



DATA MINING IN DRUG DISCOVERY AND DEVELOPMENT

Ping Zhang

pzhang@us.ibm.com

IBM T.J. Watson Research Center

USA

Lun Yang

Lun.Yang@gmail.com

GlaxoSmithKline

USA

Outline

- Introduction of Drug Discovery and Development
- Motivation of Data Mining
- Case Study: Drug Repositioning
- Case Study: Real-World Evidence
- Data Sources for Data Mining Applications
- Challenges and Summary

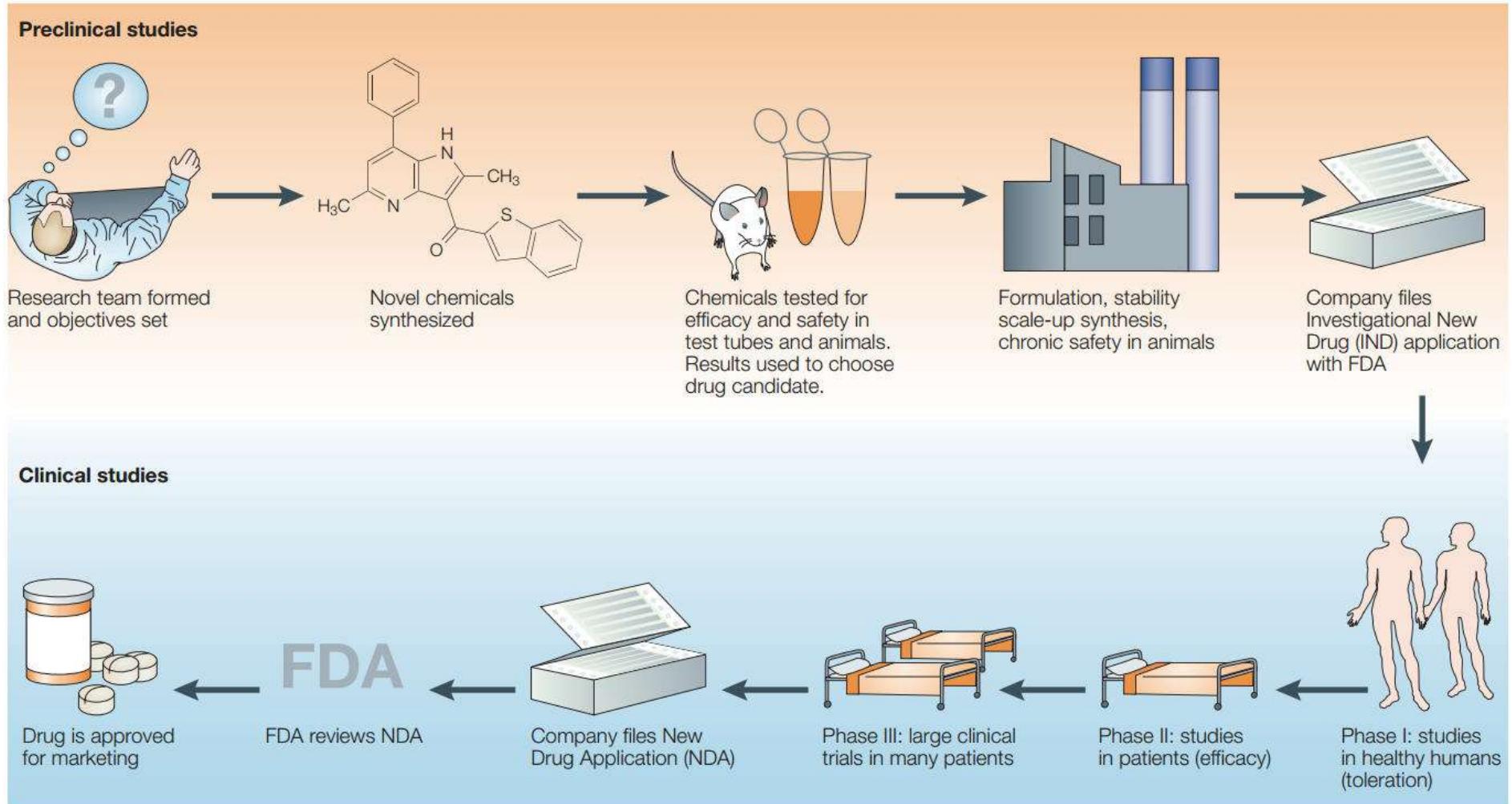
Outline

- Introduction of Drug Discovery and Development
- Motivation of Data Mining
- Case Study: Drug Repositioning
- Case Study: Real-World Evidence
- Data Sources for Data Mining Applications
- Challenges and Summary

Brief history of drug discovery and development

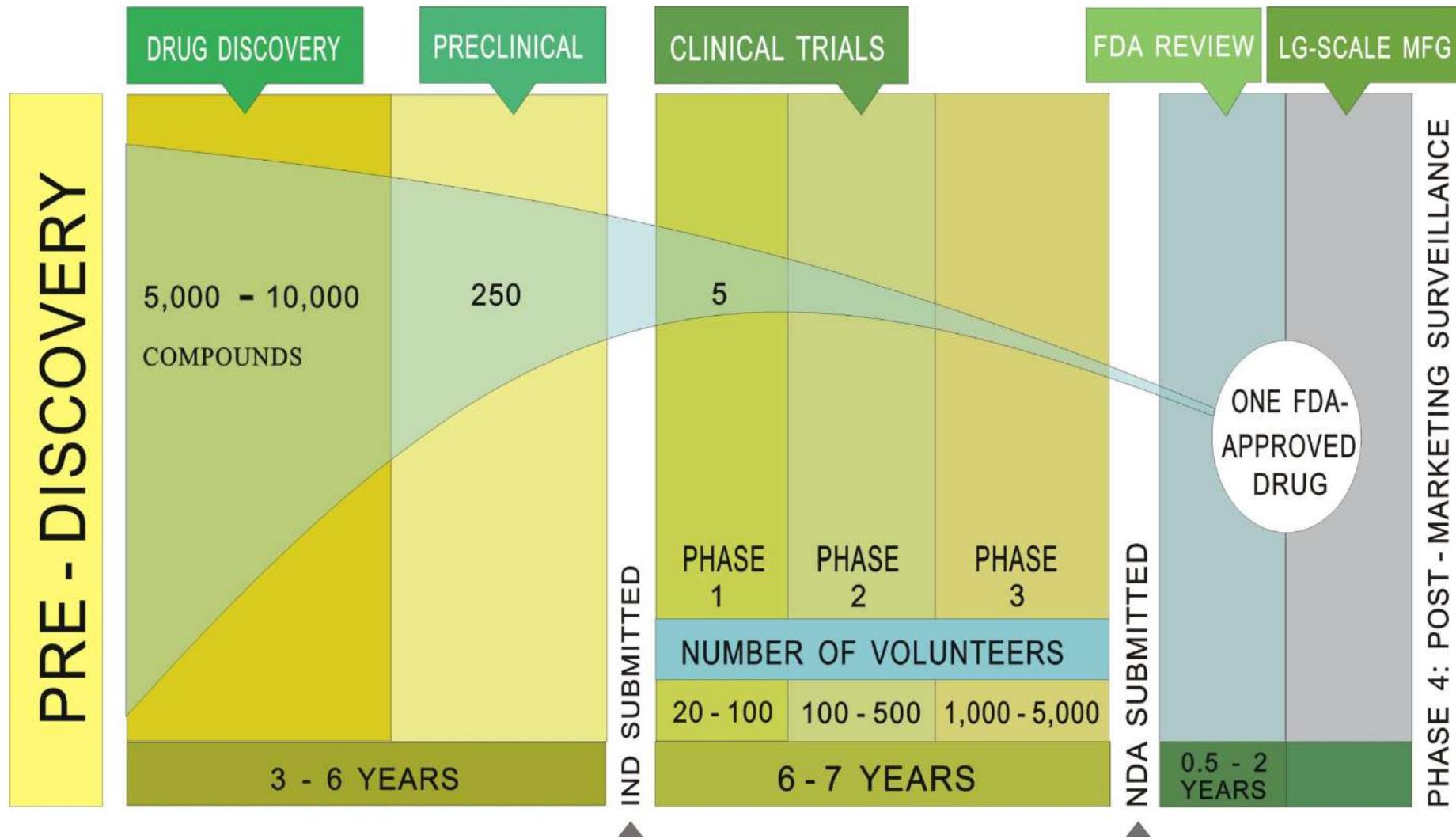
- Empirical – up until 1960's
 - 14th–11th centuries BCE: herbal drugs, serendipitous discoveries
 - Late 1800's: major pharmaceutical companies, mass production
 - 1920's, 30's: vitamins, vaccines
 - 1930-1960: major discoveries (insulin, penicillin, ...)
- Rational – 1960's to 1990's
 - Designing molecules to target protein active sites – "lock and key"
 - Computational drug discovery
 - Biggest success HIV (Reverse transcriptase, protease inhibitors)
- Big Experiment – 1990's to 2000's
 - High throughput screening
 - Microarray assays
 - Gene sequencing and human genome project
- Big Data – 2010's onwards
 - Informatics-driven drug discovery
 - Everything is connected

Stages in the drug discovery and development process



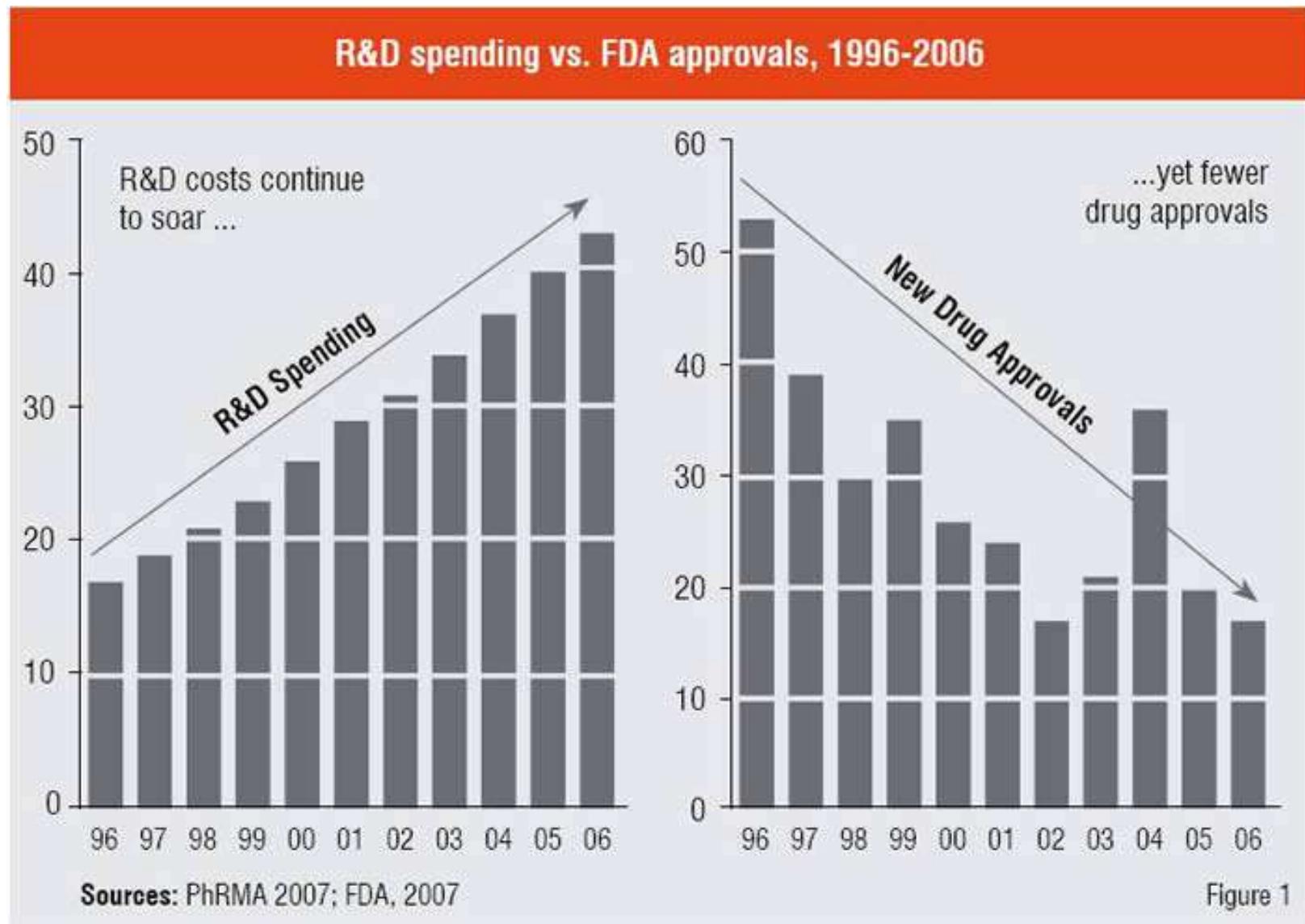
Lombardino JG, Lowe JA 3rd. Nat Rev Drug Discov. 2004 Oct;3(10):853-62.

Timescale in the drug discovery process

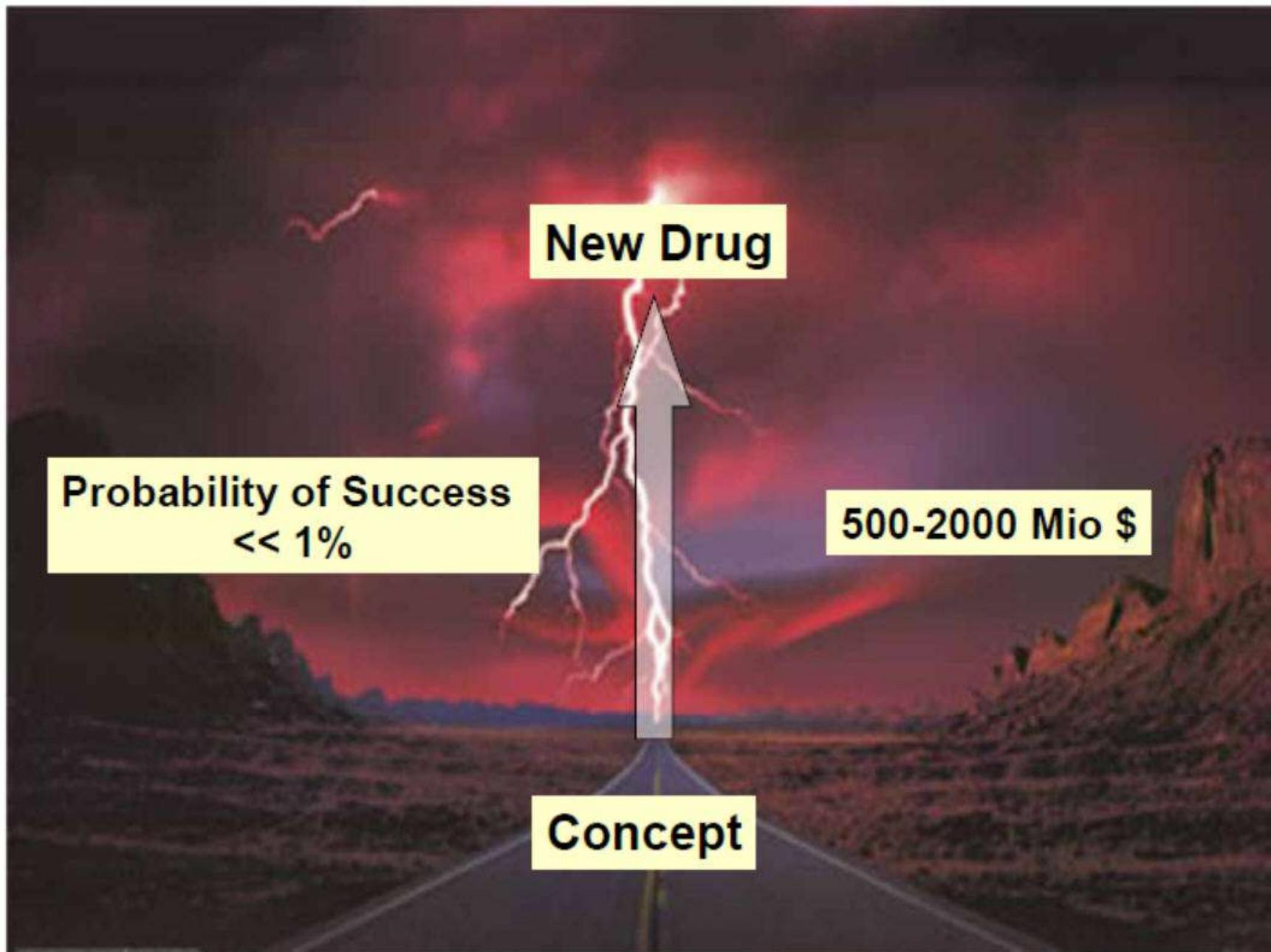


Available at http://www.innovation.org/drug_discovery/objects/pdf/RD_Brochure.pdf

Bottleneck in drug discovery



Traditional Drug Discovery Process



Outline

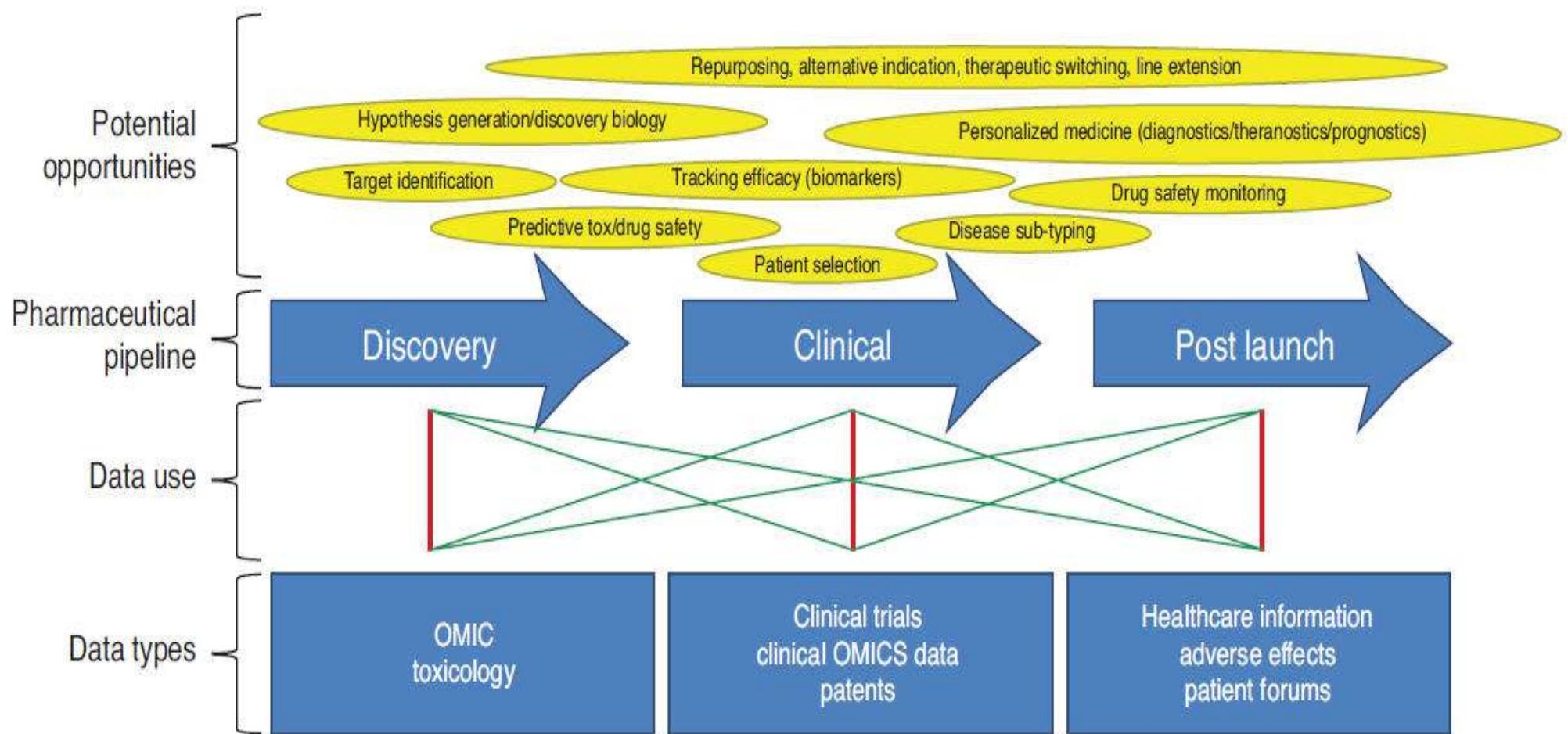
- Introduction of Drug Discovery and Development
- Motivation of Data Mining
- Case Study: Drug Repositioning
- Case Study: Real-World Evidence
- Data Sources for Data Mining Applications
- Challenges and Summary

Big Data in the public domain

- There is now an incredibly rich resource of public information relating compounds, targets, genes, pathways, and diseases. Just for starters there is in the public domain information on*:
 - 48,777,362 compounds, 127,906,628 substances, 739,657 bioassays (PubChem)
 - 1552 FDA-approved small molecule drugs, 284 biotech drugs, 6009 experimental drugs (DrugBank)
 - 542,258 manually reviewed protein sequences, 51,616,950 un-reviewed protein sequences (Swiss-Prot/UniProtKB), 95,968 3D structures (PDB)
 - 22 million life science publications – 1 million new each year (PubMed)
 - 160,781 clinical studies with locations in all 50 states and in 185 countries (ClinicalTrials.gov)
- Even more important are the relationships between these entities. For example a chemical compound can be linked to a gene or a protein target in a multitude of ways:
 - Co-occurrence in a paper abstract
 - Computational experiment (docking, predictive model)
 - System association (e.g. involved in same pathways cellular processes)
 - Statistical relationship

* All databases were accessed on 02/08/2014

Why Data Mining is appealing



Why Drug Discovery and Development is appealing

- Drug discovery is highly data driven and data are increasingly becoming public available
 - NIH has started ambitious extramural funding programs to support academic-based drug discovery programs recently
 - Pharms begin to make the trove of detailed raw data underlying its clinical trials systematically available to researchers
- Having ample data, bring challenging problems, demanding more knowledge
- Spans full data analytics cycles
 - Data collection, data cleansing, data semantics, data integration, data representation
 - Model inference, model selection, modal average, model interpretation
- We see many different data types
 - Vector, semi-structured, time-series, spatial-temporal, images, video, hypertext, literature
- Data analytics and data management challenges are from all aspects
 - Large volume, high dimensional, high noise, large amount of missing values, non iid data, structured input and output, unlabeled data
 - Multi-instance (label, class, task)

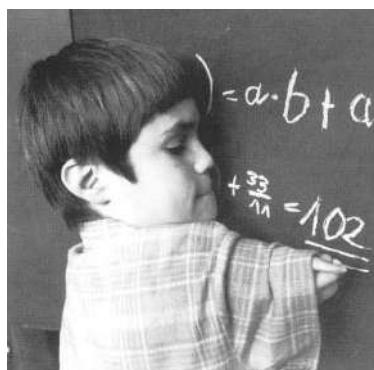
Outline

- Introduction of Drug Discovery and Development
- Motivation of Data Mining
- Case Study: Drug Repositioning
- Case Study: Real-World Evidence
- Data Sources for Data Mining Applications
- Challenges and Summary

Examples of drug repositioning

New uses for old drugs

Drug	Original indication	New indication
Viagra	Hypertension	Erectile dysfunction
Wellbutrin	Depression	Smoking cessation
Thalidomide	Antiemetic	Multiple Myeloma



The NEW ENGLAND
JOURNAL of MEDICINE

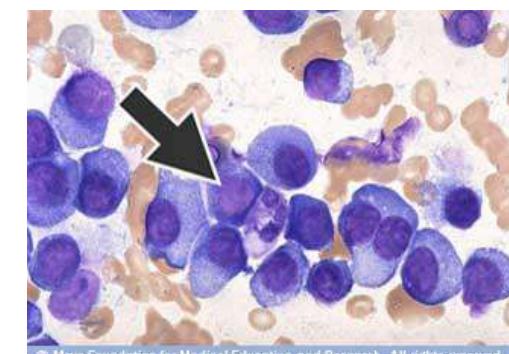
HOME

ARTICLES ▾

ISSUES ▾

SPECIALTIES & TOPICS ▾

FOR AUTHORS ▾



EDITORIAL

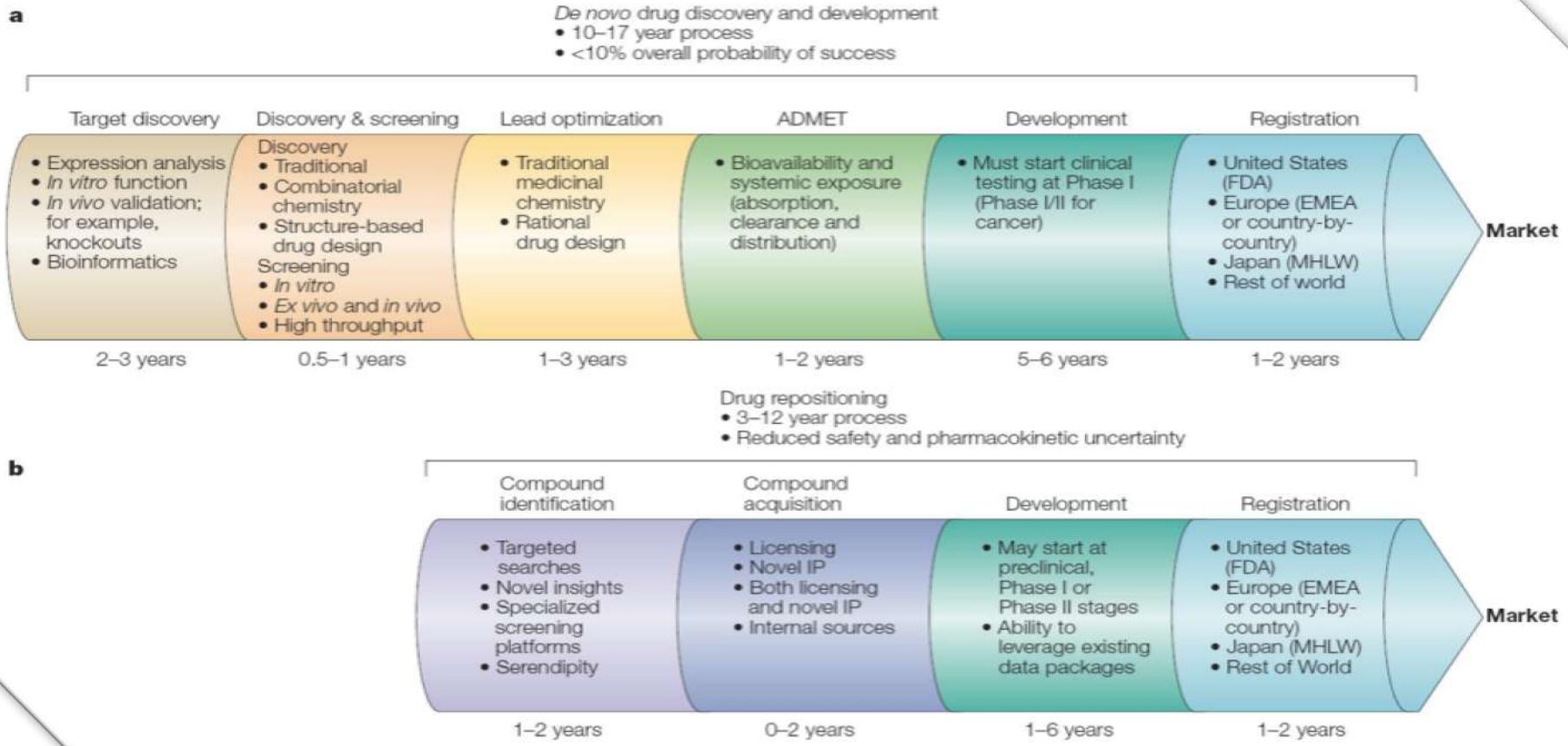
Thalidomide — A Revival Story

Noopur Raje, M.D., and Kenneth Anderson, M.D.

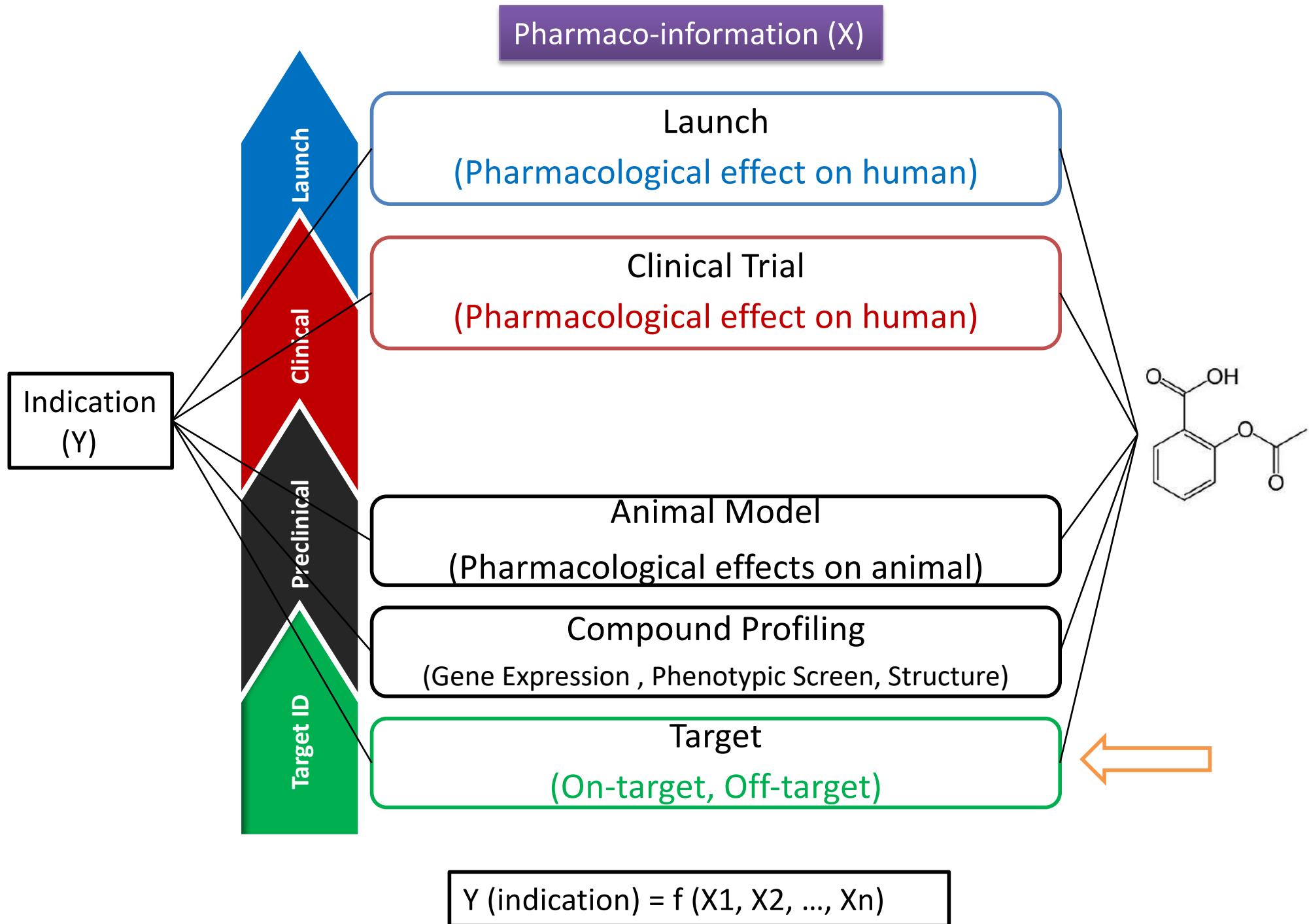
N Engl J Med 1999; 341:1606-1609 | November 18, 1999

© Mayo Foundation for Medical Education and Research. All rights reserved.

Meet the unmet medical needs efficiently



Dependent and Independent Variables in Drug Repositioning

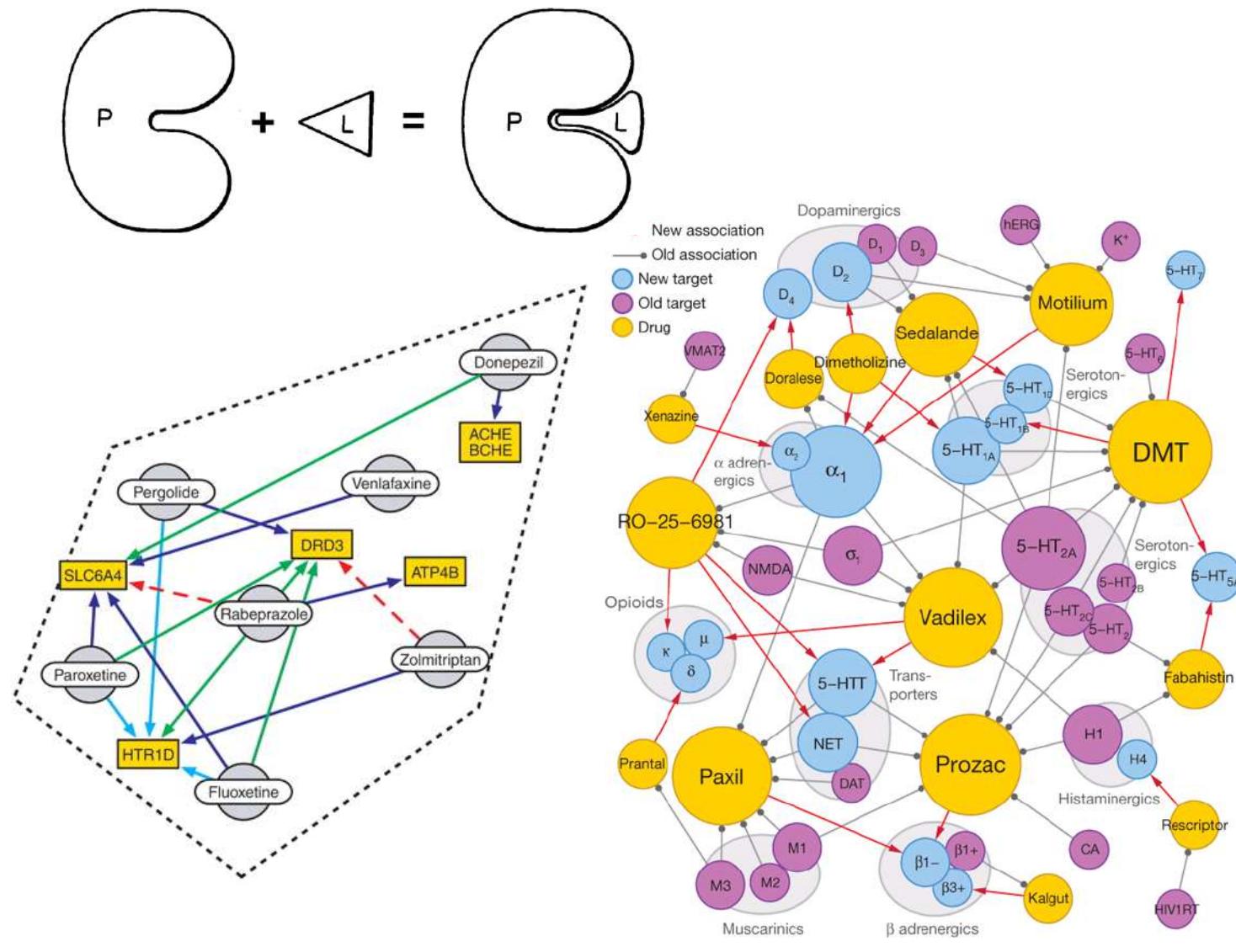
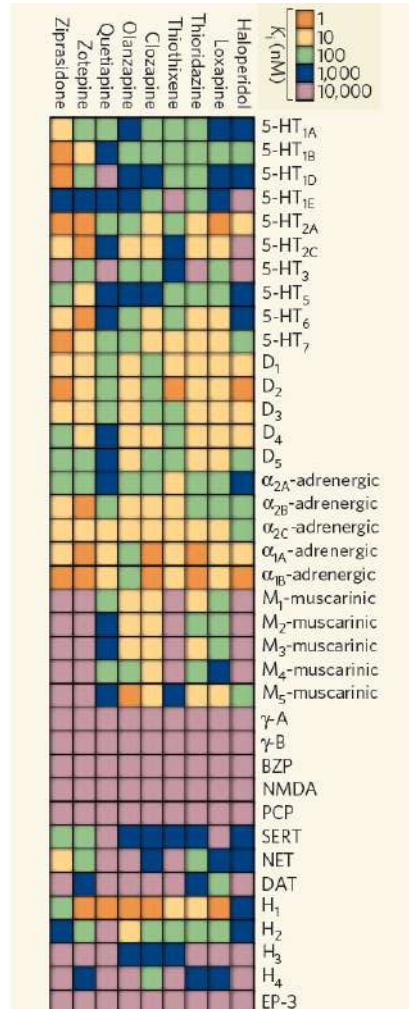


Target
(On-target, Off-target)

Chemical-Protein Interactome (CPI)

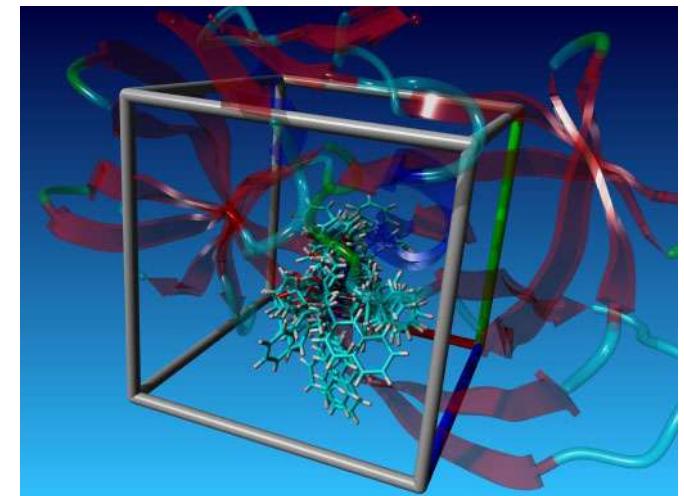
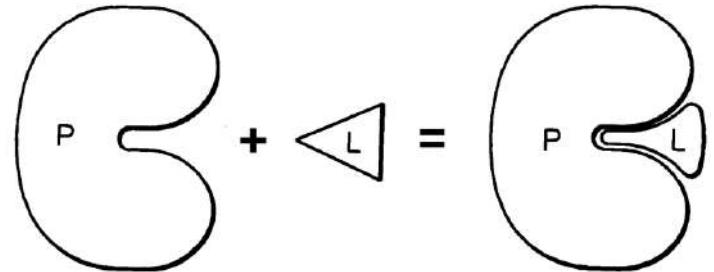
- Introduction of the CPI
- Generate CPI
- CPI data-process
- Case study
 - Drug Repositioning based on CPI

Chemical-protein interactions



The DOCK

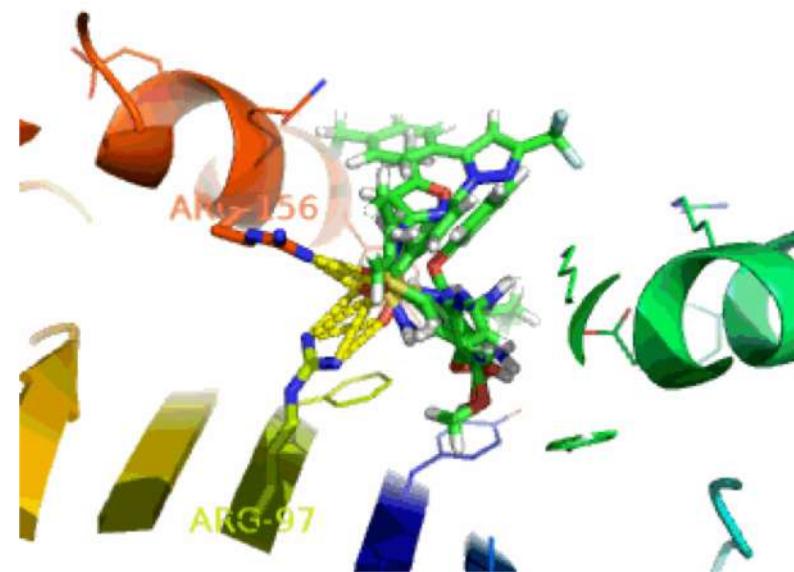
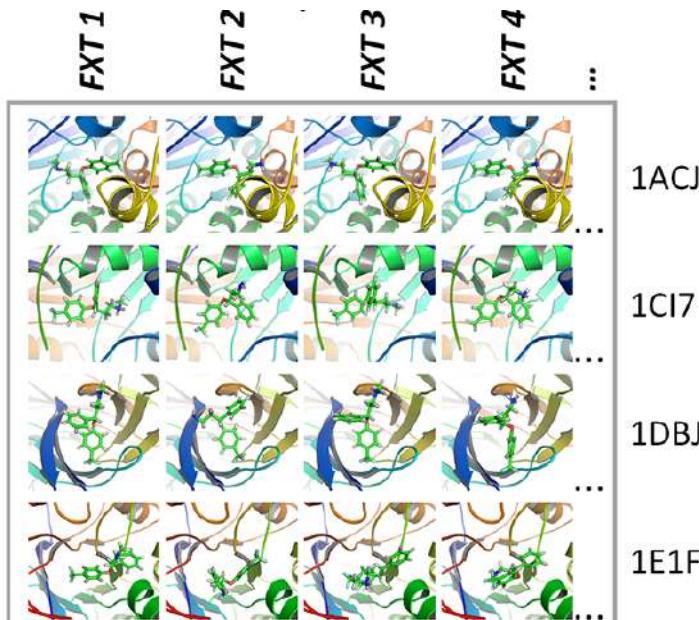
- A program used to simulate the chemical-protein interactions and to measure the interaction strength
- Provide the theoretical binding conformation of the drug's binding to protein
- A lower docking score means a higher binding strength



$$E_{\text{inter}} = \sum_{i=1}^{\text{lig}} \sum_{j=1}^{\text{rec}} \left(\frac{A_{ij}}{r_{ij}^a} - \frac{B_{ij}}{r_{ij}^b} + 332.0 \frac{q_i q_j}{Dr_{ij}} \right),$$

van der Waals and electrostatic interaction

Binding conformation in Chemical-Protein Interactome (CPI)

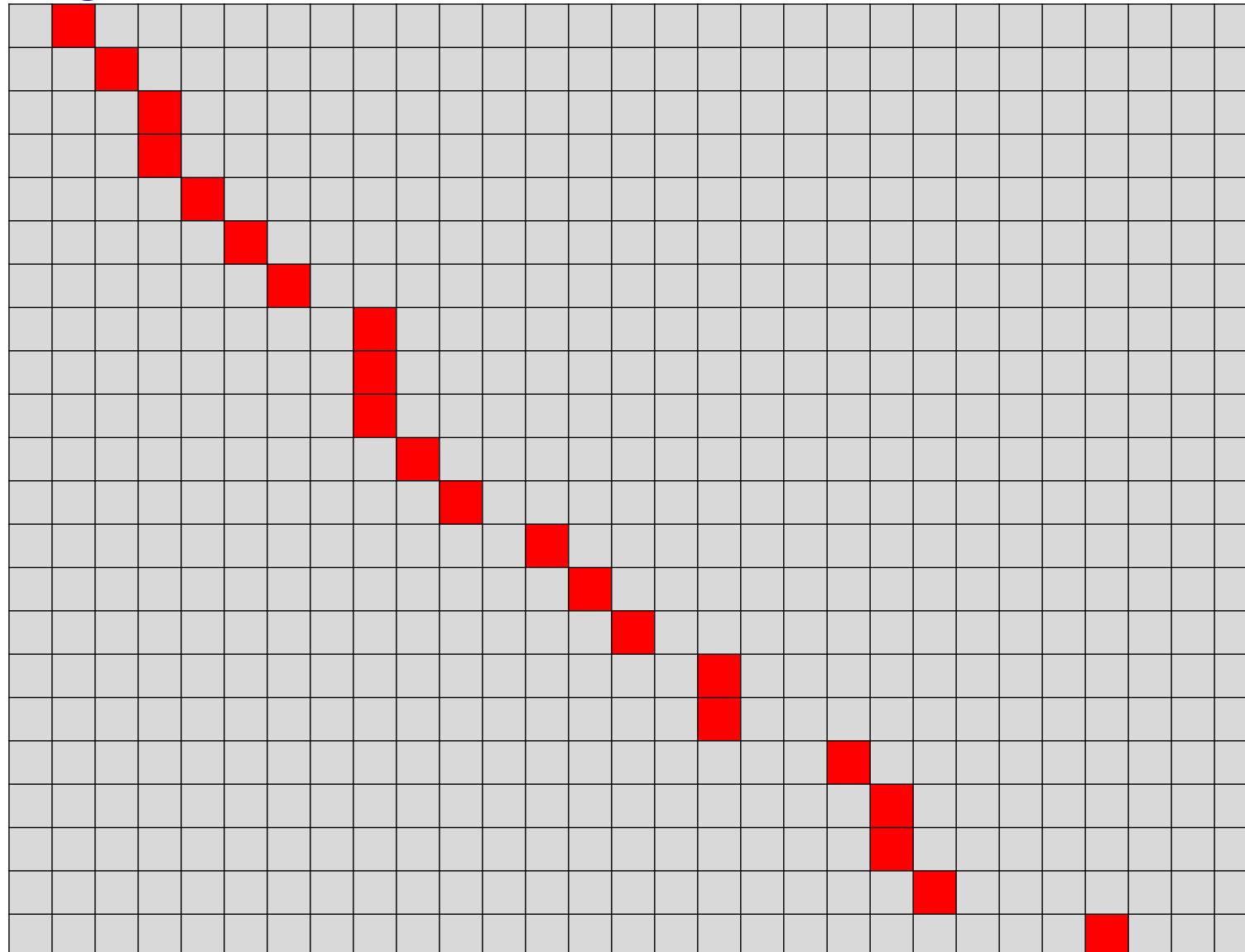


Direct binding model of sulfonamides - MHC I (Cw*4) interactions

Identify the True Drug-Protein Interactions

High Rank ————— Proteins ————— Low Rank

Drugs



■ Known Drug-target Pair

■ Negative

The figure is a heatmap illustrating drug-protein interaction data. The vertical axis is labeled "Drugs" and the horizontal axis is labeled "Proteins". An arrow at the top indicates a ranking from "High Rank" (left) to "Low Rank" (right). The grid cells are colored according to the following legend:

- Red square: Known Drug-target Pair
- Blue square: Unknown
- Grey square: Negative

The heatmap shows a sparse distribution of interactions, primarily consisting of blue ("Unknown") and grey ("Negative") cells. Red ("Known Drug-target Pair") cells are scattered across the matrix, often appearing in small clusters or along the diagonal. The overall pattern suggests a large number of unverified or unknown interactions compared to known ones.

Known Drug-target Pair

Negative

Unknown

False Positive - Tolerant MCC (FPT-MCC)

	p' (Predicted)	n' (Predicted)
P (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad |MCC| = \sqrt{\frac{\chi^2}{n}}$$



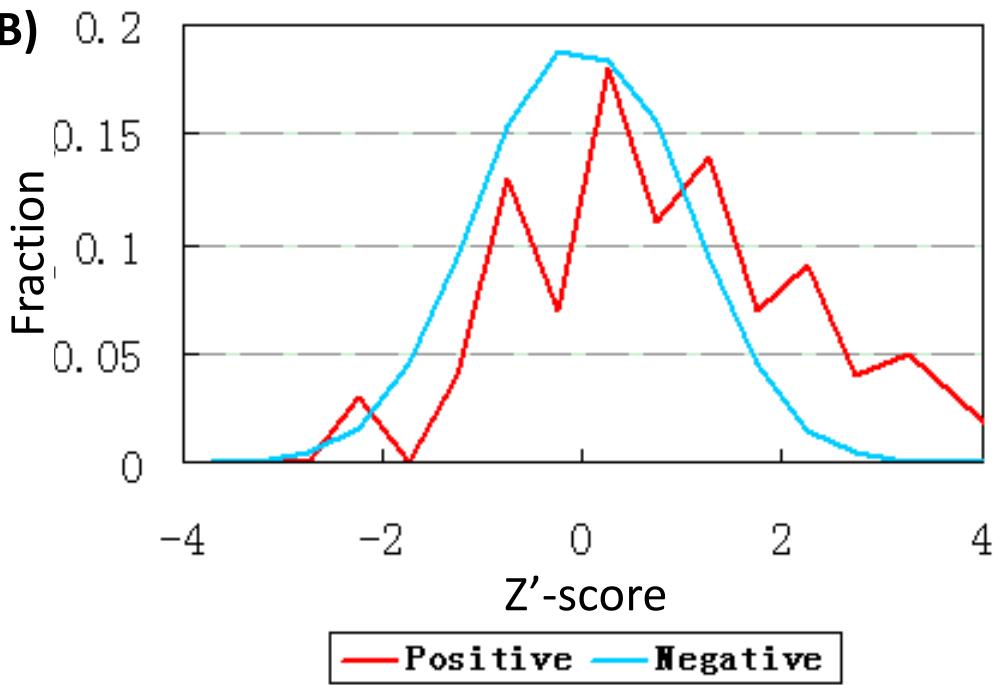
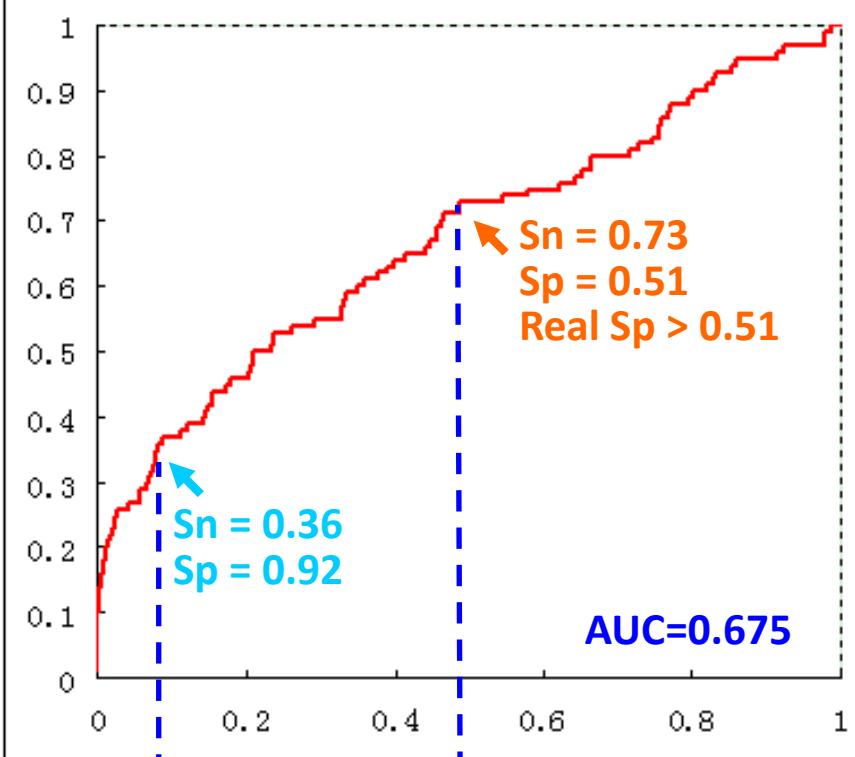
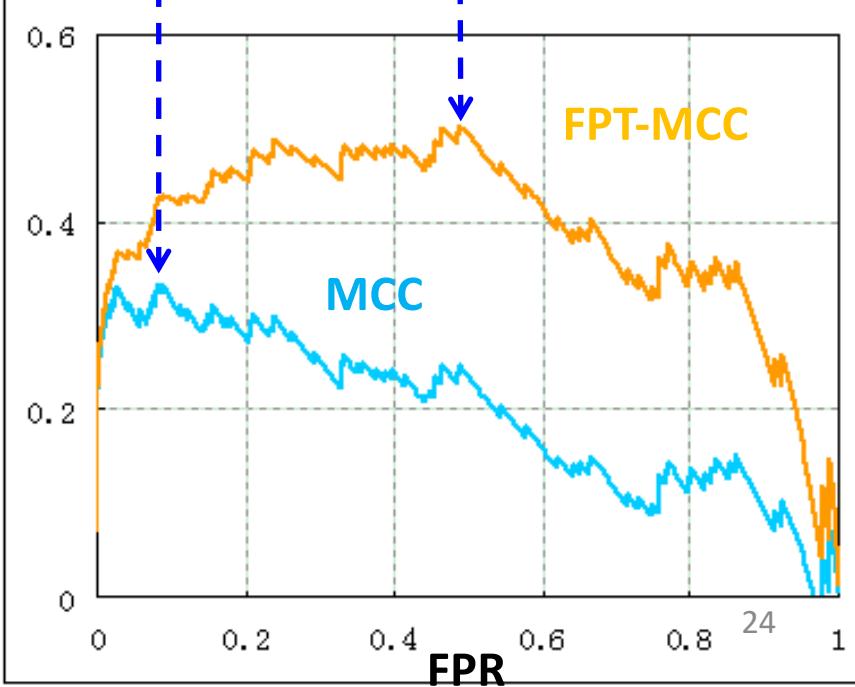
$$(FPT-MCC) = \frac{\textcolor{red}{TP'} \times TN - \textcolor{red}{FP'} \times FN}{\sqrt{(\textcolor{red}{TP'} + \textcolor{red}{FP})(\textcolor{red}{TP'} + FN)(TN + \textcolor{red}{FP})(TN + FN)}}$$

$$\begin{aligned} \textcolor{red}{TP'} &= TP + \alpha FP \\ \textcolor{red}{FP'} &= (1-\alpha)FP \end{aligned}$$

(A)

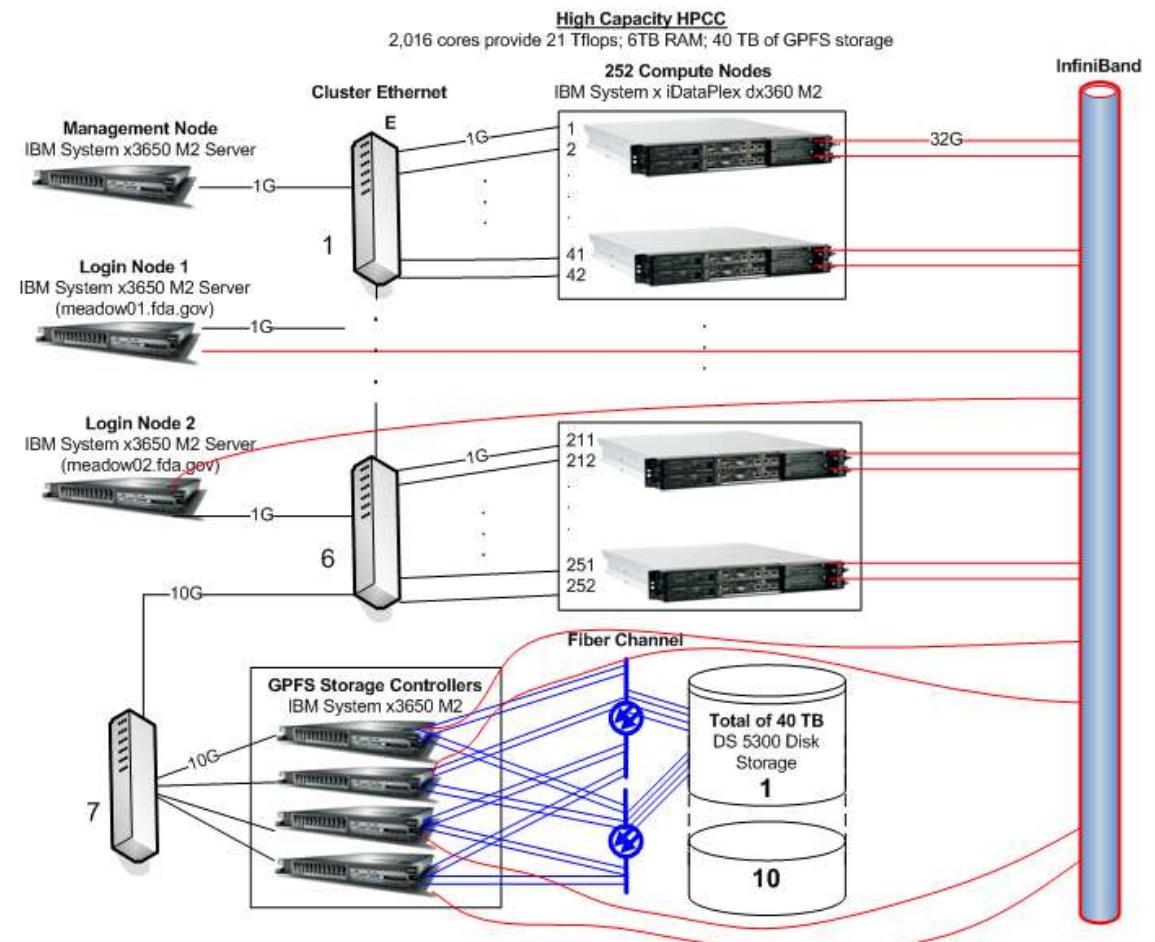
100 Chemicals – 100 Proteins

Class	Positive	Negative
Volume	100	10,000
Mean	1	0
St. Dev	1.5	1

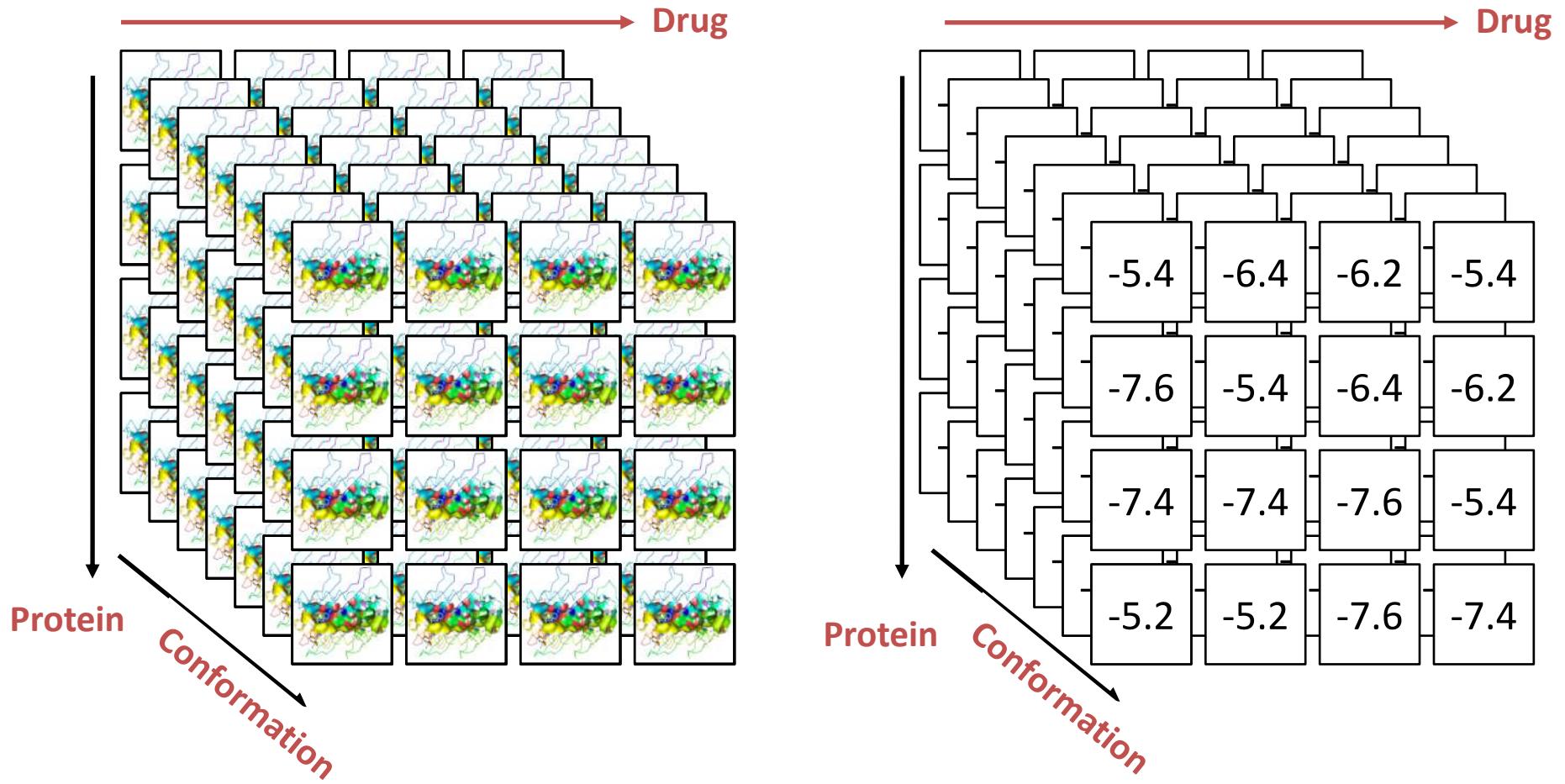
(B)**(C)****(D)**

Resource Specifications for Docking

- Blue Meadow cluster
 - IBM iDataPlex dx360 M2 Server machines & Sun Grid Engine, PBS
 - 252 nodes x 8 cores = 2016 cores
 - 6TB RAM, or 24 GB per node
 - Memory distributed between nodes & shared within nodes



CPU time for constructing CPI

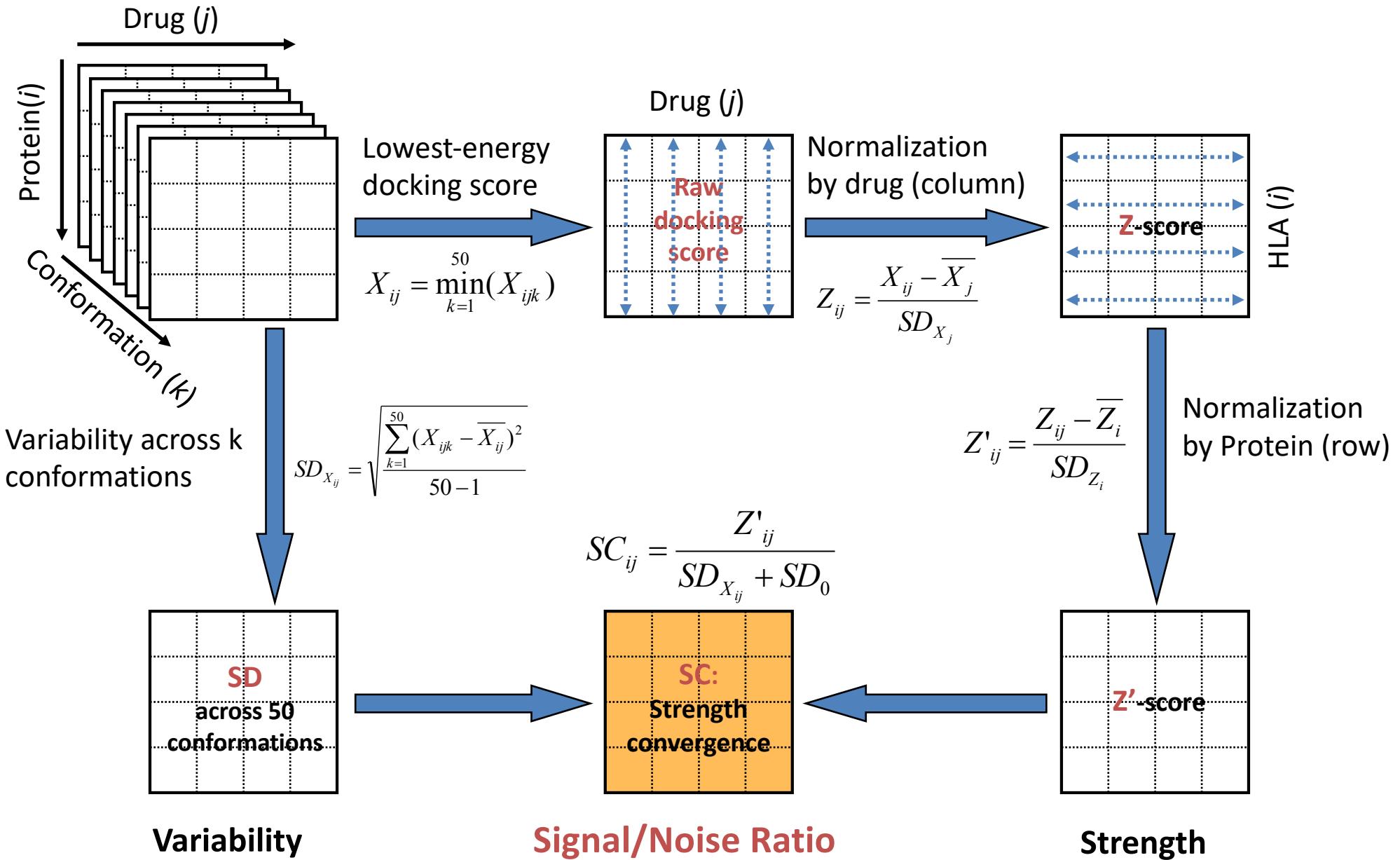


For each drug-protein pair : ~200 seconds per CPU core / 10M 3D conformation and scoring data

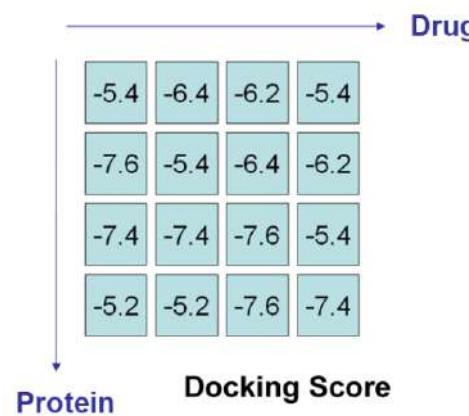
If 15,000 PDB human protein * 10,000 FDA approved drug = 150,000,000 drug-protein

If on IBM Cluster, ~ **172 days / 1,430 TB data**

Docking scores processing – two directional Z-transformation (2DIZ)



Rational of using 2DIZ



**Two Directional Z-transformation (2DIZ)
of Docking Scores X_{ij}**

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{SD_{X_j}} \quad Z'_{ij} = \frac{Z_{ij} - \bar{Z}_i}{SD_{Z_i}}$$

Linear Model of the Docking Scores

$$X_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \quad b = \frac{\sum_{q=1}^n \sum_{k=1}^m (\alpha\beta)_{kj}}{mn}$$

$$Z'_{ij} = \frac{-b\sqrt{n-1}}{\sqrt{(n-1)b^2 + [(\alpha\beta)_{ii} - b]^2}} \quad (i \neq j), \text{ when } (m \rightarrow +\infty, n \rightarrow +\infty)$$

$$Z'_{ij} = \left[(\alpha\beta)_{ij} - b \right] \sqrt{\frac{(n-1)}{(n-1)b^2 + [(\alpha\beta)_{ij} - b]^2}} \quad (i = j),$$

when $(m \rightarrow +\infty, n \rightarrow +\infty)$,

ANOVA of the chemical-protein interactive effect before and after 2DIZ

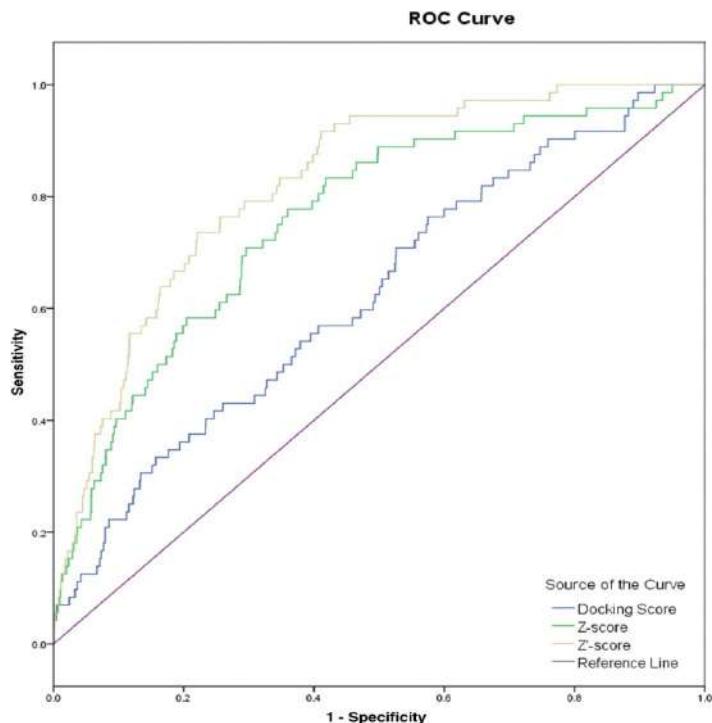
Before 2DIZ				
	Df	Sum Sq	F	p value
Protein	409	2332527	111.22	<2.2e-16
Chemical	254	10330585	793.27	<2.2e-16
Interactive	95344	4888387		

After 2DIZ				
	Df	Sum Sq	F	p value
Protein	409	0	1.37E-19	1
Chemical	254	1052	4.1776	<2.2e-16
Interactive	95344	94546		

The protein effect is huge – not a fair comparison among proteins against a drug molecule

Protein effect has been excluded

Improved performance of the docking scores after applying 2DIZ



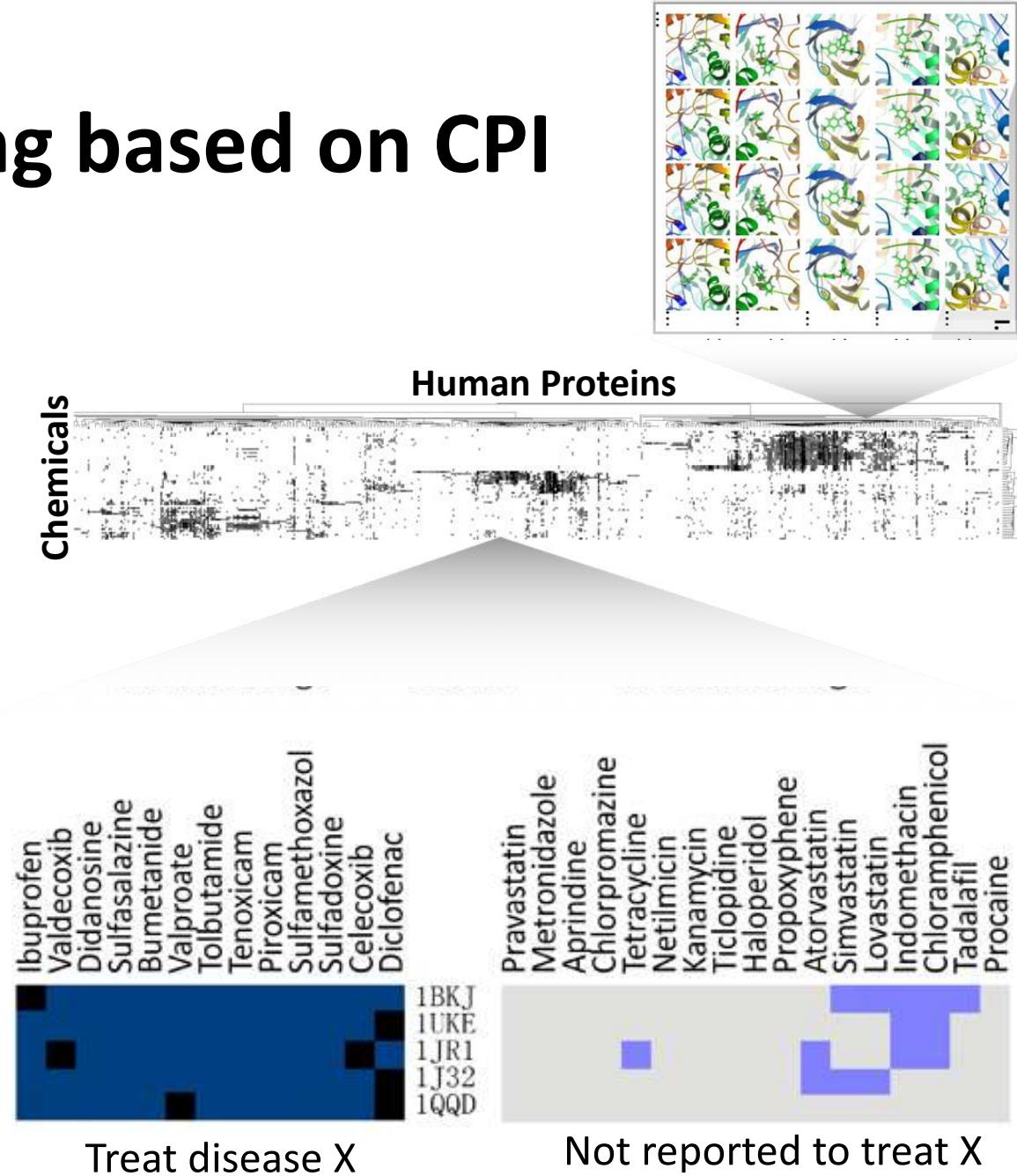
Benchmark structural model set:
100 pockets with their embedded ligands

High variability in ligand structures

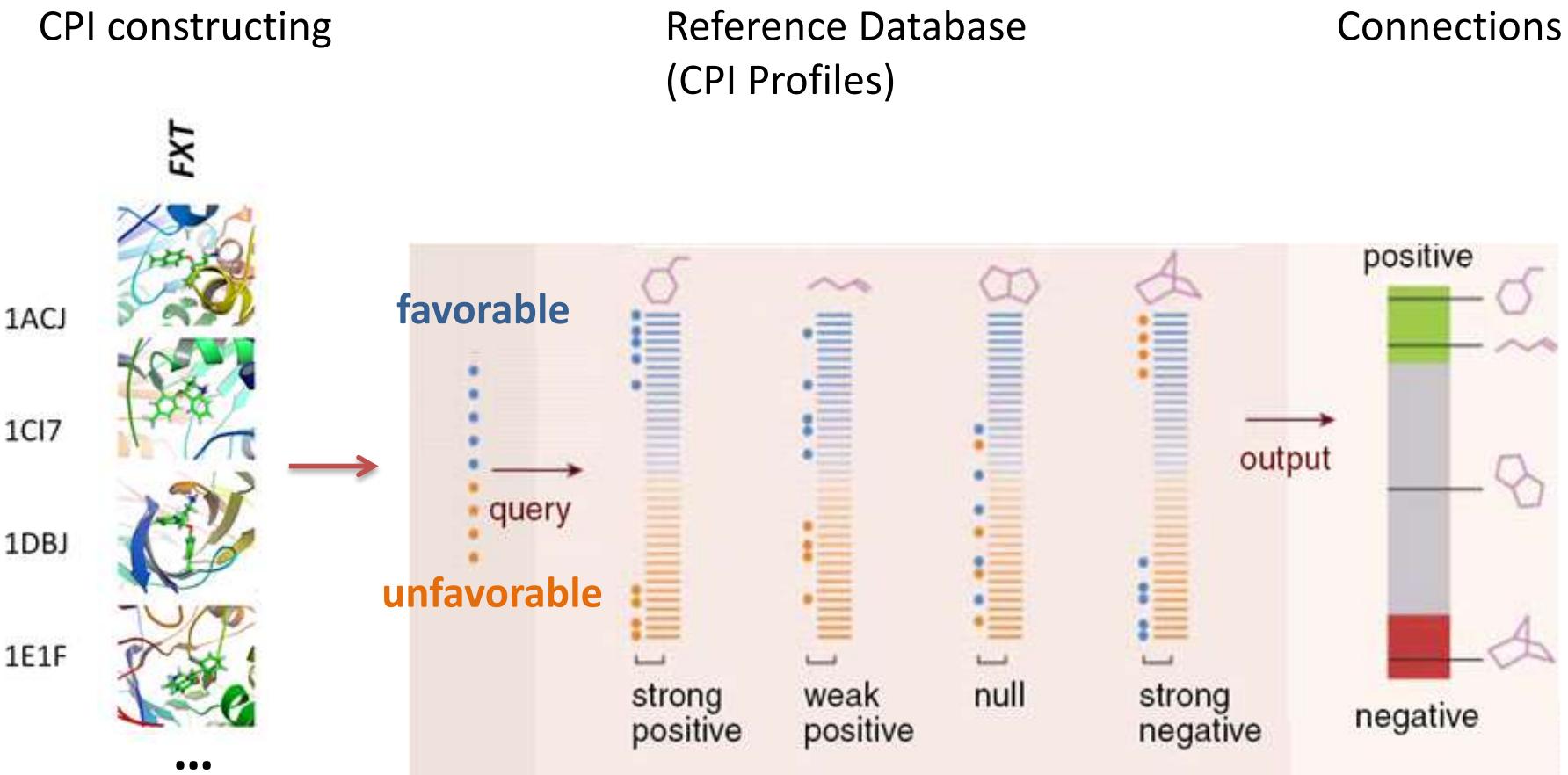
Test Result Variable(s)	AUC	Std. Error ^a	Asymptotic Sig. ^b	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Docking Score	.623	.033	.000	.558	.687
Z-score	.759	.028	.000	.703	.815
Z'-score	.823	.021	.000	.781	.865

Drug Repositioning based on CPI

- New indications are usually caused by unexpected chemical–protein interactions on off-targets
- The interaction profiles could be used as high dimensional representative of the drugs' pharmacological effect

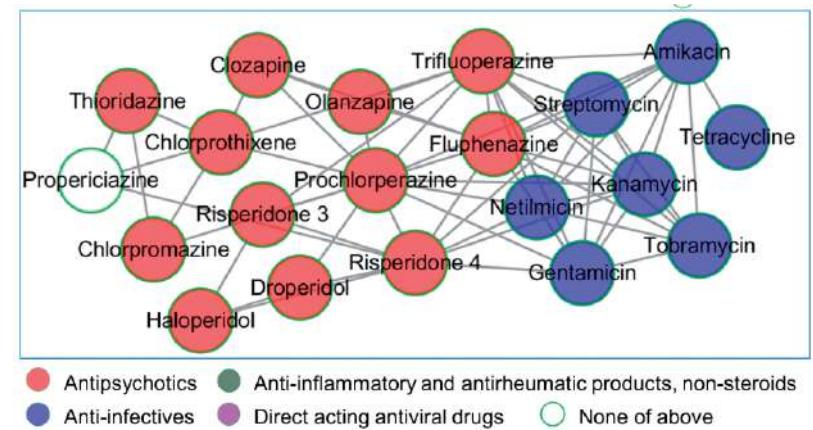
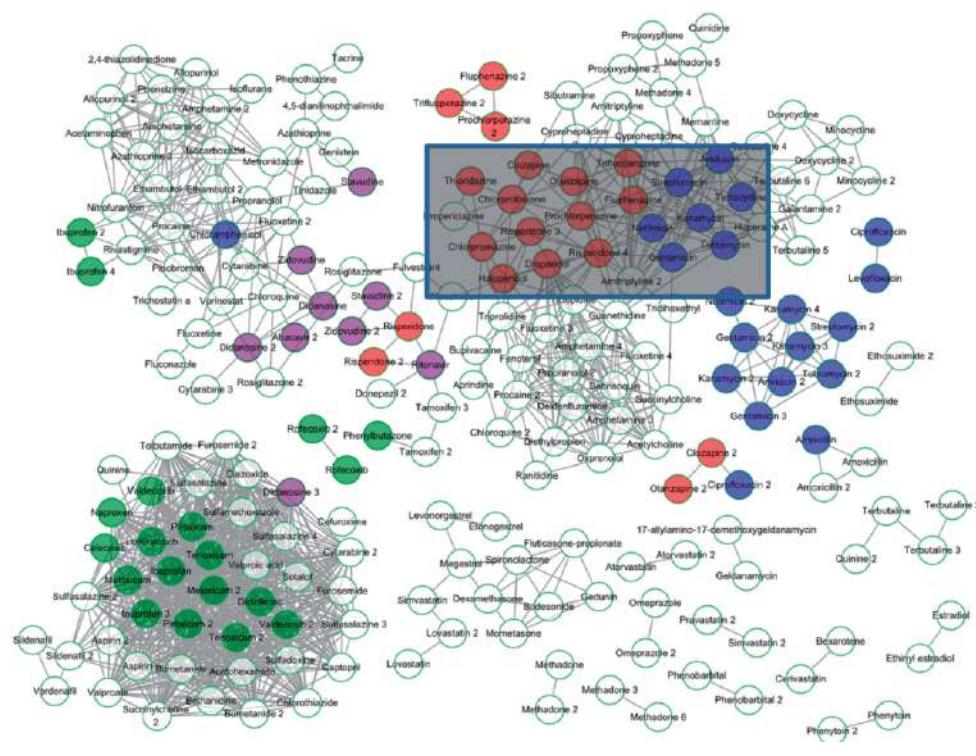


Drug Repositioning based on drug-drug connectivity



Luo, H,..., Yang, L. Nucl. Acids Res. (2011) Web Server Issue; doi: 10.1093/nar/gkr299
Figure Modified from: J. Lamb, ..., E.S. Lander, T.R. Golub. Science. 2006 313(5795):1929-35.

CPI-based drug-drug connectivity network



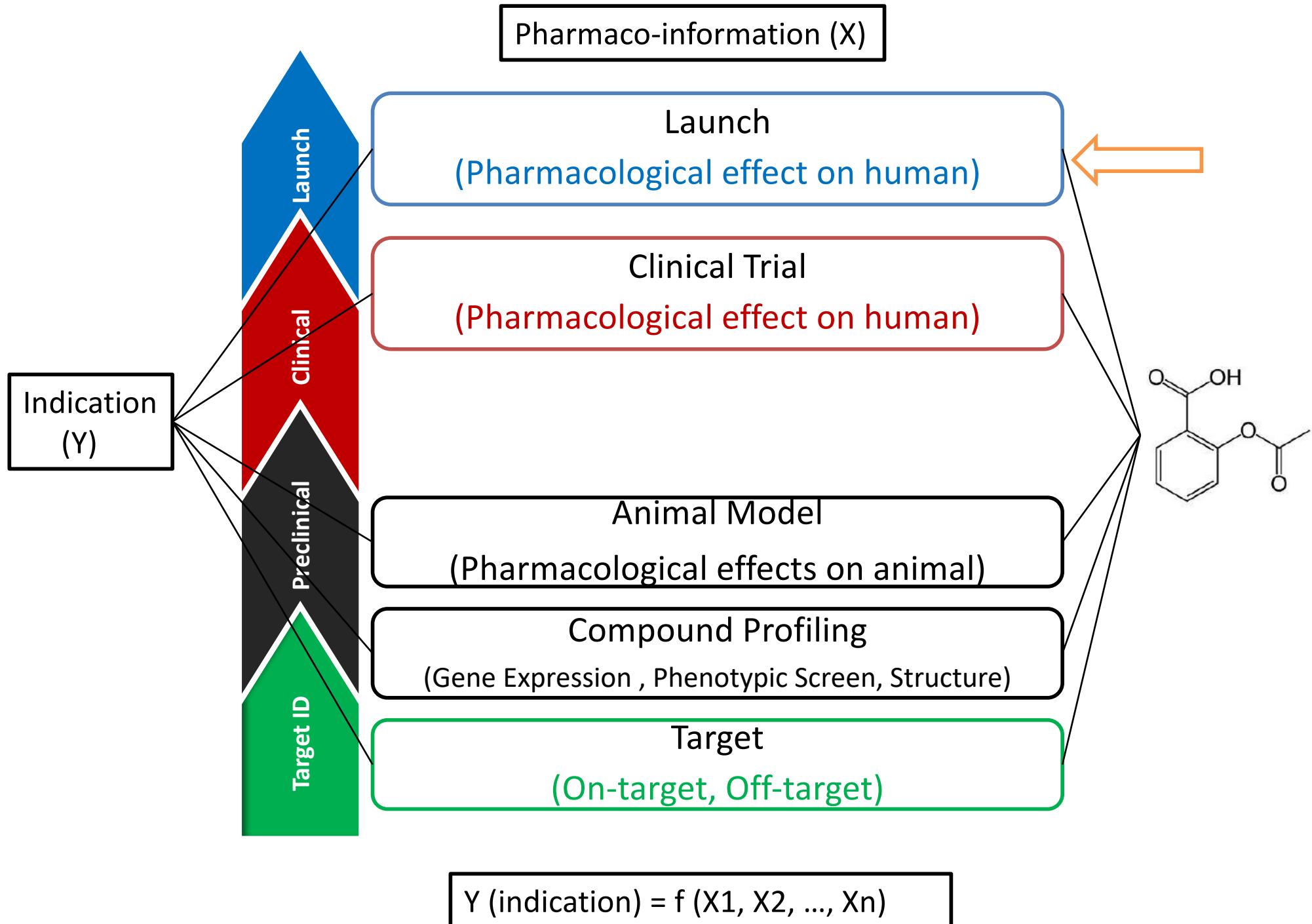
Have successfully predicted the connections between anti-psychotics and anti-infectives

Rani Basu L, et al. Microbiol. Res. 2005;160:95-100.
Chan YY, et al. Antimicrob. Agents Chemother. 2007;51:623

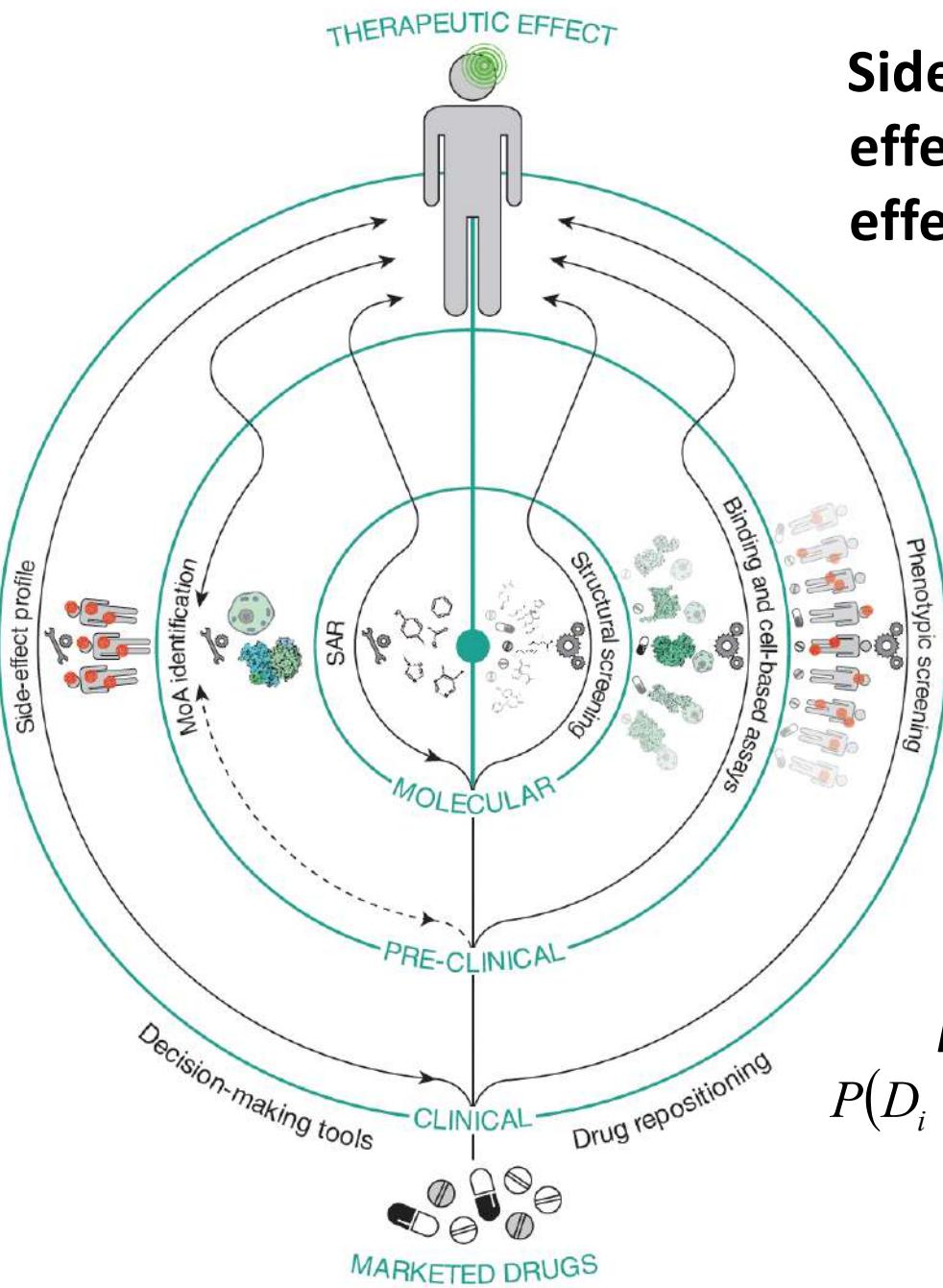
CPI related resources

- Chemical structure
 - STITCH stitch.embl.de
 - DrugBank www.drugbank.ca
- Protein Structure
 - Protein Data Bank www.pdb.org
 - PDBbind www.pdbbind-cn.org
- Docking programs
 - DOCK
 - Autodock
 - Glide
- CPI servers
 - Drug Repositioning CPI <http://cpi.bio-x.cn/drar/>
 - CPI for Drug-Drug Interaction prediction
<http://cpi.bio-x.cn/ddi/>

Dependent and Independent Variables in Drug Repositioning



Rationale of Using Pharmacological Effects in Drug Repositioning



Side-effects (SE) and therapeutic effects are clinical phenotypic effects of drug treatment

- They may associate with each other via underlying mechanism

Clinical phenotypic information comes from patients, not animals

Mimics a human phenotypic 'assay'
May have less translational issue

Quantitative Rational

$$\max(P(D_i | se_1, se_2, \dots, se_m)), i \in (|D|)$$

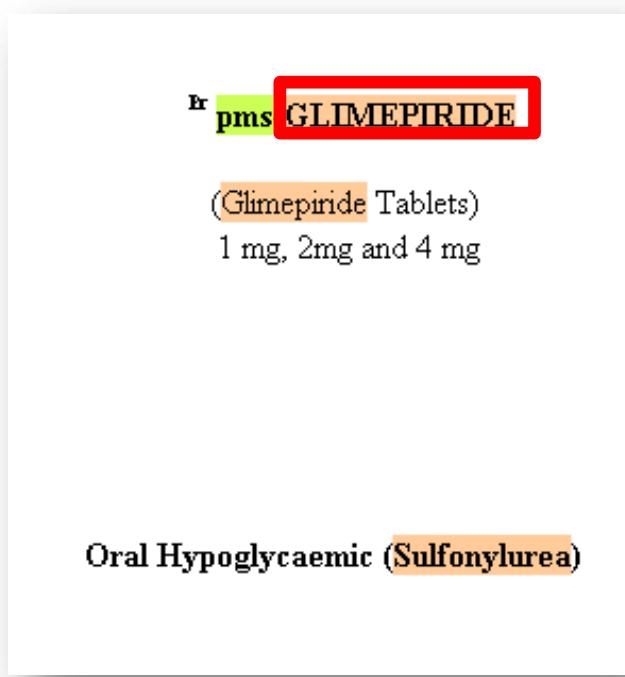
prior

$$P(D_i | se_1, se_2, \dots, se_m) = \frac{P(se_1, se_2, \dots, se_m | D_i) P(D_i)}{P(se_1, se_2, \dots, se_m)}$$

$$P(se_1, se_2, \dots, se_m | D_i) = \prod_{j=1}^m P(se_j | D_i)$$

- *Identification of the disease-side effect associations*

Retrieving side-effect/disease information from drug label and PharmGKB



A screenshot of the PharmGKB interface for the drug "glimepiride". The search bar at the top also has "glimepiride" highlighted with a red box. Below the search bar is a navigation menu with tabs: Clinical PGx, PGx Research, Overview, Properties, Pathways, Is Related To, and Downloads. Underneath the menu, there are three tabs: Related Genes and Targets, Related Drugs and Interactions, and Related Diseases. The "Related Diseases" tab is selected. A table titled "Curated Information" shows relationships between diseases and the drug. The first row shows "Diabetes Mellitus" under "Disease" and "PD" under "Relationship". The second row shows "Diabetes Mellitus, Type 2" under "Disease" and "PD PK" under "Relationship". The entire table is highlighted with a red box.

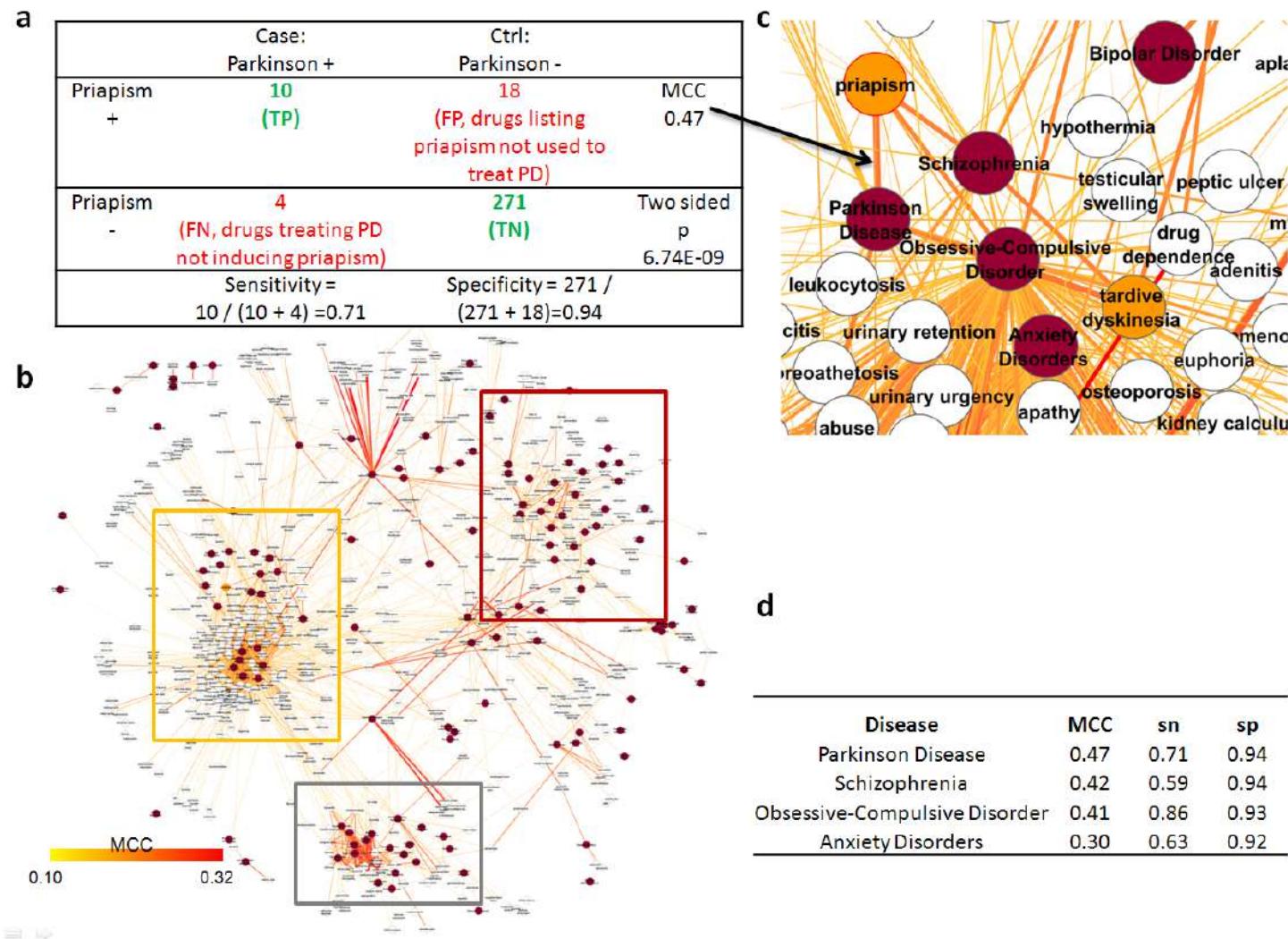
SIDE EFFECT

Skin:

Allergic skin reactions, e.g., pruritus, erythema, urticaria, vasculitis, and morbilliform or maculopapular eruptions, occur in less than 1% of treated patients. Such mild reactions may develop into serious reactions sometimes progressing to shock. These may be transient and may disappear despite continued use of glimepiride if skin reactions persist, the drug should be discontinued. Although there have been no reports for glimepiride, porphyria cutanea tarda

SE → drug → Disease ←

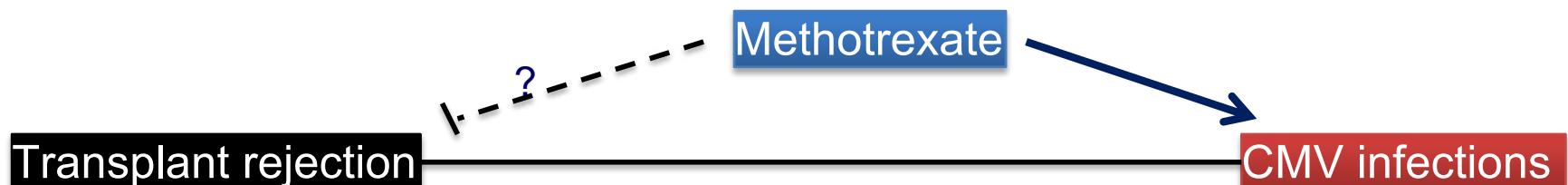
Identification of the disease-side effect associations



584 side effects; 145 diseases; 3175 informative drug-SE associations

Examples of disease-side effect associations with interpretable biological meanings

Disease Class	Disease	Side Effect	MCC	sn	sp	p value	Predictions
Circulation System	Stroke	Positive ANA	0.46	0.47	0.98	1.8E-15	statins, ramipril
Immune System	Transplant rejection	Cytomegalovirus infection	0.75	0.75	0.99	3.5E-06	methotrexate
Metabolite disease	Diabetes Mellitus	Porphyria	0.44	0.50	0.98	8.8E-06	valproic acid, pyrazinamide, naproxen, estradiol
Psychiatric disease	Depressive Disorder	Delusions	0.46	1.00	0.91	1.1E-08	cabergoline, memantine, pergolide
Psychiatric disease	Depressive Disorder	Hyperacusis	0.55	0.88	0.96	9.0E-09	phenytoin, modafinil
Neoplasms	Neoplasms	Constitutional symptoms	0.50	0.56	0.94	2.6E-18	nevirapine



- *Drug Repositioning based on Side Effects (DRoSEf) for marketed drugs*

AUCs of 10-fold cross validations across 145 diseases using multiple SE features

Disease	AUC	Disease	AUC
Amyotrophic Lateral Sclerosis	1	Influenza, Human	0.997
Anemia	1	Leukemia, Lymphocytic, Chronic, B-Cell	1
Arthritis	1	Liver Neoplasms	1
Asthma	0.959	Migraine without Aura	1
Cough	0.998	Myopathy, Central Core	1
Dementia	1	Non-small cell lung cancer	0.986
Diabetic Nephropathies	1	normal tension glaucoma	1
Diarrhea	0.982	Osteonecrosis	0.993
Esophageal Neoplasms	0.983	Osteoporosis, Postmenopausal	1
estrogen-dependent carcinogenesis	1	Pain	0.983
Gastroesophageal Reflux	0.997	Parkinson Disease	0.959
Glioblastoma	1	Peripheral Nervous System Diseases	0.957
Glomerulosclerosis	0.997	Psoriasis	0.962
Heart Diseases	1	Rectal Neoplasms	0.983
Hyperlipidemias	0.981	Rheumatic Diseases	0.994

...

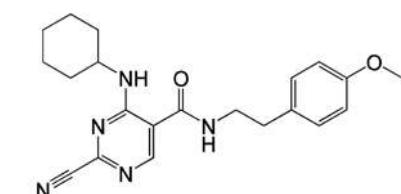
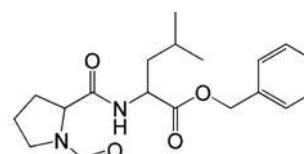
92% of the AUCs were above 0.8

- *DRoSEf for clinical molecules*

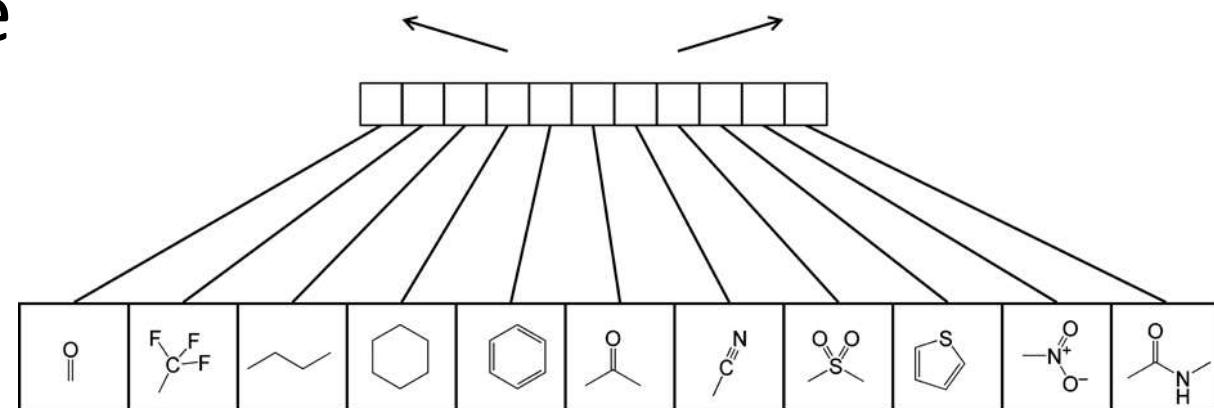
Quantitative Structure Activity Relationship (QSAR) Modeling

- Drug-like properties

- Octanol-water partition coefficient (logP)
- Hydrogen bond donors
- Hydrogen bond acceptors
- Molecular Mass



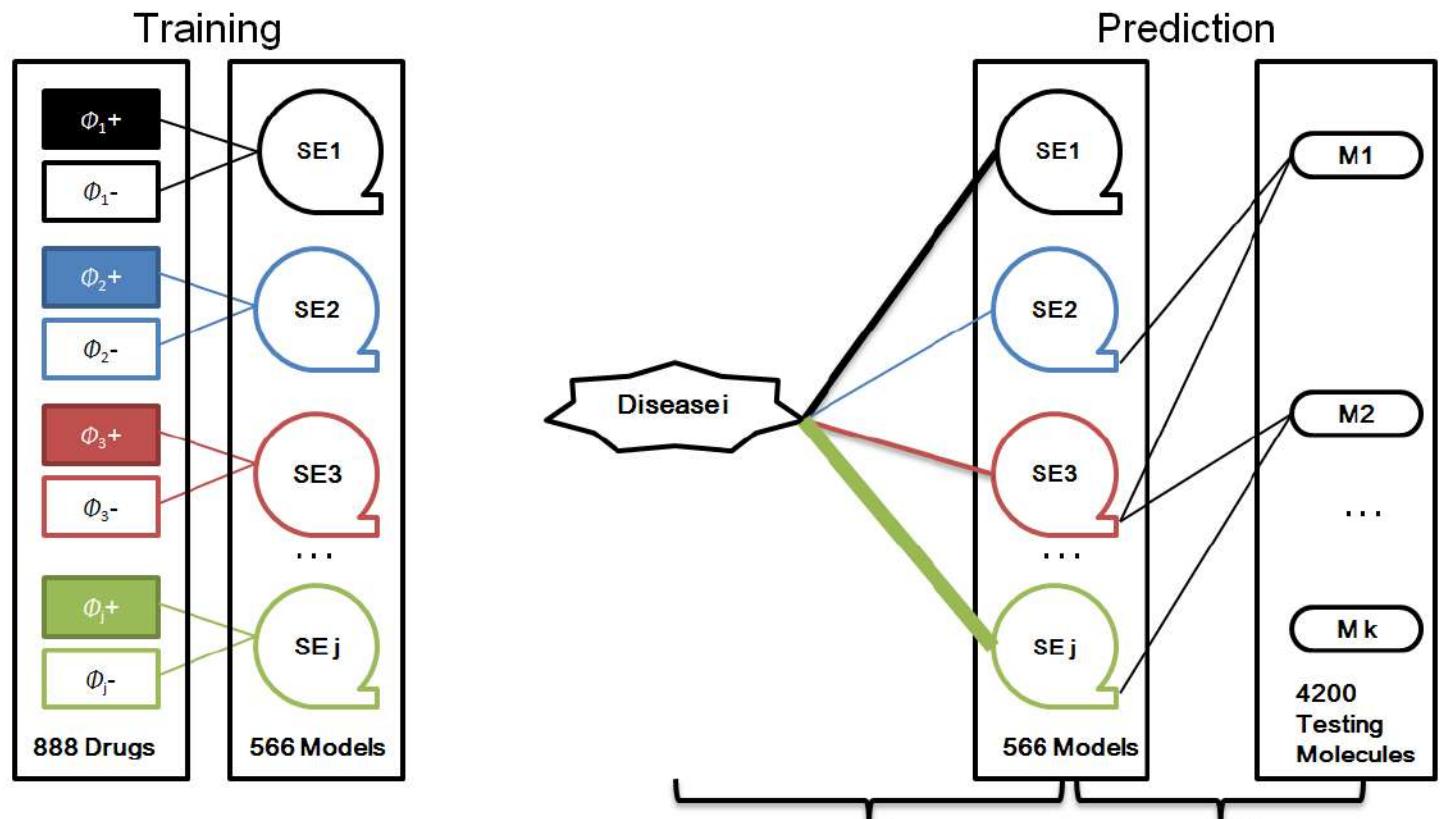
- Structural Signature



Substructure is present ■ or absent □

DRoSEf for clinical molecules

- 4,200 clinical molecules that are indicated for at least one of 101 diseases



$$DS_i = [ds_{i1}, ds_{i2}, \dots, ds_{ij}], \quad j \in [1, 566], \quad i \in [1, 101]$$

$$SM_k = [sm_{1k}, sm_{2k}, \dots, sm_{jk}], \quad j \in [1, 566], \quad k \in [1, 4200]$$

$$ds_{ij} \in \{b_{ij}, mcc_{ij}, mcc_{ij}^4, sn_{ij}, sn_{ij}^4, sp_{ij}, sp_{ij}^4\}$$

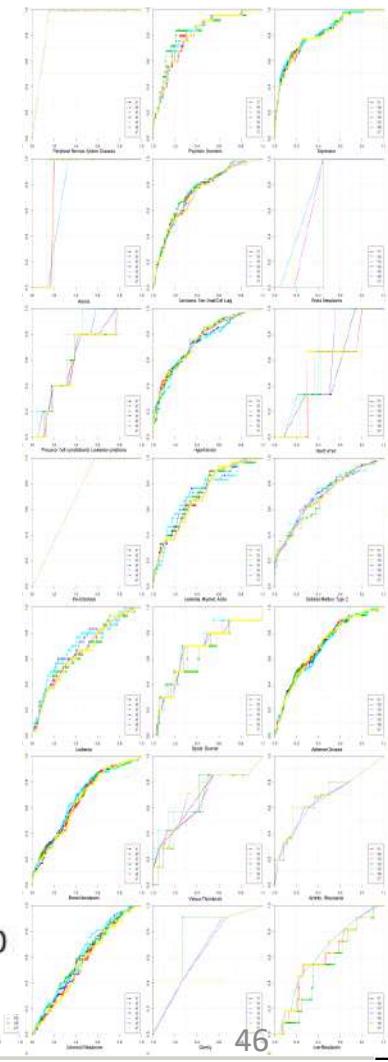
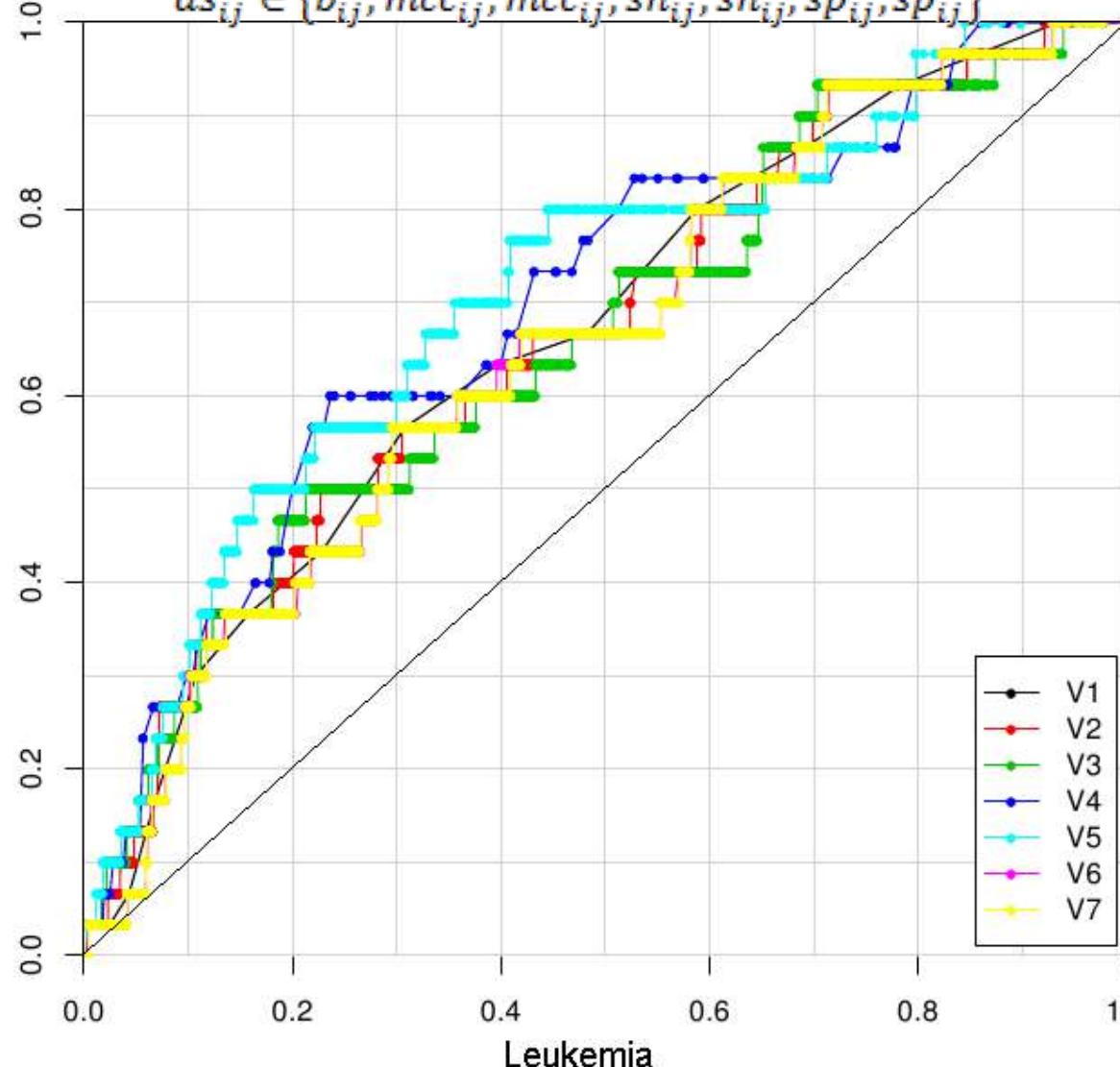
$$\Theta_{ik} = \sum_{j=1}^{566} ds_{ij} sm_{jk} \rightarrow \Theta_{i2} > \Theta_{i1}$$

Prediction results for clinical molecules

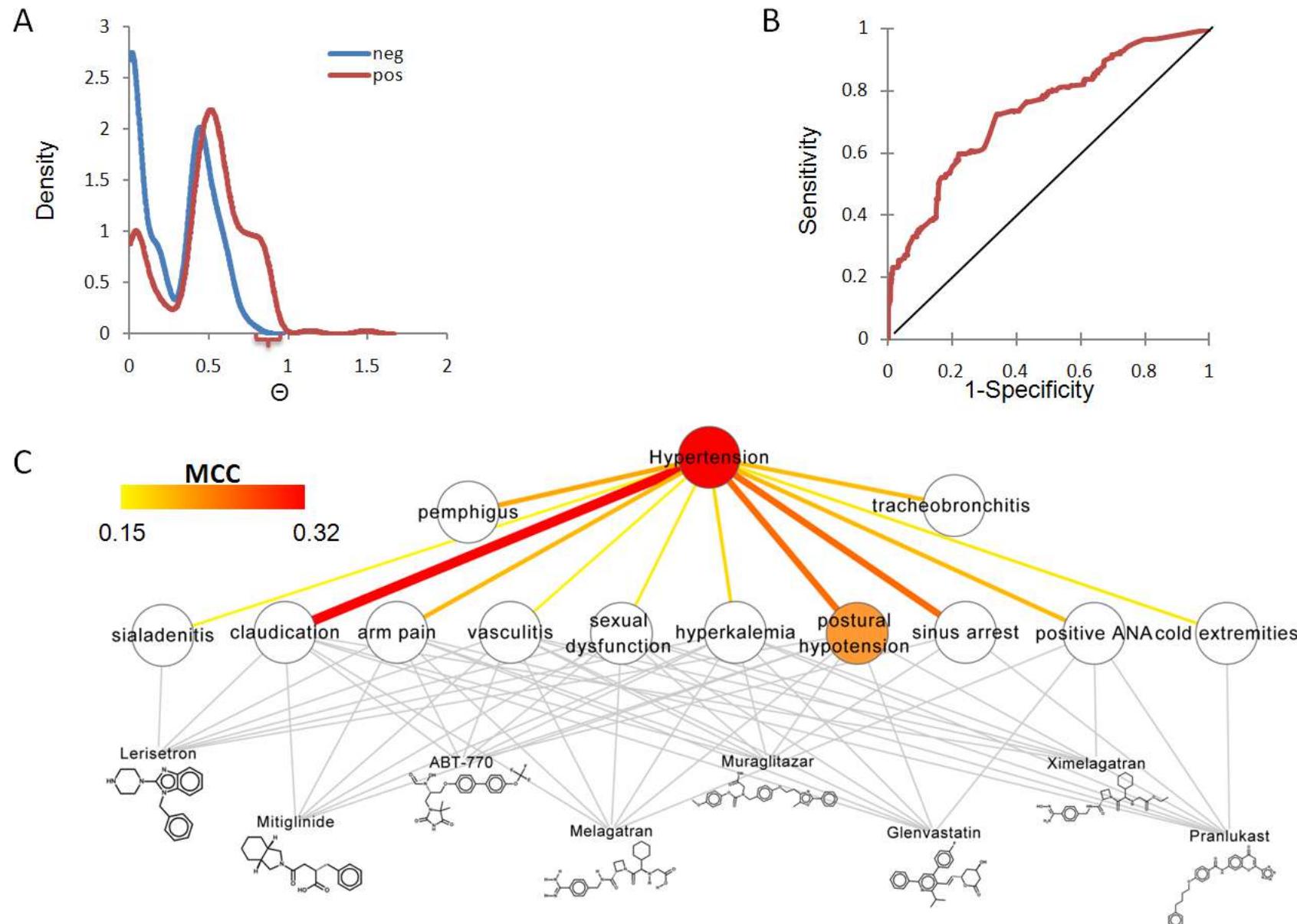
4200 Molecules * 101 Disease Endpoints

$$ds_{ij} \in \{b_{ij}, mcc_{ij}, mcc_{ij}^4, sn_{ij}, sn_{ij}^4, sp_{ij}, sp_{ij}^4\}$$

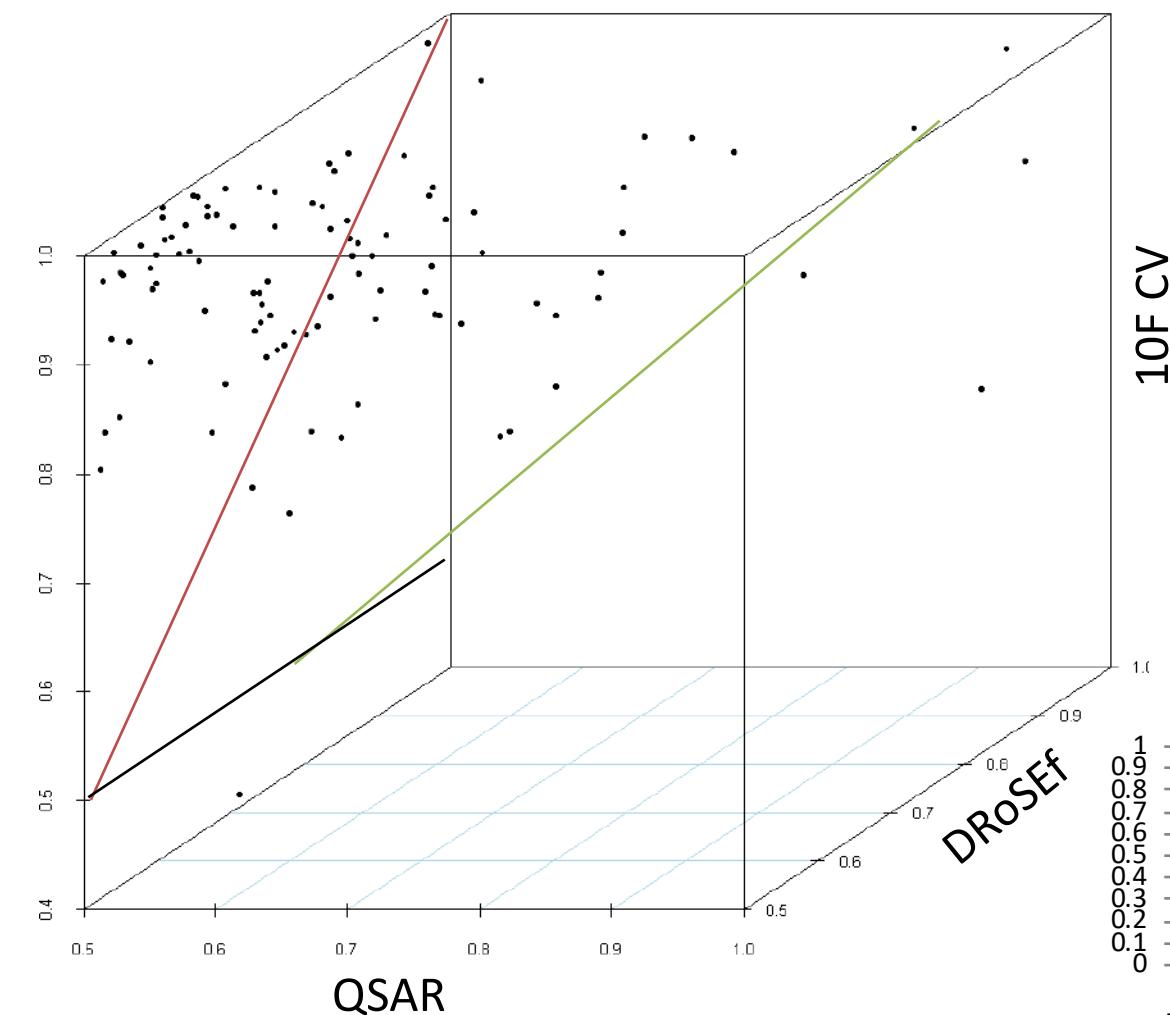
Disease category	Disease ^a
Neuropsychiatric	Depression
	Depressive Disorder
	Schizophrenia
	Depressive Disorder
	Anxiety Disorders
Neoplasms	Stomach Neoplasms
	Carcinoma, Non-Sm
	Lung Neoplasms
	Neoplasms
	Lymphoma
	Leukemia
	Head and Neck Ne
Others	Hypertension
	Diabetes Mellitus, T



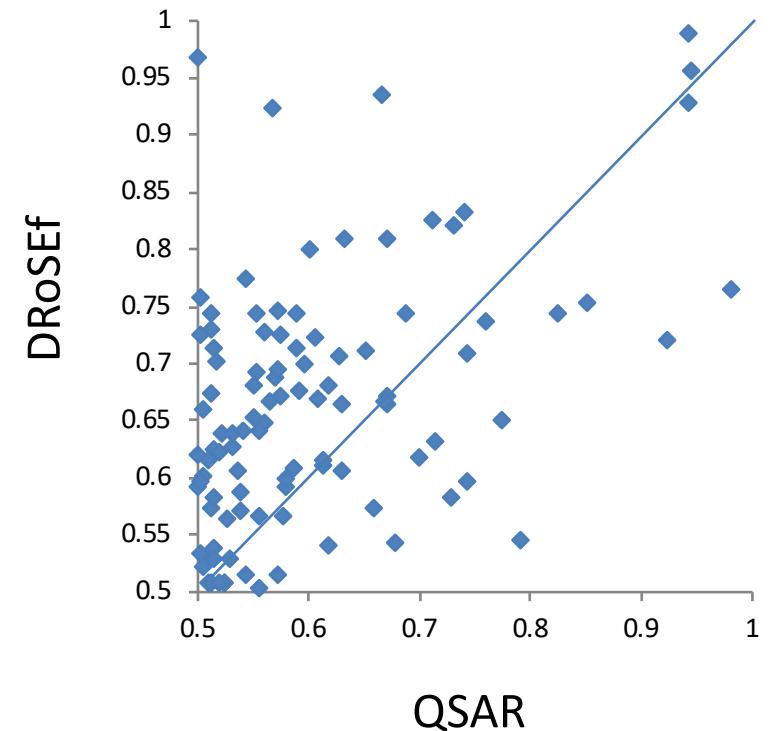
Case Study: Predict drugs' repositioning potential for hypertension



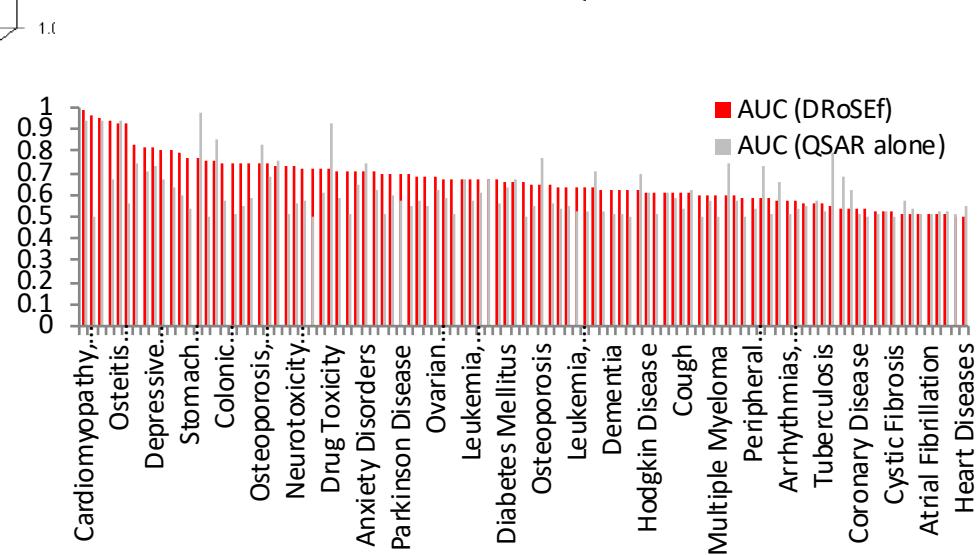
DRoSEf vs. QSAR alone



AUC comparison of different methods



QSAR



Take-home message

- Drug indication can be suggested only based on clinical side-effects
- DRoSEf may also suggest the neglected pathogenesis of disease, inspiring the basic research of the human diseases
 - For example, studying *porphyria* may help discover potential new mode of action for diabetes therapy

Summary of the data-mining in drug repositioning

- The dependent variable is disease (Y)
- Independent variables (X)
 - Chemical-protein interactome profile
 - Side Effect
- Prediction
 - The predicted results should have biomedical explanation

Side effect related resources

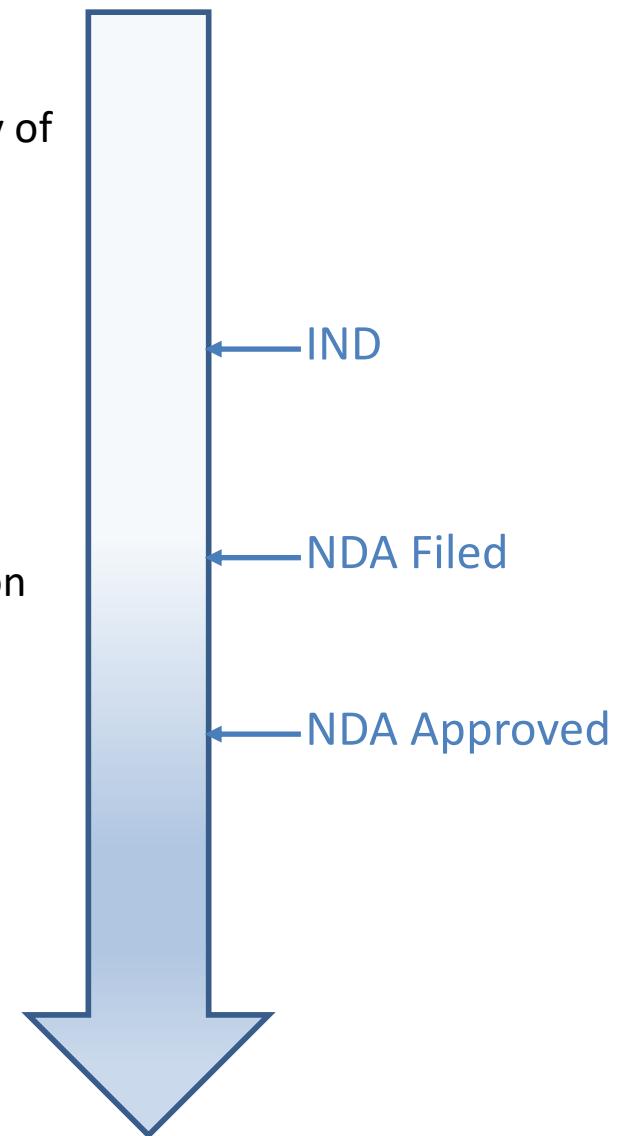
- Side effect
 - SIDER <http://sideeffects.embl.de/>
 - FAERS
- Drug-disease relationship
 - PharmGKB www.Pharmgkb.org
 - Pipeline®
 - Metabase® – Thomson Reuters
- Molecular fingerprints
 - Daylight
 - CDK
 - MACCS

Outline

- Introduction of Drug Discovery and Development
- Motivation of Data Mining
- Case Study: Drug Repositioning
- **Case Study: Real-World Evidence**
- Data Sources for Data Mining Applications
- Challenges and Summary

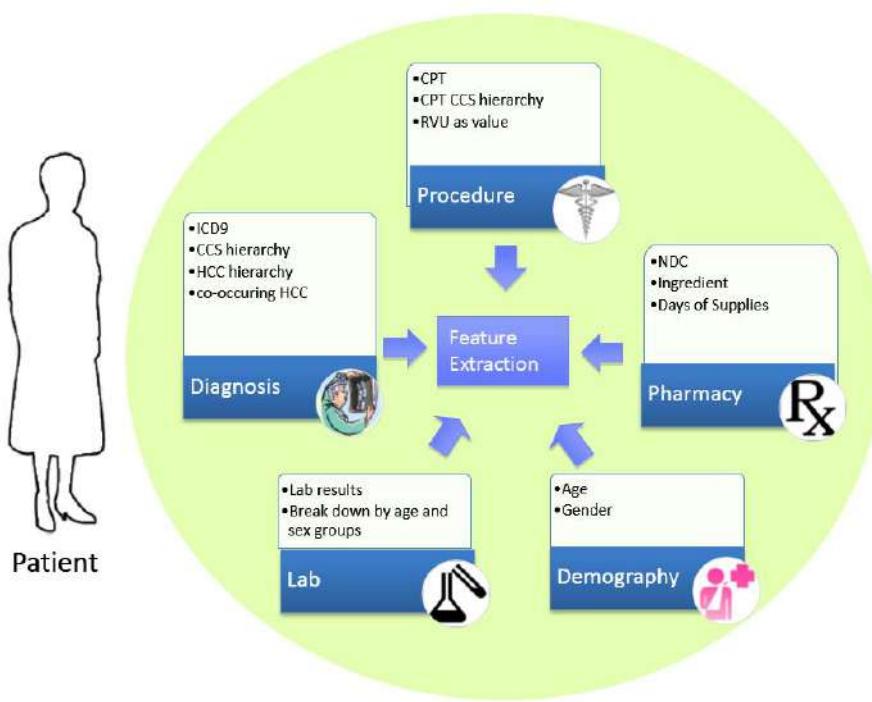
Where do “data” come from?

- Pre-clinical studies
 - Provides a first assessment of the expected safety and efficacy of a compound using proven animal models
- Phase I
 - Safety focus and the beginnings of efficacy, dose ranging, and tolerability
- Phase II
 - Demonstrate safety and efficacy in well controlled (generally masked) randomized studies sufficient for market authorization
- Phase III
 - Expanded trials in different use situations or populations
- Phase IV
 - Post marketing safety or new indications
- Real World Evidence
 - Evaluations of safety, effectiveness and outcomes in “routine” clinical practice



What is “Real World Evidence”

- RWE is clinical observations other than randomized clinical trials (RCT).
 - RCT are expensive and in far smaller scale
- RWE is observations on human in the clinical stage
 - less of a translational issue
 - “Omics“ information (genomics, proteomics, metabolomics, etc.) is not yet widely available in everyday clinical practice
 - Other than "omics", numerous external factors (e.g., environment, diet and exercise) affect response to medication
- RWE is not only vast but also varied in type and source: electronic medical records (EMR), claims data, and even social media.

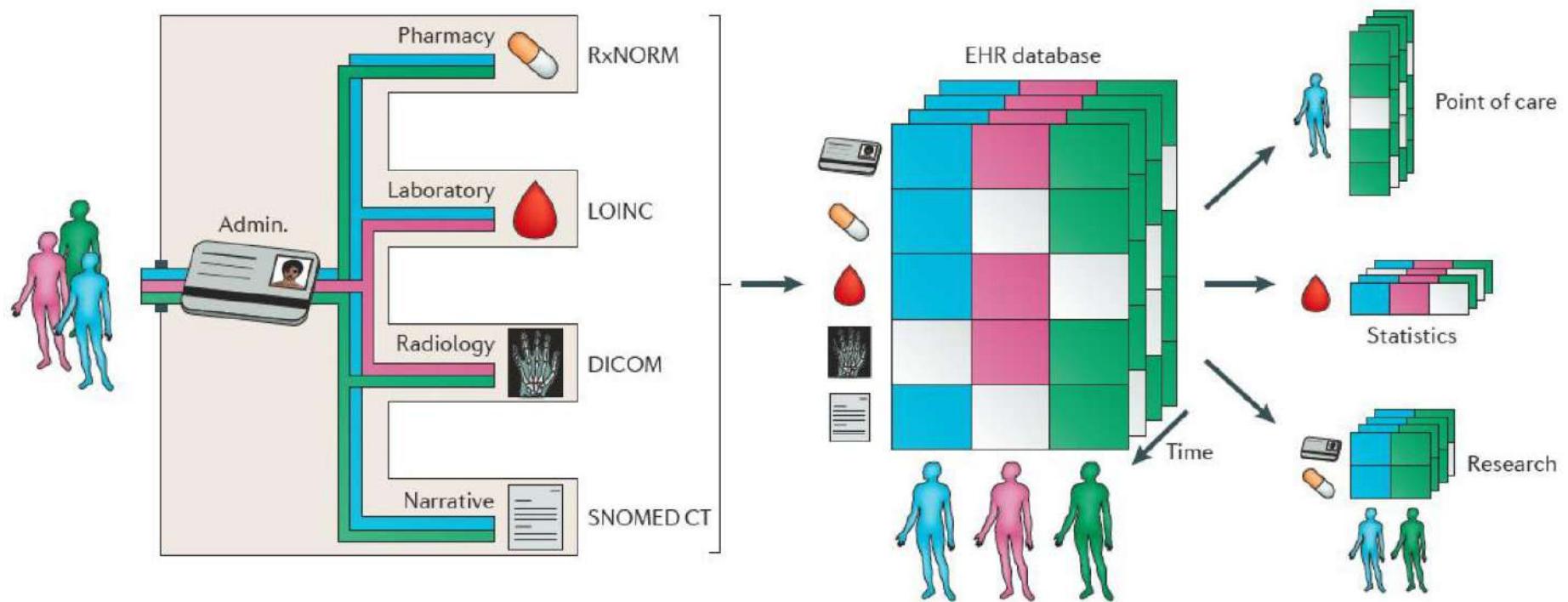


facebook

twitter



EHR data collection and analysis



Effectively integrating and efficiently analyzing various forms of healthcare data over a period of time can answer many of the impending healthcare problems.

Diagnosis data - ICD codes

- ICD stands for International Classification of Diseases
- ICD is a hierarchical terminology of diseases, signs, symptoms, and procedure codes maintained by the World Health Organization (WHO)
- In US, most people use ICD-9, and the rest of world use ICD-10
- Pros: Universally available
- Cons: medium recall and medium precision for characterizing patients

Hypertensive disease (401 - 405)

- (401) Essential hypertension
 - (401.0) Hypertension, malignant
 - (401.1) Hypertension, benign
 - (401.9) Hypertension, Unspecified
- (402) Hypertensive heart disease
- (403) Hypertensive renal disease
 - (403.0) Malignant hypertensive renal disease
 - (403.1) Benign hypertensive renal disease
- (404) Hypertensive heart and renal disease
- (405) Secondary hypertension
 - (405.0) Malignant secondary hypertension
 - (405.01) Hypertension, renovascular, malignant
 - (405.1) Benign secondary hypertension
 - (405.11) Hypertension, renovascular benign

Procedure data - CPT codes

- CPT stands for Current Procedural Terminology created by the American Medical Association
- CPT is used for billing purposes for clinical services
- Pros: High precision
- Cons: Low recall

Codes for surgery: 10021 - 69990

- (10021 - 10022) general
- (10040 - 19499) integumentary system
- (20000 - 29999) musculoskeletal system
- (30000 - 32999) respiratory system
- (33010 - 37799) cardiovascular system
- (38100 - 38999) hemic and lymphatic systems
- (39000 - 39599) mediastinum and diaphragm
- (40490 - 49999) digestive system
- (50010 - 53899) urinary system
- (54000 - 55899) male genital system
- (55920 - 55980) reproductive system and intersex
- (56405 - 58999) female genital system
- (59000 - 59899) maternity care and delivery
- (60000 - 60699) endocrine system
- (61000 - 64999) nervous system
- (65091 - 68899) eye and ocular adnexa
- (69000 - 69979) auditory system

Lab results

- The standard code for lab is Logical Observation Identifiers Names and Codes (LOINC®)
- Challenges for lab
 - Many lab systems still use local dictionaries to encode labs
 - Diverse numeric scales on different labs
 - Often need to map to normal, low or high ranges in order to be useful for analytics
 - Missing data
 - not all patients have all labs

Hematology ABG Analysis

Specimen: Arterial blood

Date and time specimen gathered: 07/21/2010 21:42pm

Blood Gases:

Acid/ Base:

pH

Results:

7.27

Reference Range:

7.35-7.45

Flag:

(L)

pCO₂

48mmHg

35-45 mmHg

(H)

pO₂

92mmHg

80-100 mmHg

HCO₃

25 mEq/L

24-26 mEq/L

O₂ sat

97%

95-100%

Medication

- Standard code is National Drug Code (NDC) by Food and Drug Administration (FDA), which gives a unique identifier for each drug
 - Not used universally by EHR systems
 - Too specific, drugs with the same ingredients but different brands have different NDC
- RxNorm: a normalized naming system for generic and branded drugs by National Library of Medicine
- Medication data can vary in EHR systems
 - can be in both structured or unstructured forms
- Availability and completeness of medication data vary
 - Inpatient medication data are complete, but outpatient medication data are not
 - Medication usually only store prescriptions but we are not sure whether patients actually filled those prescriptions



Clinical notes

- Clinical notes contain rich and diverse source of information
- Challenges for handling clinical notes
 - Ungrammatical, short phrases
 - Abbreviations
 - Misspellings
 - Semi-structured information
 - Copy-paste from other structure source
 - Lab results, vital signs
 - Structured template:
 - SOAP notes: Subjective, Objective, Assessment, Plan

Enter case note

<-Back Post the note Check spelling Client: 9019 - Elinor Dashwood

TX plans Staff: JF - Ferrara, Jessica Program: LB - Long Beach - Ocean

Enter note | Lookup previous notes | Medications | Billing information |

Activity: 003 - CLINICAL VISIT > 30 MINU | ? Date: 06/24/2004 | ?

Duration: 2 : 15 | ? Time: 1:00 PM | ?

Contact: O - Other | Location: 03 - Main site | ?

On site? Yes No Supervising physician: JF - Ferrara, Jessica | ?

Goal type: None Goal Objective Goal-library

Collaterals: 0 Status: Complete | Secondary services | Pre-fill manager

Narrative: Wiley | IMA-Write

Client arrived to discuss previously established goal:

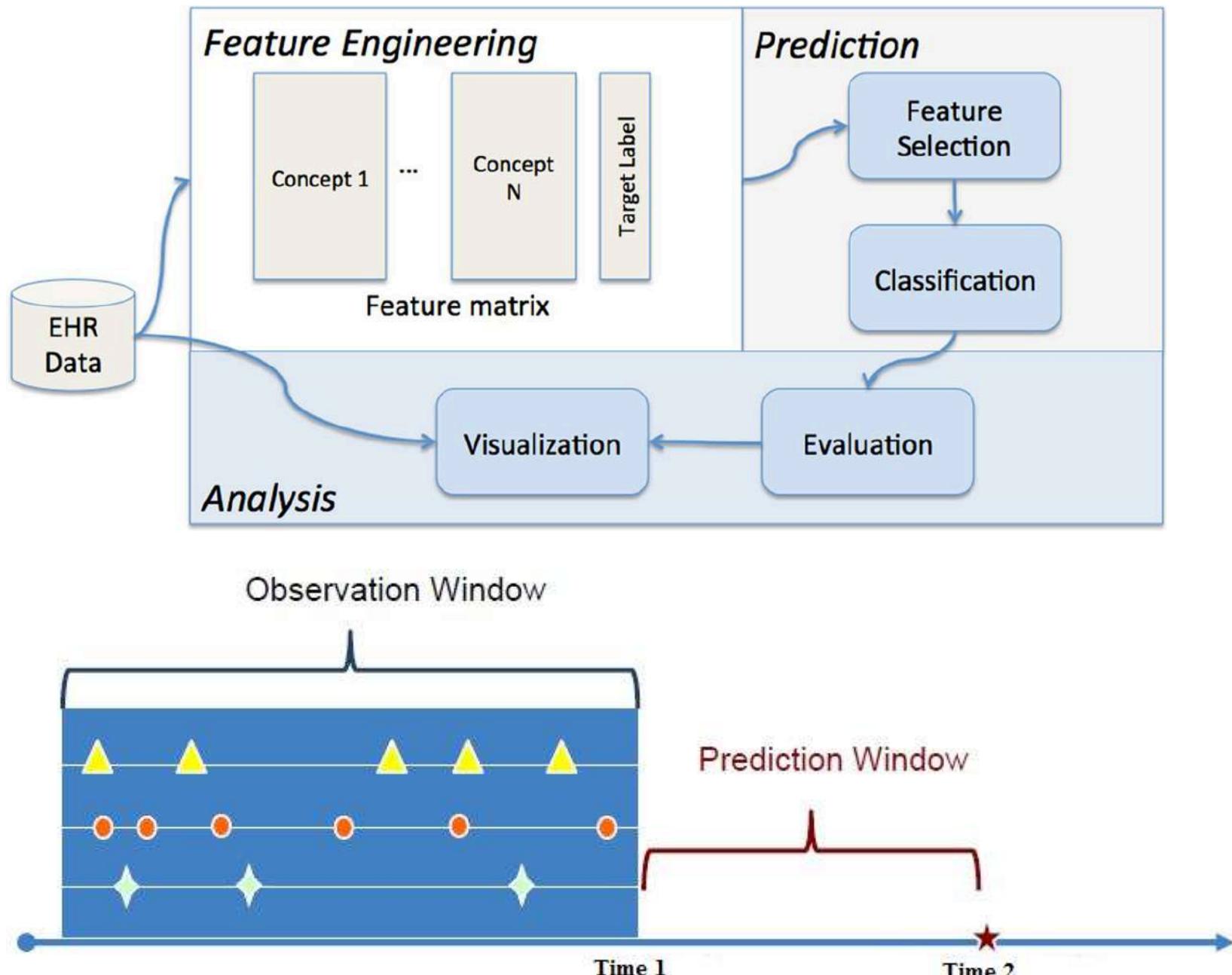
Reduce psychological energy and return to premorbid levels of activity, judgment, mood, and goal-directed behavior.

Elinor reported that her speech rate increases as she feels stressed.

Strengths and weakness of data classes within EHRs

	ICD codes	CPT codes	Laboratory Data	Medication records	Clinical Documentation
Availability in EHR systems	Near-universal	Near-universal	Near-universal	Variable	Variable
Recall	Medium	Poor	Medium	Inpatient: High Outpatient: Variable	Medium
Precision	Medium	High	High	Inpatient: High Outpatient: Variable	Medium-High
Fragmentation effect	Medium	High	Medium-High	Medium	Low-Medium
Query method	Structured	Structured	Mostly structured	Structured, text queries, and NLP	NLP, text queries, and rarely structured
Strengths	-Easy to query -Serves as a good first pass of disease status	-Easy to query -High precision	-Value depends on test -High data validity	Can have high validity	Best record of what providers thought
Weaknesses	-Disease codes often used for screening when disease not actually present -Accuracy hindered by billing realities and clinic workflow	-Most susceptible to missing data errors (e.g., performed at another hospital) -Procedure receipt influenced by patient and payer factors external to disease process	-May need to aggregate different variations of the same data elements -Normal ranges and units may change over time	-Often need to interface inpatient and outpatient records -Medication records from outside providers not present -Medications prescribed not necessarily taken	-Difficult to process automatically -Interpretation accuracy depends on assessment method -May suffer from significant "cut and paste" -Not universally available in EHRs -May be self-contradictory
Summary	Essential first element for electronic phenotyping	Helpful addition if relevant	Helpful addition if relevant	Useful for confirmation and a marker of severity	Useful for confirming common diagnoses or for finding rare ones

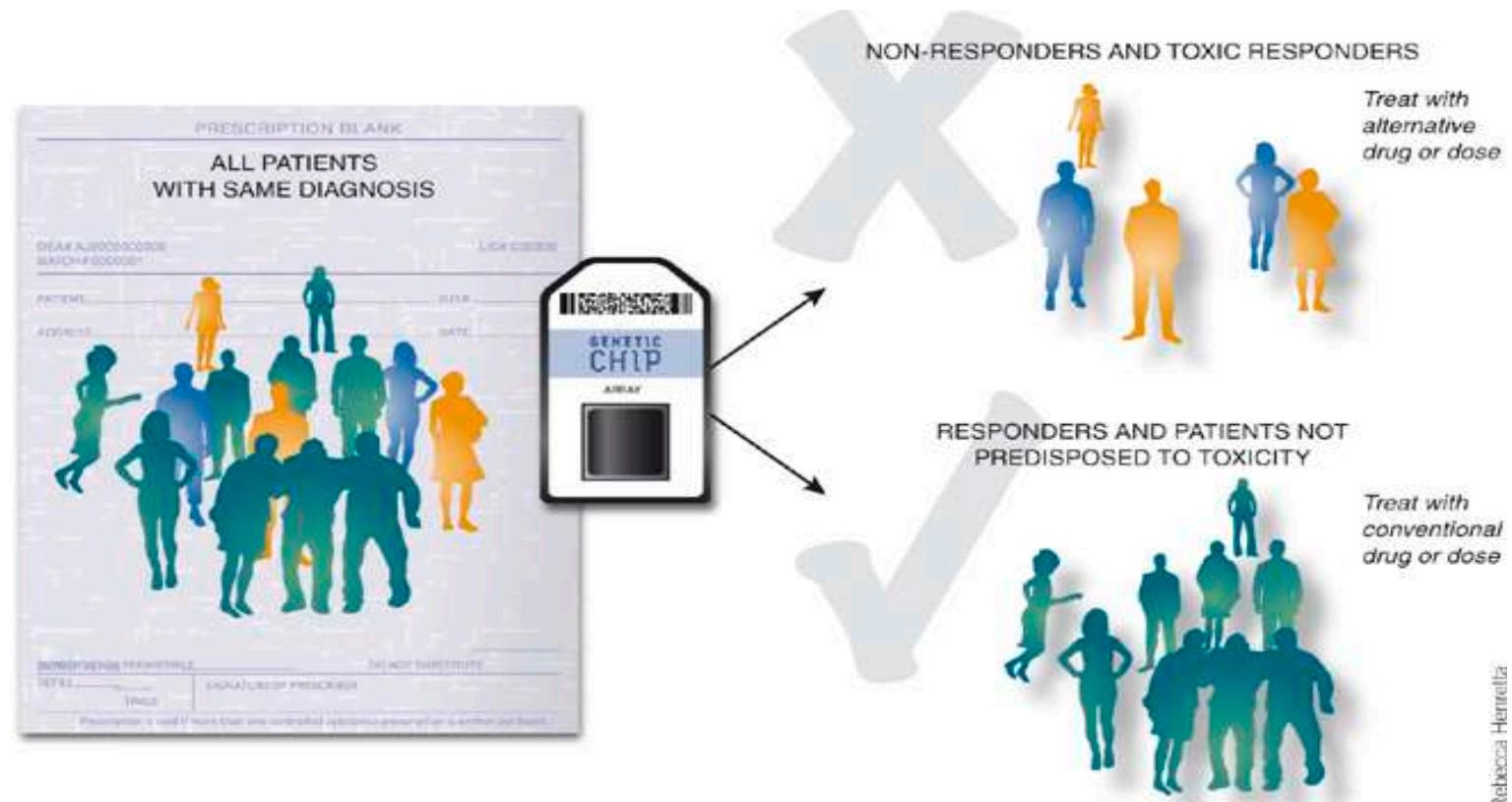
Application 1: Predictive Modeling Pipeline



More details and publications found at: <http://www.research.ibm.com/healthcare/>

Application 2: moving towards personalized medicine

- Personalized Medicine: the right patient with the right drug at the right dose at the right time.
 - (for patients) the end of one size fits all drugs would result in safer and more effective treatments
 - (for doctors) reduce wasted time for patients and resources with futile treatments
 - (for pharms) lower cost marketing due to targeted patients, faster clinical trials, less focus on animal trials



The figure is from <https://3dbiomatrix.com/a-new-dimension-in-personalized-medicine-2/>

Patient similarity and drug similarity analytics

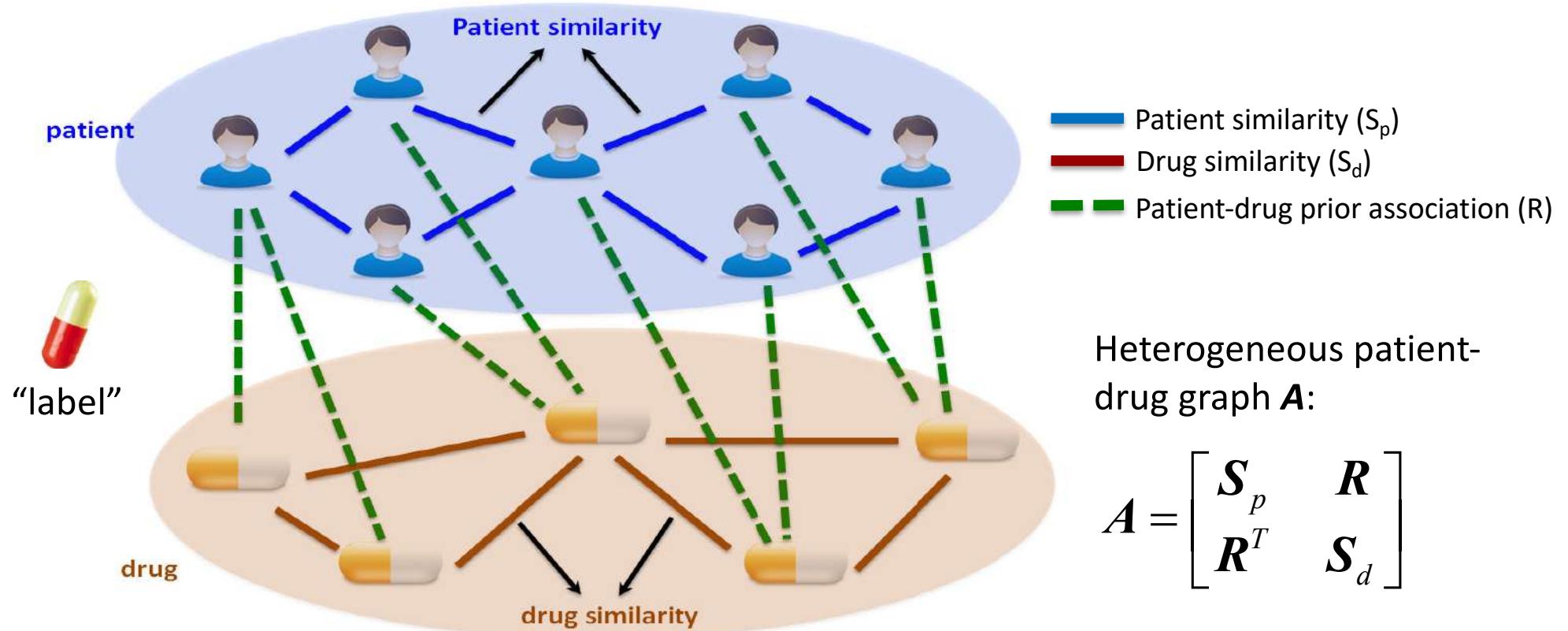


- Patient Similarity analytics: Find patients who display similar clinical characteristic to the patient of interest
- Drug Similarity analytics: Find drugs which display similar pharmacological characteristic to the drug of interest

How to leverage both patient similarity and drug similarity for personalized medicine?

Heterogeneous graph for drug personalization

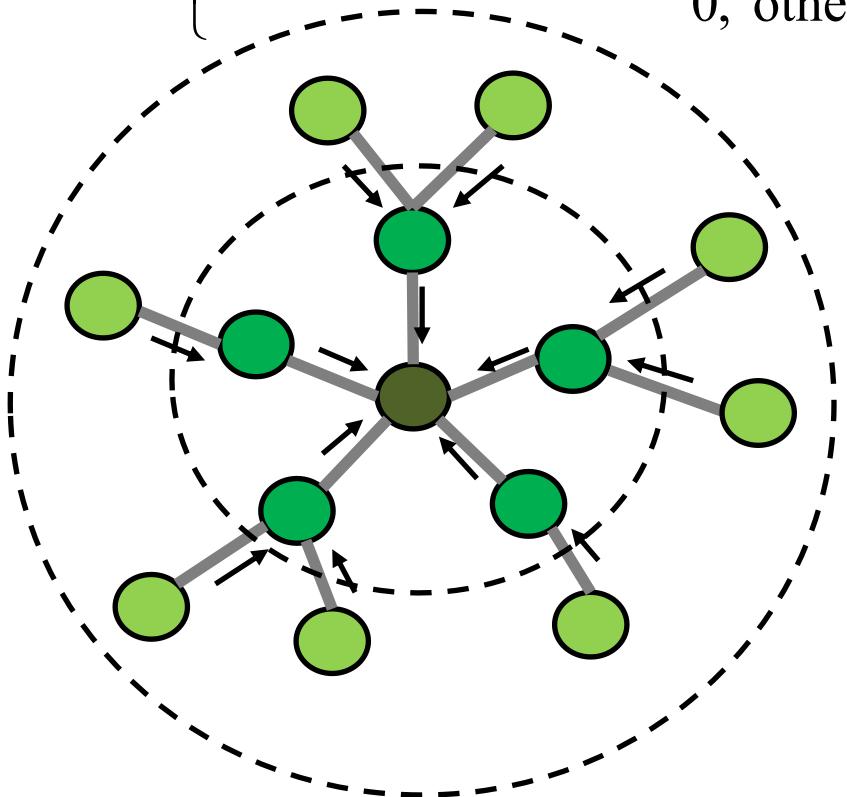
- Drug personalization problem: whether drug A is likely to be effective for specific patient B. To take into consideration the specific condition of patient B as well as the characteristics of drug A, we should leverage the information of:
 - The patients who are clinically similar to patient B
 - The drugs which are similar to drug A
 - Prior associations between patients and drugs, which are measured by diagnosis of patients and therapeutic indications of drugs



Label propagation method

- For each drug d , we constructed a corresponding effectiveness vector (i.e., ***known but not completed*** “label” vector) $\mathbf{y} = [y_1, y_2, \dots, y_n, y_{n+1}, \dots, y_{n+m}]^T$, where

$$y_k = \begin{cases} 1 & (k = 1, 2, \dots, n), \text{ if } d \text{ is an effective treatment for patient } k \\ 1 & (k = n+1, n+2, \dots, n+m), \text{ if } d \text{ is the } (k-n)\text{-th drug} \\ 0 & \text{otherwise} \end{cases}$$



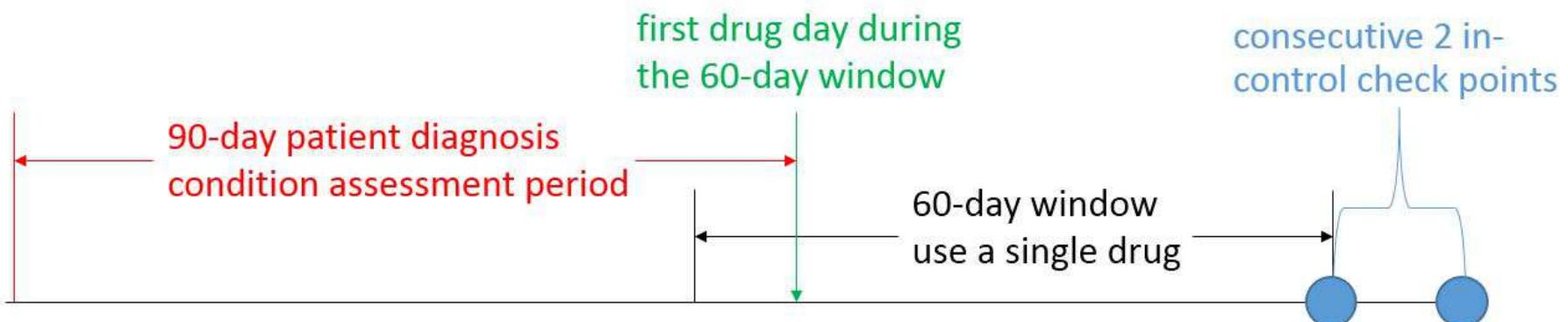
- \mathbf{W} is a normalized form of the similarity matrix \mathbf{A} .
- In each propagation iteration, the estimated score of each node “absorbs” a portion (μ) of the label information from its neighborhood, and retains a portion ($1 - \mu$) of its initial label information.
- The updating rule for node i is given by

$$f_i^{after} \leftarrow \mu \sum_{j=1}^n \mathbf{W}_{ij} f_j^{before} + (1 - \mu) y_i$$

Consider the initial condition is $f^0 = \mathbf{y}$, we have the equation $f^t = (\mu \mathbf{W})^{t-1} \mathbf{y} + (1 - \mu) \sum_{i=0}^{t-1} (\mu \mathbf{W})^i \mathbf{y}$

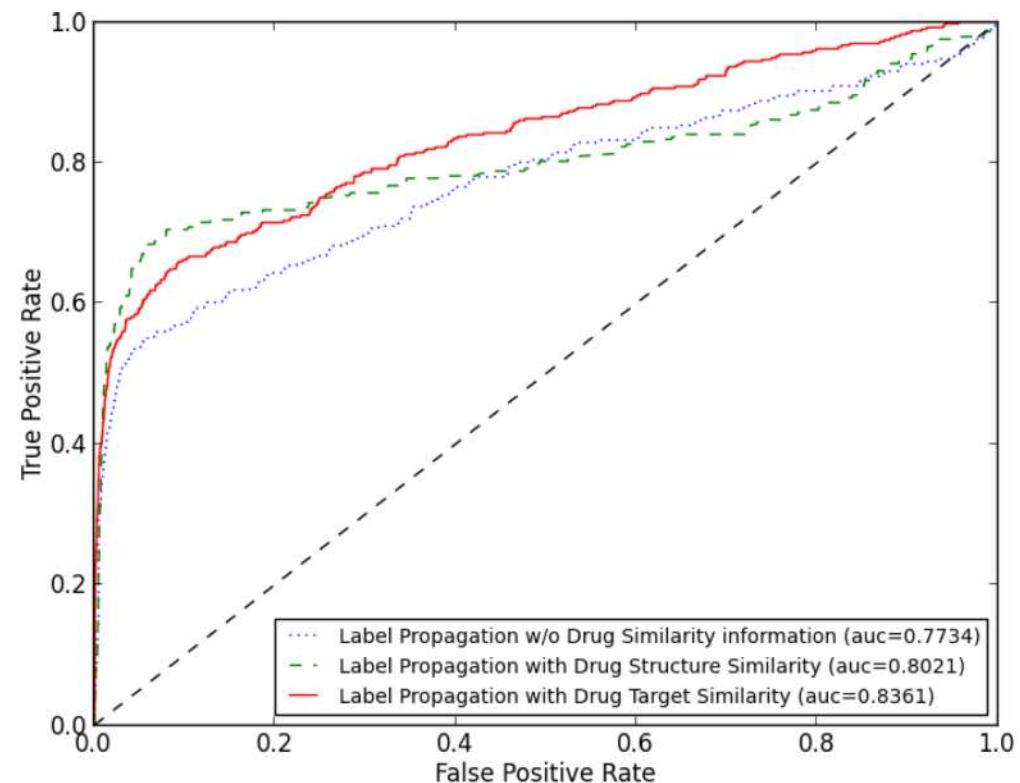
⇒ $f^* = \lim_{t \rightarrow \infty} f^t = (1 - \mu)(\mathbf{I} - \mu \mathbf{W})^{-1} \mathbf{y}$ *f – the possibility when a drug is effective for a patient*

Experimental results of personalized treatments for hyperlipidemia



Data: 1219 distinct patients and 4 statin cholesterol-lowering drugs from a real-world EHR

Drug	Patient #
Atorvastatin	97
Lovastatin	221
Pravastatin	24
Simvastatin	877



Adverse drug reactions (ADRs)

- Post-approval ADRs remain a significant source of mortality and morbidity around the world
 - 2 million potentially preventable injuries, hospitalizations, and deaths each year in US alone
 - Associated cost estimated at \$75 billion annually

The New York Times

F.D.A. Issues New Alerts About Cholesterol Drugs

By GARDINER HARRIS

Published: February 29, 2012

CORRECTION APPENDED

Federal health officials on Tuesday added new safety alerts to the prescribing information for statins, the cholesterol-reducing medications that are among the most widely prescribed drugs in the world, citing rare risks of memory loss, diabetes and muscle pain.



Statins are considered some of the safest drugs

Merck Pulls Arthritis Drug Vioxx from Market

by RICHARD KNOX

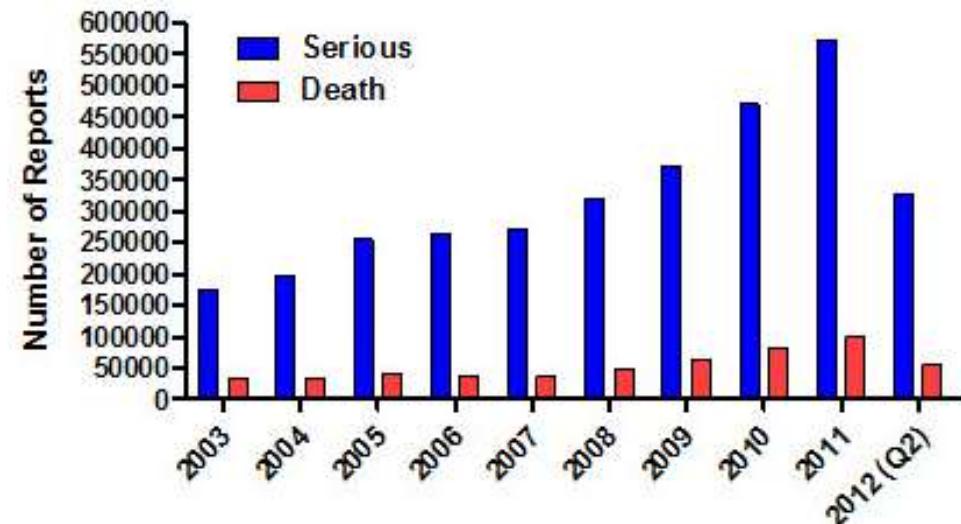
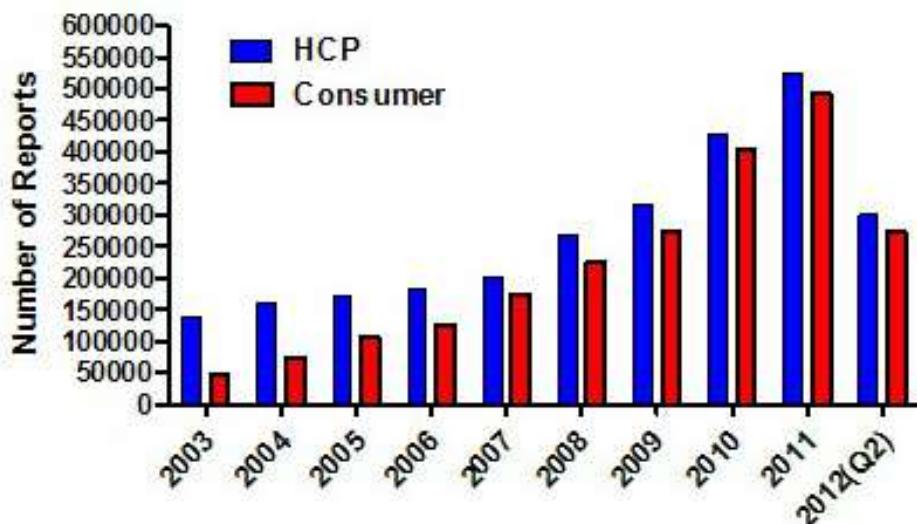
September 30, 2004 12:00 AM ET

Pharmaceutical giant Merck & Co. is pulling its arthritis drug Vioxx from the market after a study confirmed earlier concerns that it raises the risk of heart problems, including heart attacks and stroke. Vioxx is currently used by 2 million people worldwide and has been used by more than 84 million people around the world, according to Merck.

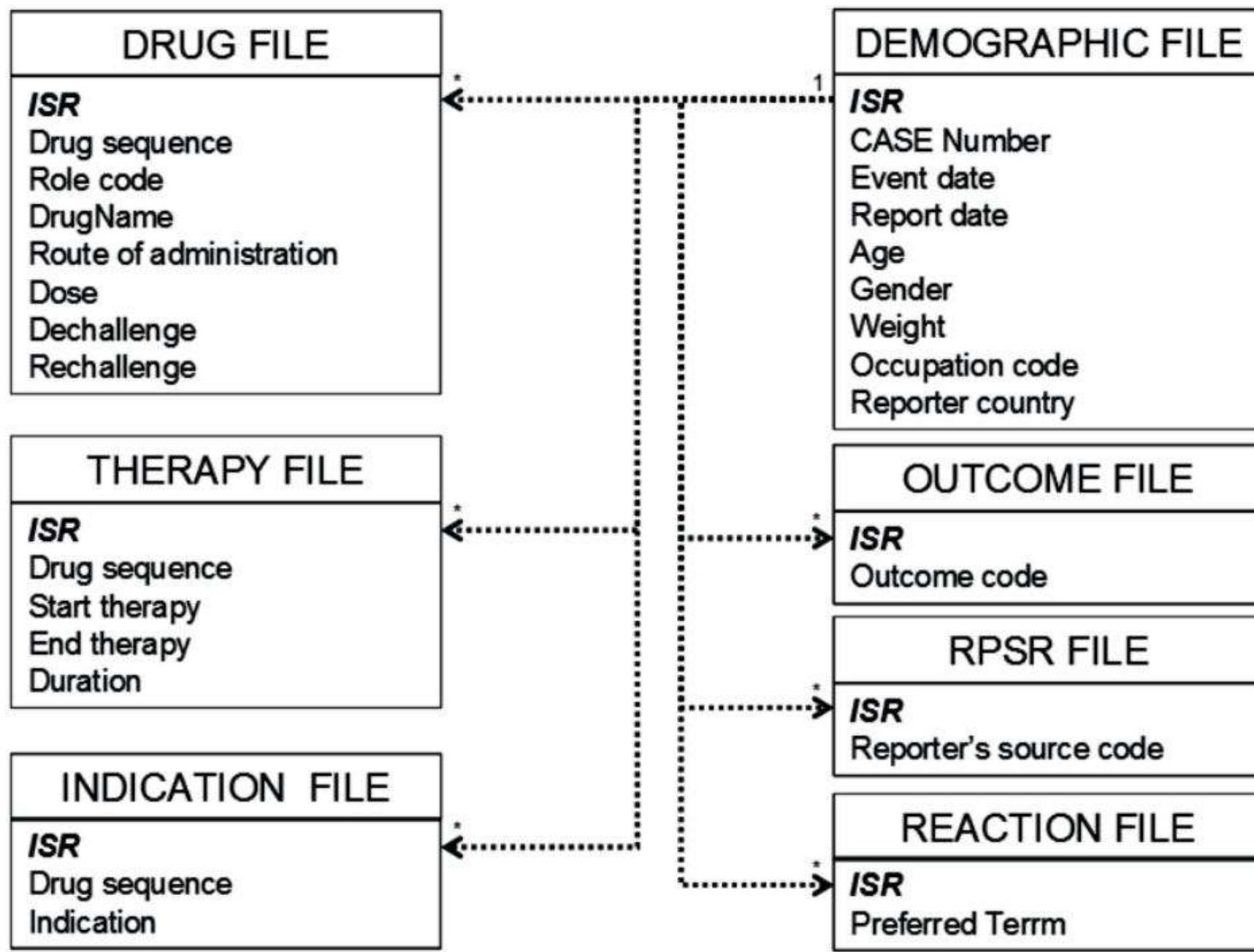
- More than 140,000 cases of serious heart disease
- \$4.85 billion for legal claims from US citizens

FDA Adverse Event Reporting System (AERS)

- FDA Adverse Event Reporting System (AERS)
 - FDA has maintained AERS since 1968
 - Spontaneous reports of suspected ADRs collected from healthcare professionals, consumers, and pharms
 - Data (from Jan 2004 to Apr 2013) is publicly available at FDA's website!
- Over 5 million reports collected so far:
 - patient: age, sex, weight, country
 - drugs they are taking
 - diseases they were being treated for
 - the adverse events that occurred to that patient

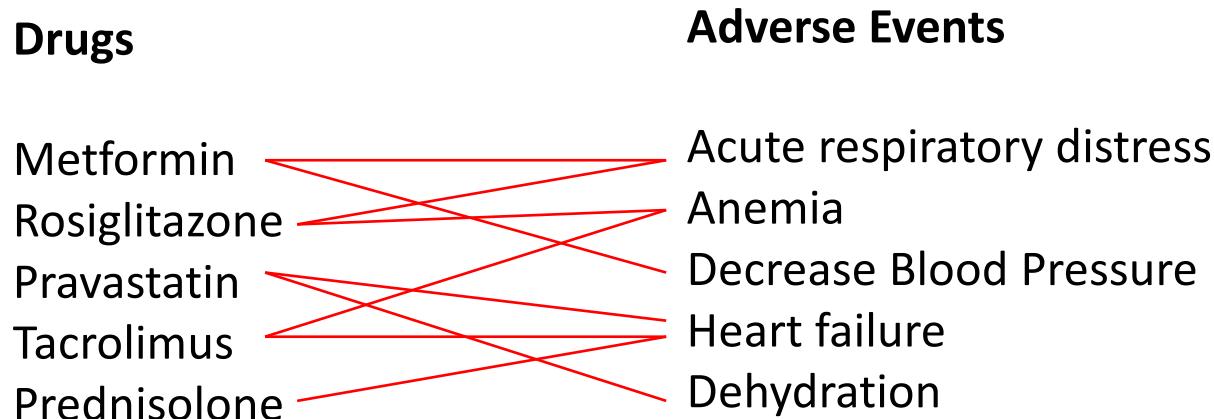


FDA AERS database structure



Interpreting those AERS reports is hard

- Many drugs, many adverse events
 - what causes what?
 - Most of these red lines are false - which are true?



- Data mining (signal detection) algorithms for AERS
 - Quantify “unexpectedness”: to identify drugs that have a greater proportion of a particular event compared to the proportion seen for other drugs
 - Sampling variance
 - Underreporting
 - Over reporting
 - Selection biases
 - Causative covariates other than drug under analysis

Disproportionality analysis

	reports w ae	reports w/o ae	Total
reports w drug	a	b	a+b
reports w/o drug	c	d	c+d
Total	a+c	b+d	a+b+c+d

Measure of association	Formula	Probabilistic interpretation
Relative reporting (RR) ¹	$\frac{a(a+b+c+d)}{(a+c)(a+b)}$	$\frac{\Pr(\text{ae} \mid \text{drug})}{\Pr(\text{ae})}$
Proportional reporting rate ratio (PRR)	$\frac{a(c+d)}{c(a+b)}$	$\frac{\Pr(\text{ae} \mid \text{drug})}{\Pr(\text{ae} \mid \sim \text{drug})}$
Reporting odds ratio (ROR)	$\frac{ad}{cb}$	$\frac{\Pr(\text{ae} \mid \text{drug}) \Pr(\sim \text{ae} \mid \sim \text{drug})}{\Pr(\sim \text{ae} \mid \text{drug}) \Pr(\text{ae} \mid \sim \text{drug})}$
Information component (IC) ²	$\log_2 \frac{a(a+b+c+d)}{(a+c)(a+d)}$	$\log_2 \frac{\Pr(\text{ae} \mid \text{drug})}{\Pr(\text{ae})}$

1. The RR, when implemented within an empirical Bayesian framework, is known as empirical Bayes geometric mean (EBGM); 2. The IC is a logarithmic RR metric that is implemented in a Bayesian framework.

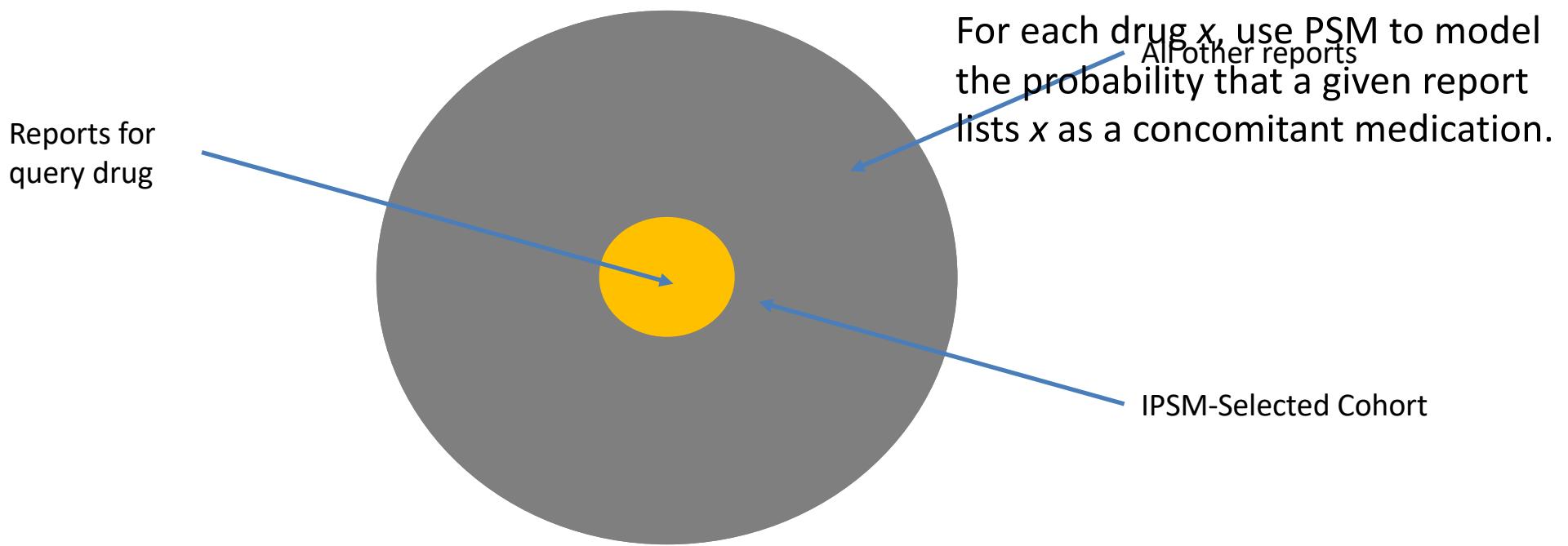
- Modern signal detection algorithms (e.g., EBGM, IC) could address sampling variance
 - Estimate confidence intervals (CIs) for disproportionality statistics
 - Dampen drug-event signals that have little evidence to support them
- How to address selection biases?

Selection biases in AERS reports

- Selection biases introduce “synthetic associations”
 - (e.g.) from concomitant drug use (co-Rx effect)
 - drugs co-prescribed with **Vioxx** more likely to be associated with **heart attacks**
 - (e.g.) from indications (indication effect)
 - drugs given to **diabetics** more likely to be associated with **hyperglycemia**
 - (e.g.) co-Rx effect and indication effect extend to other covariates
 - Patients reported to be taking a cholesterol-lowering agent are more likely to be older, and this may cause these drugs to be synthetically associated with age-related effects, such as hypertension or myocardial infarction (age bias).
- Propensity score matching (PSM) corrects for bias of MEASURED covariates
 - Identify matched controls for the studied cases in observational clinical studies
 - Model the likelihood of a case being selected based on the covariates
 - $PS = \text{Estimated } Pr(\text{Exposed}+ | \text{covariates}) \sim \text{age} + \text{sex} + \text{weight} + \dots$
 - Match each case with one or more controls with the same likelihood
 - **However, PSM requires the covariates to be both known and measured; neither parameter is guaranteed to be present in AERS**

Implicit Propensity Score Matching (IPSM)

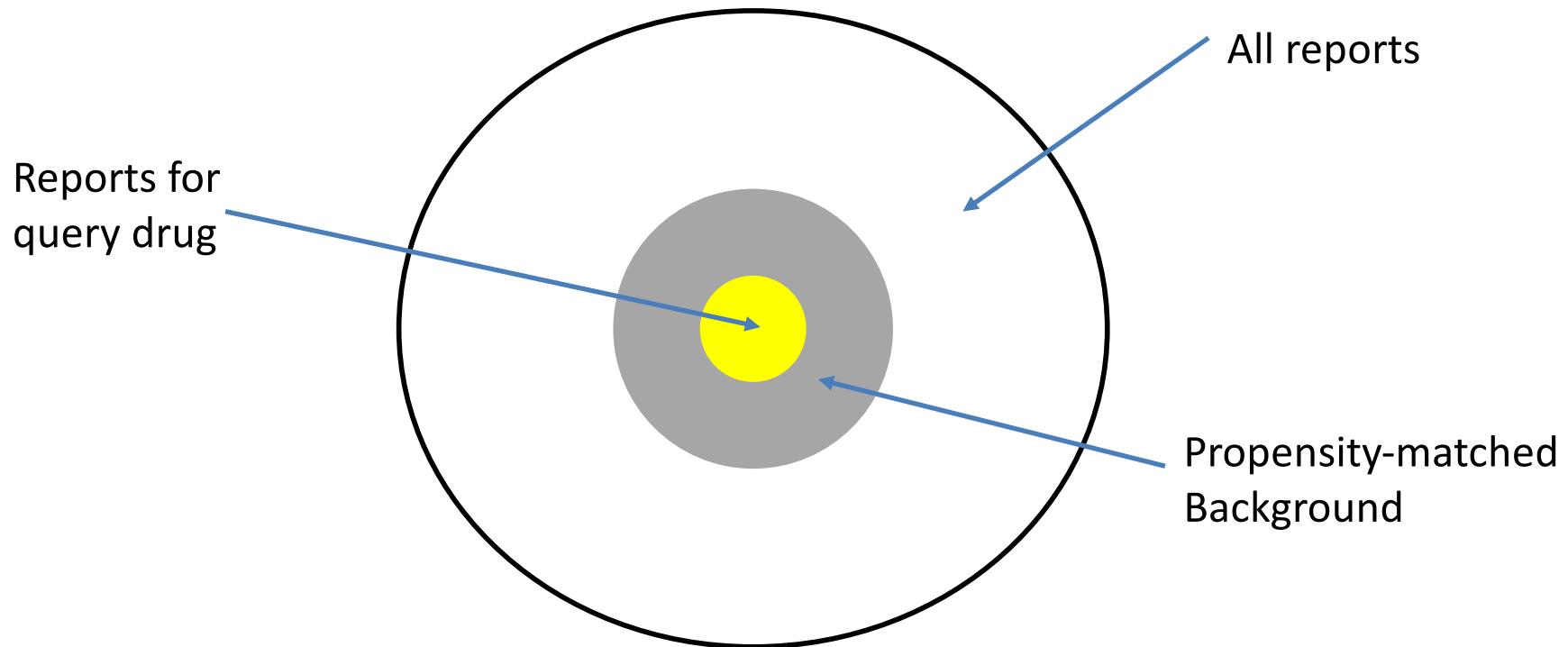
- Invented by Tatonetti NP et al. *Sci Transl Med.* 2012;4(125):125ra31.
- Assumes combination of co-reported drugs and co-indications describes all patient covariates. Hypothesize many confounders correlate with these key variables and do not need to be modeled.



- First, reduce to only those reports that have co-prescribed prescriptions listed
- Second, reduce to only those reports that have correlated indications listed

Takes advantage of co-Rx and indication variables likely to co-vary with unmeasured covariates

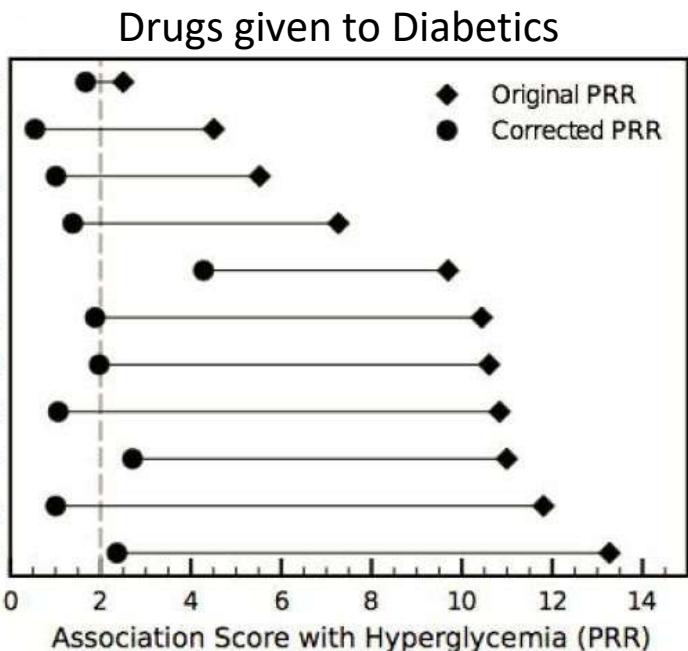
IPSM produces better estimates of expected values



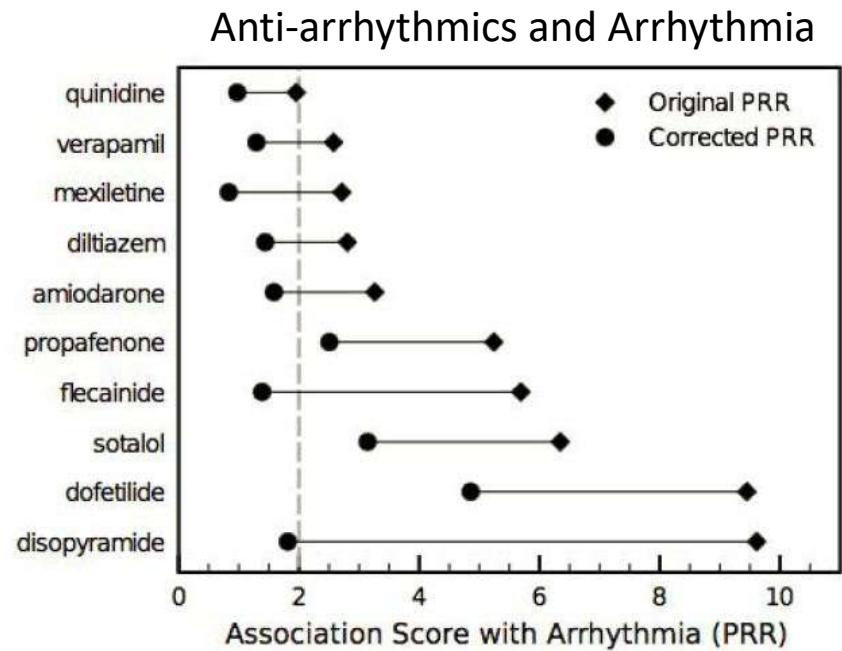
- Example: Reporting of **hyperglycemia** with **diabetes drugs**
- **Observed** reporting frequency: 17.7%
- **Expected** Estimates:
 - Entire database expected frequency: 1.5%
 - $\text{PRR} = 17.7\% / 1.5\% = 11.8!!!!$
 - IPSM-derived expected frequency: 17.6%
 - $\text{PRR} = 17.7\% / 17.6\% = 1.0 \dots$

IPSM corrects for indication and co-Rx biases

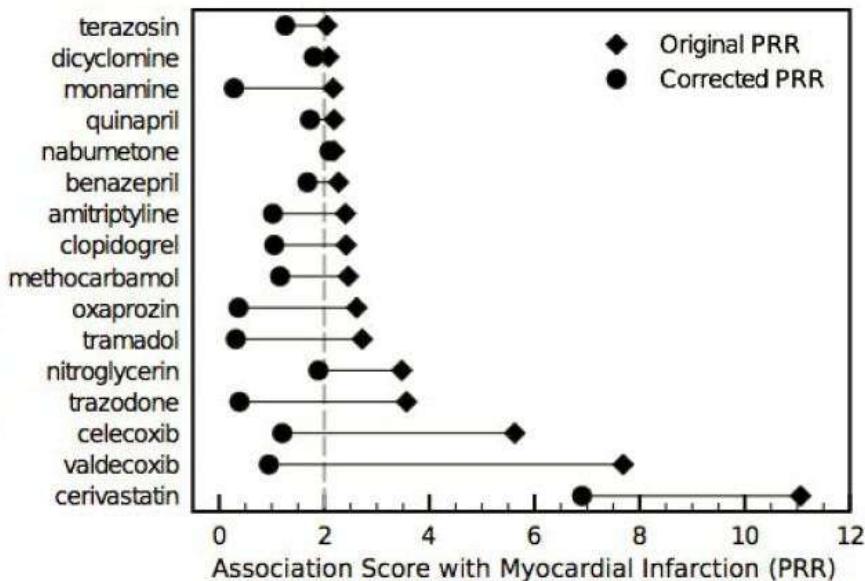
Drugs coreported with Diabetes



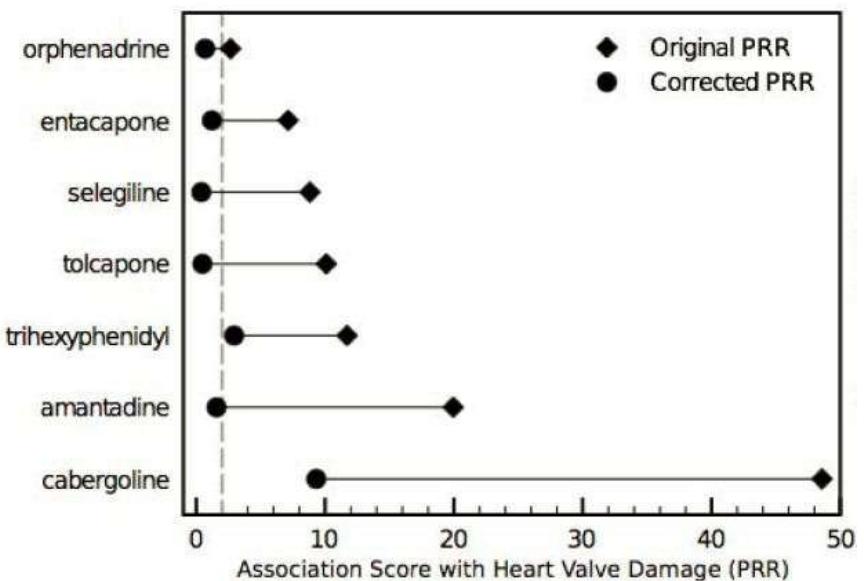
Drugs coreported with Arrhythmia



Drugs coreported with Rofecoxib (Vioxx)

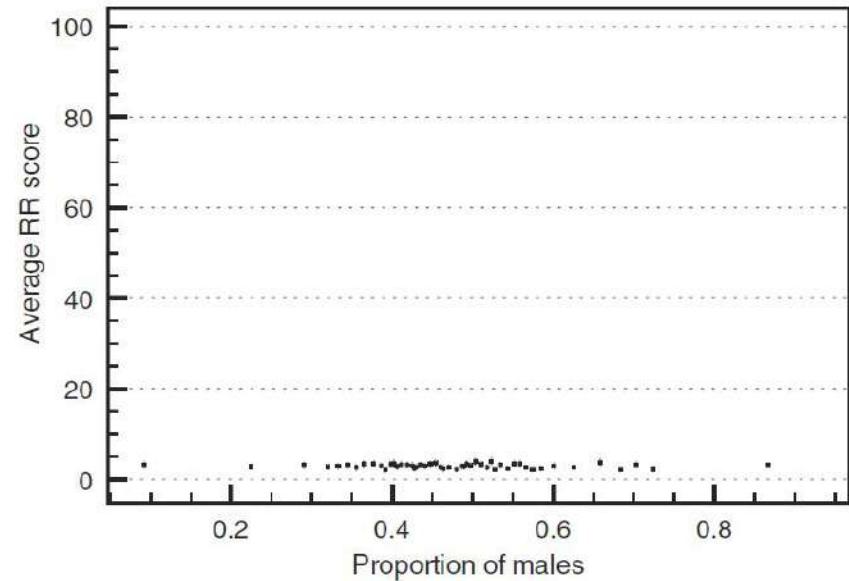
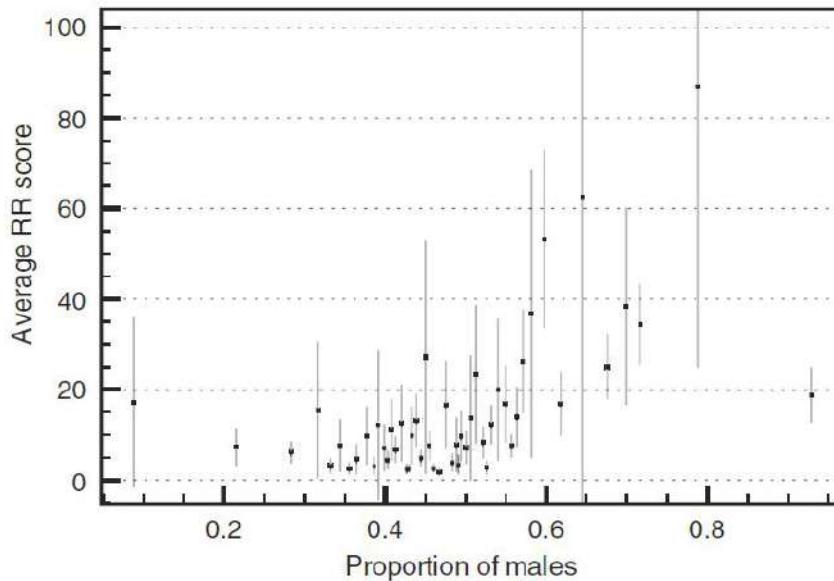


Drugs coreported with Pergolide

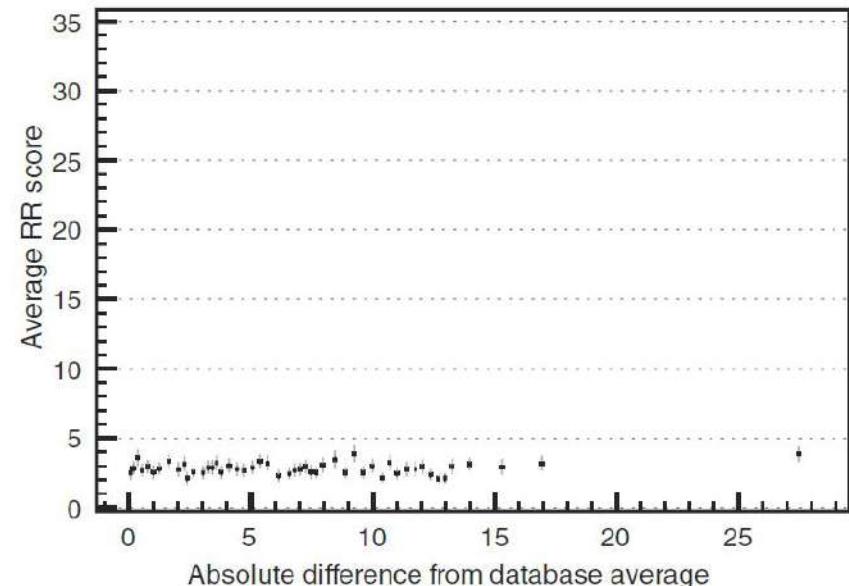
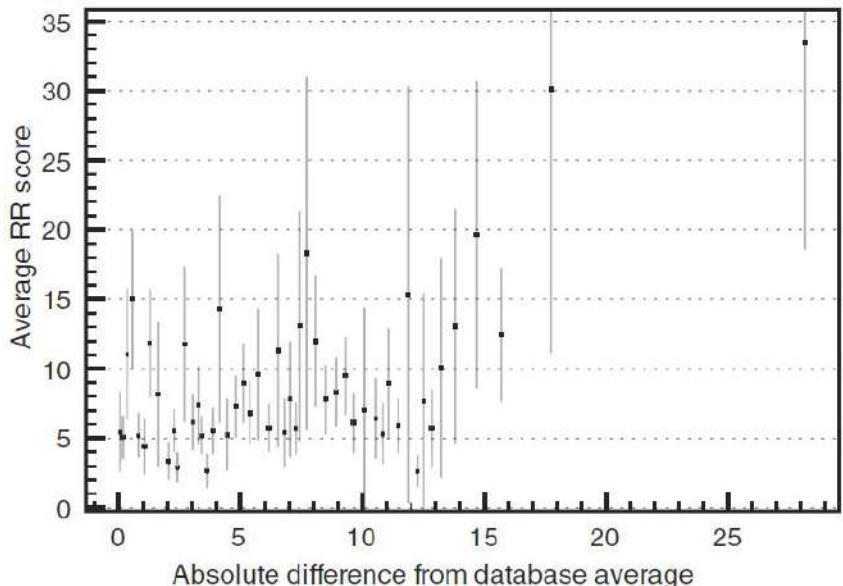


IPSM implicit correction for other biases

Drugs preferentially with males are more likely to be associated with 33 sex-related (male) effects

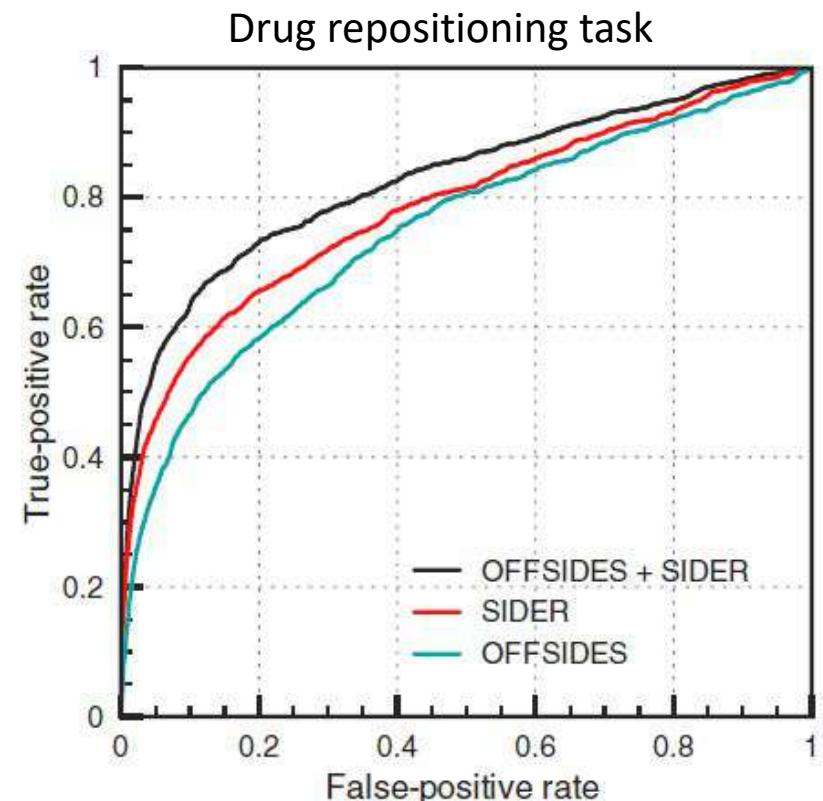
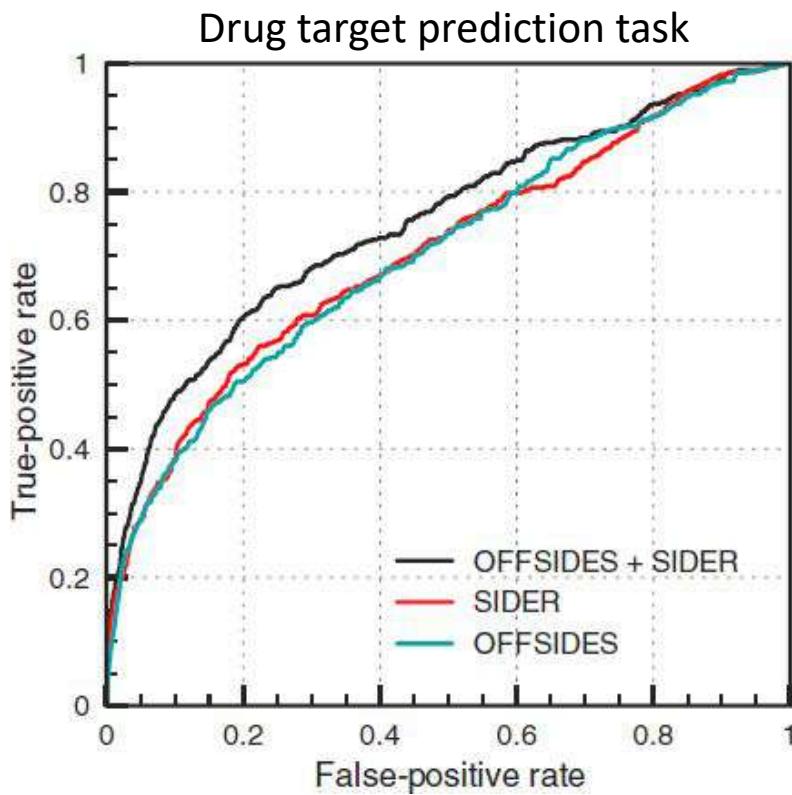


Drugs preferentially with young/old patients are more likely to be associated with 48 age-related effects



The usage of Offsides and Twosides

- IPSM corrects for biases introduced by hidden covariates
- EBGM addresses sampling variance
- Two comprehensive databases
 - Offsides: drug-AE
 - Twosides: drug1-drug2-AE



Challenges impacting real world evidence research



Outline

- Introduction of Drug Discovery and Development
- Motivation of Data Mining
- Case Study: Drug Repositioning
- Case Study: Real-World Evidence
- **Data Sources for Data Mining Applications**
- Challenges and Summary

Examples of preclinical data sources

Name	Sponsor	Description
Chemical resources		
DrugBank	U of Alberta	drug data, drug target, and drug action information
PubChem	NCBI	chemical molecules and their activities against biological assays
Genomic/Proteomic resources		
GenBank	NCBI	annotated, publicly available DNA sequences
Gene Expression Omnibus (GEO)	NCBI	publicly available gene expression profiles
Proteomics IDEntifications (PRIDE) database	EBI	publicly available proteomics data
Connectivity Map (CMap)	Broad Institute	genome-wide transcriptional expression data from cultured human cells treated with bioactive small molecules

- Pro: easy to access; high quality
- Con: translational issue

Examples of clinical data sources (1): from clinical trials to RWE

Name	Sponsor	Description
Clinical trial resources		
ClinicalTrials.gov	NIH	federally and privately supported clinical trials; provides details such as the purpose and summary results of a trial
Trialtrove	Citeline	comprehensive real-time source of pharmaceutical clinical trials (over 30,000 clinical trial data sources from more than 150 countries)
Health record resources		
STRIDE Clinical Data Warehouse	Stanford School of Medicine	1.53 million pediatric and adult patients from 1994 to now at SUMC,
National Patient Care Database (NPCD)	Veterans Health Administration	inpatient and outpatient services provided to 4 million VHA healthcare users in the USA
General Practice Research Database (GPRD)	UK Medicines Control Agency	longitudinal medical records of 5 million active patients captured from primary care provided in UK
Electronic Medical Records and Genomics (eMERGE)	NHGRI	combines DNA biorepositories with electronic medical record (EMR) systems for large-scale, high-throughput genetic research
PatientsLikeMe	PatientsLikeMe	a social-networking health site enabling members to share symptom and treatment information

- Pro: direct observation from patients
- Con: dirty; inconsistent; privacy and ethical consideration

Examples of clinical data sources (2): safety data

Name	Sponsor	Description
Safety data resources		
SIDER	EMBL	marketed medicines and their recorded adverse drug reactions extracted from package inserts capturing information collected from the post-marketing safety surveillance program for all approved drugs
Adverse Event Reporting System (AERS)	FDA	drug-effect associations/drug-drug-effect mined from the FAERS not listed on the drug package inserts
Offsides/Twosides	Columbia U	
Vaccine Adverse Event Reporting System (VAERS)	FDA/CDC	captures the reporting of adverse events following immunization
Drug Interaction DataBase (DIDB)	U of Washington	human drug interactions extracted from sources such as PubMed, NDA, and FDA
SuperToxic	Charite University	a database of toxic compounds extracted from literature and web sources that provides details of possible biological interactions

Outline

- Introduction of Drug Discovery and Development
- Motivation of Data Mining
- Case Study: Drug Repositioning
- Case Study: Real-World Evidence
- Data Sources for Data Mining Applications
- Challenges and Summary

Challenges of all

- Lack of Gold Standard for data-mining
 - Drug-disease relationships
 - Drug-side effect relationships
- Machine learning models usually not easy to explain to clinicians and biologists
- ‘Translation’ among disease names across different ontologies
 - MeSH
 - ATC Code
 - SNOMED-CT
 - MeDRA
 - ICD-9

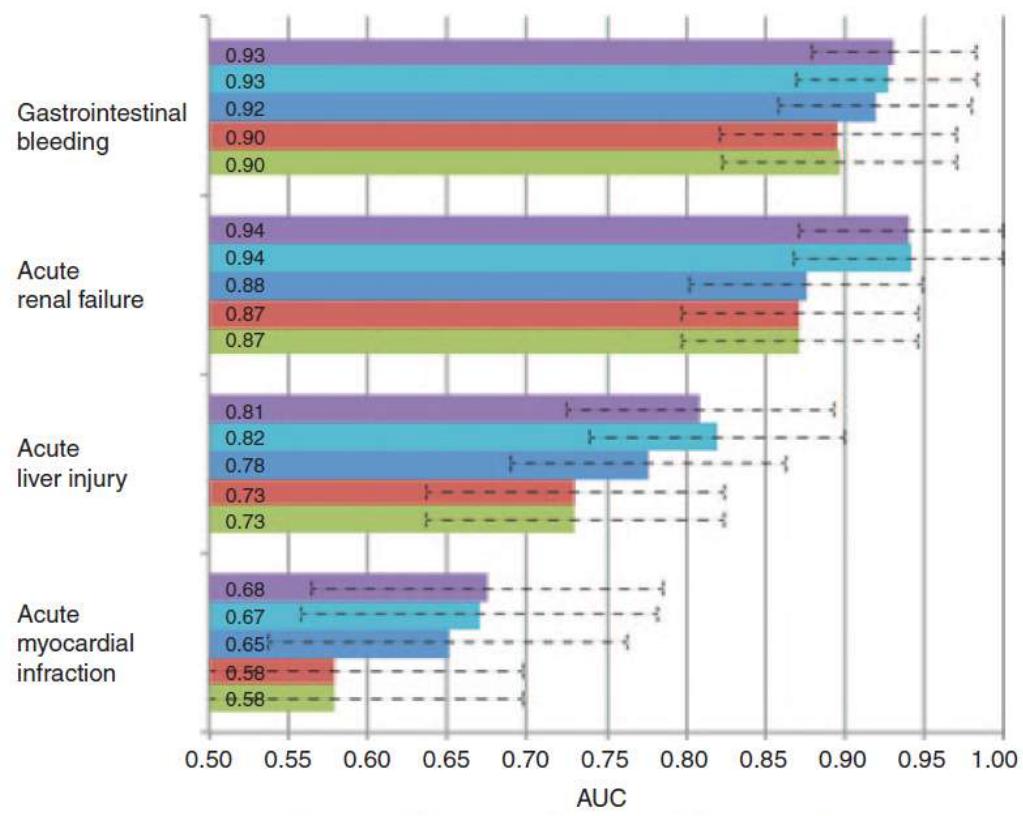
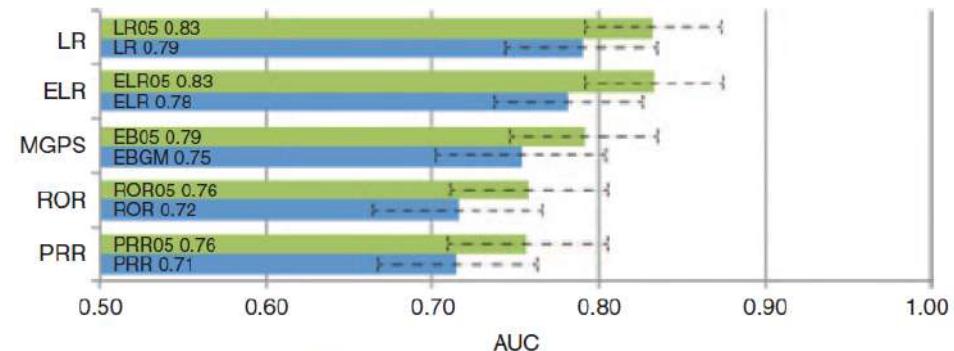
The application of biomedical gold standards - recent work

Positive Drug Set for a Side Effect:

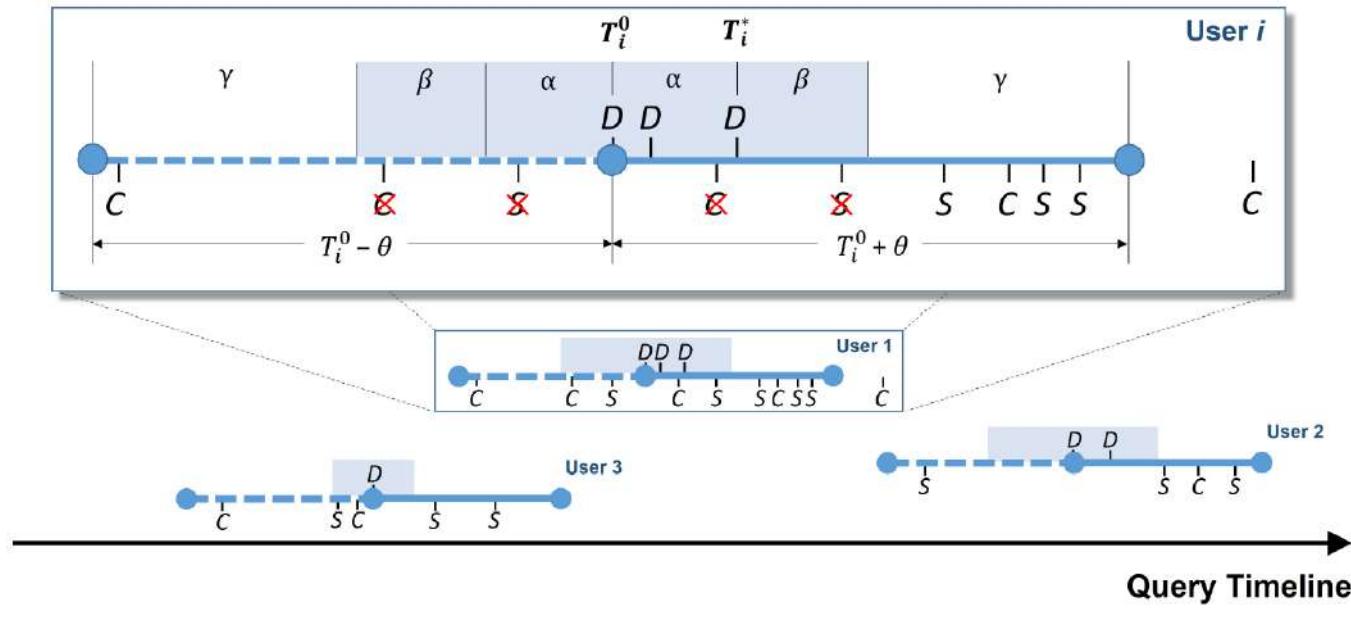
- Event listed in Boxed Warning or Warnings/Precautions section of active FDA structured product label
- Drug listed as ‘causative agent’ in Tisdale et al, 2010: “Drug-Induced Diseases”[35]
- Literature review identified no powered studies with refuting evidence of effect

Negative Set:

- Event not listed anywhere in any section of active FDA structured product label
- Drug not listed as ‘causative agent’ in Tisdale et al, 2010: “Drug-Induced Diseases”[35]
- Literature review identified no powered studies with evidence of potential positive association



Side effect detection based on search engine logs



D : query for drug of interest

C : query for condition of interest

S : query for a symptom of C

X : ignored C or S

$$\alpha = T_i^* - T_i^0$$

$$\beta = 7 \text{ days}$$

$$\gamma = 60 \text{ days}$$

$$\theta = (\alpha + \beta + \gamma)$$

surveillance window post T_i^0

surveillance window pre T_i^0

exclusion period

$$[T_i^0 - (\alpha + \beta), T_i^0 + (\alpha + \beta)]$$

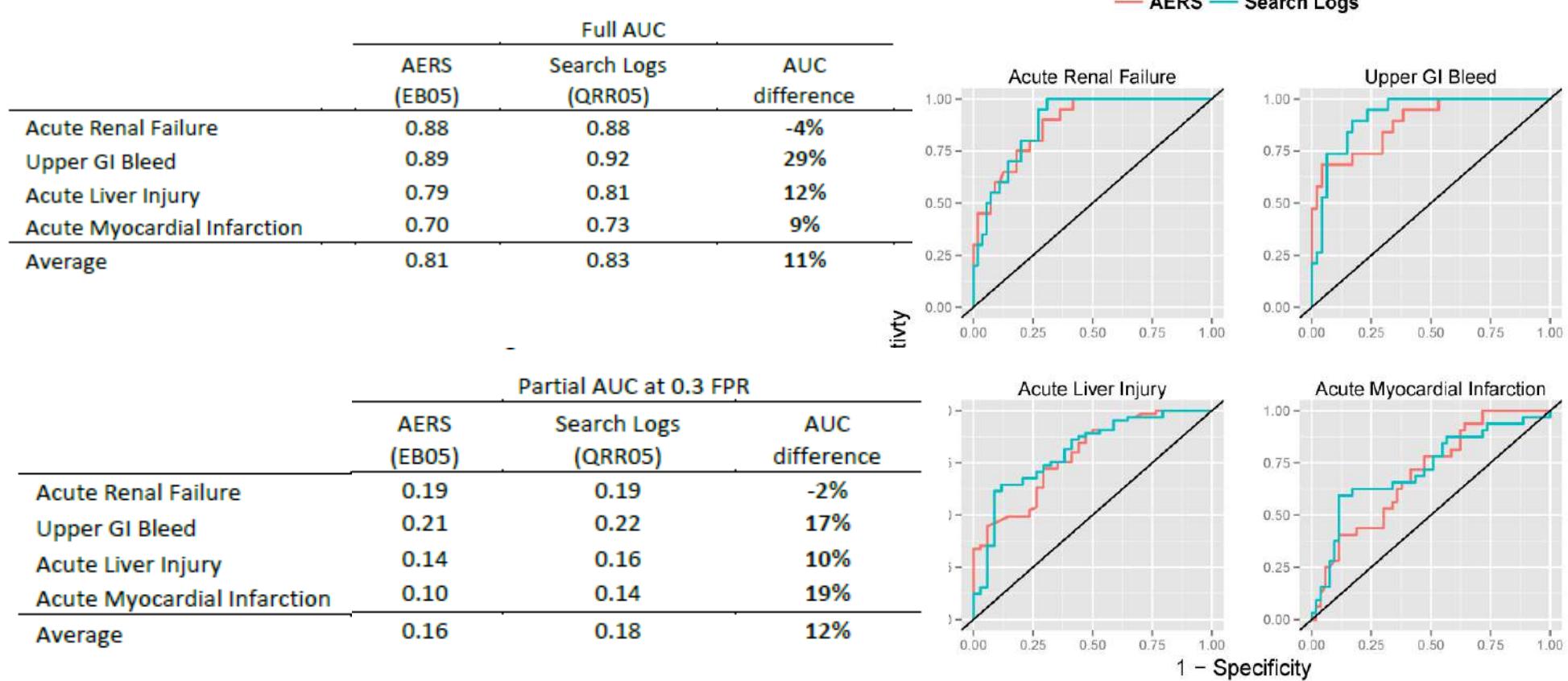
$$N_i^+ = \# \left\{ q_i^{(t)} \mid q_i^{(t)} \in C \cup S, T_i^0 + (\alpha + \beta) < t \leq T_i^0 + \theta \right\}$$

$$N_i^- = \# \left\{ q_i^{(t)} \mid q_i^{(t)} \in C \cup S, T_i^0 - \theta < t \leq T_i^0 - (\alpha + \beta) \right\}$$

$$QRR = \frac{\sum_i N_i^+}{\sum_i N_i^-}$$

$$\frac{2N^-N^+ + Z_{\alpha/2}^2(N^- + N^+) \pm \sqrt{Z_{\alpha/2}^2(N^- + N^+) (4N^-N^+ + Z_{\alpha/2}^2(N^- + N^+))}}{2(N^-)^2}$$

Comparison between AERS and search log based signal detection



Create a better life for human being

- In September, 2013, Larry Page announced his latest “moonshot,” a new venture to extend the human life span



Thank you! | Questions?



Ping Zhang: pzhang@us.ibm.com

Lun Yang: Lun.Yang@gmail.com