Data Mining for Adverse Drug Events With A Propensity Score Matched Tree-Based Scan Statistic

Shirley V. Wang¹, Judith C. Maro², Elande Baro³, Rima Izem³, Inna Dashevsky², James R. Rogers¹, Michael Nguyen⁴, Joshua J. Gagne¹, Elisabetta Patorno¹, Krista F. Huybrechts¹, Jacqueline M Major⁴, Esther Zhou⁴, Megan Reidy², Austin Cosgrove², Sebastian Schneeweiss¹, Martin Kulldorff¹

¹. Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Harvard Medical School and Brigham and Women’s Hospital; ². Department of Population Medicine, Harvard Medical School, Harvard Pilgrim Health Care Institute; ³. Office of Biostatistics, Center for Drug Evaluation and Research, U.S. FDA; ⁴. Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, U.S. FDA.
Disclosures

▪ This work supported by the U.S. Food and Drug Administration (FDA) through the Department of Health and Human Services (HHS) contract number: HHSF22301010T-0004

▪ At the time that this work was conducted, Dr. Wang was principal investigator on other grants from: U.S. HHS Agency for Healthcare Research and Quality (AHRQ), FDA Sentinel Initiative, and an investigator initiated grant from Novartis for unrelated research.

▪ Dr. Wang is a consultant to Aetion, Inc., a software company.
What is TreeScan™?

- A statistical data mining tool for signal detection
  - Utilizes tree-based scan statistics
  - Adjusts for multiple testing in evaluation of thousands of potential adverse events

What is TreeScan™?

- A statistical data mining tool for signal detection
  - Utilizes tree-based scan statistics
  - Adjusts for multiple testing in evaluation of thousands of potential adverse events

Kulldorff, M. Drug safety data mining with a tree-based scan statistic. PDS, 2013
What is TreeScan™?

- A statistical data mining tool for signal detection
  - Utilizes tree-based **scan statistics**
  - Adjusts for multiple testing in evaluation of thousands of potential adverse events

What is TreeScan™?

- A statistical data mining tool for signal detection
  - Utilizes tree-based scan statistics
  - Adjusts for multiple testing in evaluation of thousands of potential adverse events

The Tree

- Multi-level Clinical Classifications (MLCCS)
  - Includes all ICD-9 CM codes
  - Hierarchical system
  - 4 levels of clinical concepts
    - Level 1 - body systems, 18 categories
    - Level 2
    - Level 3
    - Level 4
    - Leaf

The Tree

7 Diseases of the circulatory system

7.1 Hypertension

7.1.1 Essential Hypertension

7.1.2 Hypertension with complications and secondary hypertension

7.1.2.1 Hypertensive heart and/or renal disease

7.1.2.2 Other hypertensive complications

ICD 9 codes: 40200 40201 40210 40211 40290 40291 4030 40300 40301 4031 40310 40311 4039 40390 40391 4040 40400 40401 40402 40403 4041 40410 40411 40412 40413 40414 4049 40490 40491 40492 40493

ICD 9 codes: 4010 40501 40509 40511 40519 40591 40599 4372

MLCCS

• Parent nodes are connected to children and descendants by lines
• Non-descendant nodes are on different branches

How has TreeScan been used before?

- Scanning did not perform well in **drug examples** with self-controlled design when patients were “unstable” around time of exposure initiation.

- **Propensity score (PS) matched new initiator cohort** is a powerful design that uses an active comparator selected to balance on time-varying factors around treatment initiation.
Objective

- Conduct simulation with known truth to evaluate unconditional Bernoulli TreeScan statistic with PS matched cohort design
The Scan

- \( T = \) unconditional Bernoulli scan statistic

\[
T = \max_G LLR(G)
\]

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

\( G = \) node of interest
\( c_G = \) cases in the treatment group for a given node
\( n_G = \) cases in the reference group for a given node
\( p = \) probability of being in the treatment group (for 1:1 matched this is 0.5)

Kulldorff, M. Drug safety data mining with a tree-based scan statistic. PDS, 2013
Kulldorff, M. TreeScan User Guide, version 1.2
The Scan

- $T =$ unconditional Bernoulli scan statistic

Distribution of the test statistic $T$ is unknown

∴ Use Monte Carlo based p-value = Rank/(9999+1)

1. Generate $T$ for 9999 random datasets (under the null)
2. Rank $T$
3. If observed $T \geq 1\%$ of $T$ from 9999 datasets under the null
   ⇒ alert at alpha = 0.01

Kulldorff, M. Drug safety data mining with a tree-based scan statistic. PDS, 2013
Kulldorff, M. TreeScan User Guide, version 1.2
Simulation

- “Plasmode” style simulation
  - Based on a real cohort extracted from a claims database instead of fully synthetic simulated data
  - Retains observed complexity and correlation for:
    - Baseline covariates
    - Clusters of outcomes across tree

- Permutes relationships between:
  1. Covariates and outcome
  2. Exposure and outcome

Methods and Process

1. Identify cohort* (exposure and baseline covariates)
   - New initiators Dipeptidyl peptidase 4 (DPP4) inhibitors, sulfonylureas
   - 183 day washout, allow 30 day gaps in enrollment
   - No outcome specified
   - PS based on 26 predefined covariates (caliper = 0.025)
     - Age
     - Sex
     - Combined comorbidity score
     - Chronic kidney disease
     - Hypoglycemia
     - Diabetic nephropathy
     - Diabetic neuropathy
     - Diabetic retinopathy
     - Diabetic Peripheral Circulation Disorder
     - Erectile dysfunction
     - Skin Infections
     - Diabetic complications unspecified
     - Alpha glucosidase
     - Glitazones
     - Glucagon-like peptide-1 receptors agonists
     - Insulin
     - Meglitinides
     - Metformin
     - # outpatient visits
     - # erectile dysfunction visits
     - # inpatient (IP) visits
     - # institutional stays
     - # other visits
     - # classes medication
     - # generics
     - # Rx dispensed
   - Return individual level data on unmatched cohort

* Using routine query tool Cohort Identification and Descriptive Analysis [CIDA] + PS matching on Common Data Model [CDM] for matted data
https://www.sentinelinitiative.org/sentinel/surveillance-tools/routine-querying-tools/routine-querying-system
Methods and Process

2. Pull incident outcomes within fixed window for each patient (TreeExtraction)
   – Return incident outcomes for simulation permutation
Methods and Process

3. Permute data for simulation
   – 11 scenarios
   – Maintain covariate structure for exposure and baseline covariates and clustered outcome “bundles”

<table>
<thead>
<tr>
<th>Scenario</th>
<th>True Relative Risk</th>
<th># Nodes w/ True Effect</th>
<th>Confounding?</th>
<th>Direction of Confounding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>0</td>
<td>No</td>
<td>n/a</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>3</td>
<td>Yes</td>
<td>Positive (away from the null)</td>
</tr>
<tr>
<td>3</td>
<td>2.0</td>
<td>3</td>
<td>No</td>
<td>n/a</td>
</tr>
<tr>
<td>4</td>
<td>4.0</td>
<td>3</td>
<td>Yes</td>
<td>Positive (away from the null)</td>
</tr>
<tr>
<td>5</td>
<td>1.5</td>
<td>3</td>
<td>Yes</td>
<td>Negative (toward the null)</td>
</tr>
</tbody>
</table>
Methods and Process

4. Repeat data generation 1,000 times for each simulation scenario

5. Varied degree of PS misspecification by identifying 1:1 matches based on:
   – Random sample without replacement
   – PS with random 40%, 50%, 60%, 80% of true confounders
   – PS with all confounders

6. Run TreeScan for 1,000 cohorts per simulation scenario
   – Arbitrary threshold for alerting at $p < 0.01$
### Selected nodes

With simulated elevation in risk related to exposure and/or confounding

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Diseases of the digestive system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 2</td>
<td>Gastrointestinal hemorrhage</td>
</tr>
<tr>
<td><strong>Level 3</strong></td>
<td><strong>Hemorrhage from gastrointestinal ulcer</strong></td>
</tr>
<tr>
<td>Level 4</td>
<td>--</td>
</tr>
<tr>
<td>Leaf</td>
<td>Numerous diagnosis codes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Diseases of the circulatory system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 2</td>
<td>Cerebrovascular disease</td>
</tr>
<tr>
<td><strong>Level 3</strong></td>
<td><strong>Acute cerebrovascular disease</strong></td>
</tr>
</tbody>
</table>
| Level 4 | Acute but ill-defined cerebrovascular accident  
  Intracranial hemorrhage 
  Occlusion of cerebral arteries |
| Leaf    | Numerous diagnosis codes         |

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Diseases of the genitourinary system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 2</td>
<td>Diseases of the urinary system</td>
</tr>
<tr>
<td><strong>Level 3</strong></td>
<td><strong>Acute and unspecified renal failure</strong></td>
</tr>
</tbody>
</table>
| Level 4 | Acute renal failure  
  Unspecified renal failure |
| Leaf    | Numerous diagnosis codes           |
# Results: Take-home points

<table>
<thead>
<tr>
<th>True Effect</th>
<th>Confounding</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>None</td>
<td>False positive (type 1 error) as expected</td>
</tr>
<tr>
<td>Null</td>
<td>+</td>
<td>Unadjusted → inflated type 1 100% adjusted → type 1 as expected</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>Better adjustment → recover power</td>
</tr>
<tr>
<td>+</td>
<td>None/+/-</td>
<td>PS with random 80% of true confounders performed similarly to PS with 100% of true confounders in most evaluated scenarios</td>
</tr>
<tr>
<td>+</td>
<td>None/+/-</td>
<td>Co-occurring outcomes also alerted</td>
</tr>
</tbody>
</table>

- Neither false alerts nor confounding
- Hierarchical MLCCS classification system is organ based
- Data reflect billing for multi-system disease that touch multiple branches
- Simulation retained observed bundles of co-occurring outcomes
Results:
All true effects null (Relative Risk (RR) = 1.0)
Confounding away from null (+)
Results: Take-home points

When we simulated a true effect of exposure in 3 selected nodes, co-occurring outcomes in non-descendant nodes alerted - clinically related condition?

- Example: true RR = 4.0, no confounding
- 52% of simulated datasets had alerts with p <0.01 in non-descendant nodes
  - Which nodes? (rolled up to level 3)

Nodes with simulated true effect:
- Hemorrhage, GI ulcer
- Acute cerebrovascular disease
- Acute and unspecified renal failure

<table>
<thead>
<tr>
<th>Node</th>
<th>Percent</th>
<th>MLCCS Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>08.06.01</td>
<td>32.6</td>
<td>Respiratory failure</td>
</tr>
<tr>
<td>03.08.01</td>
<td>18.6</td>
<td>Hyposmolality</td>
</tr>
<tr>
<td>06.03.01</td>
<td>17.7</td>
<td>Hemiplegia</td>
</tr>
<tr>
<td>07.01.02</td>
<td>17.5</td>
<td>Hypertension with complications</td>
</tr>
<tr>
<td>03.08.05</td>
<td>13.7</td>
<td>Other fluid and electrolyte disorders</td>
</tr>
<tr>
<td>17.01.05</td>
<td>11.0</td>
<td>Shock</td>
</tr>
<tr>
<td>10.01.03</td>
<td>10.4</td>
<td>Chronic kidney disease</td>
</tr>
<tr>
<td>Other</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Strengths

1. First **evaluation** of the unconditional Bernoulli **TreeScan** statistic to screen for unknown adverse events **when used with a PS matched cohort** design

2. Simulations retained the complexity of observed baseline covariates and “bundles” of observed outcomes within individuals
Limitations

1. Plasmode simulation based on one observational cohort
   – Baseline covariate correlation will differ in other cohorts

2. Evaluation only used MLCCS hierarchical tree
   – Primarily organ based
   – Other trees may have different properties

3. Did not address how to select covariates for PS
   – Difficult to identify risk factors for all outcomes
   – General frailty based or empirical PS may provide broad coverage
Discussion

- TreeScan with PS matching shows promise as a method for **hypothesis free screening** and **prioritization** of potential areas to pursue deeper investigation.

- Should be followed with further evaluation:
  - Patient Episode Profile Retrieval (**PEPR**) to better understand the clinical context around potential signals.
  - Targeted study to generate valid and precise estimates of effect for potential signals (confounding control tailored to specific outcome).
Questions

This work was published in Epidemiology, Aug 2018

swang1@bwh.harvard.edu