

**OBSERVATIONAL  
MEDICAL  
OUTCOMES  
PARTNERSHIP**

**Patient-centered observational analytics:  
New directions toward studying the effects of  
medical products**

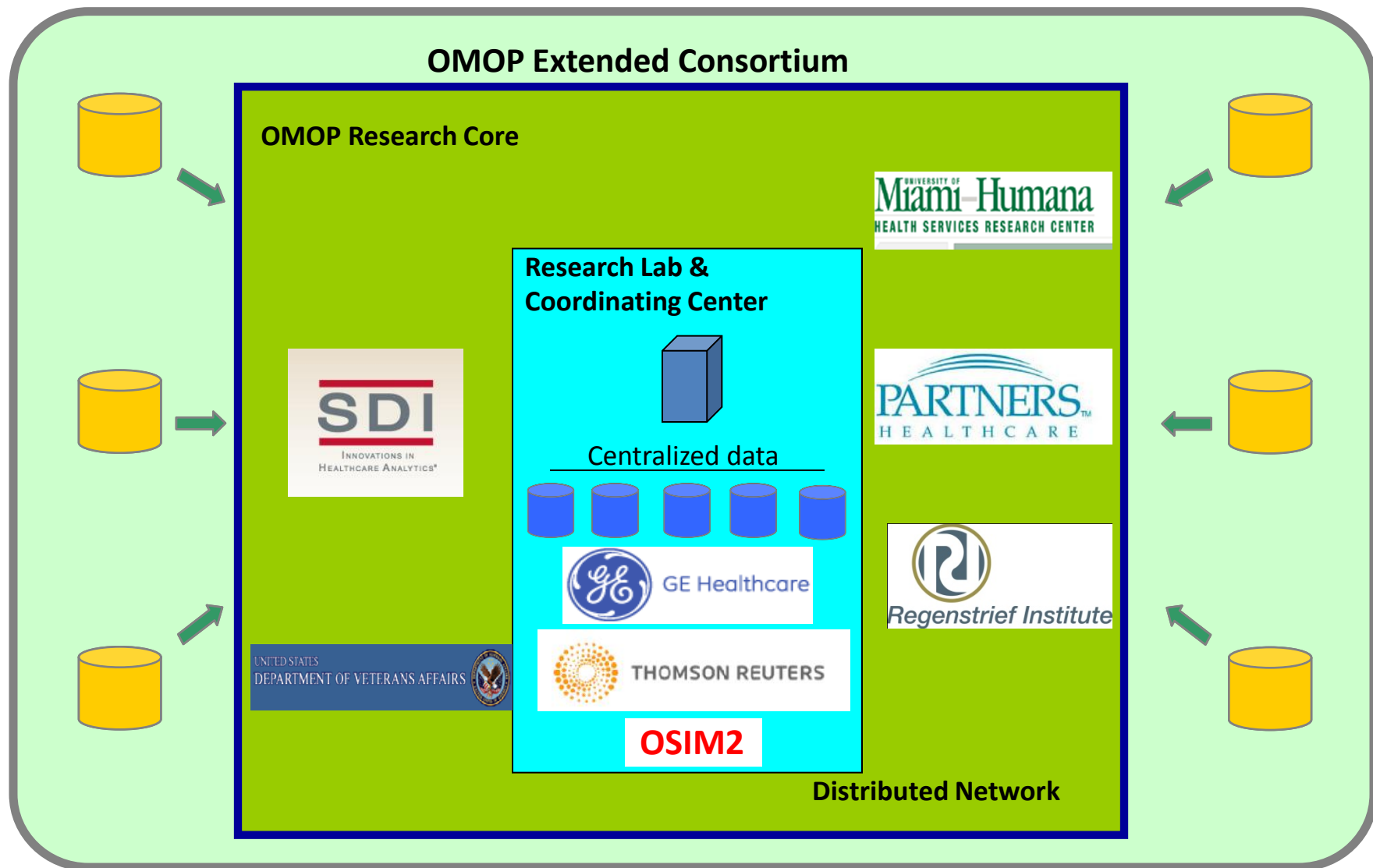
David Madigan  
Columbia University  
on behalf of OMOP Research Team  
August 27, 2012

# Observational Medical Outcomes Partnership

***Public-Private Research Partnership established to inform the appropriate use of observational healthcare databases for studying the effects of medical products:***

- Conducting methodological research to empirically evaluate the performance of alternative methods on their ability to identify true associations
- Developing tools and capabilities for transforming, characterizing, and analyzing disparate data sources across the health care delivery spectrum
- Establishing a shared resource so that the broader research community can collaboratively advance the science

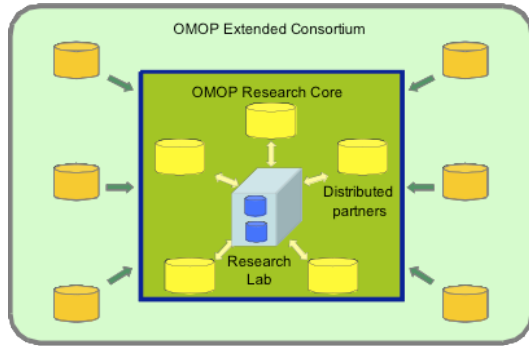
# OMOP Data Community – First Two Years



**178 million** persons with patient-level data

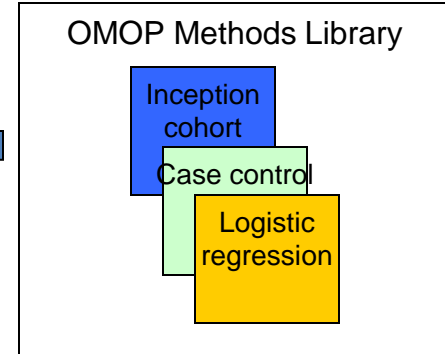
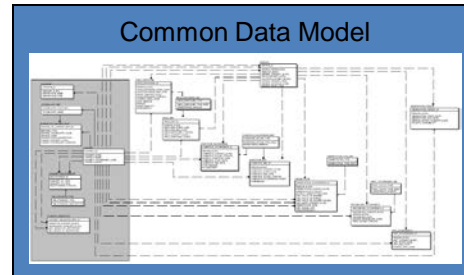
5.4 billion drug exposures, 5.8 billion procedures, 2.3 billion clinical observations

# OMOP Research Experiment



- 10 data sources
- Claims and EHRs
- 170M+ lives
- Simulated data (OSIM)

- Open-source
- Standards-based
- Systematic data characterization and quality assurance



- 14 methods implemented as standardized procedures
- Full transparency with open-source code and documentation
- Epidemiology, statistical and machine learning designs

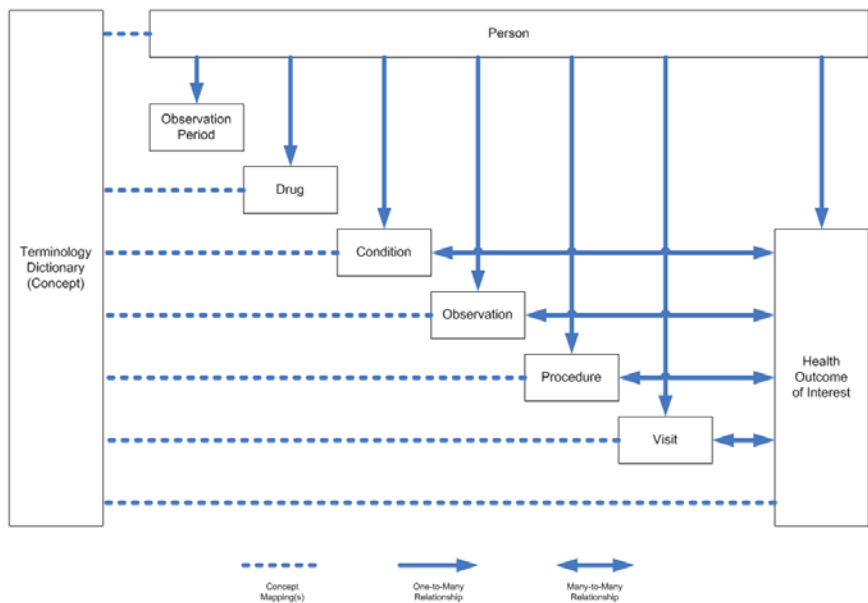
Drug

Outcome	ACE Inhibitors	Amphotericin B	Antibiotics: erythromycins, sulfonamides, tetracyclines	Antiepileptics: carbamazepine, phenytoin	Benzodiazepines	Beta blockers	Bisphosphonates: alendronate	Tricyclic antidepressants	Typical antipsychotics	Warfarin
Angioedema	Red	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue
Aplastic Anemia	Blue	Blue	Blue	Red	Blue	Blue	Blue	Blue	Blue	Blue
Acute Liver Injury	Blue	Blue	Red	Blue	Blue	Blue	Blue	Blue	Blue	Blue
Bleeding	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Red
Hip Fracture	Blue	Blue	Blue	Blue	Red	Blue	Blue	Blue	Blue	Blue
Hospitalization	Green	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue
Myocardial Infarction	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Red	Red	Blue
Mortality after MI	Blue	Blue	Blue	Blue	Blue	Green	Blue	Blue	Blue	Blue
Renal Failure	Blue	Red	Blue	Blue	Blue	Blue	Blue	Blue	Blue	Blue
GI Ulcer Hospitalization	Blue	Blue	Blue	Blue	Blue	Blue	Red	Blue	Blue	Blue

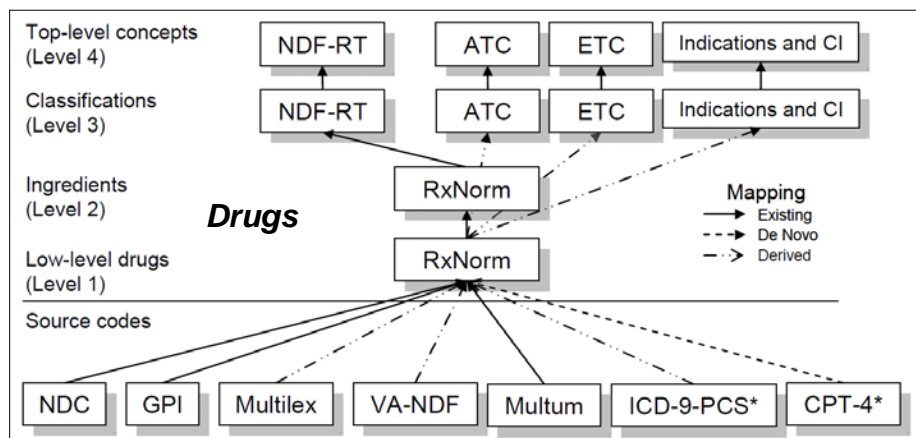
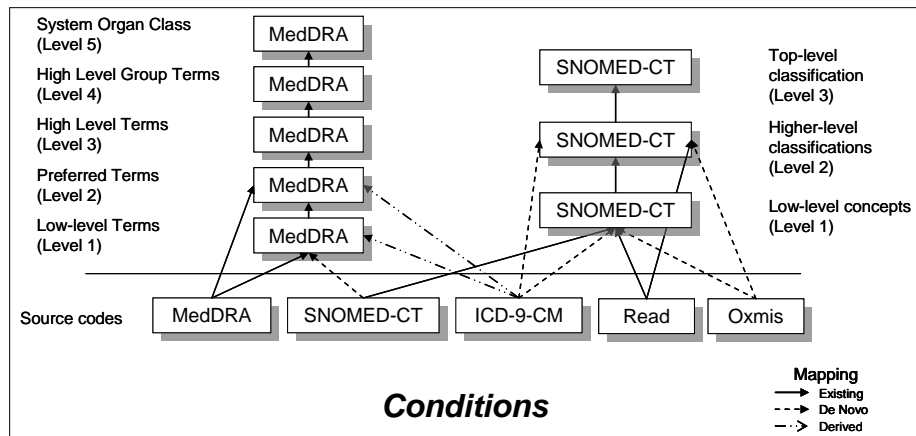
# Common Framework

## Accommodating Disparate Observational Data Sources

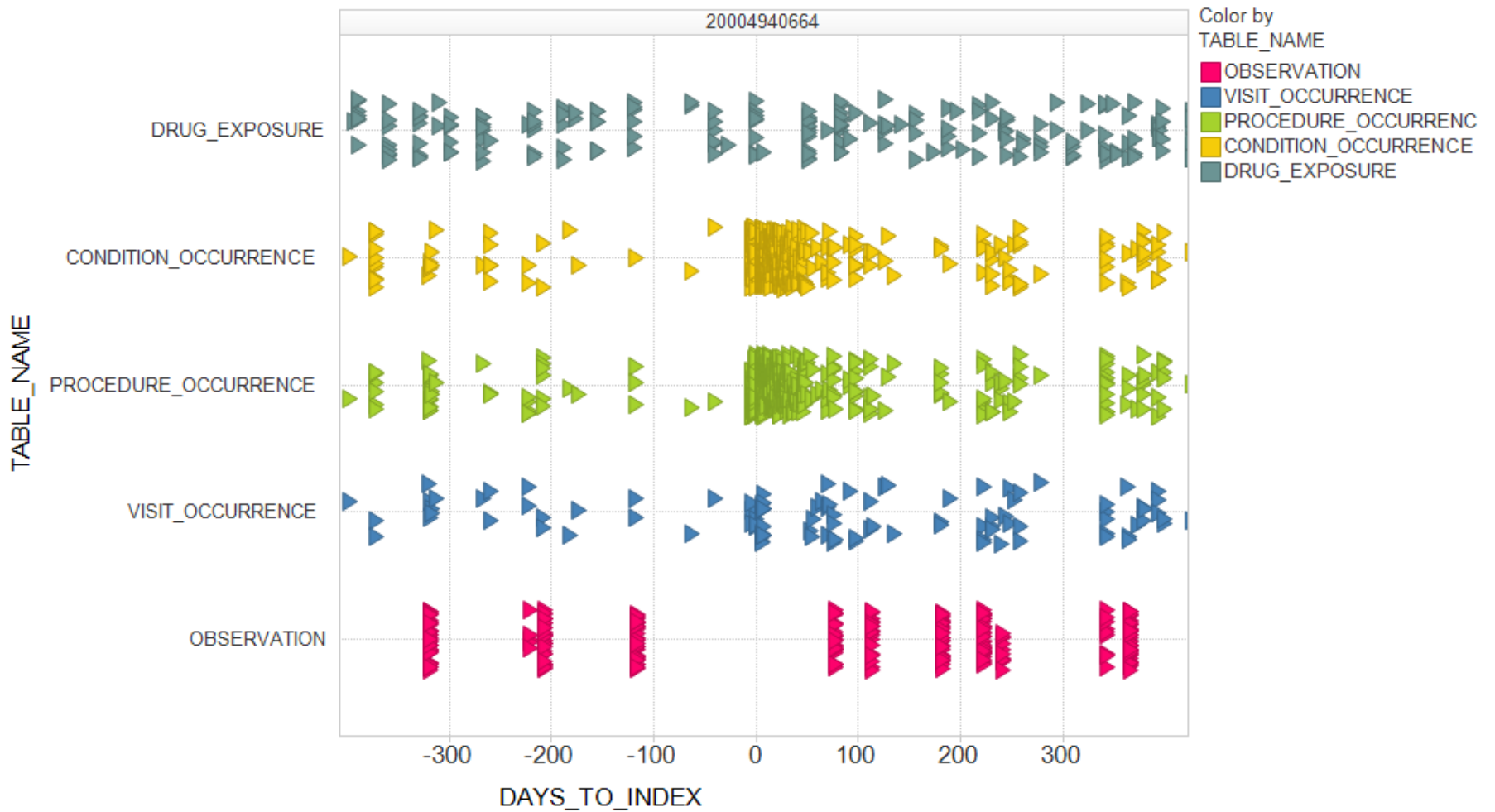
### Common Data Model

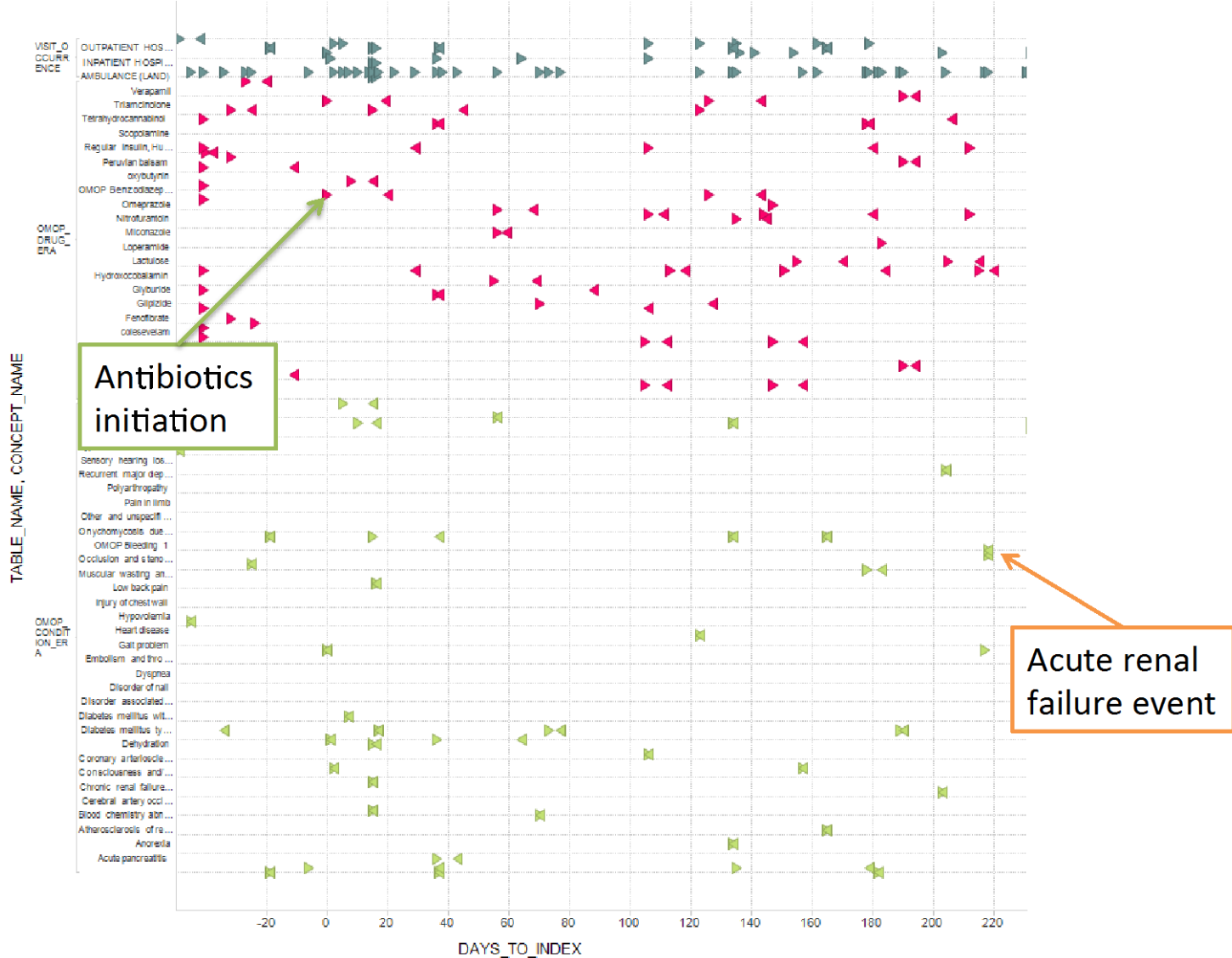


### Standardized Terminologies

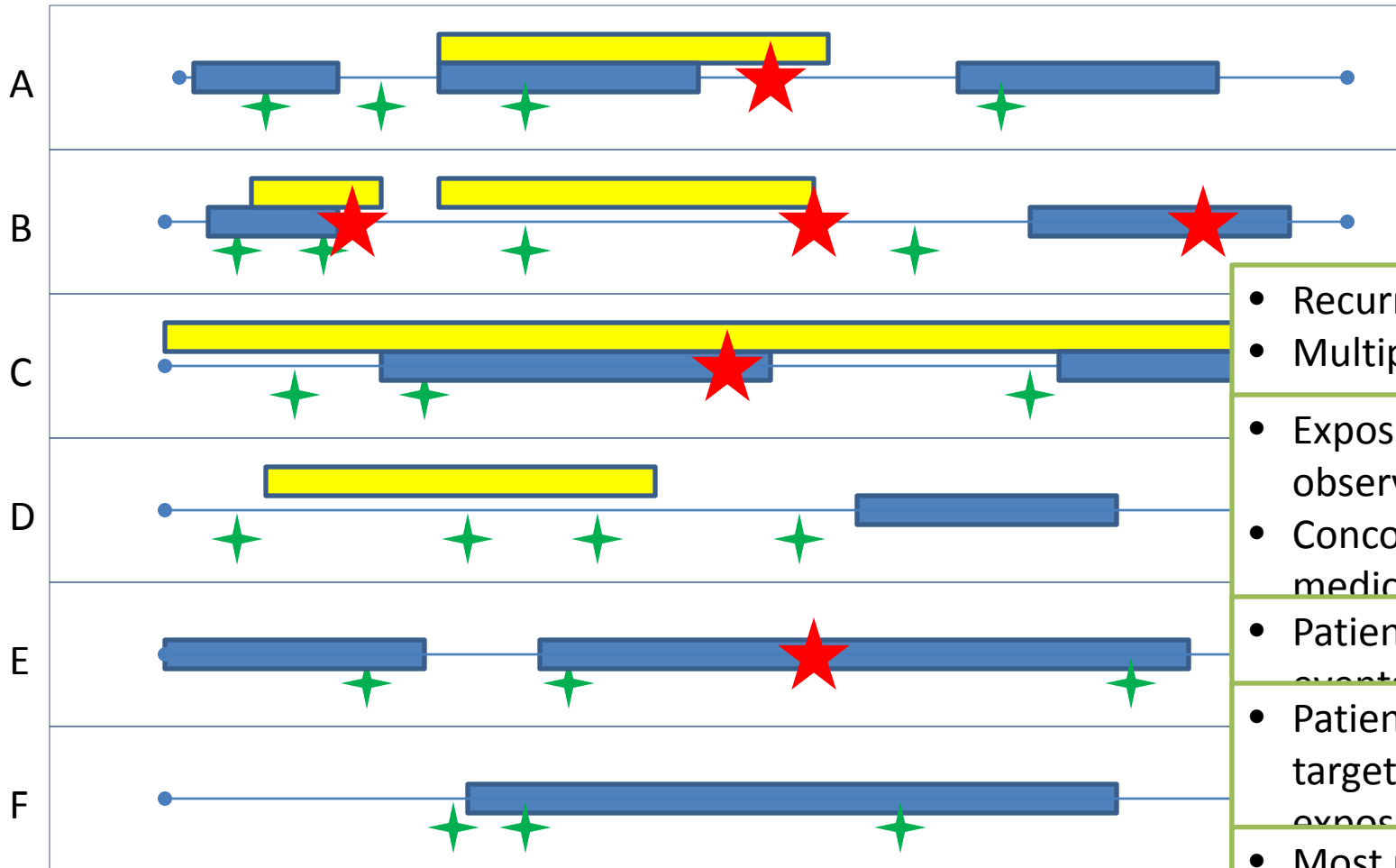


# A couple years in the life of a patient in an observational healthcare database





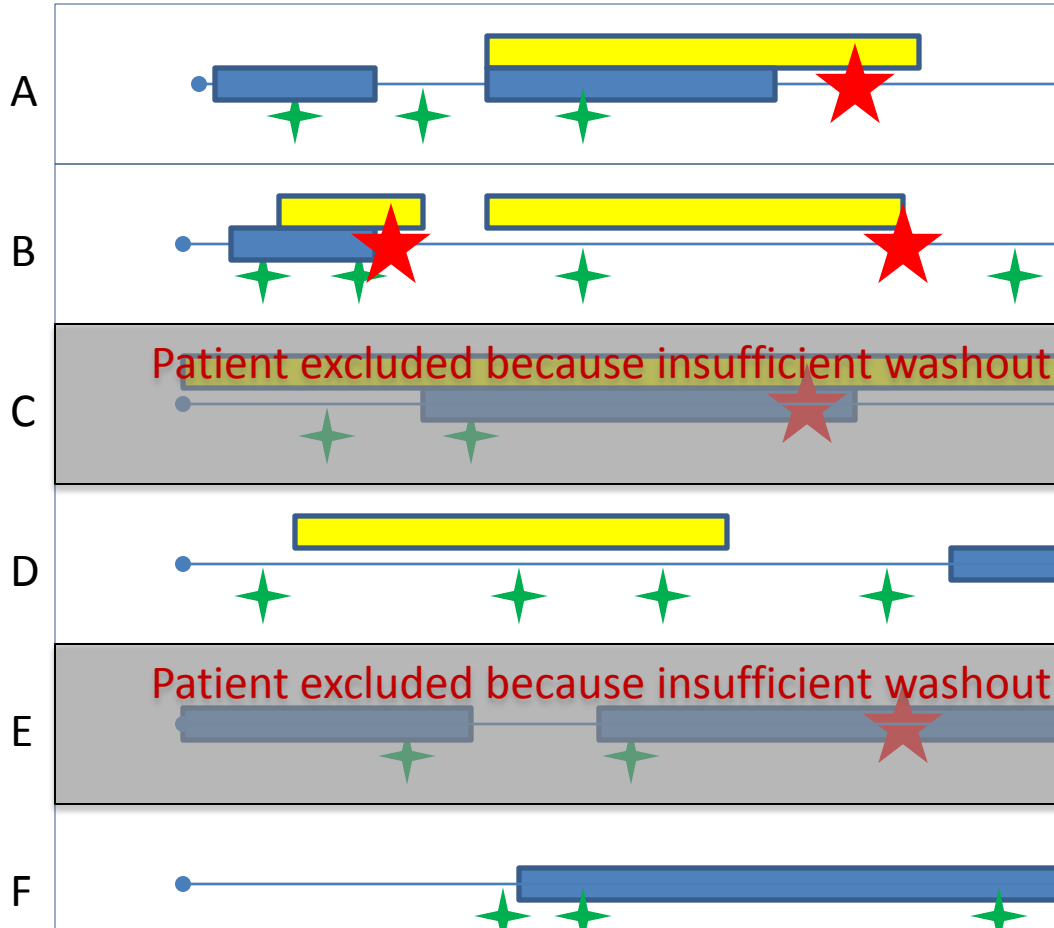
# Patient profiles in observational data when studying the effects of medical products



- Recurrent events
- Multiple periods of exposure spanning observation period
- Concomitant medications during exposure
- Patients without target drug exposure are common
- Patients without target drug exposure are common
- Most patients in the database have neither the target drug nor the target outcome



# Data used for new user cohort design to estimate average treatment effect



**New user design**

- Focus on comparing rates of events among patients exposed to target drug, relative to rates of events among patients in some referent comparator group
- Relative risk can be adjusted for baseline covariates through various strategies, including propensity score

- Define cohorts based on index exposure (first use after washout period)
- Observations prior to index may be used as covariates
- Observations on or after index, except for incident outcome, are not considered in analysis

- Target drug
- Target condition
- Other conditions
- Other drugs

CMAJ

RESEARCH

## Adverse events associated with treatment of latent tuberculosis in the general population

Benjamin M. Smith MD, Kevin Schwartzman MD MPH, Gillian Bartlett PhD, Dick Menzies MD MSc

### ABSTRACT

**Background:** Guidelines recommend treatment of latent tuberculosis in patients at increased risk for active tuberculosis. Studies investigating the association of therapy with serious adverse events have not included the entire treated population nor accounted for comorbidities or occurrence of similar events in the untreated general population. Our objective was to estimate the risk of adverse events requiring hospital admission that were associated with therapy for latent tuberculosis infection in the general population.

**Methods:** Using administrative health data from the province of Quebec, we created a historical cohort of all residents dispensed therapy for latent tuberculosis between 1998 and 2003. Each patient was matched on age, sex and postal region with two untreated residents. The observation period was 18 months (from 6 months before to 12 months after initiation of therapy). The primary outcome was hospital admission for therapy-associated adverse events.

**Results:** During the period of observation, therapy for latent tuberculosis was dispensed to 9145 residents, of whom 95% started isoni-

azid and 5% started rifampin. Pretreatment comorbid illness was significantly more common among patients receiving such therapy compared with the matched untreated cohort. Of all patients dispensed therapy, 45 (0.5%) were admitted to hospital for a hepatic event compared with 15 (0.1%) of the untreated patients. For people over age 65 years, the odds of hospital admission for a hepatic event among patients treated for latent tuberculosis infection was significantly greater than among matched untreated people after adjustment for comorbidities (odds ratio [OR] 6.4, 95% CI 2.2–18.3). Excluding patients with comorbid illness, there were two excess admissions to hospital for hepatic events per 100 patients initiating therapy compared with the rate among untreated people over 65 years (95% CI 0.1–3.87).

**Interpretation:** The risk of adverse events requiring hospital admission increased significantly among patients over 65 years receiving treatment for latent tuberculosis infection. The decision to treat latent tuberculosis infection in elderly patients should be made after careful consideration of risks and benefits.

**Competing interests:**  
None declared.

This article has been peer reviewed.

**Correspondence to:**  
Dr. Dick Menzies;  
dick.menzies@mcgill.ca

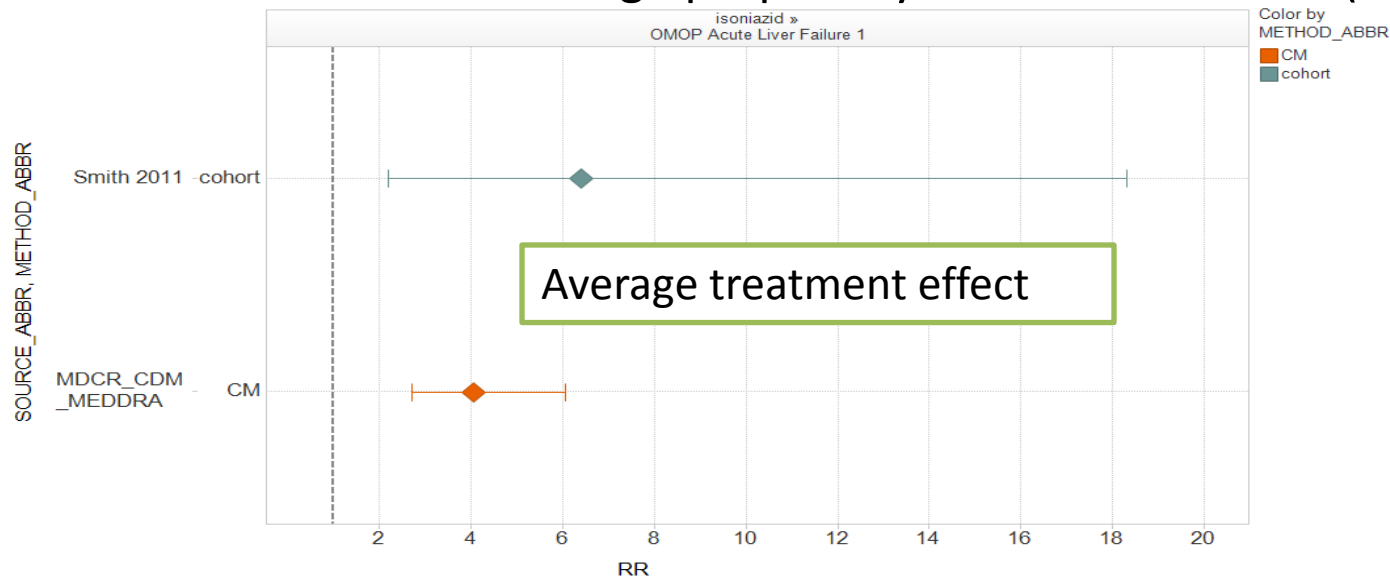
*CMAJ* 2011. DOI:10.1503/  
/cmaj.091824

Average treatment effect,  
patients > 65 years of age:  
OR = 6.4 (2.2 – 18.3)

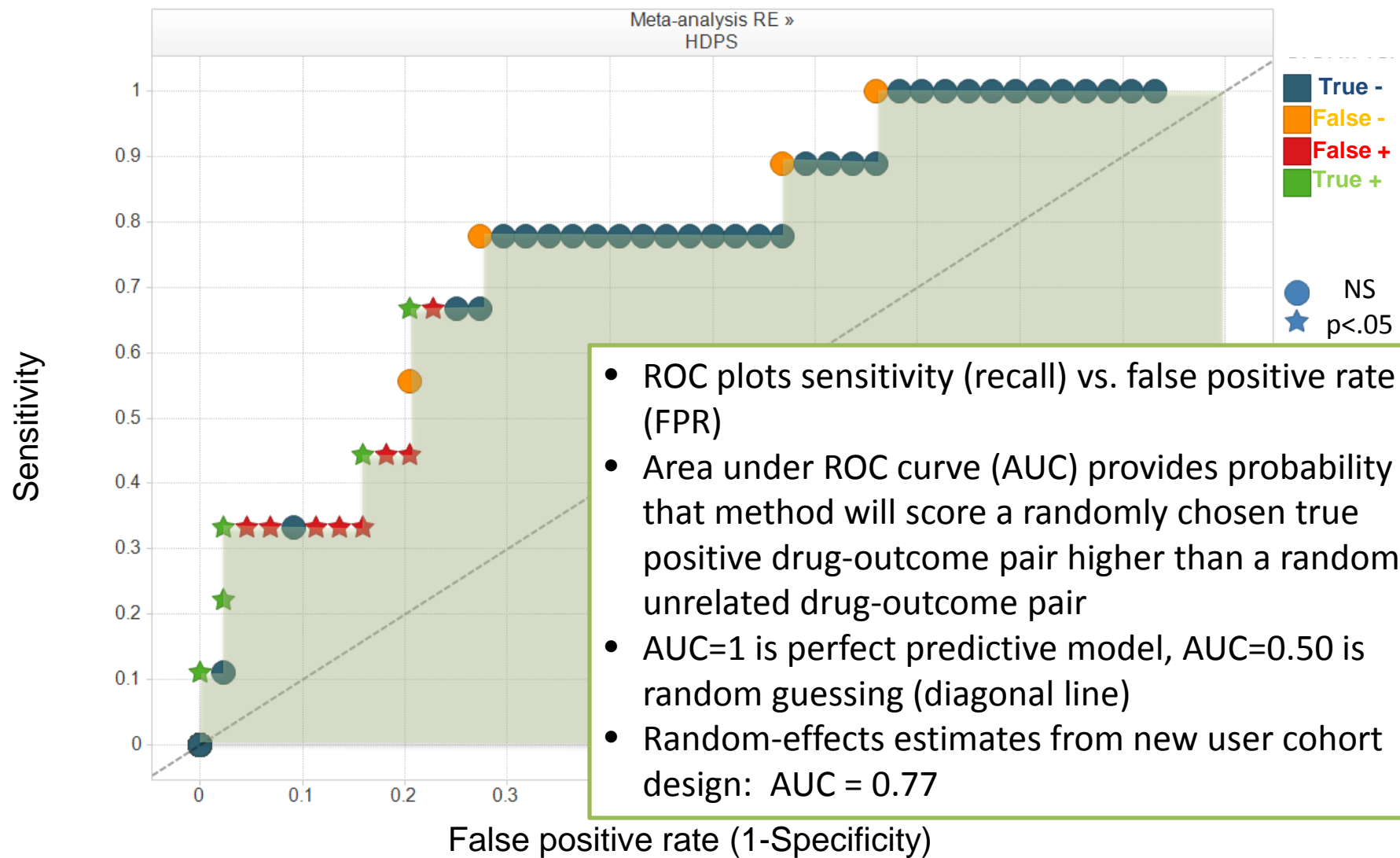
*CMAJ*, February 22, 2011, 183(3)

# OMOP replication: isoniazid – acute liver injury

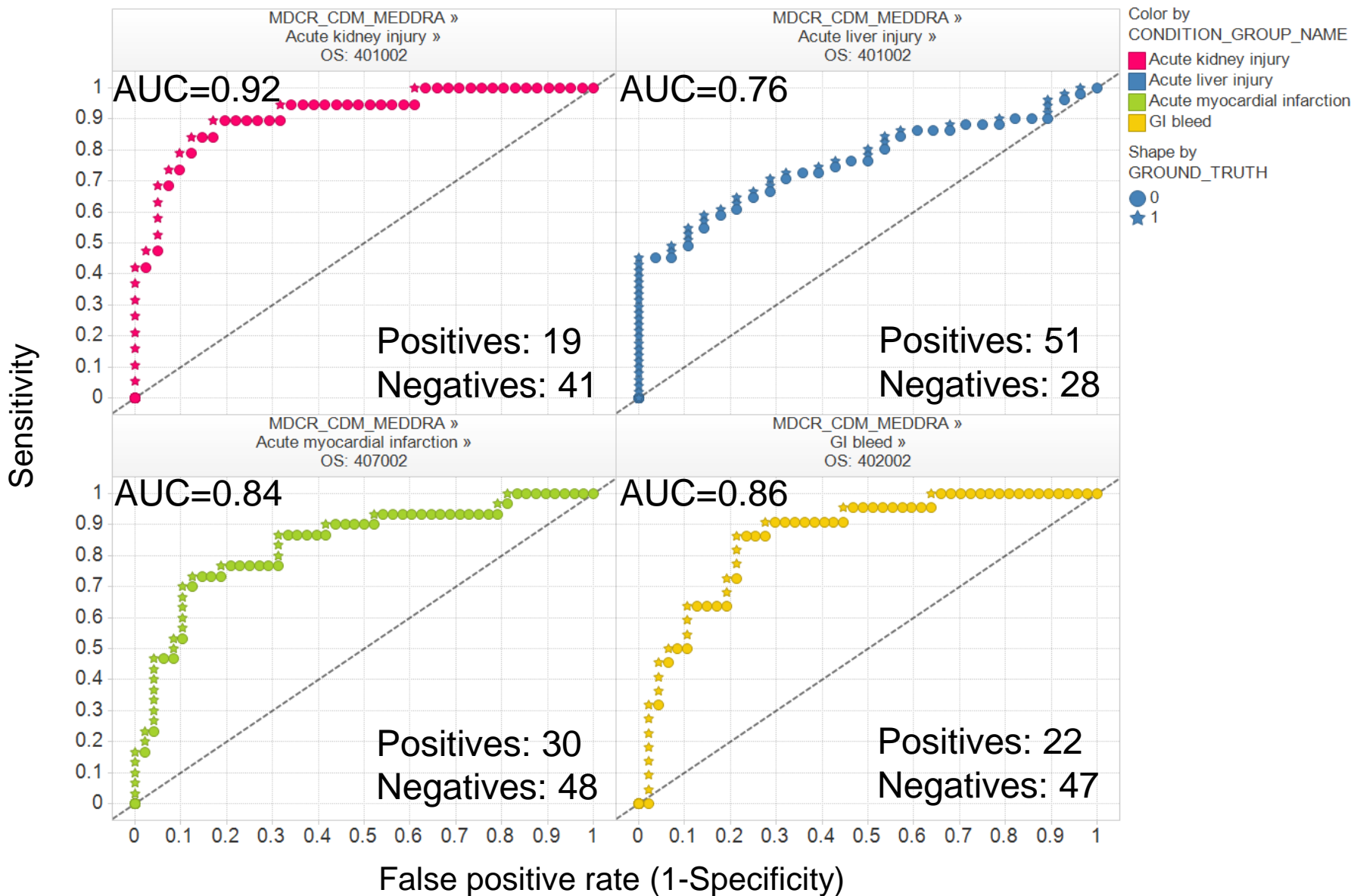
- Data source: MarketScan Medicare Beneficiaries (MDCR)
- Study design: Cohort
- Exposure: all patients dispensed new use of isoniazid, 180d washout
- Unexposed cohort: Patient with indicated diagnosis (e.g. pulmonary tuberculosis) but no exposure to isoniazid; negative control drug referents
- Time-at-risk: Length of exposure + 30 days, censored at incident events
- Covariates: age, sex, index year, Charlson score, number of prior visits, all prior medications, all comorbidities, all priority procedures
- “Odds ratio” estimated through propensity score stratification (20 strata)



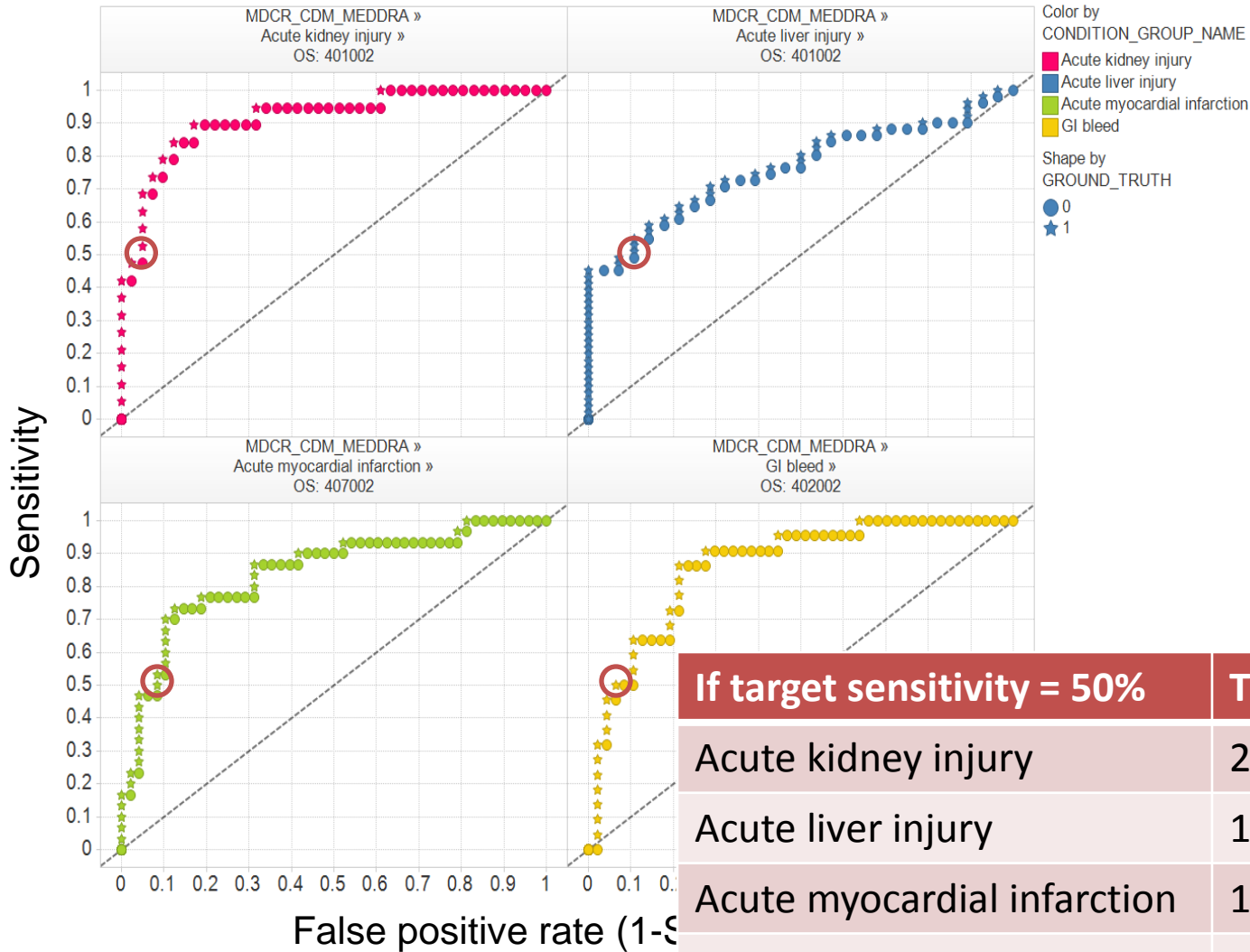
# Receiver Operating Characteristic (ROC) curve



# Tailor to outcome and database, power restriction



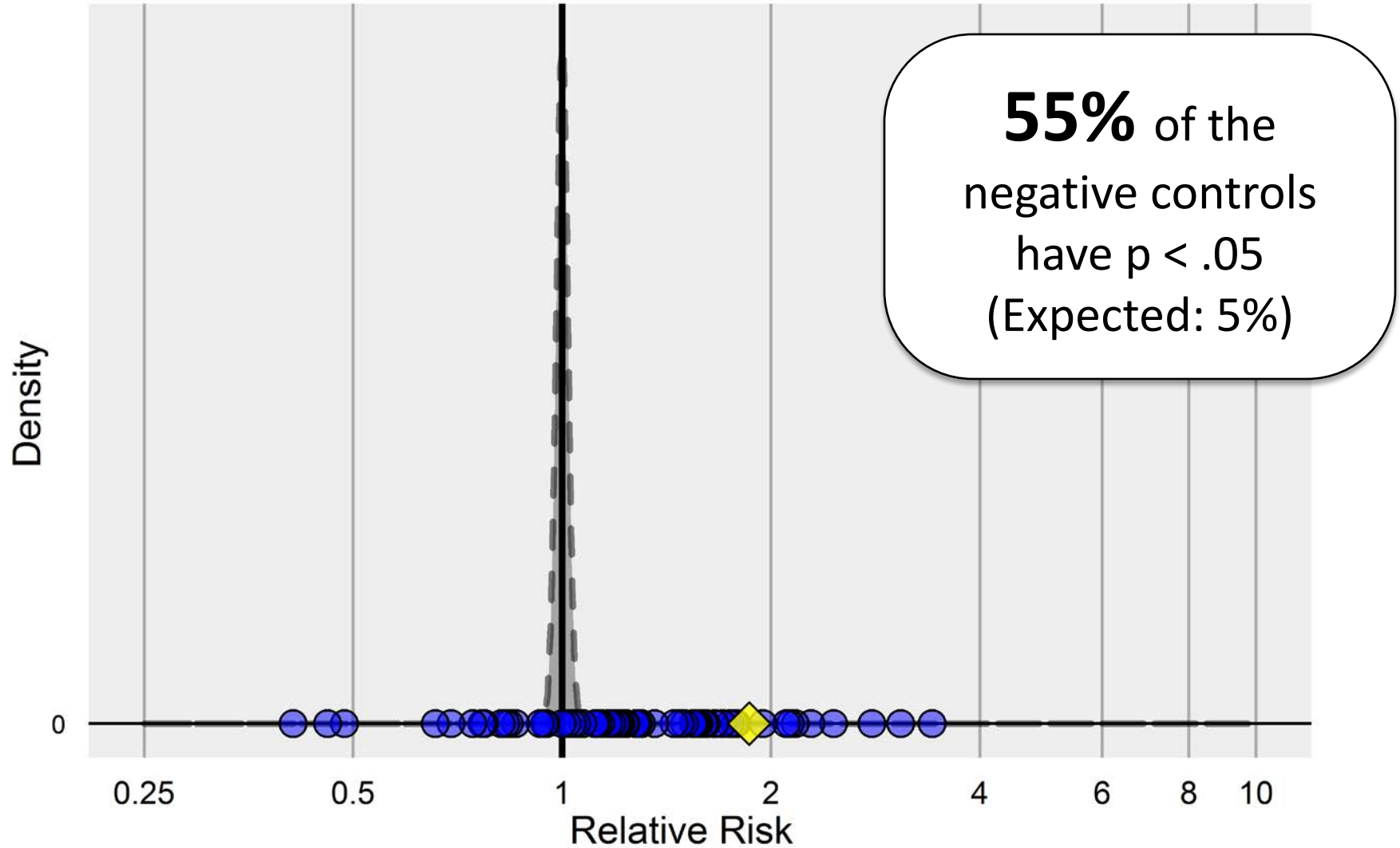
# Sensitivity-Specificity Tradeoff



If target sensitivity = 50%	Threshold	Specificity
Acute kidney injury	2.69	95%
Acute liver injury	1.51	89%
Acute myocardial infarction	1.59	92%
GI bleed	1.87	94%

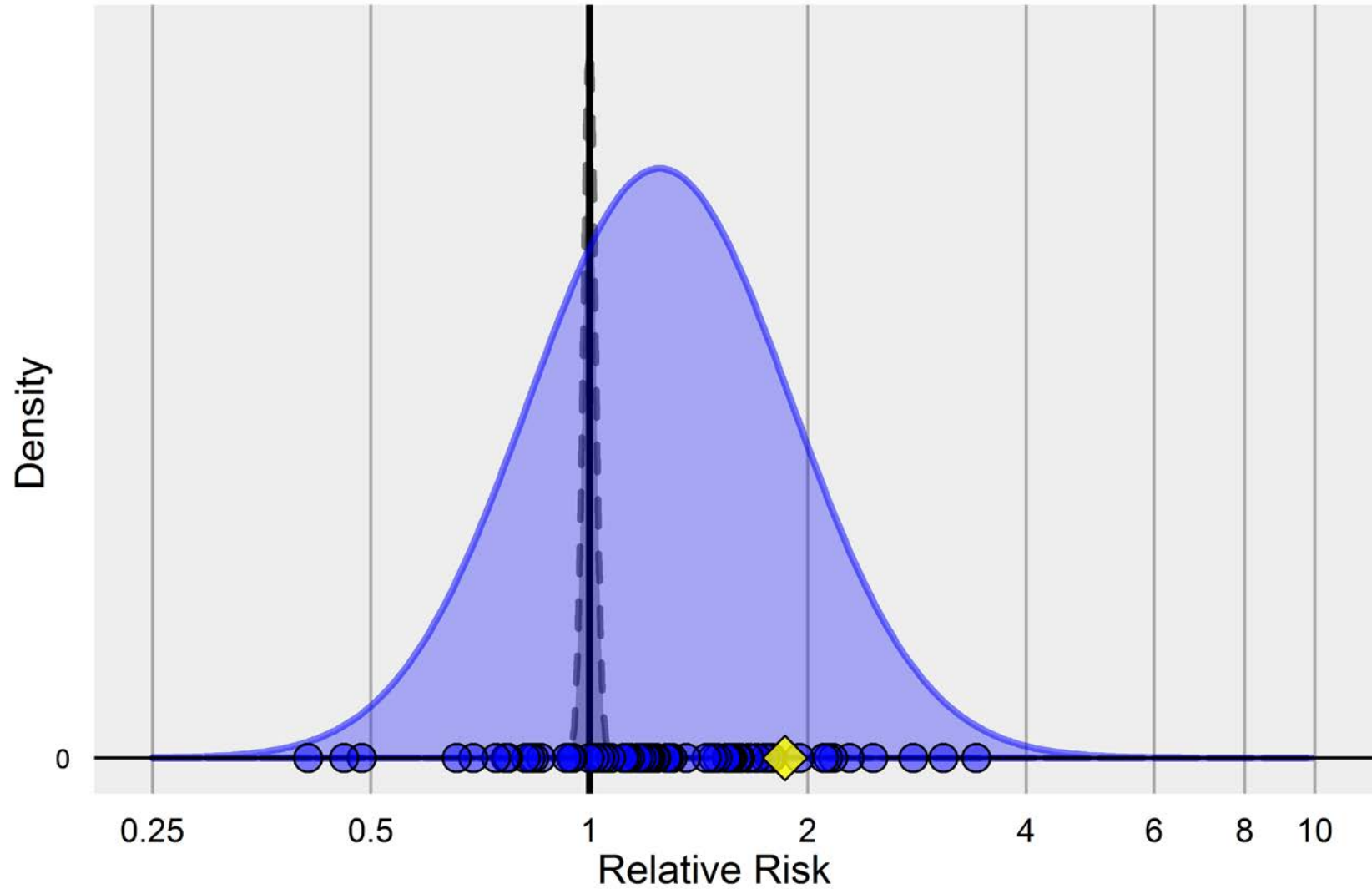
# Negative controls & the null distribution

CC: 2000314, CCAE, GI Bleed



# Negative controls & the null distribution

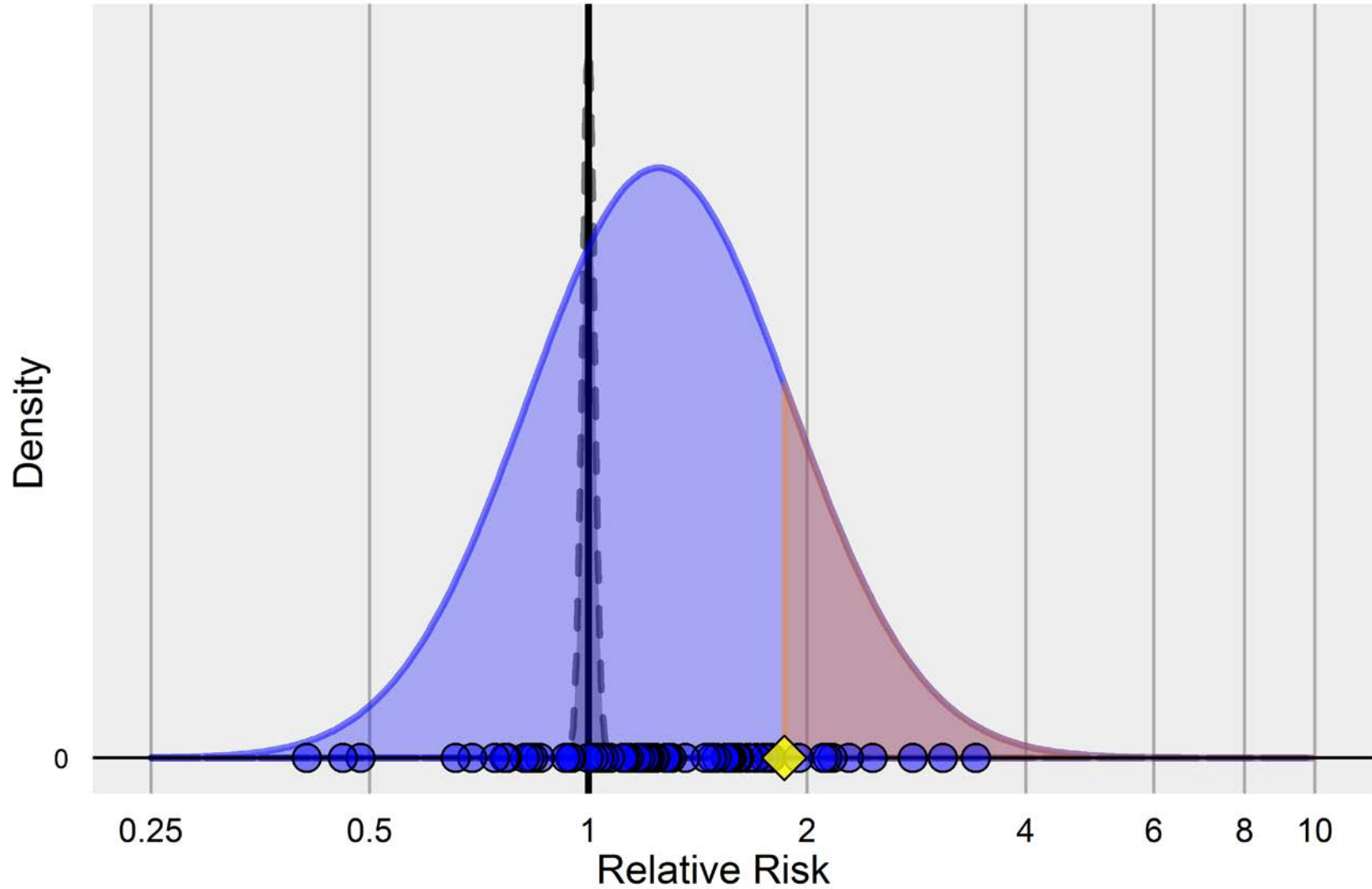
CC: 2000314, CCAE, GI Bleed



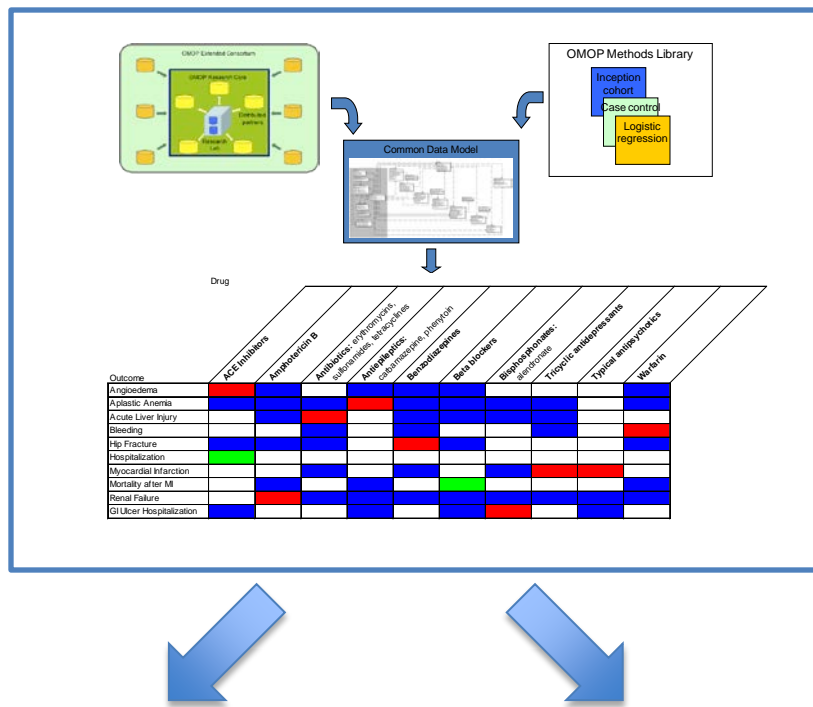


# Negative controls & the null distribution

CC: 2000314, CCAE, GI Bleed



# Where do we go from here?



Further exploration of average treatment effects

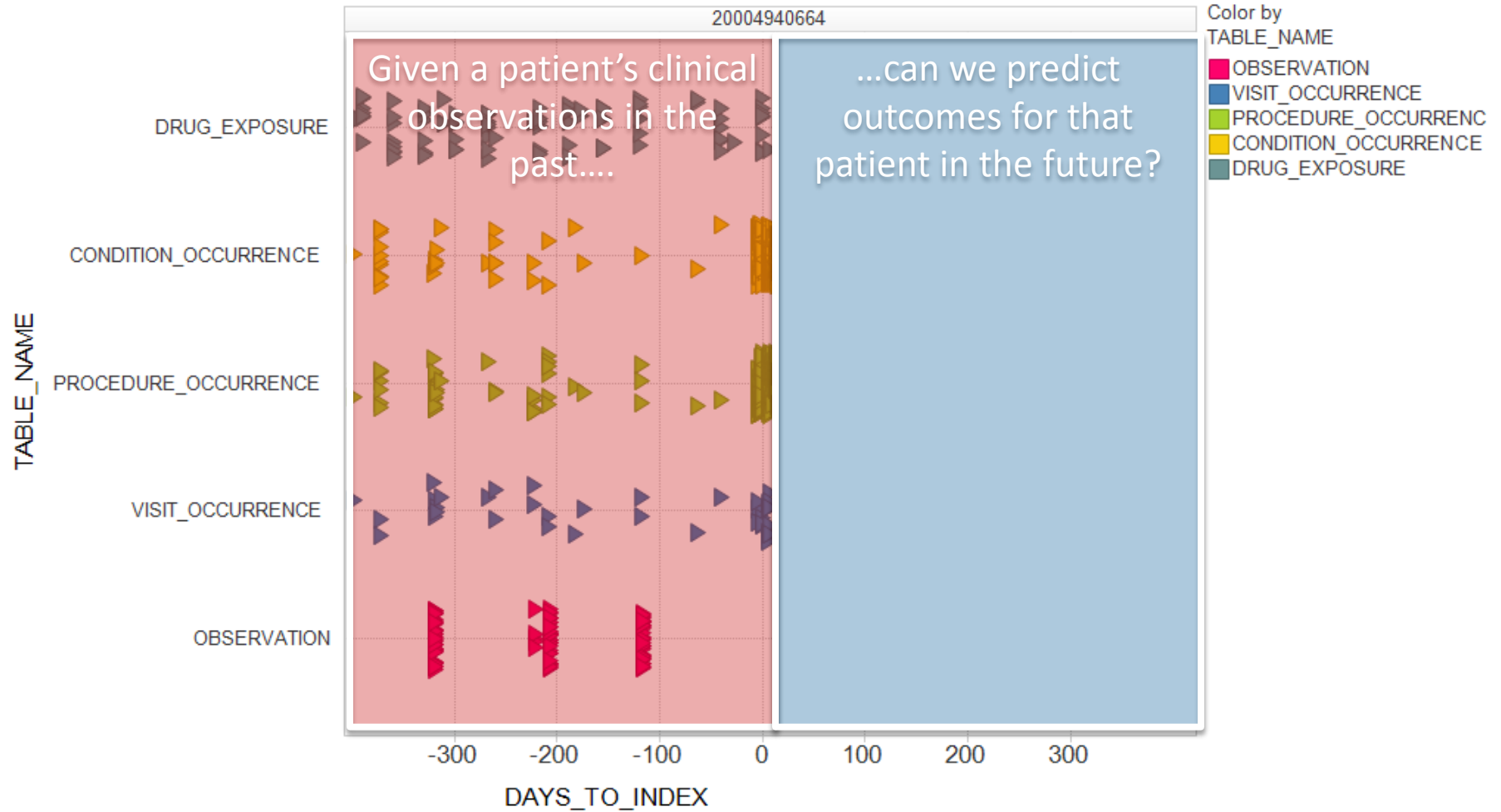
- Increased methods development
- Expansion of test cases
- Evaluate predictive accuracy

New direction:

Patient-centered predictions

- Estimate probability of future outcome, based on past clinical observations
- Evaluate predictive accuracy

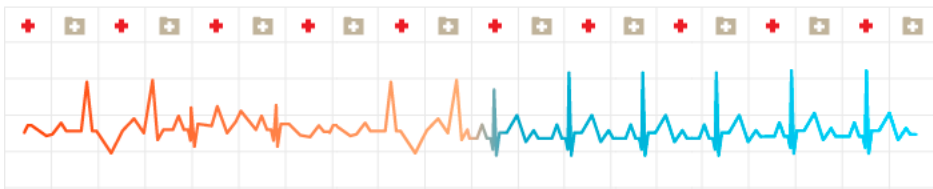
# A couple years in the life of a patient in an observational healthcare database



# Patient-centered predictive modeling on big data has big value and big interest



Information Data Forum Leaderboard



**376 discussions**  
in this [competition's forum](#)

Male Pregnancy?  
8 hours ago

Prediction intervals for your forecasts - suggested approach  
yesterday

Hospitals enlist vendors for data analytics help  
2 days ago

## Improve Healthcare, Win \$3,000,000.

COMPETITION GOAL

**Identify patients who will be admitted to a hospital within the next year, using historical claims data.**

- Description
- Evaluation
- Rules
- Dos and Don'ts
- FAQ
- Timeline

[Get the data! »](#)  
[Make a submission »](#)

More than 71 million individuals in the United States are admitted to hospitals each year, according to the latest survey from the American Hospital Association. Studies have concluded that in 2006 well over \$30 billion was spent on unnecessary hospital admissions. Is there a better way? Can we identify earlier those most at risk and ensure they get the treatment they need? The Heritage Provider Network (HPN) believes that the answer is "yes".

To achieve its goal of developing a breakthrough algorithm that uses available patient data to predict and prevent unnecessary hospitalizations, HPN is sponsoring the Heritage Health Prize Competition

<http://www.heritagehealthprize.com/>

**Leaderboard** [more »](#)

1.	Opera Solutions (171)
2.	EXL Analytics (293)
3.	Market Makers (214)
4.	jack3 (215)
5.	Dolphin (239)
6.	Edward & Willem (259)
7.	Areté Associates (70)
8.	Petterson & Caetano @ NICTA (77)
9.	SD_John_lily (113)
10.	Chris R (165)

1,032 TEAMS WITH

1 2 1 4

PLAYERS

1 4 7 7 5

ENTRIES

## About the CHD Score Sheet

This CHD score sheet can be used to estimate a man's risk of developing CHD over a 10-year period based on age, total cholesterol (TC), HDL cholesterol (HDL-C), blood pressure (BP), and cigarette smoking.

Risk estimates have been derived from the experience of NHLBI's Framingham Heart Study, a predominantly Caucasian population in Massachusetts, USA. The risk algorithm may not fit other populations quite as well.

### Step 1

AGE			
Years	Points	Years	Points
20-34	-9	55-59	8
35-39	-4	60-64	10
40-44	0	65-69	11
45-49	3	70-74	12
50-54	6	75-79	13

### Step 2

TC (mg/dL)	TOTAL CHOLESTEROL				
	Points				
	Age 20-39 y	Age 40-49 y	Age 50-59 y	Age 60-69 y	Age 70-79 y
<160	0	0	0	0	0
160-199	4	3	2	1	0
200-239	7	5	3	1	0
240-279	9	6	4	2	1
≥280	11	8	5	3	1

### Step 3

	SMOKING				
	Points				
	Age 20-39 y	Age 40-49 y	Age 50-59 y	Age 60-69 y	Age 70-79 y
Nonsmoker	0	0	0	0	0
Smoker	8	5	3	1	1

### Step 4

HDL CHOLESTEROL	
HDL-C (mg/dL)	Points
≥60	-1
50-59	0
40-49	1
<40	2

### Step 5

BLOOD PRESSURE		
Systolic BP (mm Hg)	Points If Untreated	Points If Treated
<120	0	0
120-129	0	1
130-139	1	2
140-159	1	2
≥160	2	3

### Step 6

ADDING UP THE POINTS	
(Sum from Steps 1-5)	
Age	
TC	
Smoker	
HDL-C	
BP	
<b>Point Total</b>	

### CHD RISK

DETERMINE CHD RISK FROM POINT TOTAL	
Point Total	10-year CHD Risk
<0	<1%
0	1%
1	1%
2	1%
3	1%
4	1%
5	2%
6	2%
7	3%
8	4%
9	5%
10	6%
11	8%
12	10%
13	12%
14	16%
15	20%
16	25%
≥17	≥30%

Your chance of developing  
CHD (angina or heart attack)  
over the next 10 years is:

2%

\*NCEP Expert Panel. Third report of the Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) Executive Summary. Available at [http://www.nhlbi.nih.gov/guidelines/cholesterol/tp\\_iii.htm](http://www.nhlbi.nih.gov/guidelines/cholesterol/tp_iii.htm). Accessed May 31, 2001.

48 years old  
LDL = 9  
CRP normal

# Should John have

triglycerides = 106  
LDL = 70

# an angiogram?

father died of heart d

university professor  
calcium score in 2003 = 19  
calcium score in 2008 = 42  
calcium score in 2010 = 70

# Clinical Judgment

BMI = 21.6



no diabetes  
stress test normal in 2007  
EKG unusual in 2009  
mother died of cancer (83)

aspirin  
arrhythmia in 2008  
normal heart ultrasound (2008)

# Who are we kidding

genotyping

## Risk Calculator

(Click a question number for a brief explanation, or [read all explanations.](#))

1. Does the woman have a medical history of any breast cancer or of ductal carcinoma in situ (DCIS) or lobular carcinoma in situ (LCIS)?

---

2. What is the woman's age?  
*This tool only calculates risk for women 35 years of age or older.*

---

3. What was the woman's age at the time of her first menstrual period?

---

4. What was the woman's age at the time of her first live birth of a child?

---

5. How many of the woman's first-degree relatives - mother, sisters, daughters - have had breast cancer?

---

6. Has the woman ever had a breast biopsy?

  - 6a. How many breast biopsies (positive or negative) has the woman had?
  - 6b. Has the woman had at least one breast biopsy with atypical hyperplasia?

---

7. What is the woman's race/ethnicity?

  - 7a. What is the sub race/ethnicity?

# Gail Breast Cancer Model

## Validation of the Gail et al. Model of Breast Cancer Risk Prediction and Implications for Chemoprevention

**Table 6.** Measures of discriminatory accuracy of the Gail et al. (1) model 2 in the total sample in the Nurses' Health Study and in a sample of women who reported screening within 1 year before 1992

<b>Total sample (n = 82 109; 1354 cases)</b>	<b>Recently screened sample* (n = 55 301; 941 cases)</b>
0.58 (0.56 to 0.60)	0.59 (0.57 to 0.61)

concordance coefficient



# Patient-centered predictive models are already in clinical practice

## Validation of Clinical Classification Schemes for Predicting Stroke

### Results From the National Registry of Atrial Fibrillation

Brian F. Gage, MD, MSc

Amy D. Waterman, PhD

William Shannon, PhD

Michael Boechler, PhD

Michael W. Rich, MD

Martha J. Radford, MD

**T**HE ATRIAL FIBRILLATION (AF) population is heterogeneous in terms of ischemic stroke risk.

Subpopulations have annual stroke rates that range from less than 2% to more than 10%.<sup>1-5</sup> Because the relative risk reductions from warfarin sodium (62%) and aspirin (22%) therapy are consistent across these subpopulations,<sup>2,6-8</sup> the absolute benefit of antithrombotic therapy depends on the underlying risk of stroke. Although there has been agreement that warfarin therapy is favored when the risk of stroke is high and that aspirin is favored when the risk of stroke is low,<sup>9,10</sup> there has been little agreement about how to predict the risk of stroke.<sup>11-13</sup>

Thus, an accurate, objective scheme to estimate the risk of stroke in the AF population would allow physicians and

**Context** Patients who have atrial fibrillation (AF) have an increased risk of stroke, but their absolute rate of stroke depends on age and comorbid conditions.

**Objective** To assess the predictive value of classification schemes that estimate stroke risk in patients with AF.

**Design, Setting, and Patients** Two existing classification schemes were combined into a new stroke-risk scheme, the CHADS<sub>2</sub> index, and all 3 classification schemes were validated. The CHADS<sub>2</sub> was formed by assigning 1 point each for the presence of congestive heart failure, hypertension, age 75 years or older, and diabetes mellitus and by assigning 2 points for history of stroke or transient ischemic attack. Data from peer review organizations representing 7 states were used to assemble a National Registry of AF (NRAF) consisting of 1733 Medicare beneficiaries aged 65 to 95 years who had nonrheumatic AF and were not prescribed warfarin at hospital discharge.

**Main Outcome Measure** Hospitalization for ischemic stroke, determined by Medicare claims data.

**Results** During 2121 patient-years of follow-up, 94 patients were readmitted to hospital for ischemic stroke (stroke rate, 4.4 per 100 patient-years). As indicated by a c statistic greater than 0.5, the 2 existing classification schemes predicted stroke better than chance: c of 0.68 (95% confidence interval [CI], 0.65-0.71) for the scheme developed by the Atrial Fibrillation Investigators (AFI) and c of 0.74 (95% CI, 0.70-0.76) for the Stroke Prevention in Atrial Fibrillation (SPAF) III scheme. However, a c statistic of 0.82 (95% CI, 0.80-0.84), the CHADS<sub>2</sub> index was the most accurate predictor of stroke. The stroke rate per 100 patient-years without antithrombotic therapy increased by a factor of 1.5 (95% CI, 1.3-1.7) for each 1-point increase in the CHADS<sub>2</sub> score: 1.9 (95% CI, 1.2-3.0) for a score of 0; 2.8 (95% CI, 2.0-3.8) for 1; 4.1 (95% CI, 3.1-5.1) for 2; 5.9 (95% CI, 4.6-7.3) for 3; 8.5 (95% CI, 6.3-11.1) for 4; 12.2 (95% CI, 8.2-17.5) for 5; and 18.2 (95% CI, 10.5-27.4) for 6.



**Conclusion** The 2 existing classification schemes and especially a new stroke-risk index, CHADS<sub>2</sub>, can quantify risk of stroke for patients who have AF and make the selection of antithrombotic therapy.

JAMA. 2001;285:2864-2870

www.jama.com

- CHADS2 for patients with atrial fibrillation:
- +1 Congestive heart failure
  - +1 Hypertension
  - +1 Age >= 75
  - +1 Diabetes mellitus
  - +2 History of transient ischemic attack

Settings **CHADS2 Score for...** ⓘ




 **CardioMath®**  
70 cardiology calculators on your iPhone 

**Input**

<b>CHF</b>	<input type="checkbox"/>
<b>Hypertension</b>	<input type="checkbox"/>
<b>Age &gt;=75</b>	<input type="checkbox"/>
<b>Diabetes</b>	<input type="checkbox"/>
<b>Stroke/TIA (prior)</b>	<input type="checkbox"/>

**Result**

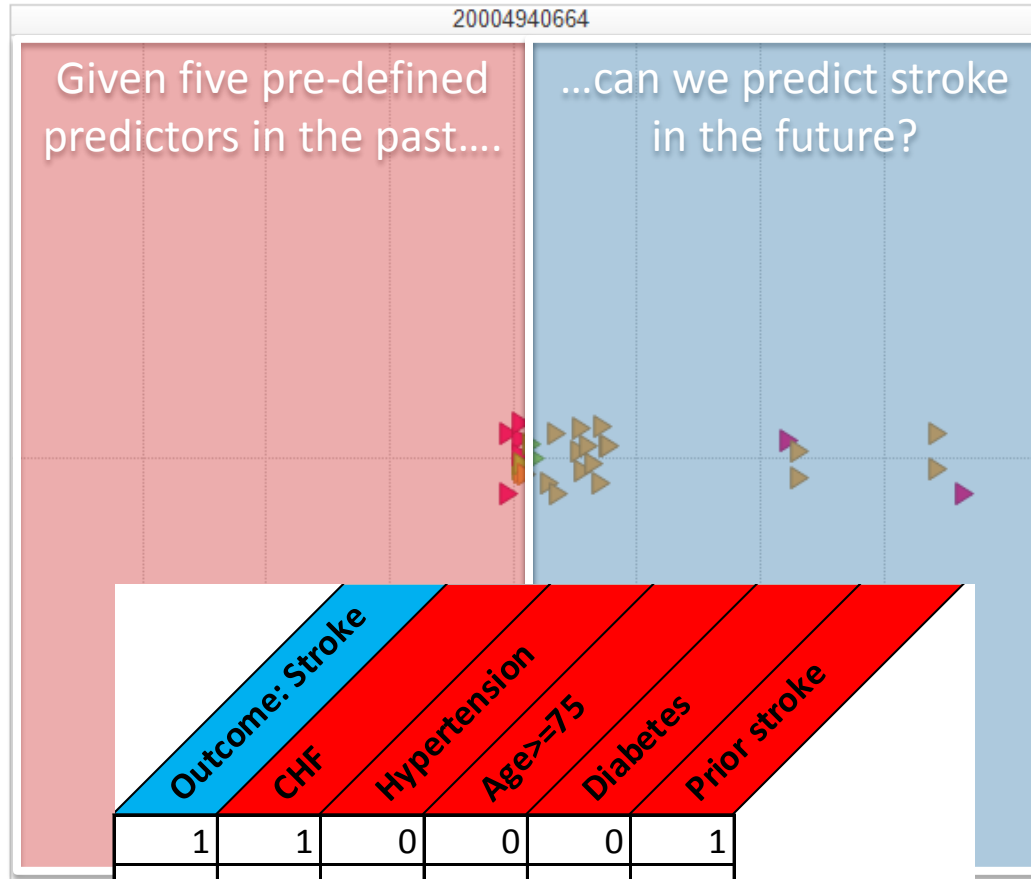
<b>CHADS2 Score</b>	0
---------------------	---

 **Calculator**     **Information**     **References**

# Applying CHADS2 to a patient

TABLE\_NAME

CONDITION\_OCCURRENCE



Color by  
 CONCEPT\_NAME

- Atrial fibrillation
- Congestive heart failure
- Essential hypertension

	Outcome: Stroke	CHF	Hypertension	Age>=75	Diabetes	Prior stroke
	1	1	0	0	0	1
	0	1	1	0	0	0
	0	1	1	1	0	1
	1	1	1	0	1	0
	0	0	1	0	0	0
	1	1	1	1	0	0
	0	0	0	1	1	0

# Evaluating the predictive accuracy of CHADS2

**Table 2.** Risk of Stroke in National Registry of Atrial Fibrillation (NRAF) Participants, Stratified by CHADS<sub>2</sub> Score\*

CHADS <sub>2</sub> Score	No. of Patients (n = 1733)	No. of Strokes (n = 94)	NRAF Crude Stroke Rate per 100 Patient-Years	NRAF Adjusted Stroke Rate, (95% CI)†
0	120	2	1.2	1.9 (1.2-3.0)
1	463	17	2.8	2.8 (2.0-3.8)
2	523	23	3.6	4.0 (3.1-5.1)
3	337	25	6.4	5.9 (4.6-7.3)
4	220	19	8.0	8.5 (6.3-11.1)
5	65	6	7.7	12.5 (8.2-17.5)
6	5	2	44.0	18.2 (10.5-27.4)

JAMA, 2001; 285: 2864-2870

AUC = 0.82 (0.80 – 0.84)

## Validation of the CHADS<sub>2</sub> clinical prediction rule to predict ischaemic stroke

### A systematic review and meta-analysis

Claire Keogh; Emma Wallace; Ciara Dillon; Borislav D. Dimitrov; Tom Fahey

Royal College of Surgeons, Dublin, Ireland

Thromb Haemost 2011; 106: 528–538

#### Summary

The CHADS<sub>2</sub> predicts annual risk of ischaemic stroke in non-valvular atrial fibrillation. This systematic review and meta-analysis aims to determine the predictive value of CHADS<sub>2</sub>. The literature was systematically searched from 2001 to October 2010. Data was pooled and analysed using discrimination and calibration statistical measures, using a random effects model. Eight data sets (n=2815) were included. The diagnostic accuracy suggested a cut-point of  $\geq 1$  has higher sensitivity (92%) than specificity (12%) and a cut-point of  $\geq 4$  has higher specificity (96%) than sensitivity (33%). Lower summary estimates were observed for cut-points  $\geq 2$  (sensitivity 79%, specificity 42%) and  $\geq 3$  (specificity 77%, sensitivity 50%). There was insufficient data to analyse cut-points  $\geq 5$  or  $\geq 6$ . Moderate pooled c statistic values were identified for the classic (0.63, 95% CI 0.52–0.75) and revised (0.60, 95% CI 0.43–0.72) view of stratification of the CHADS<sub>2</sub>. Calibration analysis in-

dicated no significant difference between the predicted and observed strokes across the three risk strata for the classic or revised view. All results were associated with high heterogeneity, and conclusions should be made cautiously. In conclusion, the pooled c statistic and calibration analysis suggests minimal clinical utility of both the classic and revised view of the CHADS<sub>2</sub> in predicting ischaemic stroke across all risk strata. Due to high heterogeneity across studies and low event rates across all risk strata, the results should be interpreted cautiously. Further validation of CHADS<sub>2</sub> should perhaps be undertaken, given the methodological differences between many of the available validation studies and the original CHADS<sub>2</sub> derivation study.

AUC = 0.63 (0.52 – 0.75)

## Is CHADS2 as good as we can do?

- What about other measures of CHADS2 predictors?
  - Disease severity and progression
  - Medication adherence
  - Health service utilization
- What about other known risk factors?
  - Hypercholesterolemia
  - Atherosclerosis
  - Anticoagulant exposure
  - Tobacco use
  - Alcohol use
  - Obesity
  - Family history of stroke
- What about other unknown risk factors?

# High-dimensional analytics can help reframe the prediction problem



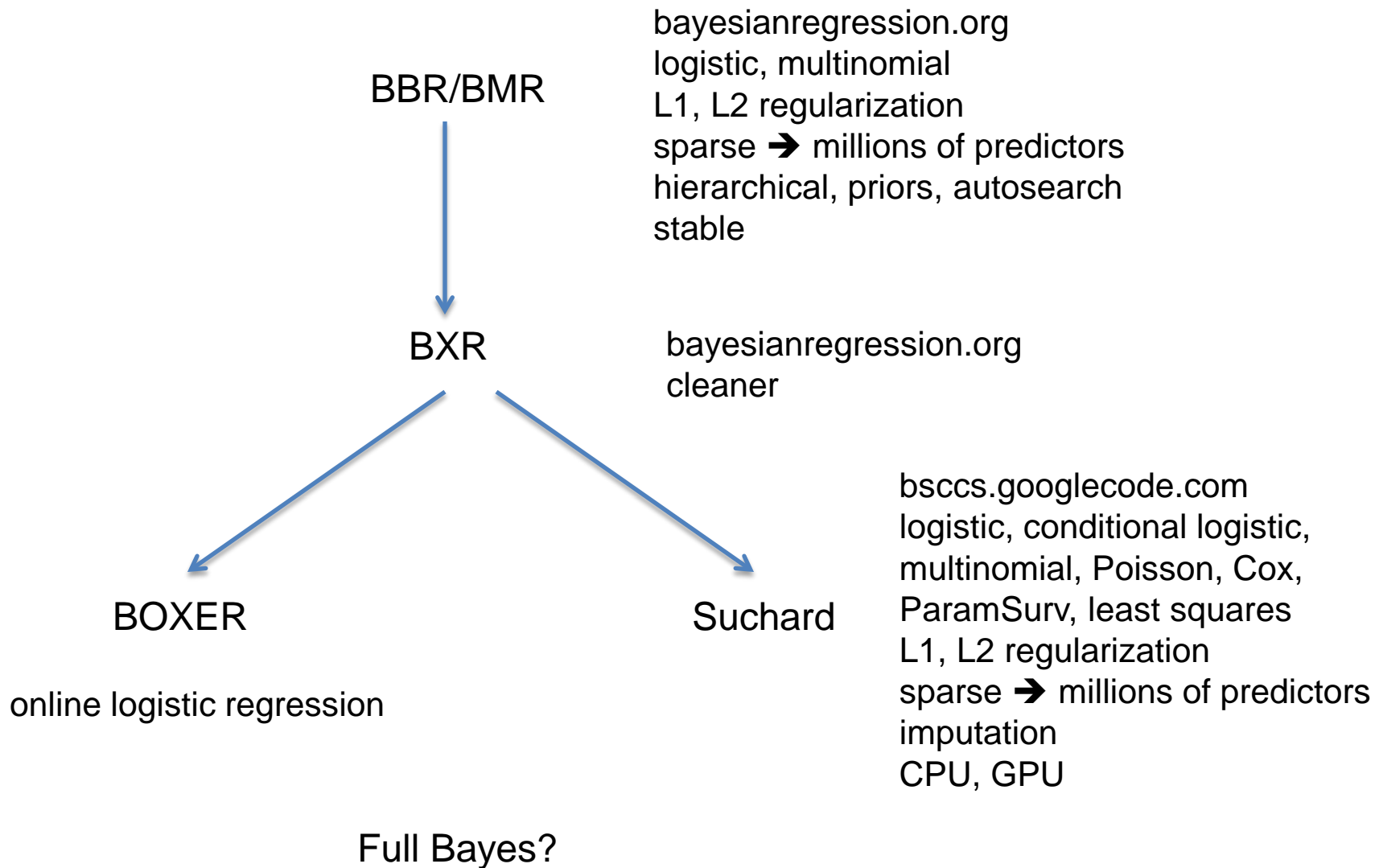
Outcome: Stroke	Age	Gender	Race	Location	Drug 1	Drug 2	...	Drug n	Condition 1	Condition 2	...	Condition n	Procedure 1	Procedure 2	...	Procedure n	Lab 1	Lab 2	...	Lab n
-----------------	-----	--------	------	----------	--------	--------	-----	--------	-------------	-------------	-----	-------------	-------------	-------------	-----	-------------	-------	-------	-----	-------

0	76	M	B	441	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	77	F	W	521	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
1	96	F	B	215	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0
1	76	F	B	646	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	64	M	B	379	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	74	M	W	627	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
1	68	M	B	348	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

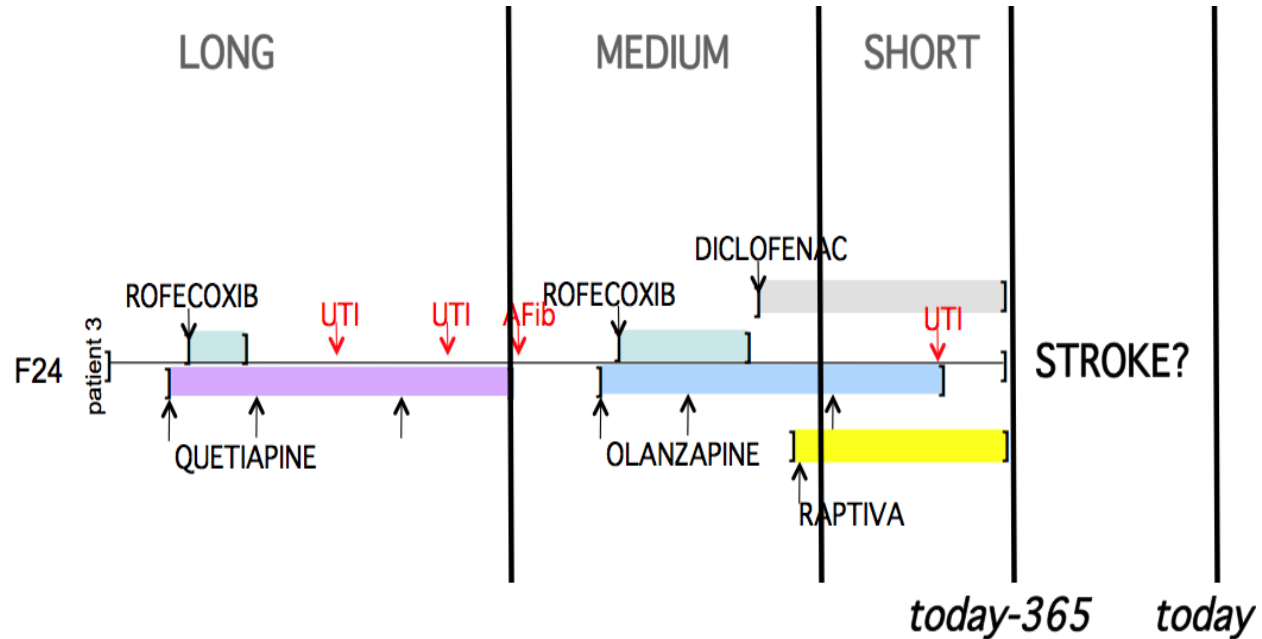
Modern predictive modeling techniques, such as Bayesian logistic regression, can handle millions of covariates. The challenge is creating covariates that might be meaningful for the outcome of interest

Demographics	All drugs	All conditions	All procedures	All lab values
--------------	-----------	----------------	----------------	----------------

# Tools for Large-Scale Regression



# Methodological Challenges



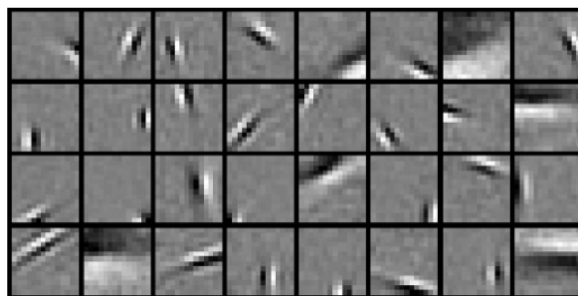
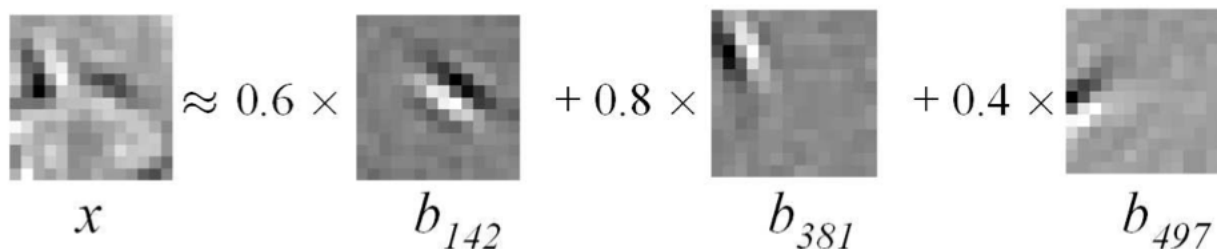
**Central challenge: how to extract features from a longitudinal health record?**



# Sparse Coding: Learning Good Features

- Express each input vector as a linear combination of basis vectors
- Learn the basis *and* the weights:

$$\operatorname{argmin}_{a,b} \sum_i \left\| x^i - \sum_j a_j^i b_j \right\|_2^2 + \beta \|a^i\|_1 \text{ such that } \|b_j\|_2 \leq 1, j = 1, \dots, s, i = 1, \dots, n.$$



- Supervised sparse coding

# Decision Tree Approach

(>-30, appendectomy, Y/N):

in the last 30 days, did the patient have an appendectomy?

(<0, max(SBP), 140):

at any time in the past did the patient's systolic blood pressure exceed 140 mmHg?

(<-90, rofecoxib, Y/N):

in the time period up to 90 days ago, did the patient have a prescription for rofecoxib?

(>-7, fever, Y/N):

in the last week, did the patient have a fever?

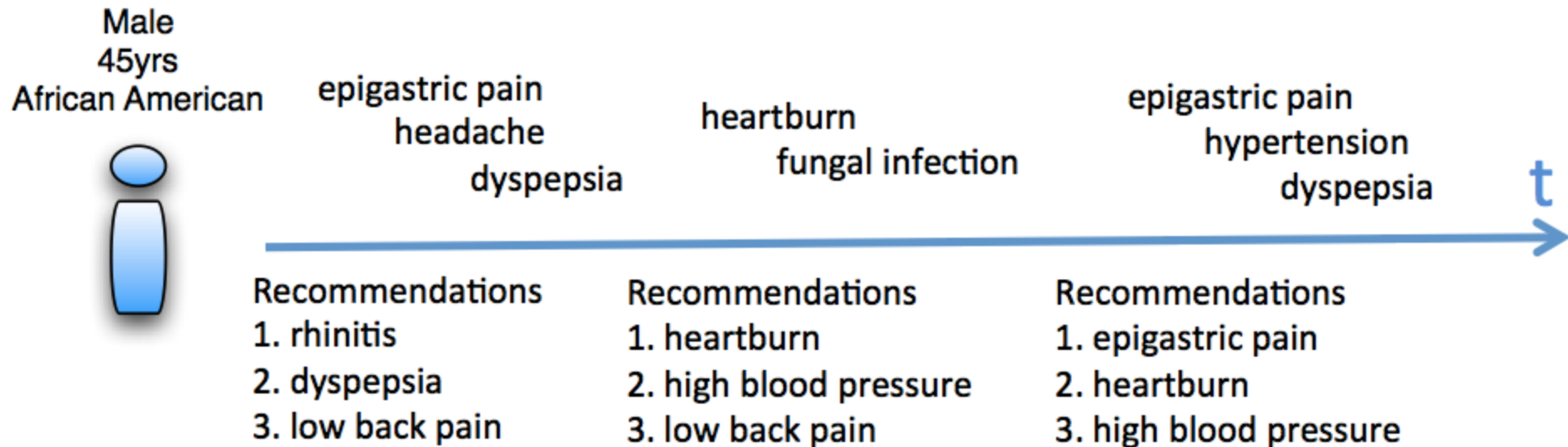
# Rule Mining

McCormick, Rudin, Madigan

- Goal: Predict next event in current sequence given sequence database
- Association Rules:
  - item 1 and item 2 → item 3
  - Recommender systems
  - Built-in explanation
- (Bayesian) Hierarchical Association Rule Mining

# Predicting Medical Conditions

- Patients visit providers periodically
- Report time-stamped series of conditions since last encounter
- Predict next condition given past sequences



- ▶ Observe  $y_{ir}$  co-occurrences (support for lhs  $\cup$  rhs) for patient  $i$  and rule  $r$
- ▶  $n_{ir}$  encounters that include the lhs
- ▶ Hierarchical Association Rule Model (HARM)

$$y_{ir} \sim \text{Binomial}(n_{ir}, p_{ir})$$

$$p_{ir} \sim \text{Beta}(\pi_{ir}, \tau_i)$$

- ▶ Model  $\pi_{ir}$  hierarchically

$$\pi_{ir} = \exp(\mathbf{M}'_i \beta_r + \gamma_i)$$

- ▶  $\mathbf{M}$  is matrix of patient characteristics,  $\gamma_i$  is patient-specific variation

# HARM

- Performed well in a number of experiments
- See Tyler's poster for details

# Why patient-centered analytics holds promise

## Average treatment effects:

- Hundreds of drug-outcome pairs
- Unsatisfactory ground truth:
  - how confident are we that drug is associated with outcome?
  - What is ‘true’ effect size?
- Questionable generalizability: who does the average treatment effect apply to?
- Final answer often insufficient:
  - Need to drilldown to explore treatment heterogeneity
  - Truth about ‘causality’ is largely unobtainable

## Patient-centered predictions:

- Millions of patients
- Explicit ground truth
  - Each patient did or did not have the outcome within the defined time interval
- Direct applicability: model computes probability for each individual
- Final model can address broader questions:
  - Which patients are most at risk?
  - What factors are most predictive of outcome?
  - How much would change in health behaviors impact risk?
  - What is the average treatment effect?

# Concluding thoughts

- Not all patients are created equally...
  - Average treatment effects are commonly estimated from observational databases, but the validity and utility of these estimates remains undetermined
  - Patient-centered predictive modeling offers a complementary perspective for evaluating treatments and understanding disease
- ...but all patients can equally benefit from the potential of predictive modeling in observational data
  - Clinical judgment may be useful, but selecting of a handful of predictors is unlikely to maximize the use of the data
  - High-dimensional analytics can enable exploration of high-dimensional data, but further research and evaluation is needed
  - Empirical question still to be answered: Which outcomes can be reliably predicted using which models from which data?