute of the Novartis Research Foundation (team "E2C"). This model achieved an AUC of 0.84 (95% CI [0.81, 0.86]), corresponding to an improvement of approximately 0.06 over the baseline model. The team performed extensive feature engineering, creating rank normalized versions of features known to demonstrate temporal coupling (e.g., phase 2 trial durations, which have shown greater mean and spread over the years). This was done because decision tree algorithms tend to be inefficient at incorporating heteroskedasticity. In addition to the core features in the dataset such as prior approval, the team created new variables to capture the impact of development history on future approvals. For example, it computed the number of past trials in which each drug had been involved, by phase, by outcome, and in aggregate, regardless of indication. It additionally made a similar computation for indications and indication groups, aggregating them over all drugs. The team also used natural language processing techniques, such as the term frequency-inverse document frequency (TFIDF) algorithm, to convert text data for trials into feature vectors. Because the set of features under consideration was large, the team performed stepwise feature selection using random forests to identify a parsimonious set of factors.

From the outset, the second-place team focused on the XGBoost model, an algorithm that has a strong track record in data science competitions. It explored multiple training-validation strategies for hyperparameter selection, eventually settling on the random five-fold cross validation approach. Like the top team, it also post-processed trial-level predictions from the XGBoost model, based on expert knowledge. For example, it reduced the predictions for trials after 2018 because team members believed that approval within two years was unlikely. It also clipped overconfident and over-pessimistic predictions to reduce the impact of outliers on the log loss scoring metric. Unlike the leading team, however, it obtained predictions for each drug-indication pair by using the maximum trial-level prediction across all trials associated with the drug-indication pair, as opposed to using penalized logistic regression. It hypothesized that the best performing trial would dominate the outcome of the drug-indication pair, regardless of any lack of evidence in other trials in support of efficacy.

Among the final set of features, the second-place team found that rank-normalized variables were generally favored over their raw, unnormalized counterparts, thus verifying the importance of normalization. Out of the top 20 most important features, 8 were novel features that were created by the team and not provided in the core dataset. It found the top features were largely consistent with those reported by Lo et al.[2], e.g., trial outcomes, trial

accrual, prior approval, and sponsor track records. Moreover, it found that drugs with strong development histories, as quantified by the percentage of past trials with positive outcomes, were more likely to be successful. Over- and under-enrollment with respect to the target accrual were also associated with lower success rates, a not entirely unexpected finding, since these signs hint at poor trial operation or a lack of efficacy. Interestingly, the team found that trials with a younger age inclusion criterion tended to be more successful. However, features created from text data did not seem to contribute meaningful predictive value.

In addition to single feature analysis, the second-place team went a step further to identify informative feature pairs. It found strong interaction effects between trial outcomes and drug development history, e.g., the historical success rate of past trials and the presence or absence of prior approval. For example, given a successful trial with its primary endpoints met, a drug with prior approval for other indications was almost twice as likely to be approved versus a new compound without any prior approval. The team also found that drugs with strong track records had higher probabilities of success in indications that had been less explored in the development process, as quantified by the cumulative number of past trials.

In addition, the team observed there was strong coupling between the success of anticancer drugs and their development history. The likelihood of success of an anticancer drug was five times greater with a prior approval than without. This effect was less pronounced in non-cancer programs, where the ratio in success rates conditional on prior approval was only twice as great. The team hypothesized that historical success rates and prior approval were especially important for anticancer drugs because it is not uncommon for effective cancer therapies to work across multiple cancer subtypes (e.g., chemotherapy), and therefore, an approval in one subtype was predictive of potential success in other subtypes.

Figure 2. Public leaderboard scores of teams over time. Each point corresponds to a submission. We use lines to trace each team's best log loss performance. We truncate the log loss axis at 1.0 for better visualization.

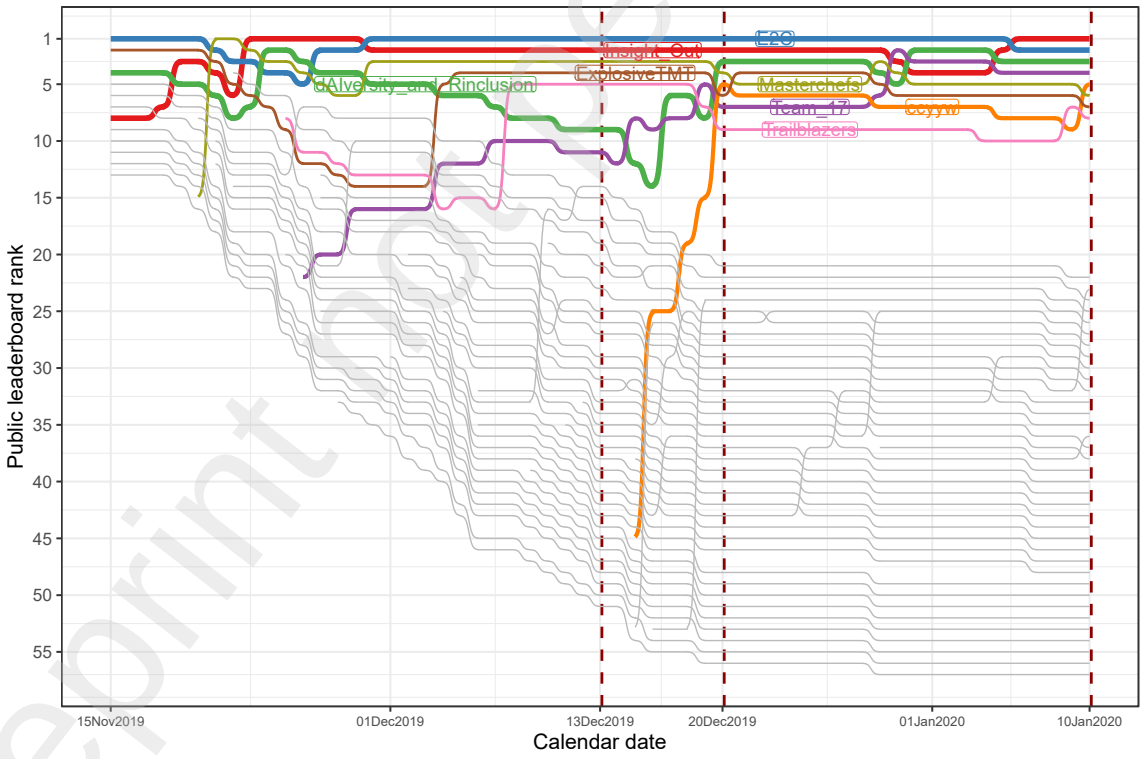

Figure 3. Public leaderboard rankings over time.

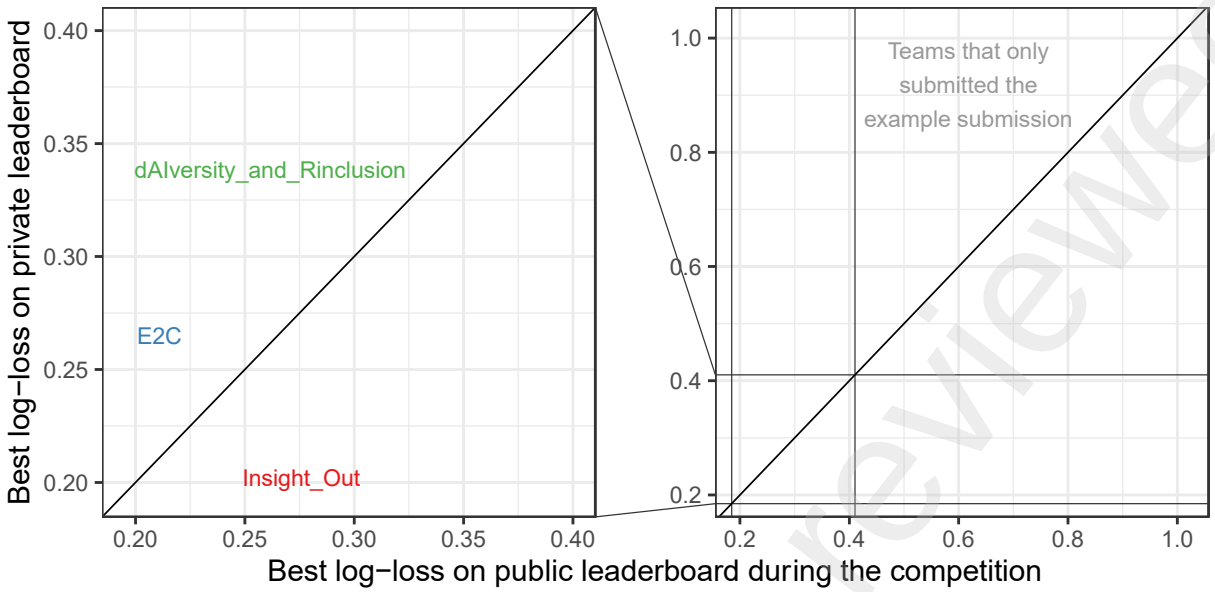

Figure 4. Scatterplot of public and private leaderboard scores. Each point corresponds to the best performing submission of each team. The points lay very close to the diagonal, which indicates that there is little evidence of overfitting in the competition.

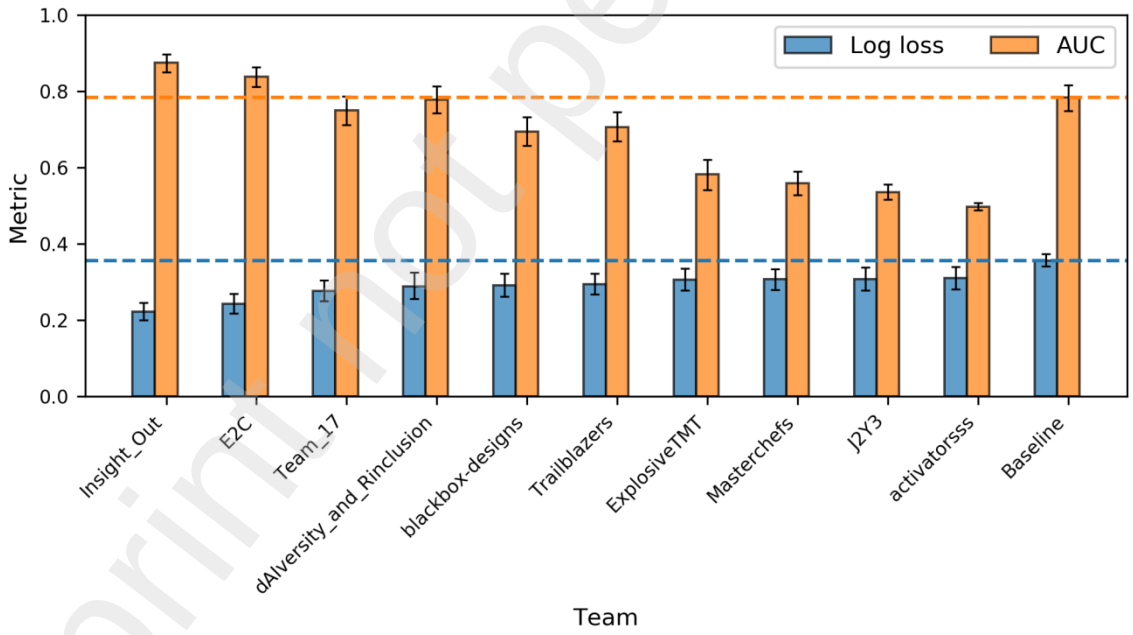

Figure 5. Private leaderboard log loss and AUC for the top ten ranking teams and the baseline model. The top two teams outperformed all other submissions in the leaderboard challenge, including the baseline model.

4 Discussion

MIT and Novartis researchers collaborated on an in-house DSAI challenge to develop machine learning models for predicting clinical development outcomes, building on Lo et al.[2], whose work used one of the largest pharmaceutical pipeline databases in the world, provided by Informa®. To the best of our knowledge, this challenge represents the first crowd-sourced collaborative competition to use pharmaceutical data for this purpose, in this case, updated snapshots of the same Informa® databases used in the earlier MIT study. In total, over 50 cross-functional teams from 25 international Novartis offices participated in the challenge. We received approximately 3,000 model submissions over a two-month period.

Internal data science competitions are both an opportunity for a company to address business problems, as well as a learning opportunity for the company's data science community. From this perspective, the large number of Novartis associates who chose to actively participate in the process and had the chance to expand their data science skillset was encouraging.

The probability of success is one of several key parameters, in combination with unmet medical need and market opportunity, which clinical researchers, biopharma investors, and portfolio managers consider when making scientific and business decisions about drug development. Accurate estimates of this parameter are therefore critical for efficient risk management and resource allocation. The top performing teams in their winning solutions delivered additional heuristics with respect to predicting the probability of success:

- Identification of novel features predictive of probability of success (as outlined above).
- Novel approaches and methodologies for feature extraction, combining domain expertise and machine learning.
- Creative ways of introducing additional data types to the problem, such as unstructured text and biochemical data. For example, several teams presented ways of connecting new data types, although this in itself did not translate into top leaderboard performance.

Additionally, the discussion about the availability of specific information at the time of decision-making about the fate of a project was also helpful for assessing the potential for target leakage in the solutions of external vendors offering similar predictive solutions.

The DSAI challenge also had several limitations. First, the P2APP dataset was split chronologically, using drug-indication pairs that failed or succeeded before 2016 as training data, and those that failed or succeeded in 2016 or later were held out as testing data. Due to the nature of drug development, however, some boundary effects were inevitably present in the last years of the testing data. Because drugs tend to fail much more quickly than those that are approved, the majority of the trials completed after 2018 ended in failure. With their experience and expertise in drug development, both teams eventually discovered this

artifact in the data, and were able to improve their model performance by adjusting their predictions for trials after 2018. While such adjustments were useful in the competition, they add little practical value for real-life application.

Second, some available features reflected a decision already taken by a company to terminate a project. These included trials that were stopped due to pipeline re-prioritization, a small-sized phase 2 program due to stopping the program after an initial small trial, and the failure to apply for an INN. Not all such information is available at the time of decision-making in practice. These limitations illustrate that in order to make data science competitions directly useful for business problems without substantial modification, it is important to align the prediction task in the competition with the real-world business problem extremely closely.

We also received feedback from knowledgeable participants that the core dataset lacked key information that decision makers typically take into consideration at the time of decision, such as the preclinical data, detailed safety and efficacy data, and the biological plausibility of the mechanism of action. Unfortunately, investigators do not usually release this information to the public domain for strategic reasons. It is therefore unsurprising that such data are not available in commercial pharmaceutical databases based on publicly available sources of information. Potentially, this limitation may be overcome with recent progress in deep learning approaches to natural language processing, which may enable information about trial protocols, development programs and drugs to be extracted from unstructured text data sources.

5 Conclusion

By tapping the power of crowdsourcing and the domain expertise of Novartis researchers working in cross-disciplinary teams, we have shown the potential for data science and artificial intelligence challenges to generate predictive models for drug development outcomes that outperform existing models from the academic literature. In addition to validating features previously associated with drug approval in the MIT study, the DSAI challenge has provided new insights into the drivers of drug approval and failure. Ultimately, these new predictive models can be used to augment human judgment to make more informed decisions in portfolio risk management. Nevertheless, there remains a clear opportunity to further improve the models in this competition. We believe that more accurate models can be developed with access to better quality and more comprehensive data, and a broader pool of challenge participants.

Acknowledgments

Research support from the MIT Laboratory for Financial Engineering is gratefully acknowledged. We thank Informa for allowing us to use their data for this project, and Jayna Cummings for editorial assistance. The views and opinions expressed in this article are those of the authors only, and do not necessarily represent the views and opinions of any institution or agency, any of their affiliates or employees, or any of the individuals acknowledged above.

References

1. Hampson, L. V *et al. Improving the assessment of the probability of success in late stage drug development*. http://arxiv.org/abs/2102.02752 (2021).
2. Lo, A. W., Siah, K. W. & Wong, C. H. Machine Learning with Statistical Imputation for Predicting Drug Approval. *Harvard Data Sci. Rev.* **1**, (2019).
3. Informa. Citeline Data Analysis Pharma Intelligence. https://pharmaintelligence.informa.com/products-and-services/data-and-analysis/citeline (2020).
4. Aridhia. Aridhia DRE Trusted Digital Research Environment. https://www.aridhia.com/ (2020).
5. AIcrowd. AIcrowd. https://www.aicrowd.com/ (2020).
6. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
7. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* vols 13-17-August-2016 785–794 (Association for Computing Machinery, 2016).
8. Roelofs, R. *et al.* A meta-analysis of overfitting in machine learning. in *Advances in Neural Information Processing Systems* vol. 32 9179–9189 (2019).
9. Goodrich B, Gabry J, Ali I & Brilleman S. Bayesian Applied Regression Modeling via Stan. *R package version 2.21.1* https://mc-stan.org/rstanarm/ (2020).
10. Brilleman SL, Crowther MJ, Moreno-Betancur M, Buros Novik J & Wolfe R. Joint longitudinal and time-to-event models via Stan. *StanCon* https://github.com/stan-dev/stancon_talks/ (2018).
11. Brest, J., Greiner, S., Bošković, B., Mernik, M. & Zumer, V. Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. in *IEEE Transactions on Evolutionary Computation* vol. 10 646–657 (2006).
12. Imai, K. & Ratkovic, M. Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **76**, 243–263 (2014).
13. Thakur, A. *Approaching (Almost) Any Machine Learning Problem - Abhishek Thakur - Google Books*. (Abhishek Thakur, 2020).
14. Gore, S. M. Biostatistics and the medical research council. *Med. Res. Counc. News* **35**, 19–20 (1987).

Supplementary Materials

A Core Dataset

We construct our datasets using two Informa® databases: *Pharmaprojects* and *Trialtrove*, two separate relational databases organized by largely different ontologies. We extract drug-specific features and drug-indication development status from *Pharmaprojects*, and clinical trial features from *Trialtrove*.

First, we identify all drug-indication pairs with known outcomes in *Pharmaprojects*. Next, we drop pairs that do not have any trials captured in *Trialtrove*. (We note that the disease coverage in Pharmaprojects and Trialtrove is slightly different.) Because missingness is present in both *Pharmaprojects* and *Trialtrove*, we impose several additional filters to make sure that all samples collected are usable for analysis.

We summarize the steps in Table 3. It is important to note that the drug, indication, and trial relationships in the databases are surjective and non-injective: different drugs may target the same indication, and some trials may involve multiple drug-indication pairs. This is to be expected, since one drug can be indicated for multiple diseases, a disease can have more than one treatment, and it is not uncommon for a trial to involve two or more related primary investigational drugs.

We extract drug compound attributes and clinical trial characteristics from *Pharmaprojects* and *Trialtrove*, respectively (see Table 4). In addition to structured features readily available in the databases, we create an augmented set of variables that captures sponsor track record and investigator experience: we quantify the track record of sponsors of a specific trial by their success in developing other drugs, using the number of prior approved and failed drug-indication developments; and in past trials for phases 1, 2, and 3 separately, using the total number of trials sponsored, the number of trials sponsored with positive and negative results, and the number of trials sponsored to completion and termination. We use the end date of the last trial of the drug-indication pair under consideration as the cutoff for considering prior experience, since the last end date will be the time of prediction. We abstract investigator experience in the same manner.

Lastly, we also construct a binary drug-indication pair feature that indicates whether a drug has previously been approved for another indication. Similarly, we use the end date of the last trial as the cutoff for considering prior approval.

Table 3. Filters for constructing P2APP.

	Rationale
Drug-indication Pairs in *Pharmaprojects*	
Trials observed in *Trialtrove*	We exclude pairs for which we do not observe any trials in *Trialtrove*.
Known approval date (if approved)	We define the approval date as the earliest date a drug-indication pair was approved in any market. We require these dates to perform time-series analysis.
Known failure date (if failed)	We define failure date as one year after the end-date of the last phase 2 or phase 3 trial (if any), whichever is latest.
Clinical Trials in *Trialtrove*	
Phase 2 trials	P2APP focuses on phase 2 trial data
Known end date	We require these dates to create sponsor track record and investigator experience, and to perform time series analysis.
Known sponsors and disease types	Trials not tagged with sponsor/disease types are typically out of *Trialtrove* commercial coverage and not maintained.

Table 4. Features extracted from *Pharmaprojects* and *Trialtrove*.

	Examples	Type
Drug-indication Pair		
Biological target	Cytokine/Growth factor; Enzyme; Ion channel; Receptor; Transporter	Multi-label
Country	China; India; Japan; United States	Multi-label
Drug-indication development status	Approved; Failed	Binary
Indication	Cancer, lung, small cell; Cancer, lung, non-small cell; Cancer, brain	Multi-class
Mechanism of action	Cell cycle inhibitor; DNA inhibitor; Ion channel antagonist; Protein kinase inhibitor	Multi-label
Medium	Capsule, hard; Capsule, soft; Powder; Solution; Suspension; Tablet	Multi-label
Name	Free text	String
Origin	Biological, protein, antibody; Biological, protein, recombinant; Chemical, synthetic	Multi-label
Prior approval of drug for another indication	True; false	Binary
Route	Inhaled; Injectable; Oral; Topical	Multi-label
Therapeutic class	Antiviral, anti-HIV; Anticancer, immunological; Antiepileptic	Multi-label
Trial		
Attribute	Biomarker/Efficacy; Biomarker/Toxicity; Pharmacogenomic - Patient Preselection/Stratification	Multi-label
Actual accrual	Integer	Numerical
Disease type	Bladder; colorectal; ovarian	Multi-label
Duration	Integer	Numerical
Exclusion criteria	Free text	String
Gender	Male, female, both	Multi-class
Investigator experience	Refer to sponsor track record	Numerical
Location	Canada; Europe; United Kingdom; United States	Multi-label
Number of identified sites	Integer	Numerical
Outcome	Completed, Negative outcome/primary endpoint(s) not met; Completed, Outcome indeterminate; Completed, Positive outcome/primary endpoint(s) met; Terminated, Safety/adverse effects	Multi-label
Patient age	Integer	Numerical
Patient population	Free text	String
Patient segment	Stage I; stage III; stage IV; second line; pediatric	Multi-label
Phase 2 end date	Date	Date
Primary endpoint	Free text	String
Sponsor	Duke University Medical Center; National institute of Health; Celgene	Multi-label
Sponsor track record	Number of prior approved drug-indication pairs; Number of prior failed pairs; Total number of phase 1 trials sponsored; Number of phase 1 trials with positive results; Number of phase 1 trials with negative results; Number of completed phase 1 trials; Number of terminated phase 1 trials; Total number of phase 2 trials sponsored; Number of phase 2 trials with positive results; Number of phase 2 trials with negative results; Number of completed phase 2 trials; Number of terminated phase 2 trials; Total number of phase 3 trials sponsored; Number of phase 3 trials with positive results; Number of phase 3 trials with negative results; Number of completed phase 3 trials; Number of terminated phase 3 trials	Numerical
Sponsor type	Academic; Industry, all other pharma; Industry, Top 20 Pharma	Multi-label
Status	Completed; terminated	Binary
Design	Free text	String
Design keywords	Cross over; Double blind/blinded; Efficacy; Multiple arm; Open label; Pharmacodynamics; Pharmacokinetics; Placebo control; Randomized; Single arm	Multi-label
Target accrual	Integer	Numerical
Therapeutic area	Autoimmune/Inflammation; Cardiovascular; CNS; Infectious Disease	Multi-label