

# **Drug Safety Data Mining with a Tree-Based Scan Statistic**

Martin Kulldorff

Department of Population Medicine  
Harvard Medical School and  
Harvard Pilgrim Health Care Institute

# Next Hour: What to Expect

- Background: Drug safety surveillance and data mining
- Method: Tree-based scan statistic
- Pilot Studies: Safety of diabetes and anti-fungal drugs; Risk of Heart Attack
- Future: Evaluating the majority of recently approved drugs

# Drug Safety Surveillance

Drug Safety: Are there pharmaceutical drugs that causes adverse events?

Surveillance: Keeping a watchful eye for adverse events.

# Data Mining

Wikipedia: Data mining is the process of extracting hidden patterns from data. As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of applications, such as marketing, fraud detection and scientific discovery.

# Data Mining

Simultaneous evaluation of a huge number of possible drug/adverse event pairs

Purpose is to identify potential problems

Any findings must be thoroughly investigated using traditional pharmacoepidemiological study designs

# Surveillance “Signals”

An adverse event ‘signal’ does not necessarily mean that there is a relationship and a relationship does not necessarily mean that there is causation

# Data Mining: Three Methodological Issues

- Calculating Expected Counts
- Adjusting for Multiple Testing
- Granularity: Is there increased risk for a very specific diagnosis (liver failure), or for a range of related diagnoses (liver problems)?

# Post-Marketing Drug Safety Data

- Phase IV randomized clinical trials
- Case-control studies
- Spontaneous adverse event reporting systems
- Routinely collected electronic health data

# The Tree-Based Scan Statistic

Kulldorff M, Fang Z, Walsh S. A tree-based scan statistic for database disease surveillance. *Biometrics*, 2003,59:323-331.

# Granularity: Nested Variables

ecotrin  $\subset$  aspirin  $\subset$  nonsteroidal

anti-inflammatory drugs  $\subset$  analgesic drugs

acute lymphoblastic leukemia  $\subset$  acute

leukemias  $\subset$  leukemia  $\subset$  cancer

# Occupational Disease Surveillance

Some professional people may be at higher risk of certain diseases.

Surveillance: Keeping a watchful eye for unsuspected relationships.

# Occupational Multiple Cause of Death Database

- National Center for Health Statistics
- Based on Death Certificates
- Occupational Classification System
- Selected States

# Occupational Multiple Cause of Death Database

- Time period: 1985-1992
- Age groups:  $\geq 25$  years
- Total deaths: 2,114,832
- Silicosis deaths: 405

# Occupational Classification System

A hierarchical structure of occupations created by the United States Bureau of the Census.

Number of occupational groups at each level:

Level:	1	2	3	4	5	6	7
	6	13	86	345	476	502	503

# Occupational Classification System

Managerial and Professional Specialty Occupations

Professional Specialty Occupations

Mathematical and Computer Scientists

Computer Systems Analysts and Scientists (064)

Operations and Systems Researchers and Analysts (065)

Actuaries (066)

Statisticians (067)

Mathematical Scientists, n.e.c. (068)

Natural Scientists

Medical Scientists (083), etc.

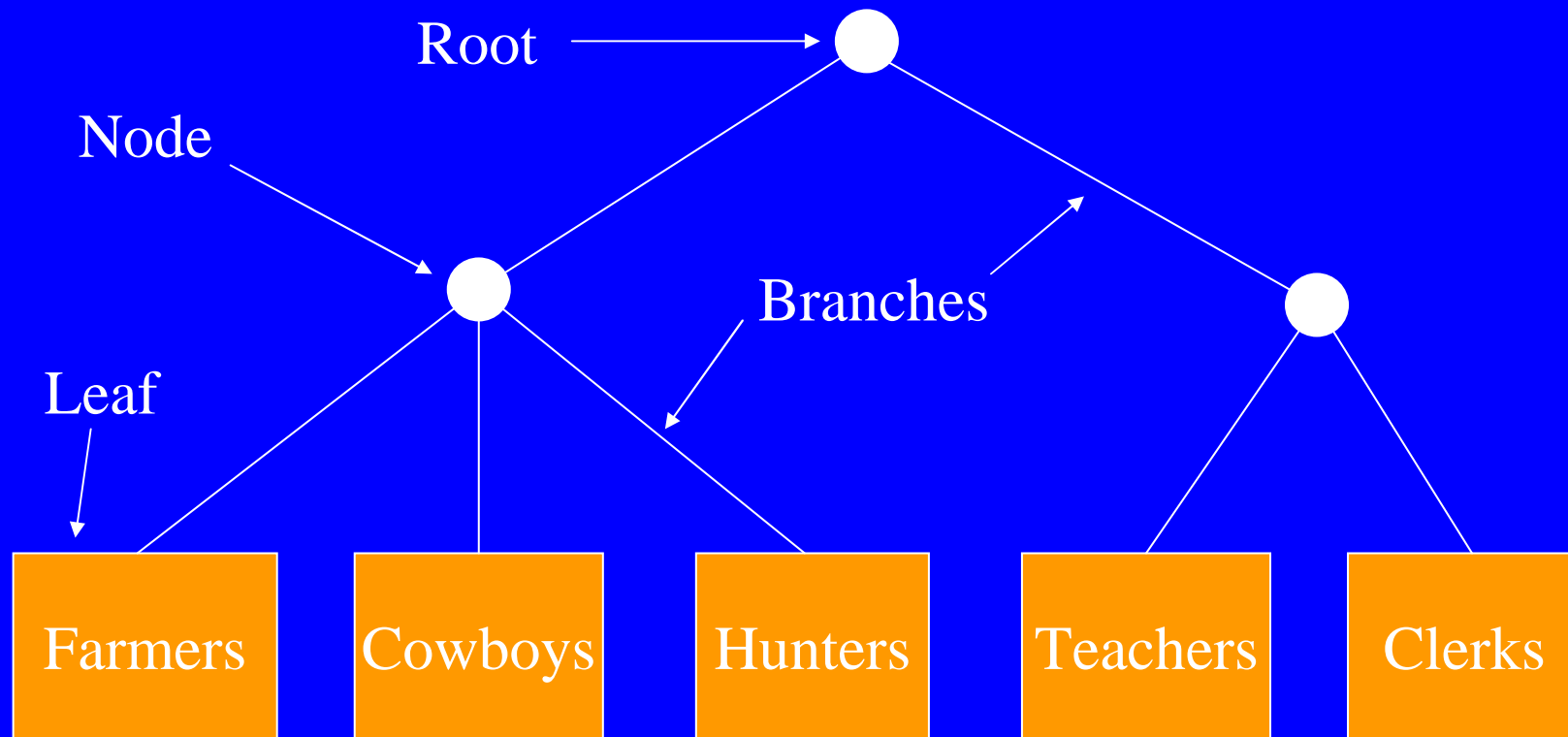
Health Diagnosing Occupations

Physicians (084), etc.

Health Assessment and Treatment Occupations

Therapists (098-105), etc.

# A Small Two-Level Tree Variable



# Silicosis

- A rare disease of the lung
- Chronic shortness of breath
- Caused by dust containing crystalline silica (quartz) particles
- No known cure

# Silicosis

Described by Agricola in 1556:

‘In the Carpathian mines, women are found who have married seven husbands, all of whom this terrible consumption has carried away’

Agricola G. (1556). *De Re Metallica*. Basel: Froben and Episopus.

# Analysis Options

- Evaluate each of the 503 occupational groups, using a Bonferroni type adjustment for multiple testing.
- Use a higher group level, such as level 3 with 86 occupational groups.

Problem: We do not know whether the disease relationships effect a smaller or larger group.

# Analysis Options

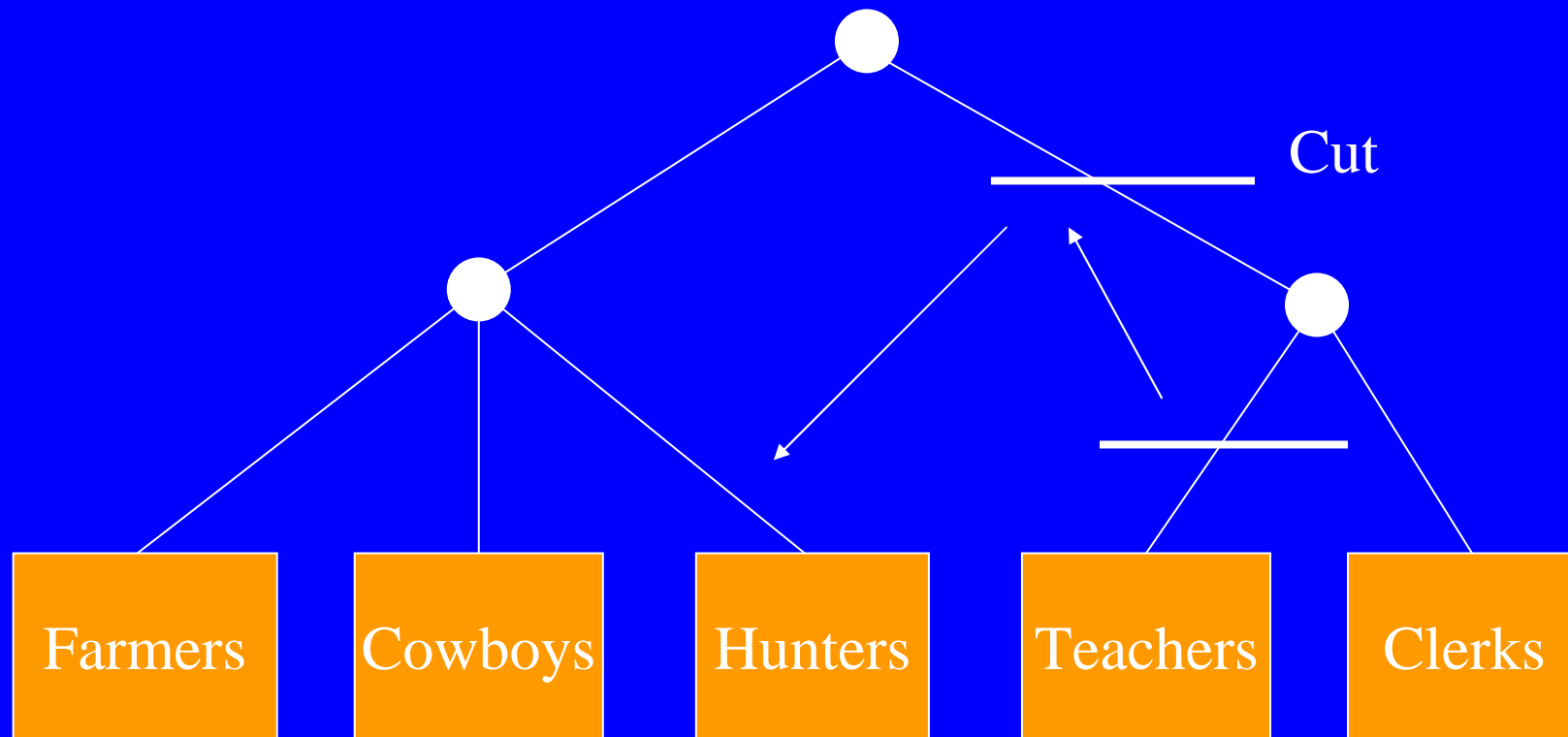
- Take the 503 occupations as a base, and evaluate all  $2^{503} - 2 = 2.6 \times 10^{151}$  combinations.

Problem: Not all combinations are of interest.

# Ideal Analytical Solution

- Use the Hierarchical Tree
- Evaluate Cuts on that Tree

# A Small Three-Level Tree Variable



# Problem

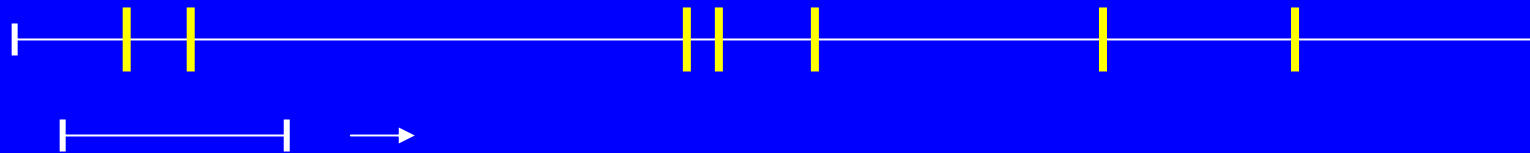
How do we deal with the multiple testing?

# Proposed Solution

Tree-Based Scan Statistic

# One-Dimensional Scan Statistic

Studied by Naus (JASA, 1965)



## Other Scan Statistics

- Spatial scan statistics using circles or squares.
- Space-time scan statistics using cylinders, for the early detection of disease outbreaks.
- Variable size window, using maximum likelihood rather than counts.
- Applied for geographical and temporal disease surveillance, and in many other fields.

# Tree-Based Scan Statistic

$H_0$ : The probability of dying from silicosis is the same for all occupations.

$H_A$ : There is at least one group of occupations (cut) for which the probability is higher.

# Tree-Based Scan Statistic

1. Scan the tree by considering all possible cuts on any branch.
2. For each cut, calculate the likelihood.
3. Denote the cut with the maximum likelihood as the most likely cut (cluster).
4. Generate 9999 Monte Carlo replications under  $H_0$ .
5. Compare the most likely cut from the real data set with the most likely cuts from the random data sets.
6. If the rank of the most likely cut from the real data set is  $R$ , then the p-value for that cut is  $R/(9999+1)$ .

# Result

## Most Likely Cut

Occupations: Mining machine operators

Observed: 56, Expected: 5.5

SPMR = 11.8,  $p=0.0001$

# Result:

## Second Most Likely Cut

Occupations: Molding and casting machine operators, Metal plating machine operators, Heat treating equipment operators, Misc. metal and plastic machine operators

Observed: 22, Expected: 1.2

SPMR = 20.5,  $p=0.0001$

# Result

## Ninth Most Likely Cut

Occupation: Heavy equipment mechanics

Observed: 5, Expected: 1.0

SPMR = 4.8,  $p=0.72$

# Extension to Complex Cuts

Consider a node with 4 branches: A, B, C, D.

Simple cuts: [A], [B], [C], [D]

Combinatorial cuts: [A], [B], [C], [D]

[AB], [AC], [AD], [BC], [BD], [CD]

[ABC], [ABD], [ACD], [BCD]

Ordinal cuts: [A], [B], [C], [D]

[AB], [BC], [CD], [ABC], [BCD]

# Result

## Most Likely Cut

Occupations: Mining machine operators,  
Mining occupations n.e.c

Observed: 59, Expected: 6.0

SPMR = 11.5,  $p=0.0001$

# Extension to Multiple Trees

There may not be one unique suitable tree.

It is trivial to extend the method to multiple trees, by simply scanning over all trees.

# Result

## Most Likely Cut

Occupations: Mining machine operators,  
Mining engineers, Mining occupations n.e.c

Observed: 60, Expected: 6.0

SPMR = 11.6,  $p=0.0001$

# Evaluated Combinations

Simple cuts:	~1,000
Mixed cuts:	~1,000,000
Two trees:	~1,000,000

$< 2.6 \times 10^{151}$

# Comparison with Computer Assisted Regression Trees (CART)

Similarity:

The letters 'T', 'R', 'E' and 'E'.

Both are Data Mining Methods

# Difference

**CART:** There are multiple continuous or categorical variables, and a regression tree is constructed by making a hierarchical set of splits in the multi-dimensional space of the independent variables.

**Tree-Based Scan Statistic:** There may be only one independent variable (e.g. occupation). Rather than using this as a continuous or categorical variable, it is defined as a tree structured variable. That is, we are not trying to estimate the tree, but use the tree as a new and different type of variable.

# Drug Safety Surveillance

- Drug safety surveillance is important, since some drugs may cause unsuspected adverse events (e.g. Thalidomide)
- Use HMO data on drug dispensings and diagnoses of potential adverse events

# Pilot Studies

The tree may be drugs or adverse events:

1. For a particular diagnosis, evaluate all drugs, using a drug tree
2. For a particular drug, evaluate all outcomes, using a diagnosis tree

# Pilot Study I: Acute Myocardial Infarction (AMI)

Co-Investigators: Jeffrey Brown, Inna Dashevsky, Richard Platt, Robert Davis, David Graham, Arnold Chan

Funding: AHRQ through HMO Research Network CERT

Status: Preliminary results

[ AMI = Heart attack ]

# Scanning Drugs

- Select a specific adverse event or a group of related adverse events
- Scan a tree of drugs

# HMO Data

- Sample of Harvard Pilgrim Health Care Data
- 376,000 patients
- Years 1999-2003
- 2755 AMI diagnoses

# Drug Tree

Based on American Society for Health-System  
Pharmacists (AHFS) Classification

Level 1, with 18 groups:

- Antihistamine Drugs (04)
- Anti-infective Agents (08)
- Antineoplastic Agents (10)
- Autonomic Drugs (12)
- Blood Formation and Coagulation Drugs (20)
- Cardiovascular Drugs (24)

etc

# Drug Tree

Level 2:

Anti-infective Agents (08)

- Amebicides (0804)
- Anthelmintics (0808)
- Antibacterials (0812)
- Antifungals (0814)
- Antimycobacterials (0816)

etc

# Drug Tree

Level 3:

Anti-infective Agents (08)

- Antibacterials (0812)
    - Aminoglycosides (081202)
    - Antifungal Antibiotics (081204)
    - Cephalosporins (081206)
    - Miscellaneous Lactams (081207)
- etc

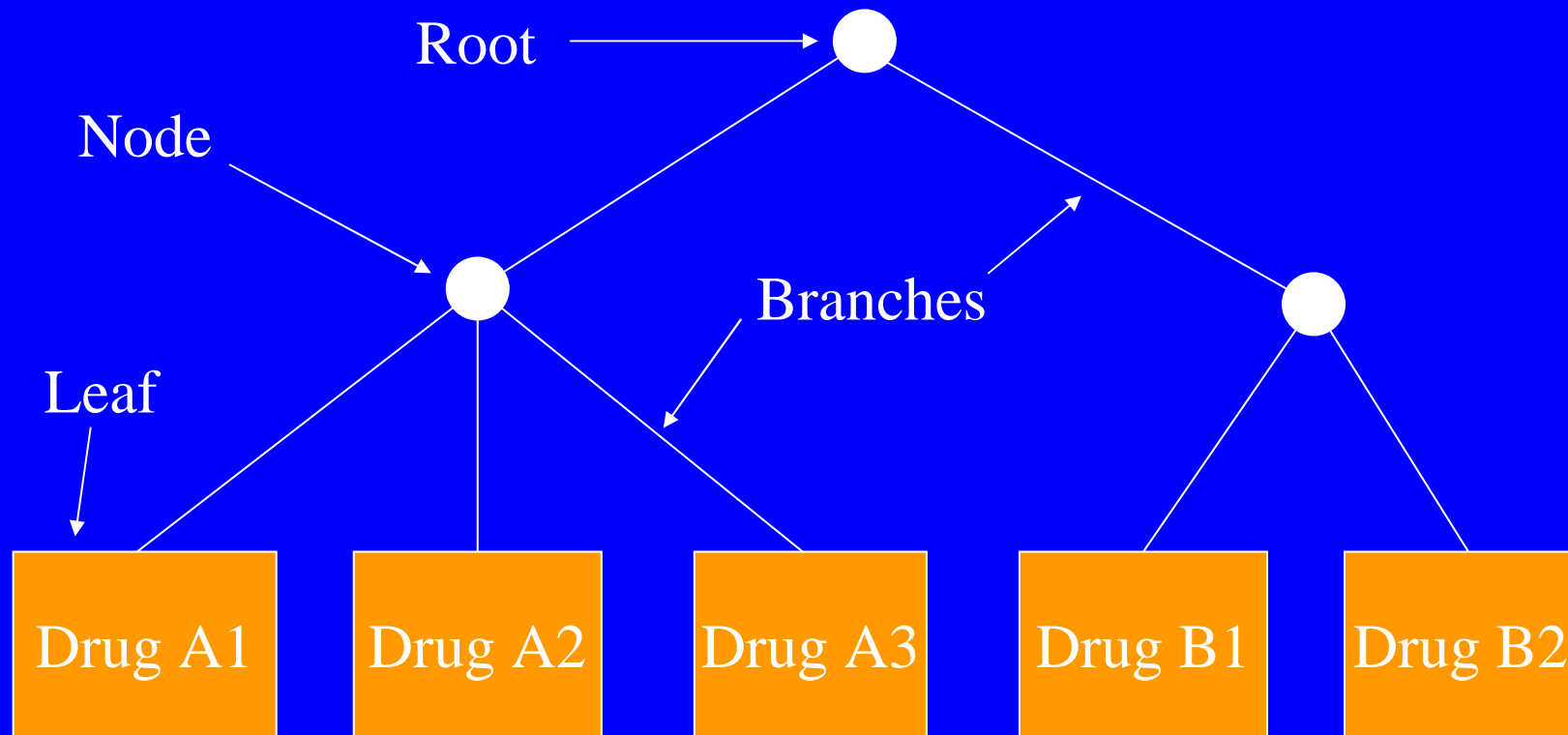
# Drug Tree

Level 5, generic drugs (1009 total):

Anti-infective Agents (08)

- Antibacterials (0812)
  - Aminoglycosides (081202)
    - Gentamicin (081202-0002)
    - Geomycin (081202-0004)
    - Tobramycin (081202-0007)

# A Small Two-Level Tree Variable



# Granularity Problem

## Analysis Options

- Evaluate each of the 1009 generic drugs, using a Bonferroni type adjustment for multiple testing.
- Use a higher group level, such as level 3 with 184 drug groups.
- Use the tree based scan statistic

# Tree-Based Scan Statistic

$H_0$ : The probability of a diagnosis after the dispensing of a drug is the same for all drugs.

$H_A$ : There is at least one group of drugs after which the probability of diagnosis is higher

. . . after various adjustments

# Tree-Based Scan Statistic

For each generic drug we have:

- observed number of diagnosed cases
- expected number of diagnosed cases, adjusted for age and gender

# Most Likely Cut

Drug(s): Nitrates and Nitrites (241208)

Observed: 98    Expected: 7.3    O/E=13.4

LLR = 165.0,    p=0.0001

# Second Most Likely Cut

Drug: Nitroglycerin (241208-0004)

Observed: 77, Expected: 6.2, O/E=12.5

LLR = 124.3, p=0.0001

# Results: Top 10 Cuts

Obs	Exp	O/E	LLR	Drug(s)
98	7.3	13.4	165.0	Nitrates and Nitrites (241208)
77	6.2	12.5	124.3	Nitroglycerin (241208-0004)
110	15.3	7.2	123.4	Vasodilating Agents (2412)
88	11.8	7.4	101.2	Adrenergic Blocking Agents (2424)
88	11.8	7.4	101.2	Adrenergic Blocking Agents (242400)
36	1.3	27.0	84.1	Clopidogrel (920000-0078)
209	74.6	2.8	83.6	Cardiovascular Drugs (24)
28	1.1	24.8	63.1	Isosorbide (241208-0003)
52	7.7	6.8	55.4	Atenolol (242400-0002)
32	2.9	10.9	47.5	Metoprolol (242400-0009)

p=0.0001, for all cuts

# Results, Tree Format

Obs	Exp	O/E	LLR	Drug(s)
209	74.6	2.8	83.6	Cardiovascular Drugs (24)
110	15.3	7.2	123.4	Vasodilating Agents (2412)
98	7.3	13.4	165.0	Nitrates and Nitrites (241208)
28	1.1	24.8	63.1	Isosorbide (241208-0003)
0	0.0002	0	-	Amyl (241208-0001)
77	6.2	12.5	124.3	Nitroglycerin (241208-0004)
5	6.7	0.7	-	<i>other 7 VA (2412xx)</i>
88	11.8	7.4	101.2	Adrenergic Block Agents (2424)
88	11.8	7.4	101.2	Adrenergic Block Agents(242400)
52	7.7	6.8	55.4	Atenolol (242400-0002)
32	2.9	10.9	47.5	Metoprolol (242400-0009)
4	1.0	3.9	-	<i>other 11 ABA (242400-xxxx)</i>
147	39.8	3.7	-	<i>other Cardiovascular Drugs (24xxxx)</i>

# Interpretation of Results

People with cardiovascular problems are often taking cardiovascular drugs and they are also at higher risk of AMI.

# Pilot Study II: Safety of Diabetes and Anti-Fungal Drugs

DPM: Martin Kulldorff, Jeffrey Brown, Inna Dashevsky, Taliser Avery, Richard Platt.

HMO Research Network: Robert Davis, Susan Andrade, Denise Boudreau, Margaret Gunter, Lisa J. Herrinton, Pam Pawloski, Marsha Raebel, Douglas Roblin.

FDA: David Graham

i3: Arnold Chan

# Pilot Study II: Safety of Anti-Fungal and Diabetes Drugs

Supported by grant HS10391 from the Agency for Healthcare Research and Quality (AHRQ) to the HMO Research Network Center for Education and Research in Therapeutics (CERT) in collaboration with the FDA through Cooperative Agreement FD-U-002068 .

# Study Population

- Sample from HMO Research Network
- Time period: 1999-2003
- Population size: 3,417,000 unique members
- Gender: 52% female
- Age: 34% under age 24, 11% above age 65
- One-year retention: ~80%

# HMO Research Network

Fallon Community Health Plan (Massachusetts)

Group Health Cooperative (Washington State)

Harvard Pilgrim Health Care (Massachusetts)

Health Partners (Minnesota)

Kaiser Permanente Colorado

Kaiser Permanente Georgia

Kaiser Permanente Northern California

Kaiser Permanente Northwest (Oregon)

Lovelace (New Mexico)

# Scanning Diagnoses

- For each analysis, select one drug or group of related drugs
- Scan a tree of diagnosis codes

# Antifungal Drugs

- Itraconazole: A triazole antifungal agent that is prescribed to patients with fungal infections.
- Terbinafine: A synthetic allylamine antifungal.

# Selection from the Multi-Level Clinical Classification Tree

07	'DISEASES OF THE CIRCULATORY SYSTEM'
07.01	'HYPERTENSION'
07.01.01	'ESSENTIAL HYPERTENSION [98.]'
07.01.01.00	'ESSENTIAL HYPERTENSION [98.]'
07.01.02	'HYPERTENSION WITH COMPLICATIONS AND SECONDARY HYPERTENSION [99.]'
07.01.02.01	'HYPERTENSIVE HEART AND/OR RENAL DISEASE'
07.01.02.02	'OTHER HYPERTENSIVE COMPLICATIONS'
07.02	'DISEASES OF THE HEART'
07.02.01	'HEART VALVE DISORDERS [96.]'
07.02.01.01	'CHRONIC RHEUMATIC DISEASE OF THE HEART VALVES'
07.02.01.02	'NONRHEUMATIC MITRAL VALVE DISORDERS'
07.02.01.03	'NONRHEUMATIC AORTIC VALVE DISORDERS'
07.02.01.04	'OTHER HEART VALVE DISORDERS'
07.02.02	'PERI-; ENDO-; AND MYOCARDITIS; CARDIOMYOPATHY (EXCEPT THAT CAUSED BY TB O
07.02.02.00	'PERI-; ENDO-; AND MYOCARDITIS; CARDIOMYOPATHY (EXCEPT THAT CAUSED BY TB O
07.02.02.01	'CARDIOMYOPATHY'
07.02.02.02	'OTHER PERI-; ENDO-; AND MYOCARDITIS'
07.02.03	'ACUTE MYOCARDIAL INFARCTION [100.]'
07.02.03.00	'ACUTE MYOCARDIAL INFARCTION [100.]'
07.02.04	'CORONARY ATHEROSCLEROSIS AND OTHER HEART DISEASE [101.]'
07.02.04.00	'CORONARY ATHEROSCLEROSIS AND OTHER HEART DISEASE [101.]'
07.02.04.01	'ANGINA PECTORIS'
07.02.04.02	'UNSTABLE ANGINA (INTERMEDIATE CORONARY SYNDROME)'
07.02.04.03	'OTHER ACUTE AND SUBACUTE FORMS OF ISCHEMIC HEART DISEASE'
07.02.04.04	'CORONARY ATHEROSCLEROSIS'
07.02.04.05	'OTHER FORMS OF CHRONIC HEART DISEASE'

# Some Diagnoses Removed

- Well-care visits
  - Accidents
  - Common infectious diseases
  - Cancer and other chronic diseases
  - Pregnancy
- etc

# Some Technical Details

- Drug exposure starts on first day of dispensing and ends 6 days after the end of supplies
- Only include incident diagnoses, without a similar diagnosis in the preceding 180 days.
- Only the first incident diagnosis per person is included
- Adjustment for age, gender and site.

# TreeScan Results: Itraconazole

## Diagnosis Based TreeScan Statistic: Itraconazole

Sample Size, # Adverse Events (Obs, Exp)		110	58			
Obs / Exp		1.90				
		Obs	Exp	Obs/Exp	Obs-Exp	TreeScan
12	<b>Skin and Subcutaneous Tissue</b>	31	11.2	2.76	19.8	0.00003
12.02	.. Other Inflammatory Conditions of Skin	9	2.1	4.23	6.9	0.02
12.04	.. Other Skin Disorders	19	8.2	2.33	10.8	0.06
13	<b>Muscoskeletal System &amp; Connective Tissues</b>					
13.01	.. Infective Arthritis & Osteomyelitis	4	0.1	56.41	3.9	0.00003
17	<b>Other</b>					
17.01.06	.... Nausea and Vomiting	6	0.9	6.37	5.1	0.03

# TreeScan Results: Terbinafine

## Diagnosis Based TreeScan Statistic: Terbinafine

Sample Size, # Adverse Events (Obs, Exp)		429	264			
		1.63				
		Obs	Exp	Obs/ Exp	Obs- Exp	TreeScan
06.07.01.00	. . . .cataract	21	9.9	2.12	11.1	-
09	<b>Digestive System</b>	63	37.4	1.68	25.6	0.008
09.08	. . Liver Disease	14	3.1	4.52	10.9	0.0001
09.08.02.04	. . . . . Other/Unspec. Liver Disorders	14	2.8	5.00	11.2	0.00002
12	<b>Skin and Subcutaneous Tissue</b>	125	52.0	2.40	73.0	0.00001
12.02.00.00	. . Other Inflammatory Conditions of Skin	25	10.7	2.35	14.3	0.01
12.04.00.00	. . Other Skin Disorders	95	37.3	2.55	57.7	0.00001
13.08.00.00	. . . .other connective tissue disease	59	42.9	1.38	16.1	-
17	<b>Other</b>					
17.01.06.00	. . . . Nausea and vomiting	10	3.9	2.56	6.1	-
17.01.09	. . . . Allergic Reactions	25	10.9	2.30	14.1	0.02

# Interpretation: Antifungal Drugs

- Out of 678 possible outcomes per drug, eight have  $p \leq 0.01$
- All signals can be explained as either known AEs (e.g., liver conditions) or confounding by indication (e.g., skin and subcutaneous tissue diagnoses).

# Diabetes Drugs

- **Glyburide** (Glibenclamide): An oral anti-diabetic drug in the sulfonylureas class.
- **Metformin**: An oral anti-diabetic drug in the biguanide class.
- **Pioglitazone** (Actos): A thiazolidinedione (TZD) with hypoglycemic action.
- **Rosiglitazone** (Avandia): A thiazolidinedione (TZD) that works as an insulin sensitizer.

# Results: Nervous System Organs

	Glyburide & Metformin		Pioglitazone		Rosiglitazone	
	Obs/Exp	P=	Obs/Exp	P=	Obs/Exp	P=
Eye Disorders	1.12	.00001	2.39	.00001	1.80	.00002
Cataract	1.09	.00008	2.25	.00001	1.88	.006
Glaucoma	1.23	.00001	2.74	.00001	1.63	..

# Interpretation

- Confounding by indication, since diabetes patients are at higher risk for these conditions.
- Higher risk for thiazolidinedione users may be because those two drugs are taken by individuals with more severe diabetes

# Selected Results: Genitourinary System

	Glyburide & Metformin		Pioglitazone		Rosiglitazone	
	Obs/Exp	P=	Obs/Exp	P=	Obs/Exp	P=
Urinary System	1.23	.00001	2.45	.00001	2.32	.00001
Nephritis, nephrosis, renal sclerosis	3.21	.00001	16.77	.00001	7.24	.02
Chronic renal failure	1.48	.01	7.66	.00001	8.27	,002
etc						

# Interpretation

- Pioglitazone has a 16-fold excess risk for “nephritis, nephrosis and renal sclerosis” and both pioglitazone and rosiglitazone have an eight-fold excess risk of chronic renal failure.
- These risks are considerably higher than for glyburide/metformin and could warrant further assessment.
- One possible explanation is confounding by indication since glyburide and metformin are contra indicated for people with renal dysfunction.

# Results: Congestive Heart Failure

- There were signals for congestive heart failure:

Glyburide & Metformin: 1.38 0.00001

Pioglitazone: 5.42 0.00001

Rosiglitazone: 5.88 0.00001

- Congestive heart failure is a known risk factor for pioglitazone and rosiglitazone

# Results: Cerebrovascular Disease

- For pioglitazone there was a signal for cerebrovascular disease [O/E=2.48, p=0.00001]
- Not present for the other drugs.
- May be appropriate for an in-depth evaluation

## Results: Other

- For all drugs, there were signals for several other cardiovascular disease codes and for chronic ulcer for leg and foot.
- These are likely due to confounding by indication

# Future: Evaluating the Majority of Recently Approved Drugs

National Library of Medicine Challenge Grant:  
Data Mining Electronic Health Records for  
Adverse Events

PopMed: Jeff Brown, Rich Platt, Inna Dashevsky,  
Beth Syat, Jessica Sturtevant, Mike Murphy

Kaiser Colorado: Marsha Raebel, David McClure

Kaiser Northern California: Bruce Fireman, Lisa  
Herrinton

# Existing Electronic Health Data

- HMO Research Network's Virtual Data Warehouse
- Harvard Pilgrim, Kaiser Colorado, Kaiser Northern California
- Time period: 2000-2008

# Data Mining Methods

- Tree-based scan statistic
- DuMouchel's Empirical Bayes gamma-Poisson shrinkage

# Covariate Adjustments

- Age
- Gender
- Site

# Drugs to be Evaluated

- A majority of the 200 new drugs approve by FDA since 2000.
- Selection criteria: Outpatient setting, taken orally, enough exposure
- Test phase: pioglitzaone, rosiglitazone, itraconazole and clopidogrel

# Expected Data Mining Signals

- We expect (hope) to find some signals due to known adverse events
- We expect to find many signals due to confounding by indication or contra-indication
- We may find signals that warrant a thorough evaluation using other methods and data

# Signal Evaluation

- Check data quality
- Stratified observed and expected counts by age, gender, site and other characteristics
- One dimensional scan statistic, looking at the time from initial exposure to the adverse event, to see if the events cluster
- Standard regression analysis using the same data with additional covariates

# Future: Data Mining for Anti-Viral Medications

- Lead investigator: Sharon Greene
- Funder: CDC, through the VSD project
- Data: Vaccine Safety Datalink
- Method: Tree-based scan statistic
- Start: Fall 2010

# Random Departing Thoughts

- HMO data shows promise for drug safety surveillance
- Calculating observed and expected counts is complex and critical
- Data mining generates signals that need to be confirmed/rejected using clinical knowledge and other methods

# Random Departing Thoughts

- The tree scan statistic can be used to solve the problems of granularity and multiple testing
- So far, this data mining work shows good promise
- Existing drug safety data mining systems have many weaknesses so the bar for success is quite low
- Many methodological fine tuning issues remain to be solved

# Random Departing Thoughts

- It is a little scary to do data mining for so many drugs simultaneously
- It's great fun!

# Random Departing Thoughts

- It is ~~a little~~ scary to do data mining for so many drugs simultaneously
- It's great fun!

# Random Departing Thoughts

- It is really scary to do data mining for so many drugs simultaneously
- It's great fun!