Machine-learning models for predicting drug approvals and clinical-phase transitions‡

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] AlphaSimplex Group, LLC, Cambridge, Massachusetts, United States of America

* Corresponding author

email: alo-admin@mit.edu (AL)

This Version: 25 July 2017

Abstract

We apply machine-learning techniques to predict drug approvals and phase transitions 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. Imputation methods are used to deal with missing data, allowing us to fully exploit the entire dataset, the largest of its kind. We achieve predictive measures of 0.74, 0.78, and 0.81 AUC for predicting transitions from phase 2 to phase 3, 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.

‡ 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 Linda Blackerby, Lara Boro, James Wade, Ellen Moore and Howard Fingert for specific comments on this manuscript. 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.

Contents

1 Introduction

While many promising breakthroughs such as immuno-therapies and gene-editing techniques offer new hope for patients, they have also made biomedical innovation riskier and more expensive. These breakthroughs generate novel therapies for investigation, each of which requires many years of translational research and clinical testing, costing hundreds of millions of dollars and yet often facing a high likelihood of failure (Fernandez, Stein, & Lo, 2012). In fact, drug discovery productivity has been declining steadily, despite scientific and technical progress, over the past 50 years. This phenomenon, termed "Eroom's Law" by Scannell et al. (2012)—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 better evaluate drug candidates and to allocate capital to high-potential opportunities more efficiently has only intensified.

Drug developers typically use general estimates of regulatory approval rates, or estimates specific only to a drug's therapeutic class, when managing their portfolio of investigational drugs. Here we propose the use of a wider range of factors and machine-learning techniques to estimate success rates. Machine-learning is a branch of computer science focused on tackling pattern recognition problems and building predictive models to make data-driven decisions, and is well suited for this application. Predictive factors include drug compound characteristics, clinical trial design, previous trial outcomes, and the sponsor track record. We hypothesize that these features contain useful signals about drug development outcomes that will allow us to forecast the outcome of pipeline developments more accurately. Our goal is to develop predictive algorithms for assessing the probability of success of drug candidates in three scenarios: advancing from phase 2 to phase 3 testing, from phase 2 to regulatory approval, and from phase 3 to regulatory approval. Such predictions may be used to evaluate the risks of different investigational drugs at different clinical stages, reducing the risk and increasing the efficiency of drug development and portfolio decision-making.

We construct three datasets, one for each scenario, from two proprietary pharmaceutical pipeline databases, *Pharmaprojects* and *Trialtrove* provided by Informa® (Informa, 2016) The phase 2 to phase 3 dataset includes more than 5,200 unique drugs for 274 indications and over 8,800 phase 2 clinical trials, while 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 1,800 unique drugs for 253 indications and over 4,500 phase 3 trials. These data cover over 15 indication groups.

To the best of our knowledge, this study is the largest of its kind. 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 et

al. (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 et al. (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 Appendix H for a comparison of our analysis to theirs).

With the FDA Amendments Act of 2007, drug and clinical trial data collection has been rapidly expanding. These data are often sparse, however, and our dataset is not an exception. Related studies (e.g. DiMasi et al., 2015) typically use only complete-case observations—discarding clinical trials with any missing information—which is highly restrictive and may lead to certain biases. In this paper, we characterize the observed patterns of missing data and propose the use of standard imputation methods—statistical procedures to infer missing data—to address this issue. We explore four common approaches to "missingness" and demonstrate their advantages and disadvantages over discarding incomplete cases.

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 curve" (AUC) metric to measure model performance (AUC, which stands for "area under the curve," is the estimated probability that a classifier will rank a positive outcome higher than a negative outcome [Fawcett, 2006]). We achieve AUCs of 0.74 for predicting transitions from phase 2 to phase 3 testing (95% confidence interval (CI): [0.71,0.76]), 0.78 for predicting phase 2 to approval (95% 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

We use two commercial pipeline databases from the commercial data vendor Informa®: *Pharmaprojects*, which specializes in drug information, and *Trialtrove*, which specializes in clinical trials intelligence (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 et al., 2016], Clinicaltrialsregister.eu [extracted from EudraCT]), and scientific conferences and publications. Using these sources, we construct three datasets of drug-indication pairs: phase 2 to phase 3 (P2P3), phase 2 to approval (P2APP) and phase 3 to approval (P3APP). Applying machine-learning algorithms to these datasets allows us to estimate: (1) whether a drug-indication pair that has concluded phase 2 testing will advance to phase 3 testing; (2) whether a pair that has concluded phase 2 testing will be approved eventually; (3) whether a pair that has concluded phase 3 testing will be approved eventually. Data cleaning procedures are outlined in Appendix A.

We consider all indications associated with a particular drug, as opposed to only the lead indication. We extract all features that could conceivably be correlated 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 Appendix 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 P2P3 dataset consists of 5,288 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 span from January 1, 1994 to December 15, 2015. In our sample, 4,168 pairs have known outcomes, while 1,120 pairs are still in the pipeline. For those pairs with known outcomes, we classify instances that successfully advanced to phase 3 testing as "successes" (18.7%), and instances that have suspended or discontinued development, or had no development reported over 18 months, as "failures" (81.3%). The P2APP dataset consists of 6,344 pairs that have ended phase 2 testing, of which 4,812 pairs have known outcomes while 1,532 pairs are still in the pipeline. The trials range from August 8, 1990 to December 15, 2015. 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 P2APP. 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 phase advancement or approval. In Section 3, 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 P2P3 and P2APP, and the second largest in P3APP (see Table 3). Industry-sponsored trials dominate all three 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 1).

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 (Malik et al., 2014; Goffin et al., 2005; El-Maraghi and Eisenhauer, 2008; DiMasi et al., 2015). 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 Appendix 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 accrual of the trial.	Continuous
Actual accrual	Actual 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 P2P3, P2APP and P3APP datasets. We consider phase 2 trial information in P2P3 and P2APP datasets, phase 3 trial information in P3APP dataset.

	Counts				
	Drug-indication Pairs	**Phase 2/3 Trials**	**Unique Drugs**	**Unique Indications**	**Unique Phase 2/3 Trials**
			P2P3		
Success	779	1,469	689	192	1,457
Failure	3,389	5,862	2,369	239	5,683
Pipeline	1,120	1,874	888	196	1,836
Total	**5,288**	**9,205**	**3,548**	**274**	**8,839**
			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*.

	Counts		
	P2P3	**P2APP**	**P3APP**
All	5,288	6,344	1,870
Anti-cancer	1,948	2,239	409
Rare Diseases	941	1,105	259
Neurological	845	1,069	444
Alimentary	620	757	249
Immunological	423	474	101
Anti-infective	381	493	177
Respiratory	367	428	134
Musculoskeletal	339	394	121
Cardiovascular	303	388	158
Dermatological	228	254	45
Genitourinary	159	210	85
Blood and Clotting	117	160	97
Sensory	116	137	41
Hormonal	14	17	4
Anti-parasitic	9	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		
	P2P3	**P2APP**	**P3APP**
All	8,839	14,584	4,552
Other Pharma	4,246	5,432	1,721
Top 20 Pharma	3,379	5,322	2,369
Academic	1,920	4,869	736
Government	1,148	1,807	314
Cooperative Group	497	958	230
Not for Profit	115	181	51
Generic	31	52	54
Contract Research Organization	31	41	17

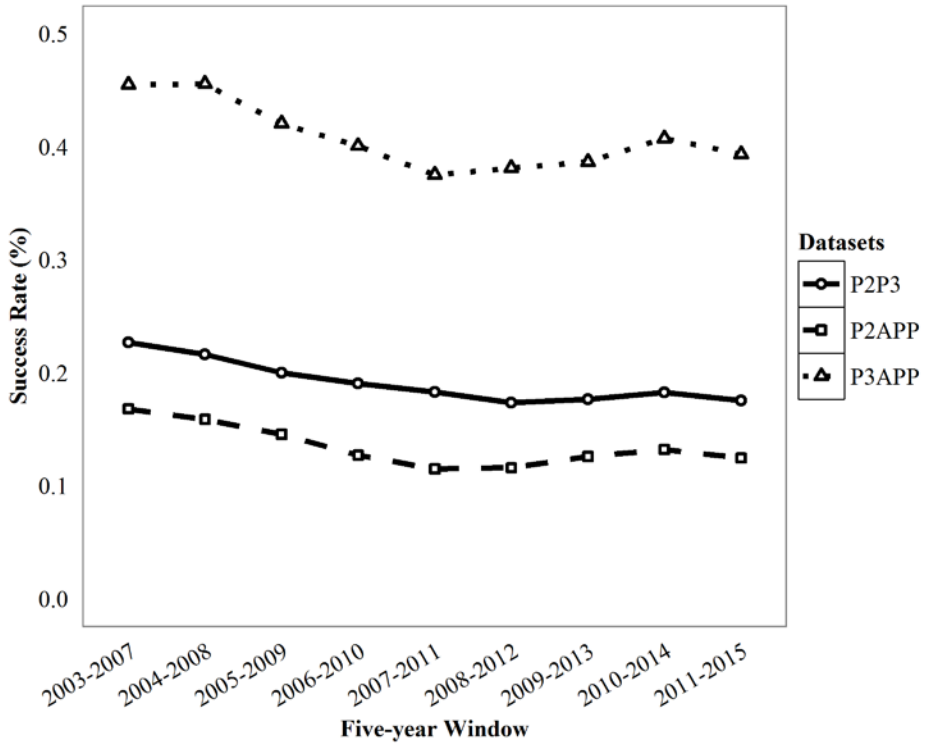

Fig 1. Success rates in P2P3, P2APP and P3APP over five year rolling windows from 2003-2015. At each time window, the P2APP success rate is strictly lesser than that of P2P3 because drug-indication pairs that successfully progress to phase 3 may still fail.

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) is usually available to the investigators but may not be released publicly, and is thus considered "missing" from our standpoint.

Fig 2, Fig 3, 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. Across the three 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 P2P3 and P2APP datasets which focus on phase 2 drugs and trials. This is expected since phase 3 trials are primarily used to support registration filings.

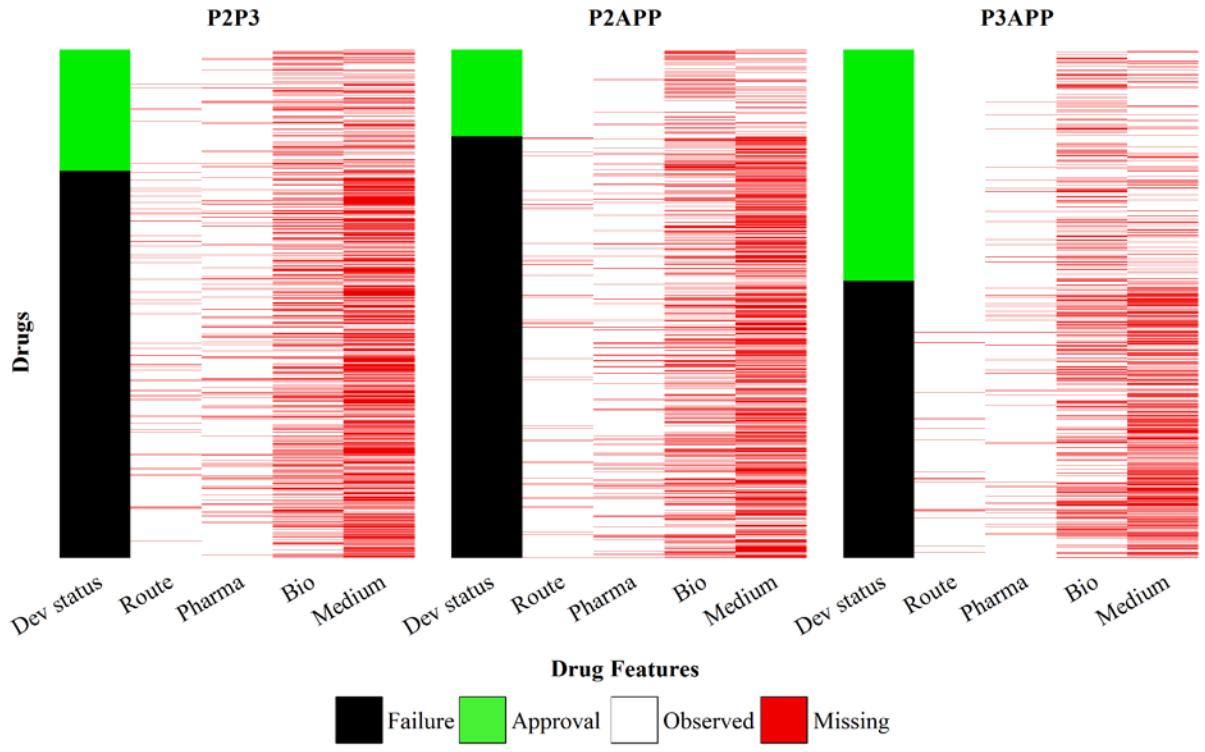

Fig 2. Missingness patterns of drug features. Each row corresponds to a unique drug. Features not included in the figure are complete and do not have missing values. Abbreviations: Dev status: highest level of development of a drug for any indication; Pharma: pharmacological target family; Bio: biological target family.

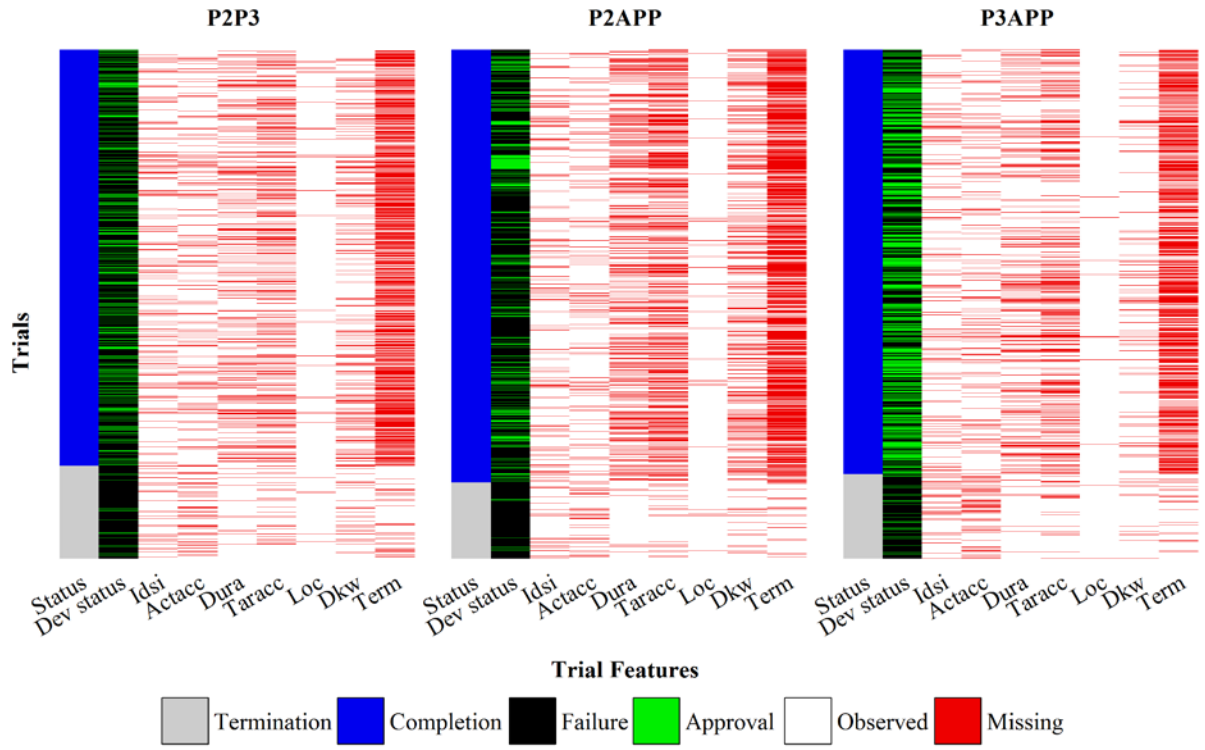

Fig 3. Missingness patterns of trial features. Each row corresponds to a unique clinical trial. Features not included in the figure are complete and do not have missing values. Abbreviations: Dev status: highest level of development of a drug for any indication; Status: trial status; Idsi: number of identified sites; Actacc: actual accrual; Dura: duration; Taracc: target accrual; Loc: locations; Dkw: trial study design keywords; Term: trial outcomes.

Table 5. Missingness in drug features with respect to unique drugs (see Fig 2). The column heading "Unconditional" refers to overall missingness without conditioning on outcomes.

	Missingness		
	Unconditional	**Success**	**Failure**
P2P3			
Route	0.04	0.01	0.05
Pharmacological target family	0.07	0.04	0.08
Biological target family	0.31	0.29	0.32
Medium	0.55	0.32	0.63
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 3).

	Missingness		
	Unconditional	**Completion**	**Termination**
P2P3			
Number of identified sites	0.12	0.13	0.11
Actual accrual	0.14	0.12	0.21
Duration	0.18	0.21	0.05
Target accrual	0.27	0.31	0.09
Locations	0.02	0.02	0.02
Study design keywords	0.16	0.17	0.10
Trial outcomes	0.56	0.67	0.10
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) reports missing data for some of their factors, and address 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 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 is 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 Appendix 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 are mainly accounted for by drug development and trial statuses respectively. This is supported by our observations in Table 5 and Table 6 where the missingness proportions for some features differ greatly depending on the outcome.

Our assumption of MAR is also 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, more data is released, and information about its characteristics becomes more readily available. 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. We note that *Pharmaprojects* and *Trialtrove* are originally meant for tracking drug and trial activities. Thus, the databases 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. 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. Both parts are performed in R version 3.2.3. The specific components of our analysis are illustrated in Fig 4.

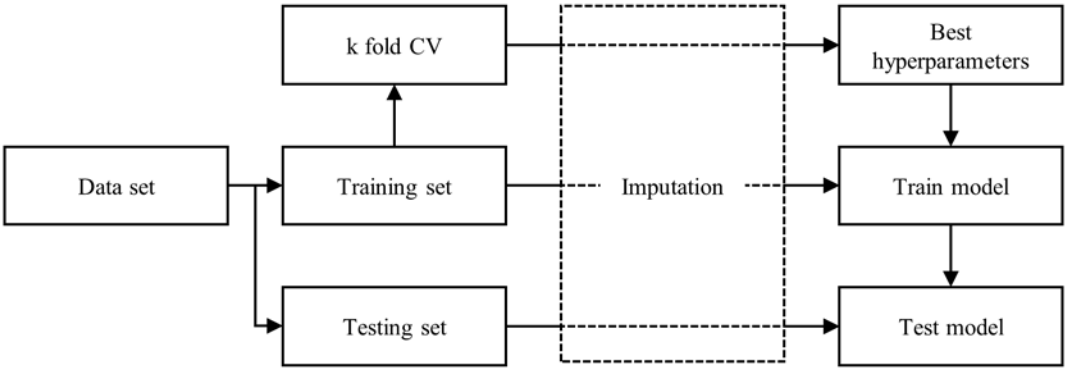

Fig 4. Modeling methodology adopted in this study. Abbreviations: CV: cross validation.

We formulate our three scenarios as supervised binary classification 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.

Feature matrices are created from the datasets by representing drug and trial features for each drug-indication pair as vectors (see Fig 5). 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 5. Consider the drug-indication pair Analiptin-diabetes type 2 in the P2APP dataset. It was observed to have two phase 2 trials in *Trialtrove*, and thus is represented by two vector rows. 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. Feature matrices for the training and testing sets are formed according to the drug-indication pairs in each set. The testing sets are meant to be out-of-sample datasets to evaluate our models. Therefore, we mask their outcomes (that is, we treat them as unknown) and will access them only at the very end to check our performance.

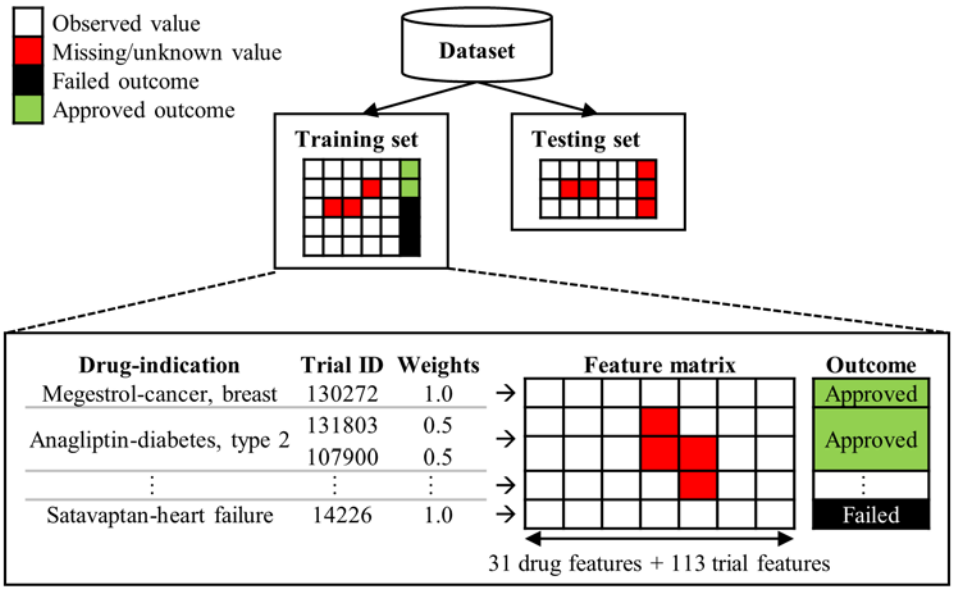

Fig 5. Feature matrix of dataset. Each row corresponds to a feature vector; each feature corresponds to an entry in the vector; each vector has a length of 144 since we have 31 drug and 113 trial features. Feature vectors of all drug-indication pairs in the dataset form the feature matrix collectively. Trial ID is a unique trial identifier in *Trialtrove*.

To deal with missing data in both training and test sets, we considered 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 Appendix 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 (Enders, 2010; Collins et al., 2001; Rubin, 1996; Schafer and 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 hypothesized that missingness is mainly accounted 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), support vector machine with radial basis functions (SVM), and decision trees C5.0. The first three algorithms are implemented in the Python scikit-learn package (Pedregosa et al., 2011) and the fourth is implemented using the C50 package in R (Kuhn et

al., 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 5). 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.

In Appendix E, we describe simulation experiments designed to evaluate the impact of our imputation methods and machine-learning algorithms. These results confirm the fact that imputation does offer improved fit and predictive power over listwise deletion. Moreover, we find kNN imputation (with k = 5), in combination with an RF classifier, to be most effective methods for our datasets and use this approach (5NN-RF) for our analysis.

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 be kept as similar to the testing process 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 three scenarios, using the standard "area under the 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]).

3 Results

Predicting phase transitions and approvals

We analyze the three datasets (P2P3, P2APP and P3APP) by first splitting each into a training set (30%) and a testing set (70%) 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 above. We repeat this experiment 100 times for robustness. Table 7 summarizes the AUC performance metrics

for the testing sets. On average, we achieve 0.74 AUC for P2P3, 0.78 AUC for P2APP and 0.81 AUC for P3APP. It seems that predicting phase transitions is more challenging than predicting drug approval.

Table 7. 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%**
P2P3										
All	0.737	0.018	0.707	0.741	0.764	-	-	-	-	-
Anti-cancer	0.745	0.028	0.700	0.745	0.788	0.779	0.027	0.739	0.780	0.818
Rare Diseases	0.752	0.041	0.685	0.755	0.818	0.747	0.035	0.692	0.743	0.807
Neurological	0.776	0.034	0.716	0.778	0.835	0.769	0.034	0.709	0.767	0.826
Alimentary	0.733	0.034	0.679	0.736	0.789	0.727	0.042	0.657	0.726	0.799
Immunological	0.715	0.067	0.604	0.723	0.826	0.764	0.058	0.675	0.765	0.859
Anti-infective	0.693	0.066	0.594	0.695	0.797	0.752	0.052	0.670	0.756	0.836
Respiratory	0.693	0.059	0.592	0.699	0.774	0.733	0.058	0.620	0.735	0.818
Musculoskeletal	0.766	0.055	0.668	0.768	0.853	0.720	0.069	0.623	0.725	0.822
Cardiovascular	0.677	0.066	0.565	0.677	0.780	0.615	0.061	0.528	0.609	0.741
Genitourinary	0.719	0.082	0.579	0.729	0.836	0.686	0.084	0.543	0.694	0.804
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 the datasets used have already been affected by backfilling. In Appendix E, we highlight the perils of relying on the MAR testing set for model validation, and suggest that the gold standard and MCAR testing sets AUCs 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 MAR and MCAR testing set AUCs 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 P2P3 and 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 statuses 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 P2P3 and P2APP; phase 3 for P3APP), be terminated (failed), or have progressed to higher phases (advanced).

Fig 6, Table 8 and Table 9 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 6, we observe peaks at the lower end of the score spectrum for failed pairs (red) for all three 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 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 P2P3 and P2APP). From Table 8, 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 9, 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. For P2APP and P3APP, 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.

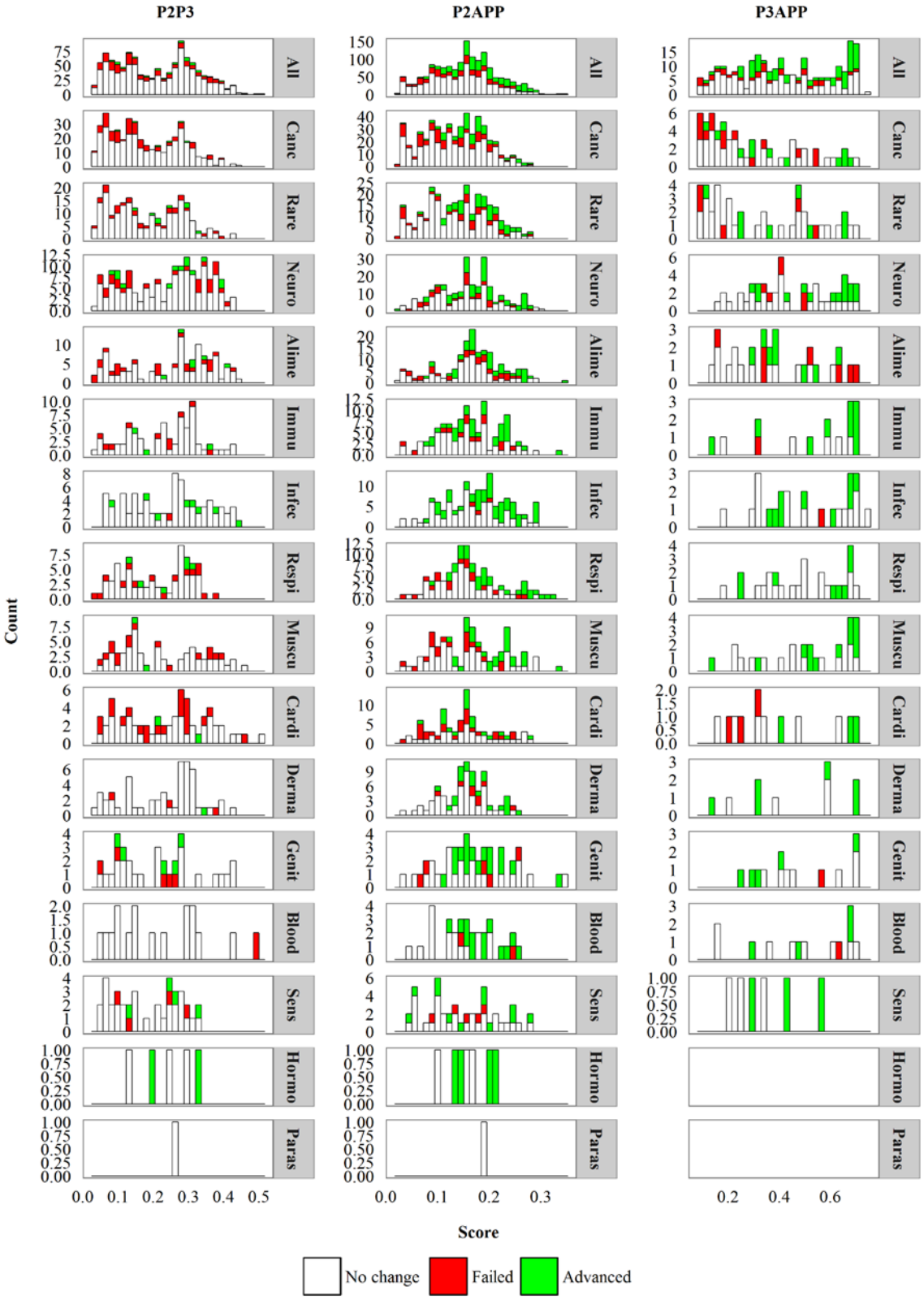

Fig 6. Distributions of prediction scores for P2P3, P2APP and P3APP. First row for all indication groups in aggregate. Subsequent rows for specific indication groups.

Table 8. Distributions of prediction scores for all indication groups in aggregate (see Fig 6). Advanced refers to progress to a higher phase from the original phase. Original phase for P2P3 and P2APP is phase 2; for P3APP is phase 3. For instance, out of 1,105 drug-indication pairs in the P2P3 testing set, 858 pairs are still pending decision in phase 2, 194 pairs have failed and 53 pairs have successfully advanced to phase 3 testing. Abbreviations: n: sample size.

	Prediction Scores					
	n	Avg	Sd	5%	50%	95%
P2P3						
Aggregate	1,105	0.209	0.109	0.054	0.211	0.387
No change	858	0.211	0.108	0.054	0.216	0.388
Failed	194	0.191	0.112	0.052	0.157	0.375
Advanced	53	0.249	0.095	0.098	0.262	0.390
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 9. Distributions of prediction scores for all indication groups in aggregate (see Fig 6). Proportion refers to the fraction of samples that advanced to a later phase from the original phase.

Scores	n	Proportion
P2P3		
< 0.1	61	0.082
0.1-0.2	66	0.167
0.2-0.3	60	0.317
≥ 0.3	60	0.300
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

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 et al., 1984). Table 10 summarizes the results.

Table 10. Top ten important variables of 5NN-RF classifiers for P2P3, P2APP and P3APP. Average and standard deviation taken across all experiments.

	Importance	
	Avg	**Sd**
P2P3		
Trial outcome – completed, positive outcome or primary endpoint(s) met	0.345	0.031
Trial status	0.120	0.015
Prior approval of drug for another indication	0.040	0.014
Actual accrual	0.040	0.010
Duration	0.038	0.011
Sponsors track record – number of positive phase 2 trials	0.037	0.009
Number of identified sites	0.035	0.009
Sponsors track record – number of positive phase 1 trials	0.032	0.012
Target accrual	0.024	0.007
Trial outcome – completed, negative outcome or primary endpoint(s) not met	0.018	0.007
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
Sponsors track record – number of positive phase 3 trials	0.023	0.006
Sponsors 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 all three 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 phase 2 to phase 3 or 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 sponsors' 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, be it research and development or regulatory prowess.

Since drugs developed for different indication groups may have very different characteristics, we might expect classifiers trained on indication-group specific data to outperform the 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 7 shows the results for selected indication groups. Unfortunately, we find that not all indication groups benefit from such specialization. There are specialized classifiers that perform significantly better than general classifiers (e.g. respiratory in P2APP), and ones that perform more poorly (e.g. musculoskeletal in P2APP).

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 Appendix 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 three datasets, P2P3, P2APP, and P3APP (see Fig 7). 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 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.

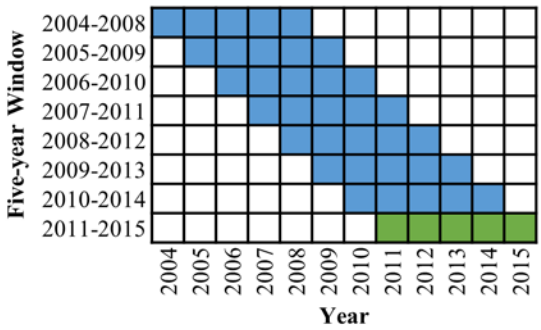

Fig 7. Time-series walk-forward analysis approach. The testing set in the last window (green) comprises drug-indication pairs in the pipeline at the time of snapshot of the databases.

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 8 summarizes the results of the time-series analysis for the first seven windows. We observe an increasing trend over the years for 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). There is a slight dip in the first few windows of P2P3, but the performance subsequently picks up in the last few periods (0.71 in the first and 0.85 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 0.82 for P2P3, 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 8, the proportions have been increasing over the years for all three datasets. This is likely due to better data reporting practices by drug developers, a possible consequence of FDAAA.

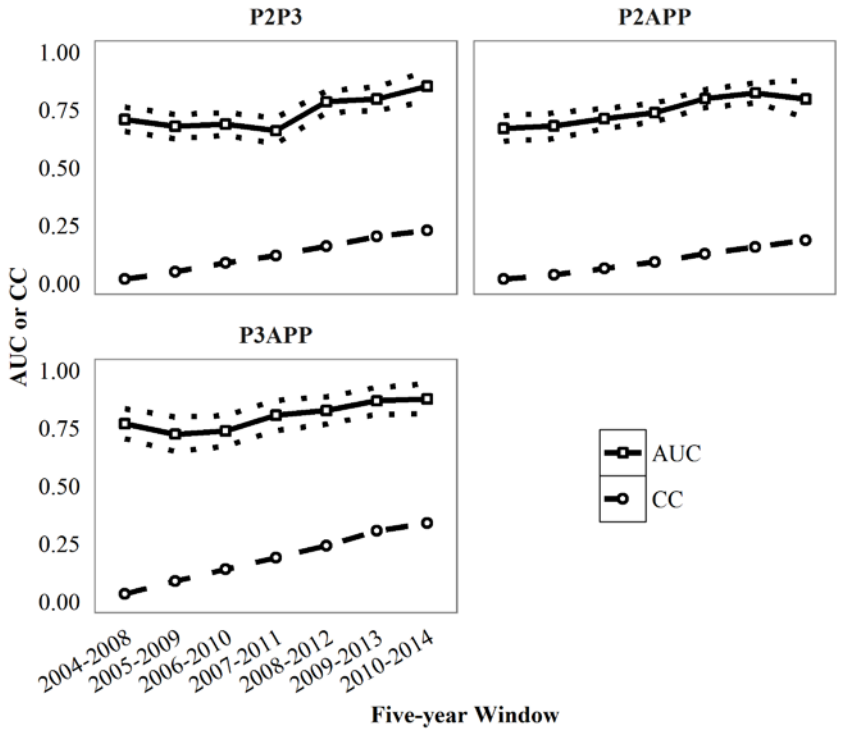

Fig 8. Time-series walk-forward analysis for P2P3, P2APP and P3APP using 5NN-RF. We use bootstrapping to determine the 95% CI for AUC (dotted lines). The dashed lines plot the corresponding proportions of complete cases in the training sets of each five-year window. Abbreviations: CC: proportion of complete cases.

Next, we examine the 2011–2015 window. Fig 9, Table 11 and Table 12 summarize the distributions of prediction scores for the P2P3, 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.

Fig 9. Distributions of prediction scores of the 2011-2015 window testing set for P2P3, P2APP and P3APP. First row for all indication groups in aggregate. Subsequent rows for specific indication groups.

Table 11. Distributions of prediction scores for all indication groups in aggregate (see Fig 9). Advanced refers to progress to a higher phase from the original phase. Original phase for P2P3 and P2APP is phase 2; for P3APP is phase 3.

	Prediction Scores					
	n	Avg	Sd	5%	50%	95%
P2P3						
Aggregate	920	0.215	0.139	0.043	0.193	0.478
No change	711	0.216	0.137	0.043	0.201	0.469
Failed	161	0.186	0.139	0.041	0.138	0.496
Advanced	48	0.297	0.136	0.095	0.314	0.500
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 12. Distribution of prediction scores for all indication groups in aggregate (see Fig 9). Proportion refers to the fraction of samples that advanced to a higher phase from the original phase.

Scores	n	Proportion
P2P3		
< 0.1	59	0.051
0.1-0.2	59	0.169
0.2-0.3	33	0.242
≥ 0.3	58	0.466
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

Table 13 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), sponsors track record and drug medium appear frequently in all three scenarios.

Table 13. Top ten important variables in 5NN-RF classifiers for P2P3, P2APP and P3APP. Average and standard deviation taken across the eight rolling windows.

	Importance	
	Avg	Sd
P2P3		
Trial outcome – completed, positive outcome or primary endpoint(s) met	0.261	0.128
Trial status	0.088	0.016
Prior approval of drug for another indication	0.085	0.080
Actual accrual	0.042	0.007
Target accrual	0.040	0.016
Number of identified sites	0.038	0.013
Trial outcome – completed, negative outcome or primary endpoint(s) not met	0.030	0.023
Duration	0.030	0.010
Sponsor type – academic	0.026	0.024
Sponsor track record – number of positive phase 2 trials	0.017	0.011
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 14, Table 15 and Table 16 summarize the results for selected indication groups in P2P3, P2APP and P3APP respectively (see Appendix 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 anti-infective indication group in P2P3). 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 14. Comparison of the general and indication-group specific classifiers for selected indication groups in P2P3. 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	1,278	420	0.708 (0.655, 0.761)	-	-	-
2005-2009	1,442	455	0.678 (0.626, 0.730)	-	-	-
2006-2010	1,634	467	0.688 (0.639, 0.737)	-	-	-
2007-2011	1,790	433	0.659 (0.602, 0.716)	-	-	-
2008-2012	1,853	447	0.784 (0.737, 0.832)	-	-	-
2009-2013	1,921	385	0.797 (0.746, 0.847)	-	-	-
2010-2014	1,933	274	0.852 (0.787, 0.917)	-	-	-
Anti-cancer						
2004-2008	1,278	134	0.688 (0.593, 0.783)	461	134	0.719 (0.630, 0.808)
2005-2009	1,442	146	0.639 (0.541, 0.738)	491	146	0.635 (0.527, 0.742)
2006-2010	1,634	151	0.684 (0.589, 0.779)	541	151	0.691 (0.595, 0.786)
2007-2011	1,790	156	0.671 (0.565, 0.777)	589	156	0.767 (0.675, 0.859)
2008-2012	1,853	154	0.756 (0.646, 0.867)	631	154	0.736 (0.629, 0.843)
2009-2013	1,921	136	0.801 (0.685, 0.917)	662	136	0.841 (0.736, 0.945)
2010-2014	1,933	119	0.898 (0.768, 1.000)	686	119	0.863 (0.703, 1.000)
Anti-infective						
2004-2008	1,278	29	0.740 (0.495, 0.986)	88	29	0.461 (0.212, 0.710)
2005-2009	1,442	36	0.716 (0.531, 0.900)	102	36	0.572 (0.376, 0.768)
2006-2010	1,634	47	0.631 (0.470, 0.793)	118	47	0.576 (0.412, 0.741)
2007-2011	1,790	40	0.685 (0.518, 0.853)	135	40	0.442 (0.258, 0.627)
2008-2012	1,853	42	0.800 (0.664, 0.936)	142	42	0.634 (0.463, 0.806)
2009-2013	1,921	29	0.717 (0.526, 0.909)	149	29	0.545 (0.317, 0.773)
2010-2014	1,933	17	0.962 (0.874, 1.000)	153	17	0.808 (0.567, 1.000)

Table 15. 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 16. 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 groups. 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 groups 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 our models. Table 17 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 (highlighted in bold) have advanced beyond phase 2 testing since our analysis, indicating predictive power of our models. Ultimately, such scores can be used by portfolio managers to rank and evaluate the potential risks and rewards of drug candidates.

Table 17. 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*

Drug	Indication	Score
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 is 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 three datasets, P2P3, P2APP and P3APP, from Informa® databases and apply 5NN imputation to make efficient use of all available data. We apply machine-learning techniques to train and validate our RF predictive models and achieve promising levels of predictive power for all three 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 all three datasets shows generally increasing trends in performance over five-year rolling windows from 2004 to 2014. We find that the classifiers' performances correlate 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 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. 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; 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 et al., 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 risk assessment and portfolio management of investigational drugs at different clinical stages.

5 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.

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.

Appendix

A Data pre-processing

We construct our datasets from two Informa® databases: *Pharmaprojects* and *Trialtrove*, two separate relational databases organized by largely different ontologies. We extract drug-specific features and drug-indication development status from *Pharmaprojects*, and clinical trial features from *Trialtrove*. We had to merge the databases through keys provided separately by Informa®.

Pharmaprojects was created earlier than *Trialtrove*, and thus the disease coverage for clinical trials is not as extensive. We start the merging process by first identifying all drug-indication pairs in *Pharmaprojects*. Subsequently, we drop pairs that do not have any trials recorded in *Trialtrove*. As highlighted in Section 2, profiles in *Pharmaprojects* and *Trialtrove* are fraught with missingness. Therefore, we impose several filters when constructing the datasets to ensure that all instances collected are usable for analysis. Table 18 summarizes the steps in the filter. We note that the drug, indication, and trial relationships in the constructed datasets are surjective and non-injective: different drugs may target the same indication, and some trials may involve multiple drug-indication pairs. This is logical because it is common that drugs treat multiple diseases, multiple drugs treat a specific disease, or trials involve two or more related primary investigational drugs. To provide some intuition for the size of these databases, we summarize, in Fig 10, Fig 11 and Fig 12 (for P2P3, P2APP and P3APP respectively), how the number of drug-indication pairs and clinical trials change as we perform the filters.

We extract drug compound attributes and clinical trial characteristics from *Pharmaprojects* and *Trialtrove*, respectively (see Table 1 and Table 19). In addition to features readily available in the databases, we create an augmented set of variables capturing sponsor track record and investigator experience. We quantify the track record of sponsors of a specific trial by their success in developing other drugs, using the number of prior approved and failed drug-indication developments; and in past trials for phases 1, 2, and 3 separately, using the total number of trials sponsored, the number of trials sponsored with positive and negative results, and the number of trials sponsored to completion and termination. We use the end date of the last trial of the drug-indication pair under consideration as the cutoff for considering prior experience. This is because the last end date will be the time of prediction. We abstract investigator experience in the same manner. Lastly, we construct a binary drug-indication pair feature, whether the drug has been approved for another indication before. Similarly, we use the end date of the last trial as cutoff for considering

prior approval. In total, our datasets have 31 drug-related features and 113 trial-related features.

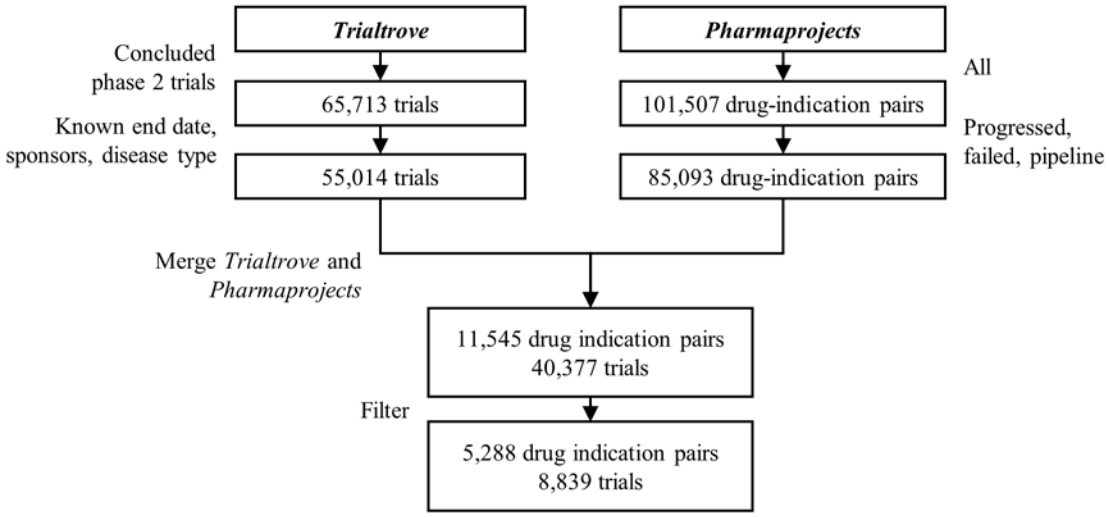

Fig 10. P2P3 data filtering.

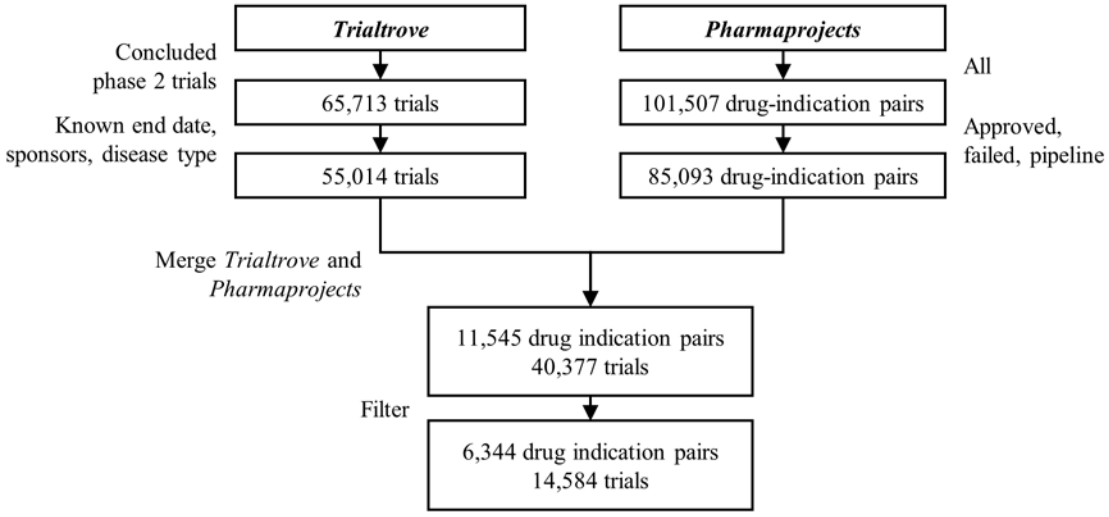

Fig 11. P2APP data filtering.

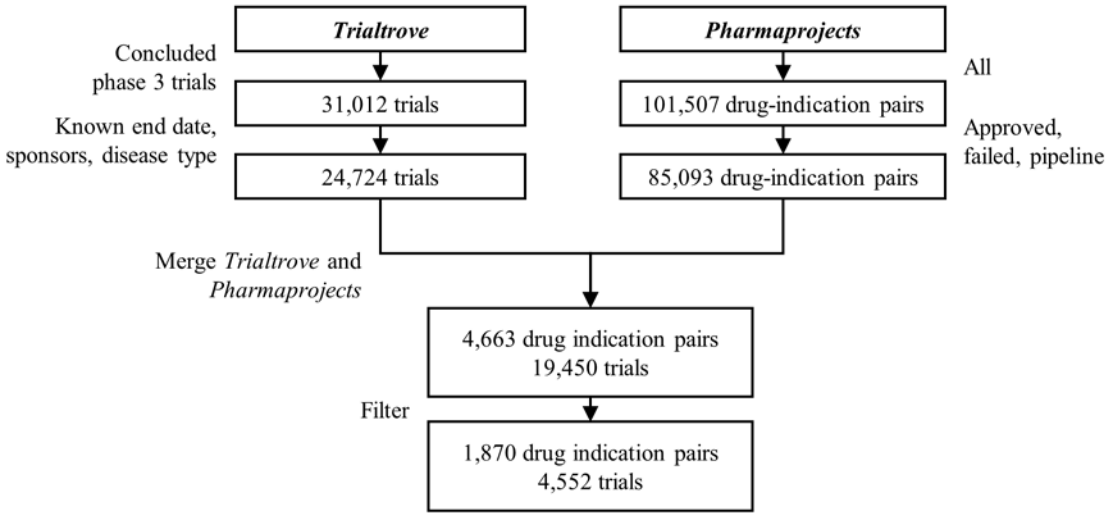

Fig 12. P3APP data filtering.

Table 18. Filters for creating datasets.

	Rationale
Drug-indication Pairs in *Pharmaprojects*	
Trials observed in *Trialtrove* (phase 2 for P2P3 and P2APP; phase 3 for P3APP)	We exclude pairs for which we do not observe any trials in *Trialtrove*.
Known approval date (for P2APP and P3APP, if approved) Approval dates are not available directly in *Pharmaprojects*. They are embedded within text blocks. We had to mine these text blocks (combination of heuristics and manual extraction) to extract the dates.	We define the approval date as the earliest date a drug-indication pair was approved in any market. We need these dates to create an augmented set of variables capturing sponsors and investigators experience, and also to perform time-series analysis.
Known phase 3 start date (for P2P3)	We define the phase 3 start date as the earliest date a drug-indication pair enters phase 3 testing.
Known failure date (if failed)	Failure dates are not directly available in *Pharmaprojects*. We define failure date as one year after the end-date of the last phase 2 or phase 3 trial (if any), whichever is latest.
Clinical Trials in *Trialtrove*	
Phase 2 for P2P3 and P2APP; phase 3 for P3APP	We are interested in predicting approvals using trial features.
Known end date	We need these dates to perform time series analysis. For approved drug-indication pairs in P2APP and P3APP, we compare the trial end date with the corresponding approval date to filter out post-approval trials. These trials may be for supplemental new drug applications (e.g. modified dosage) that are irrelevant to our analysis. Similarly, for successful drug-indication pairs in P2P3, we compare the phase 2 trial end date with the corresponding phase 3 start date to filter out trials that end post-transition.
Known sponsors and disease types	Trials not tagged with sponsor/disease types are typically out of *Trialtrove* commercial coverage and are not maintained.

Table 19. Examples of features extracted from *Pharmaprojects* and *Trialtrove*. After transforming multi-label parent features into binary child features (1 or 0), there were over 3,000 drug and trial categories in total. However, not all are useful for our analysis. For instance, trials rarely take place in Nepal, so the corresponding location feature rarely appears. Thus, this feature is unlikely to have meaningful associations with success. We remove these near zero variance factors. Also, we standardize continuous variables prior to all experiments.

	Examples	Categories
Drug Features		
Route	Inhaled; Injectable; Oral; Topical	4
Origin	Biological, protein, antibody; Biological, protein, recombinant; Chemical, synthetic	3
Medium	Capsule, hard; Capsule, soft; Powder; Solution; Suspension; Tablet	6
Biological target family	Cytokine/Growth factor; Enzyme; Ion channel; Receptor; Transporter	5
Pharmacological target family	5 Hydroxytryptamine receptor antagonist; Angiogenesis inhibitor; Apoptosis stimulant; Cell cycle inhibitor; DNA inhibitor; DNA synthesis inhibitor; Growth factor receptor antagonist; Immunostimulant; Immunosuppressant; Ion channel antagonist; Protein kinase inhibitor	11
Drug-indication development status	True; false	2
Prior approval of drug for another indication	Approved; failed	2
Trial Features		
Duration	Integer	1
Study design	Active comparator; Cross over; Dose response; Double blind/blinded; Efficacy; Multiple arm; Non-inferiority; Open label; Pharmacodynamics; Pharmacokinetics; Placebo control; Randomized; Safety; Single arm	14
Sponsor type	Academic; Cooperative Group; Government; Industry, all other pharma; Industry, Top 20 Pharma	5
Therapeutic area	Autoimmune/Inflammation; Cardiovascular; CNS; Infectious Disease; Metabolic/Endocrinology; Oncology	6
Trial status	Completed; terminated	2
Trial outcome	Completed, Negative outcome/primary endpoint(s) not met; Completed, Outcome indeterminate; Completed, Positive outcome/primary endpoint(s) met; Terminated, Business decision - Other; Terminated, Business decision - Pipeline reprioritization; Terminated, Lack of efficacy; Terminated, Poor enrollment; Terminated, Safety/adverse effects	8
Target accrual	Integer	1
Actual accrual	Integer	1
Locations	Argentina; Australia; Austria; Belgium; Brazil; Bulgaria; Canada; Chile; Czech Republic; Denmark; Europe; Finland; France; Germany; Hungary; India; Israel; Italy; Japan; Mexico; Netherlands; New Zealand; Peru; Poland; Romania; Russia; Slovakia; South Africa; South Korea; Spain; Sweden; Switzerland; Taiwan; Ukraine; United Kingdom; United States	36
Number of identified sites	Integer	1
Biomarker involvement	Biomarker/Efficacy; Biomarker/Toxicity; PGX - Biomarker Identification/Evaluation; PGX - pathogen; PGX - Patient Preselection/Stratification	5
Sponsor track record	Number of prior approved drug-indication pairs; Number of prior failed pairs; Total number of phase 1 trials sponsored; Number of phase 1 trials with positive results; Number of phase 1 trials with negative results; Number of completed phase 1 trials; Number of terminated phase 1 trials; Total number of phase 2 trials sponsored; Number of phase 2 trials with positive results; Number of phase 2 trials with negative results; Number of completed phase 2 trials; Number of terminated phase 2 trials; Total number of phase 3 trials sponsored; Number of phase 3 trials with positive results; Number of phase 3 trials with negative results; Number of completed phase 3 trials; Number of terminated phase 3 trials	17
Investigator experience	Refer to sponsor track record	17

B Missing data definitions

Missing data may be generally classified into three categories (Rubin, 1976): missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR holds when data is missing for reasons entirely unrelated to the data, that is, when the probability of missingness is the same for every data point. MAR applies when data missingness can be fully accounted for by observed variables, i.e. when the probability of missingness is the same when conditioned on groups in the observed data. Finally, MNAR comes in when neither MCAR nor MAR is appropriate, when the probability of missingness is dependent on the value of the unobserved variable, or is unknown (Van Buuren, 2012).

For a more precise definition, let Y denote a $n \times p$ data matrix (with elements y_{ij}) where the n rows represent samples and the p columns represent variables. We further partition the observed part of Y as Y_{obs} and the missing part of Y_{mis}, so collectively $Y = (Y_{obs}, Y_{mis})$. Next, let R be a $n \times p$ response indicator matrix where elements $r_{ij} = 0$ if the corresponding element y_{ij} is missing and $r_{ij} = 1$ if y_{ij} is observed. The distribution of R, known as the missing data model/missingness mechanism, may be written generally as $P(R|Y_{obs}, Y_{mis}, \xi)$. R is related in some way to the data Y and is described by some unknown parameters ξ. The missingness is said to be MCAR if $P(R|Y_{obs}, Y_{mis}, \xi) = P(R|\xi)$. This means that the probability of missingness is totally unrelated to the data. The missingness is said to be MAR if $P(R|Y_{obs}, Y_{mis}, \xi) = P(R|Y_{obs}, \xi)$. This means that the missingness does not depend on the values of the missing data when conditioned on the observed data. Finally, the missingness is said to be MNAR if $P(R|Y_{obs}, Y_{mis}, \xi) = P(R|Y_{obs}, Y_{mis}, \xi)$. The expression cannot be simplified, since the probability of missingness can depend on the unobserved underlying values of the missing data and/or of other observed variables.

Now, we let the distribution of Y, which is the data model we are interested in, be described by some parameters θ. The missingness mechanism can be further described as ignorable under two conditions (Little and Rubin, 2014): First, the missingness must be MAR. Second, the parameters θ and ξ must be distinct[2]. In many situations, the second condition is reasonable because knowing θ will provide little information about ξ and vice versa (Schafer, 1997). In general, the first MAR requirement is considered to be the more important condition. When ignorability holds, Rubin (2004) showed that $P(Y_{mis}|Y_{obs}, R) = P(Y_{mis}|Y_{obs})$. This implies that the distribution of the data is independent of the missing data model, and is identical in both the observed and unobserved groups (Van Buuren, 2012) $P(Y|Y_{obs}, R = 1) = P(Y|Y_{obs}, R = 0)$. In this case, we can model the conditional distribution $P(Y|Y_{obs}, R = 1)$ from the observed data, and use it to draw imputations for the missing data. In other words, the missing data model R is ignored and not modeled. If the missingness is nonignorable, the last equation does not hold, and the distributions are

[2] θ and ξ should be a priori independent where $P(\theta, \xi)$ factors into $P(\theta)P(\xi)$ (Little and Rubin, 2014).

not equivalent. When this happens, we need to estimate the missingness mechanism, and incorporate it into the imputation model.

C Imputation methods

Listwise deletion

In listwise deletion, we discard all observations with missing data, in which case there is no imputation. This method is generally not recommended because it is valid only under strict MCAR conditions, which rarely hold in practice. Nevertheless, we can use this as comparison against other methods.

Unconditional mean imputation

In unconditional mean imputation, we fill in the missing values of a variable with the mean/mode of the observed cases of that variable. This method is also highly discouraged because it distorts the data distribution by reducing variability and undermining relationships between variables. In this study, we implement two variants: mean/mode and median/mode imputation.

k-Nearest neighbor imputation

In kNN imputation, given an instance with missing values, we select the k most similar cases that do not have missing values in the features to be imputed. As the name suggests, the replacements for the missing values are chosen from these k nearest neighbors. In this paper, we use the Gower distance for mixed variables[3] and explore five and ten nearest neighbors. For each missing value to be imputed, we use the median/mode of the corresponding feature of the k closest neighbors as the imputation.

Multiple imputation

MI is a principled missing data method that involves three steps: imputation, analysis, and pooling. In the first step, we specify an imputation model for each incomplete variable in the form of a conditional distribution, that is, missing data conditioned on the observed data. Then we draw multiple plausible values for each missing data point according to the specified variable models, creating multiple imputed datasets from one incomplete dataset. In this study, we specify linear regression models for continuous variables and logistic

[3] Implemented in R, VIM package (Templ et al., 2015).

regression models for nominal/categorical variables. In the second step, we analyze each imputed dataset individually using standard statistical procedures. Finally, in the third step, we pool the estimates obtained from the multiple individual analyses (e.g. probability predictions, regression coefficients) using Rubin's rules (Rubin, 2004) to yield a single estimate. See Appendix D for more details on MI.

Decision tree algorithm

Decision trees are commonly used as predictive models. In contrast to most machine learning algorithms, some decision tree algorithms can handle missing values internally without the need for imputation. In this paper, we focus on the C5.0 algorithm[4]. C5.0 is a tree-based model developed by Quinlan (1998). It uses entropy as the node impurity measure. When considering a variable for a split, C5.0 uses only examples for which that variable is not missing to calculate the node impurity. When an instance sent down C5.0 encounters a split variable for which it has a missing value, it is split into the branches fractionally, according to the split proportion of the observed instances.

D Notes on multiple imputation

Multiple imputation (MI) is a principled missing data method that can provide valid statistical inferences when missingness is ignorable. It involves three steps: imputation, analysis and pooling (see Fig 13).

Imputation

Under MI, we draw multiple plausible values for each missing data point, thus creating multiple imputed datasets from one incomplete dataset. There are different strategies for multivariate multiple imputation. In this paper, we focus on Fully Conditional Specification (FCS), specifically the Multivariate Imputation by Chained Equations (MICE) algorithm[5] (Buuren and Groothuis-Oudshoorn, 2011). In MICE, we first specify an imputation model for each incomplete variable in the form of conditional distributions, missing data conditioned on the observed data. The algorithm starts with simple random draws from the observed data and imputes the incomplete data in an iterative variable-by-variable fashion according to the specified variable models. Each iteration entails one cycle through all the incomplete variables (see Fig 14). The number of iterations should be set such that

[4] Implemented in R, C50 package (Kuhn et al., 2014).
[5] Implemented in R, MICE package (Buuren and Groothuis-Oudshoorn, 2011). Van Buuren provides a comprehensive guide to MICE in Van Buuren (2012).

convergence is reached. This is typically checked by monitoring the means of imputed values and/or the values of regression coefficients and making sure they are stable over the iterations. In practice, a small number of iterations appears to be sufficient, from 10 to 20. Multiple imputed datasets can be generated by running MICE in parallel the desired number of times.

In this study, we specify linear regression models for incomplete continuous variables and logistic regression models for incomplete nominal variables. We monitor convergence by computing the mean/mode of the imputed values and making sure that they were stable over iterations. Twenty iterations appear to be sufficient.

Analysis

The analysis after a single imputation is straightforward: We apply any standard, complete-data statistical methods and end up with one set of results. In MI, we have multiple imputed compete datasets. After analyzing them individually using standard statistical procedures, we end up with multiple sets of results. The differences between the sets represent the uncertainty due to the missing data. The pooling step describes how we can combine these sets of results into a single set.

Pooling

In this step, we pool the estimates obtained from multiple individual analyses using Rubin's rules (Rubin, 2004) to yield a single estimate. Let Q be a column vector of the estimands of interest, $\tilde{Q}$ be its estimate, m be the number of imputed datasets, and $\tilde{Q}_l$ be the estimate of the l^{th} repeated analysis. The combined estimate is given by $\bar{Q} = \frac{1}{m}\sum_{l=1}^{m} \tilde{Q}_l$. Estimates that can be combined using Rubin's rules include means, regression coefficients and probability predictions.

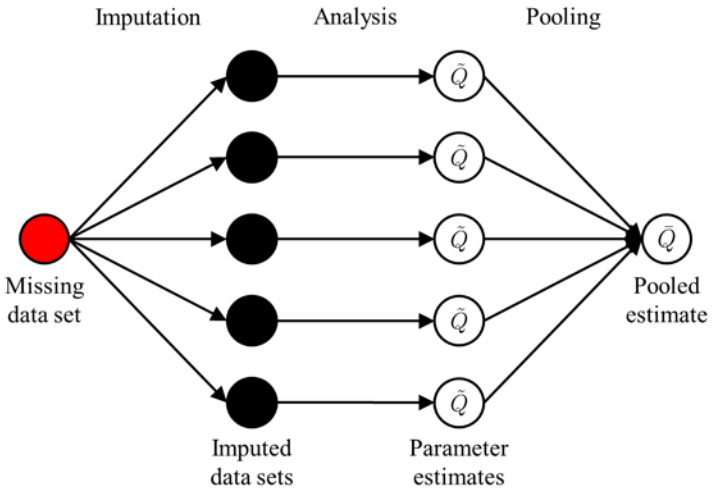

Fig 13. Multiple imputation.

Fig 14. Pseudo-code for Multivariate Imputation by Chained Equations (Van Buuren, 2012).

Algorithm: Multivariate Imputation by Chained Equations

Define Y as a $n \times p$ data matrix where rows represent samples and columns represent variables.
Data: Incomplete dataset $Y = (Y^{obs}, Y^{mis})$
Result: Incomplete dataset $Y^T = (Y^{obs}, Y^{mis,T})$ at iteration T
Define Y_j as the j^{th} feature column of Y where $Y_j = (Y_j^{obs}, Y_j^{mis})$
for $j \leftarrow 1$ to p do
| imputation model for incomplete variable $Y_j \leftarrow P(Y_j|Y_{-j}, \theta_j)$
∟ starting imputations $Y_j^{mis,0} \leftarrow$ draws from Y_j^{obs}
Define $Y_{-j}^t = (Y_1^t, Y_2^t, \cdots, Y_{j-1}^t, Y_{j+1}^{t-1}, \cdots, Y_{p-1}^{t-1}, Y_p^{t-1})$ where Y_j^t is the j^{th} feature at iteration t
for $t \leftarrow 1$ to T do
| for $j \leftarrow 1$ to p do
| | $\theta_j^t \leftarrow$ draw from posterior $P(\theta_j|Y_j^{obs}, Y_{-j}^t)$
∟ ∟ $Y_j^{mis,t} \leftarrow$ draws from posterior predictive $P(Y_j^{mis}|Y_{-j}^t, \theta_j^t)$
return Y^T

E Simulation of listwise deletion versus imputation

We design an experiment to study the effects of imputation and verify that imputation indeed offers an improvement over complete case analysis. First, we create a gold standard dataset by taking complete cases of the P2APP dataset[6] (see Table 20). Next, we randomly split the drug-indication pairs from the gold standard dataset into a training set (70%) and a testing set (30%).

To simulate the missingness present in the original dataset, we introduce missingness in the gold standard training and testing sets based on our MAR assumption and the missingness patterns observed in the P2APP dataset. When making them MAR, we ensure

[6] We exclude pipeline drug-indication pairs in this analysis because their outcomes are unknown.

that the proportions of drugs and trials with fully observed features (i.e. complete cases) are consistent with those in the parent dataset (see Appendix F for a description).

We must be cautious relying on the MAR testing set for model validation. Results may not accurately capture whether a classifier has learned the true underlying relationship between the features and the outcome. To illustrate, suppose that drug-indication pairs have only one binary feature ("0" or "1") that is unrelated to approval/failure. Thus, no classifier can do better than random guessing (0.5 AUC). Now, assume that we have MAR in the dataset: failed pairs are more likely to have missing values due to the data collection process, unrelated to the binary feature. Suppose that we impute all the missing values with 1. Intuitively, we know that this is a poor imputation method because it distorts the feature distribution of failed pairs, and it reduces the variability in the data. However, this is seemingly a "good" method because it allows the AUC of a classifier on this imputed dataset to exceed 0.5. That is, we can identify a disproportionate number of failures by guessing all pairs with feature value 1 as failures. The classifier has learned a nonexistent relationship introduced by the imputations. By predicting all 1s as failures, the classifier is implicitly exploiting its MAR-ness.

Some may argue that it is acceptable to use missingness as a signal. Unfortunately, this is inappropriate in our case, because the MAR nature of the dataset on hand is merely an artifact of data collection that would not be present during actual testing. MAR was introduced to the data due to the backfilling of information over time[7]. We believe that missingness in current test cases, e.g., drug-indication pairs currently in the pipeline, is more MCAR-like in nature because no backfilling has been performed. For example, immediately after phase 2 testing, pairs that go on to be approved are equally likely to have missing information as pairs that go on to be terminated. Clearly, missingness will not be a useful predictive factor. A classifier that relies heavily on the missingness in the dataset will fail miserably when put into production.

It is difficult to assess how good a classifier really is from the performance on a MAR testing set. Therefore, we create an additional testing set (the "MCAR testing set") in which we introduce missingness based on patterns observed in pipeline drug-indication pairs in the P2APP dataset (see Appendix F for a description). Because the drugs were still in development at the time of snapshot of the databases, they are likely to be less affected by backfilling. Consequently, the AUC on the MCAR testing set will be more reflective of a classifier's real performance. We also use the gold standard testing sets for evaluation. These two testing sets serve as a control for the backfilling artifact in the data collection process. They can help to identify non-ideal imputation methods: Poor imputation methods tend to distort the data distribution and undermine relationships between variables. This noise makes it more difficult for classifiers to learn the true underlying patterns in the data.

[7] This occurs due to a combination of reasons—some drug characteristics (e.g. mechanism of action) only become clear as the study progresses to higher phases; poor reporting practices.

These classifiers will perform poorly on the gold standard and MCAR testing sets[8]. On the other hand, applying imputation methods that are capable of preserving the data distribution will make it easier for classifiers to capture useful relationships in the data. These classifiers will perform well on the gold standard and MCAR testing sets.

We have two training sets (gold standard and MAR) and three testing sets (gold standard, MAR, and MCAR) (see Fig 15). We use five different missing data approaches, as described in Appendix C, to generate multiple complete training sets from the MAR training set. Subsequently, we use each imputed training set to build four different predictive models (PLR, RF, SVM, and C5.0) according to the methodology outlined in Section 2. We use ten-fold cross-validation to select the hyper-parameters for each model. In addition to the imputed MAR training sets, we use the gold standard training set to train gold standard classifiers: the models that would have been built if the data was complete. We impute the MAR and MCAR testing sets in a similar fashion as the training sets, and evaluate the AUC performance of all classifiers on the imputed and gold standard testing sets. We repeat the entire procedure of introducing MAR and MCAR in the dataset, imputing missingness, training models and validating performance 100 times for robustness. In addition to the AUC, we compute the biasness of the imputed values in the imputed training and testing sets with respect to their gold standard counterparts. This is a measure of accuracy of each imputation method. Finally, we use the results from the gold standard, MAR and MCAR testing sets as basis to select an imputation method and machine learning algorithm combination most suitable to the dataset on hand.

Table 22 summarizes the results. Since the training and testing sets are fixed, using the same drug-indication pairs for all methods, direct comparison across different missing data techniques and machine learning algorithms is possible. Each row corresponds to a different missing data technique used to process the training and testing sets in the experiments. Each column group corresponds to a different type of missingness introduced in the testing sets. For all four machine learning algorithms, we find that gold standard classifiers consistently outperform their complete case analysis and imputation counterparts. This is logical because useful information is invariably lost when we intentionally 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 are on average 0.04 less than those trained on imputed training sets. As expected, complete cases are ill suited for MAR data. This supports our conjecture that the use of imputation has allowed predictive models to learn useful patterns that would otherwise be lost from discarding incomplete data.

When comparing across rows, we observe that the different imputation techniques are not equally effective. In terms of imputation quality, MI and mean/mode give the most

[8] Returning to the above binary feature example, if we had tested the classifier on a gold standard testing set, we would realize that it did not learn any useful patterns.

inaccurate imputations while nearest neighbors recovers data best for both continuous and nominal variables (see Table 21). To better visualize each imputation method, Fig 16 plots the distributions of the trial feature of actual accrual, a continuous variable, in the gold standard, complete cases and imputed MAR training sets of one iteration. It is evident that mean and median imputations have distorted the variable distribution, introducing previously absent peaks at the observed mean and median respectively. In contrast, MI and nearest neighbors imputation managed to preserve the general shape of the variable distribution without introducing anomalous peaks.

We believe that the noise introduced by mean and median imputations have an adverse impact on a classifier's learning process. These effects may not be obvious from the AUC of the MAR testing sets. Indeed, for all four machine learning algorithms, we observe that mean and median imputations give the highest AUCs for the MAR testing sets. However, the trend is reversed when we look at the gold standard and MCAR testing sets. Classifiers trained on mean or median imputation performed the worst of all imputation methods on these testing sets, implying that the noise introduced by the distortions must have hindered the machine learning algorithms from fully capturing the underlying relationships in the data. It will therefore be prudent to avoid this imputation approach.

Overall, we find kNN imputation to be most suitable to the dataset[9]. It provides the least biased imputations among all missing data methods. More importantly, classifiers built on kNN-imputed training sets give the highest AUCs for the gold standard testing set for all machine learning models explored. By preserving the original data distribution while filling in missing values, kNN imputation has allowed classifiers to learn underlying patterns more effectively. In particular, the combination of 5NN with RF gives the one of the highest gold standard (0.805) and MCAR (0.780) testing set AUCs. This may be attributed to the fact that RF is a nonlinear model, and thus it is able to better capture the complex interactions between the features and regulatory approval than PLR, a linear model. We focus on the 5NN-RF combination in our analyses, since it appears that this pair is most compatible with our datasets.

[9] Note that the MI (m=10)-RF and MI (m=10)-C5.0 combinations yielded slightly better performances than kNN-RF. However, we excluded MI (m=10) from consideration because the improvement is only marginal while the imputation and analysis processes are much more time consuming, since we have ten imputed datasets in MI (m=10). Furthermore, the imputation method does not converge well (or at all) for smaller datasets. This poses an issue for the time series analysis in Section 3. In contrast, kNN imputation is relatively straightforward to implement and more stable.

Table 20. 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**

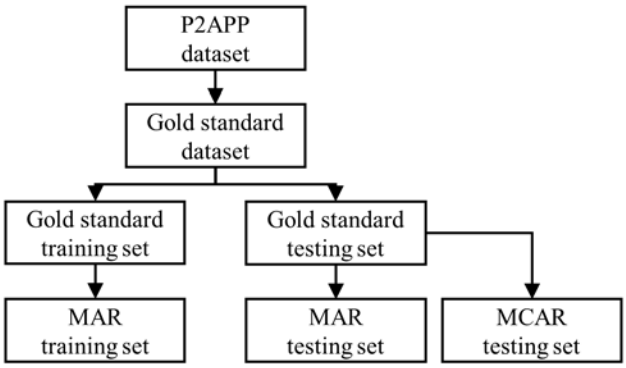

Fig 15. Datasets created in the experiment.

Table 21. Biasness of imputations with respect to gold standard. Abbreviations. Abs: absolute.

	MAR Training Set		MAR Testing Set		MCAR Testing Set	
	Bias[a]	**Wrongly Imputed[b]**	**Bias[a]**	**Wrongly Imputed[b]**	**Bias[a]**	**Wrongly Imputed[b]**
	%	**%**	**%**	**%**	**%**	**%**
Mean/mode	234.6	23.0	236.2	23.3	274.2	22.2
Median/mode	115.5	23.0	116.1	23.3	128.7	22.2
5NN	95.4	22.7	94.9	22.0	96.2	21.8
10NN	87.3	21.7	87.9	21.2	90.2	21.0
MI (m=1)	262.0	25.3	268.9	27.9	323.0	26.9
MI (m=10)	260.9	25.3	269.0	27.9	322.7	26.7

[a] Average percentage bias of imputed continuous variables. We first find the sum of the absolute percentage difference between imputed values that are continuous and their corresponding gold standard values (gold standard values as denominator), averaged over the total number of missing values that are continuous. We then take the mean over 100 iterations.

[b] Percentage of nominal variables that were wrongly imputed. We first find the number of imputed categorical values that differ from their corresponding gold standard values, averaged over the total number of missing values that are categorical. Next, we take the mean over 100 iterations.

Table 22. 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.

	Testing Set AUC														
	MAR					MCAR					Gold Standard				
	Avg	Sd	5%	50%	95%	Avg	Sd	5%	50%	95%	Avg	Sd	5%	50%	95%
PLR															
Gold Standard	-	-	-	-	-	-	-	-	-	-	0.810	0.028	0.761	0.808	0.853
Complete Cases	-	-	-	-	-	-	-	-	-	-	0.755	0.040	0.683	0.764	0.813
Mean/mode	0.786	0.028	0.746	0.785	0.829	0.751	0.029	0.702	0.753	0.794	0.778	0.031	0.729	0.779	0.823
Median/mode	0.786	0.028	0.745	0.786	0.829	0.751	0.029	0.704	0.753	0.794	0.778	0.031	0.728	0.779	0.824
5NN	0.763	0.032	0.716	0.762	0.814	0.757	0.032	0.707	0.758	0.805	0.786	0.032	0.738	0.787	0.834
10NN	0.774	0.030	0.730	0.773	0.821	0.757	0.032	0.695	0.756	0.802	0.787	0.032	0.739	0.791	0.835
MI (m=1)	0.746	0.035	0.688	0.747	0.804	0.758	0.035	0.705	0.755	0.818	0.781	0.036	0.722	0.777	0.843
MI (m=10)	0.755	0.030	0.705	0.757	0.801	0.766	0.032	0.719	0.764	0.815	0.782	0.031	0.729	0.782	0.831
RF															
Gold Standard	-	-	-	-	-	-	-	-	-	-	0.837	0.027	0.793	0.837	0.876
Complete Cases	-	-	-	-	-	-	-	-	-	-	0.764	0.048	0.685	0.772	0.830
Mean/mode	0.794	0.027	0.753	0.794	0.836	0.761	0.030	0.712	0.761	0.809	0.775	0.031	0.726	0.771	0.822
Median/mode	0.793	0.027	0.756	0.793	0.831	0.759	0.030	0.709	0.762	0.808	0.774	0.031	0.723	0.774	0.827
5NN	**0.782**	**0.031**	**0.735**	**0.783**	**0.830**	**0.780**	**0.030**	**0.734**	**0.783**	**0.828**	**0.805**	**0.033**	**0.755**	**0.805**	**0.857**
10NN	0.788	0.029	0.741	0.786	0.833	0.780	0.030	0.729	0.778	0.827	0.802	0.033	0.747	0.805	0.856
MI (m=1)	0.774	0.028	0.732	0.777	0.825	0.782	0.031	0.737	0.779	0.845	0.797	0.033	0.748	0.795	0.853
MI (m=10)	0.782	0.029	0.734	0.781	0.831	0.791	0.029	0.739	0.790	0.835	0.804	0.030	0.751	0.804	0.848
SVM															
Gold Standard	-	-	-	-	-	-	-	-	-	-	0.785	0.030	0.730	0.786	0.831
Complete Cases	-	-	-	-	-	-	-	-	-	-	0.733	0.053	0.650	0.741	0.795
Mean/mode	0.772	0.032	0.724	0.773	0.820	0.741	0.032	0.686	0.748	0.788	0.766	0.036	0.707	0.771	0.818
Median/mode	0.771	0.029	0.729	0.768	0.817	0.740	0.031	0.683	0.745	0.780	0.764	0.035	0.711	0.771	0.818
5NN	0.751	0.031	0.699	0.748	0.803	0.745	0.034	0.697	0.746	0.800	0.771	0.034	0.722	0.770	0.827
10NN	0.758	0.035	0.688	0.760	0.814	0.745	0.037	0.679	0.749	0.808	0.772	0.037	0.710	0.773	0.825
MI (m=1)	0.731	0.035	0.676	0.732	0.788	0.741	0.033	0.684	0.745	0.790	0.760	0.035	0.696	0.762	0.813
MI (m=10)	0.746	0.030	0.705	0.746	0.797	0.755	0.031	0.707	0.753	0.797	0.768	0.030	0.719	0.764	0.813
C5.0															
Gold Standard	-	-	-	-	-	-	-	-	-	-	0.800	0.033	0.758	0.800	0.844
Complete Cases	-	-	-	-	-	-	-	-	-	-	0.710	0.063	0.585	0.713	0.802
Mean/mode	0.764	0.033	0.711	0.768	0.810	0.734	0.032	0.675	0.737	0.777	0.758	0.039	0.698	0.762	0.816
Median/mode	0.764	0.038	0.708	0.761	0.825	0.735	0.041	0.676	0.736	0.797	0.754	0.043	0.679	0.751	0.823
5NN	0.756	0.036	0.703	0.753	0.816	0.749	0.038	0.695	0.745	0.805	0.772	0.038	0.715	0.772	0.843
10NN	0.759	0.035	0.696	0.762	0.807	0.747	0.037	0.687	0.749	0.799	0.770	0.035	0.710	0.771	0.822
MI (m=1)	0.733	0.038	0.672	0.731	0.795	0.741	0.036	0.680	0.740	0.800	0.758	0.037	0.701	0.754	0.819
MI (m=10)	0.786	0.030	0.738	0.786	0.836	0.793	0.031	0.738	0.797	0.842	0.807	0.031	0.756	0.808	0.857
MAR[a]	0.759	0.037	0.699	0.759	0.811	0.744	0.037	0.685	0.741	0.801	0.761	0.037	0.705	0.757	0.812

[a] For MAR, we leave the missingness as it is and rely on the decision tree algorithm to handle them internally.

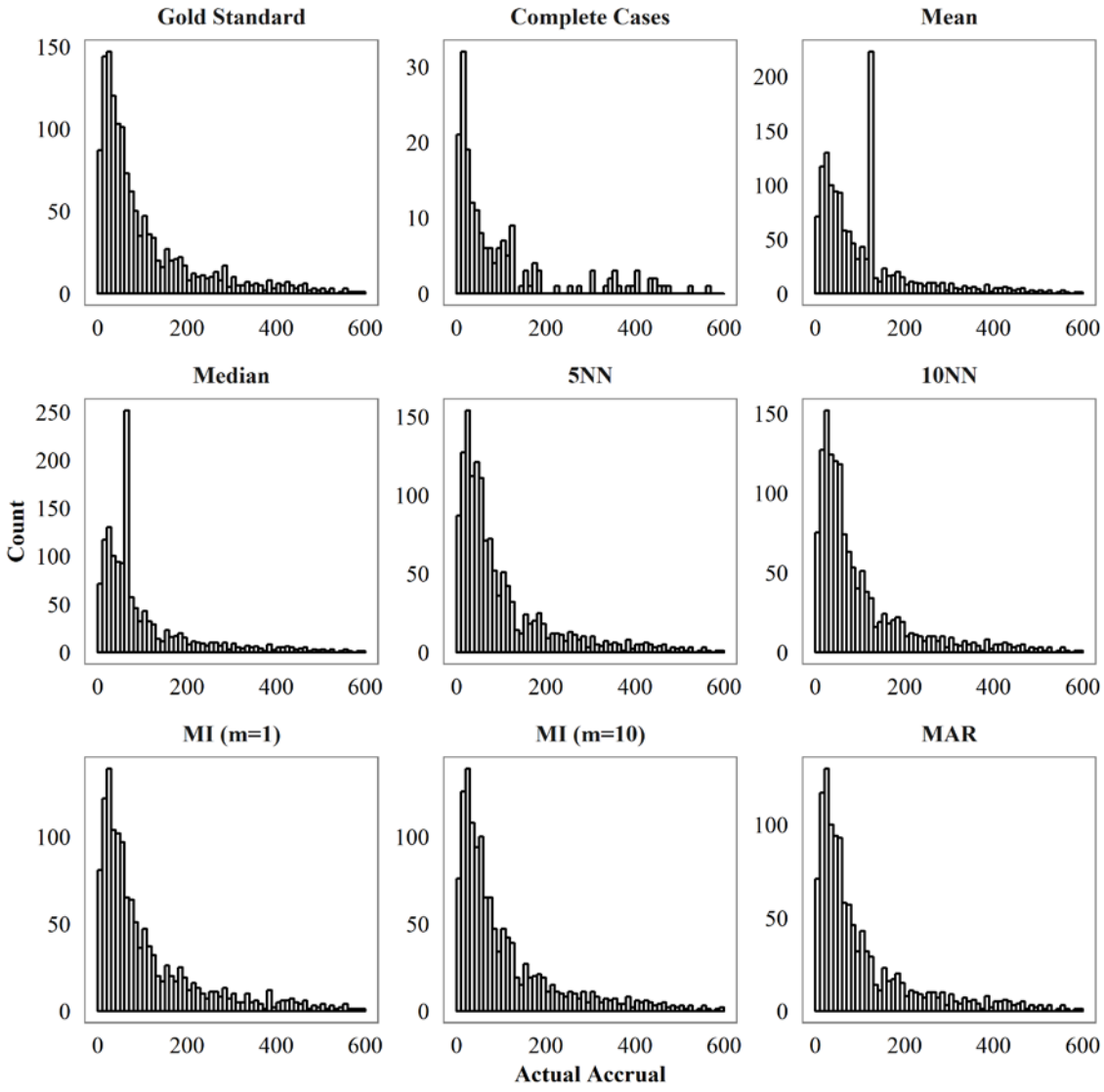

Fig 16. Gold standard, complete cases, MAR and imputed distributions of actual accrual in the training set of one of the iterations. The range of actual accrual goes up to 3,000. However, only a small number of samples go beyond 600. Thus, we truncated the histograms at 600 for better visualization. For MAR distribution, we ignored all missing values.

F Making MAR and MCAR

We simulate missingness in gold standard training and testing sets (see Table 20) based on our assumption of MAR and the missingness patterns observed in the P2APP dataset (see Table 23 and Table 24). For example, 36% of approved drugs in the P2APP dataset have some incomplete drug features. Accordingly, we randomly select 36% of approved drugs in the gold standard training set and introduce missingness in drug features according to the observed proportions to form the MAR training set, e.g. 6% of these drugs will have missing pharmacological target family values, 76% will have missing biological target family values,

and so on. We repeat this process for failed drugs, completed trials, and terminated trials. At the end, we propagate the missing drug and trial features into the training set feature matrix, so that drug-indication pairs for the same drug have the same drug features missing in their feature vectors, and drug-indication pairs with the same trial have the same trial features missing. Conversely, when making the sets MAR, we ensure that the proportions of drugs and trials with fully observed features (i.e. complete cases) are consistent with that observed in the parent dataset, e.g. 64% of approved drugs in the MAR training set have complete drug features. We repeat this procedure for the gold standard testing set to form the MAR testing set.

We simulate MCAR in the gold standard testing set in a similar fashion to form the MCAR testing set. However, here we use unconditional missingness patterns observed in the pipeline dataset (see Table 23 and Table 24), instead of the known outcomes set where backfilling has occurred.

Table 23. Breakdown of missingness in drug features in P2APP with respect to unique drugs (see Fig 2).

	Missingness[a]		
	Known Outcomes		Pipeline
	Success	Failure	Unconditional
COMPLETE CASES	0.64	0.29	0.46
INCOMPLETE CASES	0.36	0.71	0.54
Route	0.00	0.06	0.04
Pharmacological target family	0.06	0.10	0.17
Biological target family	0.76	0.45	0.63
Medium	0.43	0.86	0.69

[a] Feature missingness with respect to incomplete cases, e.g. 36% of success drugs have some incomplete drug features. 43% of these drugs have missing medium values.

Table 24. Breakdown of missingness in trial features in P2APP with respect to unique trials (see Fig 3).

	Missingness[a]		
	Known Outcomes		Pipeline
	Success	Failure	Unconditional
COMPLETE CASES	0.22	0.60	0.44
INCOMPLETE CASES	0.78	0.40	0.56
Number of identified sites	0.13	0.24	0.21
Actual accrual	0.13	0.54	0.18
Duration	0.37	0.13	0.24
Target accrual	0.54	0.21	0.37
Locations	0.02	0.04	0.02
Study design keywords	0.31	0.24	0.13
Trial outcomes	0.93	0.27	0.81

[a] Feature missingness with respect to incomplete cases.

G Comparison of general and indication-group specific classifiers

Table 25. Comparison of the general and indication-specific classifiers for selected indication group in P2P3. We use bootstrapping to determine the 95% CI for AUC. We exclude indication groups with too few samples.

	General Classifier			Specialized Classifiers		
	Train Set	Test Set	AUC (95% CI)	Train Set	Test Set	AUC (95% CI)
All						
2004-2008	1,278	420	0.708 (0.655, 0.761)	-	-	-
2005-2009	1,442	455	0.678 (0.626, 0.730)	-	-	-
2006-2010	1,634	467	0.688 (0.639, 0.737)	-	-	-
2007-2011	1,790	433	0.659 (0.602, 0.716)	-	-	-
2008-2012	1,853	447	0.784 (0.737, 0.832)	-	-	-
2009-2013	1,921	385	0.797 (0.746, 0.847)	-	-	-
2010-2014	1,933	274	0.852 (0.787, 0.917)	-	-	-
Alimentary						
2004-2008	1,278	49	0.650 (0.482, 0.818)	175	49	0.675 (0.520, 0.831)
2005-2009	1,442	63	0.637 (0.486, 0.788)	195	63	0.520 (0.360, 0.681)
2006-2010	1,634	52	0.775 (0.648, 0.902)	225	52	0.561 (0.402, 0.720)
2007-2011	1,790	46	0.824 (0.701, 0.946)	231	46	0.899 (0.812, 0.986)
2008-2012	1,853	51	0.894 (0.803, 0.986)	224	51	0.862 (0.757, 0.966)
2009-2013	1,921	43	0.679 (0.504, 0.855)	219	43	0.731 (0.561, 0.901)
2010-2014	1,933	28	0.842 (0.694, 0.991)	220	28	0.819 (0.638, 0.999)
Immunological						
2004-2008	1,278	34	0.638 (0.445, 0.830)	117	34	0.371 (0.133, 0.608)
2005-2009	1,442	40	0.600 (0.412, 0.788)	124	40	0.528 (0.332, 0.724)
2006-2010	1,634	29	0.510 (0.291, 0.729)	132	29	0.447 (0.224, 0.670)
2007-2011	1,790	31	0.477 (0.241, 0.714)	141	31	0.486 (0.268, 0.705)
2008-2012	1,853	33	0.577 (0.371, 0.782)	144	33	0.650 (0.452, 0.848)
2009-2013	1,921	26	0.762 (0.581, 0.943)	141	26	0.981 (0.936, 1.000)
2010-2014	1,933	21	0.889 (0.739, 1.000)	141	21	0.889 (0.739, 1.000)
Anti-infective						
2004-2008	1,278	29	0.740 (0.495, 0.986)	88	29	0.461 (0.212, 0.710)
2005-2009	1,442	36	0.716 (0.531, 0.900)	102	36	0.572 (0.376, 0.768)
2006-2010	1,634	47	0.631 (0.470, 0.793)	118	47	0.576 (0.412, 0.741)
2007-2011	1,790	40	0.685 (0.518, 0.853)	135	40	0.442 (0.258, 0.627)
2008-2012	1,853	42	0.800 (0.664, 0.936)	142	42	0.634 (0.463, 0.806)
2009-2013	1,921	29	0.717 (0.526, 0.909)	149	29	0.545 (0.317, 0.773)
2010-2014	1,933	17	0.962 (0.874, 1.000)	153	17	0.808 (0.567, 1.000)
Anti-cancer						
2004-2008	1,278	134	0.688 (0.593, 0.783)	461	134	0.719 (0.630, 0.808)
2005-2009	1,442	146	0.639 (0.541, 0.738)	491	146	0.635 (0.527, 0.742)
2006-2010	1,634	151	0.684 (0.589, 0.779)	541	151	0.691 (0.595, 0.786)
2007-2011	1,790	156	0.671 (0.565, 0.777)	589	156	0.767 (0.675, 0.859)
2008-2012	1,853	154	0.756 (0.646, 0.867)	631	154	0.736 (0.629, 0.843)
2009-2013	1,921	136	0.801 (0.685, 0.917)	662	136	0.841 (0.736, 0.945)
2010-2014	1,933	119	0.898 (0.768, 1.000)	686	119	0.863 (0.703, 1.000)
Musculoskeletal						
2004-2008	1,278	29	0.681 (0.475, 0.888)	92	29	0.505 (0.260, 0.750)
2005-2009	1,442	34	0.561 (0.350, 0.772)	100	34	0.451 (0.216, 0.685)
2006-2010	1,634	28	0.646 (0.416, 0.877)	108	28	0.477 (0.246, 0.708)
2007-2011	1,790	27	0.465 (0.218, 0.712)	115	27	0.629 (0.382, 0.877)
2008-2012	1,853	29	0.695 (0.484, 0.906)	120	29	0.610 (0.384, 0.837)
2009-2013	1,921	17	0.635 (0.370, 0.899)	126	17	0.808 (0.594, 1.000)
2010-2014	1,933	12	0.800 (0.539, 1.000)	120	12	0.800 (0.539, 1.000)
Neurological						
2004-2008	1,278	80	0.804 (0.704, 0.903)	190	80	0.772 (0.648, 0.895)

2005-2009	1,442	82	0.773 (0.666, 0.880)	238	82	0.698 (0.577, 0.819)
2006-2010	1,634	79	0.766 (0.657, 0.874)	285	79	0.737 (0.621, 0.853)
2007-2011	1,790	71	0.629 (0.450, 0.807)	321	71	0.610 (0.449, 0.771)
2008-2012	1,853	74	0.806 (0.675, 0.937)	322	74	0.773 (0.626, 0.920)
2009-2013	1,921	69	0.836 (0.727, 0.945)	336	69	0.848 (0.741, 0.955)
2010-2014	1,933	40	0.815 (0.666, 0.964)	325	40	0.743 (0.566, 0.920)
Rare Diseases						
2004-2008	1,278	59	0.676 (0.538, 0.815)	208	59	0.621 (0.477, 0.766)
2005-2009	1,442	72	0.688 (0.563, 0.814)	220	72	0.558 (0.420, 0.697)
2006-2010	1,634	87	0.691 (0.568, 0.813)	250	87	0.689 (0.551, 0.827)
2007-2011	1,790	73	0.689 (0.561, 0.817)	285	73	0.746 (0.619, 0.873)
2008-2012	1,853	83	0.726 (0.596, 0.856)	309	83	0.757 (0.622, 0.892)
2009-2013	1,921	63	0.853 (0.759, 0.947)	327	63	0.789 (0.619, 0.960)
2010-2014	1,933	67	0.870 (0.733, 1.000)	342	67	0.816 (0.620, 1.000)

Table 26. Comparison of the general and indication-group specific classifiers for selected indication group in P2APP. We use bootstrapping to determine the 95% CI for AUC. We exclude indication groups with too few samples.

	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)	-	-	-
Alimentary						
2004-2008	1,361	86	0.494 (0.294, 0.694)	170	86	0.502 (0.310, 0.694)
2005-2009	1,562	93	0.613 (0.440, 0.785)	197	93	0.459 (0.287, 0.630)
2006-2010	1,764	80	0.589 (0.447, 0.731)	237	80	0.491 (0.321, 0.662)
2007-2011	1,969	77	0.707 (0.592, 0.821)	257	77	0.541 (0.396, 0.686)
2008-2012	2,082	67	0.802 (0.694, 0.909)	275	67	0.402 (0.252, 0.553)
2009-2013	2,212	58	0.834 (0.715, 0.954)	279	58	0.610 (0.441, 0.780)
2010-2014	2,289	39	0.670 (0.427, 0.913)	274	39	0.656 (0.414, 0.899)
Cardiovascular						
2004-2008	1,361	39	0.515 (0.313, 0.717)	93	39	0.541 (0.310, 0.771)
2005-2009	1,562	38	0.307 (0.104, 0.509)	105	38	0.452 (0.230, 0.674)
2006-2010	1,764	46	0.613 (0.430, 0.795)	118	46	0.628 (0.449, 0.806)
2007-2011	1,969	37	0.634 (0.396, 0.872)	135	37	0.793 (0.644, 0.942)
2008-2012	2,082	42	0.640 (0.426, 0.853)	137	42	0.621 (0.425, 0.818)
2009-2013	2,212	35	0.360 (0.138, 0.582)	145	35	0.460 (0.272, 0.648)
2010-2014	2,289	19	0.529 (0.000, 1.000)	148	19	0.618 (0.000, 1.000)
Anti-infective						
2004-2008	1,361	46	0.658 (0.502, 0.815)	124	46	0.645 (0.478, 0.812)
2005-2009	1,562	44	0.695 (0.525, 0.866)	146	44	0.707 (0.551, 0.863)
2006-2010	1,764	53	0.733 (0.568, 0.897)	161	53	0.708 (0.552, 0.864)
2007-2011	1,969	44	0.648 (0.479, 0.818)	171	44	0.592 (0.420, 0.763)
2008-2012	2,082	43	0.801 (0.666, 0.936)	165	43	0.815 (0.684, 0.945)
2009-2013	2,212	32	0.658 (0.454, 0.862)	169	32	0.649 (0.435, 0.864)
2010-2014	2,289	18	0.875 (0.708, 1.000)	167	18	0.750 (0.515, 0.985)
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)
Neurological						
2004-2008	1,361	122	0.688 (0.572, 0.803)	211	122	0.768 (0.676, 0.859)
2005-2009	1,562	119	0.612 (0.471, 0.753)	271	119	0.625 (0.501, 0.748)
2006-2010	1,764	125	0.656 (0.532, 0.779)	334	125	0.673 (0.560, 0.787)
2007-2011	1,969	105	0.701 (0.580, 0.822)	375	105	0.649 (0.522, 0.776)
2008-2012	2,082	114	0.806 (0.707, 0.904)	382	114	0.695 (0.586, 0.804)
2009-2013	2,212	87	0.938 (0.857, 1.000)	417	87	0.718 (0.558, 0.879)
2010-2014	2,289	55	0.984 (0.952, 1.000)	408	55	0.860 (0.721, 0.999)
Respiratory						
2004-2008	1,361	34	0.673 (0.418, 0.927)	89	34	0.833 (0.650, 1.000)
2005-2009	1,562	42	0.842 (0.722, 0.962)	104	42	0.825 (0.670, 0.979)
2006-2010	1,764	49	0.797 (0.663, 0.931)	125	49	0.801 (0.644, 0.959)
2007-2011	1,969	36	0.694 (0.513, 0.875)	143	36	0.519 (0.323, 0.715)
2008-2012	2,082	43	0.751 (0.604, 0.899)	149	43	0.692 (0.520, 0.865)
2009-2013	2,212	37	0.827 (0.694, 0.961)	154	37	0.876 (0.764, 0.987)
2010-2014	2,289	23	0.724 (0.365, 1.000)	160	23	0.842 (0.679, 1.000)
Rare Diseases						
2004-2008	1,361	69	0.664 (0.517, 0.811)	212	69	0.521 (0.349, 0.693)
2005-2009	1,562	81	0.627 (0.471, 0.782)	231	81	0.528 (0.368, 0.687)
2006-2010	1,764	108	0.774 (0.666, 0.881)	257	108	0.691 (0.546, 0.836)
2007-2011	1,969	101	0.786 (0.698, 0.874)	303	101	0.680 (0.547, 0.812)
2008-2012	2,082	112	0.787 (0.696, 0.879)	329	112	0.600 (0.469, 0.731)
2009-2013	2,212	90	0.803 (0.702, 0.903)	358	90	0.730 (0.626, 0.834)
2010-2014	2,289	89	0.793 (0.621, 0.965)	391	89	0.779 (0.626, 0.932)

Table 27. Comparison of the general and indication-group specific classifiers for selected indication group in P3APP. We use bootstrapping to determine the 95% CI for AUC. We exclude indication groups with too few samples.

	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)	-	-	-
Alimentary						
2004-2008	472	65	0.826 (0.651, 1.000)	25	65	0.889 (0.756, 1.000)
2005-2009	559	75	0.683 (0.324, 1.000)	17	75	0.650 (0.331, 0.969)
2006-2010	604	80	0.672 (0.428, 0.915)	30	80	0.651 (0.429, 0.872)
2007-2011	664	91	0.911 (0.786, 1.000)	28	91	0.800 (0.630, 0.970)
2008-2012	677	97	0.786 (0.572, 1.000)	24	97	0.700 (0.469, 0.931)
2009-2013	740	107	0.607 (0.149, 1.000)	18	107	0.786 (0.570, 1.000)
2010-2014	734	99	0.944 (0.850, 1.000)	19	99	0.733 (0.492, 0.975)

Anti-cancer						
2004-2008	472	95	0.773 (0.618, 0.928)	34	95	0.684 (0.495, 0.874)
2005-2009	559	107	0.740 (0.543, 0.936)	28	107	0.568 (0.345, 0.791)
2006-2010	604	110	0.754 (0.599, 0.910)	50	110	0.630 (0.452, 0.809)
2007-2011	664	132	0.587 (0.333, 0.842)	24	132	0.392 (0.132, 0.651)
2008-2012	677	134	0.793 (0.549, 1.000)	40	134	0.668 (0.457, 0.879)
2009-2013	740	151	0.800 (0.480, 1.000)	29	151	0.775 (0.528, 1.000)
2010-2014	734	153	0.943 (0.842, 1.000)	26	153	0.852 (0.558, 1.000)
Neurological						
2004-2008	472	118	0.851 (0.753, 0.949)	59	118	0.837 (0.735, 0.939)
2005-2009	559	151	0.782 (0.646, 0.918)	45	151	0.784 (0.649, 0.919)
2006-2010	604	169	0.732 (0.593, 0.871)	52	169	0.759 (0.629, 0.890)
2007-2011	664	180	0.706 (0.532, 0.880)	40	180	0.698 (0.529, 0.867)
2008-2012	677	178	0.765 (0.604, 0.926)	41	178	0.743 (0.586, 0.900)
2009-2013	740	185	0.827 (0.681, 0.973)	31	185	0.805 (0.641, 0.968)
2010-2014	734	166	0.779 (0.567, 0.990)	27	166	0.900 (0.782, 1.000)
Rare Disease						
2004-2008	472	54	0.711 (0.465, 0.957)	22	54	0.620 (0.364, 0.876)
2005-2009	559	60	0.735 (0.517, 0.952)	23	60	0.606 (0.360, 0.852)
2006-2010	604	66	0.888 (0.747, 1.000)	24	66	0.825 (0.645, 1.000)
2007-2011	664	72	0.838 (0.652, 1.000)	22	72	0.735 (0.520, 0.950)
2008-2012	677	76	0.893 (0.780, 1.000)	34	76	0.700 (0.523, 0.877)
2009-2013	740	94	0.962 (0.899, 1.000)	28	94	0.932 (0.840, 1.000)
2010-2014	734	109	0.908 (0.766, 1.000)	18	109	0.985 (0.942, 1.000)

H Comparison with DiMasi et al. (2015)

The Approved New Drug Index (ANDI) algorithm was proposed by DiMasi et al. (2015) to predict regulatory approval for lead indications of cancer drugs after phase 2 testing. It is composed of a rubric of four factors to score anticancer agents (see Appendix I). The factors are based on pivotal trial characteristics and disease prevalence. Higher scores correspond to a higher probability of success. In this analysis, we apply ANDI on the oncology samples in the P2APP dataset, analyze its performance, and compare it with our 5NN-RF classifier in Appendix E.

First, we extract all cancer drugs from P2APP to form an oncology-only dataset. Since ANDI requires complete cases, we drop all examples with missing values in any of the four ANDI factors (see Table 28 for the resulting sample size). From this dataset, we draw a training set of 62 drugs with the same composition as that used by DiMasi et al. (2015): 40 failures and 22 successes. We set aside the remaining 319 drugs as a held-out testing set.

In replicating the ANDI experiment, we endeavored to follow the original proposed rubric as closely as possible. Unfortunately, two factors in the rubric are not in our dataset. We replace them with surrogate variables, and tune their cutoffs using the training set put aside earlier. The modified rubric is given in Table 29. In order to apply ANDI, we have to identify the lead indication of each oncology drug and the pivotal phase 2 trial for that drug-indication pair. However, DiMasi et al. did not provide clear instructions for identifying lead indications or pivotal trials. In this experiment, we apply heuristics which

we felt were most logical. See Appendix I for details on the proxy variables and heuristics used.

DiMasi et al. reported an impressive 0.92 AUC for ANDI on a dataset of 62 drugs. However, this figure is based on in-sample/training-set testing, i.e. the algorithm was tested on the dataset on which the scoring rubric itself was derived. Such testing naturally yields excellent results because the four factors and their cutoffs were optimized for the algorithm to do well on the dataset. However, it is nearly impossible to judge whether an algorithm will generalize well without some form of testing on held-out datasets. Unfortunately, such validation was not performed by DiMasi et al.. Furthermore, ANDI was derived from a small sample, making it even more susceptible to overfitting.

For these reasons, it is very likely that the discriminative power of ANDI is actually much lower than that implied by the reported AUC of 0.92. Knowing these issues, we augment the ANDI experiment by including an out-of-sample model validation step, using the 319 drugs set aside as the testing set. This will allow us to determine ANDI's real performance more accurately.

The receiver operating characteristic curves of the original ANDI algorithm as reported in DiMasi et al. (2015) and the modified ANDI on the oncology-only training and testing sets are shown in Fig 17. Similar to the original ANDI, our modified ANDI rubric demonstrates excellent performance on the training set with 0.94 AUC, 95% CI (0.89, 0.99). Unfortunately, this performance does not hold up on the testing set. The modified ANDI managed only 0.69 AUC on new, unseen samples. The large discrepancy between training and testing AUCs is indicative of overfitting. It is apparent that the patterns learned from the small training sample (n=62) do not generalize well, highlighting the importance of proper model validation. We believe the same holds for the original ANDI.

For a direct comparison with our classifiers, we apply the modified ANDI on oncology drugs in the gold standard testing sets in Appendix E. Fig 18 summarizes the distributions of the results and compares 5NN-RF with the modified ANDI. On this testing set subsample, we find that our classifier achieves significantly higher AUC than the modified ANDI, an average improvement of 0.1 in AUC over 100 simulations. 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 proper model validation methodology.

Lastly, we note that DiMasi et al. applied complete case analysis in their study without any characterization of the missingness in their dataset. This is dangerous because complete cases are appropriate only under strict MCAR conditions. Violation of these conditions will lead to biased estimates. Since data is rarely MCAR in reality, it is unsurprising that the modified ANDI yields an inferior performance. In practice, this limits the applicability of ANDI to only samples with complete information. Given the scattershot nature of reporting in drug development, this makes ANDI less useful.

Table 28. Sample size of the oncology-only dataset (derived from P2APP).

	Counts				
	Drug-indication Pairs	**Phase 2 Trials**	**Unique Drugs**	**Unique Indications**	**Unique Phase 2 Trials**
Success	71	178	61	28	176
Failure	668	1,345	347	40	1,213
Total	**739**	**1,523**	**381**	**40**	**1,368**

Table 29. Modified ANDI rubric in this study.

	Score		
Factor	**0**	**1**	**2**
Trial outcomes[a]	Terminated, lack of efficacy; completed, negative outcomes or primary endpoints not met	Completed, outcome indeterminate	Terminated, early positive outcomes; completed, positive outcomes or primary endpoints met
Number of patients in pivotal phase 2 trial	≤ 37	38-49	≥ 50
U.S. incidence[a]	> 100,000	10,000-100,000	< 10,000
Phase 2 duration (months)	> 44	21-44	< 21

[a] Surrogate variable.

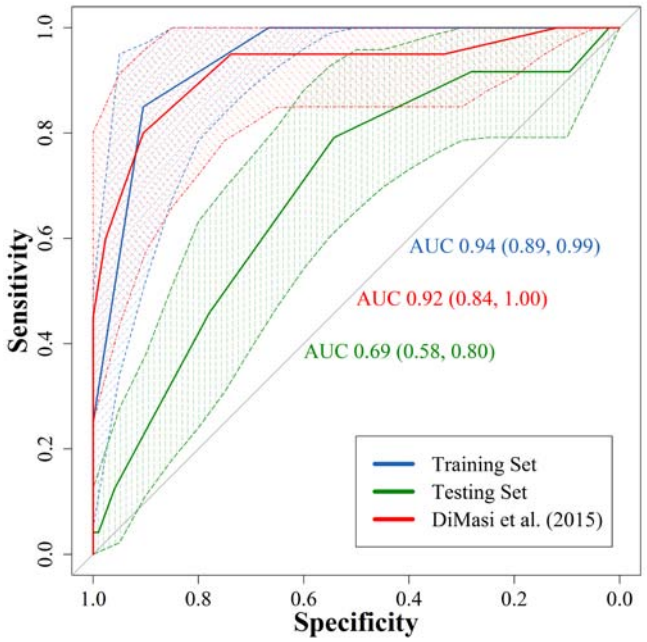

Fig 17. Receiver operating characteristic curves of the original ANDI (as reported in DiMasi et al. (2015)) and the modified ANDI on the oncology-only training and testing sets. We use bootstrapping to determine the 95% CI. We plot the receiver operating characteristic curve of the original ANDI from DiMasi et al. (2015) (red) by using the ANDI scores breakdown provided in the paper. The slight difference in the lower bound of the 95% CI between what we computed (0.84) and what DiMasi et al. reported (0.81) may be accounted by randomness in the bootstraps. Abbreviations. ROC: receiver operating characteristic curve.

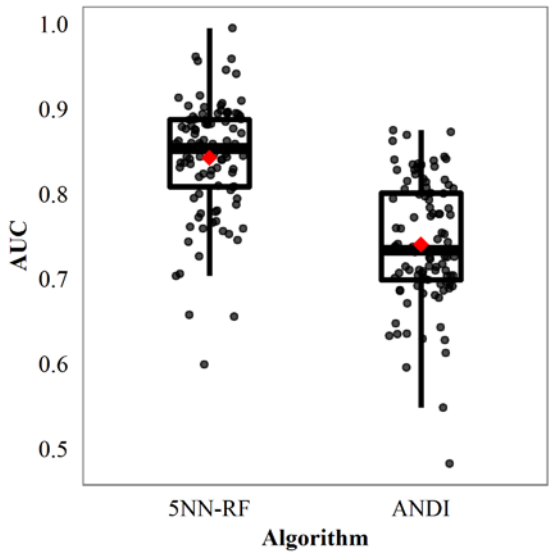

Fig 18. Distributions of AUC of 5NN-RF and the modified ANDI on oncology-only gold standard testing sets from Appendix E.

I Modified ANDI

In replicating the ANDI experiment (DiMasi et al., 2015), we endeavored to follow the original proposed rubric as closely as possible (see Table 30). Unfortunately, two factors in the rubric are not in our dataset: worldwide prevalence and activity. We replace them with surrogate variables, and tune their cutoffs using the training set placed aside earlier. The modified rubric is given in Table 29. First, we use US incidence as a proxy for worldwide prevalence. This is because the latter figure is not known accurately for many of the oncology indications in our dataset, while the US incidence is much better documented and more accessible[10]. We determine the cutoffs in a manner similar to DiMasi et al. (2015): a larger incidence has lower scores while a smaller incidence has higher scores. Second, we use the trial outcome (i.e., the results of the trial) as a proxy for activity. We set the cutoffs similarly as in the original rubric: negative results have lower scores, while positive results have higher scores.

In order to apply ANDI, we have to identify the lead indication of each oncology drug and the pivotal phase 2 trial for that drug-indication pair. Unfortunately, DiMasi et al. did not provide clear instructions for identifying lead indications or pivotal trials in the paper[11]. A fair amount of subjectivity appears to have been involved; there was no mention of any concrete criteria in the paper. This makes it difficult to replicate their study on other

[10] Sources include the American Cancer Society and the National Cancer Institute Surveillance, Epidemiology and End Results Program.

[11] They stated that they focused on what they "determined to be the lead cancer indication pursued", and they "identified what appeared to be the phase II trial that was most pivotal to the decision to proceed to large-scale phase III testing or to abandon the compound after phase II testing" (DiMasi et al., 2015).

datasets. In this experiment, we apply heuristics which we felt were most logical. For drugs with multiple indications, we take the indication with the most phase 2 trials as the lead. We hypothesize that companies will invest in more trials for the designated lead indication. For drug-indication pairs with multiple phase 2 trials, we choose the trial with the largest accrual as the pivotal trial. This is logical, since trials with larger sample size have greater statistical power. They should hold greater weight in the decision of whether to proceed to phase 3 testing. In the event of ties, with the same number of trials or an identical accrual, we randomly select one of the candidates as the lead indication or pivotal trial.

Table 30. Oncology ANDI proposed by DiMasi et al. (2015).

	Score		
Factor	**0**	**1**	**2**
Pivotal phase 2 trial activity	< 3.0% or negative randomized phase 2 trial	3.0-13.8%	< 13.8% or positive randomized phase 2 trial
Number of patients in pivotal phase 2 trial	≤ 37	38-49	≥ 50
No. of patients treated worldwide	> 302,000	50,000-302,000	< 50,000
Phase 2 duration (months)	> 44	21-44	< 21

J Simulation of random splitting versus temporal ordering

We design an experiment to study the effects of any look-ahead bias introduced by splitting drug-indication pairs into training and testing sets randomly without considering the dates of development. First, we sample five-year rolling windows between 2004 and 2014 from the P2P3, P2APP and P3APP datasets. In Section 4–Predictions over time, we note that 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. Here we disregard the temporal ordering—we aggregate the training and testing sets, and re-split them randomly before applying our machine-learning framework. To allow direct comparison with the time-series approach, we keep the new training and testing sample sizes same as those in Table 14, Table 15 and Table 16. Table 31 summarize the results.

Table 31. Comparison of classifiers trained on random splitting and temporal ordering. We use bootstrapping to determine the 95% CI for AUC.

	Sample Size		AUC (95% CI)	
	Train Set	Test Set	Random Splitting	Temporal Ordering
			P2P3	
2004-2008	1,278	420	0.731 (0.679, 0.784)	0.708 (0.655, 0.761)
2005-2009	1,442	455	0.758 (0.711, 0.804)	0.678 (0.626, 0.730)
2006-2010	1,634	467	0.754 (0.708, 0.800)	0.688 (0.639, 0.737)
2007-2011	1,790	433	0.743 (0.695, 0.790)	0.659 (0.602, 0.716)
2008-2012	1,853	447	0.742 (0.690, 0.794)	0.784 (0.737, 0.832)
2009-2013	1,921	385	0.801 (0.751, 0.851)	0.797 (0.746, 0.847)
2010-2014	1,933	274	0.764 (0.696, 0.831)	0.852 (0.787, 0.917)
			P2APP	
2004-2008	1,361	551	0.750 (0.703, 0.797)	0.669 (0.614, 0.725)
2005-2009	1,562	591	0.764 (0.720, 0.808)	0.680 (0.625, 0.735)
2006-2010	1,764	636	0.748 (0.703, 0.794)	0.712 (0.668, 0.755)
2007-2011	1,969	598	0.768 (0.727, 0.809)	0.738 (0.698, 0.777)
2008-2012	2,082	597	0.750 (0.705, 0.795)	0.799 (0.760, 0.837)
2009-2013	2,212	517	0.781 (0.732, 0.829)	0.823 (0.779, 0.867)
2010-2014	2,289	380	0.795 (0.732, 0.858)	0.797 (0.718, 0.876)
			P3APP	
2004-2008	472	196	0.720 (0.650, 0.790)	0.769 (0.704, 0.834)
2005-2009	559	177	0.748 (0.675, 0.821)	0.724 (0.650, 0.798)
2006-2010	604	211	0.771 (0.707, 0.835)	0.738 (0.671, 0.805)
2007-2011	664	174	0.810 (0.743, 0.877)	0.806 (0.740, 0.871)
2008-2012	677	197	0.805 (0.744, 0.866)	0.827 (0.768, 0.886)
2009-2013	740	153	0.820 (0.754, 0.885)	0.868 (0.809, 0.927)
2010-2014	734	110	0.849 (0.772, 0.925)	0.876 (0.811, 0.941)

We find that random splitting is indeed susceptible to overoptimistic performance (e.g. first four windows in P2APP). This may be attributed to the presence of future information in the training set, thus leading to look-ahead bias. However, we also observe overpessimistic results in some cases (e.g. last three windows in P3APP). This may occur when useful past information are set aside in the testing set. We believe that historical successes and failures contain valuable insights on the characteristics of high-potential candidates. Consider prediction for a phase 3 drug today. If we know that a drug with similar mechanism of action has been approved before, we should probably assign a higher chance of success to the pipeline drug under consideration. Conversely, if we see termination of drugs with similar mechanism of action in the past, we should lower our expectations for the pipeline drug as well. Under random allocation, the pipeline drug may be set aside in the testing set together with its historical counterpart. This prevents the model from learning from past experiences, which leads to overpessimistic performance.

The use of random splitting may be less than ideal due to the reasons noted above. It is prudent to adhere to the temporal ordering in the dataset when constructing training and testing sets in order to obtain realistic inferences.

K References

Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3).

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.

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.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Quinlan, R. (1998). C5. 0: An informal tutorial.

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

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

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

Templ, M., Alfons, A., Kowarik, A., & Prantner, B. (2015). VIM: visualization and imputation of missing values. R package version, 4.4.1.

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