TreeScan™: A Novel Data-Mining Tool for Medical Product Safety Surveillance

Martin Kulldorff

Division of Pharmacoepidemiology and Pharmacoeconomics

Brigham and Women's Hospital and Harvard Medical School

Azadeh Shoaibi

Center for Biologics Evaluation and Research, U.S. Food and Drug Administration

Rima Izem

Center for Drug Evaluation and Research, U.S. Food and Drug Administration

Judith C. Maro

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

ICPE Disclosures

- Funding source: U.S. Food and Drug Administration
  - Under contract: FDA HHSF223201400030I, Task Order: HHSF22301003T
- No relationships to disclose
- The views expressed are the authors' and not necessarily those of the Food and Drug Administration, or the Department of Health and Human Services

Agenda

- Overview of Tree-based Scan Statistics
- TreeScan in Vaccine Safety Surveillance
- TreeScan in Drug Safety Surveillance
- Q&A
- Interactive Demonstration of TreeScan™ Software
- Signal Detection Exercise
- Q&A

Overview of Tree-based Scan Statistics

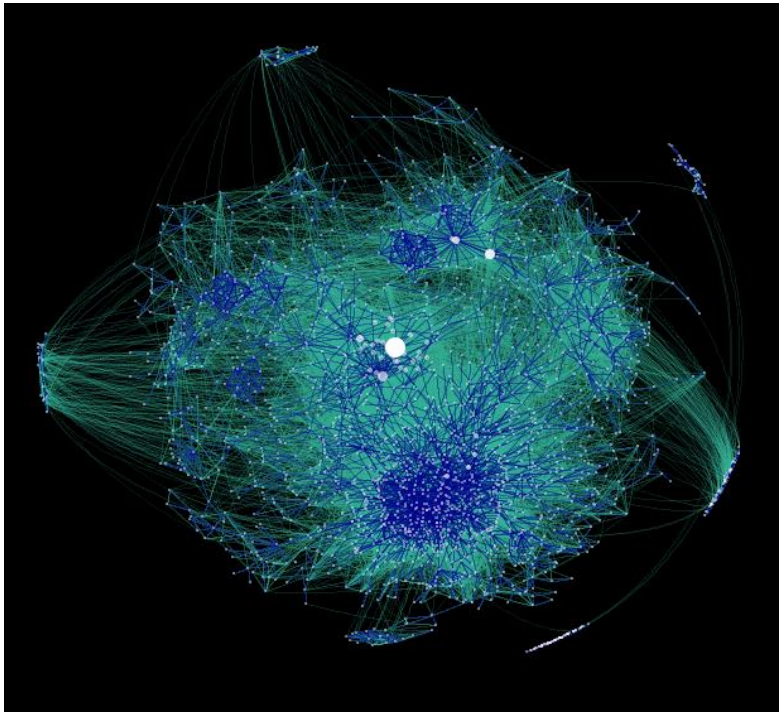
Martin Kulldorff

Division of Pharmacoepidemiology and Pharmacoeconomics

Brigham and Women's Hospital and Harvard Medical School

How can we detect unsuspected adverse reactions? How can we try to ensure that there are no unknown adverse reactions?

TreeScan Data Mining



Goal of TreeScan Method:

Close to complete ascertainment of adverse events

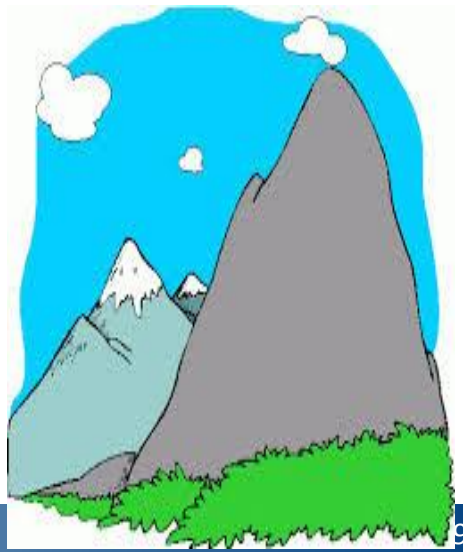
- Find known adverse reactions
- Find any additional adverse reactions , if they exist
- Few false positives, or else, easily explained false positives
- Sufficient sample size to detect very rare adverse reactions

Three Key Issues

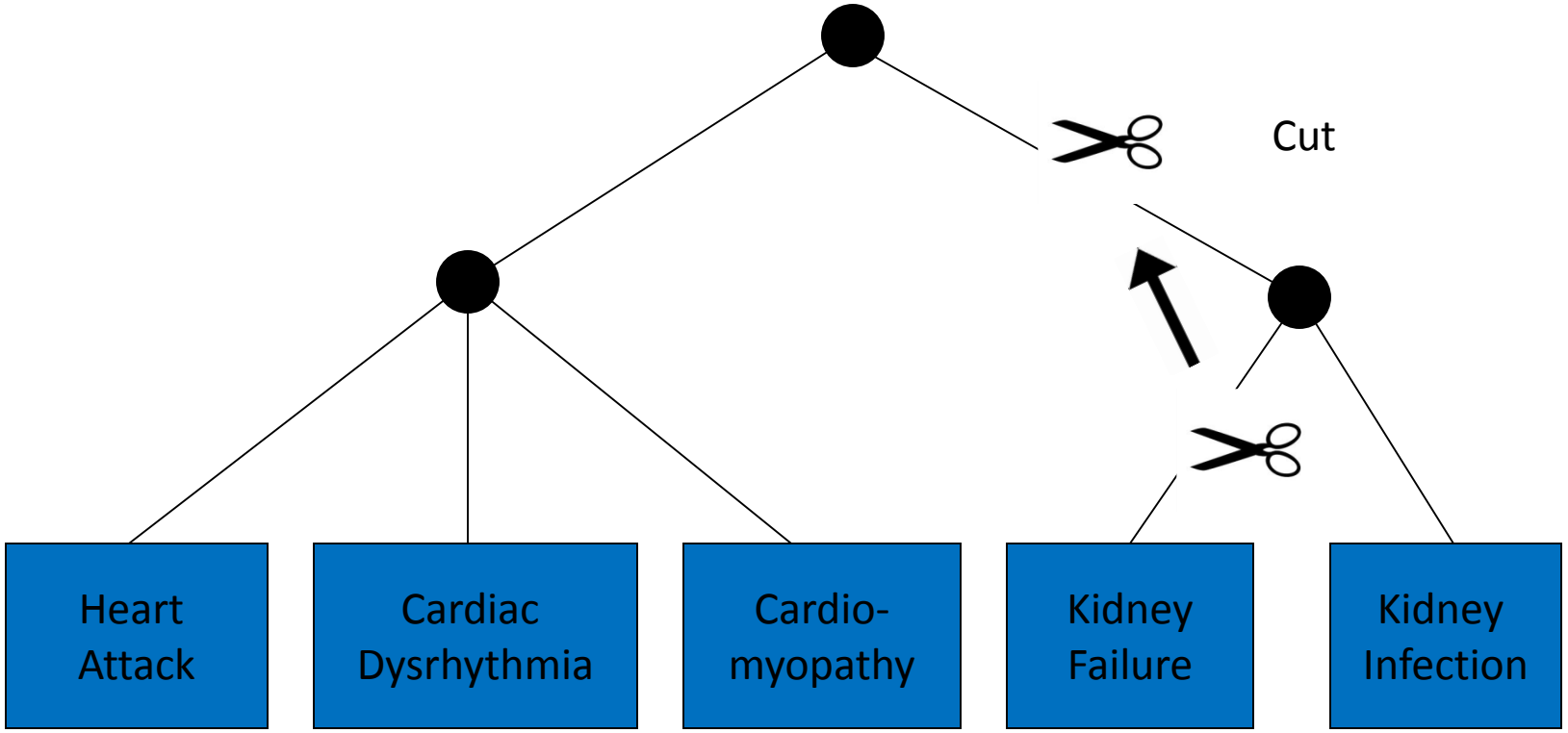
- Granularity
- Adjusting for Multiple Testing
- Choice of Comparison Group

Level of Granularity

Is there increased risk for a very specific diagnosis (acute liver failure), or for a range of related diagnoses (any liver problems)?



A Small Three-Level Tree



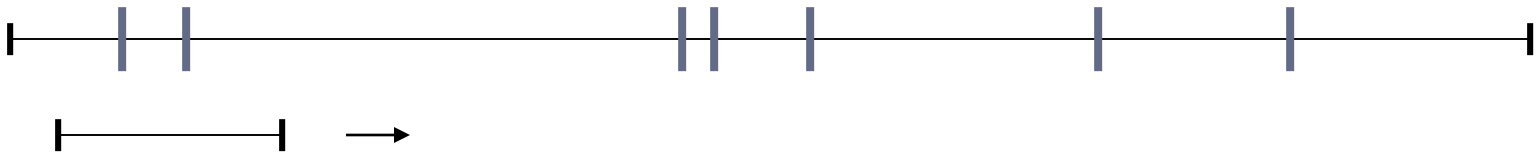
Lowest Level: ~6000 ICD-9-CM Codes

Some Diagnoses Removed

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

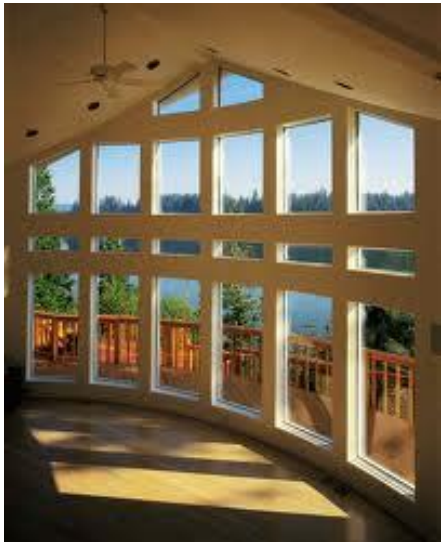
TreeScan Adjusts for Multiple Testing

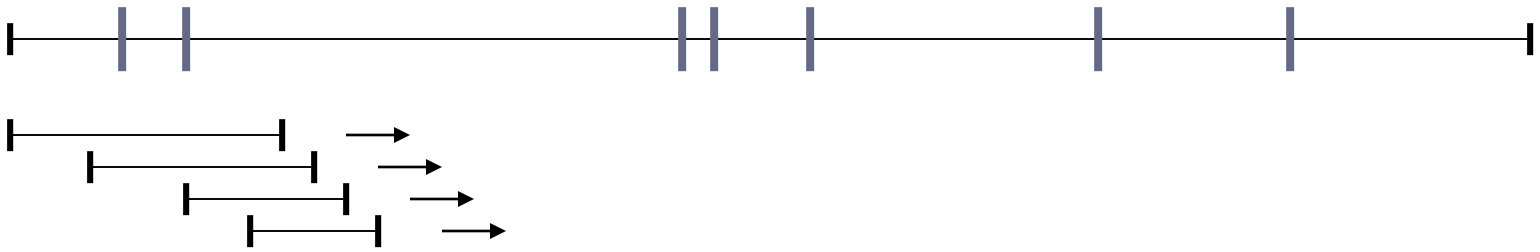
Temporal Scan Statistic
Fixed Window Size



Naus, J Am Stat Assoc, 1965

Temporal Scan Statistic
Variable Window Size





Scanning Risk Window

Follow-Up Period: 1-56 days

Risk Window Start Range: 1-28 days after vaccination

Risk Window End Range: 2-42 days after vaccination

Minimum Length: 2 days, Maximum Length: 28 days

A few of the 665 potential risk windows evaluated:
[1-5] , [2-28] , [3-4], [5-12] , [7-10] , [15-42] , [28-34]

Note: Day 0 is not included

Comparison Window

- Those days 1-56 after vaccination that are not in the risk window

Tree-Based Scan Statistic

- For each leaf, note the observed number of adverse events in each of the risk and control windows.
- For each higher level branch, add the observed number of events of its leaves.

Tree-Based Scan Statistic

1. Scan the tree by considering all possible cuts on any branch, and all possible risk wndows.
1. 2. For each cut and risk window, calculate the likelihood.
2. 3. Denote the cut/window with the maximum likelihood

3. as the most likely cut (cluster).
4. 4. Generate 9999 Monte Carlo replications under H_0.
5. 5. Compare the most likely cut from the real data set
6. with the most likely cuts from the random data sets.
7. 6. If the rank of the most likely cut from the real data

What is a TreeScan "Alert"?

- A statistically significant finding of greater than expected occurrence of an exposure-outcome pair
- Signal Detection or Screening Analysis **ONLY**
  - Produces hypotheses just as FAERS does
- Signal evaluation studies required for any further investigation

TreeScan™ in Vaccine Safety Surveillance

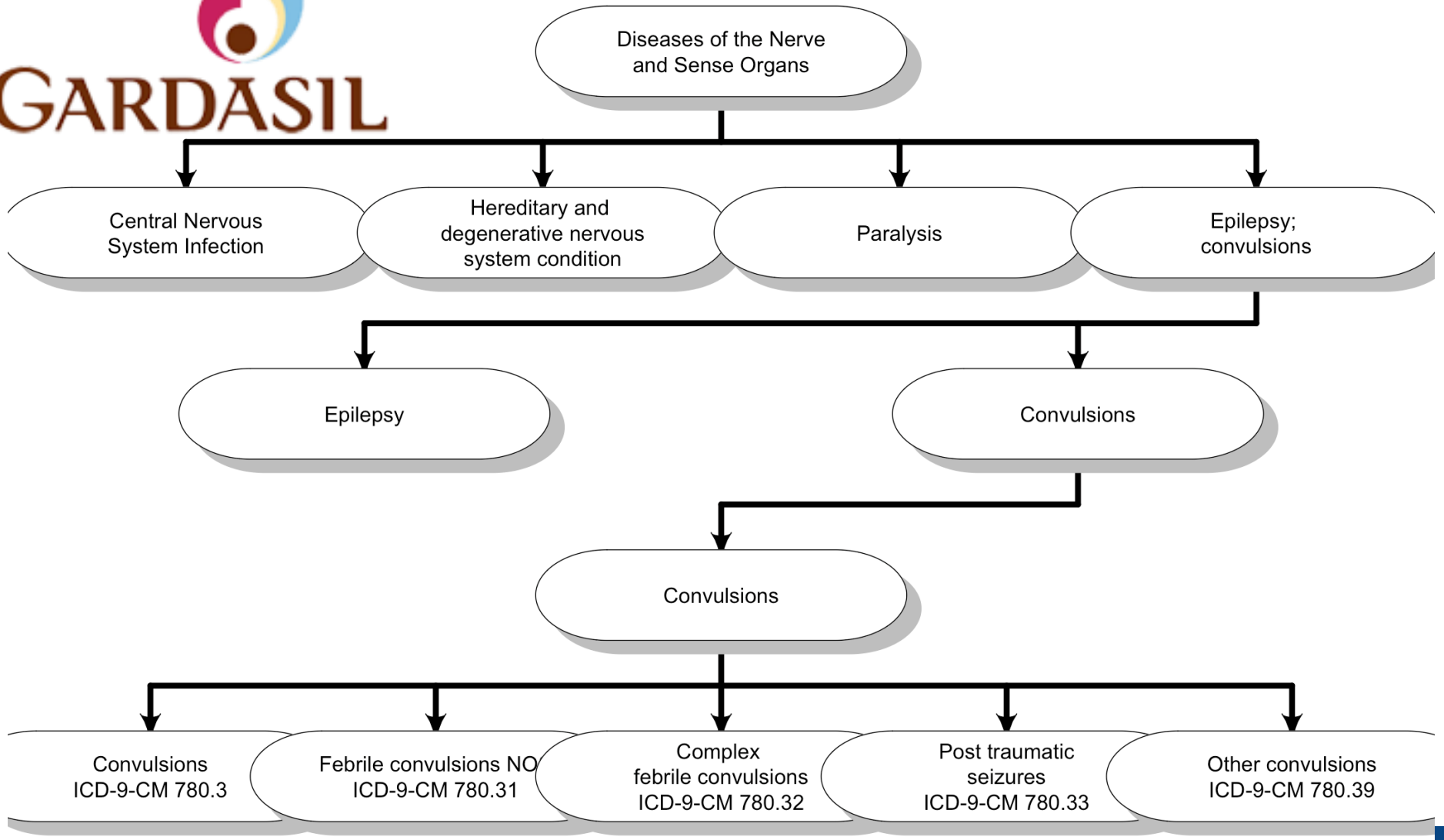
Azadeh Shoaibi

Center for Biologics Evaluation and Research, U.S. Food and Drug Administration

Data-Mining Designs with Trees

- Exposure-Oriented - 1 Exposure: N Outcomes
  - Uses Multi-Level Clinical Classification System (MLCCS) where N=6000+
- Outcome-Oriented - M Exposures: 1 Outcome
  - Uses Medi-Span Therapeutic Classification System (Drug Tree) where M=300,000+
- *Future - M Exposures: N Outcomes*

What is an Exposure-Oriented Scan?



HPV4 (Gardasil) Pilot

- Medically attended adverse events
- Conditional Tree-Temporal Scan Statistic
- Self-Controlled, adjusting for all fixed (non-time-varying) confounders
- First dose after 9th birthday or enrollment
- 1.9 million doses
- Five health plans

Results, HPV4, Dose 1

MLCCS (ICD9)	Disease Name	Win-dow	Obs	AR/ 100K	P=
12	Diseases of skin and subcutaneous tissue	2-4	214	3.8	0.0019
12.01	Skin and subcutaneous tissue infections	2-4	111	2.3	0.04
12.01.01	Cellulitis and abscess	2-4	93	2.0	0.20
… 682.3	Cellulitis and abscess of upper arm and forearm	2-3	31	1.3	0.00001
12.02	…				
… 695.9	Unspecified erythematous condition	2-3	13	0.5	0.25
16	Injury and Poisoning	1-3	48	2.2	0.00001
16.10.02.07	Other complications of surgical and medical procedures	1-3	36	1.8	0.00001
… 780.63	Post vaccination fever	1-2	4	0.2	0.31
… 999.5	Other serum reaction NEC	1-3	7	0.4	0.011
… 999.52	Other serum reaction due to vaccination	1-2	11	0.6	0.00001
… 999.9	Other and unspecified complications of medical care, NEC	1-6	12	0.6	0.0018

Cases in "Other Complications…" Signal

31 (86%) of the 36 cases received ≥ 1 other vaccine along with HPV4

Conditions	No.
With conditions identified in package insert as possible vaccine-associated adverse events*	29
No specified symptoms and no further medical visits within 60 days	3
With diverse symptoms, different in each case	4
Total	36

* e.g., headache, fever, nausea, and dizziness; local injection site reactions

Conclusions

The self-controlled tree-temporal scan statistics worked well for the HPV4 vaccine

- Known adverse reactions found
- No false alerts
- High power to detect rare adverse reactions
- Adjusts for multiple testing
- Only early onset adverse reactions evaluated
- We only looked at first dose

TreeScan™ in Drug Safety Surveillance

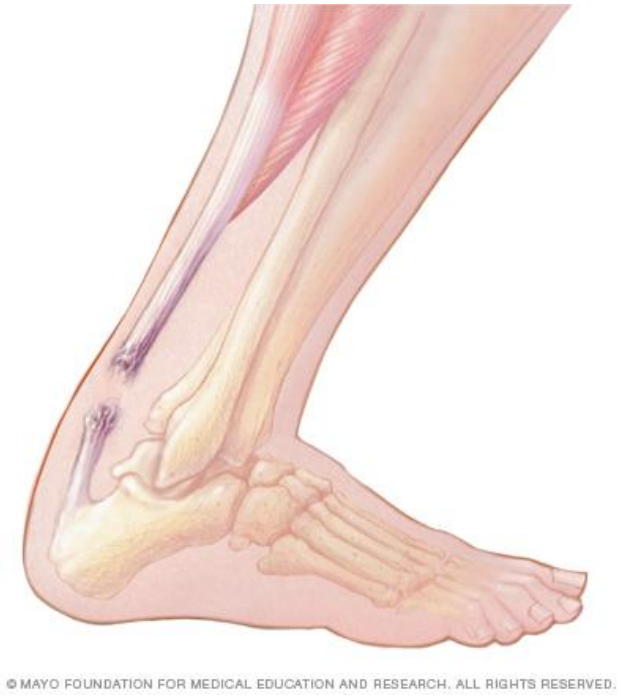
Rima Izem

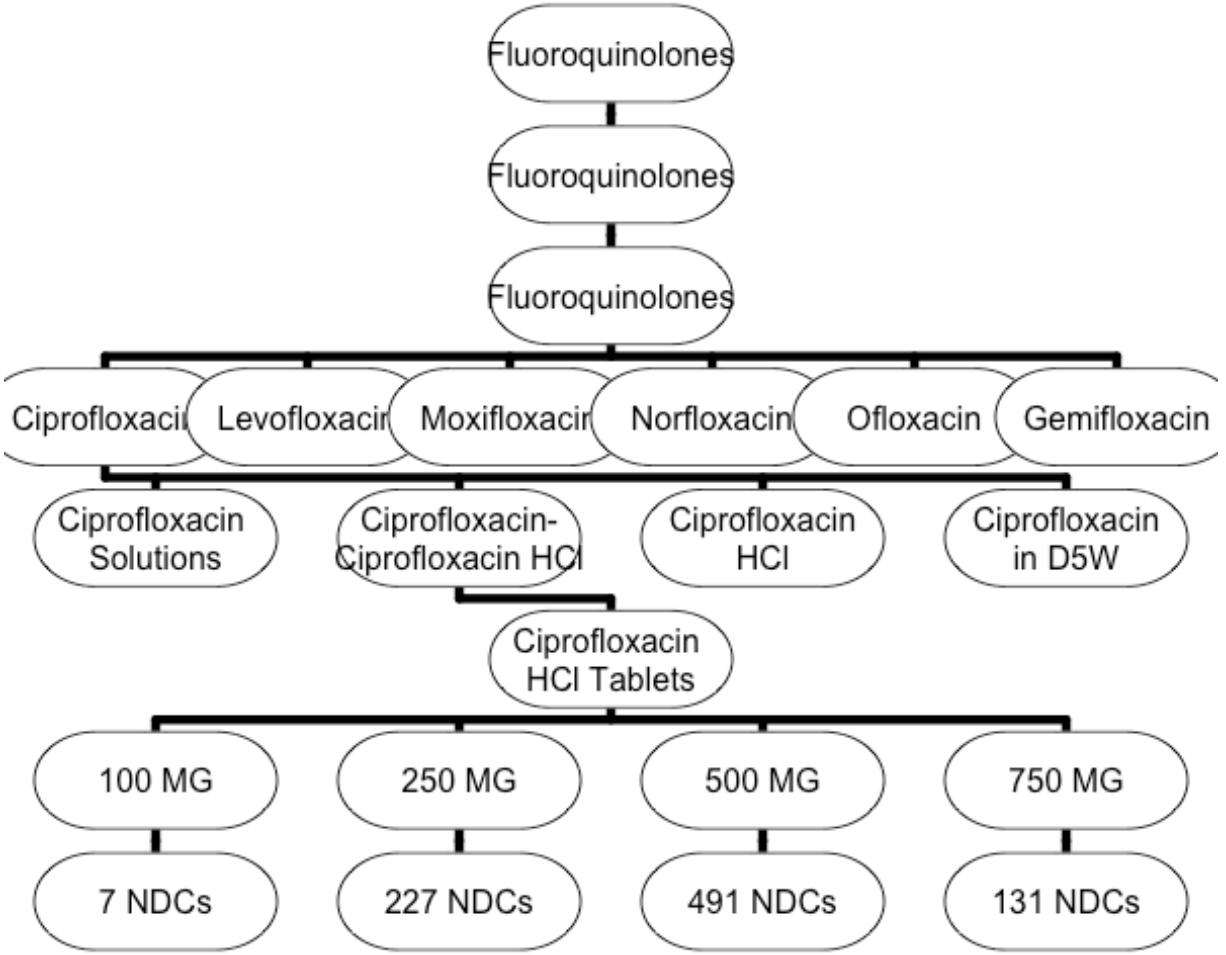
Center for Drug Evaluation and Research, U.S. Food and Drug Administration

Data-Mining Designs with Trees

- Exposure-Oriented - 1 Exposure: N Outcomes
  - Uses Multi-Level Clinical Classification System (MLCCS) where N=6000+
- Outcome-Oriented - M Exposures: 1 Outcome
  - Uses Medi-Span Therapeutic Classification System (Drug Tree) where M=300,000+
- *Future - M Exposures: N Outcomes*

What is an Outcome-Oriented Scan / DrugScan (1 Outcome: M Exposures)?





Total Nodes: 35,583 aggregate + 326,497 NDC codes

Angioedema Pilot

- Claims Data from 3 Data Partner Sites (2000-2014)
- Males and Females >=18 years with medical and drug coverage
- 45,580 incident cases of angioedema and 110,785 exposure-outcome pairs

Angioedema Results

- 28 unique alerts at 0.05 level, 20 meaningfully different
  - 9 were angioedema treatments
    - e.g., Glucocorticosteroids, Hydroxyzine, Diphenhydramine
  - Rest were known positives or likely positives
    - ACE inhibitors, Buproprion, Simvastatin, Antibiotics

- Sensitivity Analyses removed angioedema treatments from the tree
  - 13 unique, 9 meaningfully different
  - Some new antibiotics, ACEI Combos are statistically significant

Angioedema Summary

1. More misclassification of disease onset is present than expected
   - Patient Profiles show antecedent allergic reaction codes that did not rise to the level of angioedema
2. Detects known positives without too many false positives
3. Manageable number of total alerts

Question and Answer

Martin Kulldorff

Division of Pharmacoepidemiology and Pharmacoeconomics

Brigham and Women's Hospital and Harvard Medical School

Azadeh Shoaibi

Center for Biologics Evaluation and Research, U.S. Food and Drug Administration

Rima Izem

Center for Drug Evaluation and Research, U.S. Food and Drug Administration

Judith C. Maro

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

Interactive Demonstration of TreeScan Software and Signal Detection Exercise

Martin Kulldorff

Division of Pharmacoepidemiology and Pharmacoeconomics

Brigham and Women's Hospital and Harvard Medical School

Azadeh Shoaibi

Center for Biologics Evaluation and Research, U.S. Food and Drug Administration

Rima Izem

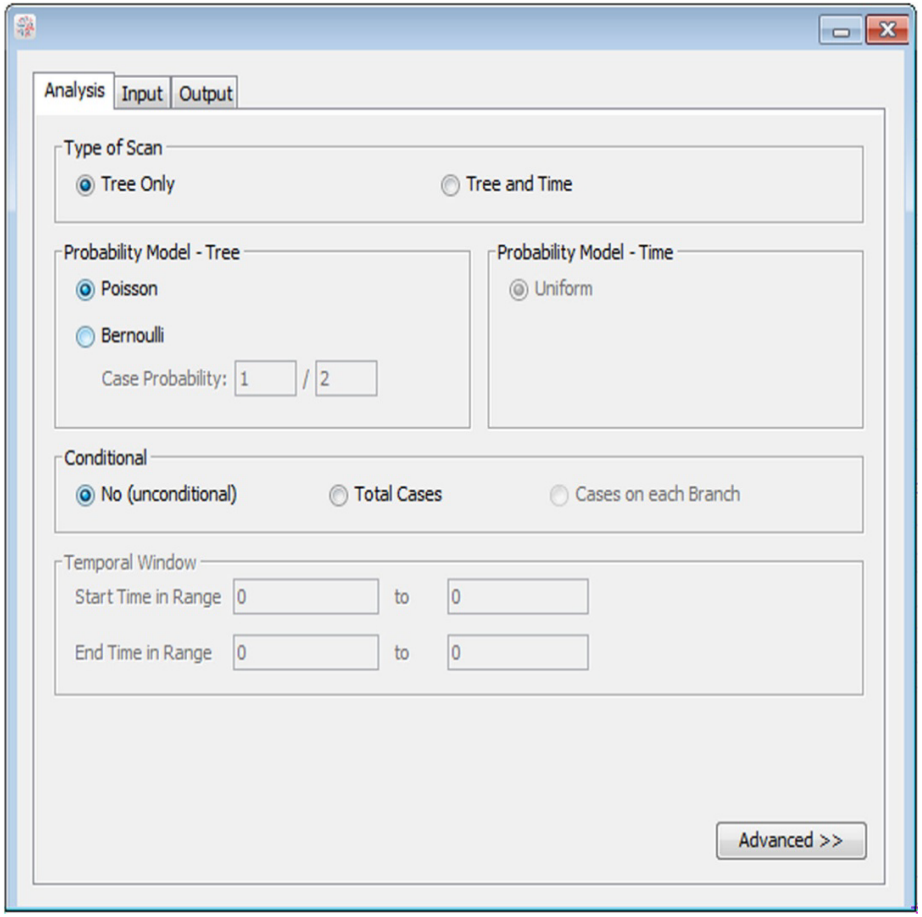
Center for Drug Evaluation and Research, U.S. Food and Drug Administration

Judith C. Maro

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

TreeScan Software

- Free
- www.treescan.org
- Windows, Mac, Linux
- User Guide



What do you need to do a TreeScan Analysis?

- Observational Design that will yield an input dataset designed to work with:
  - Data that can be analyzed using a Poisson likelihood
  - Data that can be analyzed using a Bernoulli likelihood

  - FOR TODAY: bernoulli.txt

- Hierarchical Tree Structure for Data
  - FOR TODAY: 2011dxtree.tre

Compatible Designs

- Poisson Data:
  - Set of observed outcomes compared to expected outcomes derived from expected outcome rates
  - One group monitoring
- Bernoulli Data:
  - Self-controlled Risk Interval Design (exposure-indexed with risk window and control window counts)
  - Case-crossover design (outcome-indexed with risk window and control window counts)
  - Fixed Ratio Matched Design (treatment and comparator counts)

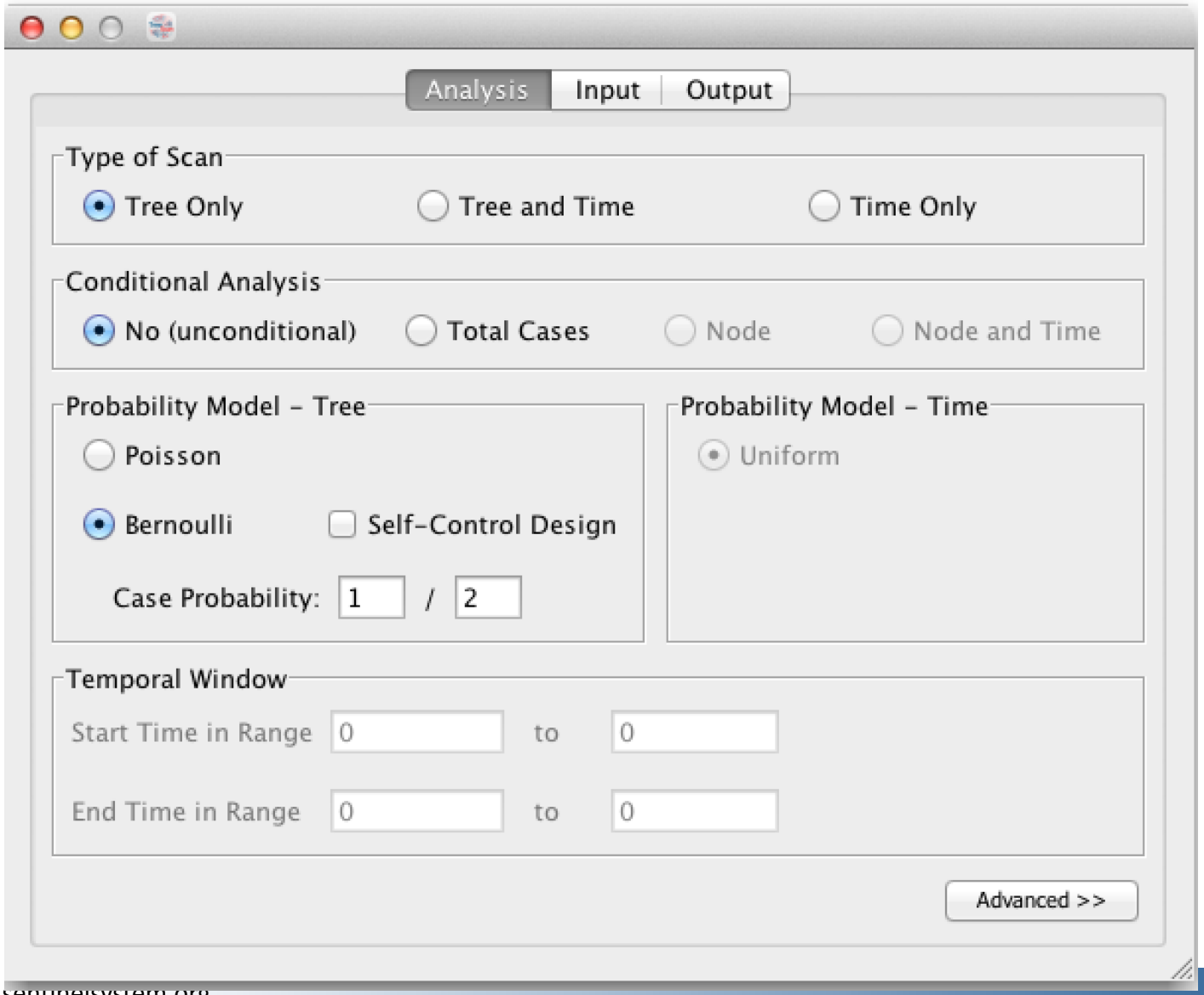
Bernoulli Simulated Problem

- Design = 1:1 Matched Design
- Exposure = Vaccine A
- Comparator = Vaccine B
- Followup Period = 28 days post-exposure
- Population = 100 million exposed persons (50M per study group)
- Tree = 2011 MLCCS Tree of ICD-9-CM codes (6162 outcomes)
- Simulated Signal = 780.2 (Syncope) at RR=2

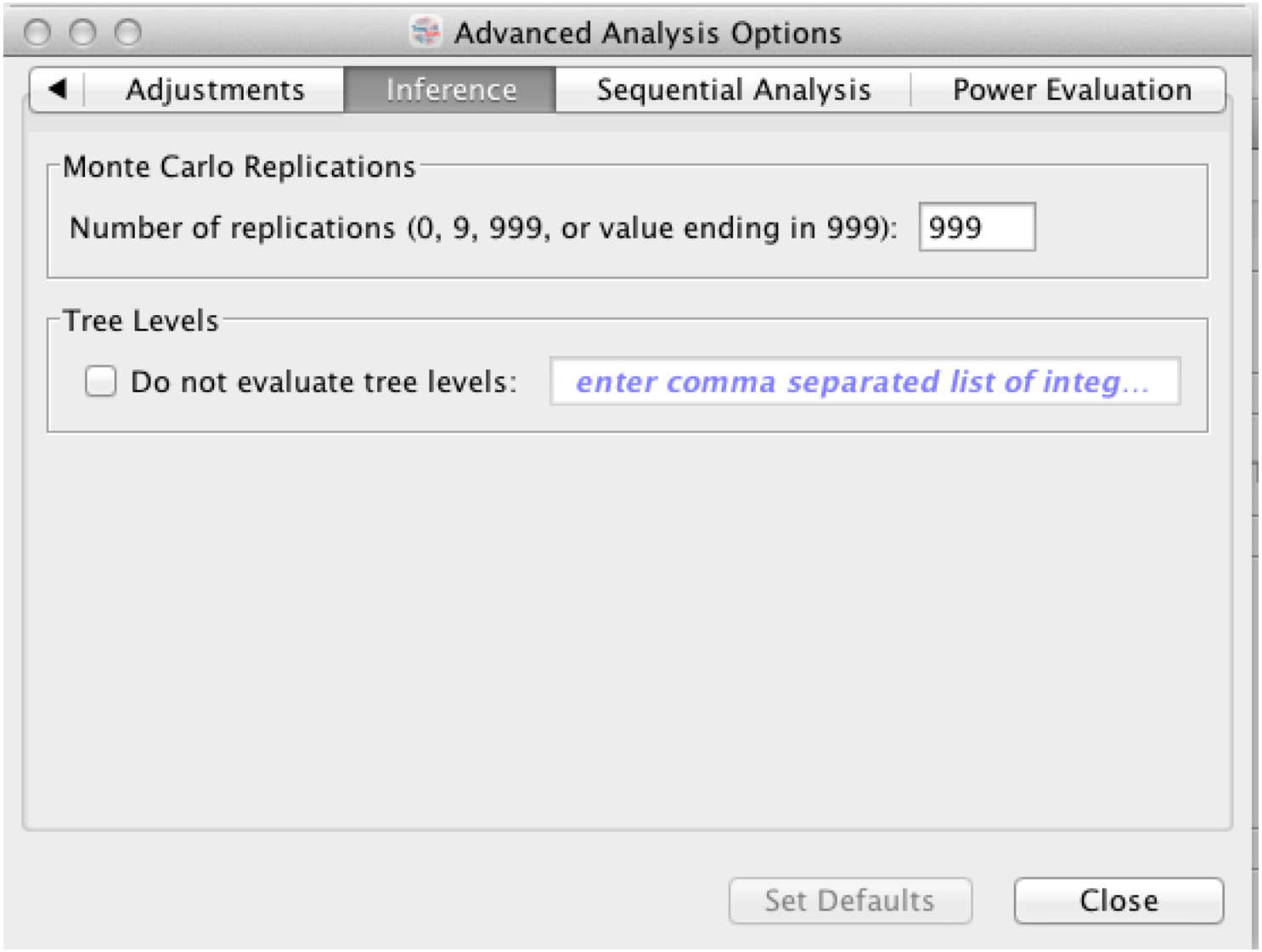
Orientation to the GUI

- Analysis Tab
  - Design Decisions
  - Advanced Features
- Input
  - Count File (Data File)
  - Tree File
- Output
  - Results File

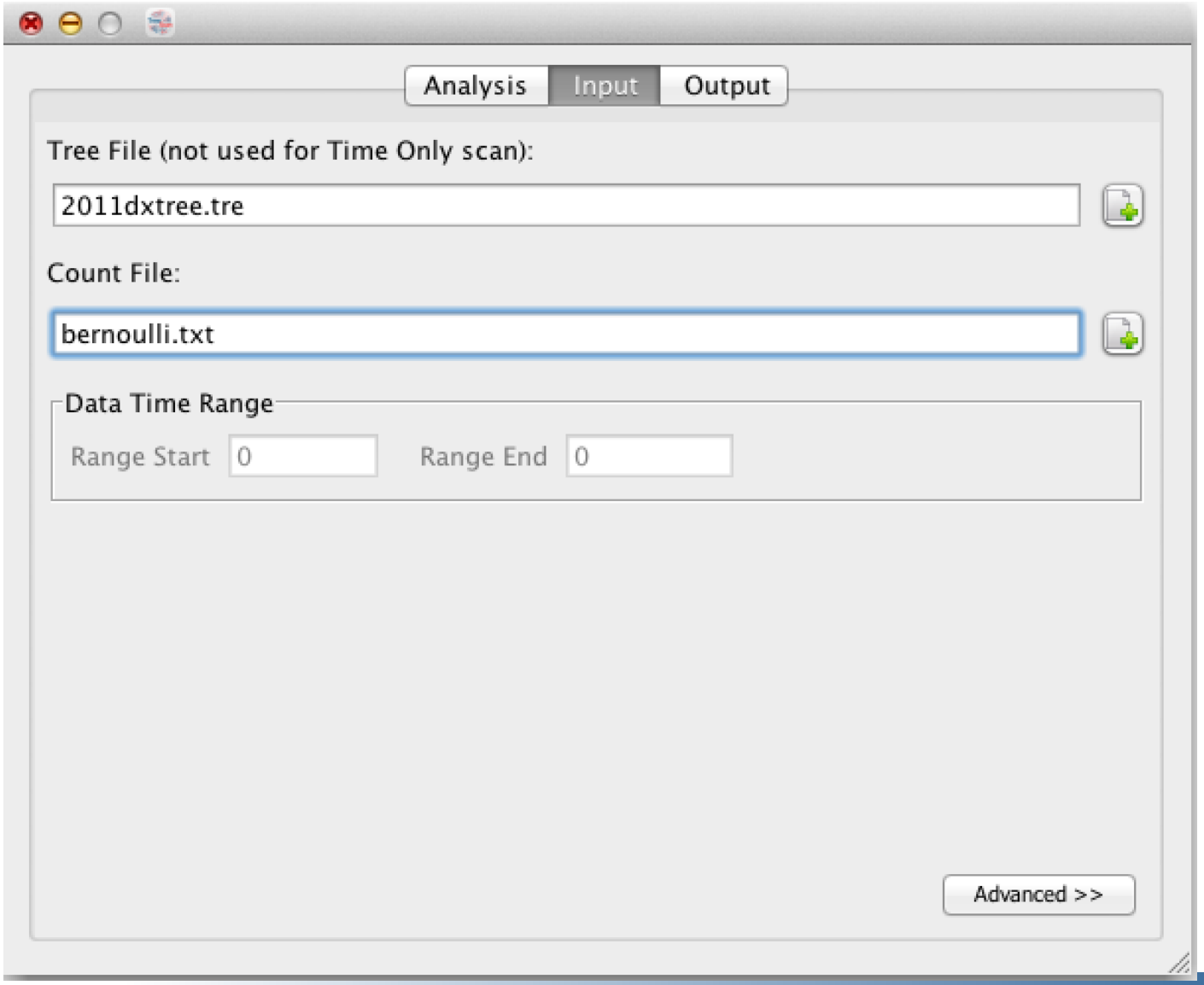
Analysis Tab



Advanced part of Analysis Tab



Input Tab


Analysis
Input
Output
Tree File (not used for Time Only scan):
2011dxtree.tre
Count File:
bernoulli.txt
Data Time Range
Range Start
0
Range End
0
Advanced >>

Tree File

```
571.42      ,01.03.02.00
571.49      ,01.03.02.00
571.41      ,01.03.02.00
571.40      ,01.03.02.00
135         ,01.04.00.00
136.1       ,01.04.00.00
087.9       ,01.04.00.00
242.81      ,03.01.01.00
242.00      ,03.01.01.00
242.01      ,03.01.01.00
242.21      ,03.01.01.00
242.20      ,03.01.01.00
242.90      ,03.01.01.00
242.80      ,03.01.01.00
242.41      ,03.01.01.00
242.11      ,03.01.01.00
242.91      ,03.01.01.00
242.40      ,03.01.01.00
```

Format

- Left Column: Child
- Right Column: Parent

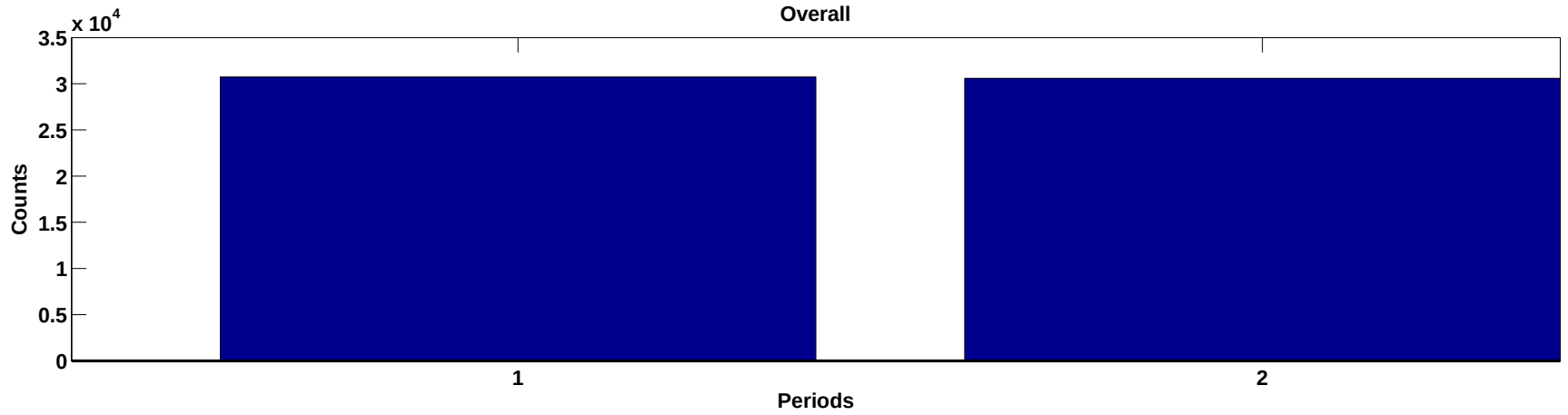
Bernoulli Training Dataset

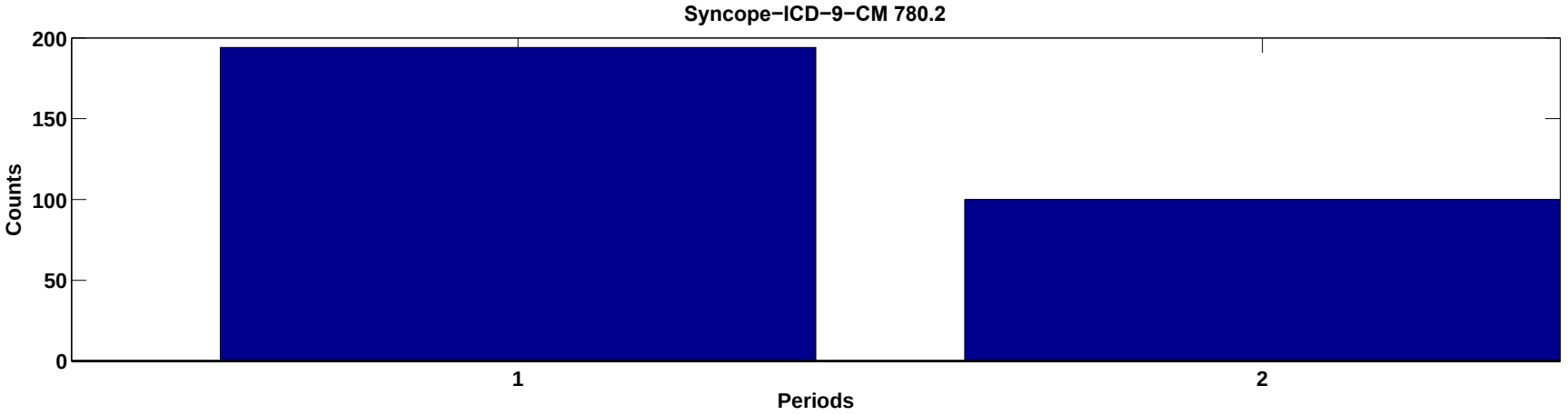
```
077.8,0,1
077.99,3,1
087.9,0,0
130.0,0,0
135,0,0
136.1,0,1
139.0,0,0
240.9,0,1
241.0,0,0
241.1,0,0
```

Format

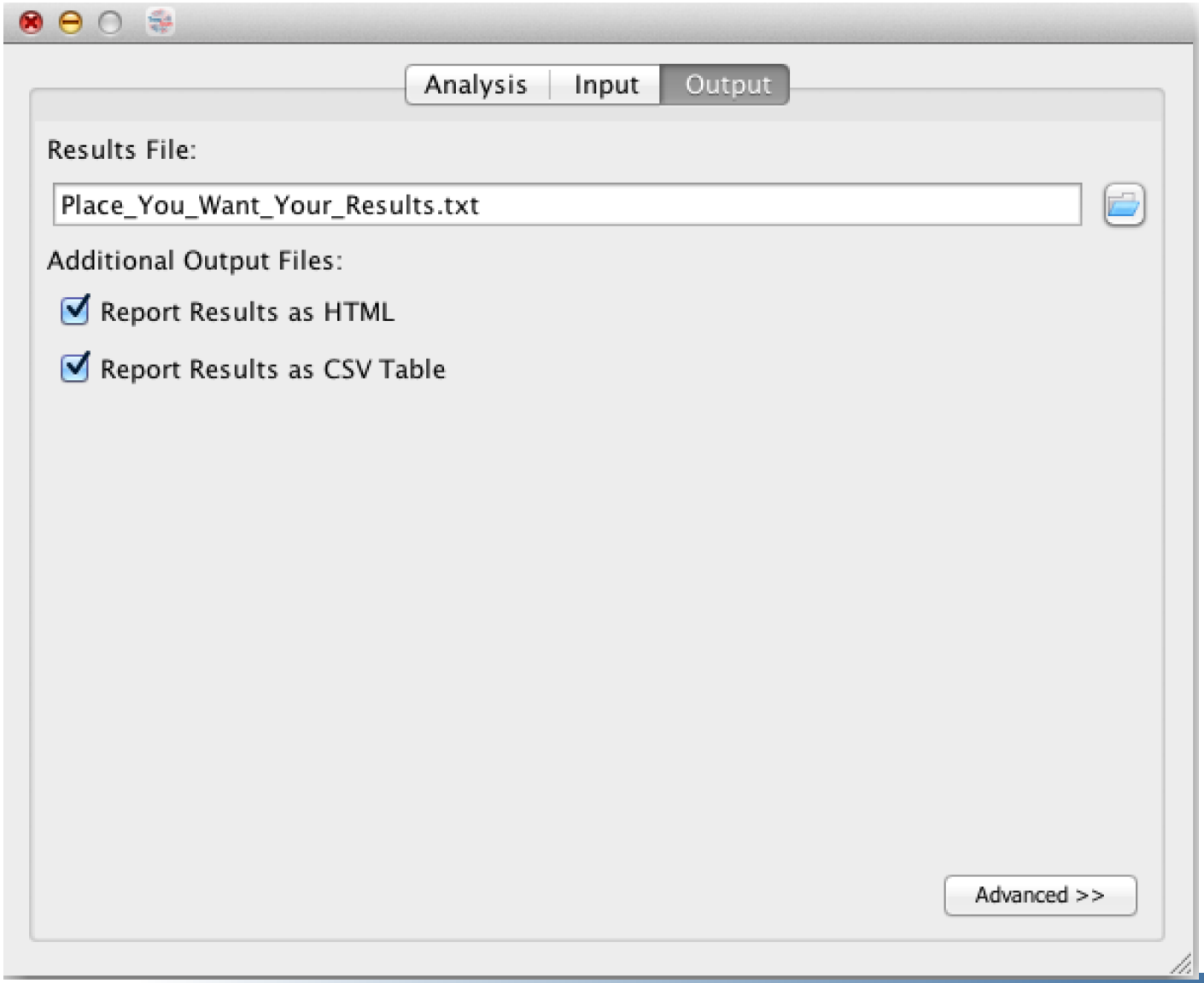
- First Column: Leaf Level Code
- Second Column: Number of Outcomes in Treatment Group
- Third Column: Number of Outcomes in Control Group

Visualization of the Bernoulli Dataset





Output Tab


Analysis
Input
Output
Results File:
Place_You_Want_Your_Results.txt
Additional Output Files:
Report Results as HTML
Report Results as CSV Table
Advanced >>

TreeScan Header

TreeScan v1.4 Alpha 1

Software for the Tree-Based Scan Statistic

Tree Only Scan with Unconditional Bernoulli Model

Total Cases:	30735
Total Observations:	61309
Number of Nodes:	6861
Number of Root Nodes:	16
Number of Nodes with Children:	699
Number of Leaf Nodes:	6162
Number of Levels in Tree:	5
Nodes per Levels:	16, 85, 239, 359, 6162

MOST LIKELY CUTS

No.	Node Identifier	Tree Level	Observations	Cases	Expected	Relative Risk	Excess Cases	L
1	780.2	5	294	194	147.00	1.32	47.00	1
2	17.01.01	3	472	272	236.00	1.15	36.00	5
3	17.01.01.00	4	472	272	236.00	1.15	36.00	5
4	17.01.05	3	20	17	10.00	1.70	7.00	5

TreeScan Method

$$LLR = \ln\left(\frac{\left(\frac{c_G}{c_G + n_G}\right)^{c_G}\left(\frac{n_G}{c_G + n_G}\right)^{n_G}}{p^{c_G}(1-p)^{n_G}}\right) I\left(\frac{c_G}{c_G + n_G} > p\right)$$

1) Solve the test statistic for the real dataset.
2) Create N simulated datasets under the null hypothesis. Calculate the T for each.
3) Rank all of those Ts and find the Monte Carlo based p-value. The winning T is your critical value for a signal to be statistically significant at the chosen p-value.

- OR
  When the null hypothesis is true, there is a (1-α)% probability that all p-values are greater than α, or in other words, that there is not a single exposure-outcome pair or grouping with p≤α.

Add Your Own Signal!

- Open up the Bernoulli text file in a Text Editing Program (Note: DON'T USE EXCEL!)
- Pick a node that suits your fancy and add in a bunch of cases.
  - Hint: Think about the total number of outcomes/observations across the node when deciding how many to add.
  - That is, if you add 5 additional outcomes to something that only occurs 10 times, you've just created a LARGE effect size.
  - Contrarily, if you add 5 additional outcomes to something that occurs 50 times, you've created a SMALLER effect size.
- Save your new file with a new name.

Back to TreeScan

- Change the input file location.
- Change the output file location.
- Run.

Question and Answer

Martin Kulldorff

Division of Pharmacoepidemiology and Pharmacoeconomics

Brigham and Women's Hospital and Harvard Medical School

Azadeh Shoaibi

Center for Biologics Evaluation and Research, U.S. Food and Drug Administration

Rima Izem

Center for Drug Evaluation and Research, U.S. Food and Drug Administration

Judith C. Maro

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

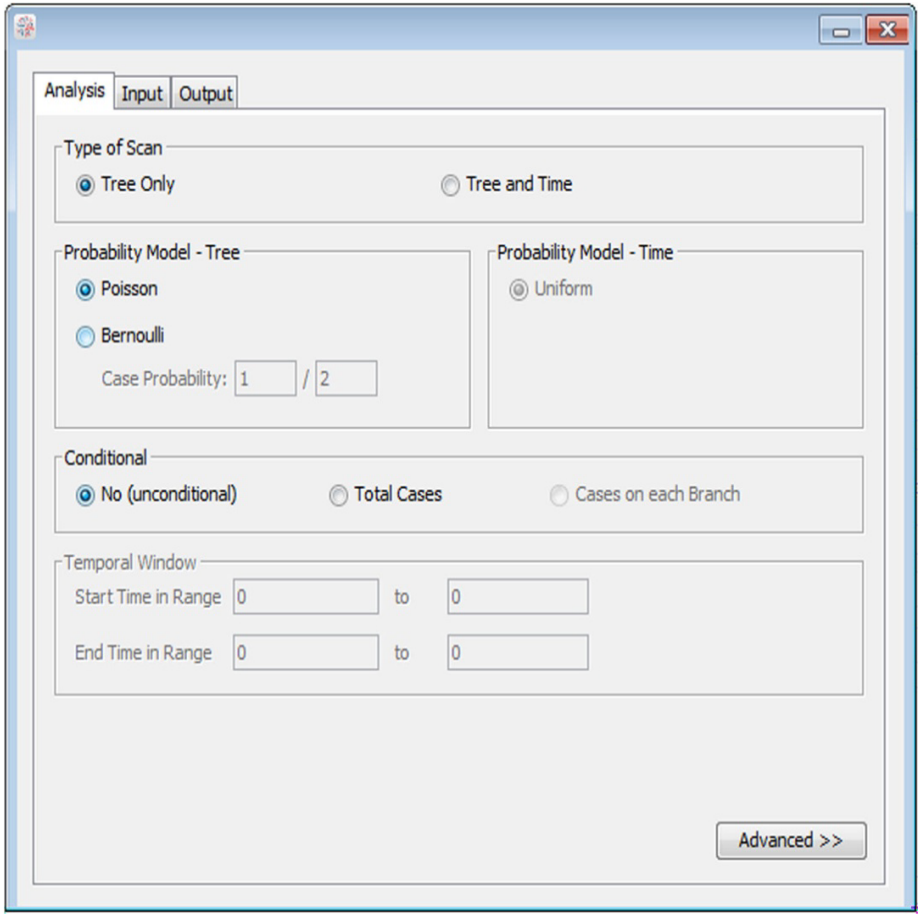
Acknowledgements

- SOC: Meghan Baker, Carolyn Balsbaugh, Jeff Brown, David Cole, Austin Cosgrove, Inna Dashevsky, Andrew Petrone, Megan Reidy, Katherine Yih
- FDA CBER: Steve Anderson, Kinnera Chada, Rositsa Dimova, Adamma Mba-Jonas, Joyce Obidi, Jawahar Tiwari
- FDA CDER: Gerald Dal Pan

- Many thanks are due to Data Partners who provided data used in the analysis.

TreeScan Software

- Free
- www.treescan.org
- Windows, Mac, Linux
- User Guide (47p)

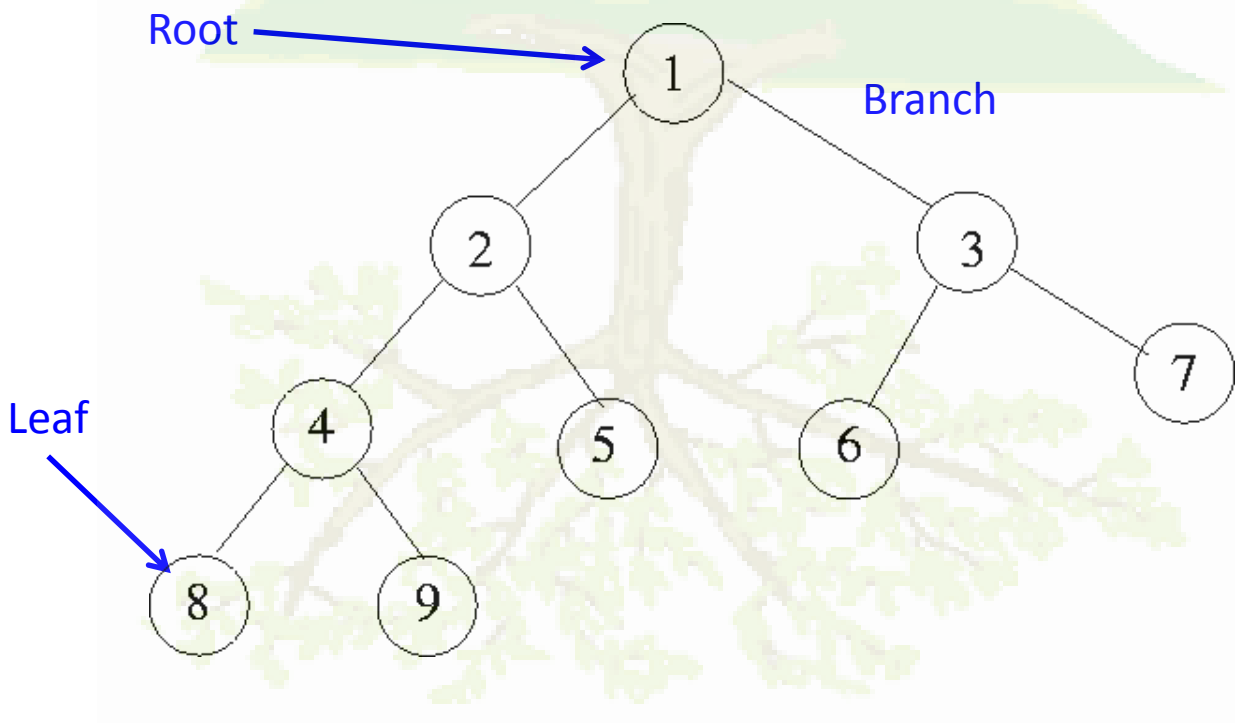


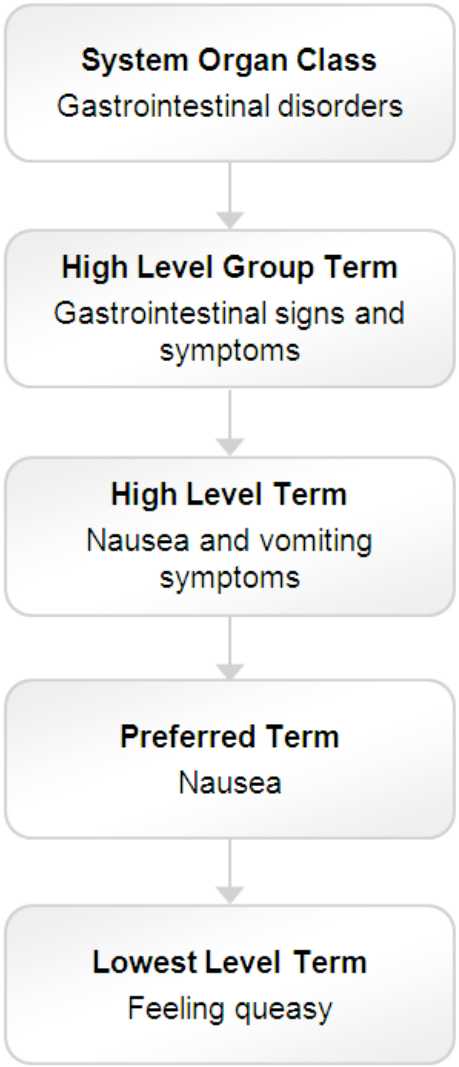
What is

- A signal detection / data-mining method
- Scans electronic health data that are grouped into hierarchical tree structures
- Automatically adjusts for multiple hypothesis testing

http://www.treescan.org

What is a Hierarchical Tree Structure?





Examples:

MedDRA reporting terms

Multi-level Clinical Classification System

Medi-Span Therapeutic Classification System

TreeScan Method

$$LLR = \ln\left(\frac{\left(\frac{c_G}{c_G + n_G}\right)^{c_G}\left(\frac{n_G}{c_G + n_G}\right)^{n_G}}{p^{c_G}(1-p)^{n_G}}\right) I\left(\frac{c_G}{c_G + n_G} > p\right)$$

1) Solve the test statistic for the real dataset.
2) Create N simulated datasets under the null hypothesis. Calculate the T for each.
3) Rank all of those Ts and find the Monte Carlo based p-value. The winning T is your critical value for a signal to be statistically significant at the chosen p-value.

- OR
  When the null hypothesis is true, there is a (1-α)% probability that all p-values are greater than α, or in other words, that there is not a single exposure-outcome pair or grouping with p≤α.