HDSR

Issue 1

Machine learning with statistical imputation for predicting drug approvals‡

Andrew W. Lo[1,2,3]*, Kien Wei Siah[1,2], Chi Heem Wong[1,2]

[1] Laboratory for Financial Engineering, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

[2] EECS and CSAIL, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

[3] Santa Fe Institute, Santa Fe, New Mexico, United States of America

* Corresponding author

email: alo-admin@mit.edu (A.W.L.)

Abstract

We apply machine-learning techniques to predict drug approvals using drug-development and clinical-trial data from 2003 to 2015 involving several thousand drug-indication pairs with over 140 features across 15 disease groups. To deal with missing data, we use imputation methods that allow us to fully exploit the entire dataset, the largest of its kind. We show that our approach outperforms complete-case analysis, which typically yields biased inferences. We achieve predictive measures of 0.78, and 0.81 AUC ("area under the receiver operating characteristic curve," the estimated probability that a classifier will rank a positive outcome higher than a negative outcome) for predicting transitions from phase 2 to approval and phase 3 to approval, respectively. Using five-year rolling windows, we document an increasing trend in the predictive power of these models, a consequence of improving data quality and quantity. The most important features for predicting success are trial outcomes, trial status, trial accrual rates, duration, prior approval for another indication, and sponsor track records. We provide estimates of the probability of success for all drugs in the current pipeline.

Keywords: Drug development; Clinical trials; Healthcare; Machine learning; Imputation

‡ We thank Informa for providing us access to their data and expertise and are particularly grateful to Christine Blazynski, Mark Gordon, and Michael Hay for many helpful comments and discussion throughout this project. We also thank them and Anna Barker, Linda Blackerby, Lara Boro, Steve Finch, Howard Fingert, Xiao-Li Meng, Ellen Moore, Jim Reddoch, Lara Sullivan, Marty Tenenbaum, James Wade, and three anonymous referees for specific comments on this manuscript, and Jayna Cummings and Paige Sammartino for editorial assistance. Research support from the MIT Laboratory for Financial Engineering is gratefully acknowledged. 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.

Author contributions: Conceptualization, data-license agreement, A.W.L.; Methodology, A.W.L., K.W.S. and C.H.W.; Investigation, K.W.S.; Writing (first draft), K.W.S.; Writing (review, editing, final draft), K.W.S. and C.H.W.; Visualization, K.W.S.; Project supervision, A.W.L.

1 Introduction

While many recent medical breakthroughs such as immuno-therapies, gene therapies, and gene-editing techniques offer new hope for patients, they have also made biomedical innovation riskier, and more complex and expensive. These breakthroughs generate novel therapies for investigation, each of which requires many years of translational research and clinical testing, costing hundreds of millions to billions of dollars and yet often face a high likelihood of failure (Fernandez, Stein, & Lo, 2012). In fact, drug development productivity—the ratio of the number of new drugs approved to R&D spending each year—has declined steadily over the past 50 years despite scientific and technical progress. This phenomenon, which Scannell, Blanckley, Boldon, and Warrington (2012) termed "Eroom's Law," as the reverse of Moore's Law, suggests that the cost of developing new drugs has doubled approximately every nine years since the 1950s. In the face of multiple uncertainties, the need to evaluate drug candidates better and allocate capital to high-potential opportunities more efficiently has only intensified.

To address these needs, in this article we apply machine-learning techniques to predict the outcomes of randomized clinical trials. Machine learning is an interdisciplinary field focused on tackling pattern recognition problems and building predictive models to make data-driven decisions, which is well-suited for this context. Successful applications of these techniques have already revolutionized a number of industries (e.g., advertising, marketing, finance and insurance, oil and gas exploration) and are poised for even greater impact via autonomous vehicles, facial-recognition authentication, and general-purpose robotics.

Drug developers have already applied machine-learning tools to the discovery process via high-throughput screening of vast libraries of chemical and biological compounds to identify drug targets. However, in managing their portfolios of investigational drugs, biopharma companies typically use unconditional estimates of regulatory approval rates based on historically observed relative frequencies. Machine learning techniques yield *conditional* estimates of success, conditioned on a host of predictive factors known to affect the likelihood of approval, including drug compound characteristics, clinical trial design, previous trial outcomes, and the sponsor track record. We show that these features contain useful signals about drug development outcomes that will allow us to forecast the outcome of pipeline developments more accurately.

Our methodology and results have several implications for stakeholders in the biomedical ecosystem. More accurate forecasts of the likelihood of success of clinical trials will reduce the uncertainty surrounding drug development, which will increase the amount of capital that investors and drug developers are willing to allocate to this endeavor. By extension, this would lower the cost of capital and increase the efficiency of the allocation process. Specifically, we predict the probability of success of drug candidates in two scenarios: (1) advancing from phase 2 to regulatory approval and (2) from phase 3 to regulatory approval (see Fig 1). Investors and drug developers may use such predictions to evaluate the risks of different investigational drugs at different clinical stages, providing them with much-needed transparency. Greater risk transparency is one source of improved financial efficiency because it facilitates more accurate matching of investor risk preferences with the risks of biomedical investment opportunities.

Machine-learning models can also offer guidance to scientists, clinicians, and biopharma professionals as to which factors are most important in determining clinical-trial success, suggesting ways to improve the drug development process and decelerate or reverse Eroom's Law.

Policymakers and regulators would also benefit from machine-learning predictions, particularly for drug-indication pairs that are predicted to fail with high likelihood—these cases highlight the most difficult challenges in biomedicine and underscore the need for greater government and philanthropic support.

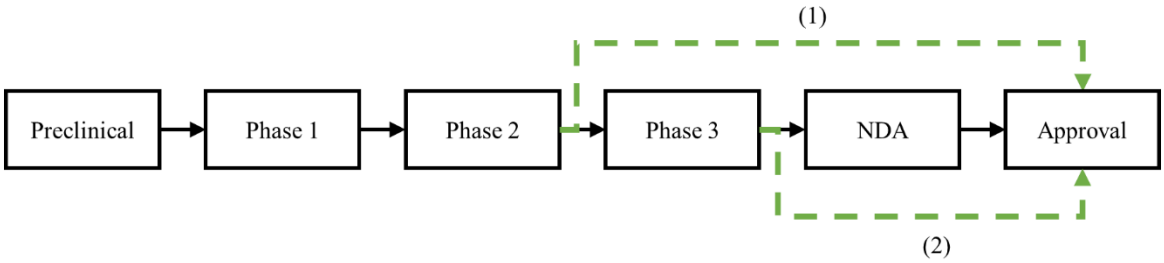

Fig 1. Predictive models for assessing the probability of approval of drug candidates in two scenarios: (1) after phase 2 testing, and (2) after phase 3 testing.

To the best of our knowledge, our study is the largest of its kind. We construct two datasets, one for each scenario, from two proprietary pharmaceutical pipeline databases, *Pharmaprojects* and *Trialtrove* provided by Informa® (Informa, 2016). The phase-2-to-approval dataset includes more than 6,000 unique drugs for 288 indications and over 14,500 phase 2 trials, and the phase-3-to-approval dataset contains more than 1,400 unique drugs for 253 indications and over 4,500 phase 3 trials. These data cover over 15 indication groups. In contrast, most published research on drug approval prediction have very small sample

sizes, are concentrated on specific therapeutic areas, and involve only one or a small number of predictive factors: Malik et al. (2014) examined the trial objective responses of 88 anticancer agents in phase 1; Goffin, Baral, Tu, Nomikos, and Seymour (2005) studied the tumor response rates of 58 cytotoxic agents in 100 phase 1 trials and 46 agents in 499 phase 2 trials; El-Maraghi and Eisenhauer (2008) looked at the objective responses of 19 phase 2 anticancer drugs in 89 single agent trials; Jardim, Groves, Breitfeld, and Kurzrock (2017) examined the response rates of 80 phase 3 oncology drugs to identify factors associated with failures; and DiMasi et al. (2015) analyzed 62 cancer drugs and proposed an approved new drug index (ANDI) algorithm with four factors to predict approval for lead indications in oncology after phase 2 testing (see Supplementary Materials H for a comparison of our analysis to theirs).

Another key difference in our approach is that we deal with missing data using statistical imputation methods. We explore four common approaches to "missingness" and demonstrate their advantages and disadvantages over discarding incomplete cases. With the FDA Amendments Act of 2007, drug and clinical trial data collection has been rapidly expanding, but these data are often sparse, and our dataset is no exception. Related studies (e.g., DiMasi et al., 2015) have typically used only complete-case observations—discarding clinical trials with any missing information—which typically eliminates large portions of the data and may also lead to certain biases.

We use machine-learning techniques to form our predictions, including cross-validation for training and a held-out testing set for performance evaluation, and use the standard "area under the receiver operating characteristic curve" (AUC) metric to measure model

performance (AUC is the estimated probability that a classifier will rank a positive outcome higher than a negative outcome [Fawcett, 2006]). We achieve AUCs of 0.78 for predicting phase 2 to approval (95% confidence interval (CI): [0.75, 0.81]) and 0.81 for predicting phase 3 to approval (95% CI: [0.78, 0.83]). A time-series, walk-forward analysis approach shows similar results. We also apply our models to the current drug pipeline—that is, all drugs still in development as of the end of our dataset—to identify the candidates that have the highest and lowest probabilities of success. We examine the latest development statuses of these pipeline drug-indication pairs—a true "out-of-sample" experiment (validation on data not used in model building)—and find that candidates with higher scores are, indeed, more likely to progress to later clinical stages. This indicates that our classifiers do discriminate between high- and low-potential candidates.

2 Materials and methods

Data

The commercial data vendor Informa® offers two databases that are used in our analysis: *Pharmaprojects*, which specializes in drug information, and *Trialtrove*, which specializes in clinical trials information (Informa, 2016). These two databases aggregate drug and trial information from over 30,000 data sources in more than 150 countries, including company press releases, government drug databases (e.g., Drugs@FDA) and trial databases (e.g., Clinicaltrials.gov [Zarin, Tse, Williams, & Carr, 2016], Clinicaltrialsregister.eu [extracted from EudraCT]), and scientific conferences and publications. Using these sources, we construct two datasets of drug-indication pairs: phase 2 to approval (P2APP) and phase 3 to approval (P3APP). We extract clinical trial features from *Trialtrove*, and augment this data using drug features from *Pharmaprojects*. Applying machine-learning algorithms to these

datasets allows us to estimate: (1) whether a drug-indication pair that has concluded phase 2 testing will be approved eventually; and (2) whether a pair that has concluded phase 3 testing will be approved eventually. Data cleaning procedures are outlined in Supplementary Materials A.

We consider all indications associated with a particular drug, as opposed to only the lead indication. We extract all features that could conceivably correlate with the likelihood of success, from drug compound attributes (31 features from *Pharmaprojects* profiles) to clinical trial characteristics (113 features from *Trialtrove*). These features are defined in Table 1 and Supplementary Materials A. In general, each dataset may be partitioned into two disjoint subsets: one with samples that have known outcomes, and another with samples that are still in the pipeline at the time of snapshot of the databases (that is, the outcomes are unknown). To provide intuition for the characteristics of the samples, we describe key summary statistics of each subset.

The P2APP dataset consists of 6,344 drug-indication pairs that have ended phase 2 testing; that is, there are no phase 2 trials in progress or planned in the database. The phase 2 trials in this dataset range from August 8, 1990 to December 15, 2015. In our sample, 4,812 pairs have known outcomes, while 1,532 pairs are still in the pipeline. In the subset with known outcomes, we define the development statuses of suspension, termination, and lack of development as "failures" (86.8%), and registration and launch as "successes" or approvals (13.2%). The P3APP dataset consists of 1,870 pairs that have ended phase 3 testing, of which 1,610 pairs have known outcomes, while 260 pairs are still in the pipeline. For those pairs with known outcomes, we define "failures" (59.1%) and "successes" (40.9%) in the same

fashion as the P2APP dataset. The phase 3 trials in P3APP span from January 1, 1988 to November 1, 2015. These figures are summarized in Table 2. Here, the use of terms "success" and "failure" is in the context of achieving approval. We note that our definition of "failures" can include drug development programs that are terminated due to factors unrelated to the performance of the drug (e.g., market conditions, business decisions). In Section 3 Results, we find that this outcome variable has significant associations with trial performance and other factors.

The datasets cover 15 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. Anti-cancer agents make up the largest subgroup in P2APP, and the second largest in P3APP (see Table 3). Industry-sponsored trials dominate both datasets (see Table 4). In aggregate, we observe a decreasing trend in success rates over five-year rolling windows from 2003 to 2015 (see Fig 2).

To the best of our knowledge, this sample is the largest of its kind. All prior published research in this literature involved fewer than 100 drugs or 500 trials (DiMasi et al., 2015; El-Maraghi & Eisenhauer, 2008; Goffin et al., 2005; Malik et al., 2014). In addition, our datasets cover a diverse set of indication groups, as opposed to a single area such as oncology.

Table 1. Description of parent features extracted from *Pharmaprojects* and *Trialtrove*. Some parent features are multi-label (e.g., a trial may be tagged with United States and United Kingdom simultaneously). We transform all multi-label parent features into binary child features (1 or 0). See Supplementary Materials A for specific examples of each feature. Note that drug-indication pairs for the same drug have the same drug features; drug-indication pairs involved in the same trial have the same trial features.

	Description	Type
Drug Features		
Route	Route of administration of the drug, the path by which the drug is taken into the body.	Multi-label
Origin	Origin of the active ingredient in the drug.	Multi-label
Medium	Medium of the drug.	Multi-label
Biological target family	Family of proteins in the body whose activity is modified by the drug, resulting in a specific effect.	Multi-label
Pharmacological target family	Mechanism of action of the drug, the biochemical interaction through which the drug produces its pharmacological effect.	Multi-label
Drug-indication development status	Current phase of development of the drug for the indication.	Binary
Prior approval of drug for another indication	Approval of the drug for another indication prior to the indication under consideration (specific to drug-indication pair).	Binary
Trial Features		
Duration	Duration of the trial (from reported start date to end date) in days.	Continuous
Study design	Design of the trial (keywords).	Multi-label
Sponsor type	Sponsors of the trial grouped by types.	Multi-label
Therapeutic area	Therapeutic areas targeted by the trial.	Multi-label
Trial status	Status of the trial.	Binary
Trial outcome	Results of the trial.	Multi-label
Target accrual	Target patient accrual of the trial.	Continuous
Actual accrual	Actual patient accrual of the trial.	Continuous
Locations	Locations of the trial by country.	Multi-label
Number of identified sites	Number of sites where the trial was conducted.	Continuous
Biomarker involvement	Type of biomarker involvement in the trial.	Multi-label
Sponsor track record	Sponsor's success in developing other drugs prior to the drug-indication pair under consideration.	Continuous
Investigator experience	Primary investigator's success in developing other drugs prior to the drug-indication pair under consideration.	Continuous

Table 2. Sample sizes of P2APP and P3APP datasets. We consider phase 2 trial information in P2APP datasets and phase 3 trial information in P3APP dataset.

	Counts				
	Drug-indication Pairs	Phase 2/3 Trials	Unique Drugs	Unique Indications	Unique Phase 2/3 Trials
P2APP					
Success	635	2,563	540	173	2,486
Failure	4,177	10,328	2,779	263	9,722
Pipeline	1,532	2,815	1,189	221	2,713
Total	**6,344**	**15,706**	**4,073**	**288**	**14,584**
P3APP					
Success	659	1,830	572	171	1,801
Failure	951	2,425	764	203	2,360
Pipeline	260	494	240	120	480
Total	**1,870**	**4,749**	**1,451**	**253**	**4,552**

Table 3. Breakdown of drug-indication pairs by indication groups. A drug-indication pair may have multiple indication group tags. For instance, renal cancer is tagged as both anti-cancer and rare disease in *Pharmaprojects*.

	Count	
	P2APP	P3APP
All	6,344	1,870
Anti-cancer	2,239	409
Rare Diseases	1,105	259
Neurological	1,069	444
Alimentary	757	249
Immunological	474	101
Anti-infective	493	177
Respiratory	428	134
Musculoskeletal	394	121
Cardiovascular	388	158
Dermatological	254	45
Genitourinary	210	85
Blood and Clotting	160	97
Sensory	137	41
Hormonal	17	4
Anti-parasitic	8	0

Table 4. Breakdown of trials by sponsor types. A trial may be sponsored by more than one party (e.g., collaboration between industry developers and academia).

	Counts	
	P2APP	P3APP
All	14,584	4,552
Other Pharma	5,432	1,721
Top 20 Pharma	5,322	2,369
Academic	4,869	736
Government	1,807	314
Cooperative Group	958	230
Not for Profit	181	51
Generic	52	54
Contract Research Organization	41	17

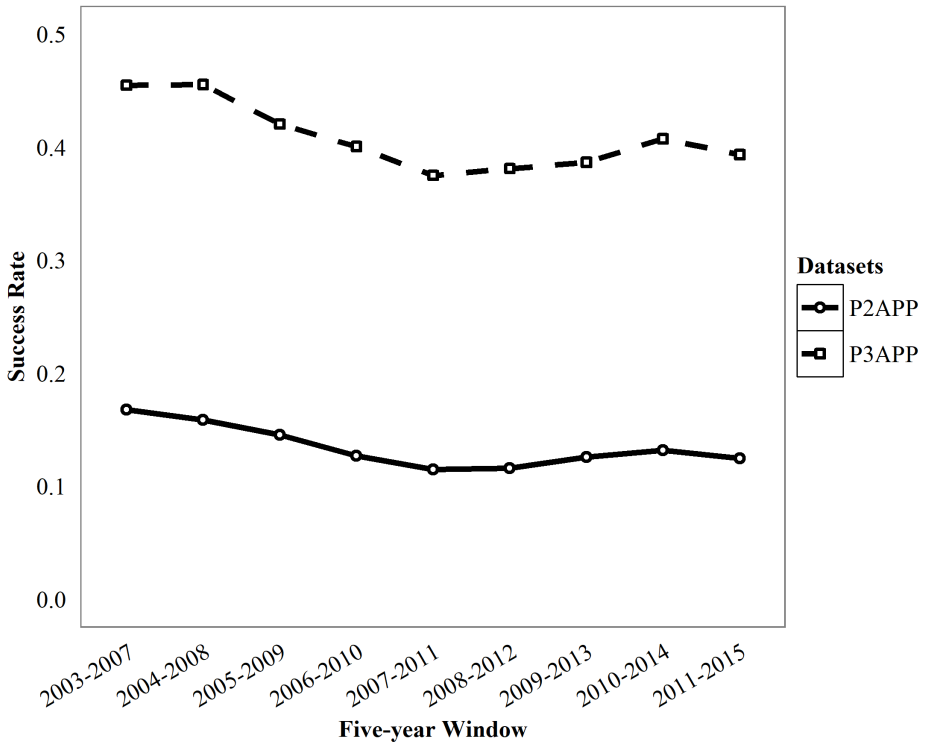

Fig 2. Success rates in P2APP and P3APP over five-year rolling windows from 2003-2015.

Missing data

Prior to the 2007 FDA Amendments Act (FDAAA), it was not uncommon for investigators to release only partial information about pipeline drugs and clinical trials to protect trade secrets or simply because there was no incentive to do more. Even today, some investigators still do not adhere to the FDAAA-mandated registration policy or submit adequate registrations. Therefore, all historical drug development databases have missing data. We note that the "missingness" here is largely related to the post-study reporting of clinical trial data as opposed to in-trial data missingness (e.g., censorship of panel data due to patients terminating trial participation prematurely). In the former case, the data (e.g., trial duration, trial outcomes) are usually available to the investigators but may not be released publicly, and are thus considered "missing" from our standpoint. Therefore, our dataset may be

considered an approximation of the actual data available to the FDA and individual drug developers.

Fig 3, Fig 4, Table 5, and Table 6 summarize the patterns of missingness in our dataset (we exclude pipeline drug-indication pairs here because their outcomes are still pending). The missing data patterns are multivariate. When conditioned on the latest level of development, for any indication, we find that successful drugs generally have lower levels of missingness compared to failed drugs. For instance, in the P2APP dataset, 61% of failed drugs have an unknown medium, while only 15% of approved drugs are missing this feature. We also observe that completed trials tend to have greater levels of missingness than terminated trials. Between two datasets, we find that the P3APP dataset, which focuses on phase 3 drugs and trials, generally has less missing data for both drug and trial features than the P2APP dataset which focuses on phase 2 drugs and trials. This is expected since phase 3 trials are primarily used to support registration filings.

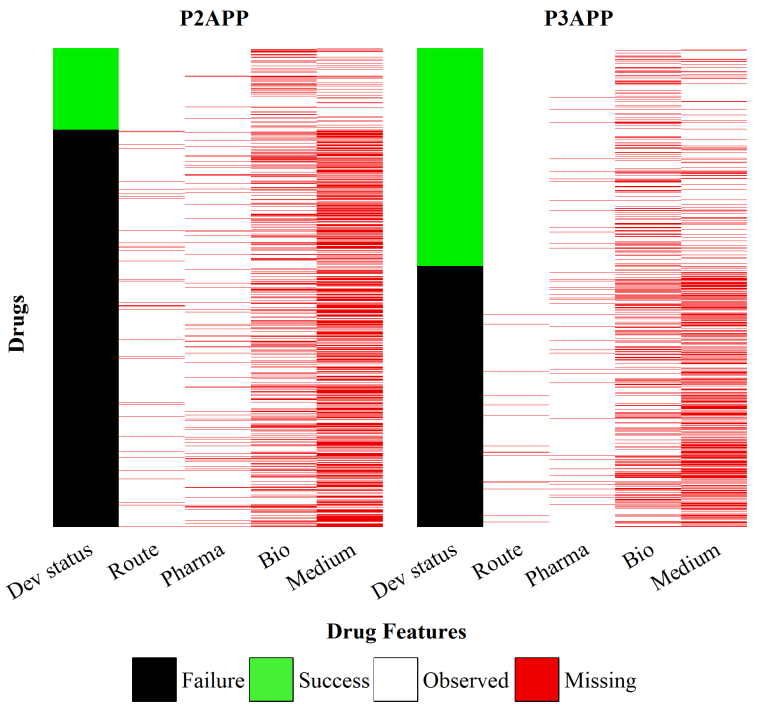

Fig 3. Missingness patterns of drug features. Each row corresponds to a unique drug. Features not included in the figure are complete and do not have missing values. Abbreviations: Dev status: highest level of development of a drug for any indication; Pharma: pharmacological target family; Bio: biological target family.

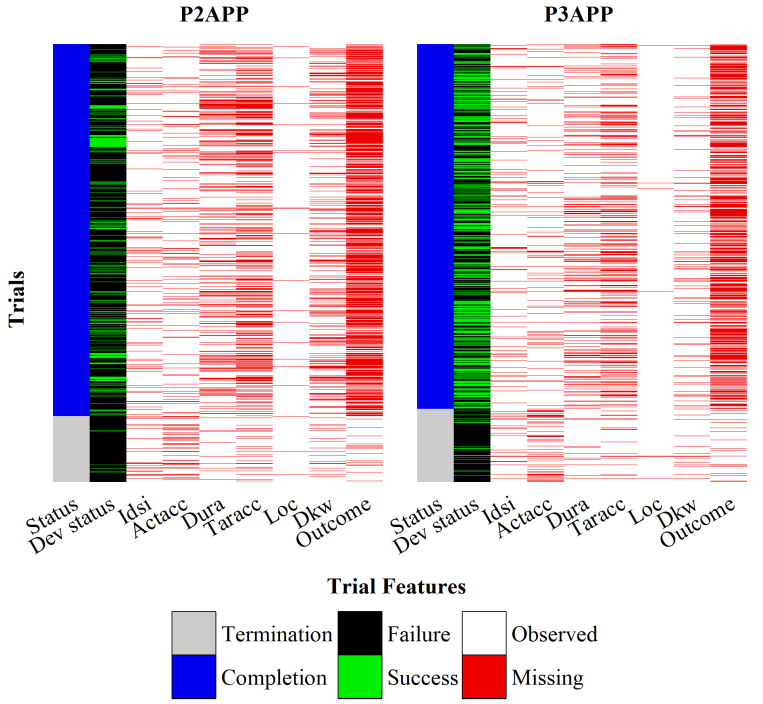

Fig 4. Missingness patterns of trial features. Each row corresponds to a unique clinical trial. Features not included in the figure are complete and do not have missing values. Abbreviations: Dev status: highest level of development of a drug for any indication; Status: trial status; Idsi: number of identified sites; Actacc: actual accrual; Dura: duration; Taracc: target accrual; Loc: locations; Dkw: trial study design keywords; Outcome: trial outcomes.

Table 5. Missingness in drug features with respect to unique drugs (see Fig 3). The column heading "Unconditional" refers to overall missingness without conditioning on outcome.

	Missingness		
	Unconditional	**Success**	**Failure**
P2APP			
Route	0.04	0.00	0.04
Pharmacological target family	0.06	0.02	0.07
Biological target family	0.32	0.27	0.32
Medium	0.53	0.15	0.61
P3APP			
Route	0.01	0.00	0.02
Pharmacological target family	0.03	0.02	0.04
Biological target family	0.27	0.24	0.30
Medium	0.35	0.14	0.54

Table 6. Missingness in trial features with respect to unique trials (see Fig 4). The column heading "Unconditional" refers to overall missingness without conditioning on trial status.

	Missingness		
	Unconditional	**Completion**	**Termination**
P2APP			
Number of identified sites	0.10	0.10	0.10
Actual accrual	0.12	0.10	0.22
Duration	0.26	0.29	0.05
Target accrual	0.37	0.42	0.09
Locations	0.02	0.02	0.02
Study design keywords	0.22	0.24	0.10
Trial outcomes	0.63	0.73	0.11
P3APP			
Number of identified sites	0.10	0.09	0.12
Actual accrual	0.12	0.09	0.26
Duration	0.17	0.19	0.06
Target accrual	0.27	0.31	0.09
Locations	0.01	0.01	0.02
Study design keywords	0.09	0.09	0.06
Trial outcomes	0.53	0.62	0.07

Most related studies do not report the extent of missing data in their samples, presumably because smaller datasets were used. DiMasi et al. (2015) reported missing data for some of their factors, and addressed it through listwise deletion—deleting all observations with any missing factors. Since statistical estimators often require complete data, this approach is the simplest remedy for missingness. However, it greatly reduces the amount of data available and decreases the statistical power of the resulting statistics. Furthermore, listwise deletion is valid only under strict and unrealistic assumptions (see below), and when such conditions

are violated, inferences are biased. In the current study, we make an effort to include in our analysis all observed examples, with or without complete features, through the use of statistical imputation.

Missing data may be classified into three categories (Rubin, 1976): missing completely at random (MCAR), missing at random (MAR), and missing not-at-random (MNAR). MCAR refers to data that are missing for reasons entirely independent of the data; MAR applies when the missingness can be fully accounted for by the observed variables; and MNAR refers to situations when neither MCAR nor MAR is appropriate, in which case the probability of missingness is dependent on the value of an unobserved variable (Van Buuren, 2012). See Supplementary Materials B for the precise definitions of each type of missingness.

If the missingness is MCAR, the observed samples can be viewed as a random subsample of the dataset. Consequently, using listwise deletion should not introduce any bias. While convenient, this assumption is rarely satisfied in practice. In most drug-development databases, failed drugs are more likely to have missing features than successful drugs (see Table 5). Clearly, MCAR does not hold.

Applying listwise deletion when the missingness is not MCAR can lead to severely biased estimates. Moreover, given the nature of drug-development reporting, a large portion of the original data may be discarded if many variables have missing values. For these reasons, the listwise-deletion approach adopted by DiMasi et al. (2015) and others is less than ideal.

Given only the observed data, it is impossible to test for MAR versus MNAR (Enders, 2010). However, our knowledge of the data-collection process suggests that MAR is a plausible starting point, and we hypothesize that the missingness in drug and trial features is mainly

accounted for by drug development and trial statuses respectively. Our observations in Table 5 and Table 6 support this approach, as the missingness proportions for some features differ greatly depending on the outcome.

Our assumption of MAR is consistent with the data-collection methodology in the Informa® databases. Drug profiles are built up over time in *Pharmaprojects*. As a drug advances to later phases, information about its characteristics becomes more readily available because investigators release more data about pipeline drugs after each phase of clinical testing. Informa® inputs this information into its databases as they become available in the public domain or through primary research. Approved drugs are more likely to have more complete profiles, while information about failed drugs tends to stay stagnant because no further studies are conducted. It is very plausible that the MAR nature of our datasets is an artifact of data collection, and by extension, so are similar pharmaceutical datasets extracted from the public domain and maintained in the same fashion. Originally intended to track drug and trial activities, *Pharmaprojects* and *Trialtrove* are not structured to keep track of information updates over time since there was no use for it. Without timestamps of the updates, we are not able to eliminate the MAR artifact from our datasets.

In our analysis, we impute the missing data under the more plausible MAR assumption to obtain complete datasets. In contrast to listwise deletion, we fill in missing values using information in the observed variables. This allows us to utilize data that would otherwise be discarded. Thereafter, we can apply all the usual statistical estimators to this imputation-completed data.

Methods

Our analysis consists of two parts that we perform in R version 3.2.3. First, we impute missing values to generate complete datasets. Next, we apply a range of machine-learning algorithms to build predictive models based on the imputed data. Illustration of the specific components of our analysis appear in Fig 5.

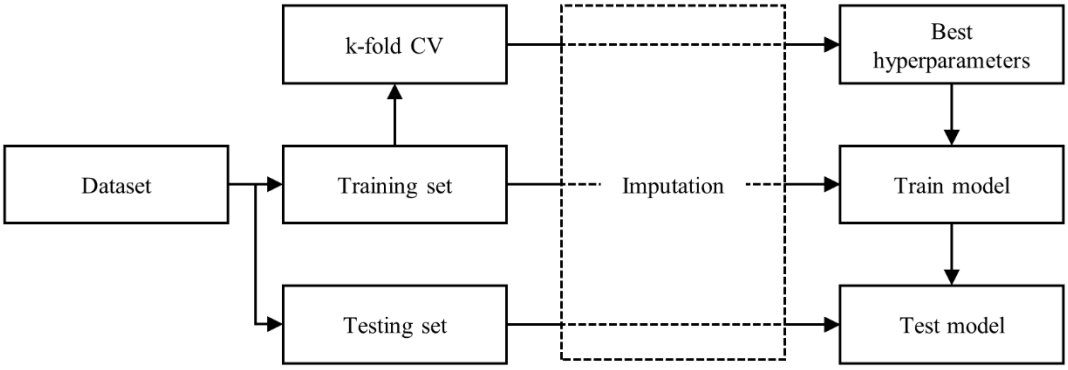

Fig 5. Modeling methodology adopted in this study. Abbreviations: CV: cross-validation.

We formulate our two scenarios as supervised bipartite ranking problems, where the goal is to predict the outcome—success or failure—of a drug-indication pair given a set of input features. Initially, we split each dataset into training and testing sets. For each scenario, we train various classifiers based on the corresponding training set, and compute the expected error of our predictive models by testing them on the held-out testing set.

We create feature matrices from the datasets by representing drug and trial features for each drug-indication pair as vectors (see Fig 6). Drug-indication pairs associated with multiple trials are represented by the same number of feature vectors, e.g., a pair with two trials has two rows. We give a concrete example in Fig 6. Consider the drug-indication pair Analiptin-diabetes type 2 in the P2APP dataset. We represent it using two vector rows since it has two phase 2 trials in *Trialtrove*. Note that the feature matrix is incomplete due to missing drug

and trial features. We also construct a column vector of labels, which contains the outcomes of the drug-indication pairs. Labels are not available for pipeline drug-indication pairs because they are still in development and their outcomes are still uncertain, hence these observations are not used to train our classifiers. However, with the trained classifiers, we can generate predictions for pipeline data.

We split each dataset (excluding pipeline drugs-indication pairs) into two disjoint sets, one training set and one testing set, and form feature matrices for both according to the drug-indication pairs in each set. The testing sets serve as out-of-sample datasets to evaluate our models. Therefore, we mask their outcomes (that is, we treat them as unknown) and access them only at the very end to check our performance.

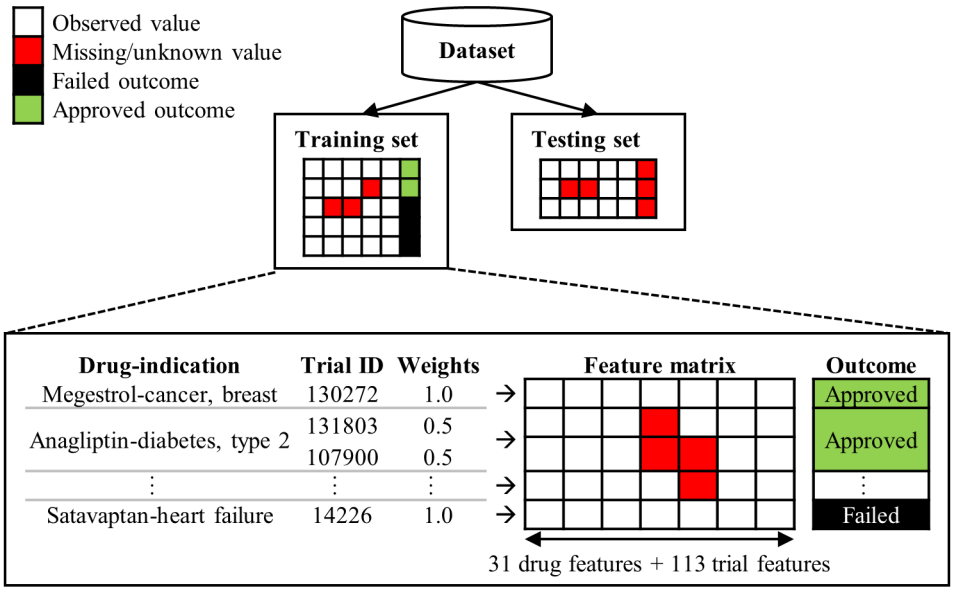

Fig 6. Feature matrix of dataset. Each row corresponds to a feature vector; each feature corresponds to an entry in the vector; each vector has a length of 144 since we have 31 drug and 113 trial features. Feature vectors of all drug-indication pairs in the dataset form the feature matrix collectively. Trial ID is a unique trial identifier in *Trialtrove*.

To deal with missing data in both training and testing sets, we consider listwise deletion and four statistical imputation techniques commonly used in social science research and biostatistics: unconditional mean imputation, k-nearest neighbor (kNN) imputation, multiple imputation (MI), and decision-tree algorithms (see Supplementary Materials C for details). We follow best practices of the missing-data literature by including as many relevant auxiliary variables as possible, as well as all variables used in subsequent models (Collins, Schafer, & Kam, 2001; Enders, 2010; Rubin, 1996; Schafer & Graham, 2002). This makes the assumption of MAR more plausible in our datasets, and helps to reduce bias in subsequent analyses (Schafer, 1997). In particular, it is necessary to include our target variable—the drug-indication development status—in our imputation model because we hypothesize that missingness is mainly accounted for by it. This is not an issue for the training sets. However, the outcomes in the testing sets are masked, and not supposed to be known. Therefore, we treat the testing set outcomes as though they were missing and impute them together with all the other missing features. After imputation, we discard the imputed testing-set outcomes, and use only the imputed feature values for predictions. We do the same when evaluating pipeline datasets.

With respect to the machine-learning algorithm, we explore several linear and non-linear classifiers commonly used in this literature: penalized logistic regression (PLR), random forests (RF), neural networks (NN), gradient boosting trees (GBT), support vector machines with radial basis functions (SVM), and decision trees C5.0. We implement the first five algorithms using the scikit-learn package in Python (Pedregosa et al., 2011) and the sixth using the C50 package in R (Kuhn, Weston, Coulter, & Quinlan, 2014). For training, we weight each feature matrix row example according to the number of trials of the corresponding

drug-indication pair. In our earlier example, the drug-indication pair Analiptin-diabetes type 2 was involved in two phase 2 trials. It is represented by two vector rows in the feature matrix (see Fig 6). Both rows are used as training examples, and each is weighted equally during training (0.5, since there are two trials in total). To obtain predictions for a drug-indication pair, we average the output probabilities and scores of the corresponding feature vector rows that are used as inputs to the classifier.

All machine-learning algorithms have hyper-parameters that affect the flexibility of the model and must be tuned to each dataset to optimize goodness of fit. Poorly-chosen hyper-parameters can lead to overfitting (attributing signal to noise) or underfitting (attributing noise to signal). We tune our parameters using k-fold cross-validation (with k = 5 or 10, depending on the sample size). Since the cross-validation process should emulate the testing process as closely as possible, we include imputation in the cross-validation loop as well. We split the training set into validation and non-validation folds. Then we treat validation fold outcomes as missing, and impute them as we would for a testing set. From here, we ignore the imputed validation fold outcomes and proceed with the standard validation process.

In the final step, we test the trained classifiers on the unseen testing sets for out-of-sample model validation. This gives the expected performance of our predictive models for each of the scenarios, using the standard AUC metric to measure model performance.

3 Results

Simulation of listwise deletion versus imputation

We study the effects of imputation using a "gold-standard" dataset derived from the complete cases of the P2APP dataset (see Table 7). To simulate the missingness present in the original dataset, we introduce missingness in the gold-standard dataset based on our MAR assumption and the missingness patterns observed in the P2APP dataset. We randomly split the drug-indication pairs into a training set (70%) and a testing set (30%), and use five different missing data approaches, as described in Supplementary Materials C, to generate complete training sets from the MAR training set. We use each imputed training set to build six different predictive models (PLR, RF, NN, GBT, SVM, and C5.0) according to the methodology outlined in Section 2. We repeat this experiment 100 times for robustness. Table 8 summarizes the AUC performance of the classifiers on the gold-standard testing sets. See Supplementary Materials E for a more detailed description and results.

Table 7. Sample size of the gold-standard dataset (derived from complete cases of P2APP).

	Counts				
	Drug-indication Pairs	**Phase 2 Trials**	**Unique Drugs**	**Unique Indications**	**Unique Phase 2 Trials**
Success	166	341	152	83	337
Failure	812	1,672	503	158	1,549
Total	**978**	**2,013**	**623**	**171**	**1,872**

Table 8. AUC of different classifiers under different missing data approaches. Abbreviations: Avg: average; Sd: standard deviation; 5%: 5th percentile; 50%: median; 95%: 95th percentile; m: number of imputations generated.

Imputation Method	Machine-learning Model	Gold-Standard Testing Set AUC				
		Avg	Sd	5%	50%	95%
Gold-Standard	PLR	0.810	0.028	0.761	0.808	0.853
Complete Cases		0.755	0.040	0.683	0.764	0.813
Mean/mode		0.778	0.031	0.729	0.779	0.823
Median/mode		0.778	0.031	0.728	0.779	0.824
5NN		0.786	0.032	0.738	0.787	0.834
10NN		0.787	0.032	0.739	0.791	0.835
MI (m=1)		0.781	0.036	0.722	0.777	0.843
MI (m=10)		0.782	0.031	0.729	0.782	0.831
Gold-Standard	RF	0.837	0.027	0.793	0.837	0.876
Complete Cases		0.764	0.048	0.685	0.772	0.830
Mean/mode		0.775	0.031	0.726	0.771	0.822
Median/mode		0.774	0.031	0.723	0.774	0.827
5NN		0.805	0.033	0.755	0.805	0.857
10NN		0.802	0.033	0.747	0.805	0.856
MI (m=1)		0.797	0.033	0.748	0.795	0.853
MI (m=10)		0.804	0.030	0.751	0.804	0.848
Gold-Standard	NN	0.800	0.032	0.754	0.799	0.849
Complete Cases		0.715	0.043	0.638	0.716	0.779
Mean/mode		0.790	0.037	0.739	0.789	0.848
Median/mode		0.789	0.036	0.740	0.792	0.849
5NN		0.794	0.032	0.743	0.798	0.842
10NN		0.797	0.036	0.737	0.798	0.851
MI (m=1)		0.780	0.036	0.719	0.781	0.838
MI (m=10)		0.795	0.030	0.750	0.795	0.838
Gold-Standard	GBT	0.820	0.028	0.776	0.821	0.868
Complete Cases		0.746	0.050	0.659	0.756	0.816
Mean/mode		0.781	0.034	0.724	0.784	0.826
Median/mode		0.778	0.033	0.719	0.783	0.823
5NN		0.796	0.029	0.737	0.798	0.837
10NN		0.796	0.028	0.748	0.798	0.838
MI (m=1)		0.796	0.031	0.747	0.796	0.847
MI (m=10)		0.804	0.031	0.757	0.803	0.854
Gold-Standard	SVM	0.785	0.030	0.730	0.786	0.831
Complete Cases		0.733	0.053	0.650	0.741	0.795
Mean/mode		0.766	0.036	0.707	0.771	0.818
Median/mode		0.764	0.035	0.711	0.771	0.818
5NN		0.771	0.034	0.722	0.770	0.827
10NN		0.772	0.037	0.710	0.773	0.825
MI (m=1)		0.760	0.035	0.696	0.762	0.813
MI (m=10)		0.768	0.030	0.719	0.764	0.813
Gold-Standard	C5.0	0.800	0.033	0.758	0.800	0.844
Complete Cases		0.710	0.063	0.585	0.713	0.802
Mean/mode		0.758	0.039	0.698	0.762	0.816
Median/mode		0.754	0.043	0.679	0.751	0.823
5NN		0.772	0.038	0.715	0.772	0.843
10NN		0.770	0.035	0.710	0.771	0.822
MI (m=1)		0.758	0.037	0.701	0.754	0.819
MI (m=10)		0.807	0.031	0.756	0.808	0.857

For all six machine-learning algorithms, we find that gold-standard classifiers—that is, the models derived from complete data—consistently outperform their complete-case analysis and imputation counterparts. This is logical because useful information is invariably lost when we introduce missingness in the datasets. In contrast, complete-case analysis often leads to inferior performance. The AUCs of classifiers trained on complete-cases training sets tend to be smaller than those trained on imputed training sets. This suggests that imputation does indeed offer improved fit and predictive power over listwise deletion.

Overall, we find kNN imputation to be most compatible with our datasets. It provides the least biased imputations among all missing data methods (see Supplementary Materials E). In particular, the combination of kNN imputation (k = 5) with RF gives one of the highest gold-standard testing set AUCs (0.81). We note a few other MI combinations that yield comparable or marginally better performance but focus on the 5NN-RF approach in subsequent analyses on the main datasets due to its ease of implementation and application. We find that SVM has the worst performance among all machine-learning models. This is not surprising because SVMs are aimed only at learning binary classifiers, and do not generally produce good class probability estimates. Consequently, such models do not necessarily give high AUCs.

We also compare our approach with the ANDI algorithm (DiMasi et al., 2015) by applying a modified version of the index on oncology drugs in the gold-standard testing sets (see Supplementary Materials H for a more in-depth description). We find that our 5NN-RF model achieves significantly higher AUC than the modified ANDI, with an average improvement of 0.1 in AUC over 100 simulations (see Fig 7). We believe that this gain can be attributed to a

larger training set with a wider range of features, a nonlinear model that can capture the complex relationships in the data, and a proper model validation methodology.

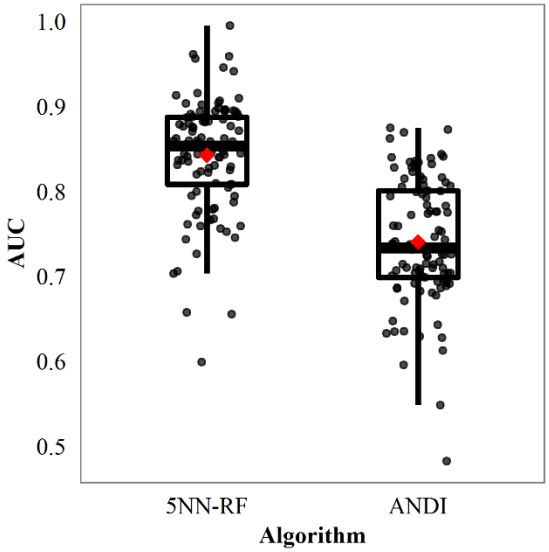

Fig 7. Distributions of AUC of 5NN-RF and the modified ANDI on oncology-only gold-standard testing sets.

Predicting approvals

We analyze the two datasets (P2APP and P3APP) by first splitting each into a training set (70%) and a testing set (30%) randomly (pipeline drug-indication pairs are omitted since their outcomes have yet to be determined). Subsequently, we train 5NN-RF models for each scenario according to the methodology outlined in Section 2. We repeat this experiment 100 times for robustness. Table 9 summarizes the AUC performance metrics for the testing sets. On average, we achieve 0.78 AUC for P2APP and 0.81 AUC for P3APP.

Table 9. Comparison of the general and indication-group specific classifiers for selected indication groups. Abbreviations: Avg: average; Sd: standard deviation; 5%: 5th percentile; 50%: median; 95%: 95th percentile.

	General Classifier					Specialized Classifiers				
	Avg	**Sd**	**5%**	**50%**	**95%**	**Avg**	**Sd**	**5%**	**50%**	**95%**
P2APP										
All	0.777	0.017	0.749	0.775	0.806	-	-	-	-	-
Anti-cancer	0.805	0.025	0.764	0.805	0.847	0.818	0.029	0.773	0.819	0.865
Rare Diseases	0.800	0.028	0.756	0.800	0.848	0.775	0.036	0.715	0.777	0.838
Neurological	0.767	0.036	0.710	0.769	0.819	0.778	0.039	0.721	0.779	0.834
Alimentary	0.749	0.045	0.672	0.751	0.817	0.732	0.048	0.651	0.734	0.807
Immunological	0.783	0.065	0.665	0.786	0.889	0.766	0.069	0.646	0.775	0.860
Anti-infective	0.735	0.043	0.673	0.736	0.800	0.750	0.047	0.684	0.746	0.832
Respiratory	0.756	0.055	0.648	0.764	0.835	0.867	0.043	0.794	0.872	0.921
Musculoskeletal	0.822	0.049	0.736	0.821	0.899	0.731	0.076	0.614	0.745	0.849
Cardiovascular	0.709	0.072	0.580	0.711	0.812	0.694	0.073	0.579	0.698	0.807
Genitourinary	0.633	0.086	0.503	0.634	0.790	0.706	0.091	0.552	0.710	0.840
P3APP										
All	0.810	0.018	0.781	0.810	0.834	-	-	-	-	-
Anti-cancer	0.783	0.047	0.699	0.779	0.853	0.707	0.054	0.612	0.714	0.786
Rare Diseases	0.819	0.054	0.727	0.822	0.896	0.786	0.058	0.687	0.793	0.875
Neurological	0.796	0.037	0.734	0.794	0.857	0.789	0.038	0.741	0.787	0.853
Alimentary	0.817	0.047	0.744	0.820	0.891	0.805	0.054	0.718	0.808	0.888
Immunological	0.811	0.074	0.680	0.815	0.910	0.757	0.099	0.586	0.765	0.892
Anti-infective	0.757	0.065	0.644	0.752	0.854	0.708	0.068	0.600	0.707	0.808
Respiratory	0.823	0.065	0.712	0.831	0.920	0.773	0.083	0.627	0.784	0.907
Musculoskeletal	0.741	0.095	0.576	0.747	0.866	0.763	0.072	0.646	0.762	0.882
Cardiovascular	0.794	0.058	0.702	0.788	0.887	0.755	0.076	0.639	0.765	0.864
Genitourinary	0.814	0.083	0.670	0.821	0.937	0.801	0.090	0.635	0.808	0.927

The observed performance is essentially the MAR testing set AUC, since backfilling has already affected the datasets used. In Supplementary Materials E, we highlight the perils of relying on the MAR testing set for model validation, and suggest that the AUCs for the gold-standard and MCAR testing sets are more reflective of a classifier's real performance. Unfortunately, we have access to neither the gold-standard nor the MCAR testing sets, because we do not know the true, underlying values of the missing features. However, our experiments indicate that the AUCs for the MAR and MCAR testing sets of the 5NN-RF combination are very close (a difference of 0.002 on average). This means that we may use the former, the only observed figure, as a reasonable estimate of the latter, which reflects real performance.

Next, we train classifiers based on the union of the training and testing sets, and use them to generate predictions for pipeline drug-indication pairs. We generate predictions for P2APP using only information from phase 2 trials and for P3APP using only information from phase 3 trials. While we cannot compute AUC scores for these samples because their outcomes are still pending, we can compare their prediction scores with their development status at the time of this writing. These pipeline drug-indication pairs may still be in the same clinical stage (no change, i.e., phase 2 for P2APP; phase 3 for P3APP), be terminated (failed), or have progressed to higher phases (advanced).

Fig 8, Table 10, and Table 11 summarize the distributions of pipeline prediction scores. We find that pairs that fail generally have lower scores than those that advance to later phases of development. In Fig 8, we observe peaks at the lower end of the score spectrum for failed pairs (red) for both datasets. In contrast, pairs that advance tend to have peaks at higher scores (green). We observe the same patterns when we disaggregate the distributions by indication groups: the green parts tend to cluster above the distribution median while the red parts cluster below. However, there are also some indication groups for which there are too few samples to make any useful remarks (e.g., hormonal products in P2APP). From Table 10, we see that the average scores of failed pairs are indeed lower than those that advance (differences ranging from 0.05 to 0.15). In Table 11, we bin drug-indication pairs that have new developments (whether failure or advancement) into four groupings, depending on their prediction scores. For each bin, we compute the proportion of samples that advance to later development stages. We find that the proportions generally increase with the score magnitude, suggesting that pairs with higher scores are more likely to advance than those with lower scores. We note that progress to later clinical stages does not always lead to

approval. However, the results are still promising because advancement is a necessary condition for approval. Our experiments indicate that our trained classifiers are able to discriminate between high- and low-potential candidates.

Table 10. Distributions of prediction scores for all indication groups in aggregate (see Fig 8). Advanced refers to progress to a higher phase from the original phase. Original phase for P2APP is phase 2; for P3APP is phase 3. For instance, out of 1,511 drug-indication pairs in the P2APP testing set, 859 pairs are still pending decision in phase 2, 244 pairs have failed and 408 pairs have successfully advanced to phase 3 testing. Abbreviations: n: sample size; Avg: average; Sd: standard deviation; 5%: 5th percentile; 50%: median; 95%: 95th percentile.

	Prediction Scores					
	n	**Avg**	**Sd**	**5%**	**50%**	**95%**
P2APP						
Aggregate	1,511	0.153	0.061	0.044	0.155	0.258
No change	859	0.143	0.060	0.041	0.147	0.246
Failed	244	0.137	0.061	0.034	0.147	0.240
Advanced	408	0.183	0.056	0.093	0.178	0.274
P3APP						
Aggregate	252	0.417	0.189	0.128	0.402	0.695
No change	142	0.392	0.185	0.129	0.384	0.693
Failed	32	0.348	0.185	0.100	0.344	0.656
Advanced	78	0.492	0.176	0.233	0.492	0.699

Table 11. Distributions of prediction scores for all indication groups in aggregate (see Fig 8). Proportion refers to the fraction of samples that advanced to a later phase from the original phase. Abbreviations: n: sample size.

Scores	**n**	**Proportion**
P2APP		
< 0.1	108	0.231
0.1-0.2	368	0.671
0.2-0.3	171	0.766
≥ 0.3	5	1.000
P3APP		
< 0.2	13	0.308
0.2-0.4	35	0.686
0.4-0.6	27	0.667
≥ 0.6	35	0.914

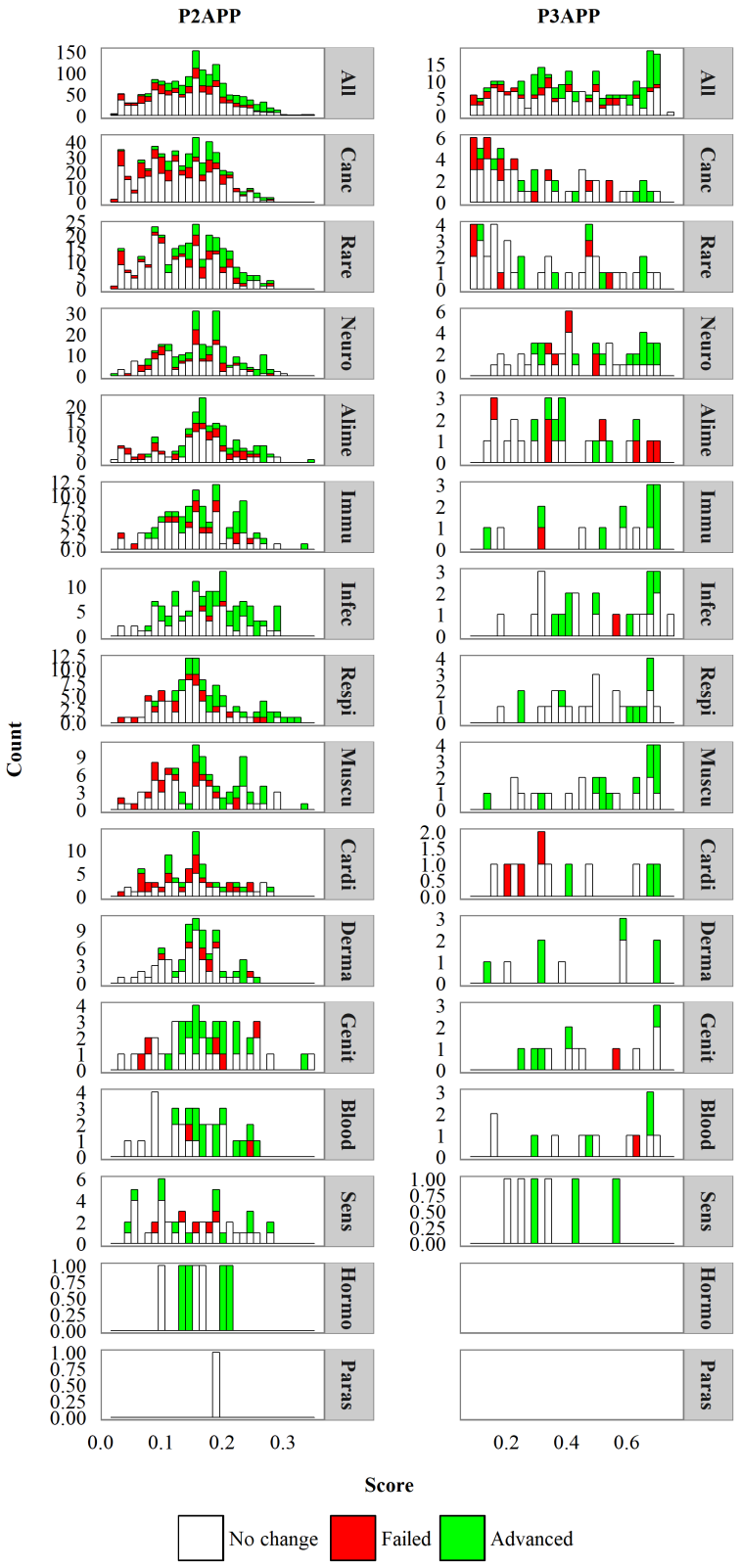

Fig 8. Distributions of prediction scores for P2APP and P3APP. First row for all indication groups in aggregate. Subsequent rows for specific indication groups.

To gain insight into the logic of our trained predictive models, we compute the average importance of features used in the 5NN-RF classifiers over all the experiments, and extract the top ten most informative variables. The RF classifier (Pedregosa et al., 2011) we used computes the importance of a variable by finding the decrease in node impurity for all nodes that split on that variable, weighted by the probability of reaching that node (as estimated by the proportion of samples reaching that node), averaged over all trees in the forest ensemble (Breiman, Friedman, Stone, & Olshen, 1984). Table 12 summarizes the results.

Table 12. Top ten important variables of 5NN-RF classifiers for P2APP and P3APP. Average and standard deviation taken across all experiments. Abbreviations: Avg: average; Sd: standard deviation.

	Importance	
	Avg	**Sd**
P2APP		
Trial outcome – completed, positive outcome, or primary endpoint(s) met	0.234	0.043
Trial status	0.160	0.026
Medium – solution	0.051	0.018
Actual accrual	0.046	0.010
Sponsor type – industry, all other pharma	0.025	0.008
Sponsor track record – number of positive phase 3 trials	0.023	0.006
Sponsor track record – number of failed drug-indication pairs	0.021	0.007
Study design – placebo control	0.019	0.009
Target accrual	0.018	0.005
Prior approval of drug for another indication	0.018	0.007
P3APP		
Trial outcome – completed, positive outcome, or primary endpoint(s) met	0.357	0.028
Trial status	0.148	0.014
Duration	0.099	0.016
Trial outcome – terminated, lack of efficacy	0.033	0.010
Trial outcome – completed, negative outcome, or primary endpoint(s) not met	0.033	0.008
Therapeutic area – oncology	0.030	0.009
Prior approval of drug for another indication	0.021	0.007
Actual accrual	0.015	0.003
Medium – powder	0.014	0.007
Medium – solution	0.012	0.006

We find that trial outcome (whether the trial was completed with its primary endpoints met) and trial status (whether the trial was completed or terminated) have significant

associations with success. These two features were consistently ranked the top two out of all variables and across both datasets. It is easy to imagine that a drug-indication pair whose trials were terminated has a low probability of success in terms of advancing from phases 2/3 to approval. In contrast, candidates that achieve positive outcomes certainly have a better shot at success. We also observe that prior approval of a drug has an effect on success for new indications or patient segmentation. It is plausible that developing an approved drug for a new indication has a greater likelihood of success than a new candidate.

In addition, trial characteristics such as accrual, duration, and the number of identified sites frequently appear in the top ten important variables. There are several possible explanations. For example, trials that end quickly without achieving primary endpoints may undermine the likelihood of success, and drugs with trials that have small accrual—and thus low statistical power—may have a lower probability of being approved.

We also find sponsor track records—quantified by the number of past successful trials (trials that achieve positive results or meet primary endpoints)—to be a useful factor for prediction. This factor has not been considered in previous related studies, but the intuition for its predictive power is clear: strong track records are likely associated with greater expertise in drug development.

Since drugs developed for different indication groups may have very different characteristics, we might expect classifiers trained on indication-group-specific data to outperform general classifiers. We build and analyze such specialized classifiers by filtering the datasets by indication group before performing the experiment described in the previous section. As a comparison, we also break down the performance of the general classifiers by indication

group. Table 9 shows the results for selected indication groups. In general, we find specialized models to give poorer performance than general models. This is likely because the former are trained on less data, which makes them less accurate and more susceptible to overfitting.

We note that the approach adopted in this section—splitting drug-indication pairs into training and testing sets randomly without considering the dates of development—may be less than ideal because of look-ahead bias. For example, if the results of a 2008 trial are included in the training set for predicting the outcome of a 2004 development path for a drug-indication pair, our model will be using future information during validation, which can yield misleading and impractical inferences. To address this issue, in the next section we apply our machine-learning framework to time-series data using rolling windows that account for temporal ordering in the construction of training and testing sets. Although this process makes use of less data within each estimation window than when the entire dataset is used, it minimizes the impact of look-ahead bias and yields more realistic inferences. We study the effects of random splitting versus temporal ordering in Supplementary Materials J.

Predictions over time

Drug development has changed substantially over time, thanks to new scientific discoveries and technological improvements. To reflect these changes in our predictive analytics, we adopt a time series, walk-forward approach to create training and testing sets for each of the two datasets, P2APP and P3APP (see Fig 9). We sample five-year rolling windows between 2004 and 2014 from each dataset. Each window consists of a training set of drug-indication pairs whose outcomes become finalized within the window, and an out-of-sample, out-of-

time testing set of drug-indication pairs that ended phase 2 or phase 3 testing, but are still in the pipeline with undetermined outcomes within the window. For example, consider the P2APP dataset. We draw the first window from 2004–2008, train our algorithm on drug-indication pairs that failed or were approved within this period as the training set, and apply the trained model to predict the outcomes of drug-indications that just ended phase 2 testing within the same window as the testing set.

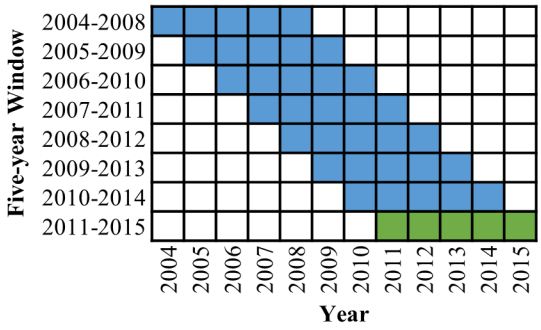

Fig 9. Time-series walk-forward analysis approach. The testing set in the last window (green) comprises drug-indication pairs in the pipeline at the time of snapshot of the databases.

We evaluate the resulting classifier by comparing its predictions with outcomes that are realized in the future (2009–2015). This rolling-window approach yields a total of eight overlapping training and testing periods where a new 5NN-RF model is trained for each period. The eighth testing period consists of drug-indication pairs in the pipeline at the time of snapshot of the databases. Unlike the first seven periods, their outcomes are still pending current development, and therefore we cannot compute a testing AUC for this window. However, we can examine the predictions and compare the scores with their development statuses at the time of this writing.

Fig 10 summarizes the results of the time-series analysis for the first seven windows. We observe an increasing trend over the years for both P2APP (0.67 in the first and 0.80 in the last window) and P3APP (0.77 in the first and 0.88 in the last window). Interestingly, we note that the proportions of complete cases in the training sets correlate well with the time series AUC (correlation coefficient 0.95 for P2APP and 0.90 for P3APP). We compute the proportion of complete cases by taking the number of feature vector rows with complete information over the total number of rows. As is apparent from Fig 10, the proportions have been increasing over the years for both datasets. This is likely due to better data reporting practices by drug developers, a possible consequence of FDAAA.

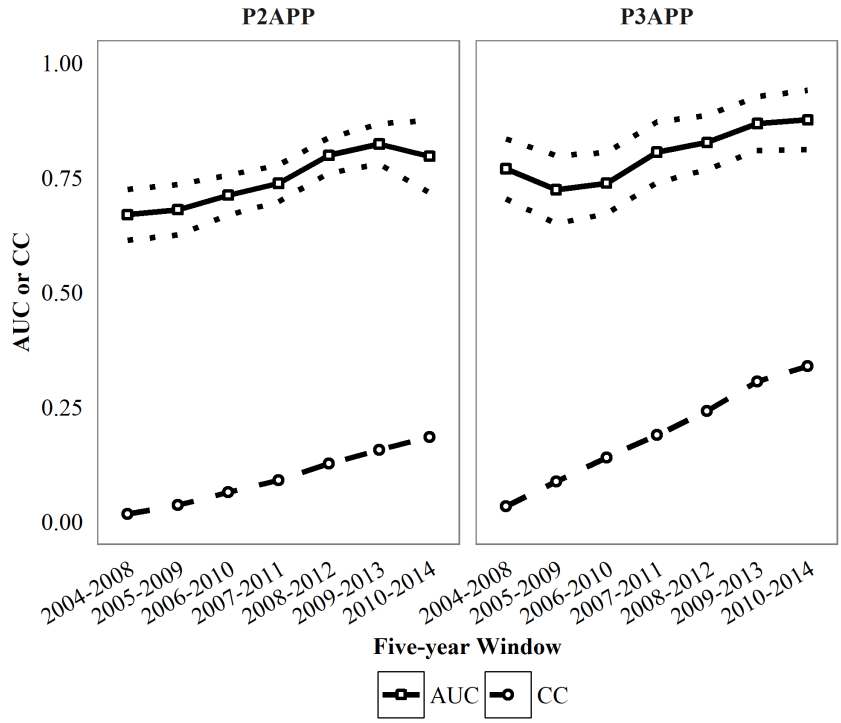

Fig 10. Time-series walk-forward analysis for P2APP and P3APP using 5NN-RF. We use bootstrapping to determine the 95% CI for AUC (dotted lines). The dashed lines plot the corresponding proportions of complete cases in the training sets of each five-year window. Abbreviations: CC: proportion of complete cases.

Next, we examine the 2011–2015 window. Fig 11, Table 13, and Table 14 summarize the distributions of prediction scores for the P2APP and P3APP datasets. We observe very similar patterns to the static pipeline predictions above. The histograms, average scores, and binning of samples indicate that pairs that fail tend to have lower prediction scores than those that advance. This shows that our classifiers are indeed able to differentiate successful candidates.

Table 13. Distributions of prediction scores for all indication groups in aggregate (see Fig 11). Advanced refers to progress to a higher phase from the original phase. Original phase for P2APP is phase 2; for P3APP is phase 3. Abbreviations: n: sample size; Avg: average; Sd: standard deviation; 5%: 5th percentile; 50%: median; 95%: 95th percentile.

	Prediction Scores					
	n	Avg	Sd	5%	50%	95%
P2APP						
Aggregate	1,190	0.158	0.080	0.036	0.173	0.290
No change	712	0.148	0.080	0.035	0.158	0.275
Failed	195	0.143	0.079	0.034	0.149	0.255
Advanced	283	0.197	0.071	0.068	0.200	0.323
P3APP						
Aggregate	218	0.431	0.211	0.113	0.476	0.689
No change	121	0.395	0.207	0.113	0.403	0.684
Failed	28	0.362	0.211	0.093	0.335	0.640
Advanced	69	0.521	0.193	0.149	0.631	0.707

Table 14. Distribution of prediction scores for all indication groups in aggregate (see Fig 11). Proportion refers to the fraction of samples that advanced to a higher phase from the original phase. Abbreviations: n: sample size.

Scores	n	Proportion
P2APP		
< 0.1	99	0.313
0.1-0.2	183	0.607
0.2-0.3	168	0.690
≥ 0.3	28	0.893
P3APP		
< 0.2	17	0.412
0.2-0.4	17	0.706
0.4-0.6	17	0.647
≥ 0.6	46	0.848

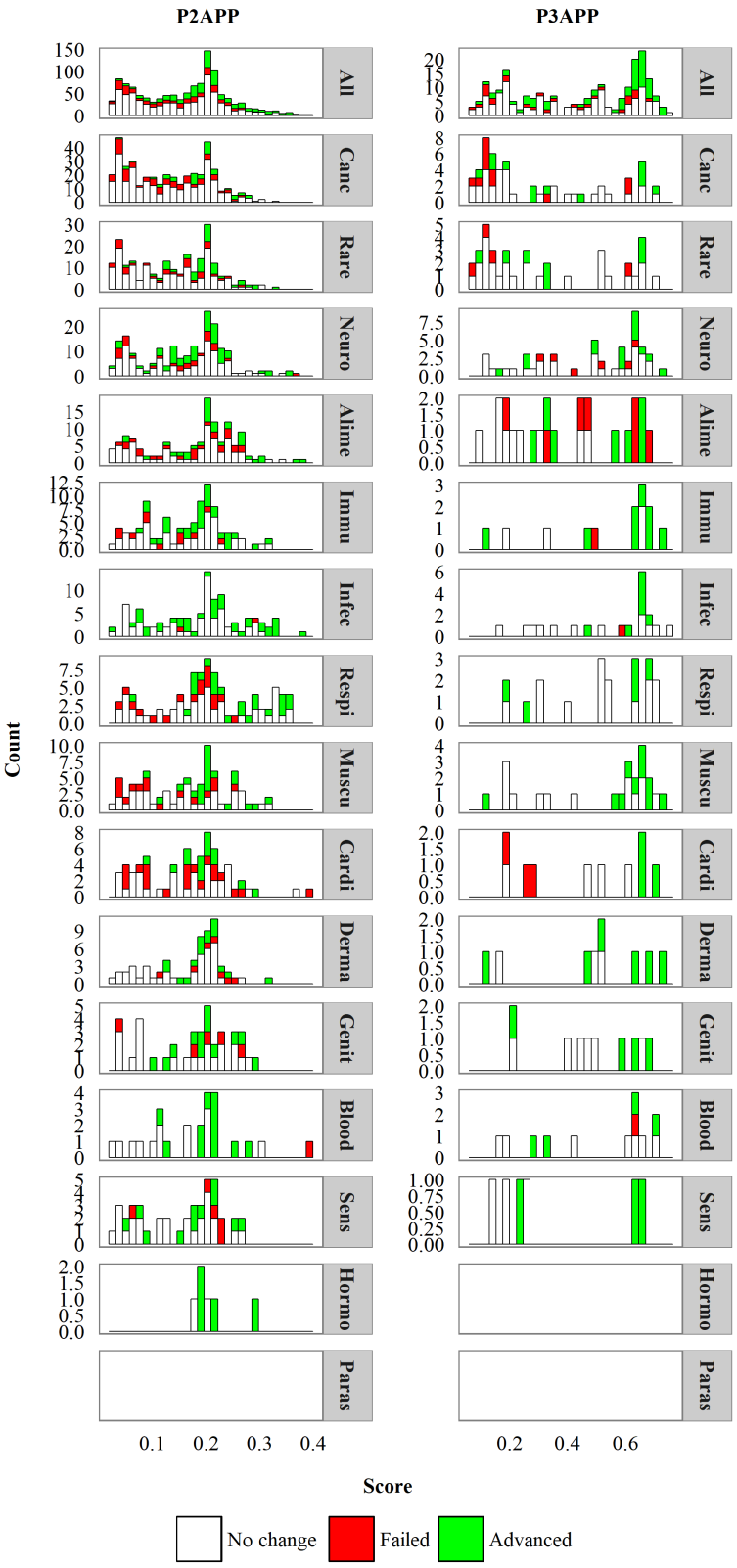

Fig 11. Distributions of prediction scores of the 2011–2015 window testing set for P2APP and P3APP. First row for all indication groups in aggregate. Subsequent rows for specific indication groups.

Table 15 summarizes the top ten most informative variables in the 5NN-RF classifiers over the eight rolling windows. We find them to be largely consistent with those observed in the static case: the trial outcome and trial status are significantly associated with success; trial characteristics (such as accrual, duration, and number of identified sites), sponsor track record, and drug medium appear frequently in both scenarios.

Table 15. Top ten important variables in 5NN-RF classifiers for P2APP and P3APP. Average and standard deviation taken across the eight rolling windows. Abbreviations: Avg: average; Sd: standard deviation.

	Importance	
	Avg	**Sd**
P2APP		
Trial outcome – completed, positive outcome, or primary endpoint(s) met	0.203	0.083
Trial status	0.102	0.033
Prior approval of drug for another indication	0.077	0.061
Actual accrual	0.039	0.015
Target accrual	0.031	0.010
Duration	0.027	0.014
Sponsor track record – number of completed phase 3 trials	0.025	0.007
Medium – suspension	0.024	0.018
Sponsor type – academic	0.023	0.017
Medium – solution	0.021	0.019
P3APP		
Trial outcome – completed, positive outcome, or primary endpoint(s) met	0.348	0.028
Trial status	0.125	0.020
Duration	0.053	0.017
Prior approval of drug for another indication	0.046	0.028
Trial outcome – completed, negative outcome, or primary endpoint(s) not met	0.033	0.026
Target accrual	0.021	0.005
Trial outcome – terminated, lack of efficacy	0.020	0.013
Actual accrual	0.019	0.004
Therapeutic area – oncology	0.017	0.013
Number of identified sites	0.012	0.002

As in the static case, we also train indication-group specific classifiers using rolling windows. Table 16 and Table 17 summarize the results for selected indication groups in P2APP and P3APP, respectively (see Supplementary Materials G for results of all other indication groups). Indication groups with small sample sizes tend to produce poor and unstable

specialized classifiers (e.g., the musculoskeletal indication group in P2APP). This is expected because models trained on small training sets are more susceptible to overfitting, especially when non-linear algorithms such as RF are used. In contrast, indication groups with larger sample sizes tend to give rise to rather good classifiers (e.g., anti-cancer in P2APP).

Table 16. Comparison of the general and indication-group specific classifiers for selected indication groups in P2APP. We use bootstrapping to determine the 95% CI for AUC.

	General Classifier			Specialized Classifiers		
	Train Set	**Test Set**	**AUC [95% CI]**	**Train Set**	**Test Set**	**AUC [95% CI]**
All						
2004–2008	1,361	551	0.669 [0.614, 0.725]	-	-	-
2005–2009	1,562	591	0.680 [0.625, 0.735]	-	-	-
2006–2010	1,764	636	0.712 [0.668, 0.755]	-	-	-
2007–2011	1,969	598	0.738 [0.698, 0.777]	-	-	-
2008–2012	2,082	597	0.799 [0.760, 0.837]	-	-	-
2009–2013	2,212	517	0.823 [0.779, 0.867]	-	-	-
2010–2014	2,289	380	0.797 [0.718, 0.876]	-	-	-
Anti-cancer						
2004–2008	1,361	137	0.665 [0.528, 0.803]	456	137	0.683 [0.533, 0.833]
2005–2009	1,562	163	0.739 [0.618, 0.861]	494	163	0.635 [0.512, 0.758]
2006–2010	1,764	188	0.774 [0.702, 0.846]	546	188	0.726 [0.635, 0.816]
2007–2011	1,969	193	0.830 [0.773, 0.887]	618	193	0.746 [0.661, 0.831]
2008–2012	2,082	198	0.805 [0.717, 0.894]	682	198	0.760 [0.665, 0.855]
2009–2013	2,212	177	0.852 [0.783, 0.922]	736	177	0.786 [0.696, 0.876]
2010–2014	2,289	173	0.815 [0.691, 0.938]	791	173	0.803 [0.666, 0.940]
Musculoskeletal						
2004–2008	1,361	35	0.765 [0.597, 0.933]	96	35	0.704 [0.512, 0.896]
2005–2009	1,562	38	0.716 [0.489, 0.944]	109	38	0.674 [0.472, 0.876]
2006–2010	1,764	35	0.634 [0.439, 0.830]	111	35	0.509 [0.276, 0.742]
2007–2011	1,969	37	0.737 [0.571, 0.903]	119	37	0.677 [0.493, 0.860]
2008–2012	2,082	36	0.884 [0.773, 0.995]	127	36	0.683 [0.462, 0.904]
2009–2013	2,212	26	0.792 [0.573, 1.000]	133	26	0.667 [0.429, 0.904]
2010–2014	2,289	19	0.882 [0.724, 1.000]	128	19	0.882 [0.706, 1.000]

Table 17. Comparison of the general and indication-group specific classifiers for selected indication groups in P3APP. We use bootstrapping to determine the 95% CI for AUC.

	General Classifier			Specialized Classifiers		
	Train Set	Test Set	AUC [95% CI]	Train Set	Test Set	AUC [95% CI]
All						
2004–2008	472	196	0.769 [0.704, 0.834]	-	-	-
2005–2009	559	177	0.724 [0.650, 0.798]	-	-	-
2006–2010	604	211	0.738 [0.671, 0.805]	-	-	-
2007–2011	664	174	0.806 [0.740, 0.871]	-	-	-
2008–2012	677	197	0.827 [0.768, 0.886]	-	-	-
2009–2013	740	153	0.868 [0.809, 0.927]	-	-	-
2010–2014	734	110	0.876 [0.811, 0.941]	-	-	-
Anti-cancer						
2004–2008	472	34	0.773 [0.618, 0.928]	95	34	0.684 [0.495, 0.874]
2005–2009	559	28	0.740 [0.543, 0.936]	107	28	0.568 [0.345, 0.791]
2006–2010	604	50	0.754 [0.599, 0.910]	110	50	0.630 [0.452, 0.809]
2007–2011	664	24	0.587 [0.333, 0.842]	132	24	0.392 [0.132, 0.651]
2008–2012	677	40	0.793 [0.549, 1.000]	134	40	0.668 [0.457, 0.879]
2009–2013	740	29	0.800 [0.480, 1.000]	151	29	0.775 [0.528, 1.000]
2010–2014	734	26	0.943 [0.842, 1.000]	153	26	0.852 [0.558 ,1.000]
Rare Diseases						
2004–2008	472	22	0.711 [0.465, 0.957]	54	22	0.620 [0.364, 0.876]
2005–2009	559	23	0.735 [0.517, 0.952]	60	23	0.606 [0.360, 0.852]
2006–2010	604	24	0.888 [0.747, 1.000]	66	24	0.825 [0.645, 1.000]
2007–2011	664	22	0.838 [0.652, 1.000]	72	22	0.735 [0.520, 0.950]
2008–2012	677	34	0.893 [0.780, 1.000]	76	34	0.700 [0.523, 0.877]
2009–2013	740	28	0.962 [0.899, 1.000]	94	28	0.932 [0.840, 1.000]
2010–2014	734	18	0.908 [0.766, 1.000]	109	18	0.985 [0.942, 1.000]

For comparison, we disaggregate performance by indication group. We find that these classifiers do not lose out to their specialized counterparts. In fact, our results show that the former tend to exhibit more stable performance across the seven windows, particularly on indication groups with small sample sizes. We hypothesize that classifiers trained on all data benefit from having access to larger datasets with greater diversity, and are thus able to make more informed predictions. This suggests that it may be more appropriate to rely on general classifiers, rather than specialized ones, for predictions over time where samples are spread out over multiple windows, since further filtering by indication group results in even smaller sample sizes.

Finally, we extract the top five P2APP pipeline drug candidates with the highest scores in each indication group as predicted by the 2011–2015 rolling-window model. Table 18 summarizes the results. We include only candidates that are still outstanding at the time of writing (neither discontinued nor approved). It is encouraging that many of these candidates (indicated in italics) have advanced beyond phase 2 testing since our analysis, indicating the predictive power of our models. We include an interactive version in Fig 12 where readers can filter our pipeline predictions by indication group and probability of approval. Ultimately, all biopharma stakeholders can use such scores to rank and evaluate the potential risks and rewards of drug candidates.

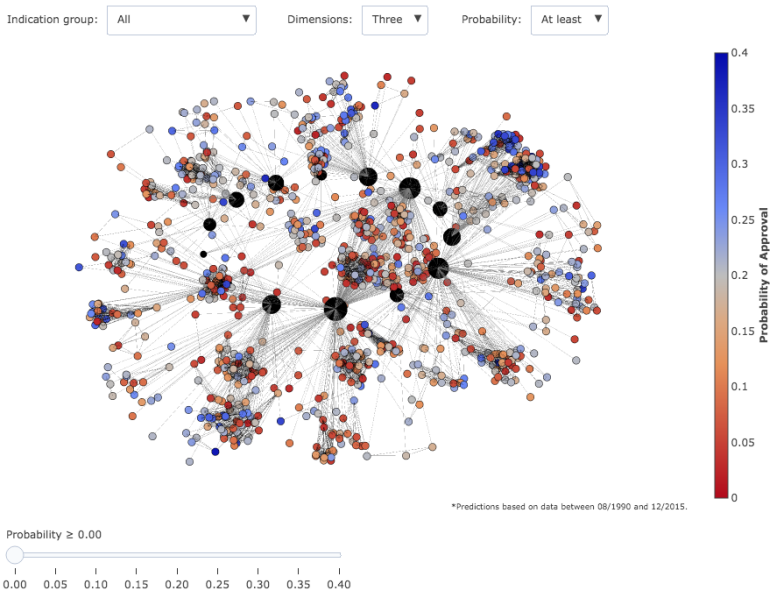

Fig 12. Network graph of P2APP pipeline drug candidates. Black nodes correspond to indication groups. Colored nodes correspond to drug-indication pairs. Each drug-indication pair node is connected to its parent indication group and also other drug-indication pairs that have the same indication. They are colored according to their respective probability of approval as predicted by our model—blue for higher scores and red for lower scores. Hover over nodes for details of each drug-indication pair. Black indication group nodes are sized based on the number of connections.

The information presented through this research and in the included figure are made available solely for general informational purposes. The authors do not warrant the accuracy, completeness or usefulness of this information. Any reliance you place on such information is strictly at your own risk. The authors expressly disclaim all liability and responsibility arising from any reliance placed on such information by you, or by anyone who may be informed of this information.

Table 18. Top five P2APP pipeline drug candidates with the highest scores in each indication group as predicted by our model. We include only candidates that are still outstanding at the time of writing (neither discontinued nor approved). Drug-indication pairs in italics are those that have advanced beyond phase 2 testing since our analysis.

Drug	Indication	Score
Anti-cancer		
ontecizumab	Cancer, colorectal	0.34
calmangafodipir	Radio/chemotherapy-induced injury, bone marrow, neutropenia	0.31
tivantinib	Cancer, sarcoma, soft tissue	0.30
pidilizumab	Cancer, colorectal	0.29
NK-012	Cancer, colorectal	0.28
Rare Diseases		
surotomycin	*Infection, Clostridium difficile*	*0.34*
tivantinib	Cancer, sarcoma, soft tissue	0.30
VP-20621	Infection, Clostridium difficile prophylaxis	0.30
KHK-7580	*Secondary hyperparathyroidism*	0.29
nitric oxide, inhaled	*Hypertension, pulmonary*	0.29
Neurological		
dasotraline	*Attention deficit hyperactivity disorder*	*0.35*
idalopirdine	*Alzheimer's disease*	*0.35*
GRC-17536	Neuropathy, diabetic	0.34
caprylic triglyceride	*Alzheimer's disease*	*0.32*
levodopa	*Parkinson's disease*	*0.31*
Alimentary		
ibodutant	*Irritable bowel syndrome, diarrhoea-predominant*	*0.37*
GRC-17536	Neuropathy, diabetic	0.34
mesalazine + N-acetylcysteine	Colitis, ulcerative	0.31
apabetalone (tablet)	*Diabetes, Type 2*	*0.31*
phosphatidylcholine	*Colitis, ulcerative*	*0.31*
Immunological		
tofacitinib	*Arthritis, psoriatic*	*0.31*
ixekizumab	Arthritis, rheumatoid	0.31
anti-BLyS/APRIL antibody fusion protein	Arthritis, rheumatoid	0.31
sirukumab	*Arthritis, rheumatoid*	*0.29*
dimethyl fumarate	Psoriasis	0.27
Anti-infective		
delafloxacin	*Infection, skin and skin structure, acute bacterial*	*0.39*
surotomycin	*Infection, Clostridium difficile*	*0.34*
delafloxacin	*Infection, pneumonia, community-acquired*	*0.33*
plazomicin	*Infection, urinary tract, complicated*	*0.33*
Ypeginterferon alpha-2b	*Infection, hepatitis-C virus*	*0.33*
Respiratory		
fluticasone + salmeterol	*Asthma*	*0.36*
fluticasone furoate + umeclidinium + vilanterol	*Chronic obstructive pulmonary disease*	*0.36*
fluticasone furoate + umeclidinium	Chronic obstructive pulmonary disease	0.36
beclometasone + formoterol	Chronic obstructive pulmonary disease	0.35
fluticasone propionate DPI	*Asthma*	*0.35*
Musculoskeletal		
tofacitinib	*Arthritis, psoriatic*	*0.31*
ixekizumab	Arthritis, rheumatoid	0.31
anti-BLyS/APRIL antibody fusion protein	Arthritis, rheumatoid	0.31
sirukumab	*Arthritis, rheumatoid*	*0.29*
romosozumab	*Osteoporosis*	*0.28*
Cardiovascular		
K-134	Peripheral vascular disease	0.37
nitric oxide, inhaled	*Hypertension, pulmonary*	*0.29*
TY-51924	Infarction, myocardial	0.28
s-amlodipine + telmisartan	*Hypertension, unspecified*	*0.27*
tirasemtiv	Peripheral vascular disease	0.24
Dermatological		
tofacitinib	*Arthritis, psoriatic*	*0.31*
dimethyl fumarate	Psoriasis	0.27
pefcalcitol	*Psoriasis*	*0.24*
Benvitimod	*Psoriasis*	*0.22*
calcipotriol monohydrate + betamethasone dipropionate	Psoriasis	0.22
Genitourinary		
etonogestrel + estradiol (vaginal ring), next generation	*Contraceptive, female*	*0.30*
drospirenone + estradiol	Contraceptive, female	0.28
finerenone	*Nephropathy, diabetic*	*0.27*
afacifenacin fumarate	Overactive bladder	0.26
GKT-137831	Nephropathy, diabetic	0.26
Blood and Clotting		
calmangafodipir	Radio/chemotherapy-induced injury, bone marrow, neutropenia	0.31
balugrastim	*Radio/chemotherapy-induced injury, bone marrow, neutropenia*	*0.27*
eflapegrastim	*Radio/chemotherapy-induced injury, bone marrow, neutropenia*	*0.25*
pegfilgrastim	*Radio/chemotherapy-induced injury, bone marrow, neutropenia*	*0.22*
lexaptepid pegol	Radio/chemotherapy-induced anaemia	0.20
Sensory		
AR-13324 + latanoprost	*Glaucoma*	*0.27*
S-646240	Macular degeneration, age-related, wet	0.27
netarsudil	*Glaucoma*	*0.26*
fenofibrate, micronized-2	Oedema, macular, diabetic	0.25
LX-7101	Glaucoma	0.21
Hormonal		
KHK-7580	*Secondary hyperparathyroidism*	*0.29*
somatropin prodrug, pegylated	*Growth hormone deficiency*	*0.21*
2MD	Secondary hyperparathyroidism	0.21
velcalcetide	*Secondary hyperparathyroidism*	*0.19*
tesamorelin acetate	Growth hormone deficiency	0.18

4 Discussion

Drug development is an extremely costly process and the accurate evaluation of a candidate drug's likelihood of approval is critical to the efficient allocation of capital. Historical successes and failures contain valuable insights on the characteristics of high-potential candidates. Unfortunately, such data are often incomplete due to partial reporting by investigators and developers. Most analytic methods require complete data, however, and prior studies on estimating approval rates and predicting approvals are typically based on a small number of examples that have complete information for just a few features.

In this paper, we extract two datasets, P2APP and P3APP, from Informa® databases and apply 5NN statistical imputation to make efficient use of all available data. We use machine-learning techniques to train and validate our RF predictive models and achieve promising levels of predictive power for both datasets. When applied to pipeline drugs, we find that candidates with higher scores are indeed more likely to advance to higher clinical phases, indicating that our 5NN-RF classifiers are able to discriminate between high- and low-potential candidates.

A time-series analysis of the datasets shows generally increasing trends in performance over five-year rolling windows from 2004 to 2014. We find that the classifiers' performance correlates well with the proportions of complete cases in the training sets: as completeness increases, the classifier learns better and achieves higher AUCs. This highlights the importance of data quality in building more accurate predictive algorithms for drug development.

Finally, we compute feature importance in the predictive models and find that trial outcomes, trial status, trial accrual rates, duration, prior approval for another indication, and sponsor track records are the most critical features for predicting success. Because the 5NN-RF classifiers are non-linear, there is no simple interpretation of the incremental contribution of each predictor to the forecast. However, the intuition behind some of these factors is clear: drug-indication pairs with trials that achieve positive outcomes certainly have a better chance of approval; candidates sponsored by companies with strong track records and greater expertise in drug development should have higher likelihood of success; and approved drugs may have higher chances of approval for a second related indication. Many of these factors contain useful signals about drug development outcomes but have not been considered in prior studies.

These results are promising and raise the possibility of even more powerful drug development prediction models with access to better quality data. This can be driven by programs such as Project Data Sphere (Green et al., 2015) and Vivli (Bierer, Li, Barnes, & Sim, 2016) that promote and facilitate public sharing of patient-level clinical trial data. Ultimately, such predictive analytics can be used to make more informed data-driven decisions in the risk assessment and portfolio management of investigational drugs at all clinical stages.

References

Bierer, B. E., Li, R., Barnes, M., & Sim, I. (2016). A global, neutral platform for sharing trial data. *New England Journal of Medicine*, 374(25), 2411-2413.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330.

DiMasi, J. A., Hermann, J. C., Twyman, K., Kondru, R. K., Stergiopoulos, S., Getz, K. A., & Rackoff, W. (2015). A tool for predicting regulatory approval after phase II testing of new oncology compounds. *Clinical Pharmacology & Therapeutics*, 98(5), 506-513.

El-Maraghi, R. H., & Eisenhauer, E. A. (2008). Review of phase II trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase III. *Journal of Clinical Oncology*, 26(8), 1346-1354.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition letters*, 27(8), 861-874.

Fernandez, J. M., Stein, R. M., & Lo, A. W. (2012). Commercializing biomedical research through securitization techniques. *Nature Biotechnology*, 30(10), 964-975.

Goffin, J., Baral, S., Tu, D., Nomikos, D., & Seymour, L. (2005). Objective responses in patients with malignant melanoma or renal cell cancer in early clinical studies do not predict regulatory approval. *Clinical Cancer Research*, 11(16), 5928-5934.

Green, A. K., Reeder-Hayes, K. E., Corty, R. W., Basch, E., Milowsky, M. I., Dusetzina, S. B., Bennett, A.V., & Wood, W. A. (2015). The project data sphere initiative: accelerating cancer research by sharing data. *The Oncologist*, 20(5), 464-e20.

Informa - Pharmaceutical Clinical Trial Intelligence Products. (2016). *Informa*. Retrieved 5 December 2016, from https://pharmaintelligence.informa.com/products-and-services/data-and-analysis/citeline-joins-informas-pharma-intelligence

Jardim, D. L., Groves, E. S., Breitfeld, P. P., & Kurzrock, R. (2017). Factors associated with failure of oncology drugs in late-stage clinical development: A systematic review. *Cancer Treatment Reviews*, 52, 12-21.

Kuhn, M., Weston, S., Coulter, N., & Quinlan, R. (2014). C50: C5. 0 decision trees and rule-based models. R package version 0.1. 0-21.

Malik, L., Mejia, A., Parsons, H., Ehler, B., Mahalingam, D., Brenner, A., Sarantopoulos J., & Weitman, S. (2014). Predicting success in regulatory approval from Phase I results. *Cancer Chemotherapy and Pharmacology*, 74(5), 1099-1103.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825-2830.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473-489.

Scannell, J., Blanckley, A., Boldon, H. & Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery* 11, 191–200.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC Press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2), 147-177.

Van Buuren, S. (2012). *Flexible imputation of missing data*. CRC Press.

Zarin, D. A., Tse, T., Williams, R. J., & Carr, S. (2016). Trial reporting in clinicaltrials.gov—the final rule. *New England Journal of Medicine*, 375(20), 1998-2004.