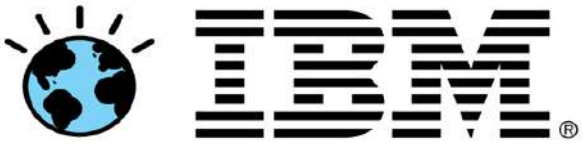


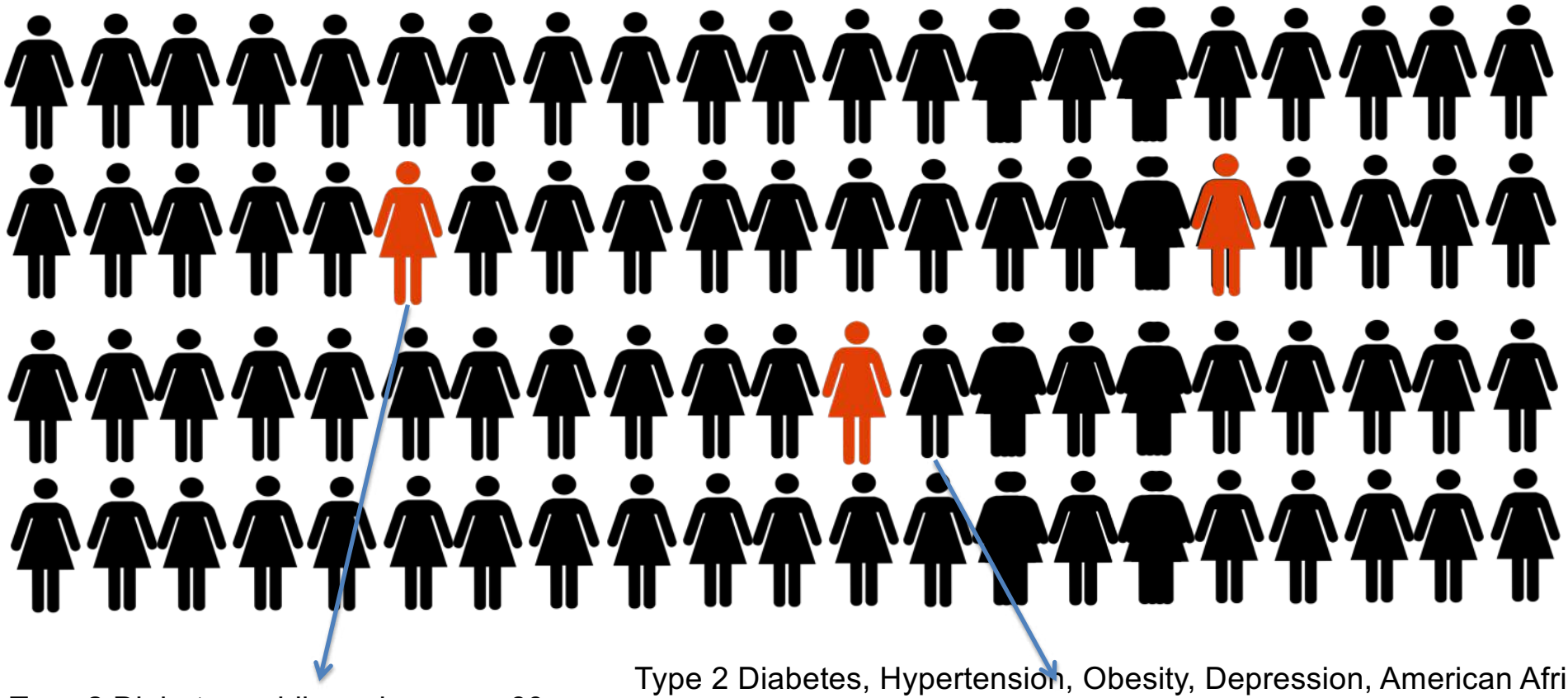
Towards Large-Scale Drug Safety Surveillance: A Big Data Perspective

Ying Li, PhD
liying@us.ibm.com

Ping Zhang, PhD
pzhang@us.ibm.com

Center for Computational Health
IBM T.J. Watson Research Center

Can Big Data Tell Us What Clinical Trials Don't?



Type 2 Diabetes, while male, age < 60

Type 2 Diabetes, Hypertension, Obesity, Depression, American African female, age > 70

Purpose of Post-marketing Safety Monitoring

- To learn about new risks
- To learn more about known risks
- To learn about medication errors
- To learn about how patterns of use may contribute to unsafe use

Historical Perspectives

- 1961 – 1962: Thalidomide tragedy
- If adequate post-market monitoring had been in place in Europe in the 1950's, it is believed that teratogenicity due to thalidomide would have been detected much earlier
- Post-marketing Adverse Event Reporting in USA
  - Begin in late 1950's after registration of cases of aplastic anemia due to chloramphenicol
  - Expanded in 1962 when industry was required to report adverse drug reactions to FDA
  - Since 1969 reports have been computerized
  - 1993 "MedWatch" expanded and facilitated reportings

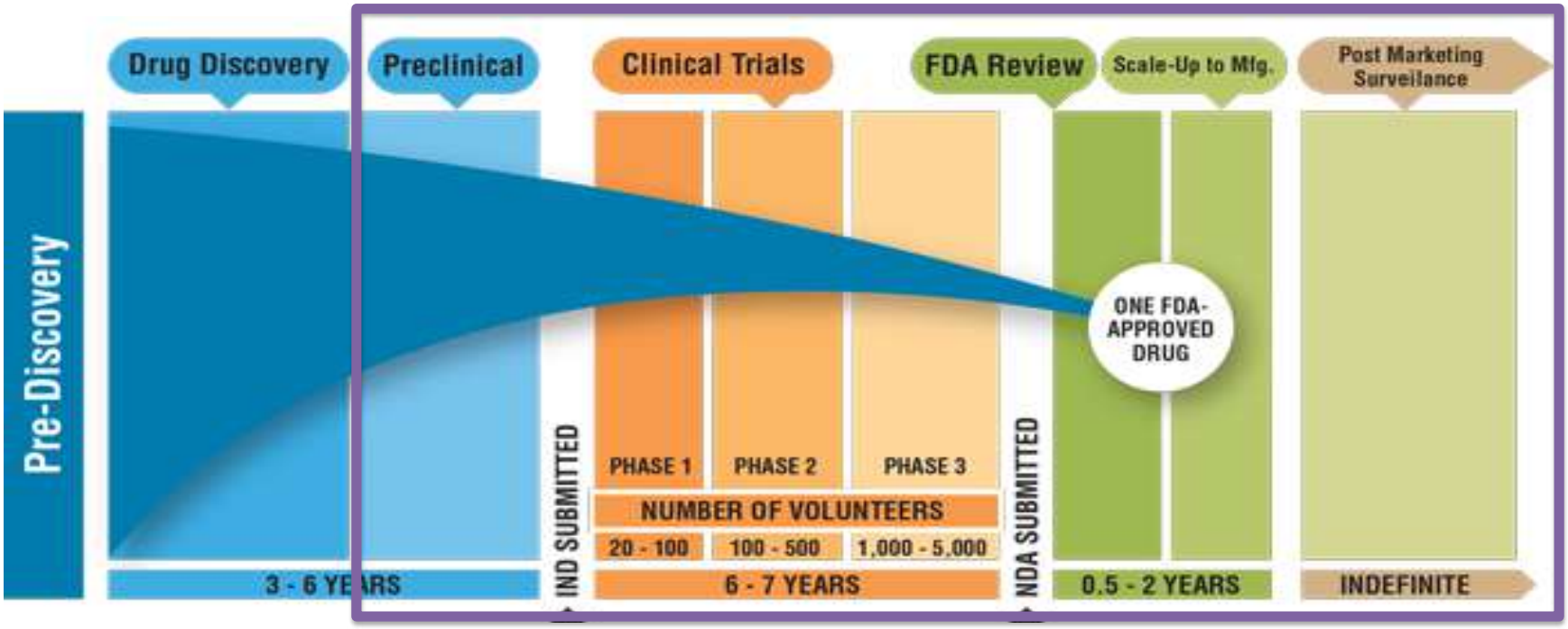
What is an adverse drug reaction?

- Adverse drug reaction (ADR) is a noxious and unintended response to a drug at normal doses during normal use (WHO)
  - Teratogenicity <- Thalidomide
  - Side effect == Adverse drug reaction == adverse event
- Public Health
  - 4th - 6th leading cause of death
  - > 10% of hospitalization
- Financial Burden
  - $5.6 billion annually

Classen DC 1997, Cullen DJ 1995, 1997;

Drug safety (pharmacovigilance) happens from the time a drug is discovered throughout it's approval and release to the market

- Side effects are collected during animal studies conducted during the "preclinical phase.[1]" Adverse events reported during clinical trials before FDA / EMA review help form the drug's label or approved claims.[2] Side effects reported after approval are collected in a process called "post marketing surveillance"



FDA web site. Animal study How FDA Evaluates Regulated Products: Drugs Wikipedia - Post Marketing Surveillance

Late discovery of safety signals during post marketing is a real challenge

AVASTIN® bevacizumab
Solution for intravenous infusion

Approved August, **2004**: Brain cancer, Colorectal cancer, Lung Cancer, etc.,
Warning added **2011**: Ovarian Failure

Approved August, **2001**: heart burn
Warning added **2016**: Kidney failure

ABILIFY (aripiprazole)
2, 5, 10, 15, 20, 30 mg Tablets

Approved August **2002**: Depression
Warning added **2016**:: Binge eating, shopping

LEVAQUIN

Approved **1996**: Pneumonia
Warning added May **2016**: Central Nervous system damage

onglyza® (saxagliptin) 5 mg tablets

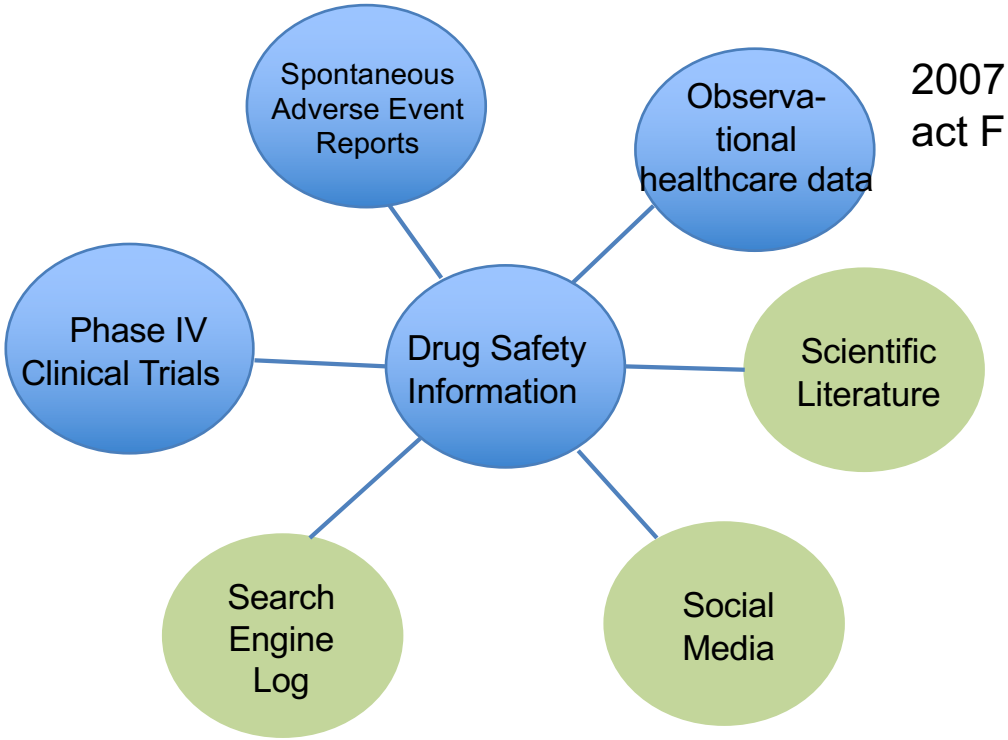
Approved August **2009**: Type II Diabetes
Warning added April **2016** Heart Failure

CHANTIX® (varenicline) TABLETS

Approved **2006**: smoking cessation
Warning added March, **2015**: alcohol interaction, Mood alterations, rare seizures

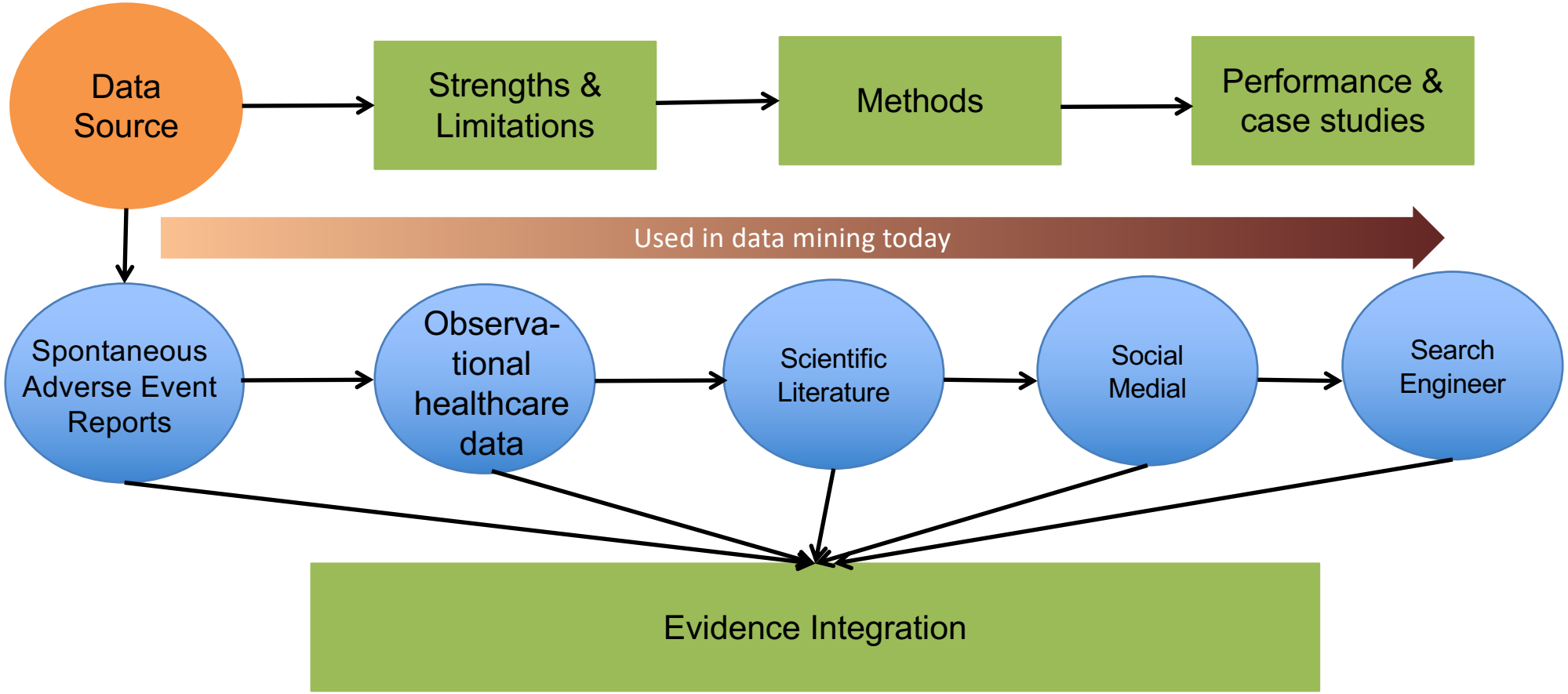
Abilitfy gets potential for binge eating; Astra and Merck Diabetes Drugs Get Warnings; PPIs get new warnings; Doctors didn't Know this common antibiotic was deadly; FDA issues warnings for Chantix

Data sources of drug safety information in post market stage



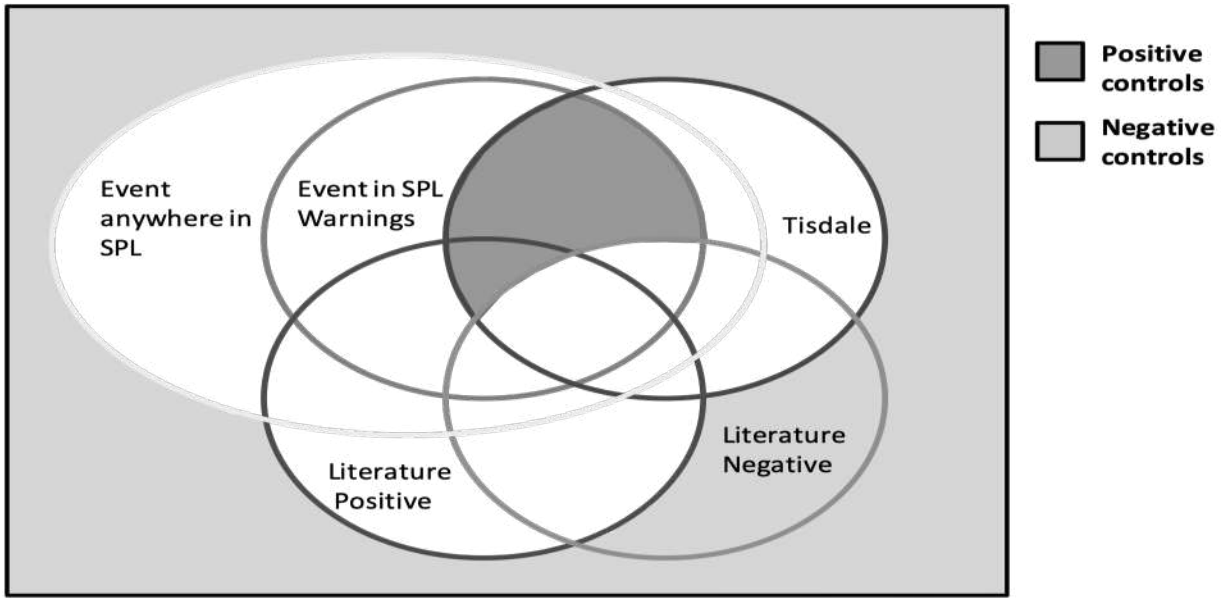
2007 FDA Amendments act FDAAA

Outline


Data Source
Strengths & Limitations
Methods
Performance & case studies
Used in data mining today
Spontaneous Adverse Event Reports
Observa-tional healthcare data
Scientific Literature
Social Medial
Search Engineer
Evidence Integration

Reference Standard – benchmark

- What ADR to monitor?
  - Acute myocardial infarction
  - Acute renal failure
  - Acute liver failure
  - Upper gastrointestinal bleeding



SPL: Structured Product Label
Tisdale: Tisdale's literature review.
Positive literature indicates the set of cases with at least one article confirming the existence of a causal relationship.
Negative literature indicates the set of cases with at least one published study that was sufficiently powered but found no relationship between the drug and outcome.

Ryan, Patrick B., et al., Drug safety 36.1 (2013): 33-47.
http://dailymed.nlm.nih.gov

OMOP Reference Standard

Acute Myocardial Infarction				
Positive controls				
amlodipine	**ketorolac[54]**	almotriptan	factor VIIa	rizatriptan
Negative controls				
benzonatate	**ramelteon**	chlorothiazide	methenamine	stavudine

Statistics for reference standard

Event	Positive Cases	Negative Case	Total
Gastrointestinal Bleeding	24	67	91
Acute Liver Injury	80	37	117
Acute Myocardial Infarction	36	66	102
Acute Renal Failure	24	64	88
Total	164	234	398

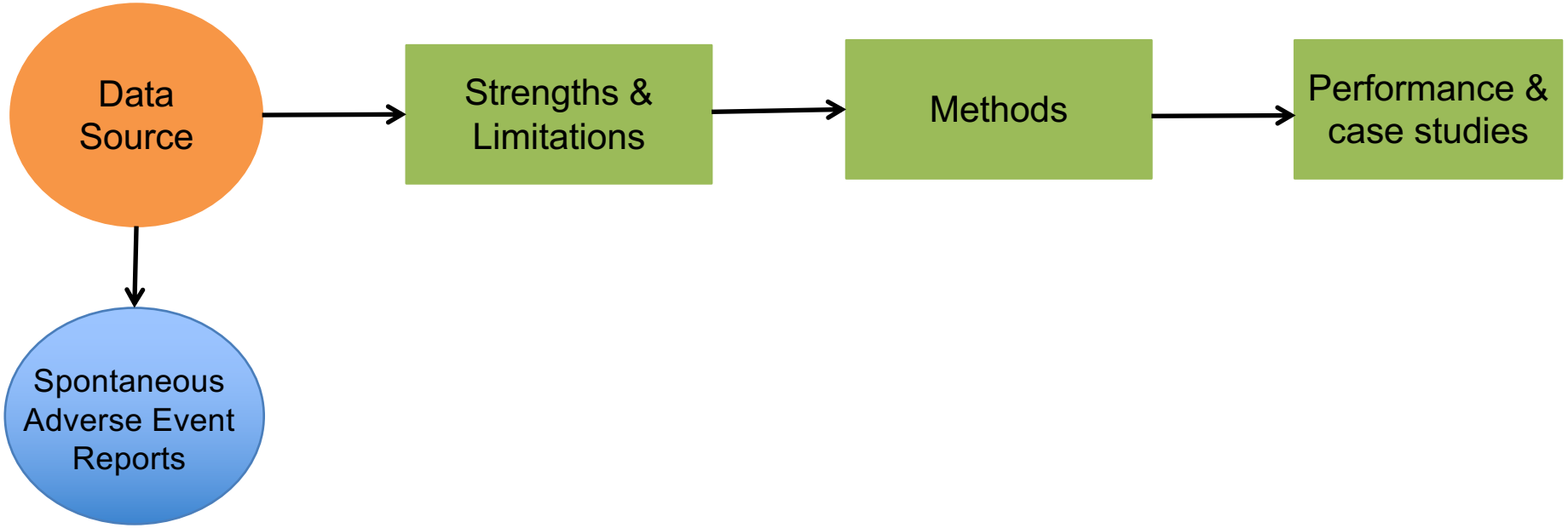
Other reference standards

- SIDER : Side Effect Resource
  - Automatic extraction from FDA structured product label (SPL)
- Time-index reference standard (2013)

EVENT	DRUG	MONTH	APPROVED	BW	W	AR	AR_POSTMARKETING
Taste disorders	Pantoprazole	12	2000				1
Hematopoietic disorders	Pantoprazole	12	2000				1
Anaphylaxis	Dalfampridine	1	2010		1	1	
Anaphylaxis	Mesalamine	12	1993/2007				1
Anaphylaxis	Ketoconazole	7	1981		1		1
Angioedema	Fidaxomicin	4	2011		1		1
Atrial fibrillation	Solifenacin	10	2004				1
Bradycardia	Lacosamide	2	2008		1		1
Biliary tract disorders	Sunitinib	8	2006				1
Coronary Heart Disease	Niacin	2	1997/2008		1		
Drug reaction with eosinophil	Terbinafine	6	1996		1		1
Drug reaction with eosinophil	Mesalamine	12	1993/2007				1
Drug reaction with eosinophil	Clopidogrel	9	1997				1
Dysphonia	Levalbuterol	9	1999				1

http://sideeffects.embl.de/, Harpaz, R. et al. Sci. Data 1:140043 doi: 10.1038/sdata.2014.43 (2014).

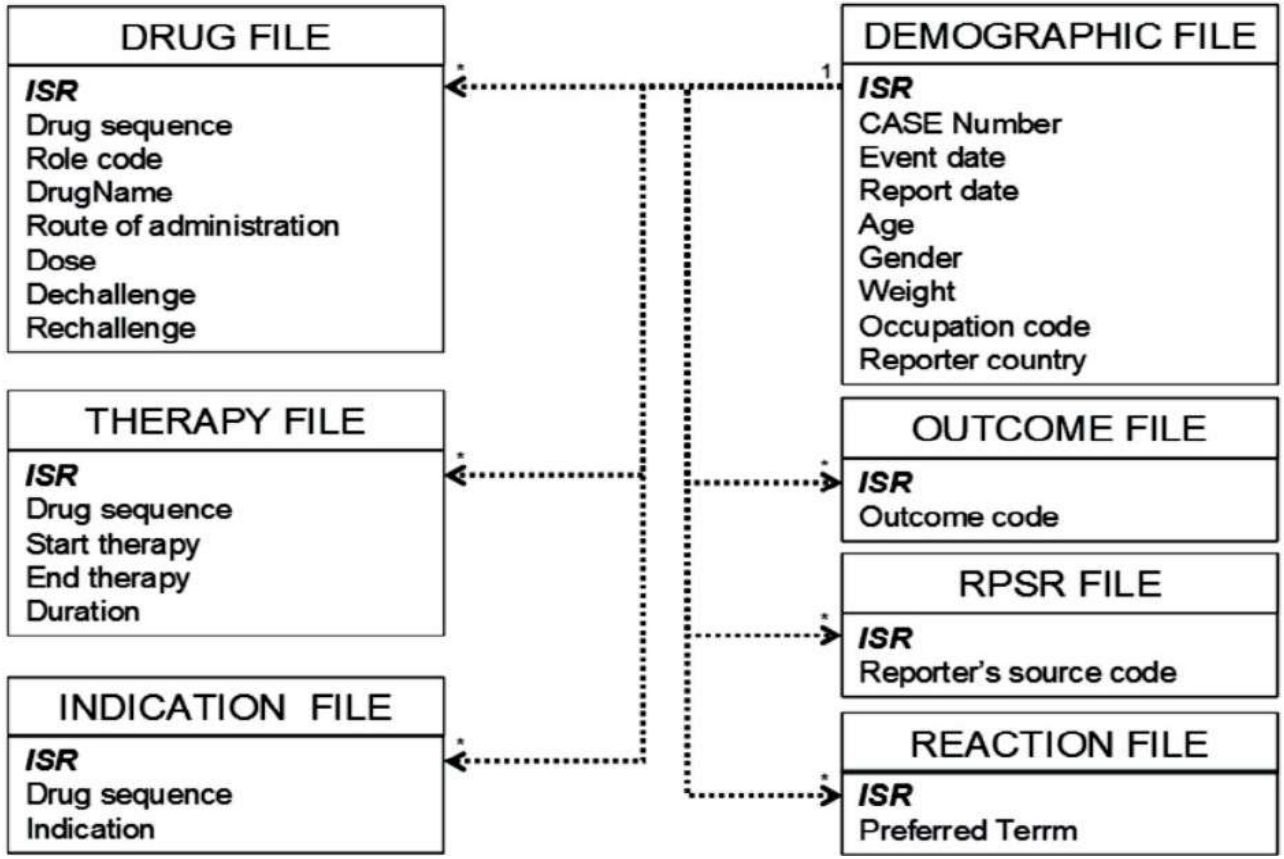
Outline


Data Source
Strengths & Limitations
Methods
Performance & case studies
Spontaneous Adverse Event Reports

Spontaneous reporting systems

Strengths
- Detect rare adverse events
  - Acute liver failures
  - Stevens Johnson syndrome
  - Torsade de pointes

Limitations
- Under and bias reporting
- Lack of accurate “denominators”
- Difficulty detecting events with long latency and with high background rate



Examples of SRSs

SRS	Organization	Number of reports	Availability	Update frequency
FDA Adverse Events Reporting System (FAERS)	US FDA	>9 million (1969-present)	Public (back to 2004)	Quarterly
Vigibase	WHO Programme for International Drug Monitoring	>13 million (1968-present)	Health professionals can request access Public may use VigiAccess for summary statistics	Continuous as received (countries report at least quarterly)
MedEffect	Health Canada	~ 480,000 (1973-2015)	Public	Quarterly

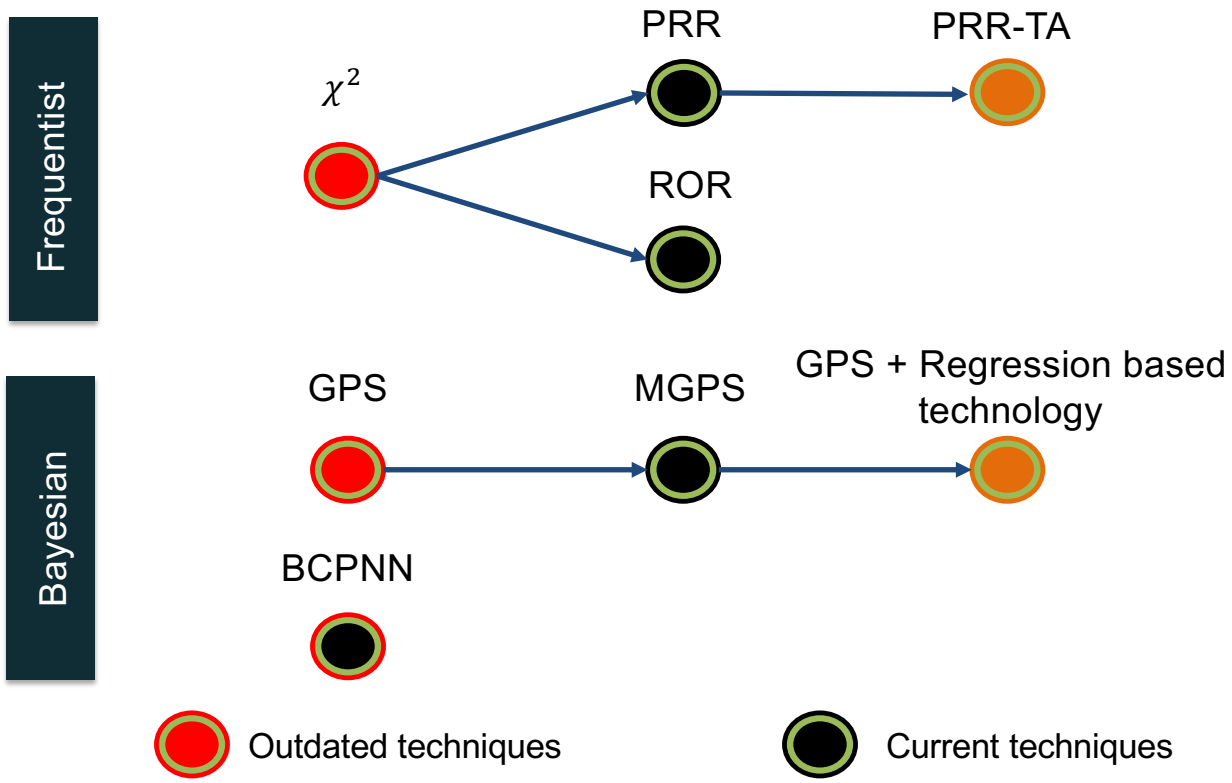
Method - Disproportionality Analysis

- A 2 × 2 Table for Disproportionality Calculation

	Reports with AE	Reports Without AE	Total
Reports with drug	a	b	a+b
Reports without drug	c	d	c+d
Total	a+c	b+d	a+b+c+d

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

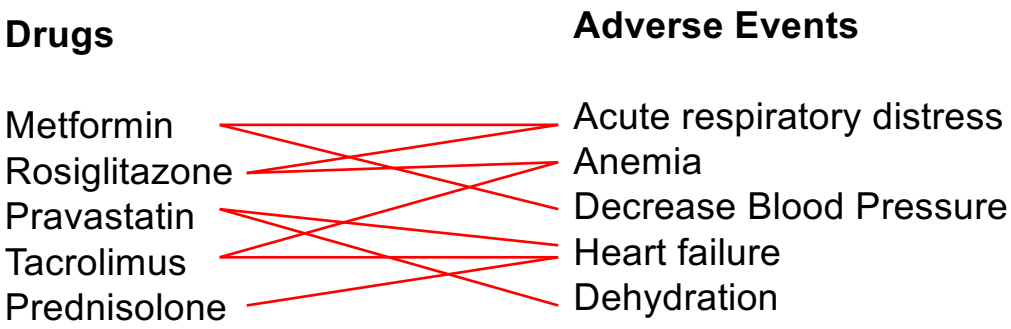
Evolution of disproportionality signal detection methods



- GPS (Gamma Poisson Shrinker) is the simpler precursor to MGPS
- PRR-TA (PRR by therapeutic area) restricts background to therapeutic area of interest, so far seems superior to simple PRR
- GPS + Regression based technology

Interpreting FAERS reports is hard

- Many drugs, many adverse events
  - What causes what?
  - Most of these red lines are false - which are true?
- Is primary suspected information always right?



The Confounding Effect poses many challenges for ADR detection of real world events

Co-Prescription Confounders

Mary has hypertension and arthritis. She has been taking both Aspirin and Vioxx. Which drug caused her heart attack?

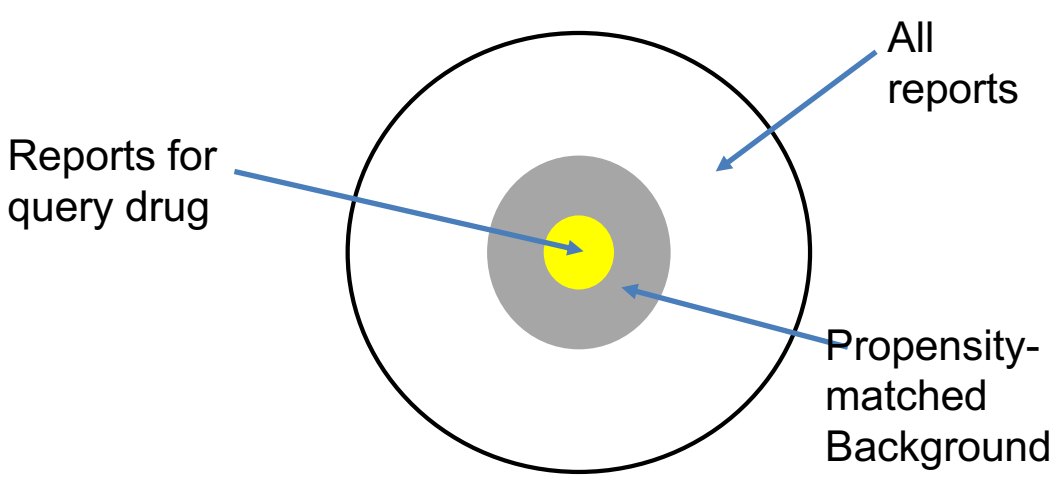
Drug Indicator Confounders

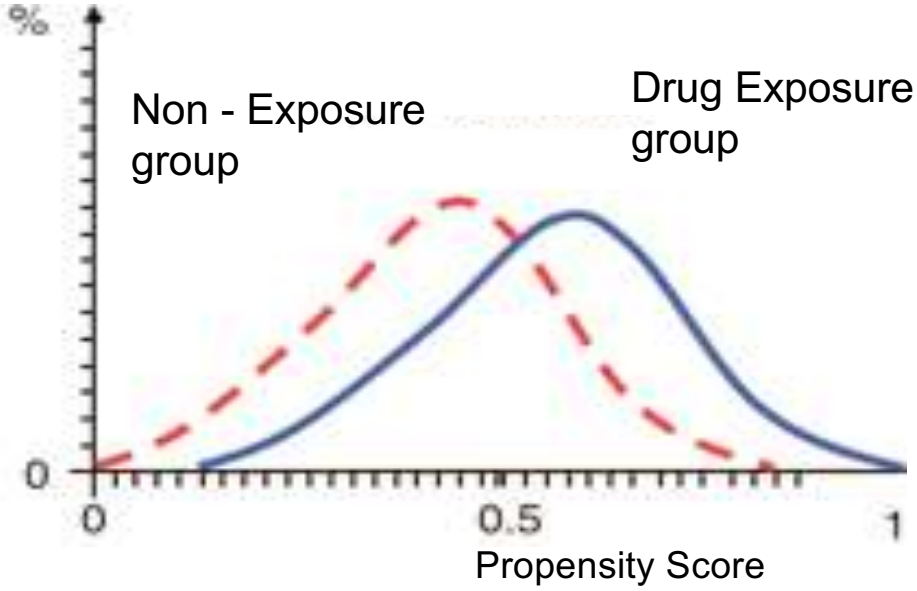
Pancreatitis

Joe is an alcoholic who develops Pancreatitis. He has been drinking daily and taking Naltrexone. What caused the Pancreatitis?

Implicit Propensity Score Matching (IPSM)

$$\boldsymbol{logit}\big(\boldsymbol{P}(\boldsymbol{Drug}=\mathbf{1})\big) = \boldsymbol{\alpha} + \sum_{\boldsymbol{i}=\mathbf{1}}^{\mathbf{200}} \boldsymbol{\delta_i Rx_i} + \sum_{\boldsymbol{j}=\mathbf{1}}^{\mathbf{200}} \boldsymbol{\gamma_j Dx_j}$$

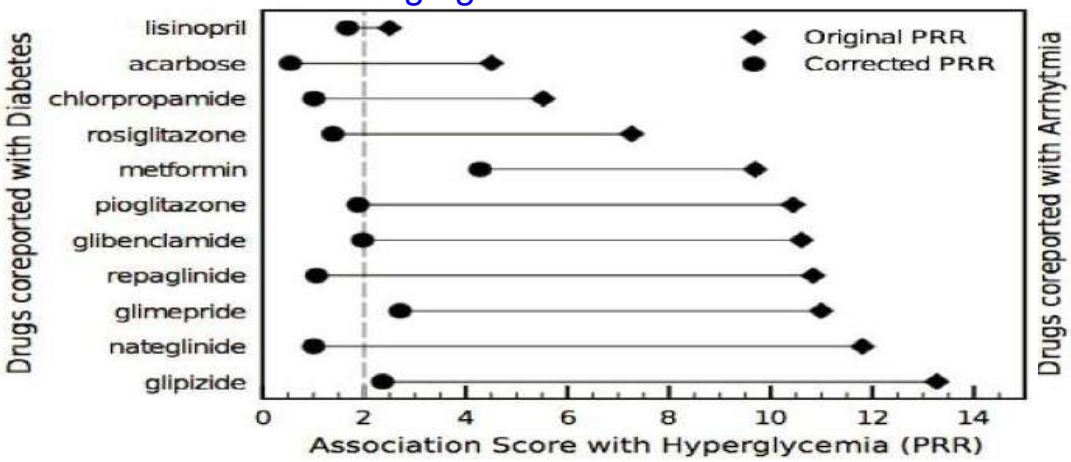




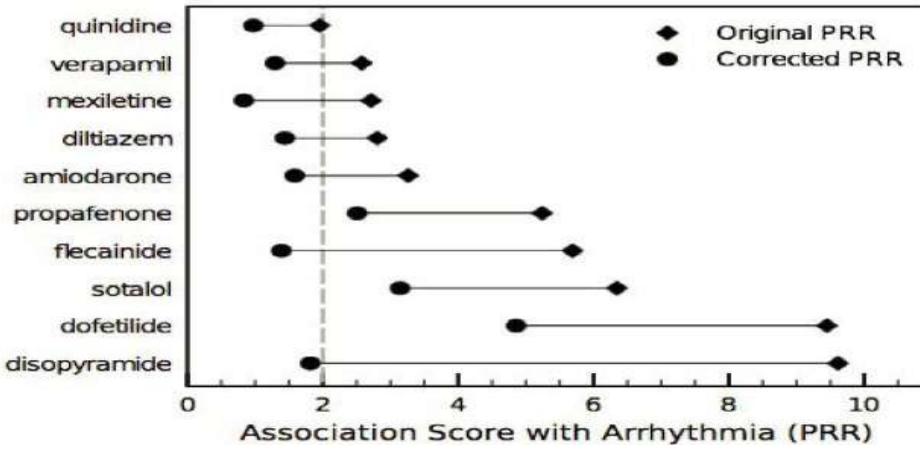
Tatonetti NP et al. Science translational medicine. 2012 Mar 14;4(125):125; Rosenbaum, Paul R., and Donald B. Rubin. Biometrika 70.1, 1983;

IPSM corrects for indication and co-Rx biases

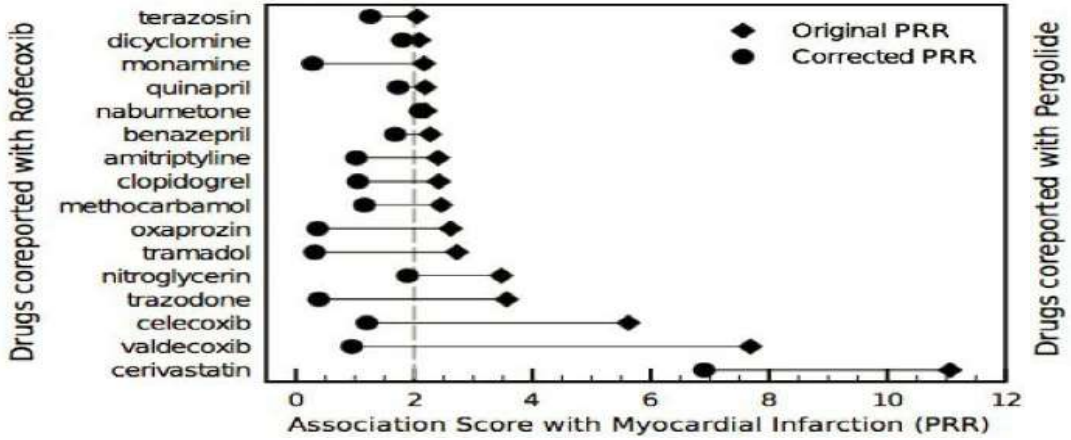
Drugs given to Diabetics



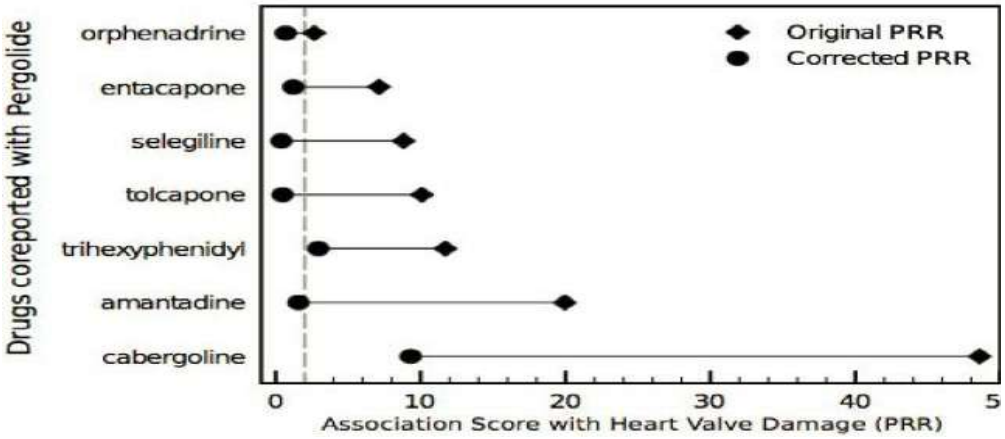
Anti-arrhythmics and Arrhythmia



Drugs co-reported with rofecoxib (Vioxx)

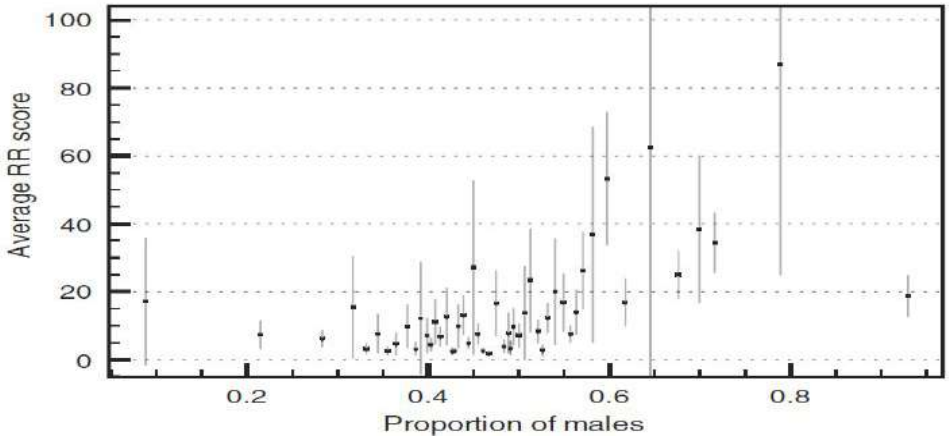


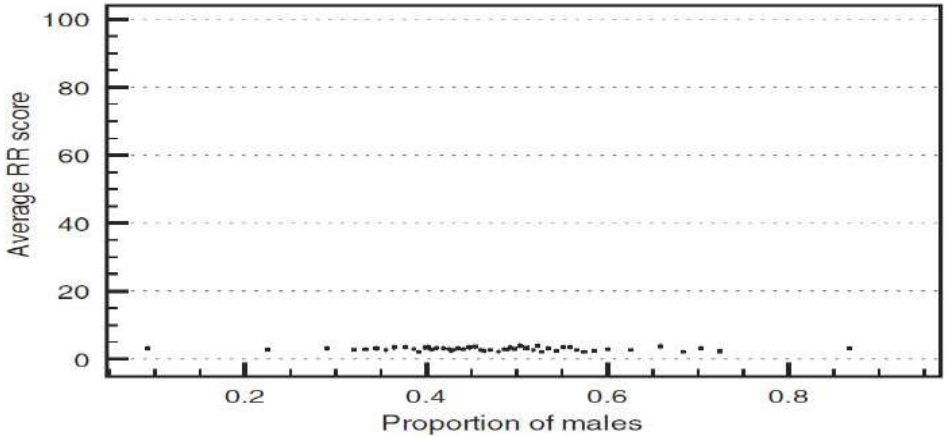
Drugs co-reported with pergolide



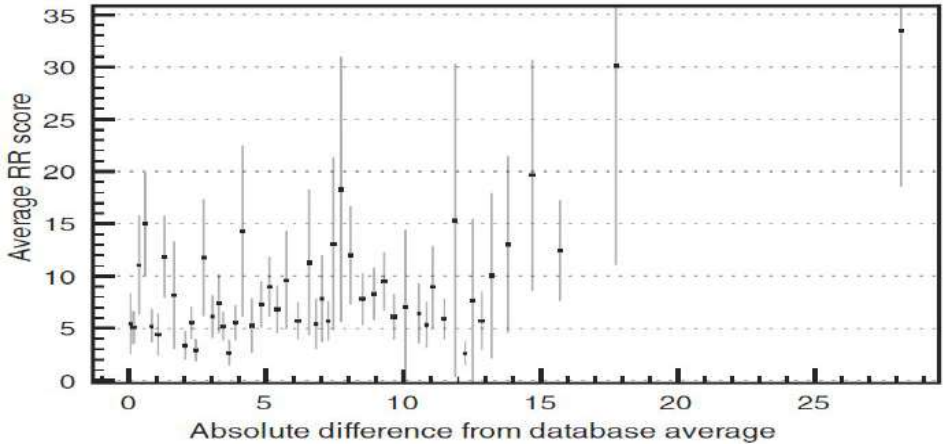
IPSM implicit correction for other biases

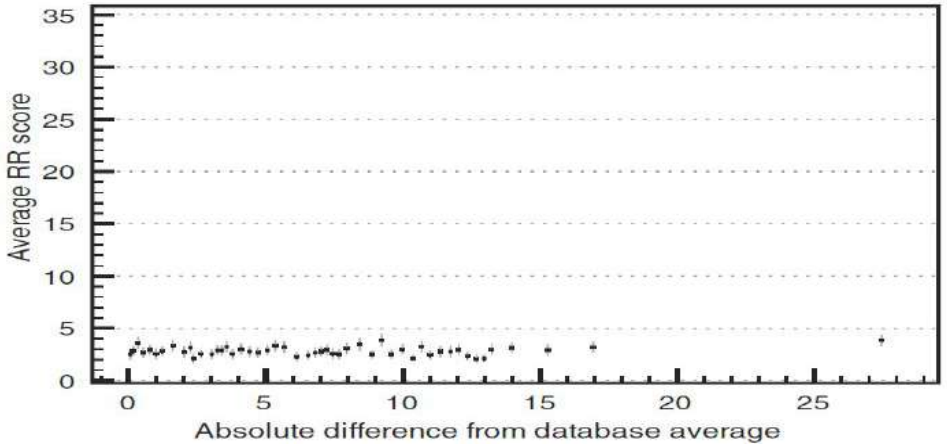
Drugs preferentially associated with males are more likely to be associated with 33 sex-related (male) effects



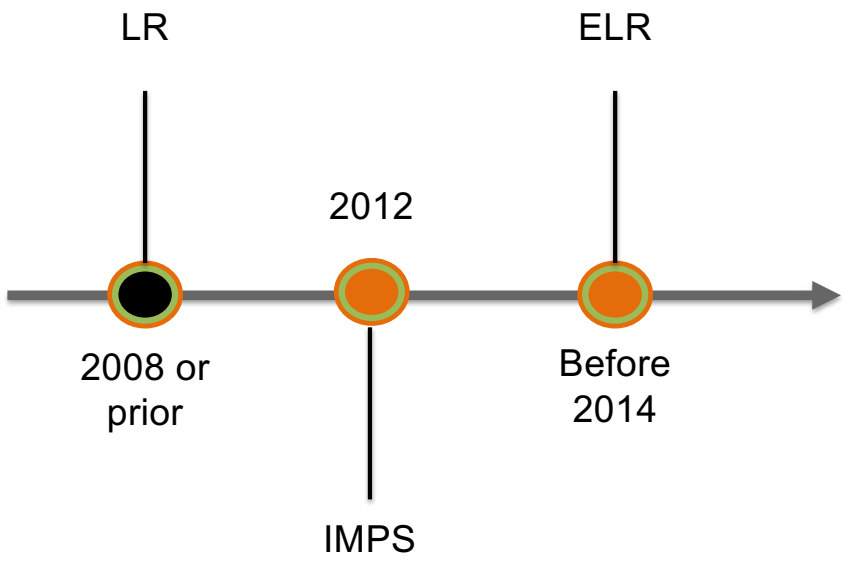


Drugs preferentially associated with young/old patients are more likely to be associated with 48 age-related effects





Evolution of regression based signal detection



- LR (logistic regression) computes odds ratios to measure strength of association between a drug and event while controlling for confounding effect
- ELR (extended logistic regression) is a modification of LR for rare events

Outdated techniques

Current techniques

Emerging techniques

Performance of Pharmacovigilance Signal-Detection Algorithms for the FDA Adverse Event Reporting System

- Data: FAERS data covered the period from 1968 through 2011 Q3, totaling 4,784,337 reports.

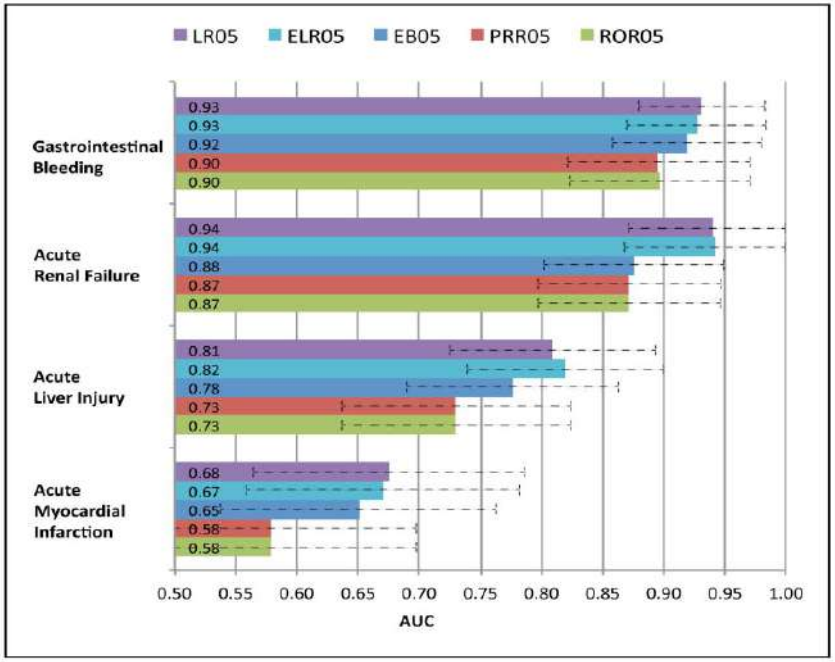
	Method name	Signal score computed
Disproportionality Analysis	Multi-item Gamma Poisson Shrinker (MGPS)	EBGM (empirical Bayes geometric mean): a centrality measure of the posterior distribution of the true observed-to-expected in the population EB05: lower 5th percentile of the posterior observed-to-expected distribution
	Proportional Reporting Ratio (PRR)	PRR: point estimate (mean) of the relative risk reporting ratio distribution PRR05: lower 5th percentile of the relative risk reporting ratio distribution
	Reporting Odds Ratio (ROR)	ROR: point estimate (mean) of the reporting odds ratio distribution ROR05: lower 5th percentile of the reporting odds ratio distribution
Multivariate Modeling	Logistic Regression (LR)	LR: point estimate of the odds ratio distribution obtained from logistic regression LR05: lower 5th percentile of the odds ratio distribution obtained from logistic regression
	Extended Logistic Regression (ELR)	ELR: point estimate of the odds ratio distribution obtained from extended logistic regression ELR05: lower 5th percentile of the odds ratio obtained from extended logistic regression

Harpaz, Rave, et al. "Performance of Pharmacovigilance Signal-Detection Algorithms for the FDA Adverse Event Reporting System." Clinical Pharmacology & Therapeutics 93.6 (2013): 539-546.

Performance of Pharmacovigilance Signal-Detection Algorithms for the FDA Adverse Event Reporting System

Reference Standard

Event	Positive Cases	Negative Case	Total
Gastrointestinal Bleeding	24	67	91
Acute Liver Injury	80	37	117
Acute Myocardial Infarction	36	66	102
Acute Renal Failure	24	64	88
Total	164	234	398



Harpaz, Rave, et al. 2013, CPT; Ryan, Patrick B., et al., 2013, Drug Safety

Summary - strengths and weaknesses of notable signal detection methods

	PRR	ROR	MGPS	BCPNN	LR
Simple to use					
Applicable to low event counts					
Easy to interpret					
Usable with SRS data					
Accounts for confounding factors					
Sensitivity					
Specificity					
					

Notes: The ROR can be incorporated into a logistic regression analysis. A kind of de-confounding can be done with PRR and ROR by splitting the data inputs into separate contingency tables, but is not inherent to the algorithm.

Triaging to select signals and follow up

QUANTITATIVE "RULES"

- Apply fixed thresholds
  - EB05 ≥2; EBGM ≥2; EBGM ≥4;
  - PRR ≥2; a number of reports (N) ≥3; a Chi-square ≥4
  - Lower 95% CI of PRR ≥1
  - Lower 95% CI of ROR ≥1
  - IC025 > 0
- Apply flexible thresholds
  - Estimate the false discovery rate (FDR) to decide threshold on a signal-by-signal basis

QUALITATIVE "RULES"

- Novel
  - Not currently known and on drug label
  - New adverse event or new drug ("early warning")
- High potential relevance
  - Public health issue – e.g. important drug (serious indication, widely used), serious reaction, many cases
  - Change in merit/harm
- Strong evidence
  - Exposure-response relationship (site, time-to-onset, dose, reversibility in dechallenge/rechallenge)
  - Reasonable from a biological mechanism perspective
- Time trend
  - Surge in recent reporting, notable increase in reporting over time

Meyboom RH, et al. Drug safety. 2002 May 1;25(6):459-65.

Unsupervised method - Biclustering

Table 1. Contingency table specifying the number of reports mentioning a specific drug and a specific adverse effect (AE)

	Target AE	All other AEs	Total
Target drug	a	b	n=a+b
All other drugs	c	d	c+d
Total	m=a+c	b+d	t=a+b+c+d

$$b_{ij} = \begin{cases} 1 & if\ a_{ij} \geq T \\ 0 & if\ a_{ij} < T \end{cases}$$

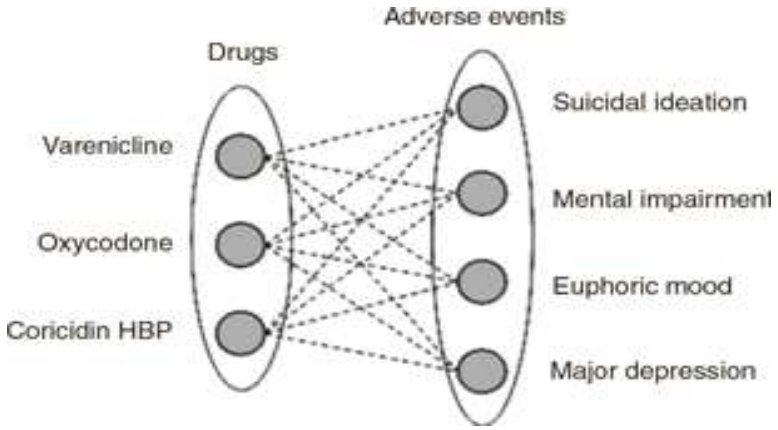
a_{ij} contains GPS' EBGM association strength value computed for the *i*-th drug and the *j*-th AE pair.

	e1	e2	e3	e4	e5
d1	1	1	0	1	1
d2	0	1	1	1	0
d3	1	1	0	1	1
d4	0	0	0	0	0
d5	0	1	1	1	0
d6	1	1	0	1	1

	e1	e5	e2	e4	e3
d1	1	1	1	1	0
d3	1	1	1	1	0
d6	1	1	1	1	0
d4	0	0	0	0	0
d2	0	0	1	1	1
d5	0	0	1	1	1

Binary inclusion-maximal biclustering

Case Study

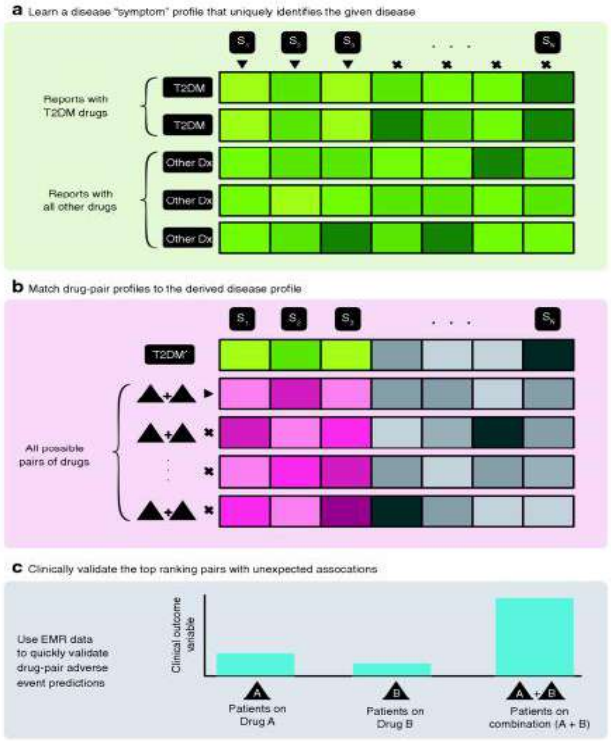


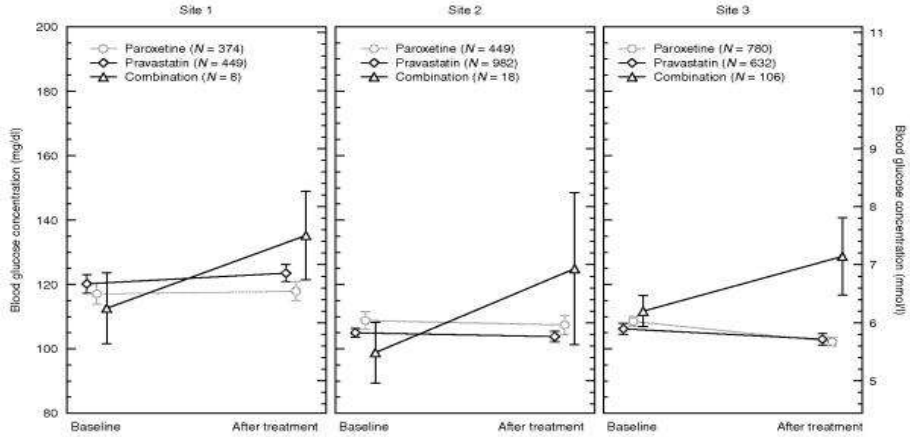
Harpaz, Rave, et al., Clinical Pharmacology & Therapeutics 89.2 (2011): 243-250.

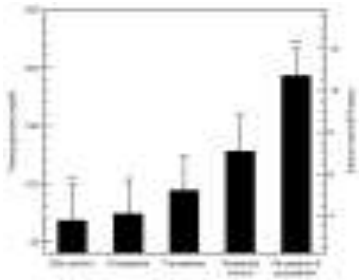
Beyond ADR detection

Common drug combo increases diabetes risk

Hypothesis generation based on FAERS → Signal validation based on EHR databases → Mice model validation





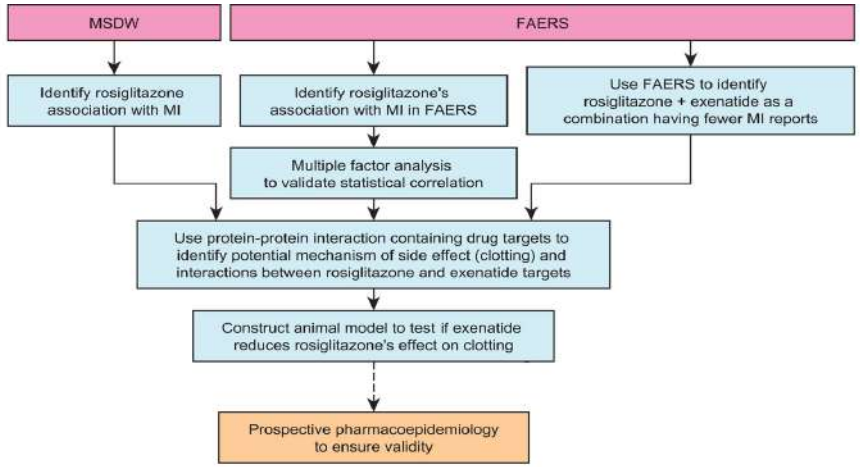


Tatonetti, Nicholas P., et al. Clinical pharmacology and therapeutics 90.1 2011

Beyond ADR detection

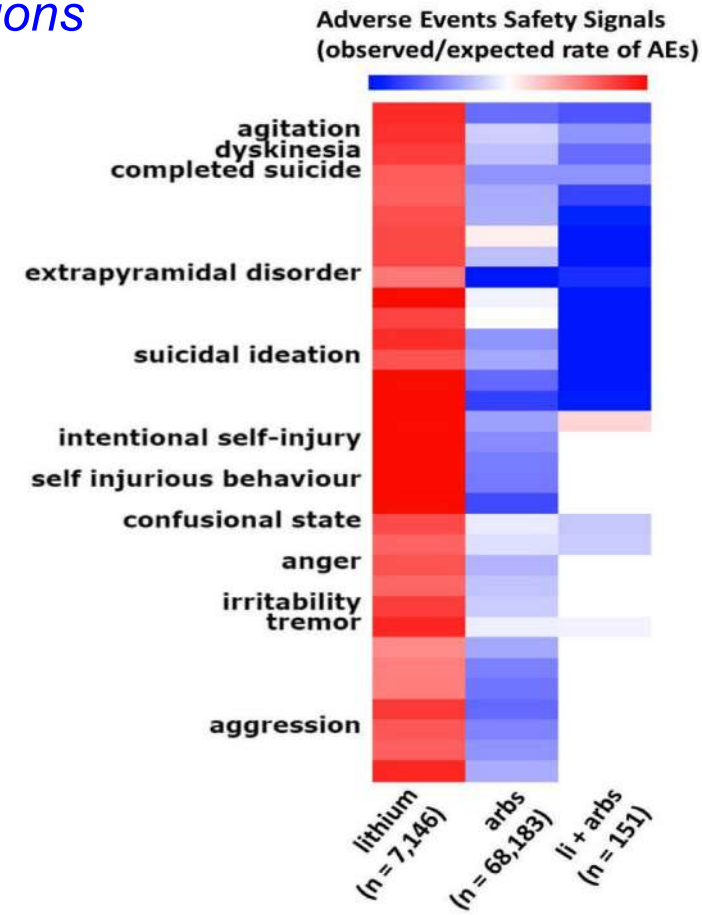
Common drug combo decreases adverse drug reactions

A



B

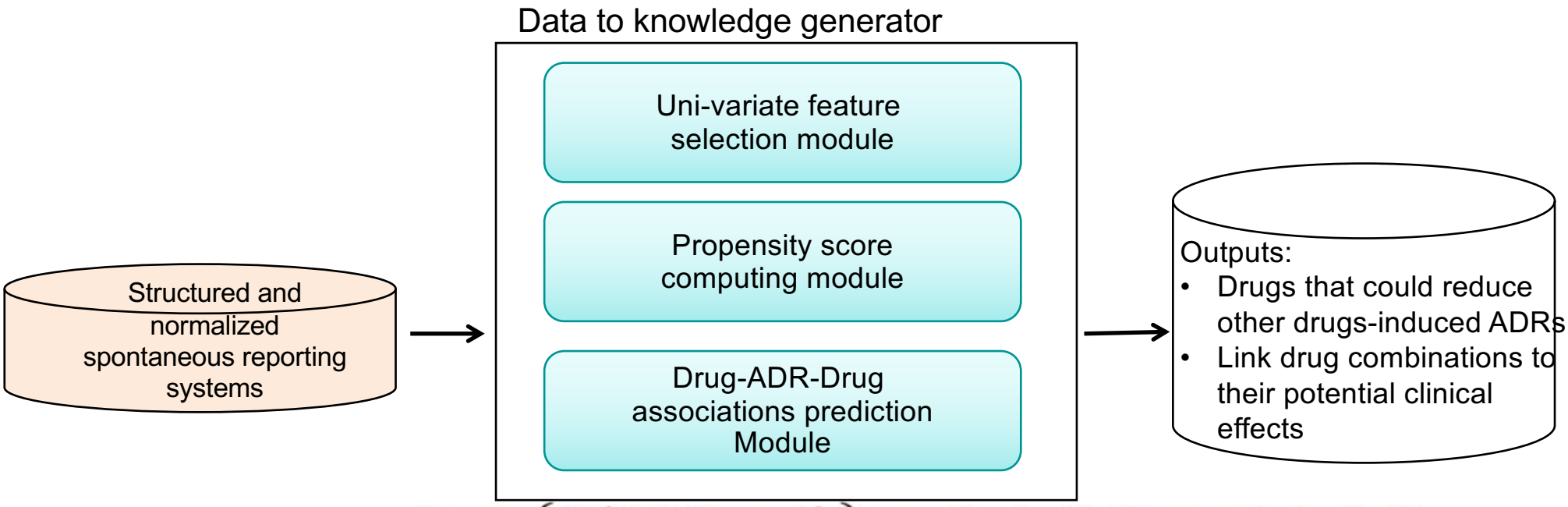
Drug B	Rosiglitazone MI without drug B (%)	Rosiglitazone MI with drug B (%)	Odds ratio	95% CI	*P*
Exenatide	19,910/58,604 (33.97)	95/4,460 (2.13)	0.04	(0.03-0.05)	7.0E-203
Metformin	17,481/51,266 (34.10)	2,524/11,798 (21.39)	0.53	(0.50-0.55)	6.0E-154
Aspirin	19,107/57,813 (33.05)	898/5,251 (17.10)	0.42	(0.39-0.45)	1.8E-118
Furosemide	19,439/59,555 (32.64)	566/3,509 (16.13)	0.40	(0.36-0.43)	4.5E-87
Acetaminophen	19,623/60,635 (32.36)	382/2,429 (15.73)	0.39	(0.35-0.44)	1.5E-62
Levothyroxine	19,591/60,660 (32.30)	414/2,404 (17.22)	0.44	(0.39-0.49)	5.8E-52
Warfarin	19,824/61,666 (32.15)	181/1,398 (12.95)	0.31	(0.27-0.37)	2.2E-47
Glibenclamide	19,117/59,052 (32.37)	888/4,012 (22.13)	0.59	(0.55-0.64)	1.1E-40
Quetiapine	19,954/62,368 (31.99)	51/696 (7.33)	0.17	(0.13-0.22)	1.9E-34
Gliclazide	19,971/62,409 (32.00)	34/655 (5.19)	0.12	(0.08-0.16)	3.1E-34



Sarangdhar, Mayur, et al. Nature Biotechnology, 2016; Zhao, Shan, et al. Science translational medicine (2013)

arbs: angiotensin II receptor blockers

Data-Driven Prediction of Beneficial Drug Combinations in Spontaneous Reporting Systems



$$logit(P(ADR = 1)) = \beta_0 + \beta_1 DrugA + \beta_2 P_1 + \beta_3 DrugB + \beta_4 P_2 + \beta_5 DrugA * DrugB + \lambda|\beta|_1$$

Our novel regularized logistic regression is able to reveal two different mechanism of drug combinations

- （$\beta_3 + \beta_5$）: the degree that a patient who is on Drug A could benefit or suffer from taking Drug B for the ADR of interest
- β_5: the degree that the interaction effect between Drug B and Drug A on the ADR

Clinical validation

Pamidronate is used to treat high blood calcium levels

List of 15 predicted beneficial drug combinations and their ADR reduction

Drug A name	ADRs associated with drug A	Drug B name	Predicted beneficial score	Common ATC code	Evidence for combined use
benazepril	DIZZINESS	amlodipine besylate	-0.57	yes	F
atovaquone	PYREXIA	proguanil	-0.36	yes	F
rofecoxib	MYOCARDIAL INFARCTION	pamidronate	-0.33	yes	
rosiglitazone	MYOCARDIAL INFARCTION	exenatide	-0.32	yes	
progesterone	BREAST CANCER	adalimumab	-0.27	no	
trimethoprim	PYREXIA	sulfamethoxazole	-0.17	yes	F
exemestane	ARTHRALGIA	everolimus	-0.16	yes	III
amoxicillin	DIARRHOEA	clavulanic acid	-0.15	yes	IV
ampicillin	PYREXIA	sulbactam	-0.15	yes	F
desmopressin	HYPONATRAEMIA	somatropin	-0.15	yes	
sertraline	ANXIETY	nicotinic acids	-0.14	no	
sumatriptan	MIGRAINE	naproxen	-0.14	no	F
olanzapine	DIABETES MELLITUS	biperiden	-0.13	yes	
clindamycin	DIARRHOEA	benzoyl	-0.13	yes	F
fluticasone	DYSPNOEA	salmeterol	-0.13	yes	F

F: FDA approved drug combination; III: phase III clinical trial; IV: phase IV clinical trial

a NSAID. On September 30, 2004, Merck withdrew rofecoxib from the market because of concerns about increased risk of heart attack and stroke associated with long-term, high-dosage use.

From Passive to Active Surveillance

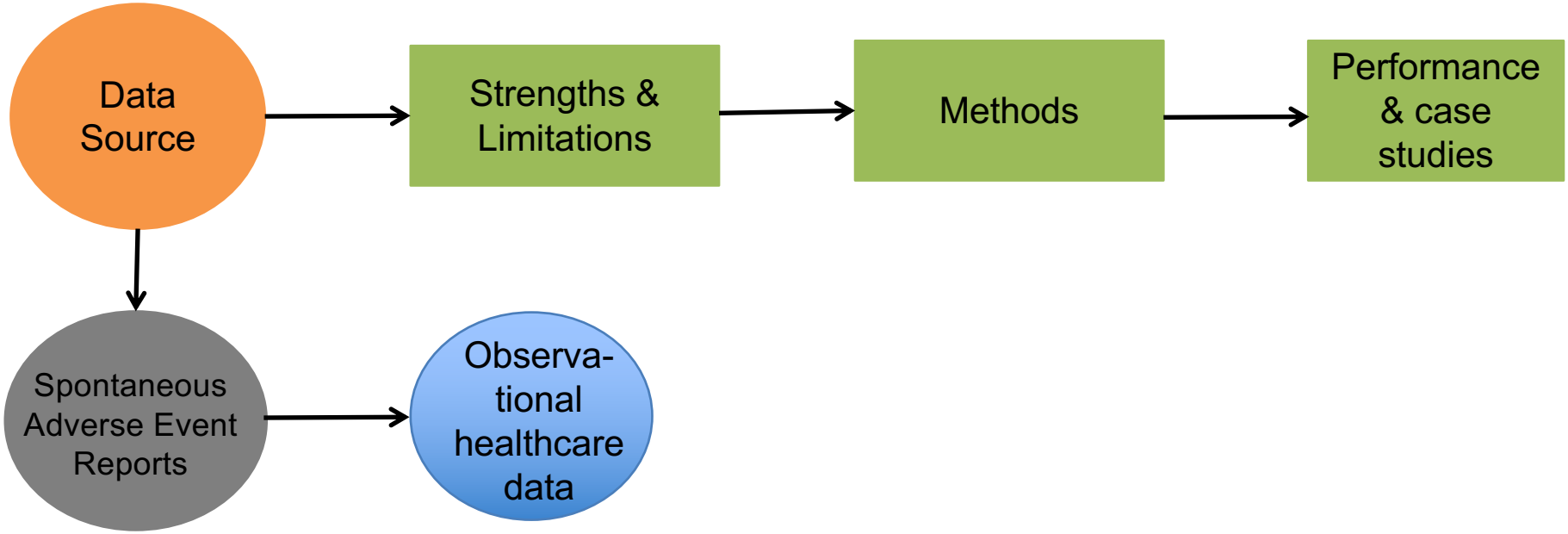
Regulatory Agencies

Transforming how we monitor the safety of FDA-regulated products

Academic and Nonprofit Organizations

http://www.mini-sentinel.org/

Outline


Data Source
Strengths & Limitations
Methods
Performance & case studies
Spontaneous Adverse Event Reports
Observa-tional healthcare data

Observational healthcare databases (OHD)

Subtype
- EHR
- Claims

Strength
- No reporting biases
- Events with high background rate
- Information with exposed patients
- Comprehensive and longitudinal patient information

Limitations
- Biases due to secondary use
- Confounding
- False positive discovery
- Missing and irregular data
- Not publicly available

Patient Demographics
- Age
- Race
- Ethnicity
- Gender
- Zip 3
- Payer
- Status
- Tenure

Provider Demographics
- Specialty
- Role

Therapeutics
- Ambulatory & Inpatient
- Drug - Brand and Class
  - SNOMED, NDC, RxNorm
- Medication start & end dates
- Select Reasons for Stopping
- Dosage, refills, & quantity

Device
- Implant site & type
- Date of implant
- Manufacturer
- Model no.

Financial
- Billing 837/835
- Claims

PROs
- HOOS
- KOOS
- PHQ2/9

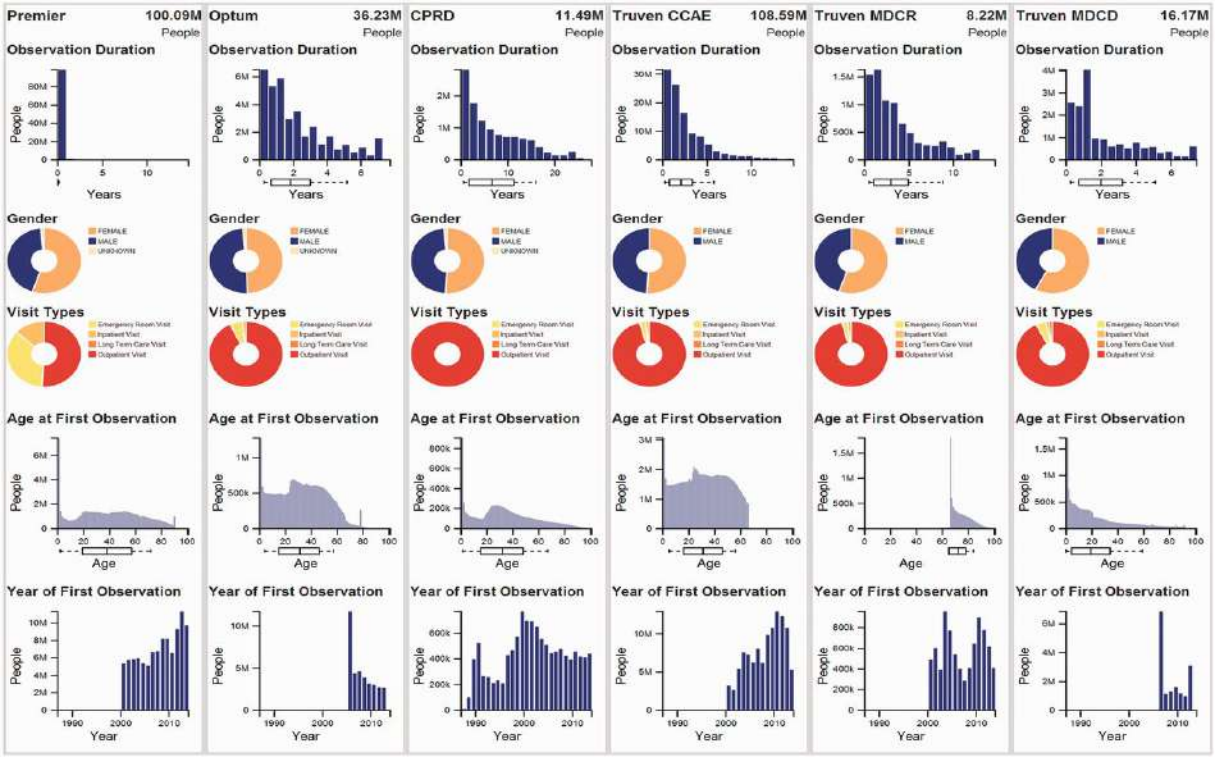
Clinical
- EMR and Billing Diagnoses
- Problem list
  - w/Start & end dates
- Allergies
- Immunizations
- Procedures
  - CPT, HCPCS, ICD-9/10
- Medical & Social History
- Surgical history

Vitals & Biometrics
- BP
- BMI
- Body temp
- Heart rate
- Respiratory rate
- BSA

Utilization
- Site of care & service dates
- Encounters, admissions, and discharges
  - Inpatient, ambulatory, ED, SNF, etc.
- IDN and Community (CINs)
- Length of Stay and Discharge Disposition
- Appointments
  - Missed, Cancelled, Scheduled, Left w/o seen

Laboratory *(representative only)*
- CBC
- Fibrinogen
- Hemoglobin A1C
- BMP & CMP
- DHEA
- PSA
- Homocysteine
- C-reactive protein
- TSH & T4
- Testosterone
- Estradiol
- Amylase
- PT (Protime)
- Electrolytes
- ESR
- Glucose
- hCG
- Lipid profile
- Liver panel
- Microalbumin
- Sodium
- BNP

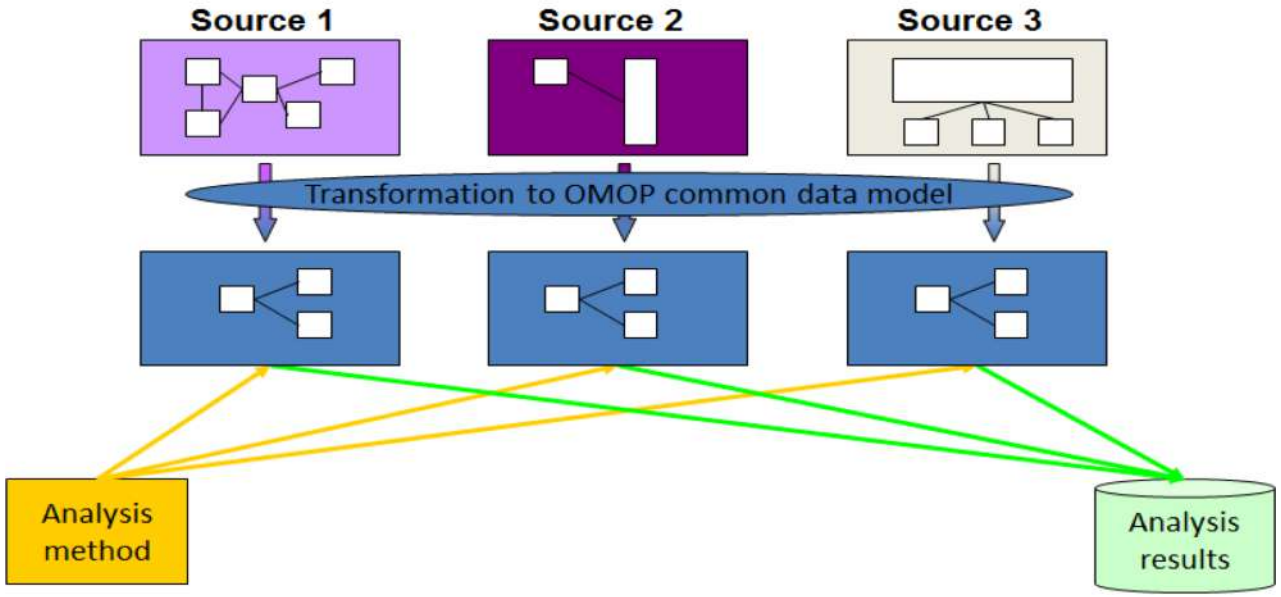
Summary statistics for OHD



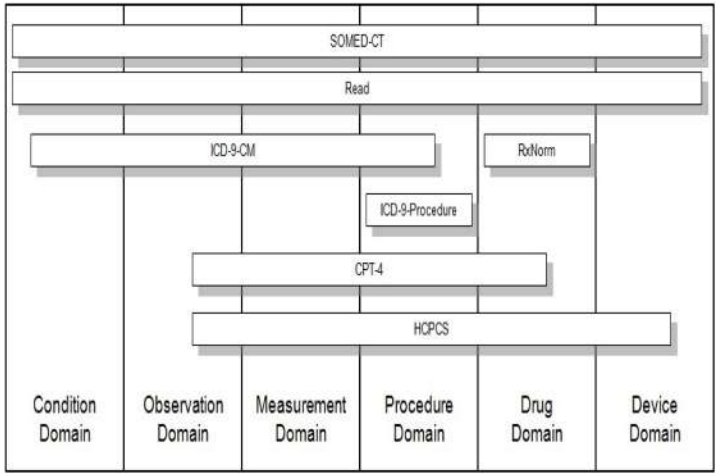
CCAE : MarketScan Commercial Claims and Encounters
MDCD : MarketScan Multi-State Medicaid
MDCR : MarketScan Medicare Supplemental Beneficiaries
MSLR : MarketScan Lab Supplemental

Voss, Erica A., et al. "Feasibility and utility of applications of the common data model to multiple, disparate observational health databases." Journal of the American Medical Informa
(2015): 553-564.

Common Data Model



Medical Terminologies



Mini-sential Common Data Model; I2B2 common data model; PCORnet Common Data Model (CDM) - PCORnet

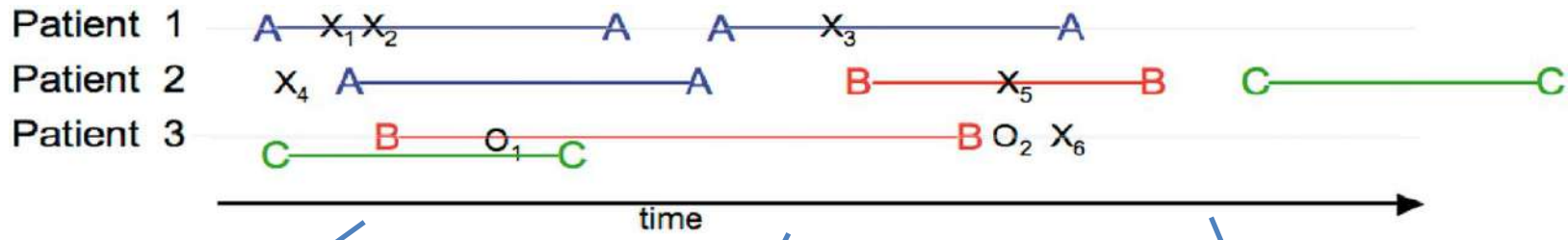
http://www.ohdsi.org/data-standardization/the-common-data-model/

Overview of methods based on OHD

- Disproportionality methods
- Longitudinal Gamma Poisson Shrinker
- Observational screen
- Multiple self-controlled case series
- High-dimensional Propensity Score

Disproportionality methods – How to count

Prevalence based



Prevalent distinct patients

	Event X	Not X
A	1 (patient 1)	1 (patient 2)
Not A	1 (patient 3)	0

Prevalent mimic SRS

	Event X	Non X
A	3 (A+X1, A+X2, A+X3)	0
Not A	1 (B+X5)	2 (B+O1, C+O1)

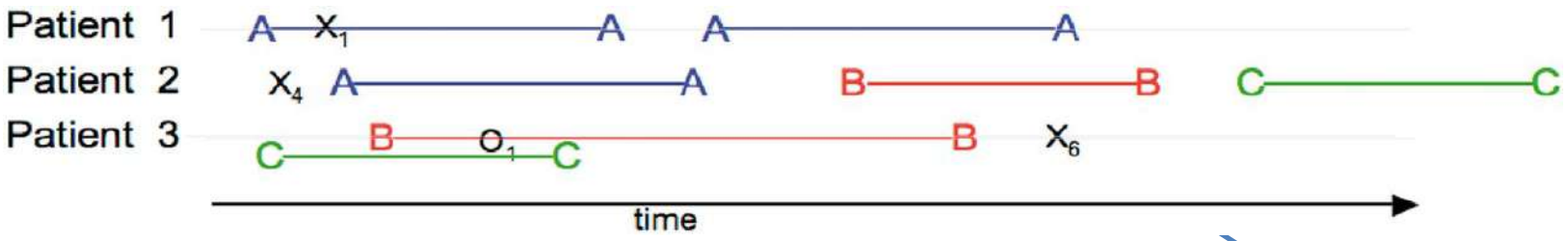
Prevalent mimic modified SRS

	X	Non X + not X
A	3 (A+X1, A+X2, A+X3)	1 (patient 2's A era)
Not A	3 (X4, B+X5, X6)	4 (patient 2's C era, B+O1, C+O1, O2)

Ivan Zorych et al. Stat Methods Med Res 2011;22:39-56

Disproportionality methods – How to count (cont')

Incidence based



Incident distinct patients

	Event X	Not X
A	1 (patient 1)	1 (patient 2)
Not A	1 (patient 3)	0

Incident mimic SRS

	Event X	Non X
A	1 (patient 1)	0
Not A	0	2 (B+O1, C+O1)

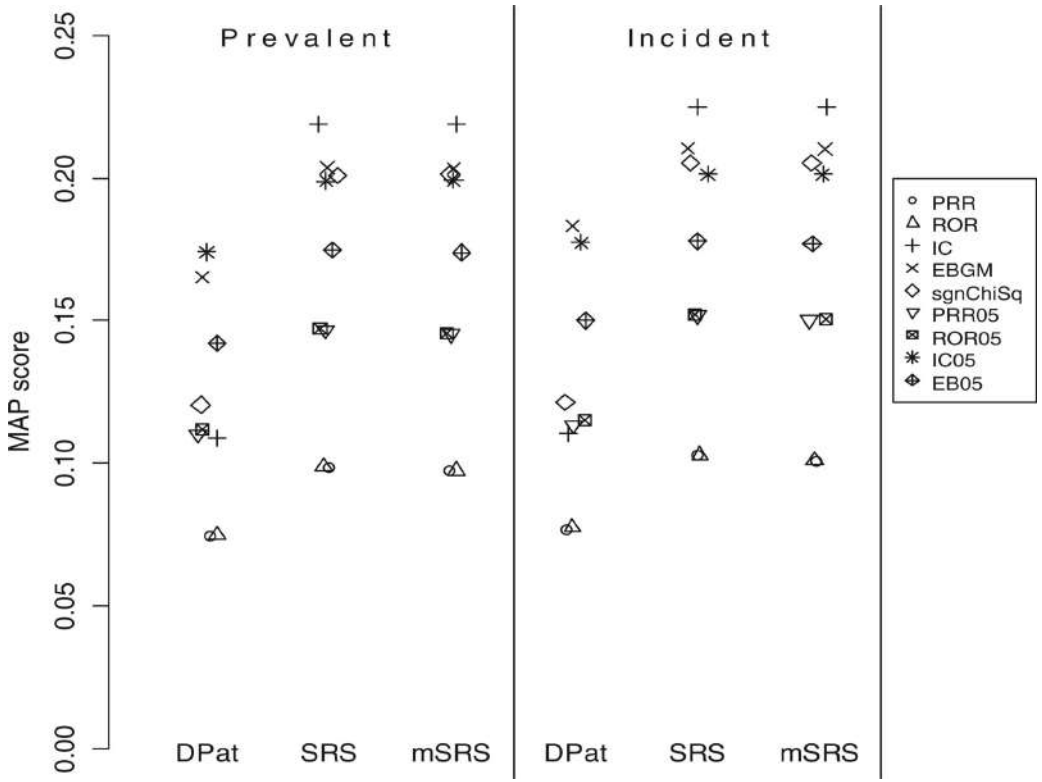
Incident mimic modified SRS

	Event X	Non X + not X
A	1 (A+X1)	2 (patient 1,2's A era)
Not A	2 (X4, X6)	4 (patient 2's B, C, B+O1, C+O1)

Ivan Zorych et al. Stat Methods Med Res 2011;22:39-56

Disproportionality methods - Results

MAP Scores for DP Methods (simulated data).

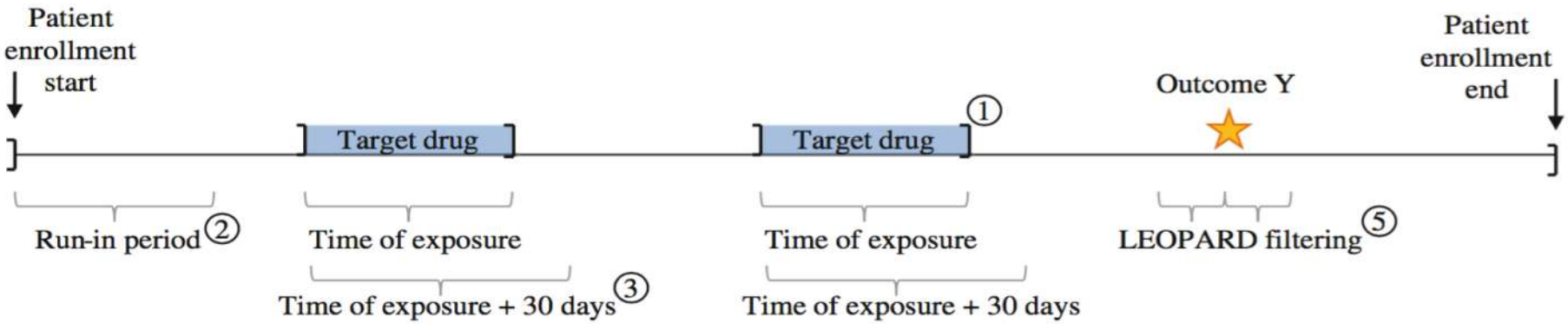


Ivan Zorych et al. Stat Methods Med Res 2011;22:39-56

Take home messages

- Shrinkage measures, IC and EBGM performs best
- Derivative shrinkage measures, EB05 and IC05 and signed chi-square test, have the second best performance
- SRS and modified SRS are better representations than distinct patients

Longitudinal Gamma Poisson Shrinker (LGPS)



$$E = t_1 \frac{O_0}{t_0} \quad \Leftarrow \quad E = \frac{(a+b)*(a+c)}{(a+b+c+d)}$$

	Reports with AE	Reports Without AE	Total
Reports with drug	a	b	a+b
Reports without drug	c	d	c+d
Total	a+c	b+d	a+b+c+d

Schuemie, Martijn J. Pharmacoepidemiology and drug safety 20.3 (2011): 292-299.

Observational screen

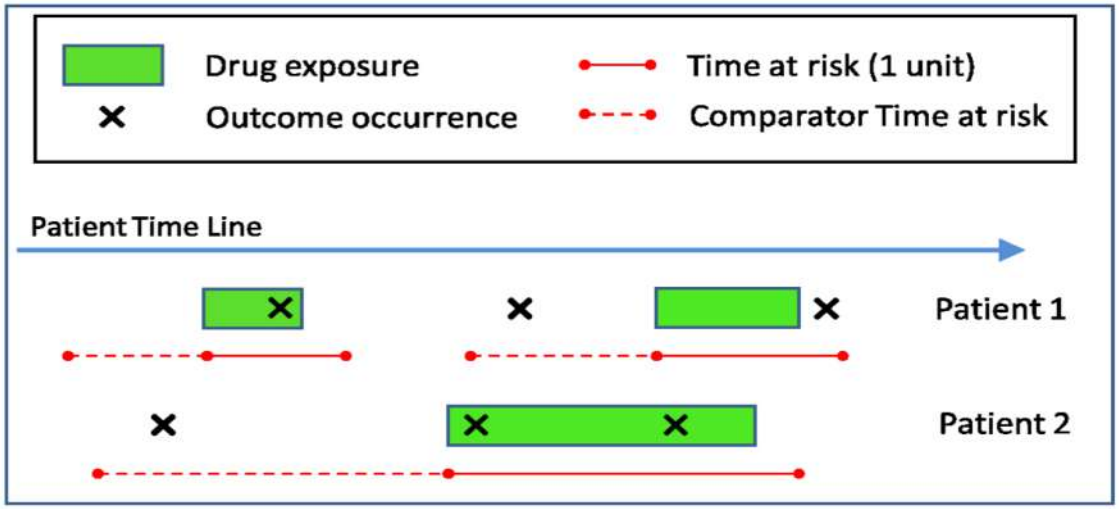


Figure 1: Illustration of the components used in the calculation performed by the Observation Screening signal detection method.

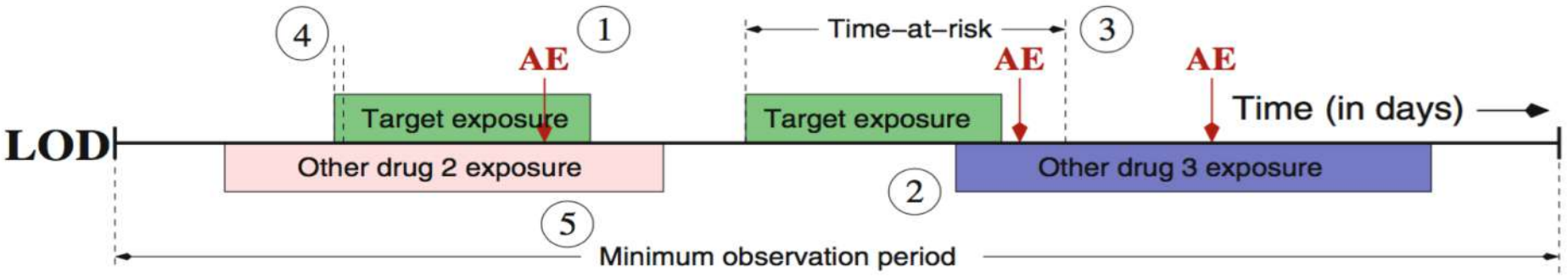
$$\textbf{Screening Rate (SR)} = \frac{\textbf{\# of outcome}}{\textbf{Total time at risk}}$$

$$\textbf{Screening Rate Ratio (SRR)} = \frac{\textbf{SR of exposed group}}{\textbf{SR of unexposed group}}$$

Specifically
SR of exposed group=(1+1+2)/(2+3+5)
SR of unexposed group=(1+1)/(3+5)
SRR=(4/10)/(2/8)=1.6

Harpaz, Rave, et al. KDD, 2013.

Multiple self-controlled case series



$$p(y_i|x_i) = \prod_{d=1}^{t_i} P(y_{id}|x_{id}) = \prod_{d=1}^{t_i} \frac{e^{-\lambda_{id}} \lambda_{id}^{y_{id}}}{y_{id}!}$$

$$= \exp\left(\phi_i n_i - e^{\phi_i} \sum_d e^{x_{id}'\boldsymbol{\beta}}\right) \prod_{d=1}^{t_i} \frac{(e^{x_{id}'\boldsymbol{\beta}})^{y_{id}}}{y_{id}!}$$

$$x_{id} = (x_{id1,\ldots,} x_{idJ})'$$

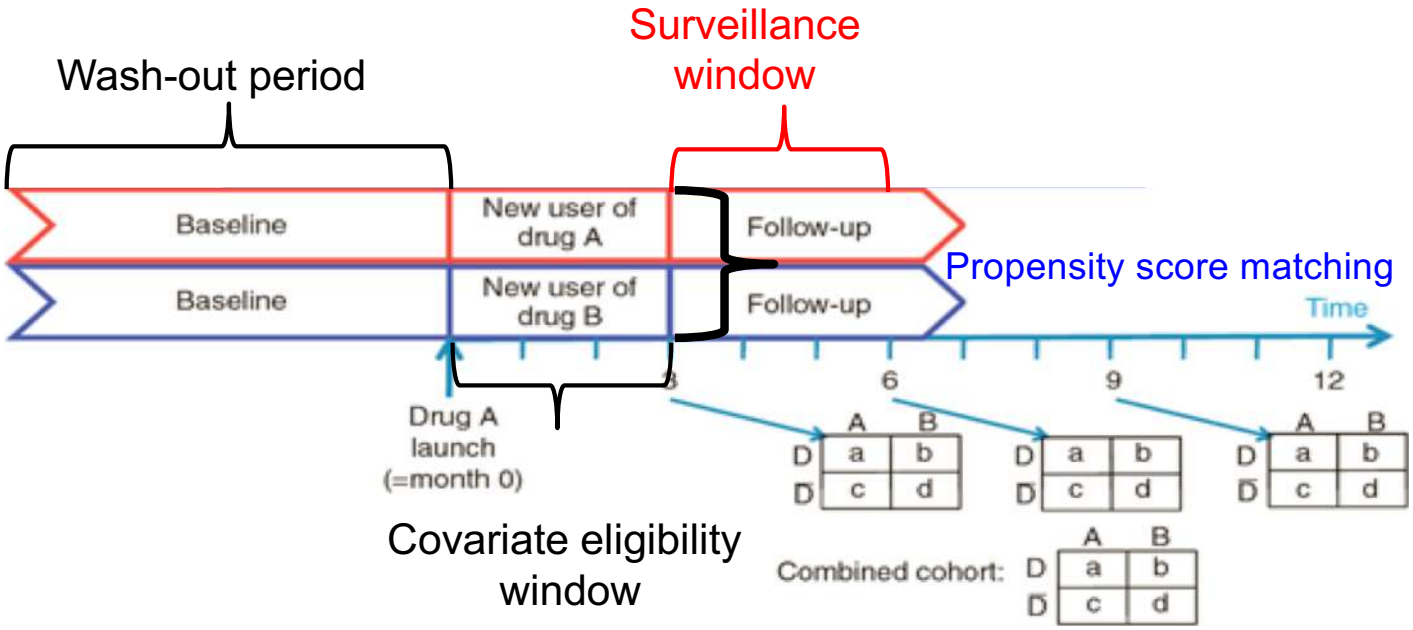
$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left[\sum_{d=1}^{t_i} y_{id}\, x_{id}'\boldsymbol{\beta} - n_i \log\left(\sum_{d=1}^{t_i} e^{x_{id}'\boldsymbol{\beta}} \right) \right] - f(\boldsymbol{\beta})$$

$$f(\boldsymbol{\beta}) = \begin{cases} \lambda_1 \sum_{j=1}^{J} |\beta_j| \;\; under\ an\ L_1\ norm \\ \lambda_2 \sum_{j=1}^{J} \beta_j^2 \;\; under\ an\ L_2\ norm \end{cases}$$

i=1,2,….,n, index patients; d index days; ti is the total number of days for a patient observed in a database; (i,d) identifies their dth day of observation; j = 1,2,…J are J drugs of interest;

Simpson, Shawn E., et al, Biometrics 69.4 (2013): 893-902. Suchard, Marc A., et al. , Drug safety 36.1 (2013)

High-dimensional Propensity Score + New user cohort design



$$logit\big(P(Drug_A = 1)\big) = \alpha + \sum_{i=1}^{200} \delta_i Rx_i + \sum_{j=1}^{200} \gamma_j Dx_j$$

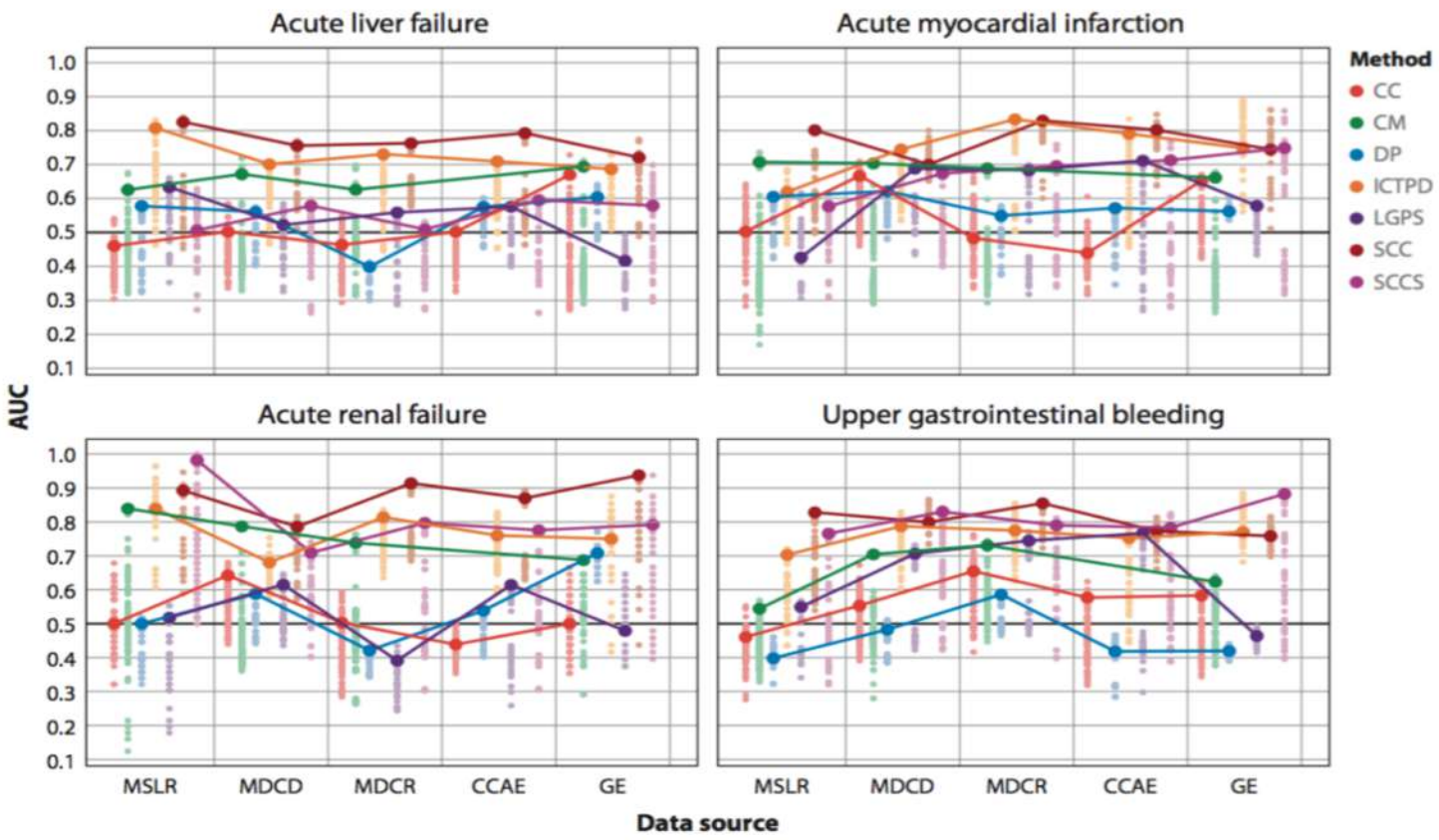
$$logit\big(P(Drug_B = 1)\big) = \alpha + \sum_{i=1}^{200} \delta_i Rx_i + \sum_{j=1}^{200} \gamma_j Dx_j$$

Parameters:

Washout period: 180 d;
Surveillance window: 30 d from exposure start; exposure+30d; all time from exposure start
Covariate eligibility window: 30 d prior to exposure
# of confounders: 100, 200, 500
Propensity strata: 5, 20 strata
Analysis strategy: Mantel-Haenszel stratification, propensity adjusted, propensity strata adjusted
Comparator cohort: drugs with same indication, not in same class; most prevalent drug with same indication, not in same class

Schneeweiss, Sebastian, et al. Epidemiology (Cambridge, Mass.) 20.4 (2009): 512.

A systematic statistical approach to evaluating evidence from observational studies



CC, case control;
CM, cohort method-propensity score method
DP, disproportionality analysis;
ICTPD, information component temporal pattern discovery;
LGPS, longitudinal gamma Poisson shrinker;
SCC, self-controlled cohort, observational screening
SCCS, self-controlled case series.
MSLR, MarketScan Lab Supplemental;
MDCD, MarketScan Multi-State Medicaid;
MDCR, MarketScan Medicare Supplemental Beneficiaries;
CCAE, MarketScan Commercial Claims and Encounters;
GE, GE Centricity;

Madigan, David, et al. Annual Review of Statistics and Its Application 1 (2014): 11-39.

Clinical Notes?

Input text

PAST MEDICAL/SURGICAL HISTORY: Positive for atrial fibrillation. The patient had AVR 6 years ago. Peripheral arterial disease with hypertension, peripheral neuropathy, atherosclerosis, hemorrhoids, proctitis, CABG, and cholecystectomy.

a

FAMILY HISTORY: Positive for atherosclerosis, hypertension, autoimmune diseases in the family.

REVIEW OF SYSTEMS: Weight loss of 25 pounds within the last 6 months, shortness of breath, constipation, bleeding from hemorrhoids, increased frequency of urination, muscle aches, dizziness and faintness, focal weakness and numbness in both legs, knees and feet.

LABORATORY DATA AND RADIOLOGICAL RESULTS: The patient had a chest x-ray, which showed cardiomegaly with atherosclerotic heart disease, pleural thickening and small pleural effusion, a left costophrenic angle which has not changed when compared to prior examination, COPD pattern. The patient also had a head CT, which showed atrophy with old ischemic changes. No acute intracranial findings.

DISCHARGE DIAGNOSIS: Syncope.

DISCHARGE MEDICATIONS: The patient was discharged on the following medications; Cardizem 90 mg p.o. thrice daily, digoxin 0.125 mg p.o. once daily, allopurinol 100 mg two times daily, Coumadin 4 mg p.o. q.h.s., and Remeron 15 mg p.o. q.h.s.

Natural Language Processing

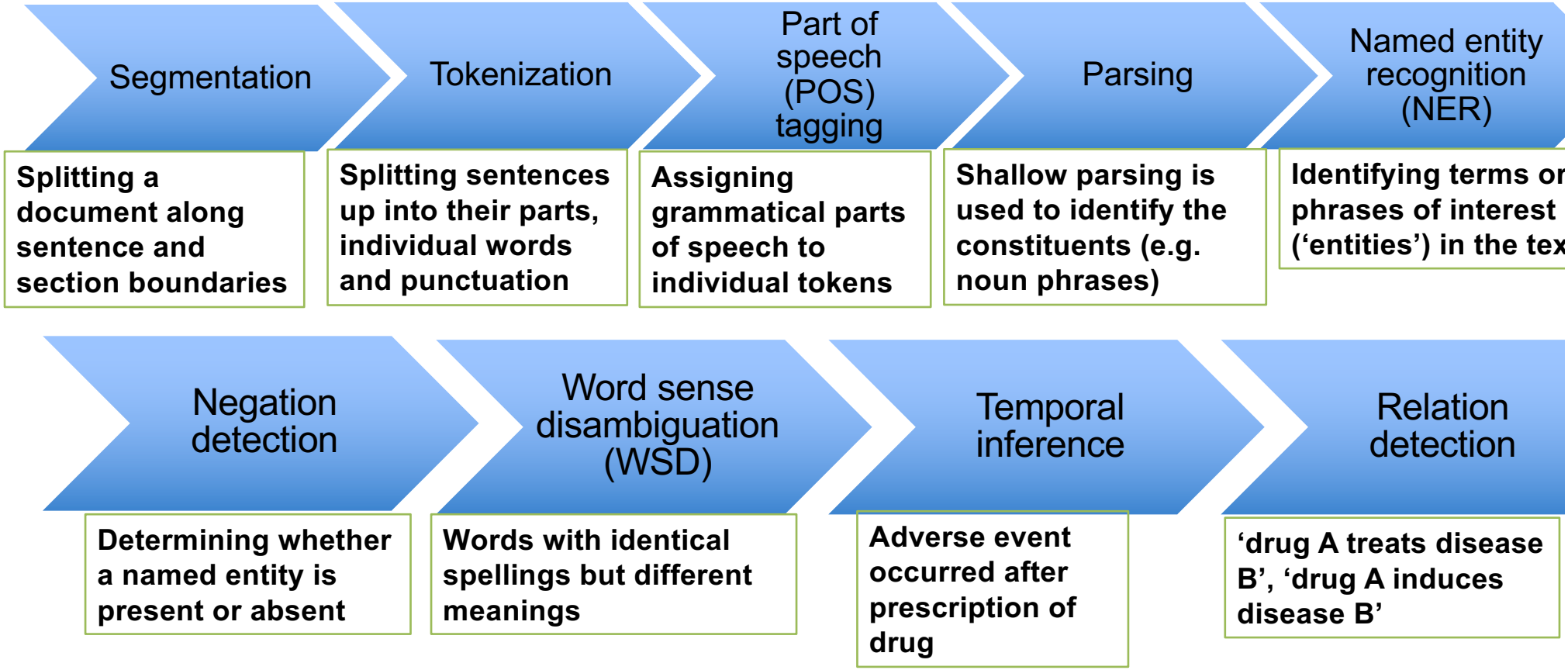
Biomedical Terminology System

"His past medical history is significant for asthma"

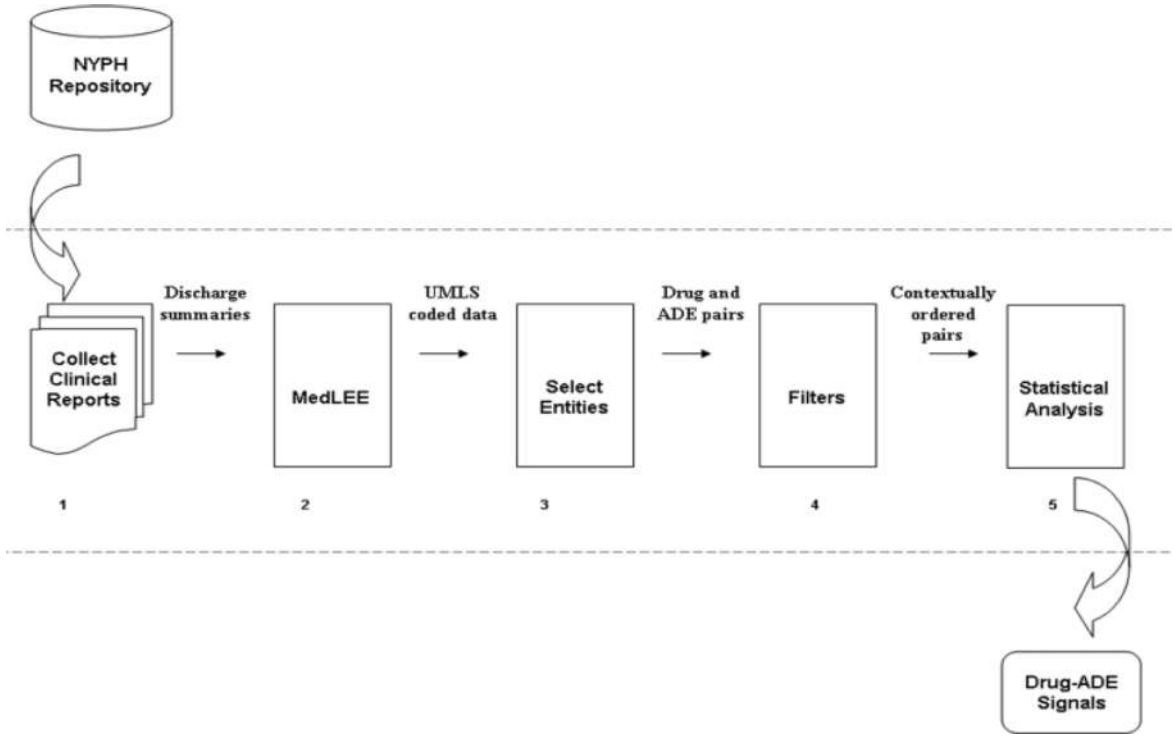
<**problem** v = "asthma" code = "UMLS:C0004096_asthma">
  <certainty v = "high certainty"></certainty>
  <parsemode v = "mode1"></parsemode>
  <sectname v = "report past history item"></sectname>
  <sid idref = "s2"></sid>
  <status v = "past history"></status>
  <**code** v = "**UMLS:C0004096_asthma**"></code>
</problem>

Wang, Xiaoyan, et al. Journal of the American Medical Informatics Association 16.3 (2009): 328-337.

Natural Language Processing



Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study



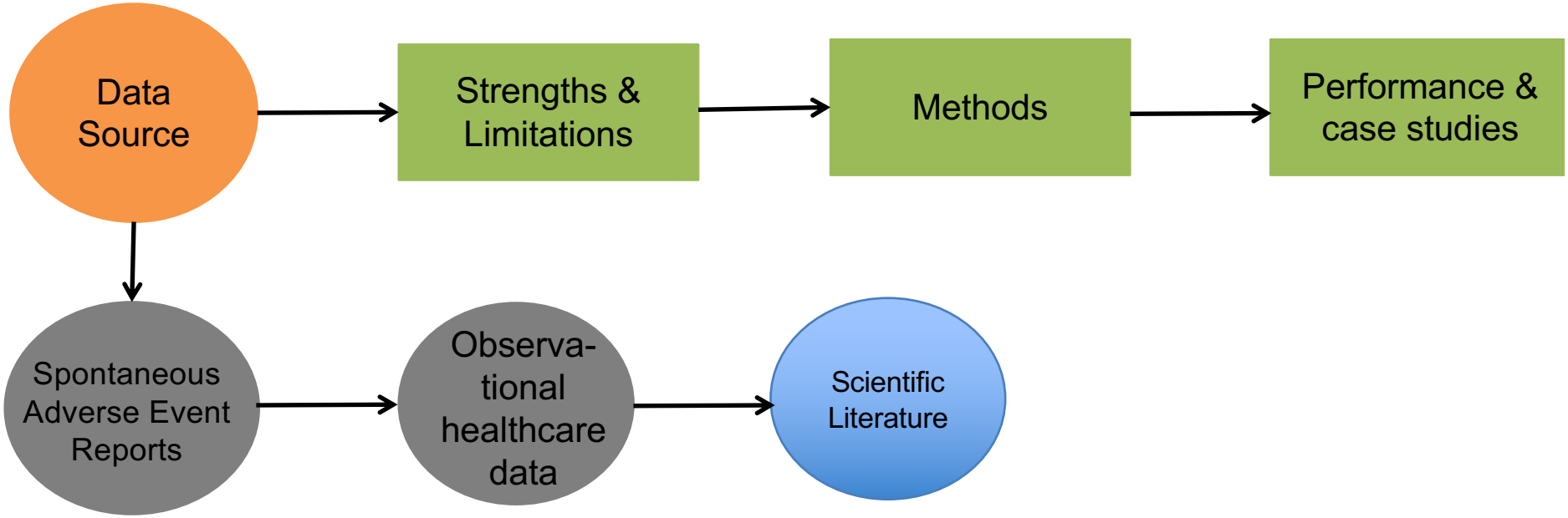
Drug	Total Documents	2 × 2 Tables*	Cutoff†
Ibuprofen	583	125	21
Morphine	490	128	22
Warfarin,	2040	189	10
Bupropion	188	124	32
Paroxetine	468	137	16
Rosiglitazone	287	119	10
ACE inhibitors	2482	257	14

ACE = angiotensin-converting enzyme.
*2 × 2 tables reflect number of potential drug-ADE associations for each drug.
†The cut-off represents the total number of potential drug-ADE associations selected as possible signals when ordered by $\epsilon(\chi^2)$.

Recall = 75%
Precision = 31%

Wang, Xiaoyan, et al. JAMIA (2009): 328-337.

Outline

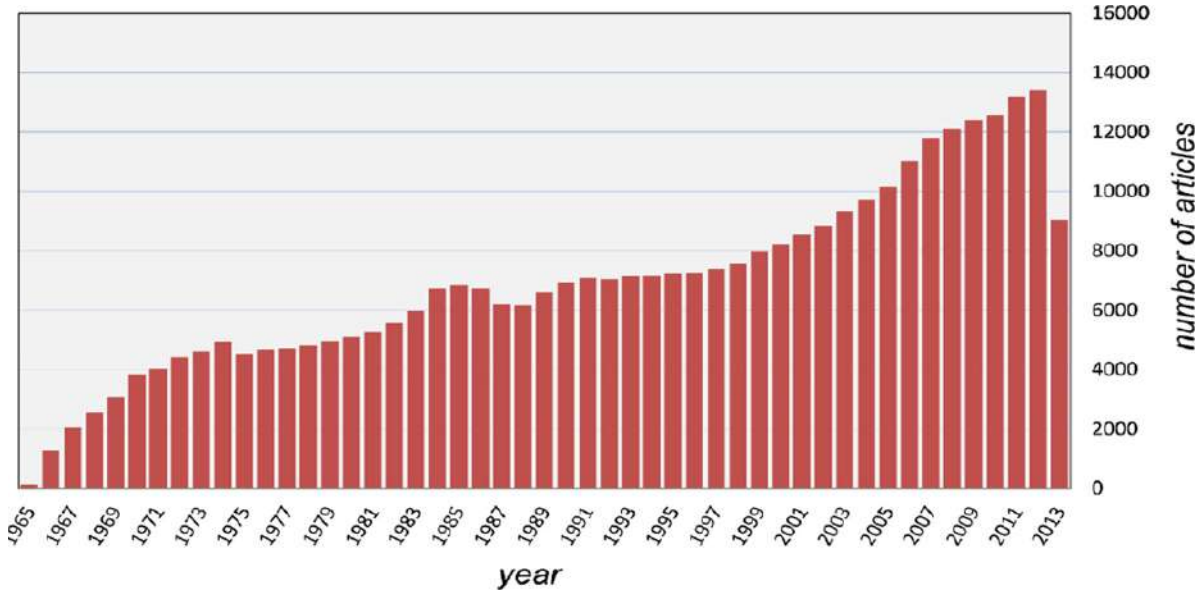

Data Source
Strengths & Limitations
Methods
Performance & case studies
Spontaneous Adverse Event Reports
Observa-tional healthcare data
Scientific Literature

Biomedical Literature

Subtypes
Research article
Review
Case study

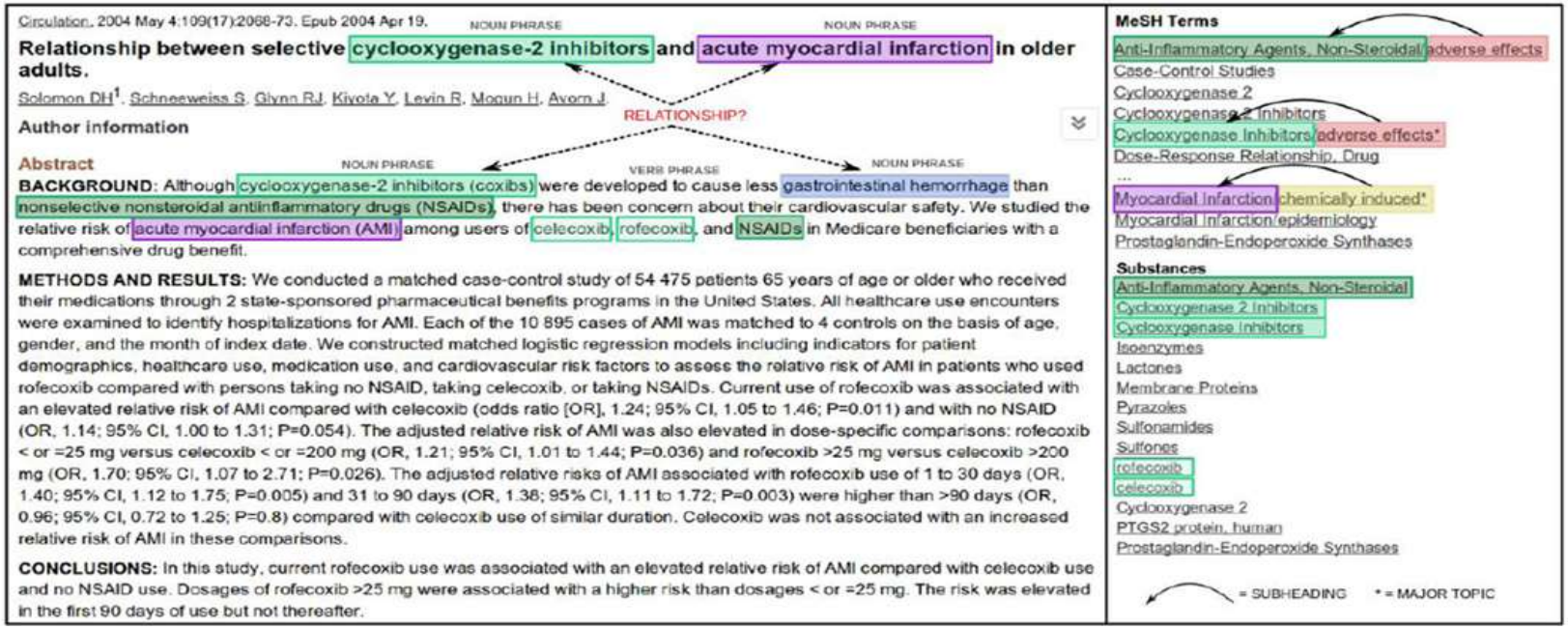
Strengths
Provide biological/physiological insights

Limitations
Delay for drug surveillance



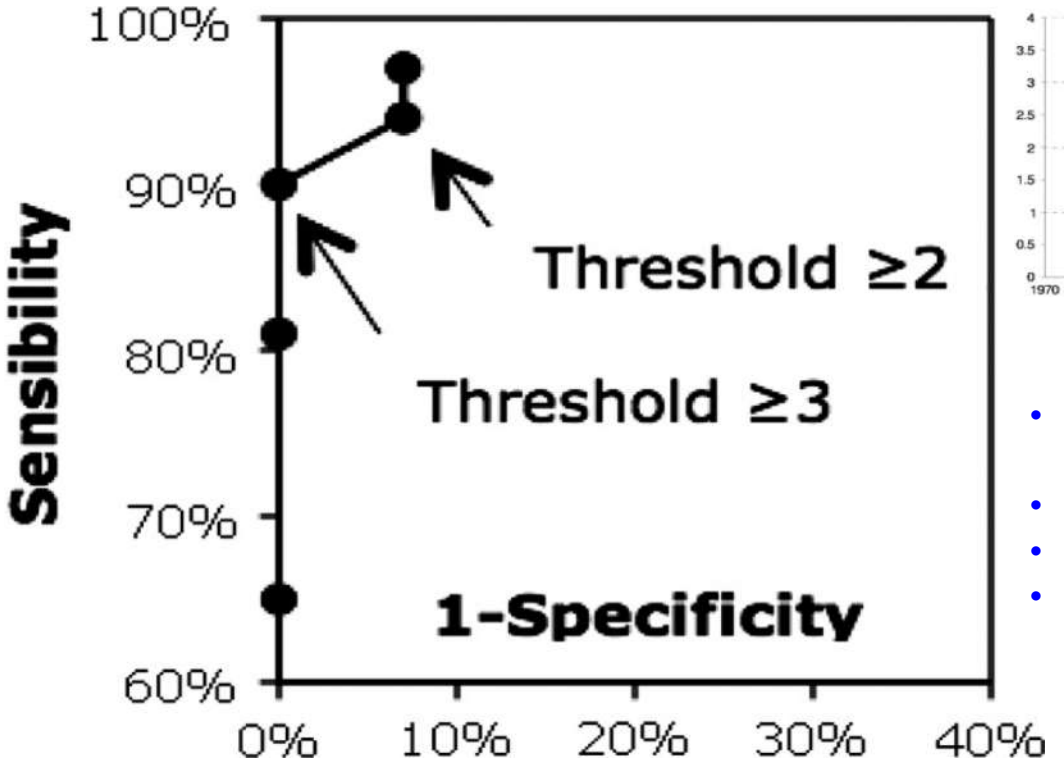
Data source	Amount of data	ADR Specific articles	Frequency
Medline	>26 million articles, all time	340,000	13,000 new ADR-related articles each year

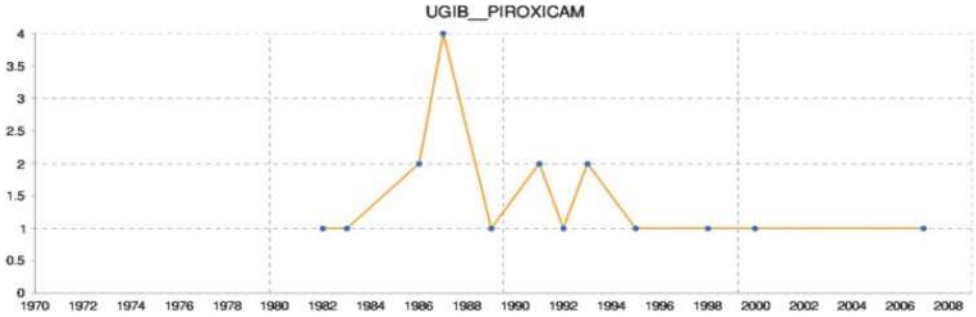
An example



NLM indexers select the most appropriate MeSH descriptors and subheadings (or qualifiers) to resume the full content of an article after reading the full text.

Design and validation of an automated method to detect known adverse drug reactions in MEDLINE



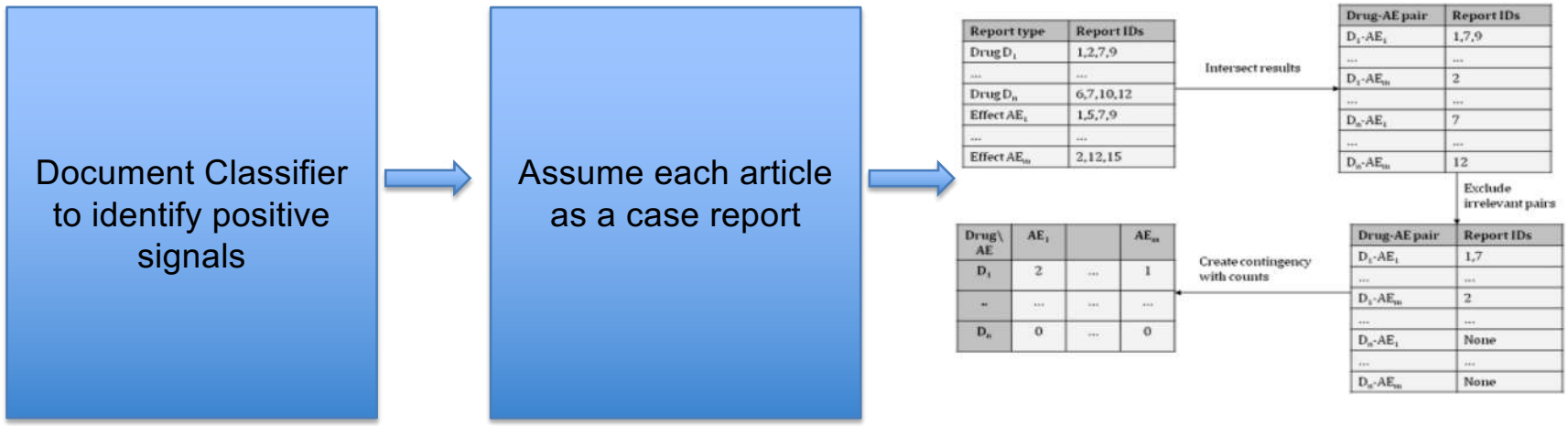


- Using a threshold of three or more publications containing adverse event and drug co-occurrences
- Sensitivity of 90%
- Specificity of 100%
- Precision of up to 93%.

Avillach, Paul, et al. , JAMIA (2013): 446-452.

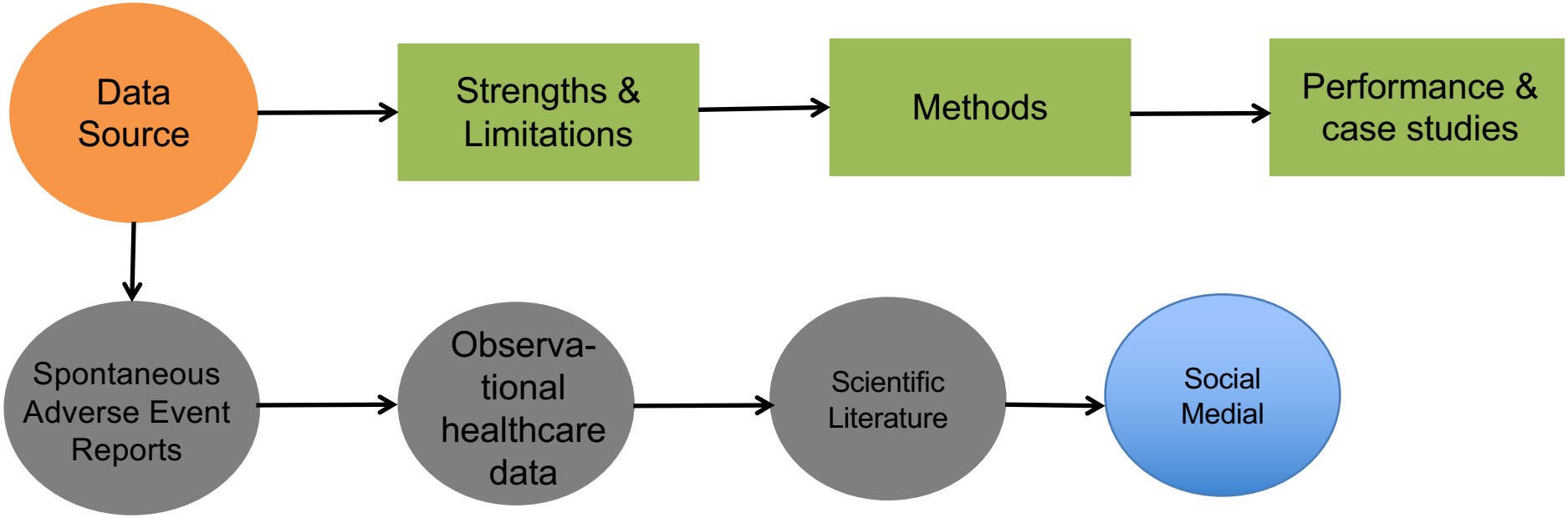
Using information mining of the medical literature to improve drug safety

Disproportionality analysis



Shetty, Kanaka D., and Siddhartha R. Dalal., JAMIA (2011)

Outline


Data Source
Strengths & Limitations
Methods
Performance & case studies
Spontaneous Adverse Event Reports
Observa-tional healthcare data
Scientific Literature
Social Medial

Social Media

Subtypes
- Patient web forums
- Twitter/facebook

Strengths
- Internet-based
- Patient-generated
- Unsolicited
- Up to date

Limitations
- Discrepancy in language (Non-medical, descriptive terms)
- Highly subjective, duplicates, hear-say information

twitter

facebook

Challenges

- No-medical, descriptive terms
  - Messed up my sleeping patterns -> sleep disturbance
  - Feeling need of deep breaths -> short of breath
- Complicated drug-condition relationship
  - Adverse effect: A reaction to the drug experienced by the patient, which the user considered negative
  - Beneficial effect: A reaction to the drug experienced by the patient, which the user considered positive
  - Indication: The condition for which the patient is taking the drug
  - Other: A disease or reaction related term not characterizable as one of the above

Complicated drug-condition relationship

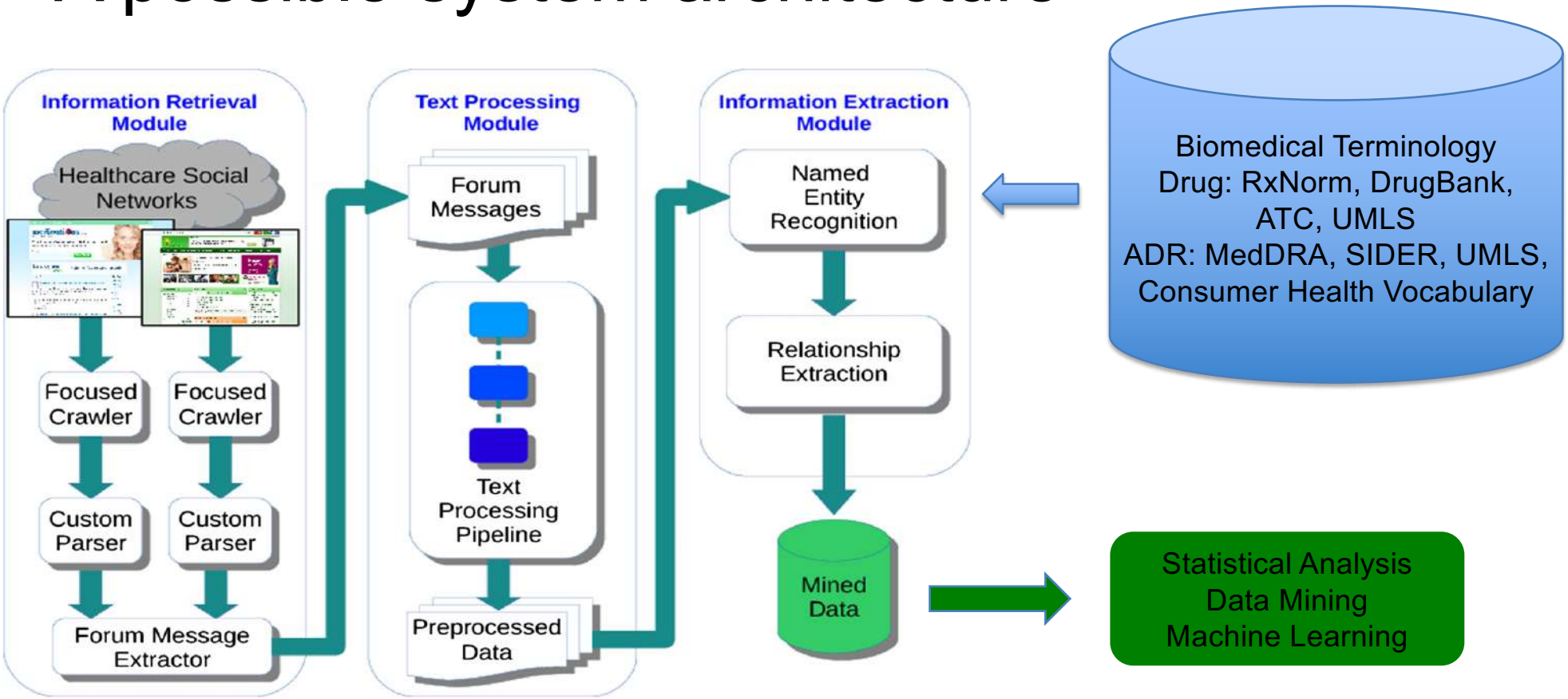
Sample Comments	Annotations
This has helped take the edge off of my constant sorrow. It has also perked up my appetite. I had lost a lot of weight and my doctor was concerned.	“constant sorrow” - depression: indication; “perked up my appetite” - appetite increased: beneficial effect; “lost a lot of weight” - weight loss: other
Works to calm mania or depression but zonks me and scares me about the diabetes issues reported.	“mania” - mania: indication; “depression” - depression: indication; “zonks me” - somnolence: adverse effect; “diabetes” - diabetes: other (hearsay)
Twitter Example: #Schizophrenia #Seroquel did not suit me at all. Had severe tremors and weight gain	“schizophrenia” – schizophrenia: indication; “tremors” – tremors: adverse effect; “weight gain” – weight gain: adverse effect

Leaman, Robert, et al. Proceedings of the 2010 workshop on biomedical natural language processing. ACL, 2010.

Challenges: Own experience or hearsay

Category	Example
Personal experience	I had memory problems with Simvastatin also to the point that I forgot where I was while driving.
An experience of a close family member or a friend	My step-dad was on Effexer, taking supplements for energy and drinking like a fish when he shot my daughter and me
Hearsay	There are more people out here having memory loss problems from statin drug that anyone can count.

A possible system architecture



Sampathkumar, Hariprasad, Xue-wen Chen, and Bo Luo. BMC medical informatics and decision making 14.1 (2014): 1

ADR Relation Extraction

- Co-occurance
  - Association rule mining
  - Disproportionality analysis
- Semi/supervised learning based approach
  - Hidden Markov Model
  - Conditional Random Field
    - POS, semantic type, word2vec, topic modeling

Case study: statins label change on 2012

Data (2003-2011)

Forum	No. of unique messages	No. of sentences	No. of unique usernames
medhelp.org	1,887	14,276	647
exchanges.webmd.com	5,492	32,693	854
healthboards.com	32,665	207,765	3,250
ehealthforum.com	1,042	7,150	562

#	Class Name	Head Drugs
1	Bile Acid Sequestrants	Welchol, Questran, Colestid
2	Cholesterol Absorption Inhibitors	Zetia
3	Fibric Acid Derivatives (Fibrates)	Tricor, Lopid, Trilipix, Atromid-S
4	Misc. Antihyperlipidemic Agents	Niacin, Vascepa, Choloxin, Juxtapid, Kynamro
5	Antihyperlipidemic Combinations	Vytorin, Advicor, Simcor, Caduet, Pravigard Pac, Juvisync, Liptruzet
6	Statins	Altoprev, Crestor, Lescol, Lipitor, Mevacor, Pravachol, Zocor, Livalo, Baycol

Relation Extraction

Drug-ADR in the same sentence

I took Lipitor and {I} suffered muscle weakness and memory loss.

Figure 1: An example of a MPR candidate. (Curly brackets denote an implicit word in the sentence.)

Drug-ADR in the adjacent sentence

My husband took statins for 9 years, the last one was Lipitor. Side effects included severe neck and shoulder pain, muscle atrophe, loss of muscle strength and both short term and long term memory loss

Figure 2: An example of a MPRE candidate.

Co-occurrence + filters

Feldman, Ronen, et al, KDD. ACM, 2015.

Case study: statins label change on 2012

Statistical Analysis

- *classic-induced lift*:

$$\frac{\Pr(\text{message has } D - S \text{ relation})}{\Pr(\text{message has D entity}) \times \Pr(\text{message has S entity})}$$

- *relation-driven lift*:

$$\frac{\Pr(D_0 - S_0 \text{ relation})}{\sum_i \Pr(D_i - S_0 \text{ relations}) \times \sum_i \Pr(D_0 - S_i \text{ relations})}$$

* Chi-square test statistics

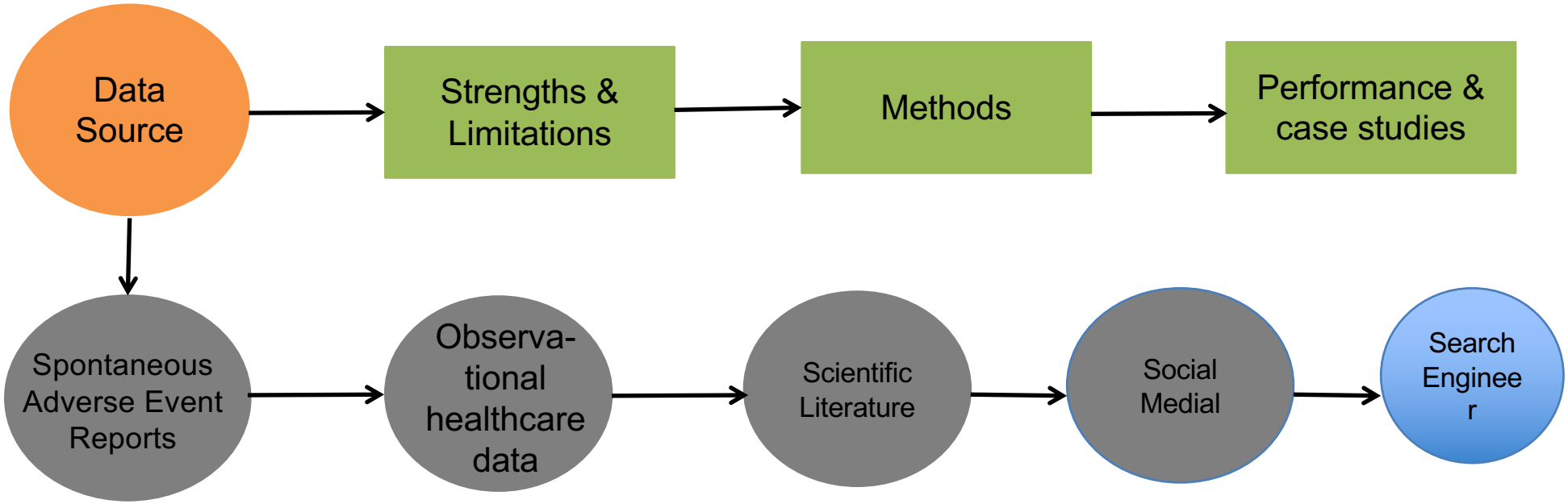
Results 1. Lifts and respective chi-square values preceded the relevant FDA label change

Year	Relation-driven lift	Chi-square value	Classic-induced lift	Chi-square value
2011	1.20	**13.33**	1.99	**49.28**
2010	1.21	**13.24**	1.94	**42.21**
2009	1.22	**13.35**	1.97	**40.03**
2008	1.21	**10.70**	1.89	**31.42**
2007	1.20	**9.95**	2.00	**36.46**
2006	1.21	**10.30**	1.89	**28.20**
2005	1.20	**6.63**	2.04	**25.12**
2004	1.25	3.46	2.18	**12.93**
2003	1.27	1.55	2.16	**5.79**

Results 2. Lifts and respective chi-square values

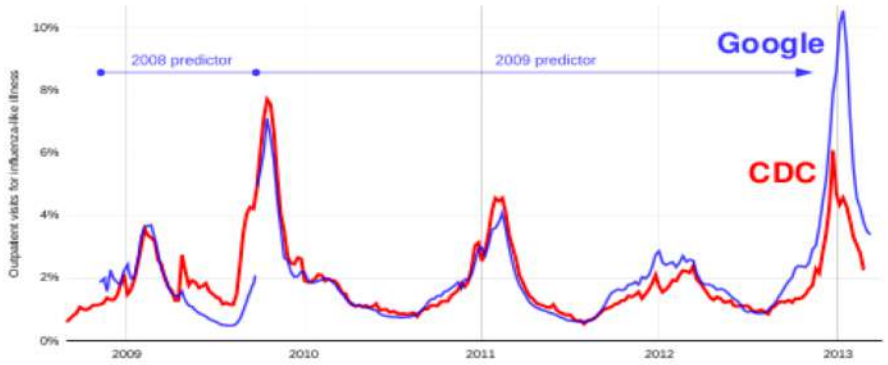
Relation-Driven Lift	1	2	3	4	5	6	1	2	3	4	5	6
pain	1.1	1.3	0.8	0.7	0.2	1.2	0.1	0.3	0.0	0.0	0.0	50.3
muscle pain	0.9	0.0	1.1	1.5	0.2	1.1	0.0	0.0	0.0	8.1	0.0	18.6
flushing	0.5	0.0	0.0	0.2	7.3	0.1	0.0	0.0	0.0	0.0	1207.0	0.0
heart attack	0.2	0.0	0.7	0.4	0.5	1.2	0.0	0.0	0.0	0.0	0.0	17.3
muscle damage	0.3	1.8	1.1	2.4	0.3	1.0	0.0	0.8	0.1	22.2	0.0	0.2
feeling weak	0.3	0.0	1.3	2.0	0.2	1.1	0.0	0.0	0.5	11.7	0.0	3.4
allergic reaction	1.0	2.4	0.5	1.1	1.2	1.0	0.0	1.8	0.0	0.0	1.2	0.0
liver failure	1.6	0.0	2.8	0.0	2.9	0.7	0.9	0.0	12.4	0.0	61.1	0.0
diabetes	2.5	3.0	0.6	0.8	1.3	0.9	5.9	2.8	0.0	0.0	1.1	0.0
cognitive impairment	0.4	0.0	1.3	0.3	0.2	1.2	0.0	0.0	0.3	0.0	0.0	13.3
leg pain	2.3	0.0	1.6	1.3	0.6	1.0	3.9	0.0	1.4	0.4	0.0	0.0
muscle problems	1.3	2.3	0.9	2.0	0.2	1.0	0.1	0.8	0.0	4.4	0.0	0.3
infection	0.6	0.0	0.9	0.3	1.0	1.1	0.0	0.0	0.0	0.0	0.0	1.2
leg cramps	2.0	4.8	1.0	1.1	0.2	1.1	1.6	6.2	0.0	0.0	0.0	0.9
muscle weakness	0.0	5.0	3.0	0.8	0.0	1.1	0.0	6.5	8.3	0.0	0.0	1.9

Outline


Data Source
Strengths & Limitations
Methods
Performance & case studies
Spontaneous Adverse Event Reports
Observa-tional healthcare data
Scientific Literature
Social Medial
Search Enginee r

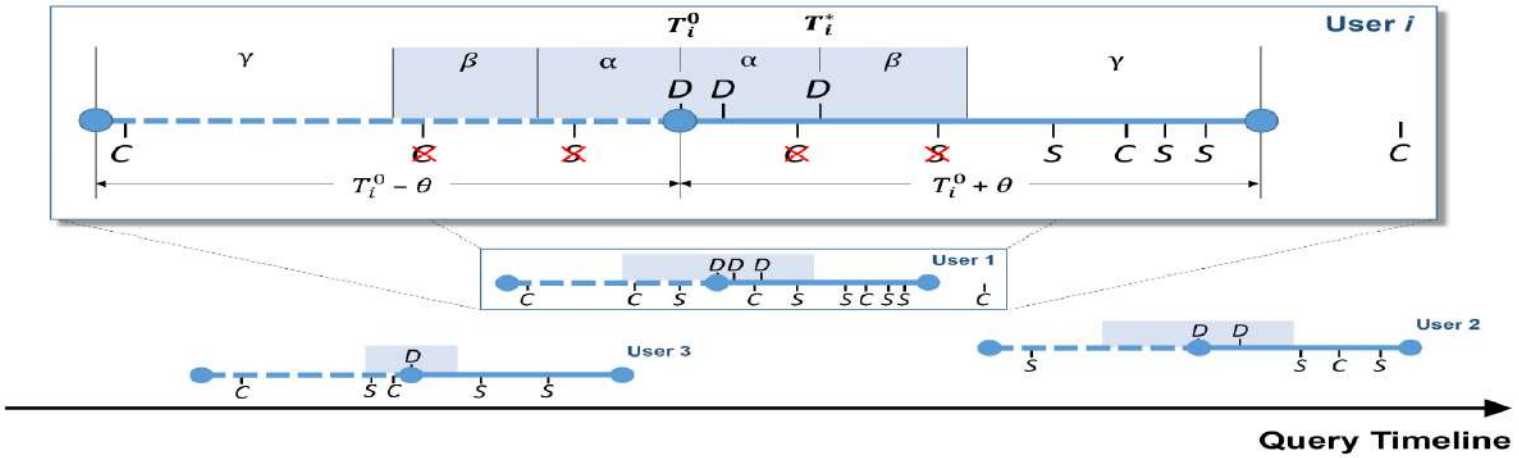
Search engine logs - Google Flu Trend

Second divergence in 2012–2013 for U.S.



Ginsberg, Jeremy, et al. Nature 457.7232 (2009): 1012-1014.

Side effect detection based on search engine logs



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

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

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

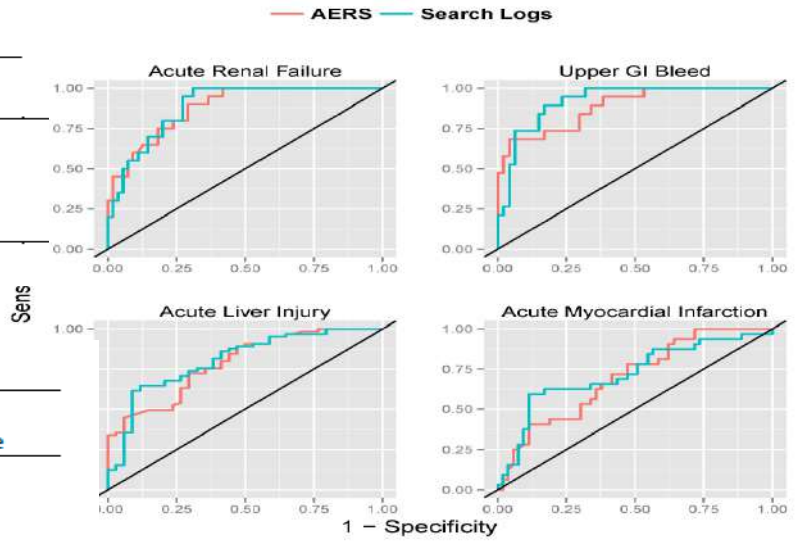
$$\frac{2N^-N^+ + Z_{\alpha/2}^2(N^- + N^+) \pm \sqrt{Z_{\alpha/2}^2(N^- + N^+)\left(4N^-N^+ + Z_{\alpha/2}^2(N^- + N^+)\right)}}{2(N^-)^2}$$

Ryen W, et al. **Clinical Pharmacology and Therapeutics** 2014. doi:10.1038/clpt.2014.77

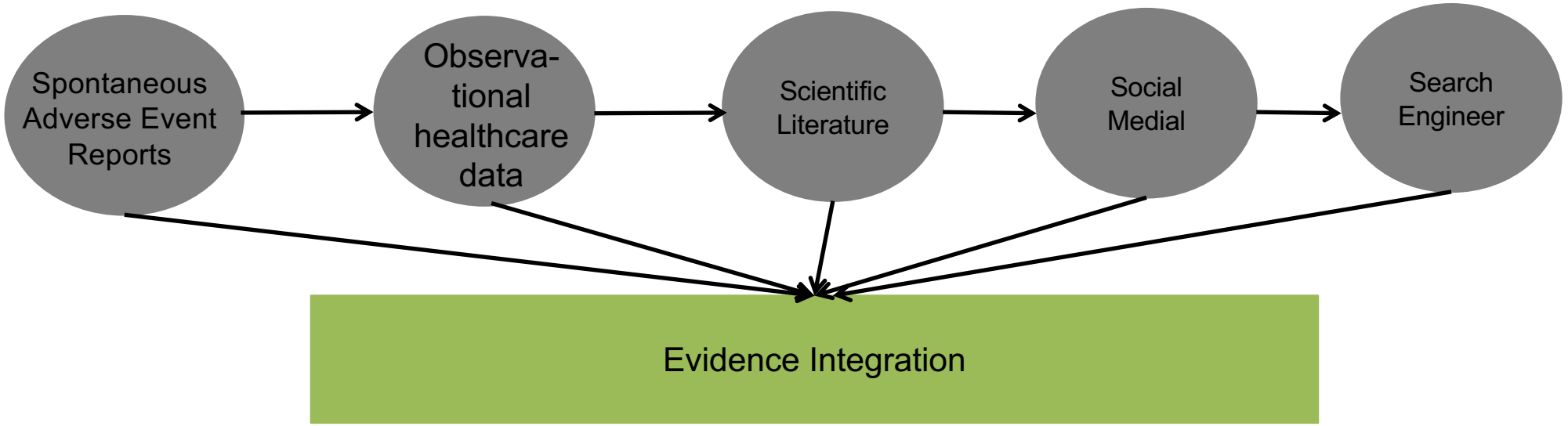
Comparison between FAERS and search log based signal detection

	Full AUC		
	AERS (EB05)	Search Logs (QRR05)	AUC difference
Acute Renal Failure	0.88	0.88	-4%
Upper GI Bleed	0.89	0.92	29%
Acute Liver Injury	0.79	0.81	12%
Acute Myocardial Infarction	0.70	0.73	9%
Average	0.81	0.83	11%

	Partial AUC at 0.3 FPR		
	AERS (EB05)	Search Logs (QRR05)	AUC difference
Acute Renal Failure	0.19	0.19	-2%
Upper GI Bleed	0.21	0.22	17%
Acute Liver Injury	0.14	0.16	10%
Acute Myocardial Infarction	0.10	0.14	19%
Average	0.16	0.18	12%



Evidence Integration

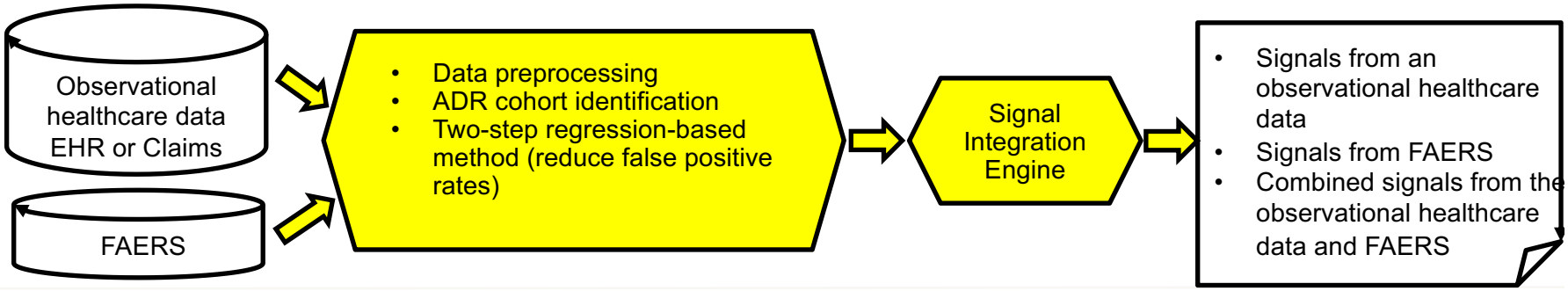


Literature review

- Combine SRS and search logs
- Combine SRS and literature
- Combine observational health data and literature
- Combine SRS and observational health data

ADR detection based on SRS and EHR/Claims

Methods



Results

AUCs of signal detection performance for FAERS, healthcare data and combined systems

	Combining FAERS and GE EHR		
ADR	FAERS	GE	Combined
Acute renal failure	0.91	0.68	**0.92**
Acute liver injury	0.71	0.63	**0.76**
Acute myocardial infarction	0.72	0.80	**0.82**
Upper GI bleeding	0.80	0.77	**0.87**
Total	0.76	0.76	**0.82**
	Combining FAERS and MarketScan claims		
ADR	FAERS	Claims	Combined
Acute renal failure	0.91	0.83	**0.93**
Acute liver injury	0.72	0.69	**0.79**
Acute myocardial infarction	0.71	0.77	**0.82**
Upper GI bleeding	0.81	0.83	**0.86**
Total	0.76	0.78	**0.82**

- Significant improvement over signal detection from single data source

* Evaluated based on known drugs which cause or do not cause the specific ADR
* Combined signals perform significantly better than signals acquired from each individual data source

Li, Ying, et al. Drug safety 38.10 (2015): 895-908.

Real world scenario

OHD

FAERS

	Positive	Negative
Positive	Exhibit in both sources	Appear in SRS but not in OHD
Negative	Appear in OHD but not in SRS	The lack of a signal in either source

GE EHR

FAERS

	Positive	Negative
Positive	NA (25/0)	0.73/0.78/**0.89** (29/11)
Negative	0.60/**0.68**/**0.68** (38/23)	0.71/0.69/**0.75** (61/152)

FAERS/GE/Combined AUCs (TP/TN)

Li, Ying, et al. Drug safety 38.10 (2015): 895-908.

Detecting Drugs that Could Possibly Cause Acute Myocardial Infarction (AMI)

- Drugs in red are known to cause AMI
- Drugs in green are known to not cause AMI
- None of the six drugs passed the signal threshold of <0.05 based on either EHR or FAERS
- Combined evidence from EHR and FAERS strength the signals with signal score <0.05

EHR based evidence

Drug	AMI Signal Score in EHR
amoxapine	0.118
diflunisal	0.192
eletriptan	0.072
nabumetone	0.494
nelfinavir	0.263
zolmitriptan	0.381

FAERS based evidence

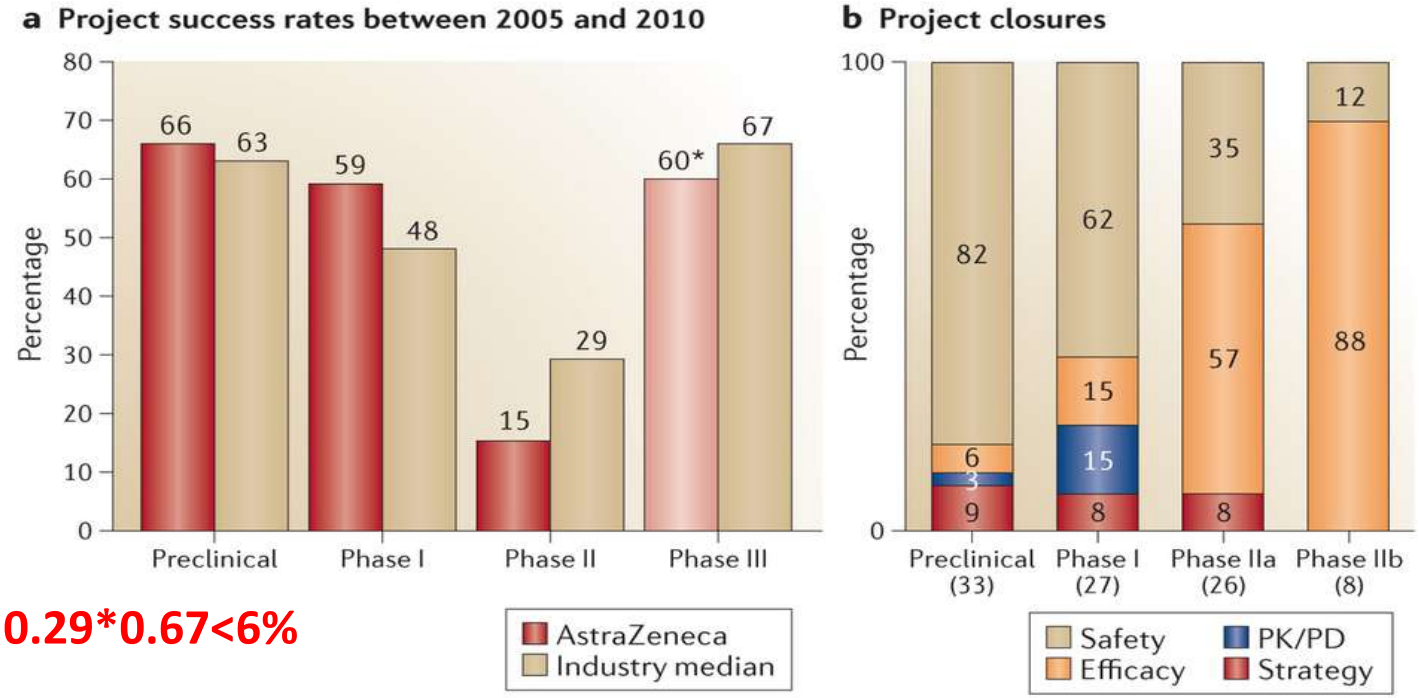
Drug	AMI Signal Score in FAERS
amoxapine	0.076
diflunisal	0.109
eletriptan	0.682
nabumetone	0.079
nelfinavir	0.292
zolmitriptan	0.224

Combined evidence

Drug	Combined AMI Signal Score
amoxapine	**0.007**
diflunisal	**0.007**
eletriptan	**0.034**
nabumetone	**0.035**
nelfinavir	**0.044**
zolmitriptan	**0.034**

ng, et al. Drug safety 38.10 (2015): 895-908.

Why drugs fail in clinical trial?

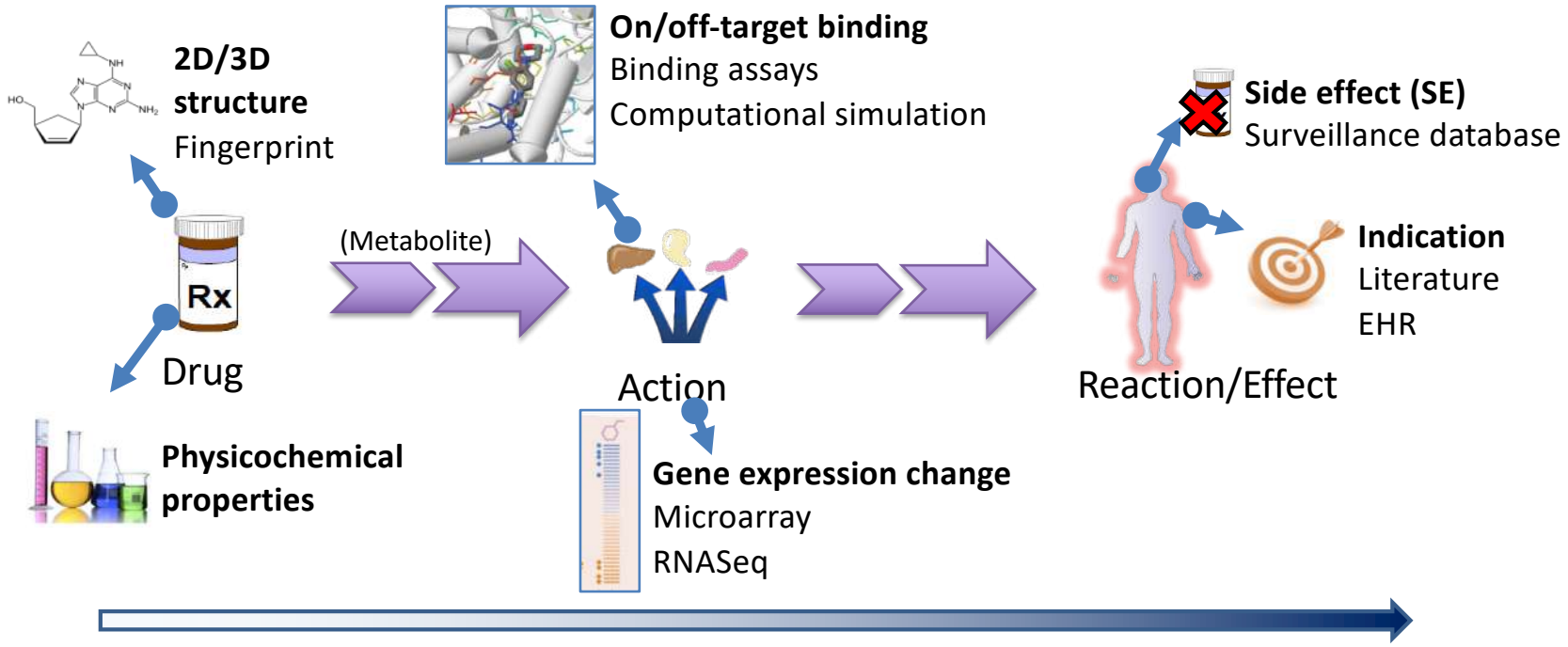


0.63*0.48*0.29*0.67<6%

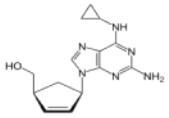
Cook D et al. "Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework." Nat Rev Drug Discov. 2014 Jun;13(6):419-431.

- Safety (toxicology or clinical safety) and efficacy (failure to achieve sufficient efficacy) are two major reasons for which a drug fails clinical trials.
- Can predictive modelling techniques help to generate hypothesis on efficacy and safety profiles of drugs?

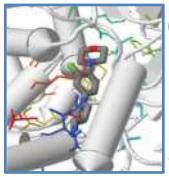
Pharmacology 101: A Simplified Path from Drug to Effect



Free big data in the domain



Drug structures
2,198 approved drugs
5,022 experimental drugs
(DrugBank)

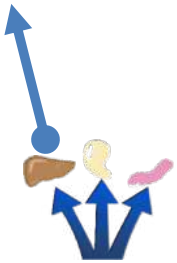


On/off-target binding
1,154,431 BioAssays (PubChem)
118,748 crystal structures (RSCB PDB)
551,193 reviewed protein sequences
62,148,086 not reviewed (UniProt)

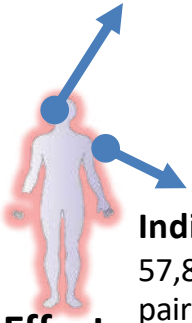
Side effect (SE)
5,868 side effects
139,756 drug-SE pairs (SIDER)
6,503,071 reports (FAERS)

Drug

(Metabolite)



Action



Effect

Indication
57,805 drug-indication pairs (NDF-RT)
215,433 clinical trials (ClinicalTrials.gov)
22,000,000+ articles (PubMed)

Chemical structures
89,124,716 compounds
219,712,379 substances
(PubChem)

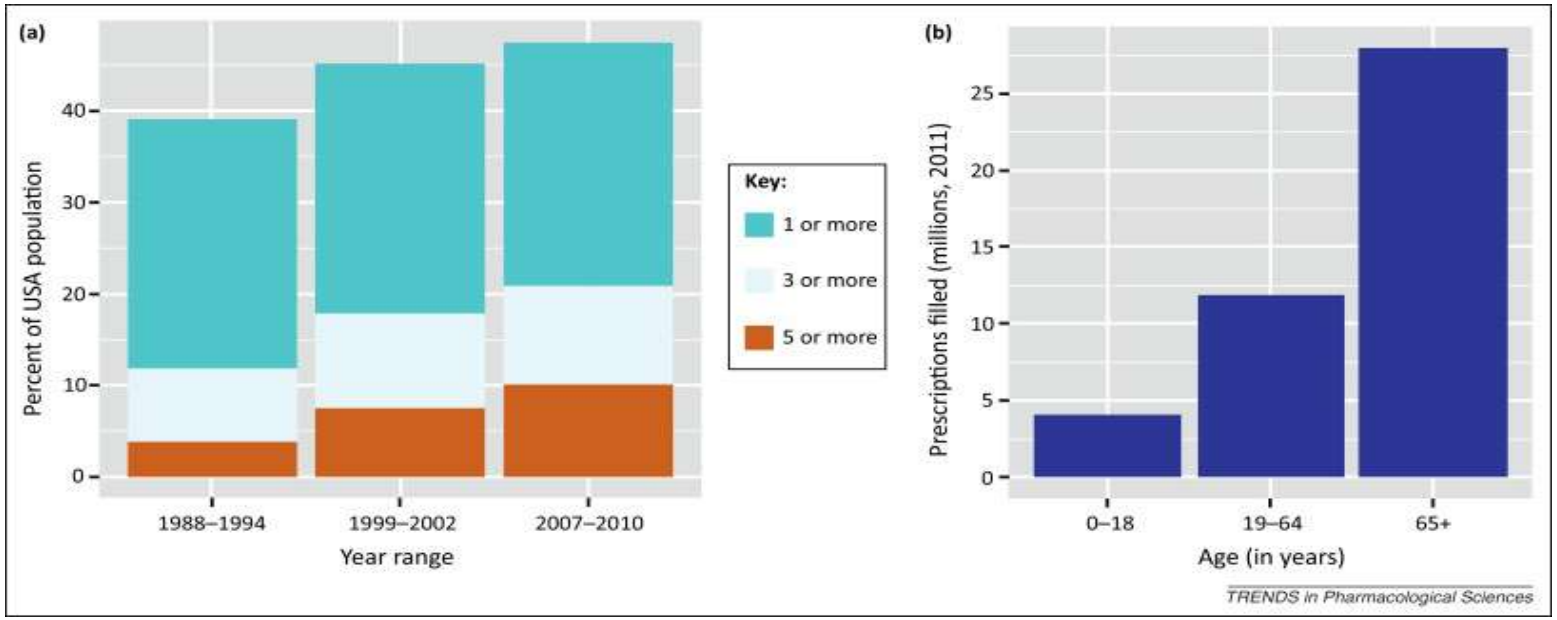
Gene expression change
3,775 human genomes (1000 genome)
15,819 sequencing platforms (GEO)
68,503 gene expression series (GEO)
1,801,592 gene expression samples (GEO)

By May 2016

From Surveillance to Prediction: A Few Case Studies

- Predicting drug-drug interactions through implementing the chemical-protein interactome
- Predicting drug-drug interactions through large-scale similarity-based link prediction
- Predicting drug repositioning opportunities through integrating multiple aspects of drug similarity and disease similarity

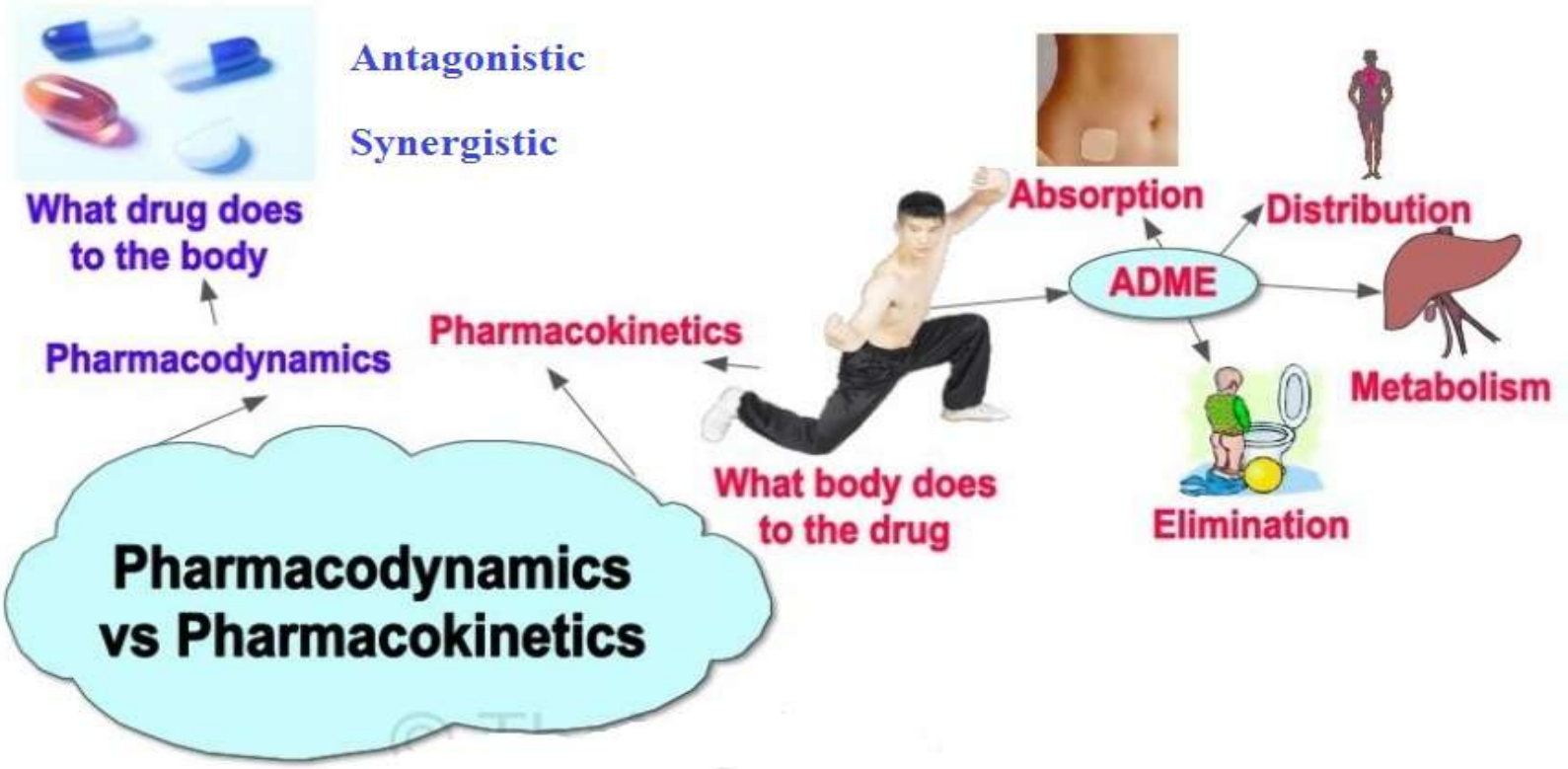
Statistics of Prescriptions in USA and Drug-Drug Interactions (DDIs)



- (a) Number of prescription drugs used in the past 30 days by percentage of the USA population
- (b) Average number of prescriptions filled in 2011 in the USA by age

- DDIs may happen unexpectedly when more than one drugs are co-prescribed, causing serious ADRs.
- DDIs are serious health threats that can result in significant morbidity and mortality - causing nearly 74,000 emergency room visits and 195,000 hospitalizations each year in the USA.

Pharmacokinetic (PK) and Pharmacodynamic (PD): Another Definition of DDIs



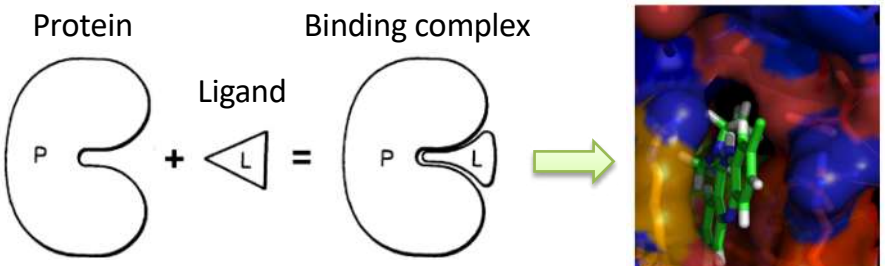
- PK and PD properties of one drug affect either the PK or PD of another drug

Types of DDIs

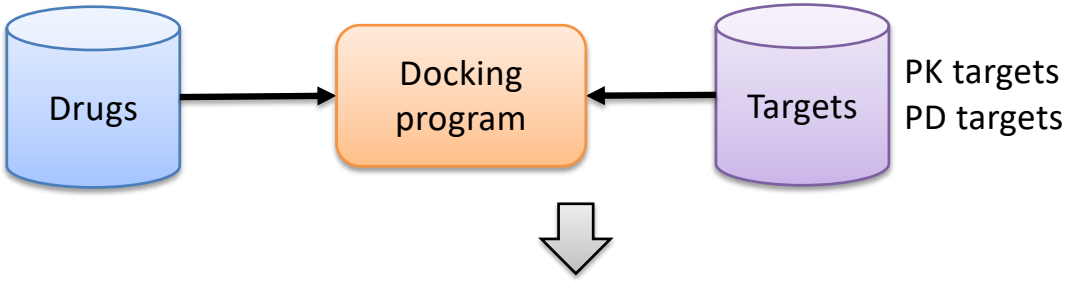
- Potentiation: Drugs with similar actions cause an additive effect. e.g.,
  - Coumadin and aspirin taken together cause excessive bleeding
  - Sedatives and alcohol cause excessive sedation
- Interference: One drug accelerates or slows the metabolism or excretion of another drug. e.g., Erythromycin taken with
  - Digoxin = elevated blood levels of digoxin
  - Coumadin = enhanced action of Coumadin
- Antagonism: One drug decreases the effectiveness of another drug because of divergent actions
  - Oral ketoconazole (Nizoral) is absorbed in an acidic environment
  - H2-receptor antagonists or proton pump inhibitors decrease acidity in the stomach
- Displacement: Two drugs compete for protein binding sites
  - One drug "wins" (is bound to protein)
  - Displaced drug is active in greater quantities
  - Same effect as taking a higher dose of the displaced drug!

A major cause of DDIs

Molecular docking and chemical-protein interactome (CPI)

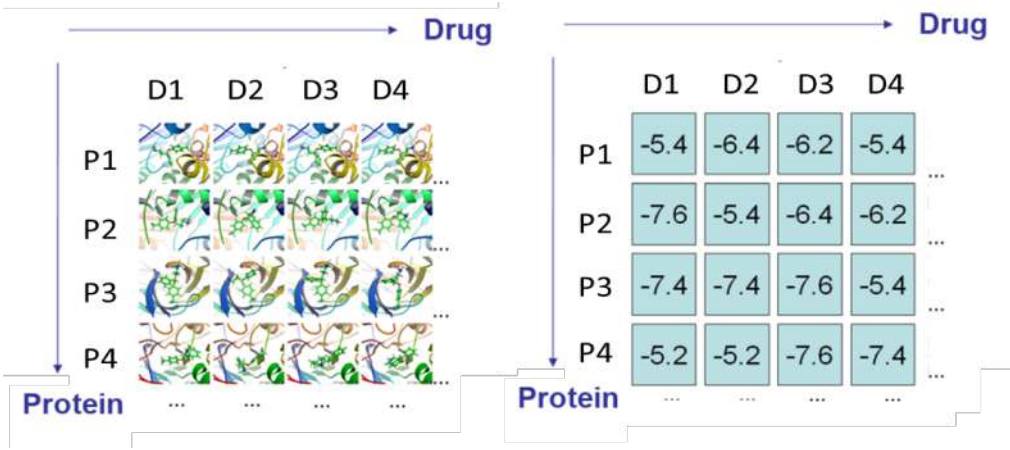


Drug binding inside the protein



Use AutoDock Vina to simulate the binding between a small molecule and a protein target.

- Provide the theoretical binding conformation (i.e., free energy) of the drug's binding to protein
- A lower docking score means a higher binding strength



Simulation of a CPI

Biological rationale of DDI-CPI

- Biological rationale
  - Competition between protein resources (e.g., metabolizing enzyme, transporter, or unexpected off-targets) can cause DDIs.
  - MOAs are simple in explanation, such as which PK/PD proteins may be involved in this DDI; and are there any comparable strong CPI for this protein.

- Preparation of the library drugs and targets
  - 2515 library drug molecules (85% are FDA approved drugs)
  - 611 representative collection of PK/PD proteins (239 human PK proteins and 372 PD proteins)

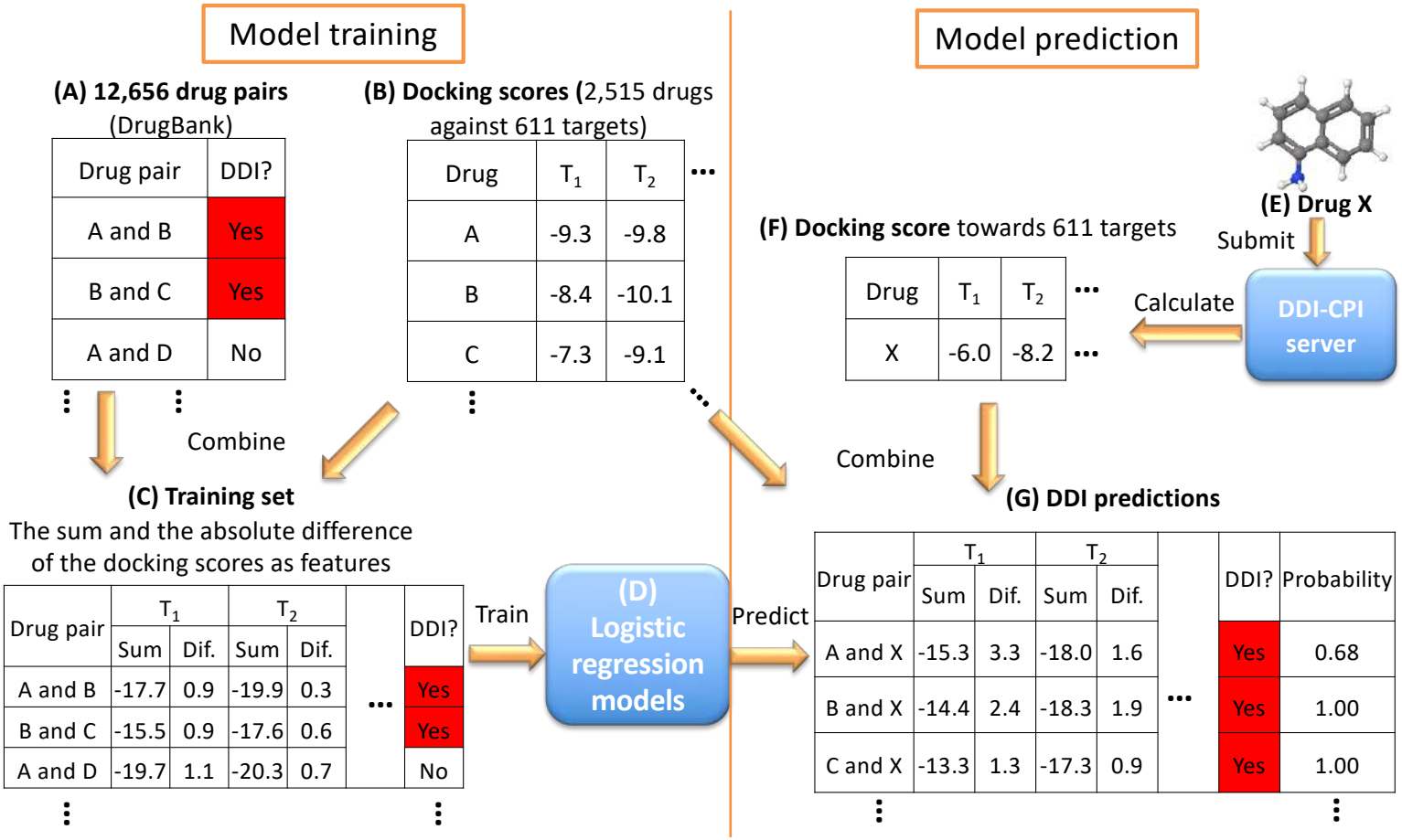
PK proteins: PDB with all available metabolite enzymes

PD proteins: PDBBind database with binding pocket information

- all proteins have X-ray crystal structures
- all structures have better resolution than 3.4 A
- binding pockets were identified around the embedded ligands in the crystal structure

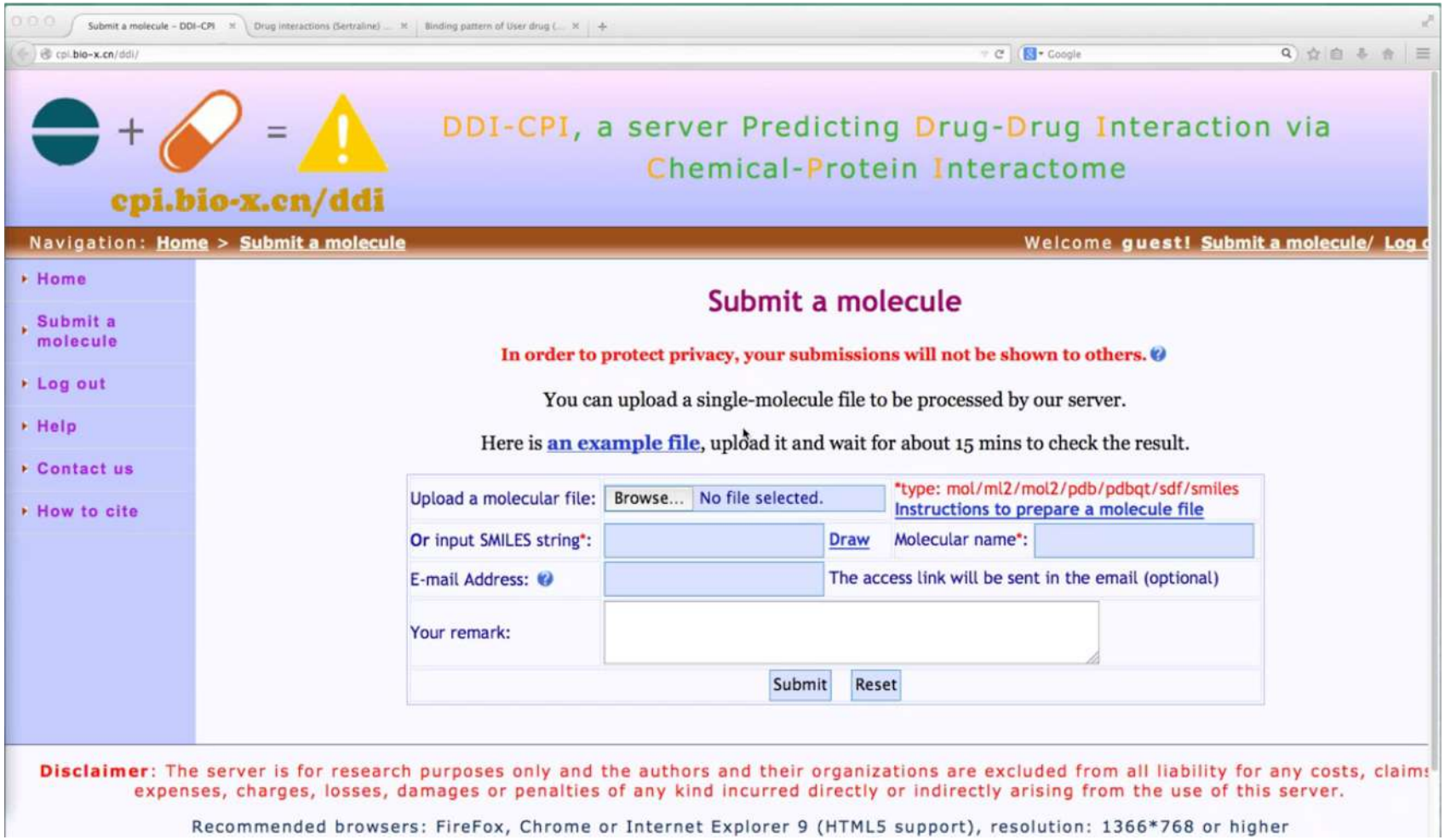
239 PK proteins and 372 PD proteins

Workflow of DDI-CPI server

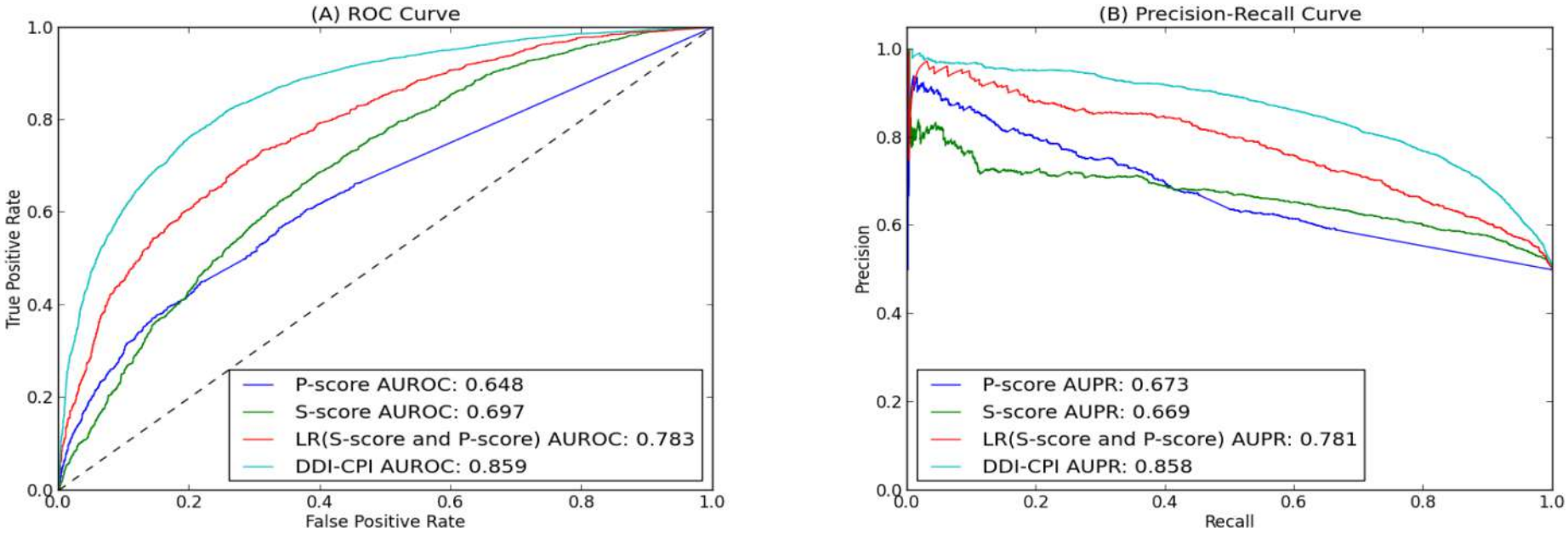


Luo#, Zhang#, et al. DDI-CPI, a server that predicts drug-drug interactions through implementing the chemical-protein interactome. *Nucleic acids res.* (2014): gku433

Demo: DDI-CPI


Submit a molecule - DDI-CPI
cpi.bio-x.cn/ddi/
cpi.bio-x.cn/ddi
DDI-CPI, a server Predicting Drug-Drug Interaction via Chemical-Protein Interactome
Navigation: Home > Submit a molecule
Welcome guest! Submit a molecule/ Log
Home
Submit a molecule
Log out
Help
Contact us
How to cite
Submit a molecule
In order to protect privacy, your submissions will not be shown to others.
You can upload a single-molecule file to be processed by our server.
Here is an example file, upload it and wait for about 15 mins to check the result.
Upload a molecular file:
Browse...
No file selected.
*type: mol/ml2/mol2/pdb/pdbqt/sdf/smiles
Instructions to prepare a molecule file
Or input SMILES string*:
Draw
Molecular name*:
E-mail Address:
The access link will be sent in the email (optional)
Your remark:
Submit
Reset
Disclaimer: The server is for research purposes only and the authors and their organizations are excluded from all liability for any costs, claims, expenses, charges, losses, damages or penalties of any kind incurred directly or indirectly arising from the use of this server.
Recommended browsers: FireFox, Chrome or Internet Explorer 9 (HTML5 support), resolution: 1366*768 or higher

Model evaluation and comparison



The ROC and precision-recall curve comparison for different DDI prediction methods based on independent validation

P-score: uses side-effect similarities to predict target sharing (Campillos, et al. Science (2008), 321, 263-266.)
S-score: uses drug-target network to predict DDIs (Huang, et al. PLoS Comput Biol (2013), 9, e1002998)
LR(S-score and P-score): integrates P-score and S-score by a Bayesian probabilistic model
DDI-CPI: predicts DDI using machine learning models via CPI

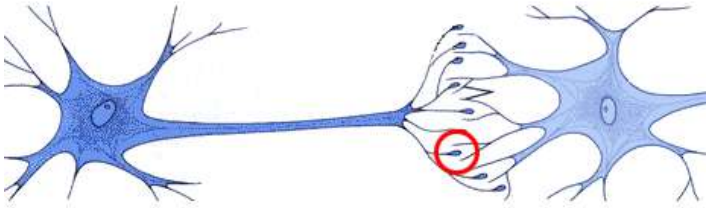
MAOI: Monoamine oxidase inhibitor
SSRI: Selective serotonin reuptake inhibitor

Case study - MAO-A inhibitors

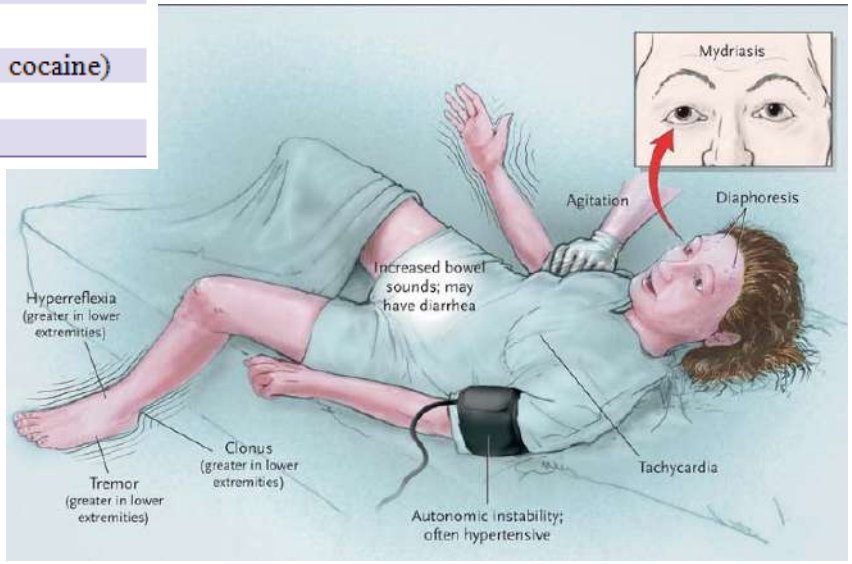
Table 3 (adapted from reference 6,7)

Drugs to Avoid When Taking MAOIs

Amphetamines	Bupropion
Cyclobenzaprine	Dextromethorphan
Linezolid	Meperidine
Methadone	Mirtazapine
SSRIs/SNRIs	TCA's
Triptans	Tramadol
Vasoconstrictors (psuedoephedrine, phenylephrine, cocaine)	
Chlorpeniramine, brompheniramine	
St. John's Wort	General anesthesia



- SSRI with MAOI results in high extracellular serotonin (5-HT) concentration – serotonin syndrome.



Source: pharmacytimes.org, Terry Gotham, dancesafe.org

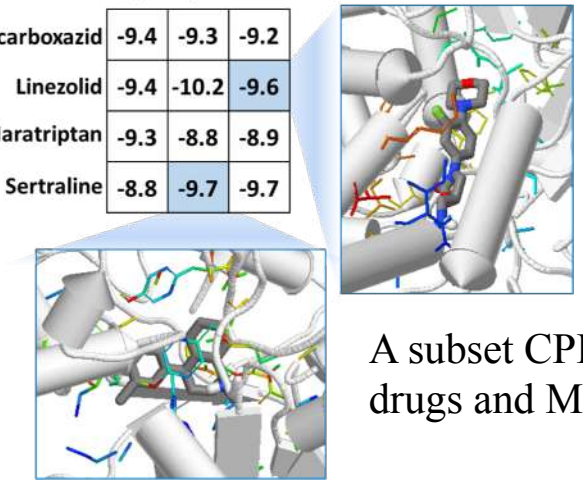
Case study - MAO-A inhibitors

Monoamine oxidase inhibitors (MAOIs)

Selective serotonin reuptake inhibitor (SSRI)

MAO-A targets

	2BXR	2Z5X	2Z5Y
Isocarboxazid	-9.4	-9.3	-9.2
Linezolid	-9.4	-10.2	-9.6
Naratriptan	-9.3	-8.8	-8.9
Sertraline	-8.8	-9.7	-9.7



A subset CPI visualization between drugs and MAO-A targets

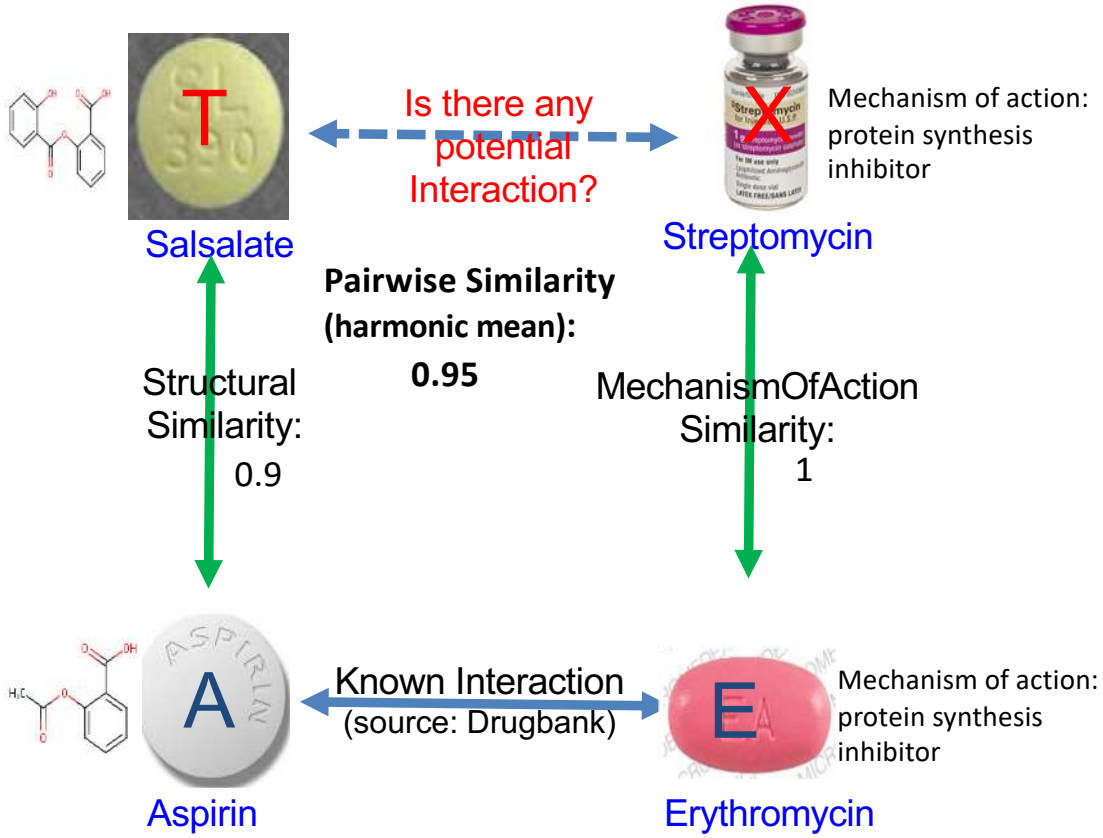
- The server predicts that sertraline may interact with isocarboxazid, linezolid, and naratriptan
- All of the predicted drugs can rank the monoamine oxidase A (MAO-A) targets to the top 20% – possible mechanism suggested

From Surveillance to Prediction: A Few Case Studies

- Predicting drug-drug interactions through implementing the chemical-protein interactome
- Predicting drug-drug interactions through large-scale similarity-based link prediction
- Predicting drug repositioning opportunities through integrating multiple aspects of drug similarity and disease similarity

Similarity-based Drug-Drug Interaction (DDI) Predictions

- Inspired from content-based recommender systems: Predict the existence of an DDI through comparisons with known DDIs
- Drug T might interact with drug X based on T's similarity to drug A and X similarity to drug E:
  - A-E already known to interact
- Limitation of prior arts
  - Skewed distribution
  - Appropriate evaluation metrics
  - Incompleteness of similarity measures



Overview of DDI-SIM

Ingest

Drug1	Drug2	Sim
Salsalate	Aspirin	.9
Dicoumarol	Warfarin	.7

sim_N ⋮

Drug1	Drug2	Sim
Salsalate	Aspirin	.7
Dicoumarol	Warfarin	.6

Known DDIs

Drug1	Drug2
Aspirin	Gliclazide
Aspirin	Dicoumarol

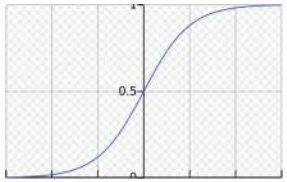
Build features

Candidate Features

Drug1	Drug2	Feature Vector
Salsalate	Gliclazide	[.9, .., .7]
Salsalate	Warfarin	[.7,..., .4]

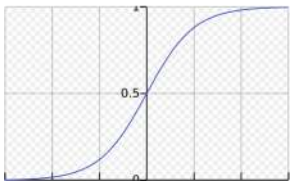
Build adjusted logistic regression models

Logistic Regression Model (λ_1)



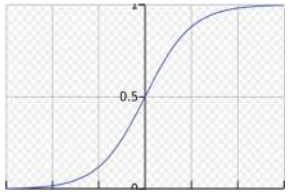
...

Logistic Regression Model (λ_8)



Select model & threshold

Logistic Regression Model (λ, η)



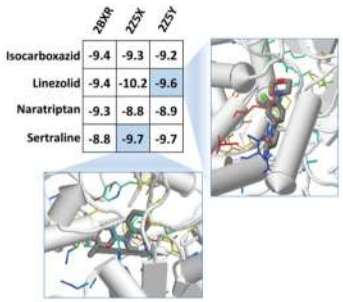
Predict

DDI Predictions

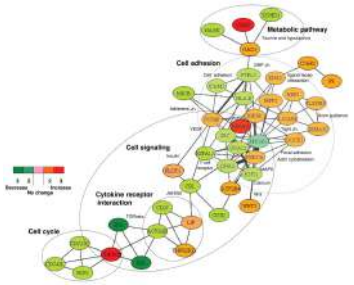
Drug1	Drug2	score
Salsalate	eltrombopag	0.98
Salsalate	colesevelam	0.94

Fokoue, A., Sadoghi, M., Hassanzadeh, O., Zhang, P. Predicting Drug-Drug Interactions through Large-Scale Similarity-Based Link Prediction. ESWC, 2016.

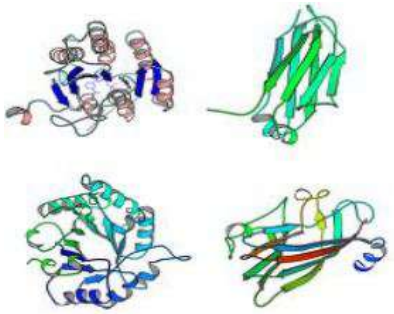
13 Drug Similarity Measures



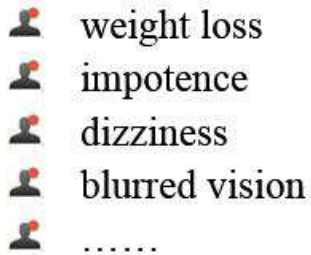
Chemical-Protein Interactome (CPI)



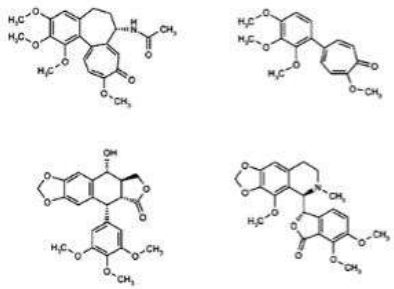
Pathway



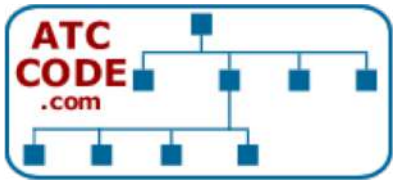
Target



Side Effects



Molecular Structure

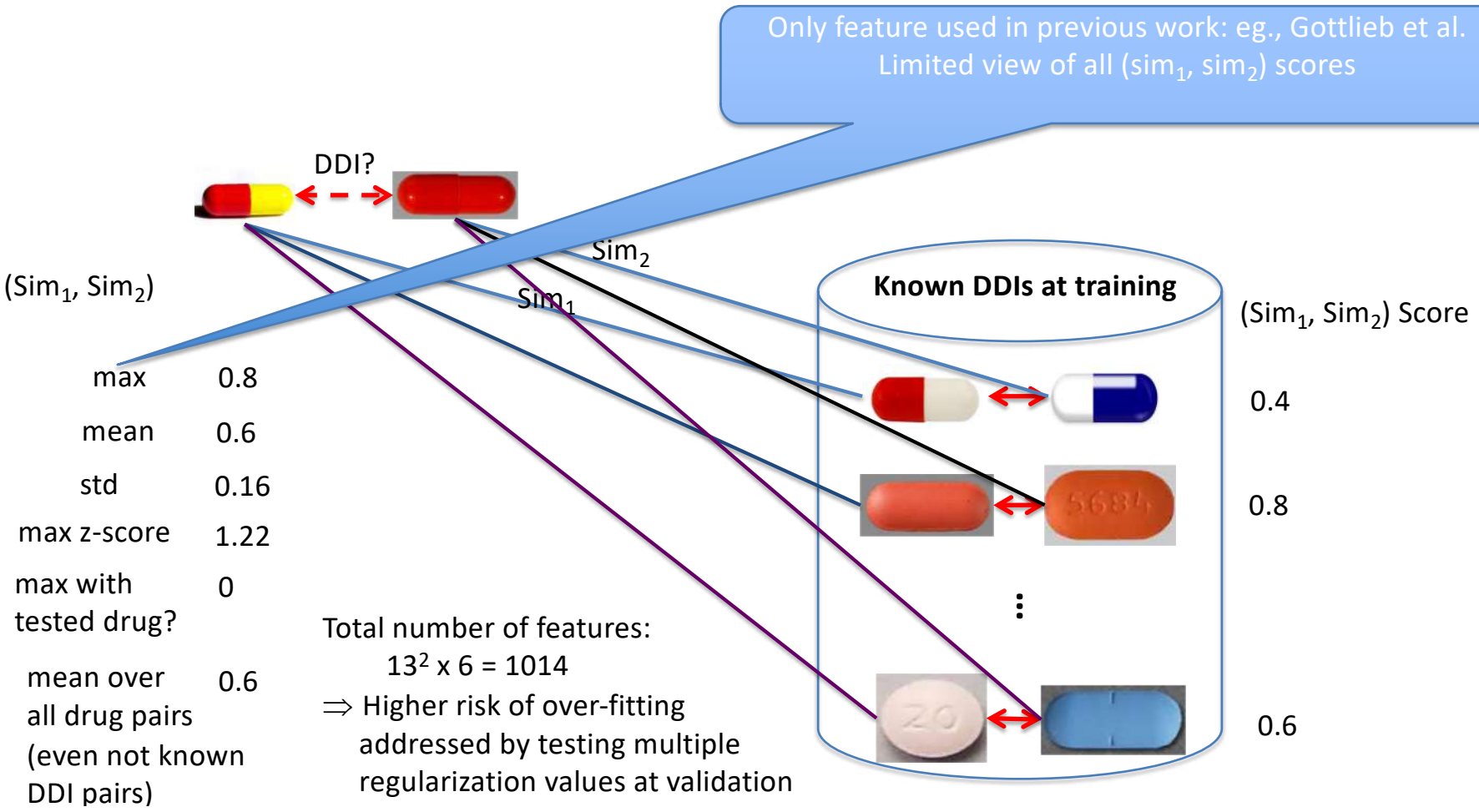


Therapeutic classification

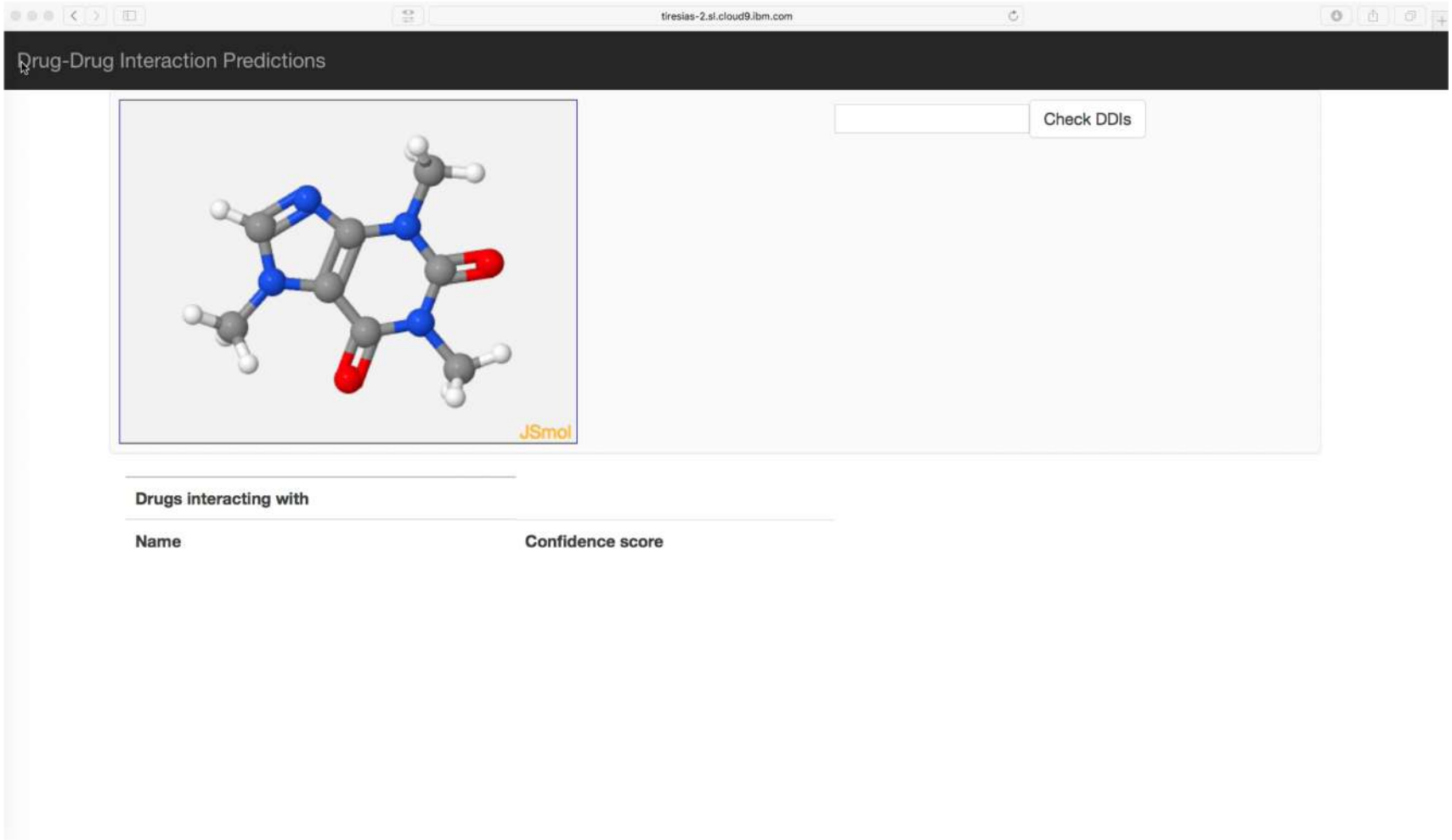
And others such as:

- Mechanism of Action
- Physiological Effect
- Metabolizing Enzyme
- MeSH term
- ……

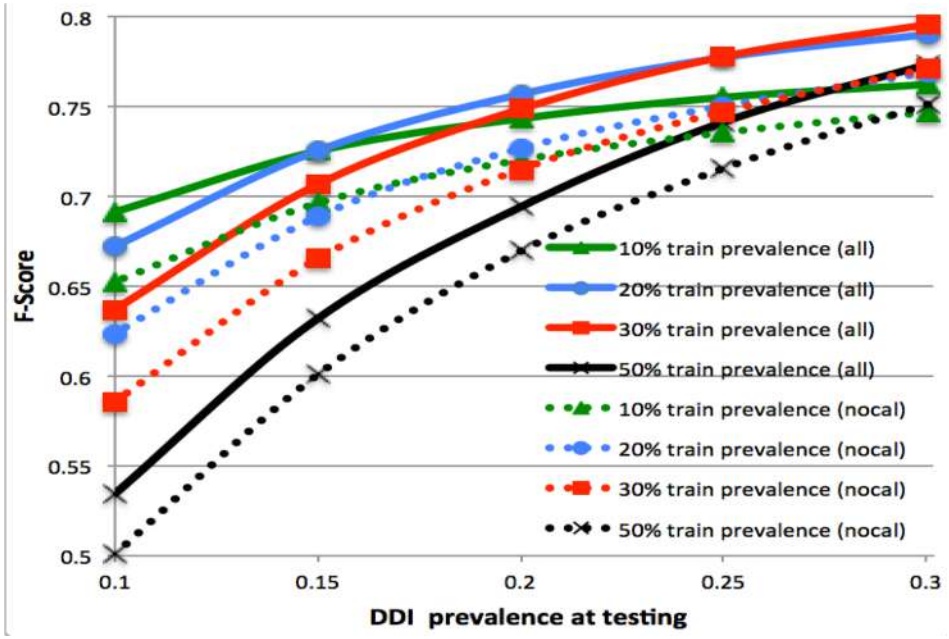
Feature Generation

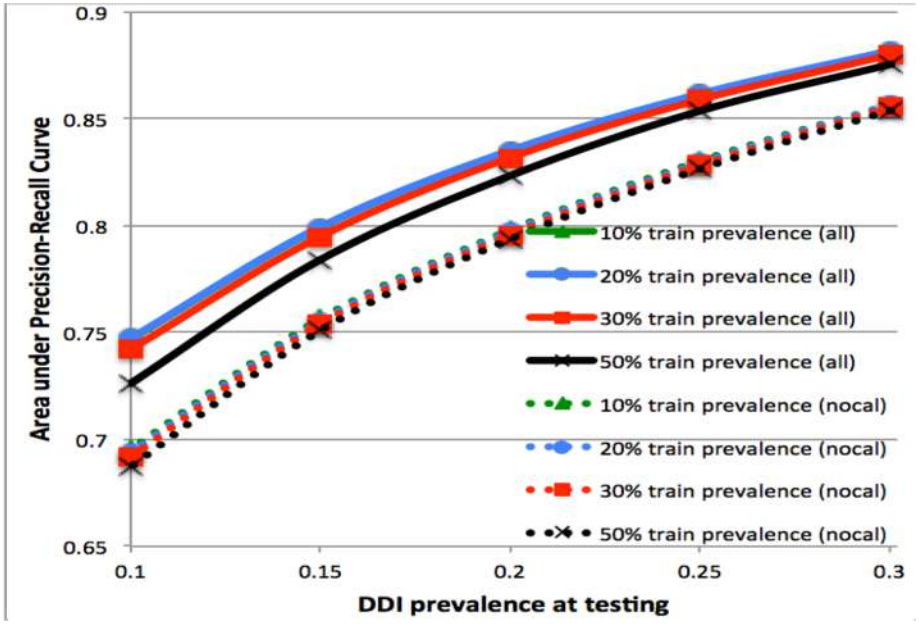


Demo: DDI-SIM


tiresias-2.sl.cloud9.ibm.com
Drug-Drug Interaction Predictions
Check DDIs
JSmol
Drugs interacting with
Name
Confidence score

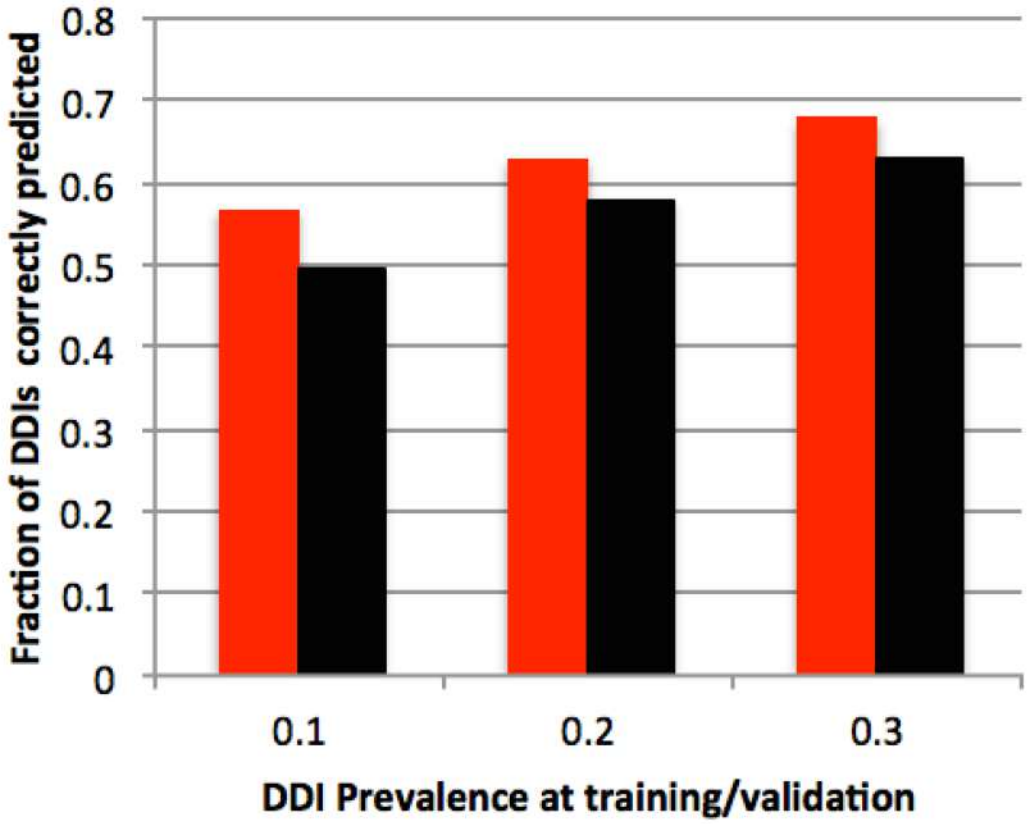
Experimental Evaluation: 10-fold cross validation





1) Using calibration features and unbalanced training/validation data significantly outperforms the baseline
2) For a fixed DDI prevalence at training/validation, using calibration features is always better
3) No similarity measure by itself has good predictive power (ATC is the best with 0.58 F-Score and 0.56 AUPR), removing any given similarity measure has limited impact on the quality of the predictions (< 1% decrease)

Experimental Evaluation: Retrospective Analysis (Predicting new DDIs in DrugBank 4.0 based on DrugBank 3.0)



Predictions using only known DDIs as of 2011

Predict up to 68% of DDIs found after 2011

with calibration

no calibration

From Surveillance to Prediction: A Few Case Studies

- Predicting drug-drug interactions through implementing the chemical-protein interactome
- Predicting drug-drug interactions through large-scale similarity-based link prediction
- Predicting drug repositioning opportunities through integrating multiple aspects of drug similarity and disease similarity

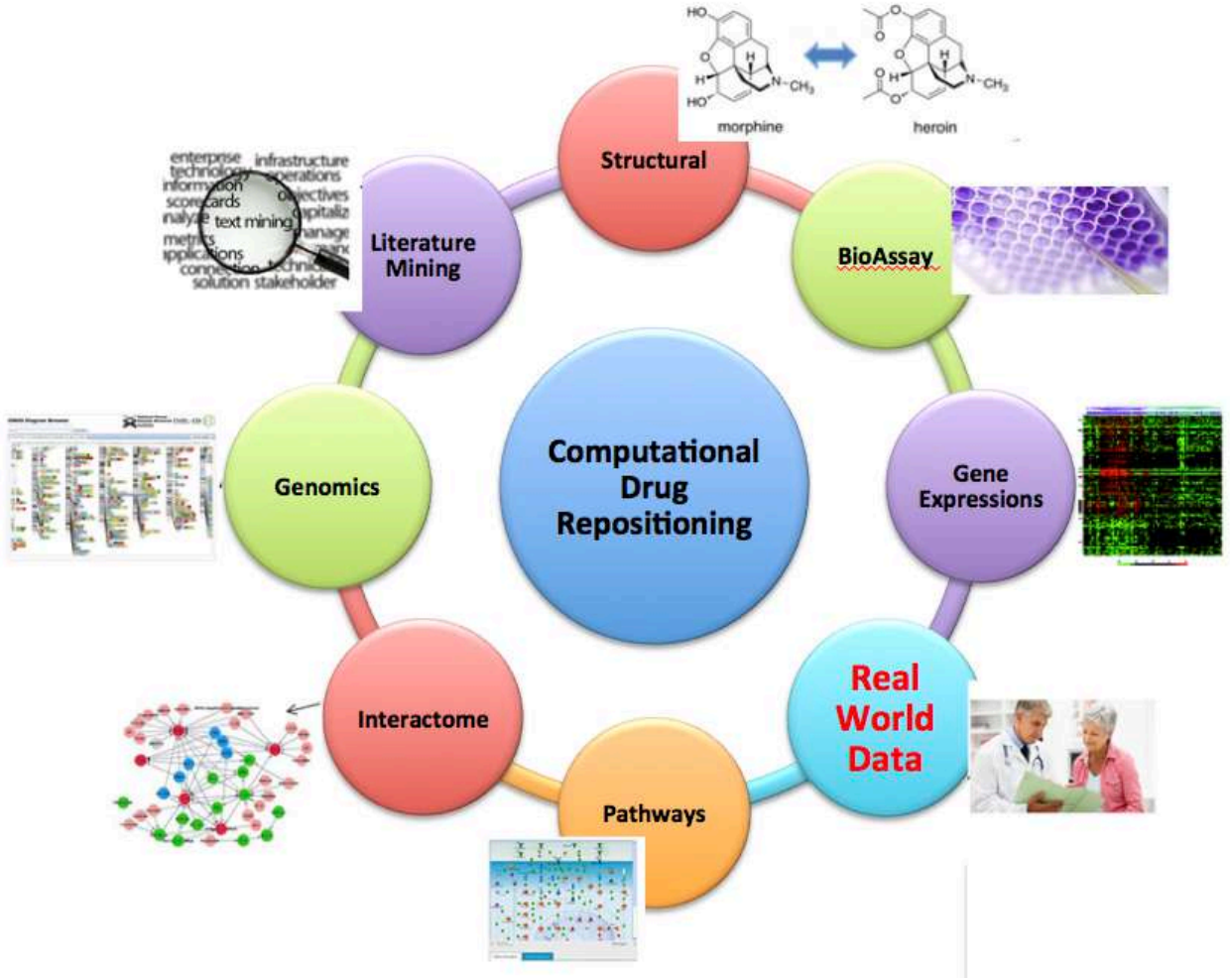
Drug repositioning

- **Drug repositioning** (also known as **Drug repurposing**, **Drug re-profiling**, **Therapeutic Switching** and **Drug re-tasking**) is the application of known drugs and compounds to new indications (i.e., new diseases).

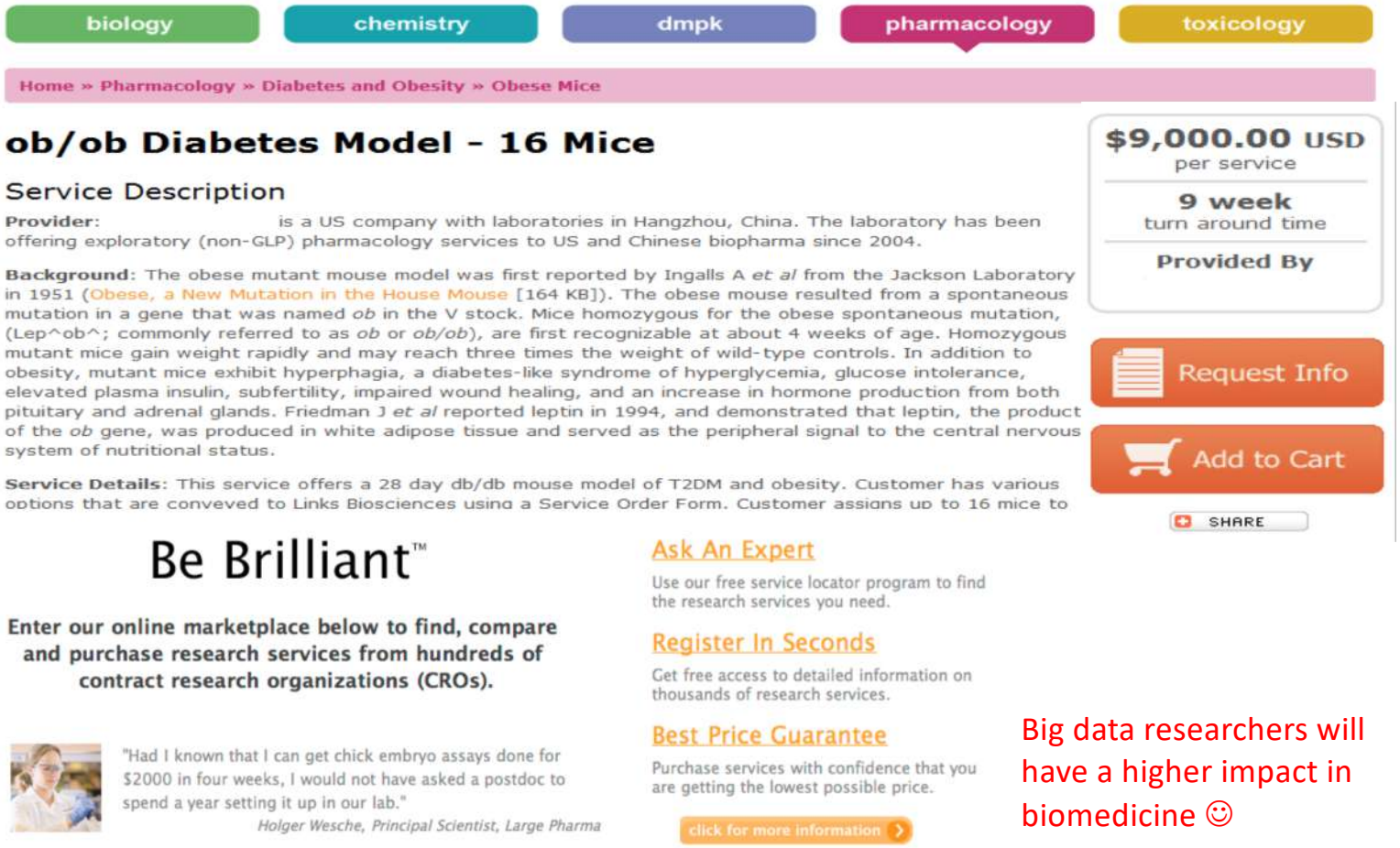
Drug	Original indication	New indication
Viagra	Hypertension	Erectile dysfunction
Wellbutrin	Depression	Smoking cessation
Thalidomide	Antiemetic	Multiple Myeloma

- The repositioned drug has already passed a significant number of toxicity and other tests, its safety is known and the risk of failure for reasons of adverse toxicology are reduced.

Next: Multi-channel detailed computational hypothesis generation



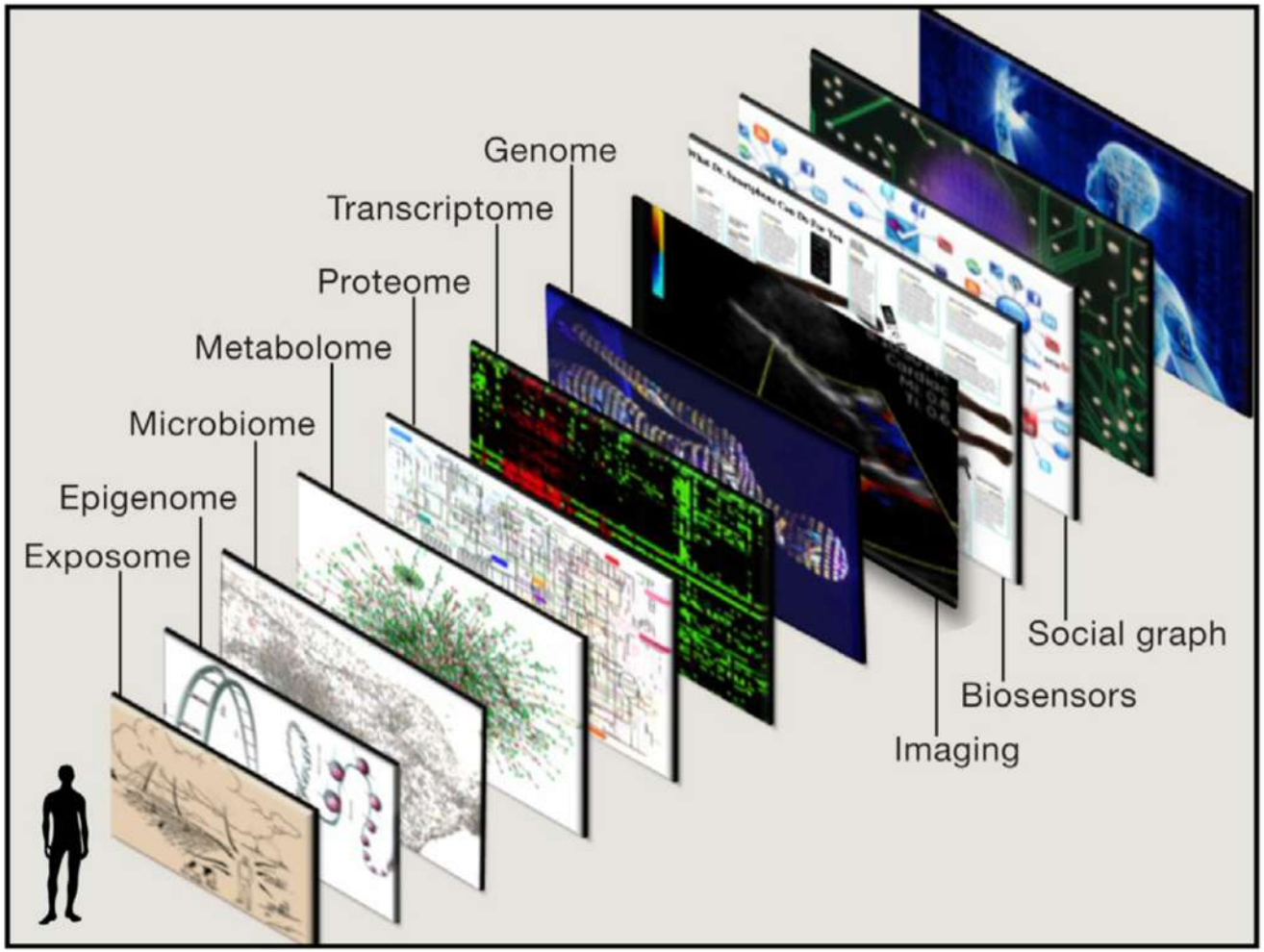
And even beyond the hypothesis generation…



Big data researchers will have a higher impact in biomedicine ☺

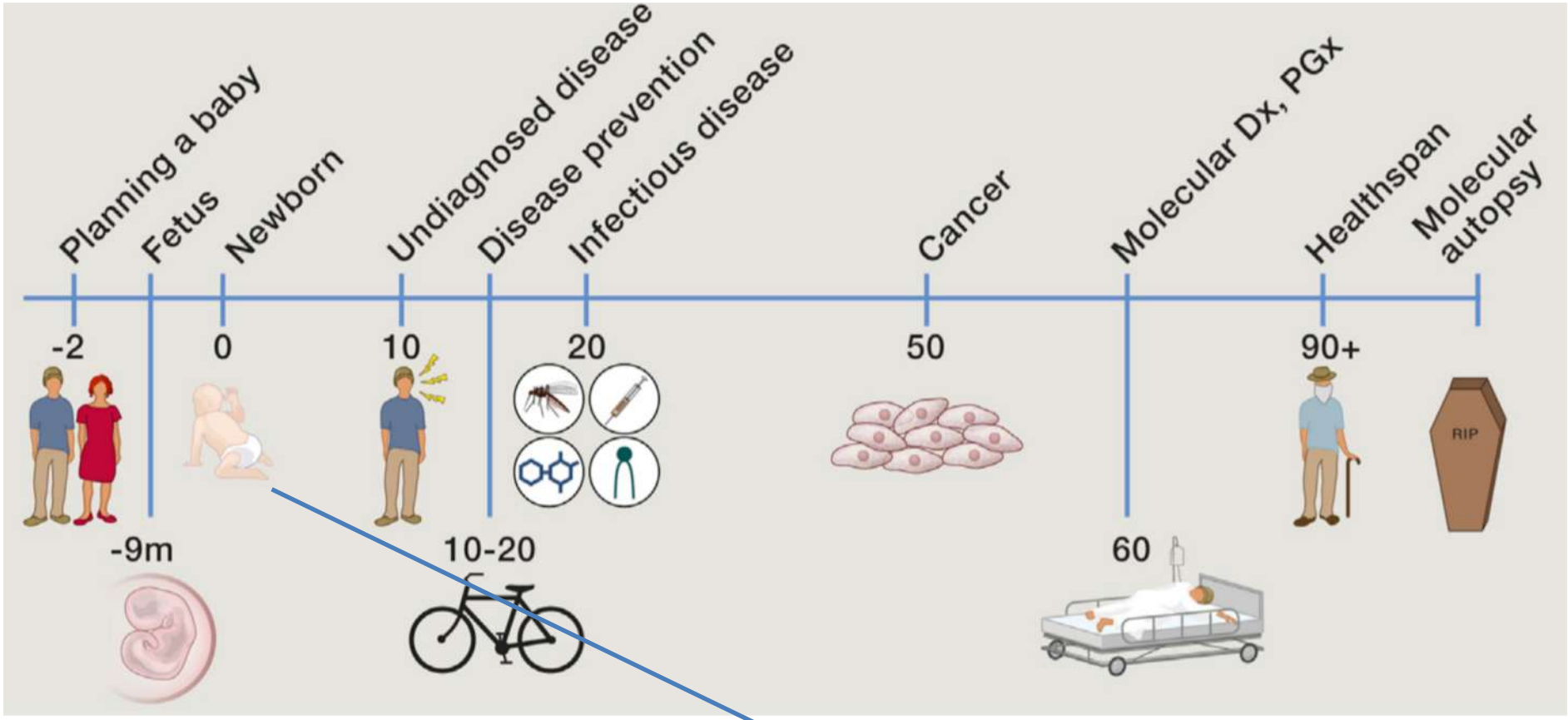
Validation methods are increasingly commoditized

Challenges and opportunities: multiscale networks instead of a diagnosis



Topol E. Individualized Medicine from Prewomb to Tomb. Cell 157, 2014.

Dynamic network: timeline of individualized genomic medicine

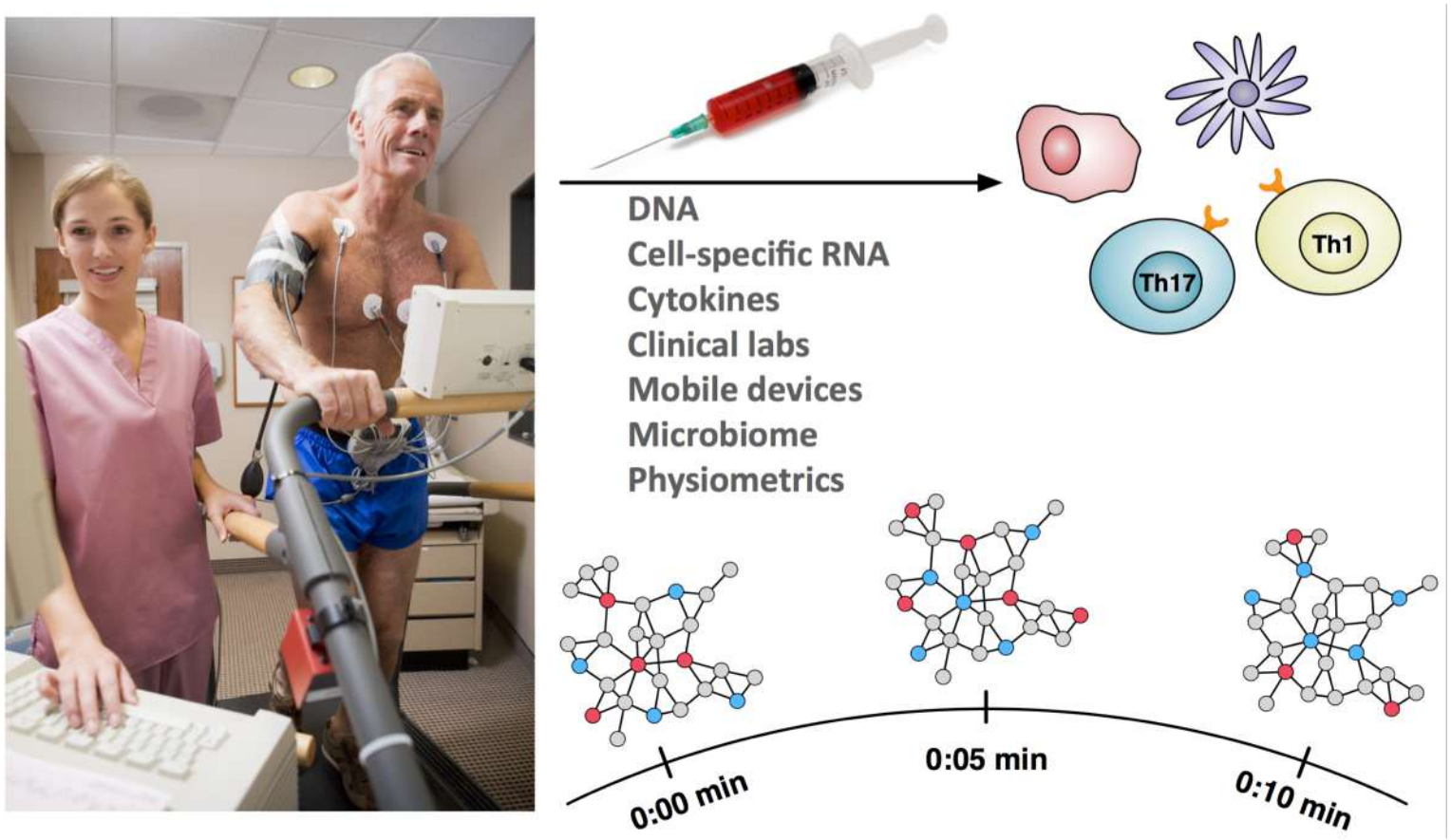


During an individual's lifespan: ***from prewomb to tomb***

Boland MR et al. Birth Month Affects Lifetime Disease Risk: A Phenome-Wide Method. JAMIA 2015.

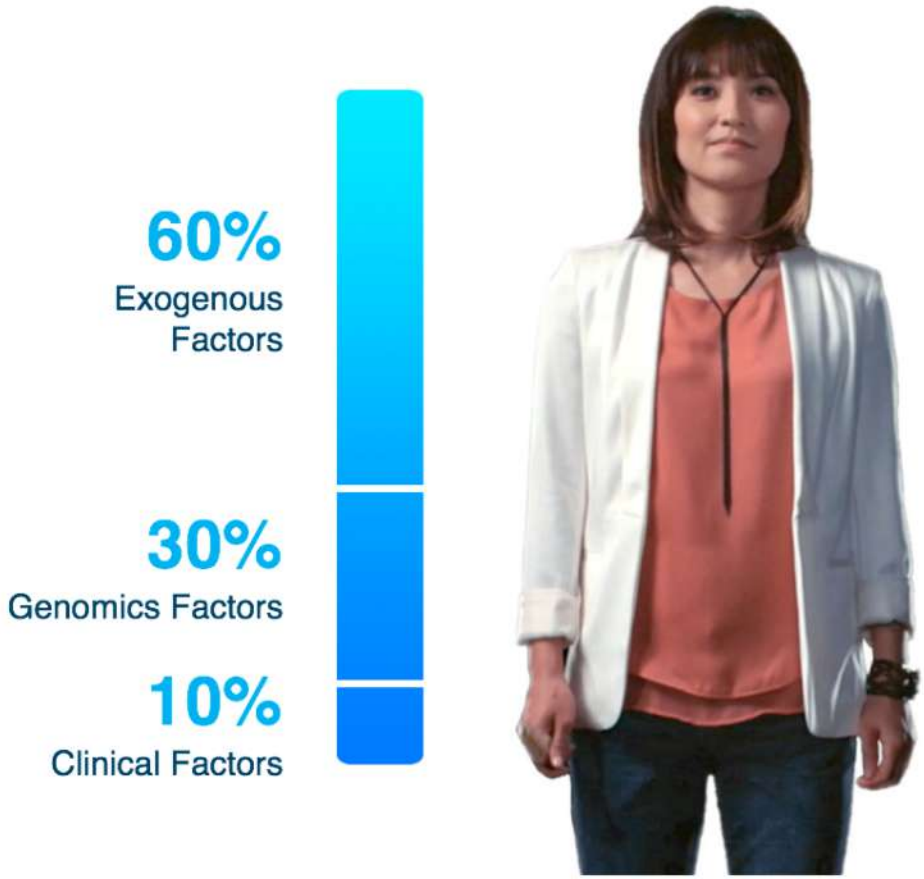
Topol E. Individualized Medicine from Prewomb to Tomb. Cell 157, 2014.

Personalized multiscale networks to model dynamics of complex disease



Dudley J. Big data in biology and medicine. Retrieved at www.aaas.org

Healthcare is really a big data industry



1,100 Terabytes
Generated per lifetime

6 Terabytes
Per lifetime

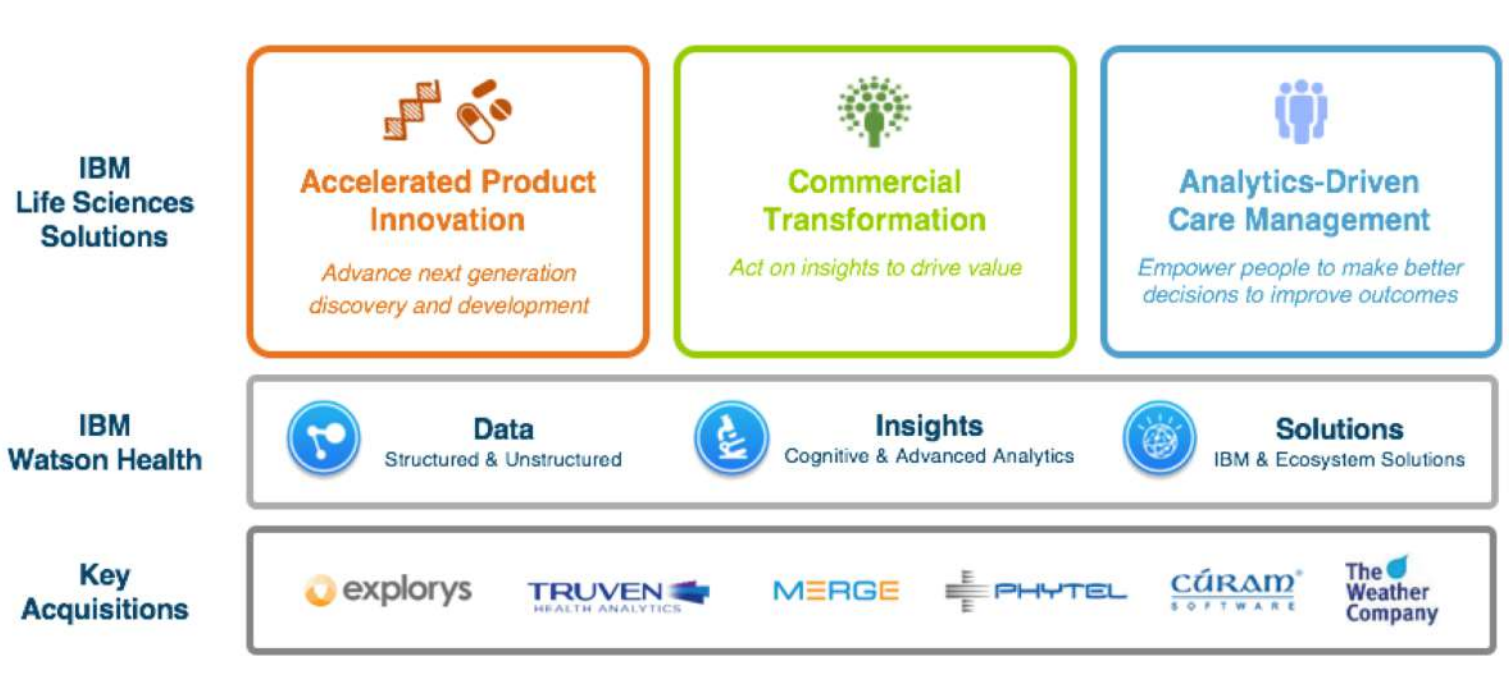
0.4 Terabytes
Per lifetime

Help people live longer and feel better

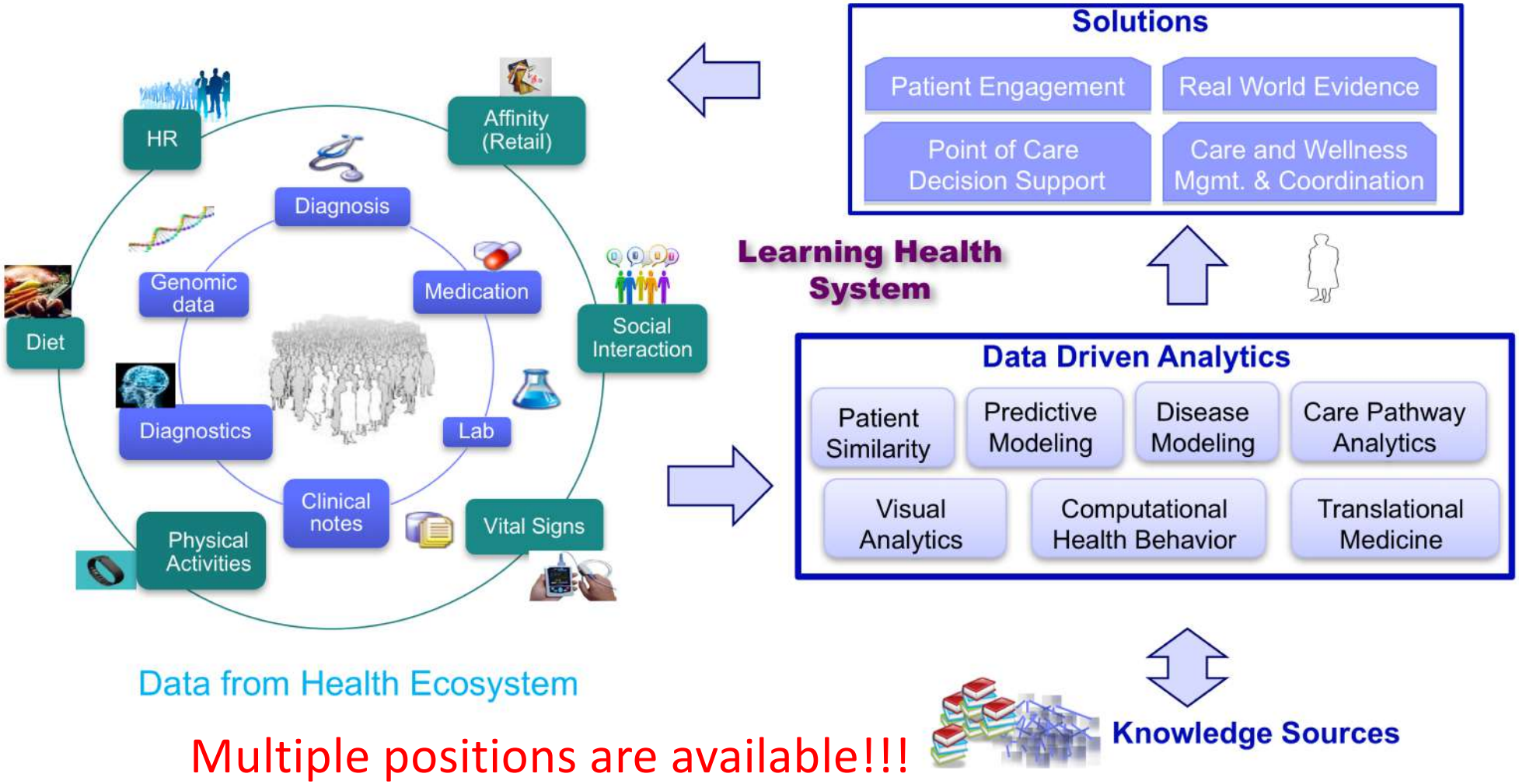
Our commitment to Health – IBM Moonshot

"I'm telling you, our moonshot will be the impact we will have on Healthcare. It has already started. We will change and do our part to change the face of Healthcare. I am absolutely positive about it. And that, to me, while we do many other things, that will be one of the most important."

Ginni Rometty
IBM Chairman, President and CEO
April 16, 2015



Center for Computational Health @ IBM Research



Multiple positions are available!!!

Acknowledgement

- IBM Watson Health
  - Elenee Argentinis
  - Brandon Fiegoli
  - Katherine Shen
- IBM Research
  - Achille Fokoue
  - Heng Luo
- Vanderbilt University
  - Zhijun Yin

And all of our co-authors!!!

Questions?

liying@us.ibm.com
pzhang@us.ibm.com