15.482 Healthcare Finance
Spring 2017

Andrew W. Lo, MIT

Unit 9, Part 3: Predictive Analytics for Drug Approvals and Clinical-Phase Transitions

Unit Outline

- Risk and Return in the Biopharma Industries, 1930-2015
- Estimating Clinical Success Rates
- Predicting Phase Transitions and Approvals
- Patient-Centered Clinical Trials

Predictive Analytics for Drug Approvals and Clinical-Phase Transitions

Kien Wei Siah, Chi Heem Wong, Andrew W. Lo (2017)

What Is BigData?

- A consequence of cheap computing and storage
- Massively large datasets contain novel insights
- First application: supermarket UPC barcodes
- Other applications include healthcare, marketing, supply chain management, etc.
- Economics and finance have been slow to adopt

The Target Story

The New York Times

February 16, 2012

How Companies Learn Your Secrets

By CHARLES DUHIGG

Andrew Pole had just started working as a statistician for Target in 2002, when two colleagues from the marketing department stopped by his desk to ask an odd question: "If we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that? "

- Consider a 23-year-old women who buys:
  - Cocoa-butter lotion, large purse, zinc and magnesium supplements, bright blue rug
- 87% she's pregnant and due in five months
- Increased probability if this pattern is new
- Send her coupons for diapers, baby clothes, etc.

Big Data for Consumer Credit

Anonymized Data from Large U.S. Commercial Bank

Transaction Data

Transaction Count
Total Inflow
Total Outflow

By Channel:

ACH (Count, Inflow and Outflow)
ATM (Count, Inflow and Outflow)
BPY (Count, Inflow and Outflow)
CC (Count, Inflow and Outflow)
DC (Count, Inflow and Outflow)
INT (Count, Inflow and Outflow)
WIR (Count, Inflow and Outflow)

By Category		
Mortgage payment	Hotel expenses	Bar Expenses
Credit car payment	Travel expenses	Fast Food Expenses
Auto loan payment	Recreation (golf	Total Rest/Bars/Fast-Food
Student loan payment	Department Stores Expenses	Healthcare related expenses
All other types of loan payment	Retail Stores Expenses	Health insurance
Other line of credit payments	Clothing expenses	Gas stations expenses
Brokerage net flow	Discount Store Expenses	Vehicle expenses
Dividends net flow	Big Box Store Expenses	Car and other insurance
Utilities Payments	Education Expenses	Drug stores expenses
TV	Total Food Expenses	Government
Phone	Grocery Expenses	Treasury (eg. tax refunds)
Internet	Restaurant Expenses	Pension Inflow
Collection Agencies	Unemployment Inflow	Collection Agencies

Balance Data

Checking Account Balance
Brokerage Account Balance
Saving Account Balance
CD Account Balance
IRA Account Balance

Credit Bureau Data

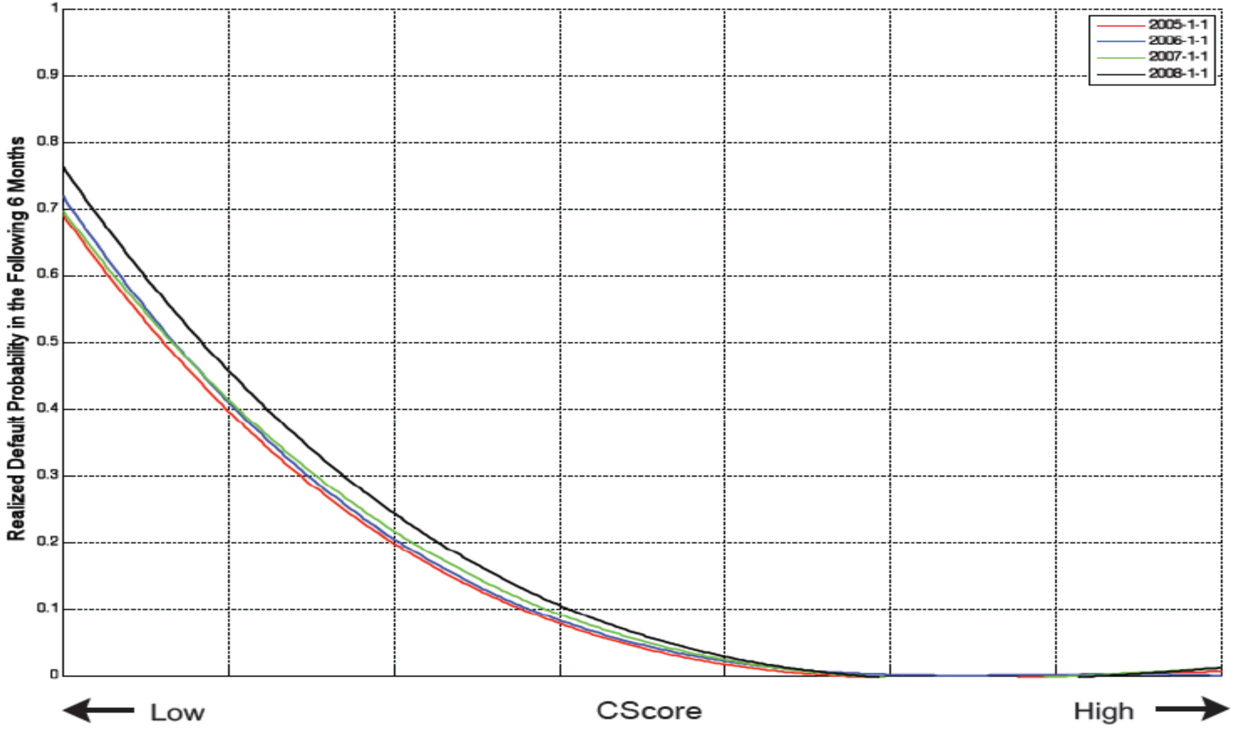
File Age
Credit Score
Open/Closed Flag & Date of Closure
Bankruptcy (Date & Code)
MSA & Zip

Type (CC, MTG, AUT, etc)
Age of Account
Balance
Limit if applicable
Payment Status
48-Month Payment Status History

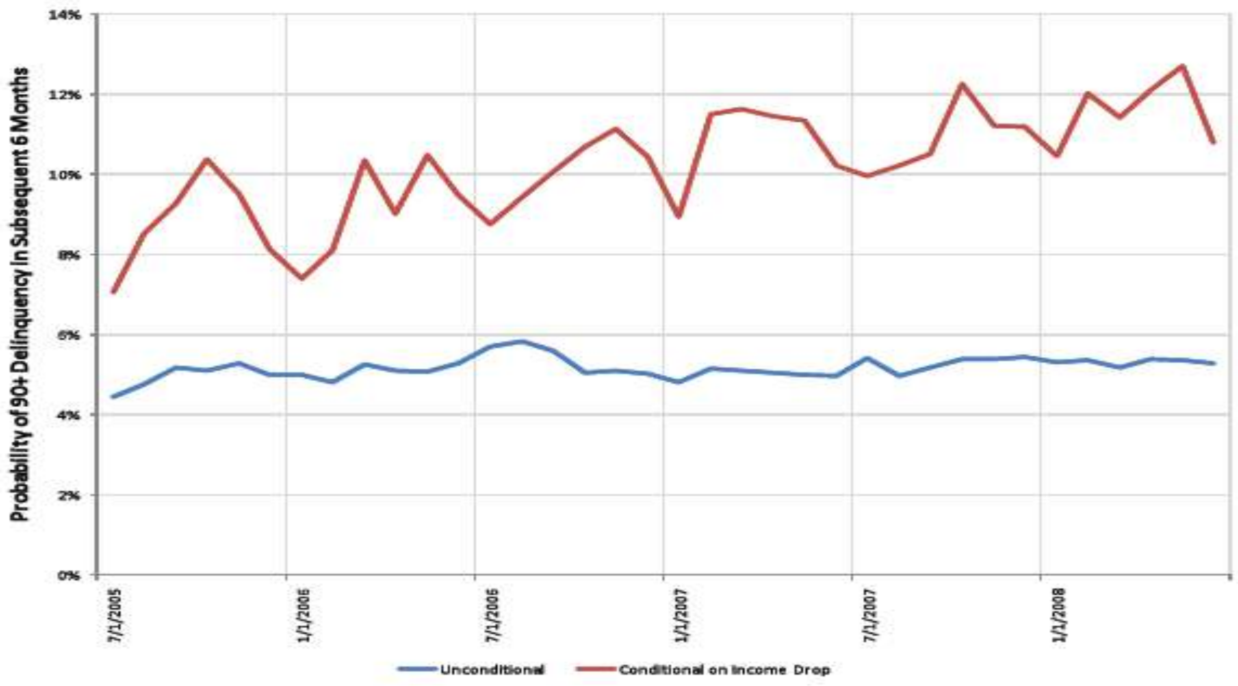
1% Sample = 10 Tb!

Big Data for Consumer Credit

Standard Credit Scores Are Too Insensitive



Big Data for Consumer Credit



Big Data for Consumer Credit

Inputs describing consumer j

consumer level categorical expenditures

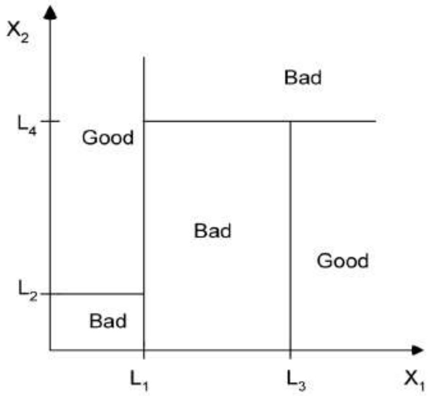
consumer credit history & financial behaviors

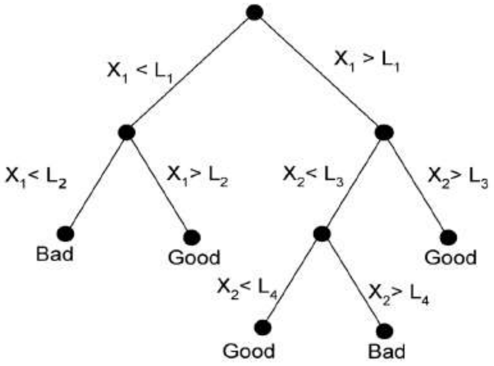
Forecast Model

$F^*(x)$

Credit Risk Forecast of consumer j

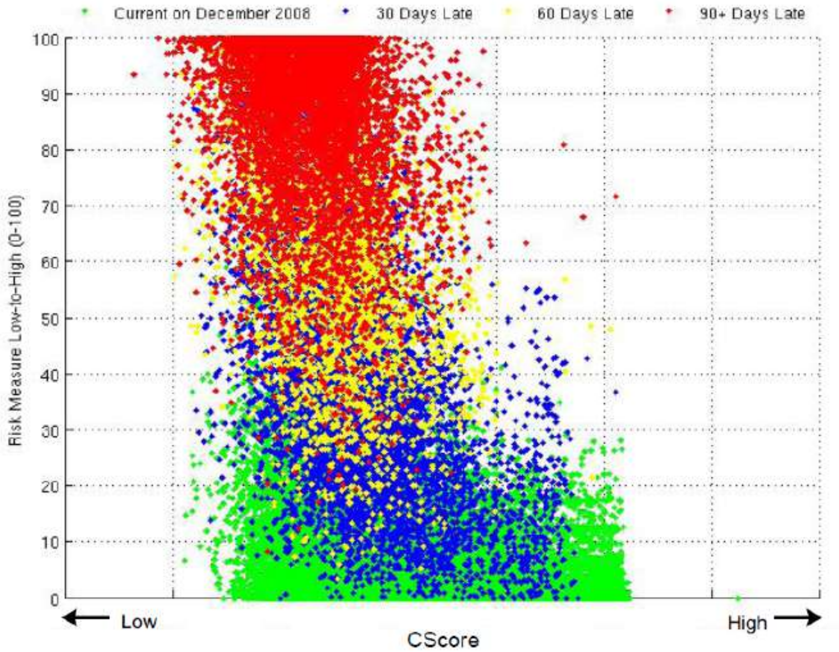
$P(X_j)$: probability of consumer j becoming 90+ days delinquent within next 3 months





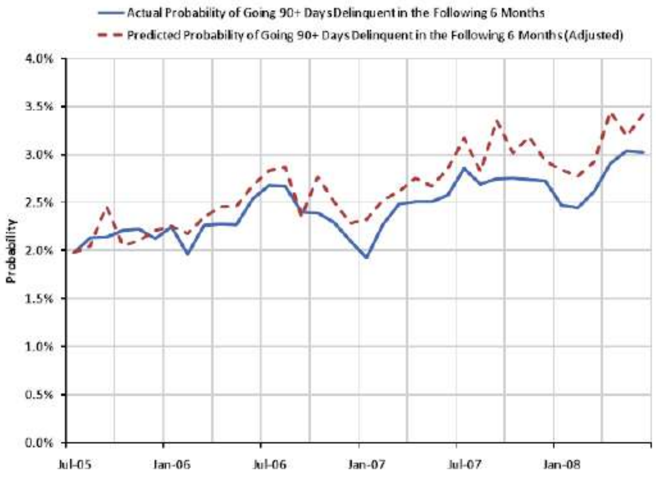
Big Data for Consumer Credit

- Khandani, Kim, and Lo (2010)
- 600,000 credit cards per month; 40-hour runtime

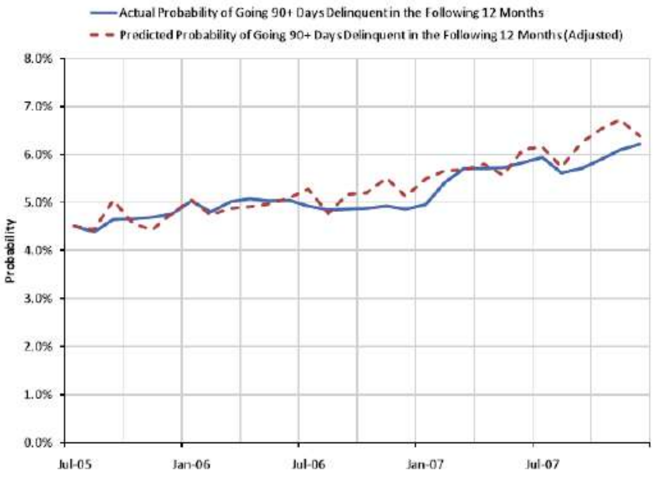


Big Data for Consumer Credit

Credit Forecasts Over Time



(a) Time series of actual and predicted 90-days-or-more delinquency rates (6-month)



(b)Time series of actual and predicted 90-days-or-more delinquency rates (12-month)

Big Data for Consumer Credit

Contents lists available at ScienceDirect

Journal of Banking and Finance 72(2016), 218–239.

journal homepage: www.elsevier.com/locate/jbf

ELSEVIER

Risk and risk management in the credit card industry☆

Florentin Butaru[a], Qingqing Chen[a], Brian Clark[a,e], Sanmay Das[b], Andrew W. Lo[c,d,*], Akhtar Siddique[a]

[a] U.S. Department of the Treasury, Office of the Comptroller of the Currency, Enterprise Risk Analysis Division, United States
[b] Washington University in St. Louis, Department of Computer Science & Engineering, United States
[c] Massachusetts Institute of Technology, Sloan School of Management, Computer Science and Artificial Intelligence Laboratory, Electrical Engineering and Computer Science, United States
[d] AlphaSimplex Group, LLC, United States
[e] Rensselaer Polytechnic Institute (RPI), Lally School of Management, United States

- 6 largest banks from Jan 2008 to Dec 2013
- Macro and institution-specific factors (137), 25 Tb of data
- Models varied greatly across institutions
- Used to gauge quality of risk management across institutions

Big Data for Consumer Credit

Category	Attribute	Bank 1	Bank 2	Bank 3	Bank 4	Bank 5	Bank 6
Utilization	MonthUtilization1MoChange	6.2	6.5	1.5	4.5	4.9	5.3
Utilization	CycleUtilization1MoChange	1.8		5.4		1.4	1.6
Utilization	MonthUtilization					1.0	
Utilization	MonthUtilization3MoChange		2.5				0.0
Utilization	CycleUtilization						1.3
Utilization	Dum1IfTotBal_TotLmtAllOpenBankCardAcctsEQ0			0.1			
delinquency Status	Dum1IfGT0Acct60DPD	5.6	2.3	4.8	5.2	0.8	3.4
Delinquency Status	DaysPastDue	5.4	5.7		5.5	3.8	2.2
Delinquency Status	Dum1IfGT0Acct90DPD	4.2	4.5	4.4	3.2		1.4
Delinquency Status	NumOfAcc60DPD	4.1	2.8	2.5	4.3	0.2	-0.6
Delinquency Status	Dum1IfGT0Acct30DPD	3.8	4.3		2.4		1.5
Delinquency Status	NumAccts30DPD	3.0	2.6		2.4		
Delinquency Status	NumOfAcc90DPD	2.4		2.9	1.8		
Delinquency Status	TotNumAcc60DaysPastDue12MoVerif			-0.1		-0.7	
Delinquency Status	TotNumOpenBankCard60DPD12MoVer					-0.2	
Delinquency Status	Dum1IfGT0BankCardAcct60DPD12MoVer		2.9	-0.5		0.2	
Borrower Payment behavior	ActualPmtAmt_TotPmtDue	5.0	4.0	3.8	4.9	2.0	0.6
Borrower Payment behavior	PaymentEqDueLast3MoFlag	3.9	1.7	3.3	2.3	0.7	-0.8
CardCharacteristics	CurrentCreditLimit	2.4			3.9	0.1	0.8
CardCharacteristics	MonthEndBalance	2.2	2.6	0.1		-0.6	1.7
CardCharacteristics	ProductType	1.8					
CardCharacteristics	CycleEndBalance			0.3	6.5	0.9	2.2
CardCharacteristics	TotNumberOfAccounts			-0.5			
CardCharacteristics	TotNumberGoodAccounts		3.1		2.9		
CardCharacteristics	TotNonMortgBalAllAcc12MoVerif						-0.6
CardCharacteristics	MaxTotAmt60DPDAllAcctsOrTotBalOpenBankCards60DPD		5.7			1.5	
CardCharacteristics	TotCredLmtBankCardAccts			-0.2			
CardCharacteristics	Dum1IfTotCredlmtAllRvlvgAcctsGT012MoVer		2.3	-0.2			
CardCharacteristics	CreditCardType				3.8		
BorrowerCharacteristics	3MoChangeRefreshedFICO	3.5		-0.4			
BorrowerCharacteristics	BehevScore	2.3	3.1	0.7	4.6	1.9	2.9
BorrowerCharacteristics	RefreshedFICO	1.9	1.6		1.7	0.4	1.9
BorrowerCharacteristics	6MoChangeBehevScore						-0.9
AccountStatus	chg1Mo_LineFrozenFlag_0	2.4			1.5	1.8	
AccountStatus	LineFrozenFlag	2.4	1.5				
AccountStatus	LineDecreaseFlag				3.5		
AccountStatus	TotalPaymentDue				2.1	-0.4	2.0
AccountStatus	OverLimitLast3MoFlag					0.4	
Macro	MACROd3hrs_wkly_private	1.5	2.9	0.6	2.7		
Macro	MACROd3num_total_private_nsa		2.5				
Macro	MACROl12hrs_wkly_leisure						0.0
Macro	MACROd12index_sa			-0.3			

Informa® Databases

Pharmaprojects Database

- Tracks drug development pipelines (e.g. development status, route of administration, drug medium, pharmacological target family, …)
- Detailed profiles on 101,507 drug-indication pairs

Drug	Indication	Status	Route of Administration	Medium	Pharmacological Target
Carperitide	Myocardial infarction	Launched	Injectable	???	Natriuretic peptide agonist
Levocetirizine	Perennial allergic rhinitis	Launched	Oral	Tablet, solution, hard capsule	Histamine receptor antagonist

Informa® Databases

Trialtrove Database

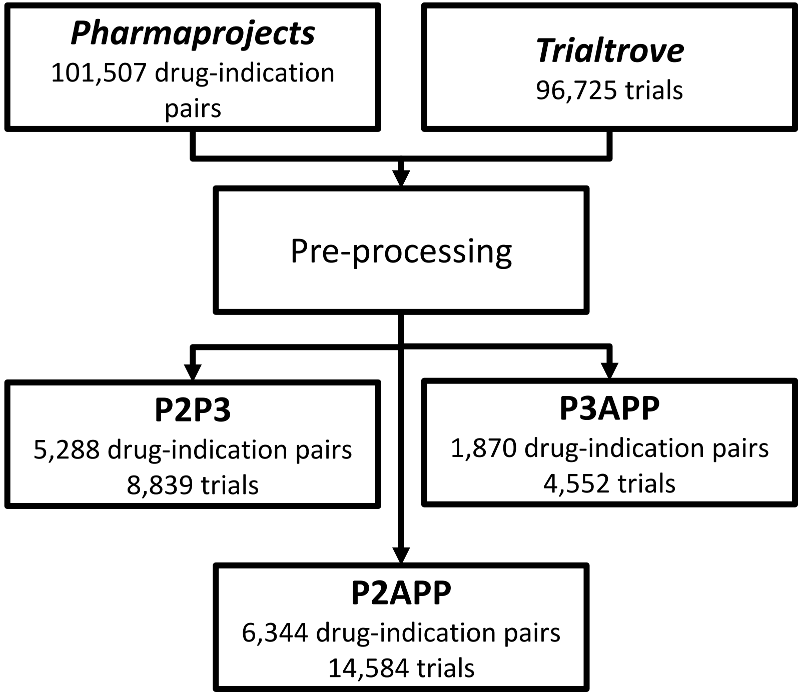
- Tracks clinical trial information (e.g. status, sponsor, accrual, locations, start and end dates, outcomes)
- Aggregates information from nearly 40,000 sources (e.g. company press releases, government drug and trial databases, scientific conferences, …)

Trial ID	Drug	Indication	Phase	Status	Actual Accrual	Locations	End Date	Sponsor Type	Outcome
75404	Carperitide	Myocardial infarction	2	Completed	124	Japan	9/1/2007	Academic	???
65284	Dirucotide	Multiple sclerosis	3	Terminated	580	Canada, Spain, Germany, …	7/27/2009	Industry, top 20 pharma	Terminated, business decision – pipeline reprioritization

Goal

Develop predictive algorithms for assessing probability of success of drug candidates

- Phase 2 to phase 3 (P2P3)
- Phase 2 to approval (P2APP)
- Phase 3 to approval (P3APP)

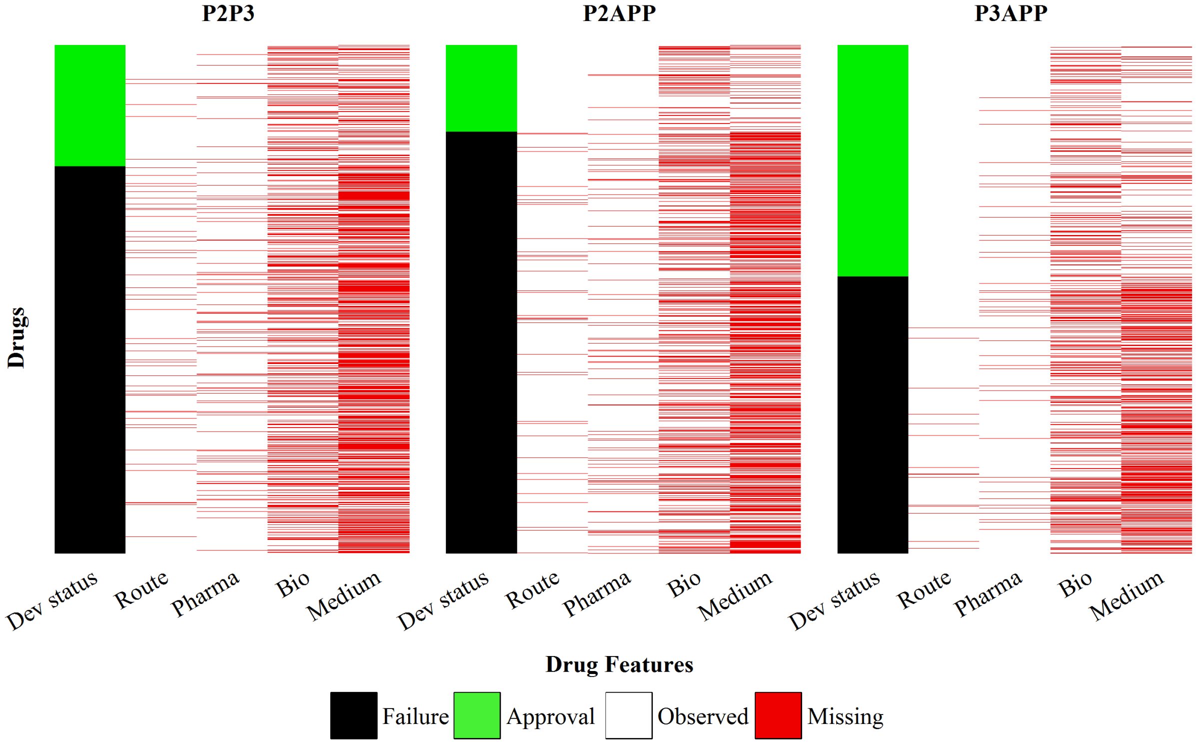


Features

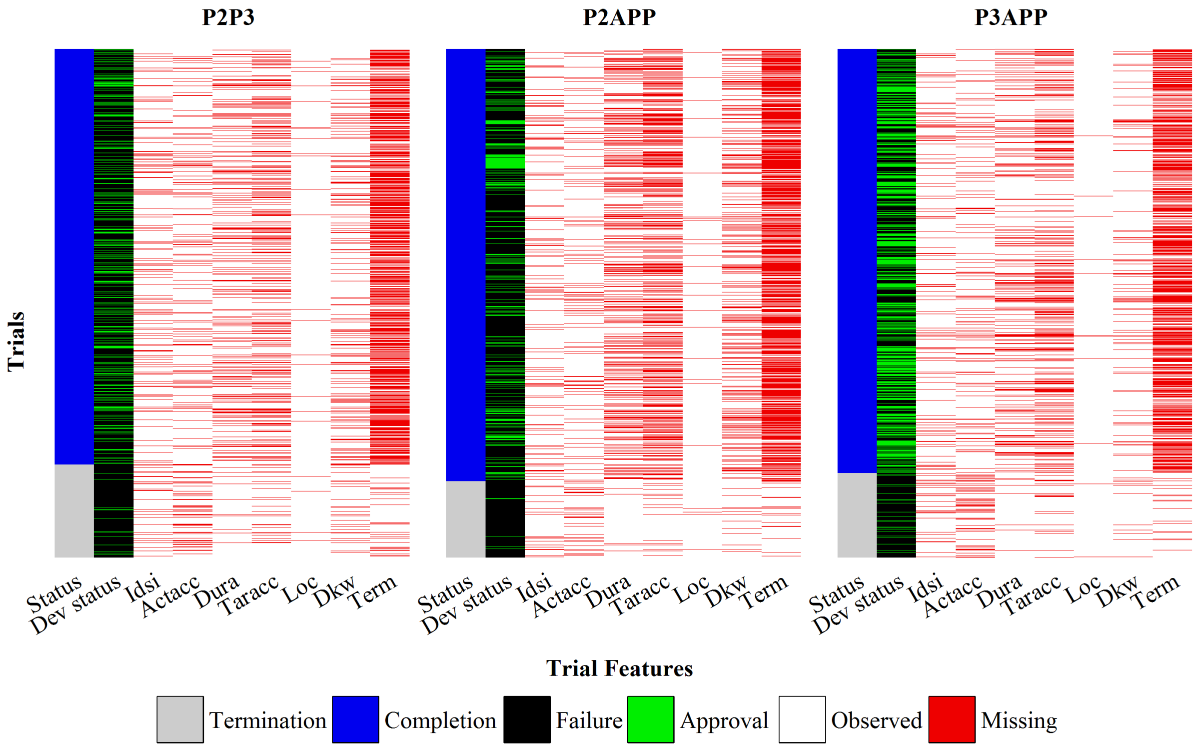
Drug Features
Route
Origin
Medium
Biological target family
Pharmacological target family
Prior approval of drug for another indication
Drug-indication pair development status

Trial Features
Duration
Study design
Sponsor type
Therapeutic area
Trial status
Trial outcome
Target accrual
Actual accrual
Locations
Number of identified sites
Sponsor track record
Investigator experience

Missing Data



Missing Data



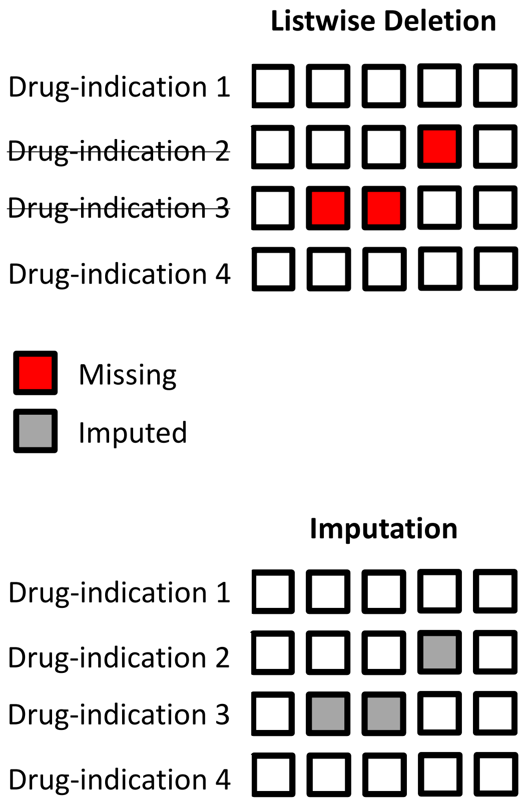
Imputation

Most related studies do not report extent of missingness or use listwise deletion

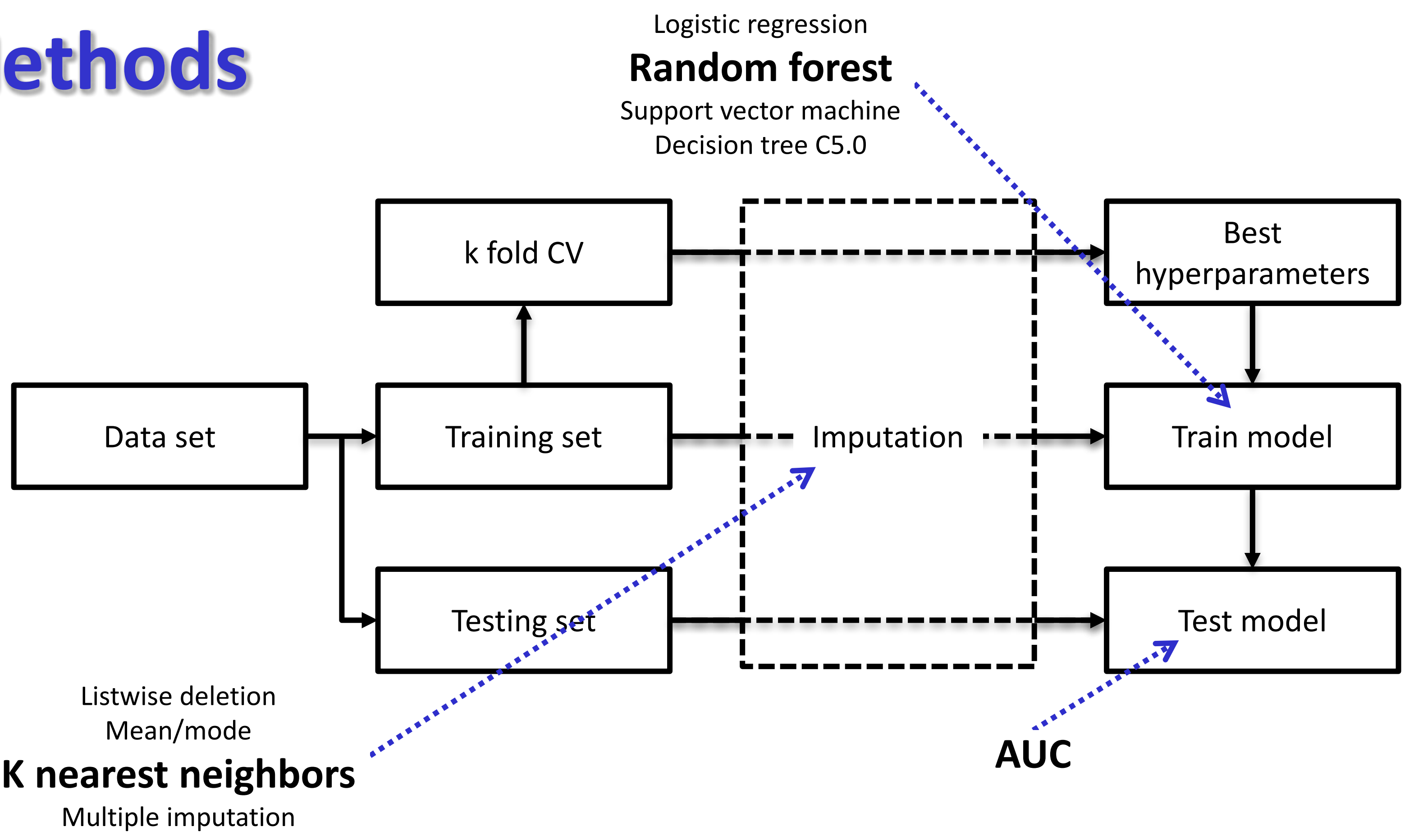
- Simplest remedy
- Biased inferences

We impute missing values using observed data

- Improvement over complete cases



Methods


Logistic regression
Random forest
Support vector machine
Decision tree C5.0
k fold CV
Best hyperparameters
Data set
Training set
Imputation
Train model
Testing set
Test model
Listwise deletion
Mean/mode
K nearest neighbors
Multiple imputation
AUC

Predicting Approvals

- Five nearest neighbors imputation
- Random forest classifier
- AUC as metric

	Avg AUC	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
Rare Diseases	0.752	0.041	0.685	0.755	0.818
Neurological	0.776	0.034	0.716	0.778	0.835
Alimentary	0.733	0.034	0.679	0.736	0.789
Immunological	0.715	0.067	0.604	0.723	0.826
Anti-infective	0.693	0.066	0.594	0.695	0.797
Respiratory	0.693	0.059	0.592	0.699	0.774
Musculoskeletal	0.766	0.055	0.668	0.768	0.853
Cardiovascular	0.677	0.066	0.565	0.677	0.780
Genitourinary	0.719	0.082	0.579	0.729	0.836

Predicting Approvals

- Five nearest neighbors imputation
- Random forest classifier
- AUC as metric

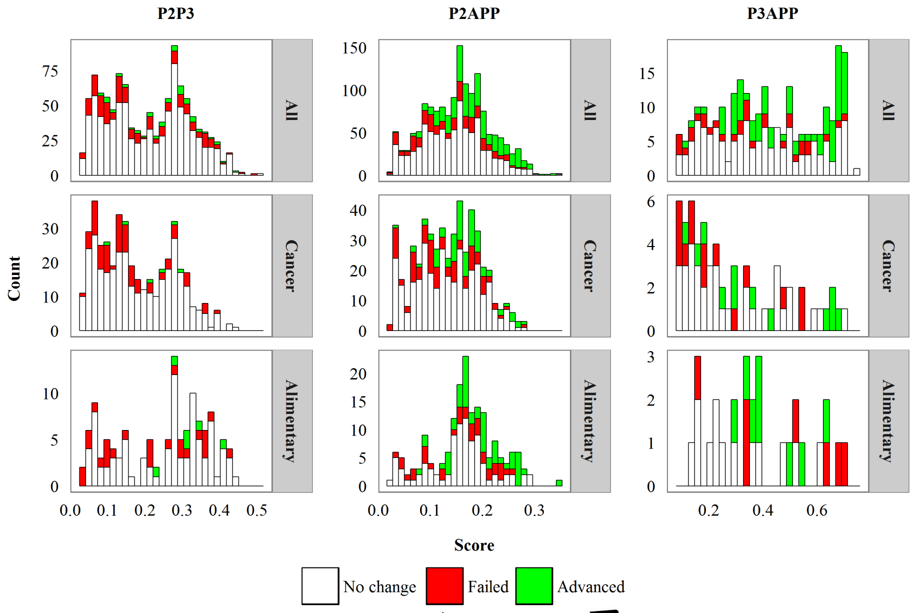
	Avg AUC	Sd	5%	50%	95%
			P2APP		
All	0.777	0.017	0.749	0.775	0.806
Anti-cancer	0.805	0.025	0.764	0.805	0.847
Rare Diseases	0.800	0.028	0.756	0.800	0.848
Neurological	0.767	0.036	0.710	0.769	0.819
Alimentary	0.749	0.045	0.672	0.751	0.817
Immunological	0.783	0.065	0.665	0.786	0.889
Anti-infective	0.735	0.043	0.673	0.736	0.800
Respiratory	0.756	0.055	0.648	0.764	0.835
Musculoskeletal	0.822	0.049	0.736	0.821	0.899
Cardiovascular	0.709	0.072	0.580	0.711	0.812
Genitourinary	0.633	0.086	0.503	0.634	0.790

Predicting Approvals

- Five nearest neighbors imputation
- Random forest classifier
- AUC as metric

	Avg AUC	Sd	5%	50%	95%
			P3APP		
All	0.810	0.018	0.781	0.810	0.834
Anti-cancer	0.783	0.047	0.699	0.779	0.853
Rare Diseases	0.819	0.054	0.727	0.822	0.896
Neurological	0.796	0.037	0.734	0.794	0.857
Alimentary	0.817	0.047	0.744	0.820	0.891
Immunological	0.811	0.074	0.680	0.815	0.910
Anti-infective	0.757	0.065	0.644	0.752	0.854
Respiratory	0.823	0.065	0.712	0.831	0.920
Musculoskeletal	0.741	0.095	0.576	0.747	0.866
Cardiovascular	0.794	0.058	0.702	0.788	0.887
Genitourinary	0.814	0.083	0.670	0.821	0.937

Distribution of Predictions for Pipeline Drug-Indication Pairs



No progress

Terminated

Progressed to higher phases

Distribution of Predictions for Pipeline Drug-Indication Pairs

- Pairs that fail generally have lower scores than those that advance

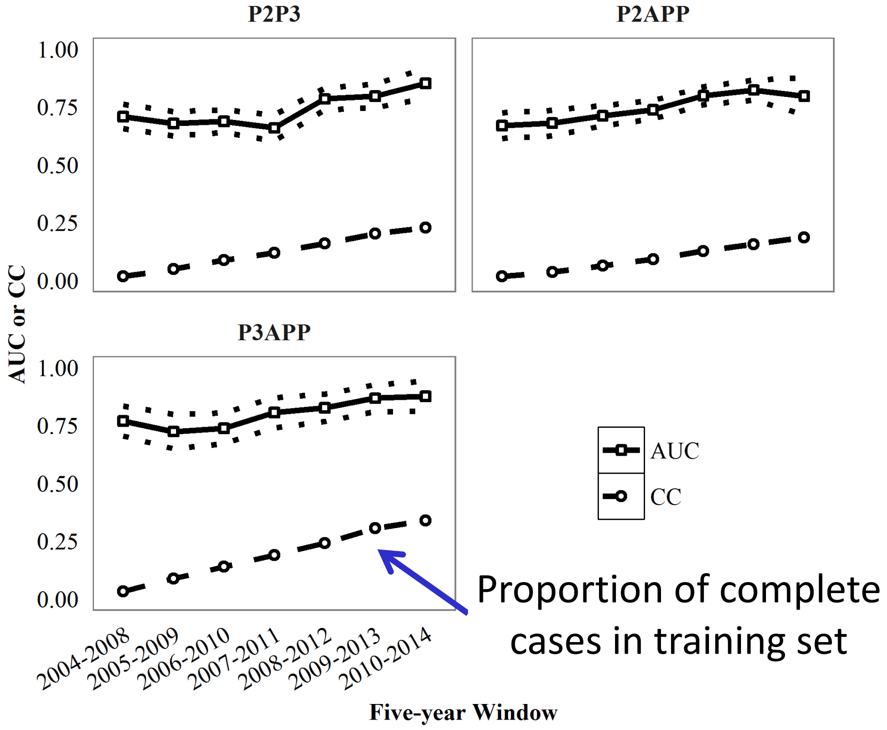
	n	Avg	Sd	5%	50%	95%
			P2P3			
Aggregate	1,105	0.209	0.109	0.054	0.211	0.387
No change	858	0.211	0.108	0.054	0.216	0.388
Failed	194	0.191	0.112	0.052	0.157	0.375
Advanced	53	0.249	0.095	0.098	0.262	0.390
			P2APP			
Aggregate	1,511	0.153	0.061	0.044	0.155	0.258
No change	859	0.143	0.060	0.041	0.147	0.246
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

Important Variables

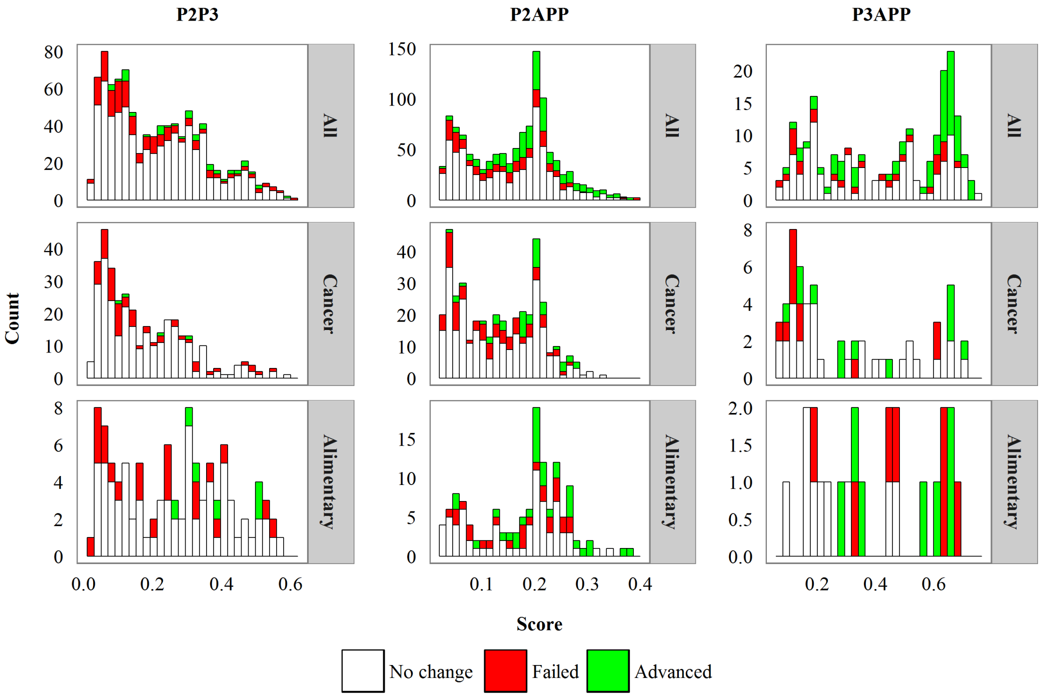
- Trial outcome (primary endpoints met)*
- Trial status (completed or terminated)*
- Prior approval of drug for another indication
- Trial characteristics (accrual, duration, …)
- Sponsor track records (quantified by number of past successful trials)

Predictions Over Time

- Increasing trend over five-year rolling windows
- Proportion of complete cases in training sets correlate well with time series AUC (0.82-0.95)



Distribution of Predictions for Pipeline Drug-Indication Pairs (2011-2015)



Top Five P2APP Pipeline Drug Candidates with Highest Scores

- Candidates still outstanding at the time of writing (neither discontinued nor approved)

Group	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

Discussion

- Large datasets from *Pharmaprojects* and *Trialtrove*
- Imputation and machine-learning approach for analysis
- Classifiers with promising levels of predictive power, able to discriminate between high- and low-potential candidates
- Insights into important variables not considered in prior studies
- Possibility of more powerful prediction models with better quality data