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

Heart Attack
Cardiac Dysrhythmia
Cardiac Myopathy
Kidney Failure
Kidney Infection

Cut
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 windows.

2. For each cut and risk window, calculate the likelihood.

3. Denote the cut/window 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 less than or equal to 100, then the null hypothesis is rejected.
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?

- Diseases of the Nerve and Sense Organs
  - Central Nervous System Infection
  - Hereditary and degenerative nervous system condition
  - Paralysis
  - Epilepsy; convulsions

  - Epilepsy
  - Convulsions

  - Convulsions

- Convulsions
  - Febrile convulsions NO ICD-9-CM 780.31
  - Complex febrile convulsions ICD-9-CM 780.32
  - Post traumatic seizures ICD-9-CM 780.33
  - Other convulsions ICD-9-CM 780.39
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 9\textsuperscript{th} birthday or enrollment
- 1.9 million doses
- Five health plans
<table>
<thead>
<tr>
<th>MLCCS (ICD9)</th>
<th>Disease Name</th>
<th>Window</th>
<th>Obs</th>
<th>AR/100K</th>
<th>P=</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Diseases of skin and subcutaneous tissue</td>
<td>2-4</td>
<td>214</td>
<td>3.8</td>
<td>0.0019</td>
</tr>
<tr>
<td>12.01</td>
<td>Skin and subcutaneous tissue infections</td>
<td>2-4</td>
<td>111</td>
<td>2.3</td>
<td>0.04</td>
</tr>
<tr>
<td>12.01.01</td>
<td>Cellulitis and abscess</td>
<td>2-4</td>
<td>93</td>
<td>2.0</td>
<td>0.20</td>
</tr>
<tr>
<td>... 682.3</td>
<td>Cellulitis and abscess of upper arm and forearm</td>
<td>2-3</td>
<td>31</td>
<td>1.3</td>
<td>0.00001</td>
</tr>
<tr>
<td>12.02</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>... 695.9</td>
<td>Unspecified erythematous condition</td>
<td>2-3</td>
<td>13</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>16</td>
<td>Injury and Poisoning</td>
<td>1-3</td>
<td>48</td>
<td>2.2</td>
<td>0.00001</td>
</tr>
<tr>
<td>16.10.02.07</td>
<td>Other complications of surgical and medical procedures</td>
<td>1-3</td>
<td>36</td>
<td>1.8</td>
<td>0.00001</td>
</tr>
<tr>
<td>... 780.63</td>
<td>Post vaccination fever</td>
<td>1-2</td>
<td>4</td>
<td>0.2</td>
<td>0.31</td>
</tr>
<tr>
<td>... 999.5</td>
<td>Other serum reaction NEC</td>
<td>1-3</td>
<td>7</td>
<td>0.4</td>
<td>0.011</td>
</tr>
<tr>
<td>... 999.52</td>
<td>Other serum reaction due to vaccination</td>
<td>1-2</td>
<td>11</td>
<td>0.6</td>
<td>0.00001</td>
</tr>
<tr>
<td>... 999.9</td>
<td>Other and unspecified complications of medical care, NEC</td>
<td>1-6</td>
<td>12</td>
<td>0.6</td>
<td>0.0018</td>
</tr>
</tbody>
</table>
### Cases in “Other Complications...” Signal

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

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

* 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, Bupropion, 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

Type of Scan
- Tree Only
- Tree and Time
- Time Only

Conditional Analysis
- No (unconditional)
- Total Cases
- Node
- Node and Time

Probability Model – Tree
- Poisson
- Bernoulli
- Self-Control Design

Case Probability: 1 / 2

Probability Model – Time
- Uniform

Temporal Window
- Start Time in Range: 0 to 0
- End Time in Range: 0 to 0

Advanced >>
Advanced part of Analysis Tab

Monte Carlo Replications

Number of replications (0, 9, 999, or value ending in 999): 999

Tree Levels

Do not evaluate tree levels: enter comma separated list of integers...
Input Tab

Tree File (not used for Time Only scan):
2011dxtree.tre

Count File:
bernoulli.txt

Data Time Range
Range Start 0  Range End 0
## Tree File

<table>
<thead>
<tr>
<th>571.42</th>
<th>01.03.02.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>571.49</td>
<td>01.03.02.00</td>
</tr>
<tr>
<td>571.41</td>
<td>01.03.02.00</td>
</tr>
<tr>
<td>571.40</td>
<td>01.03.02.00</td>
</tr>
<tr>
<td>135</td>
<td>01.04.00.00</td>
</tr>
<tr>
<td>136.1</td>
<td>01.04.00.00</td>
</tr>
<tr>
<td>087.9</td>
<td>01.04.00.00</td>
</tr>
<tr>
<td>242.81</td>
<td>03.01.01.00</td>
</tr>
<tr>
<td>242.00</td>
<td>03.01.01.00</td>
</tr>
<tr>
<td>242.01</td>
<td>03.01.01.00</td>
</tr>
<tr>
<td>242.21</td>
<td>03.01.01.00</td>
</tr>
<tr>
<td>242.20</td>
<td>03.01.01.00</td>
</tr>
<tr>
<td>242.90</td>
<td>03.01.01.00</td>
</tr>
<tr>
<td>242.80</td>
<td>03.01.01.00</td>
</tr>
<tr>
<td>242.41</td>
<td>03.01.01.00</td>
</tr>
<tr>
<td>242.11</td>
<td>03.01.01.00</td>
</tr>
<tr>
<td>242.91</td>
<td>03.01.01.00</td>
</tr>
<tr>
<td>242.40</td>
<td>03.01.01.00</td>
</tr>
</tbody>
</table>

### Format
- **Left Column: Child**
- **Right Column: Parent**
Bernoulli Training Dataset

Format

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

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
242, 0, 0, 0
Visualization of the Bernoulli Dataset

Overall

Counts
Periods
1
2

Syncope–ICD–9–CM 780.2

Counts
Periods
1
2

Output Tab

Results File:

Place_You_Want_Your_Results.txt

Additional Output Files:

- Report Results as HTML
- Report Results as CSV Table
# TreeScan Header

**TreeScan v1.4 Alpha 1**

Software for the Tree-Based Scan Statistic

<table>
<thead>
<tr>
<th>Tree Only Scan with Unconditional Bernoulli Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Cases:</strong></td>
</tr>
<tr>
<td><strong>Total Observations:</strong></td>
</tr>
<tr>
<td><strong>Number of Nodes:</strong></td>
</tr>
<tr>
<td><strong>Number of Root Nodes:</strong></td>
</tr>
<tr>
<td><strong>Number of Nodes with Children:</strong></td>
</tr>
<tr>
<td><strong>Number of Leaf Nodes:</strong></td>
</tr>
<tr>
<td><strong>Number of Levels in Tree:</strong></td>
</tr>
<tr>
<td><strong>Nodes per Levels:</strong></td>
</tr>
</tbody>
</table>

## MOST LIKELY CUTS

<table>
<thead>
<tr>
<th>No.</th>
<th>Node Identifier</th>
<th>Tree Level</th>
<th>Observations</th>
<th>Cases</th>
<th>Expected</th>
<th>Relative Risk</th>
<th>Excess Cases</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>780.2</td>
<td>5</td>
<td>294</td>
<td>194</td>
<td>147.00</td>
<td>1.32</td>
<td>47.00</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>17.01.01</td>
<td>3</td>
<td>472</td>
<td>272</td>
<td>236.00</td>
<td>1.15</td>
<td>36.00</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>17.01.01.00</td>
<td>4</td>
<td>472</td>
<td>272</td>
<td>236.00</td>
<td>1.15</td>
<td>36.00</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>17.01.05</td>
<td>3</td>
<td>20</td>
<td>17</td>
<td>10.00</td>
<td>1.70</td>
<td>7.00</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TreeScan Method

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

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-\(\alpha\))% probability that all p-values are greater than \(\alpha\), or in other words, that there is not a single exposure-outcome pair or grouping with \(p \leq \alpha\).
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

System Organ Class
Gastrointestinal disorders

High Level Group Term
Gastrointestinal signs and symptoms

High Level Term
Nausea and vomiting symptoms

Preferred Term
Nausea

Lowest Level Term
Feeling queasy
TreeScan Method

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

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 \( (1-\alpha) \)% probability that all p-values are greater than \( \alpha \), or in other words, that there is not a single exposure-outcome pair or grouping with \( p \leq \alpha \).