Predicting pharmaceutical drug approval using earnings call transcripts

Raymond Chiasson – chiass35 - 1002728337
Zeeshan Hassan – hassa290 - 1002479472
Christopher Holiday – holidayc - 1001345760

University of Toronto
ECO3400
April 12, 2020

Introduction

Accurately predicting the successful development of new pharmaceutical drugs is incredibly valuable for drug developers, investors, medical professionals, and patients. Drugs have unique molecular structures and interact with large biological systems which makes predicting their safety and efficacy very challenging. Only around 10% of drugs that enter clinical development are ultimately approved [1] [2] and it takes on average at least 10 years and $2.5B in development costs to bring a single drug through to regulatory approval [3]. A large part of the cost is paid by the developer and it escalates throughout development [4].

We build on the work of others to try to predict the probability of drug approvals from the perspective of someone outside the company with access only to public information. Accurate predictions of the probability of a drug's success are valuable to three major parties: patients and physicians, competing pharmaceutical companies, and investors. Physicians and their patients are often forward looking; they benefit from knowing what the future standard of care for a condition will be. Physicians may want to refer patients to clinical trials for treatments that they believe will be effective and, for serious diseases, the treatment approach that a physician prescribes could vary if they are aware of a new potentially more effective drug that is likely to be approved.

Pharmaceutical companies compete by promoting their own products or developing new competing therapies. Promotion is a major cost for pharmaceutical companies: marketing costs in the US exceeded $25 billion on total pharmaceutical drug sales of $320 billion in 2016 [5]. A company may look to cut costs when competitors are unlikely to make it to market or alternatively ramp up efforts to preempt threats from new products. Competitors also need to make critical decisions about the development of their own competing products where the direct costs are large and escalating. Costs are on order of $5 million for phase 1 clinical trials, $10-15 million for phase 2 and $20-25 million for phase 3 [6]. Afterwards, there is a $2 million fee to have the FDA evaluate an application and the potential for post-approval studies that can also cost tens of millions of dollars. When two similar products are in development, the developers may want to run larger and more compelling clinical trials to obtain superior clinical data and gain a competitive edge in market. Alternatively,

if a company believes that competing products are likely to succeed and dominate the market before its own, it may be optimal to discontinue development and avoid additional costs. These are large investment decisions that depend directly on beliefs about the likelihood of competing drugs succeeding.

Lastly, investors allocate resources to drug developers according to their ability to successfully develop and market new drugs. Many publicly traded biotech companies have a market capitalization of hundreds of millions of dollars based on only a single product in late-stage clinical development. A company's stock price implies probabilities that their products will succeed, and investors can profit by having better information about the probabilities reflected in prices. More generally, capital will be more efficiently allocated if investors can more accurately predict the success of companies' products.

The standard approach to predicting drug development success has used historical success rates stratified by basic features such as the product's stage of development and disease area [1] [2] [7] [8]. Recently, researchers have turned to machine learning to incorporate additional features and employ more complex models. Initial studies built predictive models with small sets of drugs using numerical or textual descriptions of a drug's clinical trial results [9] [10]. The most extensive effort to date comes from Lo et al. 2019 which incorporates over 140 features from a large proprietary database of over 6,000 unique drugs [11]. A feature that has not yet been incorporated by previous research is the information disclosed during earnings calls and medical conferences by the company developing a given drug. Previous work shows that during these events, executives communicate important information about their companies and products both directly and indirectly [12] [13] [14]. Executives have access to non-public information related to a drug's prospect of success (such as the results of a clinical trial that has not yet been published) and sometimes disclose this information for the first time in these meetings. Even when they do not disclose previously unknown information, they may indirectly communicate their optimism or pessimism though their word choice or sentiment. In this paper we use natural language processing techniques to analyze information conveyed during earnings calls and conference presentations. We use word frequency, sentiment

scores, and a set of basic drug features to classify which drugs succeeded or failed clinical development. Figure 1 illustrates the approach used in this paper.

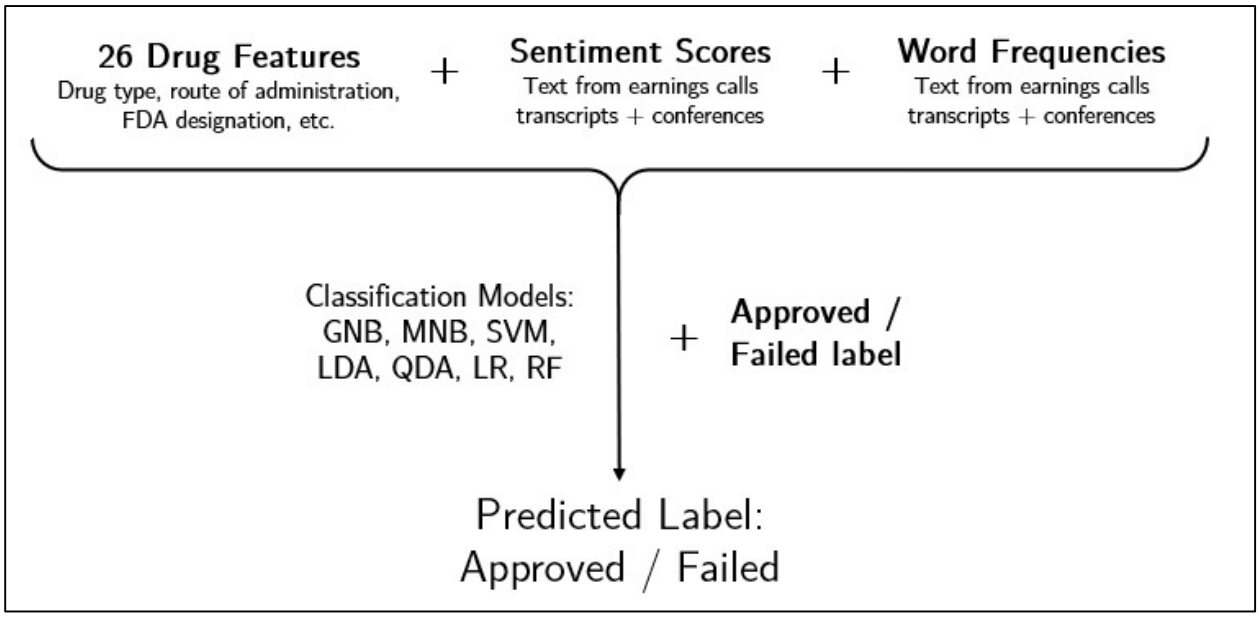

Figure 1. Illustration of the dataset and approach used in this paper.

Literature

Predicting Drug Approval

Classical methods for estimating drug approval probability rely on historical rates of success [1] [2] [7] [8]. They group drugs by a few broad characteristics such as current stage of development, disease area, drug type, and FDA regulatory designations and calculate the expected rate of success. For example, 14.7% of allergy drugs have advanced from phase 1 to approval according to BIO 2016 [2]. These analyses benefit from simple methodology and the availability of data on thousands of drugs that have entered clinical development in recent decades.

In the last several years, researchers have employed machine learning techniques to incorporate more features and build more complex models. Heinemann et al. 2016 use ML-classifiers to predict the approval of 116 cancer drugs in phase 2 and 3 clinical trials. The authors use textual analysis to mine for keywords related to approval or failure from online medical publications. Using both a random forest and decision tree classification methods, they conclude that article count, overall

commitment by researchers, and a strong therapeutic focus are strongly predictive of regulatory approval. DiMasi et al. 2017 predict approval after phase 2 testing in a sample of 62 cancer drugs using logistic regression and other classification techniques. They incorporate features derived from the phase 2 trials: measures of effectiveness (such as response rate), the number of subjects enrolled, the number of subjects treated, and the length of the trials.

Most recently, researchers have used this approach on datasets with thousands of drugs and built additional features. Munos et al. 2021 analyze 4,500 unique drugs and 37 features with 8 classification models [15]. In the largest study to date, Lo et al. 2019 analyze over 6,000 unique drugs with over 140 features. Examples of features include drug medium (capsule, powder, tablet), route of administration (inhaled, oral, injectable), biochemical-specific indicators, and information from clinical trials.

Textual Analysis of Earnings Calls

There is an established literature focusing on textual sentiment in finance (see [16] for a survey). Sentiment analysis has been applied to earnings call transcripts and similar executive engagements to predict a wide variety of financial indicators. For example, firms that take an abnormally positive tone have poorer future earnings [12], firms that call on questions from bullish analysts underperformed peers [13], and firms that excessively mention regulation are unsurprisingly more likely to be impacted by future regulatory decisions [14]. Sentiment analysis has been applied to the pharmaceutical industry in the context of patient drug reviews and overall firm reputation, particularly on social media [17] [18] [19] [20]. To the best of our knowledge, textual analysis has not been applied to earnings calls and medical conferences for the purpose of predicting drug approval.

Data

We begin with a list of 75 drugs that meet the criteria listed in

Table 1. We label a drug as approved if it is on the FDA's list of approved drugs for the given indication. A drug is labelled a failure if any of three events occur. First, if the drug fails to meet one of its primary endpoints in a clinical trial, it is considered a failure. Second, if the drug has completed a clinical trial but has not progressed to the next stage of development after three years have passed, it is labelled as a failure. Lastly, if a drug is rejected by the FDA when it is submitted for approval it is considered a failure and the date of rejection is recorded as its outcome date. We estimate an "outcome date" which corresponds to the date after which the drug is known to have succeeded or failed in clinical development. For approved drugs, this is the date on which approval is announced. For drugs that failed, we use the end date of its last clinical trial. Predictions are made using only information available before the outcome date.

Table 1. Criteria for drugs to be used in this paper.

Criteria	**Basis**
The outcome of clinical development (approved or failed) is known.	Drug label is known.
Developed primarily for a single indication (disease).	Simplify labeling and other analysis.
Developed by a publicly traded company.	Ensures there are earnings call transcripts to analyze.
Analyzed by Lo et al. 2019.	Allow for a direct model performance comparison.

We manually add basic features of the drugs from clinicaltrials.gov for treatments that failed and use the corresponding FDA label for drugs that were approved. This includes the type of drug: small molecule, monoclonal antibody, etc., the route of administration: oral, intravenous, etc., as well as regulatory milestones such as Orphan Drug Designation and Fast-Track Designation. These are designations assigned to drugs the FDA that often make their development eligible for funding and assistance due to their predicted high effectiveness or their indication for a high priority disease

area such as cancers [21]. The last basic feature added for a drug is the market capitalization of the company for the year corresponding to the outcome date. This is given by Bloomberg Professional services and is CPI adjusted to 2021 US dollars.

For word frequency and sentiment analysis we collected a total of 7,346 transcripts from earnings calls and medical conferences from Bloomberg Professional Services. To identify speech most likely to be related to a given drug, we search the transcripts for mentions of the corresponding drug name. We extract any sentence containing the drug name and the following two sentences as shown in Figure 2. These sets of sentences form the corpus used for preprocessing and analysis. We also note the date of the transcript in relation to the drug's outcome date. To test their predictive value over different periods we group transcripts within four time periods: (i) all available transcripts up to the outcome date (whether approval or failure), and all transcripts up to (ii) 1 year, (iii) 3 years, and (iv) 5 years before the outcome date which roughly correspond with the stages of drug development from pre-approval, to phase 3, phase 2 and phase 1.

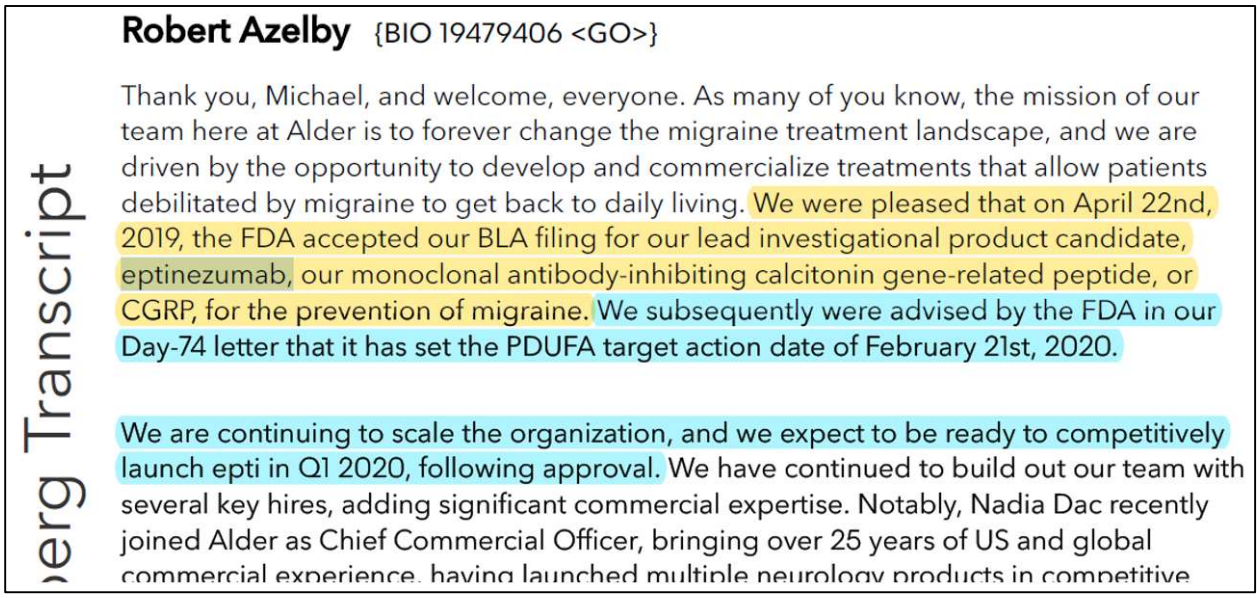

Figure 2. Illustration of text extraction process from a sample earnings call. The sentence containing the drug's name (eptinezumab) is shown in yellow and the following two sentences are shown in blue. The yellow and blue sentences are extracted for further processing.

Methods

To convert the sentences extracted from the transcripts into features we generate word frequencies and sentiment scores. To obtain word frequencies, we split all sentences in the data into individual words dropping both punctuation and common stop-words. Each word is lemmatized to avoid arbitrary distinction between similar words. The frequency of occurrence of these lemmatized words are counted for each treatment in our sample and each becomes a feature for our model.[1] To account for high frequency words which appear across many drugs,[2] each word frequency feature is standardized using term-frequency inverse-document frequency (TF-IDF). This places higher weight on the occurrence of a word if it appears in the sentences corresponding to only a few treatments.

We evaluate sentence sentiment using polarity scores from the Afinn, TextBlob, and VADER lexicons. We calculate sentiment score for the entire text for a given drug and divide by the number of sentences to adjust for the fact that some drugs are discussed more than others.

Prior to running our predictive models, we normalize all features in the data by setting the mean to zero and standard deviation to one. This ensures that our models will not overweight features according to their units during the training and testing steps.

We test 7 common machine learning algorithms on the dataset: Gaussian naïve Bayes (GNB), Multinomial naïve Bayes (MNB), support vector machines (SVM), k-nearest neighbors (KNN), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic regression (LR), random forest (RF). Our dataset contains 3 groups of features: basic drug features (FDA designations, market cap, etc.), word frequencies, and sentiment scores. We test each model on various combinations of feature sets and on word frequencies for the four different time periods. CNB and MNB were used exclusively with the datasets consisting only of word frequencies. KNN

[1] For the full dataset, 1 year prior, 3 years and 5 years, the number of unique lemmatized words are 9635, 8362, 6351 and 4708, respectively.

[2] See Figure A5 and Figure A6 for total word frequencies across drugs which failed or were approved.

was ineffective on the high dimensional feature sets containing word frequencies so it was only trained on a few other feature sets. Otherwise, each model was tested on 13 different feature sets.

We evaluate performance using the 5 common metrics: accuracy, f1 score, precision, recall, and AUC ROC (area under curve of the receiving operator characteristic) using 5-fold cross validation. For models with hyperparameters (SVM, KNN, RF), we perform grid search over a wide range of parameter values, optimize for accuracy, and compute performance using the optimal parameters.

Results

Table 2 shows the performance of models using the entire feature set and transcripts up to the outcome date. Table 3 shows our main result, the performance of models using the entire feature set with transcripts up to one year before the outcome date. These represent predictions that could be made one year before a drug is known to have failed or succeeded. Table 4 and Table 5 show performance on the same time periods but using only the word frequency features. The full results on each model can be found in the appendix Table A7 - Table A14. Figure 3 analyzes model performance for transcripts available at different time periods and Figure 4 compares performance using the basic features with and without sentiment scores.

Table 2. Performance for feature set with basic features, sentiment scores, and word frequencies for all years up to the outcome date.

	Support Vector Machines	**Logistic Regression**	**Random Forest**	**Linear Discriminant Analysis**	**Quadratic Discriminant Analysis**
Accuracy	0.79 (0.08)	0.71 (0.11)	0.84 (0.03)	0.67 (0.12)	0.56 (0.14)
F1	0.73 (0.06)	0.51 (0.17)	0.76 (0.03)	0.41 (0.27)	0.62 (0.11)
Precision	0.73 (0.18)	0.77 (0.29)	0.87 (0.16)	0.51 (0.36)	0.46 (0.11)
Recall	0.79 (0.12)	0.45 (0.18)	0.71 (0.08)	0.35 (0.23)	0.96 (0.08)
ROC AUC	0.80 (0.08)	0.75 (0.08)	0.84 (0.03)	0.64 (0.16)	0.82 (0.12)

Table 3. Performance for feature set with basic features, sentiment scores, and word frequencies for up to 1 year prior to the outcome date.

	Support Vector Machines	**Logistic Regression**	**Random Forest**	**Linear Discriminant Analysis**	**Quadratic Discriminant Analysis**
Accuracy	0.73 (0.06)	0.68 (0.07)	0.80 (0.04)	0.67 (0.08)	0.44 (0.08)
F1	0.60 (0.09)	0.40 (0.11)	0.70 (0.04)	0.39 (0.18)	0.54 (0.06)
Precision	0.70 (0.17)	0.78 (0.27)	0.82 (0.16)	0.61 (0.26)	0.38 (0.05)
Recall	0.57 (0.15)	0.31 (0.11)	0.63 (0.10)	0.33 (0.19)	0.92 (0.10)
ROC AUC	0.64 (0.05)	0.68 (0.06)	0.80 (0.04)	0.59 (0.11)	0.73 (0.07)

Table 4. Performance for feature set using only word frequencies for all years up to the outcome date.

	Support Vector Machines	**Logistic Regression**	**Random Forest**	**Linear Discriminant Analysis**	**Quadratic Discriminant Analysis**	**Gaussian Naïve Bayes**	**Multinomial Naïve Bayes**
Accuracy	0.79 (0.08)	0.73 (0.11)	0.83 (0.05)	0.69 (0.12)	0.40 (0.11)	0.68 (0.11)	0.76 (0.12)
F1	0.73 (0.06)	0.53 (0.28)	0.72 (0.08)	0.57 (0.14)	0.53 (0.06)	0.60 (0.13)	0.60 (0.20)
Precision	0.73 (0.18)	0.62 (0.38)	0.87 (0.17)	0.62 (0.24)	0.37 (0.06)	0.58 (0.18)	0.76 (0.27)
Recall	0.79 (0.12)	0.53 (0.30)	0.63 (0.03)	0.55 (0.09)	0.92 (0.10)	0.67 (0.13)	0.51 (0.18)
ROC AUC	0.81 (0.07)	0.83 (0.05)	0.83 (0.05)	0.73 (0.08)	0.74 (0.14)	0.68 (0.11)	0.65 (0.20)

Table 5. Performance for feature set using only word frequencies for up to 1 year prior to the outcome date.

	Support Vector Machines	**Logistic Regression**	**Random Forest**	**Linear Discriminant Analysis**	**Quadratic Discriminant Analysis**	**Gaussian Naïve Bayes**	**Multinomial Naïve Bayes**
Accuracy	0.72 (0.05)	0.72 (0.05)	0.77 (0.07)	0.68 (0.05)	0.64 (0.03)	0.68 (0.09)	0.72 (0.10)
F1	0.56 (0.09)	0.49 (0.05)	0.63 (0.10)	0.48 (0.07)	0.00 (0.00)	0.48 (0.15)	0.53 (0.16)
Precision	0.70 (0.17)	0.80 (0.24)	0.80 (0.17)	0.63 (0.22)	0.00 (0.00)	0.60 (0.24)	0.70 (0.24)
Recall	0.53 (0.18)	0.37 (0.03)	0.53 (0.10)	0.41 (0.05)	0.00 (0.00)	0.41 (0.14)	0.44 (0.14)
ROC AUC	0.70 (0.03)	0.73 (0.02)	0.77 (0.07)	0.66 (0.10)	0.31 (0.03)	0.68 (0.11)	0.64 (0.16)

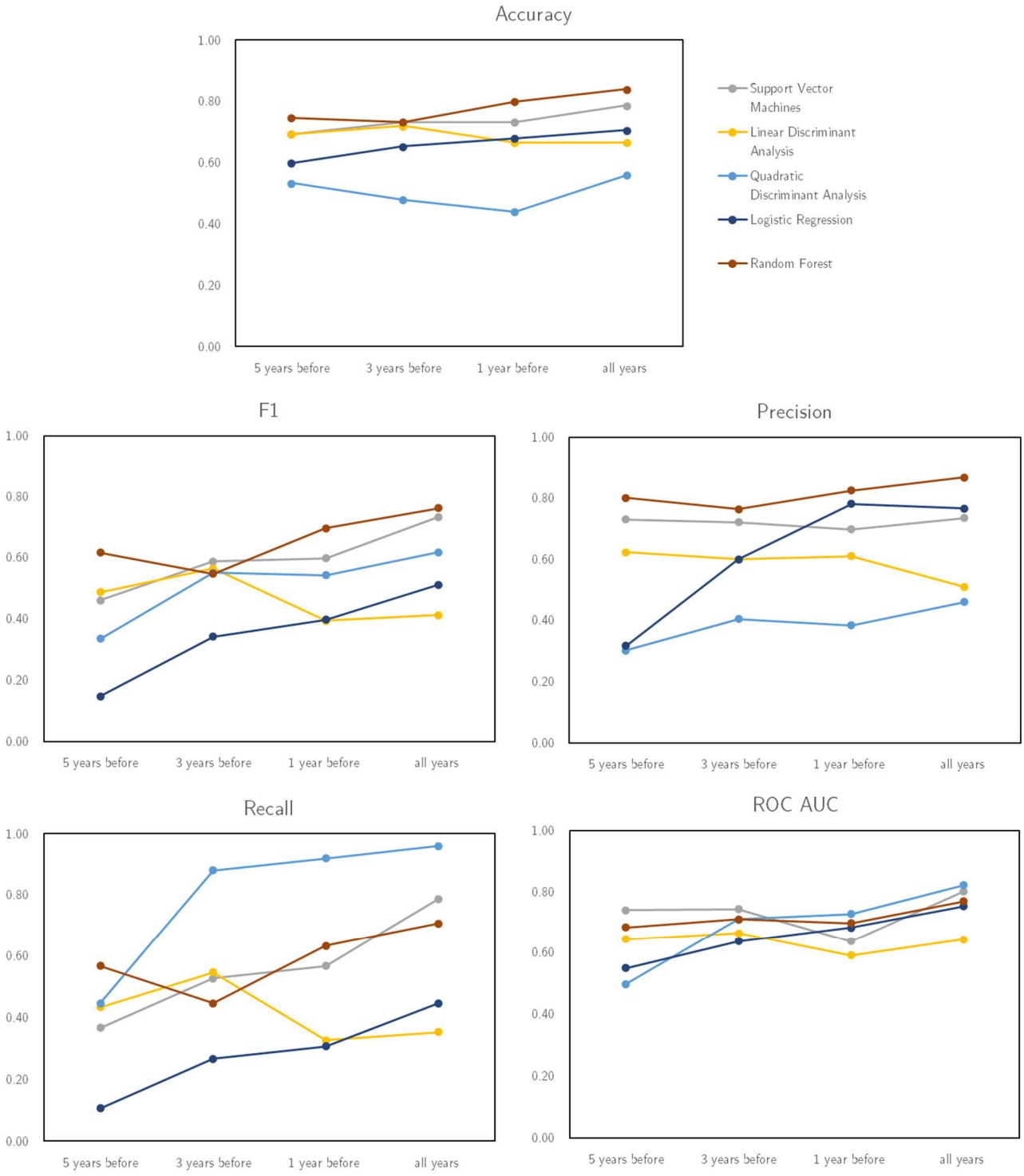

Figure 3. Comparison of performance metrics for transcripts in different time periods. Transcripts with a date up to the specified label are used. The feature set contains basic features, sentiment scores, and word frequencies up the date specified on the chart label.

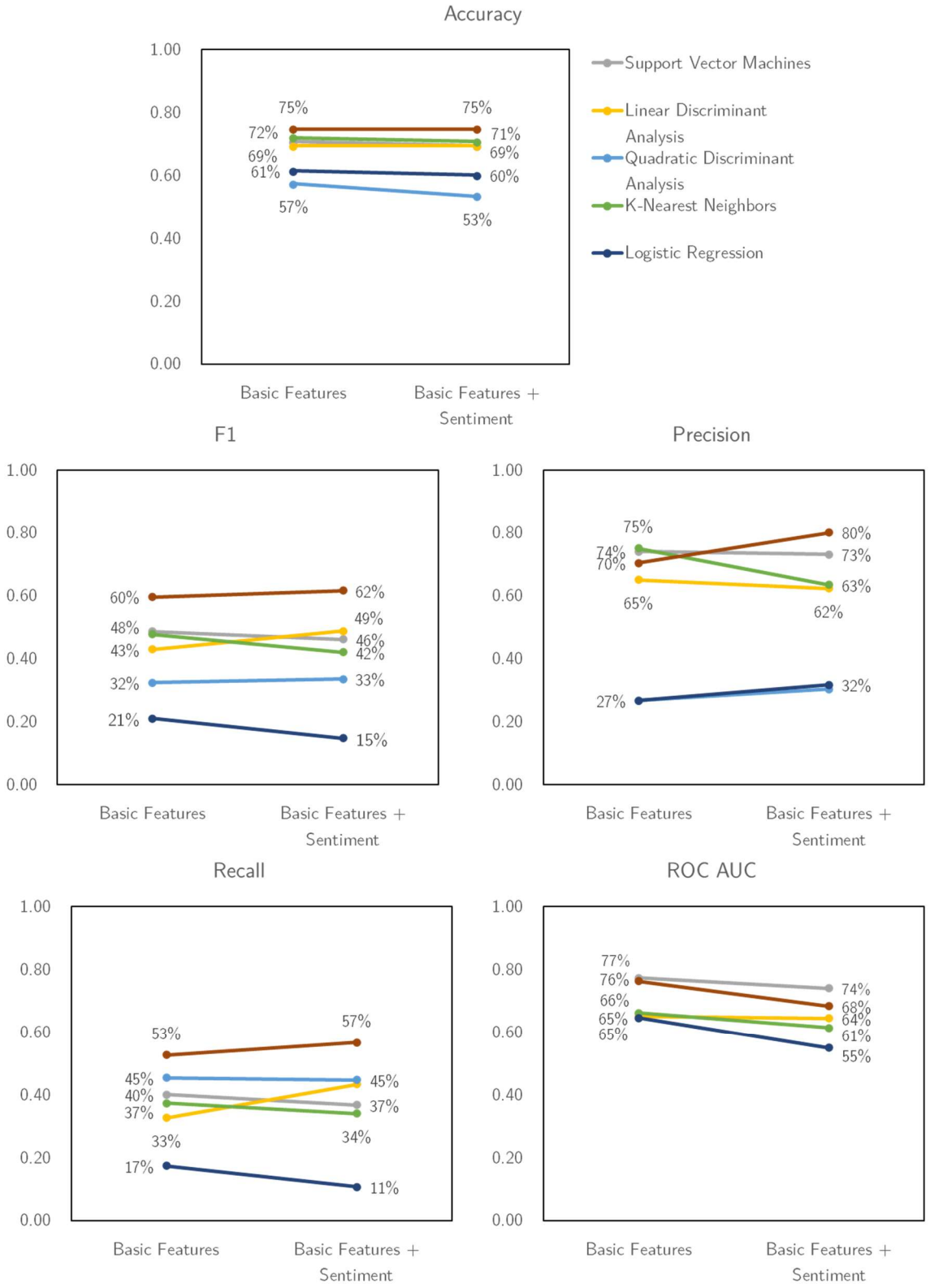

Figure 4. Comparison of performance metrics with of models trained on feature sets containing basic features with and without sentiment scores.

Discussion

The performance of our models on the full feature set (Table 2 and Table 3) and using only word frequencies (Table 4 and Table 5) supports our hypothesis that executives of pharmaceutical companies reveal a significant amount of information pertaining to whether a drug will succeed. Our primary measure for comparing model performance is AUC ROC. For the full feature set up until one year before the outcome date, the highest average AUC ROC of 0.80 was given by RF, followed by QDA with 0.73. Linear models and SVM do not perform as well suggesting that the relationship between all features available up until a year before the outcome date and the likelihood of approval is highly non-linear. This non-linear relationship is something SVM is not able to capture effectively. Interestingly, the QDA classifier achieves a high AUC ROC with a very low accuracy and precision meaning it correctly identifies many drugs that will be approved but also incorrectly classifies drugs that failed. The strong performance of RF is most likely due to its ability to handle such a high dimensional sparse feature space and model key interactions between word frequencies, sentiment scores and basic features.

There is a considerable increase in model accuracy when each model is trained on the entire set of available transcripts (Figure 3). This is expected because there is less information the further away from the outcome date and the information available in earlier periods is likely less predictive.

We hypothesized that the sentiment commentary on a drug would correlate with its likelihood of success. We found, however, that adding sentiment scores to the basic feature set had a generally neutral or negative impact (Figure 4). This could mean that our hypothesis is incorrect, or that our measurements of sentiment were inaccurate or imprecise. Perhaps the single-word sentiment scoring we employed was insufficient and could be improved with more advanced techniques that consider other elements of sentence structure. Alternatively, it is possible that executive sentiment expressed in these transcripts is not genuine as executives might have reason to falsely express neutral or positive sentiment. It could also be that sentiment is not very predictive even when perfectly quantified as it represents an indirect measure of executive opinion which may not be correlated

with approval. Executives may express genuine positive sentiment regarding their own products in development based on inaccurate beliefs on their probability of success.

Table A6 shows the performance scores from other studies that predicted drug approvals with machine learning. Our performance scores are comparable to those produced by Lo et al. 2019 and somewhat lower than Munos et al. 2021 and Dimasi et al. 2017. Most importantly, our results compare favorably to the performance of predictions from Lo et al. 2019 on the same 75 drugs used in our models. The corresponding AUC ROC is 0.72 on this subset of drugs (Figure 7). Note that our data set is concentrated on a considerably different set of features than those used in other papers in Table A6. These models include large amounts of direct information about the drug and its development, while ours contains only basic drug information coupled with the indirect information from textual analysis. The left side of charts in Figure 4 show that models trained on only our basic features (excluding textual features) have good scores on Accuracy, Precision and AUC ROC but low scores on F1 and Recall. This compares unfavorably with studies in Table A6 which have a much richer set of drug features such as clinical trial results. Directly encoding the results of a clinical trial as features (as other studies were able to do) is surely a less noisy signal than a paragraph of text describing the same results.

Conclusion

Our results confirm the hypothesis that transcripts of earnings calls contain valuable information about whether a drug will succeed or fail. The performance of the models showed that the relative frequency of words used in these transcripts were reasonably predictive even in the absence of basic features of the drugs. We find that flexible models which can capture both linear and non-linear relationships are best suited for generating predictions. As expected, the predictive value of these transcripts increases with proximity to the approval or failure date, reflecting the increased amount and quality of available information. We found that sentiment, at least measured using individual word score, is not predictive in the context of our models. Our models performed reasonably well compared to other studies that had access to a much richer set of features and many more observations. In particular, our results compare favorably to predictions on the same set of drugs. Our work shows that earnings calls form a valuable feature set that is publicly available to interested third parties (patients and physicians, competing pharmaceutical companies, and investors). The ideal extension of this work would be to combine it with a rich feature set such as that in Lo et al. 2019 or Munos et al. 2021 to deliver valuable improvements in predictions.

References

[1] M. Hay, D. W. Thomas, J. L. Craighead, C. Economides and R. J., "Clinical development success rates for investigational drugs," *Nature Biotechnology,* vol. 32, no. 11, pp. 40-51, 2014.

[2] BIO, "Clinical Development Success Rates 2006-2015," Biotechnology Innovation Organization, 2016.

[3] J. DiMasi, H. Grabowski and R. Hansen, "Innovation in the pharmaceutical industry: New estimates of R&D costs," *Journal of Health Economics,* vol. 47, pp. 20-33, 2016.

[4] C. Matraves, "Market Structure, R&D and Advertising in the Pharmaceutical Industry," *The Journal of Industrial Economics (47 No. 2),* pp. 169-194, June 1999.

[5] L. Schwartz and S. Woloshin, "Medical marketing in the United States, 1997-2016," *Journal of the American Medical Association,* vol. 321, no. 1, pp. 80-96, 2019.

[6] A. Sertkaya, H. Wong, A. Jessup and T. Beleche, "Key cost drivers of pharmaceutical trials in the United States," *Clinical Trials,* vol. 13, no. 2, pp. 114-126, 2016.

[7] J. DiMasi, L. Feldman, A. Seckler and A. Wilson, "Trends in Risks Associated With New Drug Development: Success Rates for Investigational Drugs," *Clinical Pharmacology & Therapeutics,* vol. 87, no. 3, pp. 272-277, 2010.

[8] K. Smietana, M. Siatkowski and M. Møller, "Trends in clinical success rates," vol. 15, no. 6, pp. 379-380, 2016.

[9] F. Heinemann, T. Huber, C. Meisel, M. Bundschus and U. Leser, "Reflection of successful anticancer drug development processes in the literature," *Drug discovery today,* vol. 21, no. 11, pp. 1740-1744, 2016.

[10] J. A. DiMasi, J. C. Hermann, K. Twyman, R. K. Kondru, S. Stergiopoulos, K. A. Getz and W. Rackoff, "A tool for predicting regulatory approval after phase II testing of new oncology compounds," *Clinical Pharmacology & Therapeutics,* vol. 98, no. 5, pp. 506-513, 2015.

[11] A. W. Lo, K. W. Siah and C. H. Wong, "Machine Learning with Statistical Imputation for Predicting Drug Approvals," *Harvard Data Science Review,* vol. 1, no. 1, 2019.

[12] X. Huang, S. H. Teoh and Y. Zhang, "Tone management," *The Accounting Review,* vol. 89, no. 3, pp. 1083-1113, 2014.

[13] L. Cohen, D. Lou and C. J. Malloy, "Casting conference calls," *Management Science,* vol. 66, no. 11, pp. 5015-5039, 2020.

[14] C. W. Calomiris, H. Mamaysky and R. Yang, "Measuring the cost of regulation: A text-based approach," National Bureau of Economic Research, 2020.

[15] B. Munos, J. Niederreiter and M. Riccaboni, "Improving the Prediction of Clinical Success Using Machine Learning," *medRxiv (published online),* 2021.

[16] C. Kearney and S. Liu, "Textual sentiment in finance: A survey of methods and models," *International Review of Financial Analysis,* vol. 33, pp. 171-185, 2014.

[17] F. Grasser, S. Kallumadi, H. Malberg and S. Zaunseder, "Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning," pp. 121-125, 2018.

[18] H. Isah, P. Trundle and D. Neagu, "Social media analysis for product safety using text mining and sentiment analysis," *2014 14th UK workshop on computational intelligence (UKCI),* 2014.

[19] K. Mahboob and F. Ali, "Sentiment analysis of pharmaceutical products evaluation based on customer review mining," *J. Comput. Sci. Syst. Biol.,* vol. 11, pp. 190-194, 2018.

[20] V. Pampulevski, J. R. Giaquinto, M. Rametta, M. Toscani, J. Barone and J. C. Nadal, "Sentiment of Media Coverage and Reputation of the Pharmaceutical Industry," *Therapeutic innovation & regulatory science,* vol. 54, no. 1, pp. 220-225, 2020.

[21] US FDA, "Fast Track, Breakthrough Therapy, Accelerated Approval, Priority Review," U.S. Food & Drug Administration, 23 February 2018. [Online]. Available: https://www.fda.gov/patients/learn-about-drug-and-device-approvals/fast-track-breakthrough-therapy-accelerated-approval-priority-review. [Accessed 12 April 2021].

Appendix – Figures

Figure A5: Sum of word frequencies across all drugs that were approved. Larger words indicate higher frequency. Drug names are those corresponding to other drugs mentioned by executives not the drugs in the dataset.

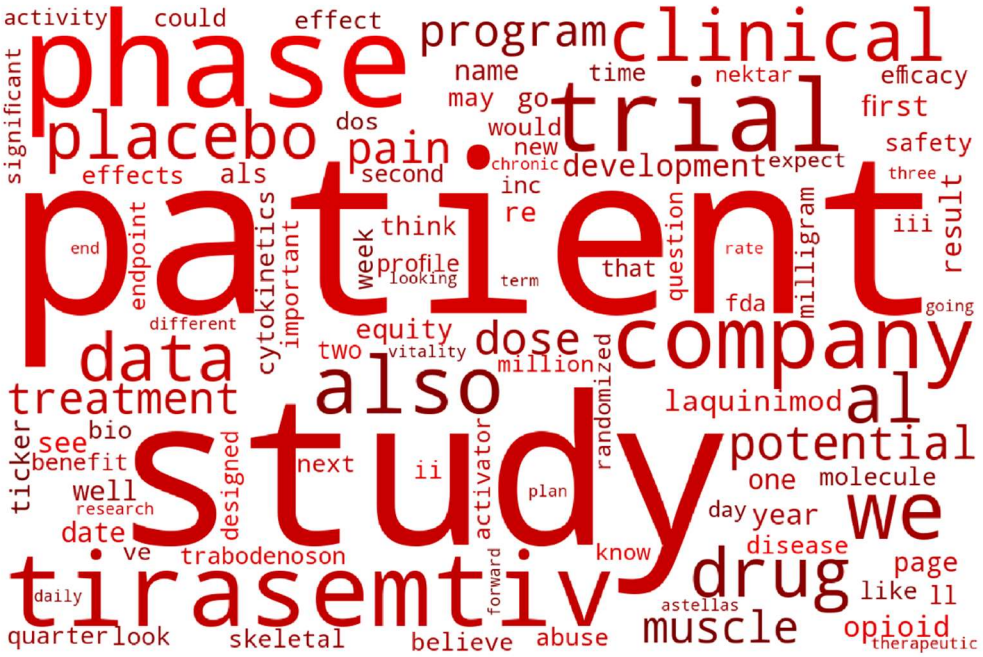

Figure A6: Sum of word frequencies across all drugs that failed. Larger words indicate higher frequency. Drug names are those corresponding to other drugs mentioned by executives not the drugs in the dataset.

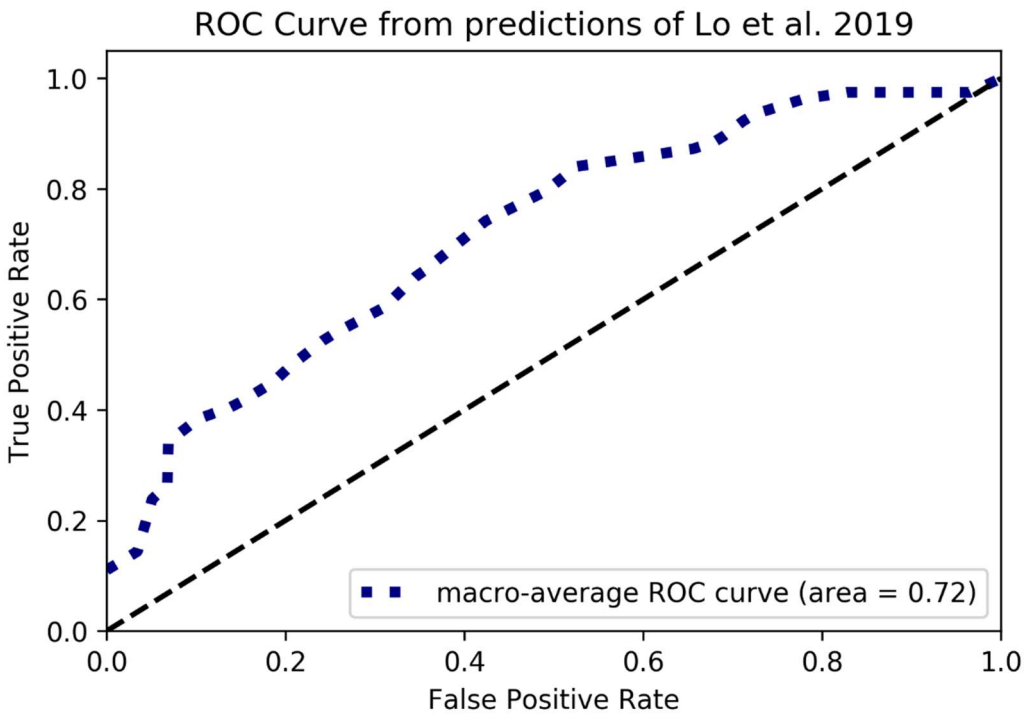

Figure 7: Receiving operator characteristic for predictions from Lo et al. 2019 [11]. Predictive values of eventual approval are rounded two decimal place values from zero to one and are generated based off information available when the drug was in phase II development.

Appendix - Tables

Table A6. Performance of models trained to predict drug approval in the literature.

Authors	Method	Data	AUC	Accuracy	Precision	F1	Recall
Munos et al. (2021)	Bayesian additive regression tree (BART)	Phase I	0.93	0.94	0.89	0.90	0.76
		Phase II	0.96	0.82	0.90	0.86	0.95
		Phase III	0.94	0.88	0.92	0.78	0.84
	Linear Discriminant Analysis (LDA)	Phase I	0.85	0.89	0.82	0.85	0.56
		Phase II	0.88	0.65	0.79	0.71	0.91
		Phase III	0.84	0.87	0.82	0.84	0.60
Lo et al. (2019)	Random Forest with 5 Nearest Neighbors Imputation (5NN-RF)	Phase II	0.78				
		Phase III	0.81				
	Penalized Logistic Regression (PLR)	Phase III	0.81				
	Gradient Boosting Trees (GBT)	Phase III	0.82				
	Support Vector Machine (SVM)	Phase III	0.76				
DiMasi et al. (2017)	ANDI scoring algorithm (cut-off ≥ 5)	Phase III (oncology only)	0.92	0.95			0.74

Table A7. Logistic Regression model performance over all feature sets.

Feature Set														
	Basic features	✓	✓	✓	✓	✓	✓	✓	✓	✓				
	Sentiment scores	All	1 year before	3 years before	5 years before		All	1 year before	3 years before	5 years before				
	Word frequencies	All	1 year before	3 years before	5 years before	5 years before	5 years before				All	1 year before	3 years before	5 years before
Metrics	Accuracy	0.71 (0.11)	0.68 (0.07)	0.65 (0.03)	0.60 (0.07)	0.61 (0.09)	0.65 (0.10)	0.64 (0.10)	0.64 (0.10)	0.63 (0.10)	0.73 (0.11)	0.72 (0.05)	0.67 (0.06)	0.64 (0.03)
	F1	0.51 (0.17)	0.40 (0.11)	0.34 (0.08)	0.15 (0.12)	0.21 (0.22)	0.32 (0.21)	0.37 (0.14)	0.33 (0.14)	0.28 (0.20)	0.53 (0.28)	0.49 (0.05)	0.29 (0.19)	0.00 (0.00)
	Precision	0.77 (0.29)	0.78 (0.27)	0.60 (0.20)	0.32 (0.37)	0.27 (0.28)	0.43 (0.28)	0.57 (0.28)	0.57 (0.28)	0.47 (0.36)	0.62 (0.38)	0.80 (0.24)	0.59 (0.39)	0.00 (0.00)
	Recall	0.45 (0.18)	0.31 (0.11)	0.27 (0.11)	0.11 (0.09)	0.17 (0.18)	0.25 (0.18)	0.29 (0.13)	0.25 (0.12)	0.21 (0.16)	0.53 (0.30)	0.37 (0.03)	0.23 (0.20)	0.00 (0.00)
	ROC AUC	0.75 (0.08)	0.68 (0.06)	0.64 (0.07)	0.55 (0.04)	0.65 (0.14)	0.76 (0.09)	0.69 (0.11)	0.68 (0.12)	0.64 (0.12)	0.83 (0.05)	0.73 (0.02)	0.66 (0.10)	0.43 (0.06)

Table A8. Linear Discriminant Analysis model performance over all feature sets.

Feature Set														
	Basic features	✓	✓	✓	✓	✓	✓	✓	✓	✓				
	Sentiment scores	All	1 year before	3 years before	5 years before		All	1 year before	3 years before	5 years before				
	Word frequencies	All	1 year before	3 years before	5 years before	5 years before	5 years before				All	1 year before	3 years before	5 years before
Metrics	Accuracy	0.67 (0.12)	0.67 (0.08)	0.72 (0.11)	0.69 (0.09)	0.69 (0.07)	0.68 (0.11)	0.65 (0.10)	0.65 (0.08)	0.65 (0.07)	0.69 (0.12)	0.68 (0.05)	0.71 (0.07)	0.63 (0.03)
	F1	0.41 (0.27)	0.39 (0.18)	0.56 (0.21)	0.49 (0.17)	0.43 (0.14)	0.48 (0.21)	0.41 (0.23)	0.43 (0.11)	0.36 (0.19)	0.57 (0.14)	0.48 (0.07)	0.34 (0.20)	0.00 (0.00)
	Precision	0.51 (0.36)	0.61 (0.26)	0.60 (0.19)	0.62 (0.24)	0.65 (0.20)	0.55 (0.22)	0.48 (0.28)	0.55 (0.15)	0.51 (0.33)	0.62 (0.24)	0.63 (0.22)	0.70 (0.40)	0.00 (0.00)
	Recall	0.35 (0.23)	0.33 (0.19)	0.55 (0.24)	0.43 (0.21)	0.33 (0.12)	0.43 (0.20)	0.37 (0.20)	0.37 (0.10)	0.29 (0.15)	0.55 (0.09)	0.41 (0.05)	0.23 (0.14)	0.00 (0.00)
	ROC AUC	0.64 (0.16)	0.59 (0.11)	0.66 (0.18)	0.64 (0.14)	0.65 (0.14)	0.62 (0.16)	0.58 (0.16)	0.66 (0.08)	0.64 (0.10)	0.73 (0.08)	0.66 (0.10)	0.64 (0.11)	0.45 (0.10)

Table A9. Quadratic Discriminant Analysis model performance over all feature sets.

Feature Set														
	Basic features	✓	✓	✓	✓	✓	✓	✓	✓	✓				
	Sentiment scores	All	1 year before	3 years before	5 years before		All	1 year before	3 years before	5 years before				
	Word frequencies	All	1 year before	3 years before	5 years before	5 years before	5 years before				All	1 year before	3 years before	5 years before
Metrics	Accuracy	0.56 (0.14)	0.44 (0.08)	0.48 (0.13)	0.53 (0.12)	0.57 (0.12)	0.63 (0.08)	0.64 (0.08)	0.68 (0.09)	0.65 (0.07)	0.40 (0.11)	0.64 (0.03)	0.64 (0.03)	0.40 (0.10)
	F1	0.62 (0.11)	0.54 (0.06)	0.55 (0.10)	0.33 (0.22)	0.32 (0.27)	0.27 (0.04)	0.40 (0.12)	0.37 (0.18)	0.32 (0.21)	0.53 (0.06)	0.00 (0.00)	0.00 (0.00)	0.41 (0.21)
	Precision	0.46 (0.11)	0.38 (0.05)	0.40 (0.09)	0.30 (0.18)	0.27 (0.23)	0.62 (0.32)	0.66 (0.29)	0.71 (0.29)	0.50 (0.33)	0.37 (0.06)	0.00 (0.00)	0.00 (0.00)	0.28 (0.14)
	Recall	0.96 (0.08)	0.92 (0.10)	0.88 (0.16)	0.45 (0.36)	0.45 (0.39)	0.19 (0.02)	0.38 (0.23)	0.29 (0.19)	0.29 (0.23)	0.92 (0.10)	0.00 (0.00)	0.00 (0.00)	0.80 (0.40)
	ROC AUC	0.82 (0.12)	0.73 (0.07)	0.71 (0.13)	0.50 (0.16)	0.00 (0.00)	0.47 (0.14)	0.61 (0.13)	0.52 (0.18)	0.48 (0.18)	0.74 (0.14)	0.31 (0.03)	0.00 (0.00)	0.00 (0.00)

Table A10. Gaussian Naïve Bayes model performance over all feature sets.

Feature Set	Word frequencies	All	1 year before	3 years before	5 years before
Metrics	Accuracy	0.68 (0.11)	0.68 (0.09)	0.39 (0.07)	0.39 (0.10)
	F1	0.60 (0.13)	0.48 (0.15)	0.41 (0.13)	0.49 (0.09)
	Precision	0.58 (0.18)	0.60 (0.24)	0.31 (0.09)	0.35 (0.07)
	Recall	0.67 (0.13)	0.41 (0.14)	0.62 (0.23)	0.81 (0.13)
	ROC AUC	0.68 (0.11)	0.62 (0.10)	0.43 (0.09)	0.48 (0.09)

Table A11. Multinomial Naïve Bayes model performance over all feature sets.

Feature Set	Word frequencies	All	1 year before	3 years before	5 years before
Metrics	Accuracy	0.76 (0.12)	0.72 (0.10)	0.68 (0.05)	0.61 (0.07)
	F1	0.60 (0.20)	0.53 (0.16)	0.48 (0.10)	0.28 (0.17)
	Precision	0.76 (0.27)	0.70 (0.24)	0.65 (0.20)	0.38 (0.22)
	Recall	0.51 (0.18)	0.44 (0.14)	0.45 (0.16)	0.23 (0.15)
	ROC AUC	0.65 (0.20)	0.64 (0.16)	0.63 (0.07)	0.53 (0.07)

Table A12. K-nearest Neighbor model performance and optimal hyperparameters over all feature sets.

Feature Set	Basic features	✓	✓	✓	✓	✓	✓	✓	✓	✓				
	Sentiment scores	All	1 year before	3 years before	5 years before		All	1 year before	3 years before	5 years before				
	Word frequencies	All	1 year before	3 years before	5 years before	5 years before	5 years before				All	1 year before	3 years before	5 years before
Metrics	Accuracy	0.64 (0.03)	0.65 (0.05)	0.71 (0.05)	0.71 (0.09)	0.72 (0.03)	0.75 (0.10)	0.71 (0.14)	0.73 (0.12)	0.69 (0.07)	0.64 (0.03)	0.64 (0.03)	0.64 (0.03)	0.64 (0.03)
	F1		0.07 (0.13)	0.38 (0.10)	0.42 (0.24)	0.48 (0.10)	0.52 (0.22)	0.47 (0.27)	0.48 (0.27)	0.47 (0.19)				
	Precision		0.20 (0.40)	0.83 (0.21)	0.63 (0.34)	0.75 (0.13)	0.72 (0.19)	0.60 (0.34)	0.67 (0.38)	0.58 (0.09)				
	Recall		0.04 (0.08)	0.26 (0.09)	0.34 (0.22)	0.37 (0.11)	0.44 (0.25)	0.41 (0.26)	0.41 (0.26)	0.44 (0.25)				
	ROC AUC	0.58 (0.08)	0.56 (0.07)	0.58 (0.13)	0.61 (0.10)	0.66 (0.11)	0.71 (0.13)	0.66 (0.20)	0.65 (0.18)	0.67 (0.13)				
Hyper-parameters	n_neighbors	2	4	3	5	7	3	3	3	3	2	2	2	2

Table A13. Random Forest model performance and optimal hyperparameters over all feature sets.

Feature Set	Basic features	✓	✓	✓	✓	✓	✓	✓	✓	✓				
	Sentiment scores	All	1 year before	3 years before	5 years before		All	1 year before	3 years before	5 years before				
	Word frequencies	All	1 year before	3 years before	5 years before	5 years before	5 years before				All	1 year before	3 years before	5 years before
Metrics	Accuracy	0.84 (0.03)	0.80 (0.04)	0.73 (0.07)	0.75 (0.08)	0.75 (0.05)	0.88 (0.05)	0.83 (0.07)	0.77 (0.09)	0.73 (0.07)	0.83 (0.05)	0.77 (0.07)	0.69 (0.09)	0.64 (0.03)
	F1	0.76 (0.03)	0.70 (0.04)	0.55 (0.10)	0.62 (0.09)	0.60 (0.08)	0.78 (0.12)	0.70 (0.12)	0.60 (0.16)	0.57 (0.17)	0.72 (0.08)	0.63 (0.10)	0.37 (0.26)	0.00 (0.00)
	Precision	0.87 (0.16)	0.82 (0.16)	0.76 (0.23)	0.80 (0.24)	0.70 (0.06)	1.00 (0.00)	0.90 (0.12)	0.82 (0.15)	0.65 (0.10)	0.87 (0.17)	0.80 (0.17)	0.60 (0.39)	0.00 (0.00)
	Recall	0.71 (0.08)	0.63 (0.10)	0.45 (0.09)	0.57 (0.15)	0.53 (0.10)	0.66 (0.16)	0.59 (0.14)	0.49 (0.18)	0.53 (0.21)	0.63 (0.03)	0.53 (0.10)	0.31 (0.27)	0.00 (0.00)
	ROC AUC	0.84 (0.03)	0.80 (0.04)	0.73 (0.07)	0.75 (0.08)	0.75 (0.05)	0.88 (0.05)	0.83 (0.07)	0.77 (0.09)	0.73 (0.07)	0.83 (0.05)	0.77 (0.07)	0.69 (0.09)	0.64 (0.03)
Hyper-parameters	Criterion	gini	gini	entropy	gini	entropy	gini	entropy	entropy	gini	entropy	entropy	entropy	entropy
	n_estimators	10	3	4	10	10	5	7	3	7	4	4	5	1
	max_depth	sqrt	None	None	None	None	sqrt	None	None	None	None	None	sqrt	sqrt
	max_features	50	50	10	10	50	100	10	10	500	300	100	10	10

Table A14. Support Vector Machines model performance and optimal hyperparameters over all feature sets.

Feature Set	Basic features	✓	✓	✓	✓	✓	✓	✓	✓	✓				
	Sentiment scores	All	1 year before	3 years before	5 years before		All	1 year before	3 years before	5 years before				
	Word frequencies	All	1 year before	3 years before	5 years before	5 years before	5 years before				All	1 year before	3 years before	5 years before
Metrics	Accuracy	0.79 (0.08)	0.73 (0.06)	0.73 (0.11)	0.69 (0.08)	0.71 (0.09)	0.76 (0.12)	0.72 (0.14)	0.72 (0.08)	0.71 (0.09)	0.79 (0.08)	0.72 (0.05)	0.68 (0.08)	0.64 (0.03)
	F1	0.73 (0.06)	0.60 (0.09)	0.59 (0.16)	0.46 (0.11)	0.49 (0.14)	0.58 (0.24)	0.60 (0.19)	0.52 (0.12)	0.47 (0.14)	0.73 (0.06)	0.56 (0.09)	0.42 (0.23)	0.00 (0.00)
	Precision	0.73 (0.18)	0.70 (0.17)	0.72 (0.24)	0.73 (0.25)	0.74 (0.25)	0.73 (0.20)	0.60 (0.17)	0.70 (0.16)	0.75 (0.23)	0.73 (0.18)	0.70 (0.17)	0.53 (0.33)	0.00 (0.00)
	Recall	0.79 (0.12)	0.57 (0.15)	0.53 (0.16)	0.37 (0.10)	0.40 (0.15)	0.55 (0.30)	0.63 (0.25)	0.41 (0.10)	0.37 (0.13)	0.79 (0.12)	0.53 (0.18)	0.39 (0.25)	0.00 (0.00)
	ROC AUC	0.80 (0.08)	0.64 (0.05)	0.74 (0.15)	0.74 (0.10)	0.77 (0.13)	0.82 (0.06)	0.76 (0.11)	0.76 (0.15)	0.75 (0.12)	0.81 (0.07)	0.70 (0.03)	0.65 (0.08)	0.52 (0.08)
Optimal hyper-parameters	Gamma	1	10	10000	10	10	1000	10000	0.001	0.001	1	10	1	0.001
	C	0.01	0.01	0.001	0.001	0.001	10	1	0.001	0.001	0.01	0.01	10	0.001
	Kernel	rbf	rbf	rbf	linear	linear	rbf	rbf	linear	linear	rbf	rbf	rbf	linear

Table A15. Illustrative extracted sentences and corresponding sentiment scores.

Transcript Text	**Afinn polarity**	**TextBlob polarity**	**TextBlob sentiment**
We see the launch of PCSK9, or you see the launch of ertugliflozin, which we believe is a best-in-class molecule and coming into a revitalized positioning in the marketplace.	3.0	0	0
30 seconds on this slide -- huge market potential for anything that comes out to help patients, including an agent such as Azeliragon, which is being targeted as disease modifying.	2.0	0.13	0.80
Now, early in the study what was noted for the 20-milligram dose was concentration-related reversible cognitive worsening.	-3.0	0.43	0.73
We have strong support from both FDA and EMA in terms of the appropriateness of the program we have launched	5.0	0.43	0.73
But all in all, a very, very strong neurology franchise in the United States, which is, of course, also the basis for us being able to launch our psychiatry products in the U.S. market and in the longer perspective, also being able to execute a launch of our Alzheimer's product idalopirdine a few years from now, assuming data supports an approval.	7.0	0.34	0.58
And I can't comment specifically on why they have chosen to stop the development.	-1.0	0	0
As for the progress status of compounds under development since July, JTT-851, a compound for Type 2 diabetes mellitus entered into phase II overseas.	2.0	0	0
And crisaborole's known mechanism of action, already used systemically in other agents, has the potential to have a very favorable safety profile, which is particularly important in pediatric populations, which is a very big part of the overall atopic dermatitis population.	6.0	0.08	0.42