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

Shirley V. Wang[1], Judith C. Maro[2], Elande Baro[3], Rima Izem[3], Inna Dashevsky[2], James R. Rogers[1], Michael Nguyen[4], Joshua J. Gagne[1], Elisabetta Patorno[1], Krista F. Huybrechts[1], Jacqueline M Major[4], Esther Zhou[4], Megan Reidy[2], Austin Cosgrove[2], Sebastian Schneeweiss[1], Martin Kulldorff[1]

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

Maro, J et al. Using tree-based scan statistics to evaluate outcomes following incident antibiotic use. Sentinel Methods Protocol.
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

Maro, J et al. Using tree-based scan statistics to evaluate outcomes following incident antibiotic use. Sentinel Methods Protocol.
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

Maro, J et al. Using tree-based scan statistics to evaluate outcomes following incident antibiotic use. Sentinel Methods Protocol.
Kulldorff, M. Drug safety data mining with a tree-based scan statistic. PDS, 2013

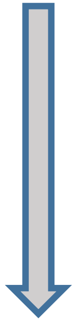
What is TreeScan™?

- 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

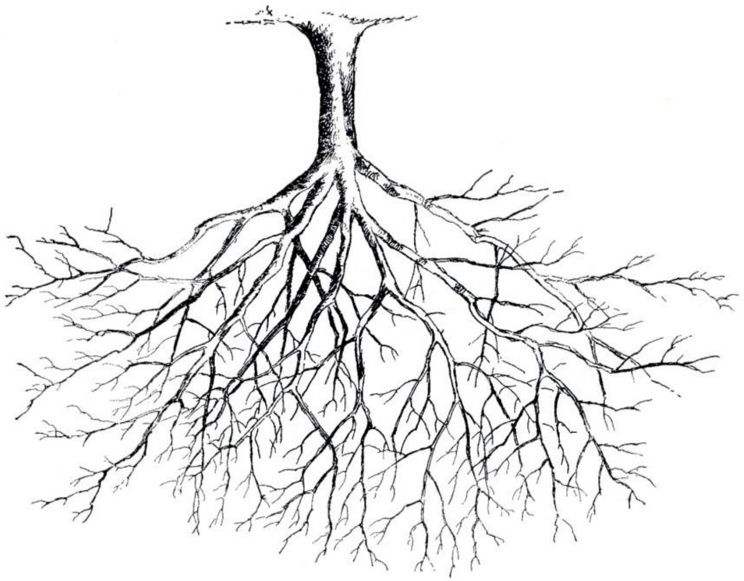
Maro, J et al. Using tree-based scan statistics to evaluate outcomes following incident antibiotic use. Sentinel Methods Protocol.
Kulldorff, M. Drug safety data mining with a tree-based scan statistic. PDS, 2013

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

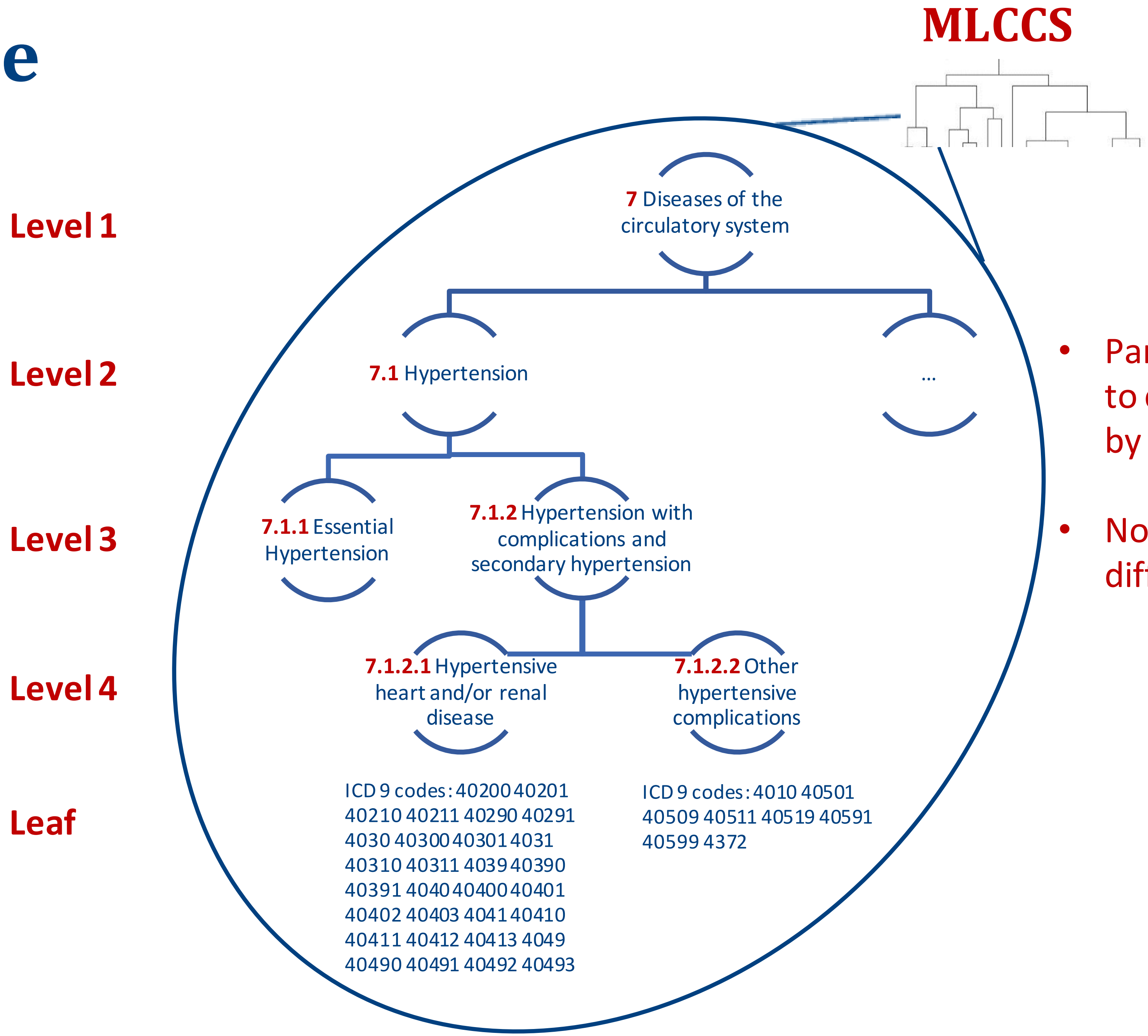


Greater specifity



Clinical Classifications software (CCS 2015). https://www.hcup-us.ahrq.gov/toolssoftware/ccs/CCSUsersGuide.pdf

The Tree



MLCCS

Level 1

7 Diseases of the circulatory system

Level 2

7.1 Hypertension

…

Level 3

7.1.1 Essential Hypertension

7.1.2 Hypertension with complications and secondary hypertension

Level 4

7.1.2.1 Hypertensive heart and/or renal disease

7.1.2.2 Other hypertensive complications

Leaf

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 4049 40490 40491 40492 40493

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

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

Clinical Classifications software (CCS 2015). https://www.hcup-us.ahrq.gov/toolssoftware/ccs/CCSUsersGuide.pdf

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

$$T = \max_G LLR(G)$$

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

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)

Maro, J et al. Using tree-based scan statistics to evaluate outcomes following incident antibiotic use. Sentinel Methods Protocol.
Kulldorff, M. Drug safety data mining with a tree-based scan statistic. PDS, 2013
Kulldorff, M. TreeScan User Guide, version 1.2

The Scan

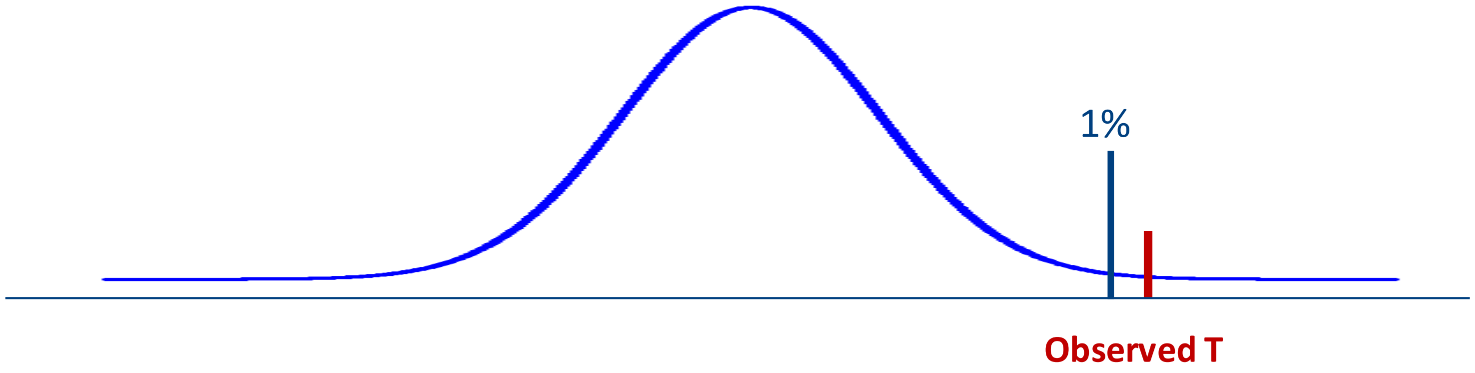
- T = unconditional Bernoulli scan statistic

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 ≥1% of T from 9999 datasets under the null

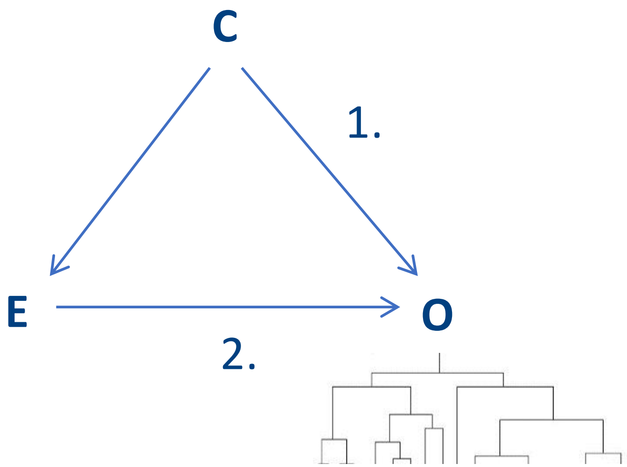
→ alert at alpha = 0.01



Maro, J et al. Using tree-based scan statistics to evaluate outcomes following incident antibiotic use. Sentinel Methods Protocol.
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



Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational statistics & data analysis.* Apr 2014;72:219-226

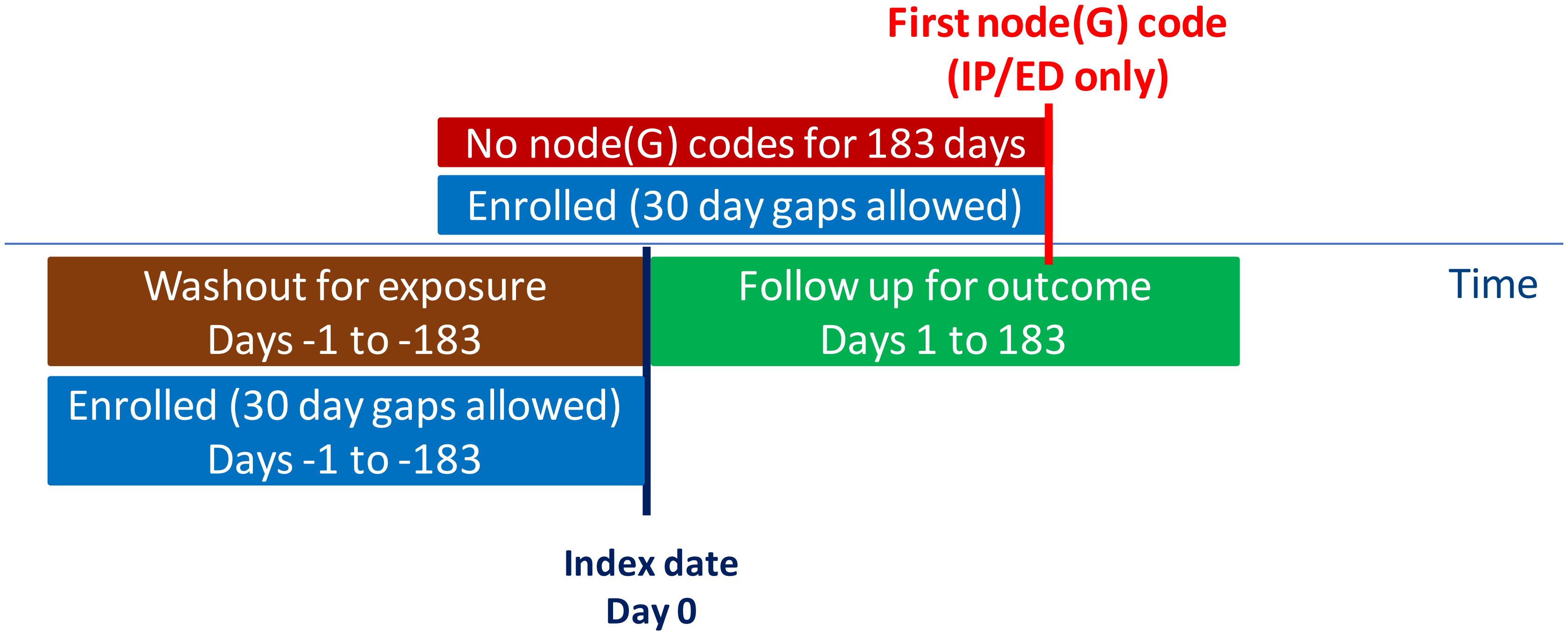
Methods and Process

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] formatted 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



https://www.sentinelinitiative.org/sentinel/surveillance-tools/software-toolkits/treeextraction-documentation

Methods and Process

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

Scenario	True Relative Risk	# Nodes w/ True Effect	Confounding?	Direction of Confounding
1	1.0	0	No	n/a
2			Yes	Positive (away from the null)
3	1.5	3	No	n/a
4	2.0			
5	4.0			
6	1.5	3	Yes	Positive (away from the null)
7	2.0			
8	4.0			
9	1.5	3	Yes	Negative (toward the null)
10	2.0			
11	4.0			

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

Level 1	Diseases of the digestive system
Level 2	Gastrointestinal hemorrhage
Level 3	**Hemorrhage from gastrointestinal ulcer**
Level 4	--
Leaf	*Numerous diagnosis codes*

Level 1	Diseases of the circulatory system
Level 2	Cerebrovascular disease
Level 3	**Acute cerebrovascular disease**
Level 4	Acute but ill-defined cerebrovascular accident Intracranial hemorrhage Occlusion of cerebral arteries
Leaf	*Numerous diagnosis codes*

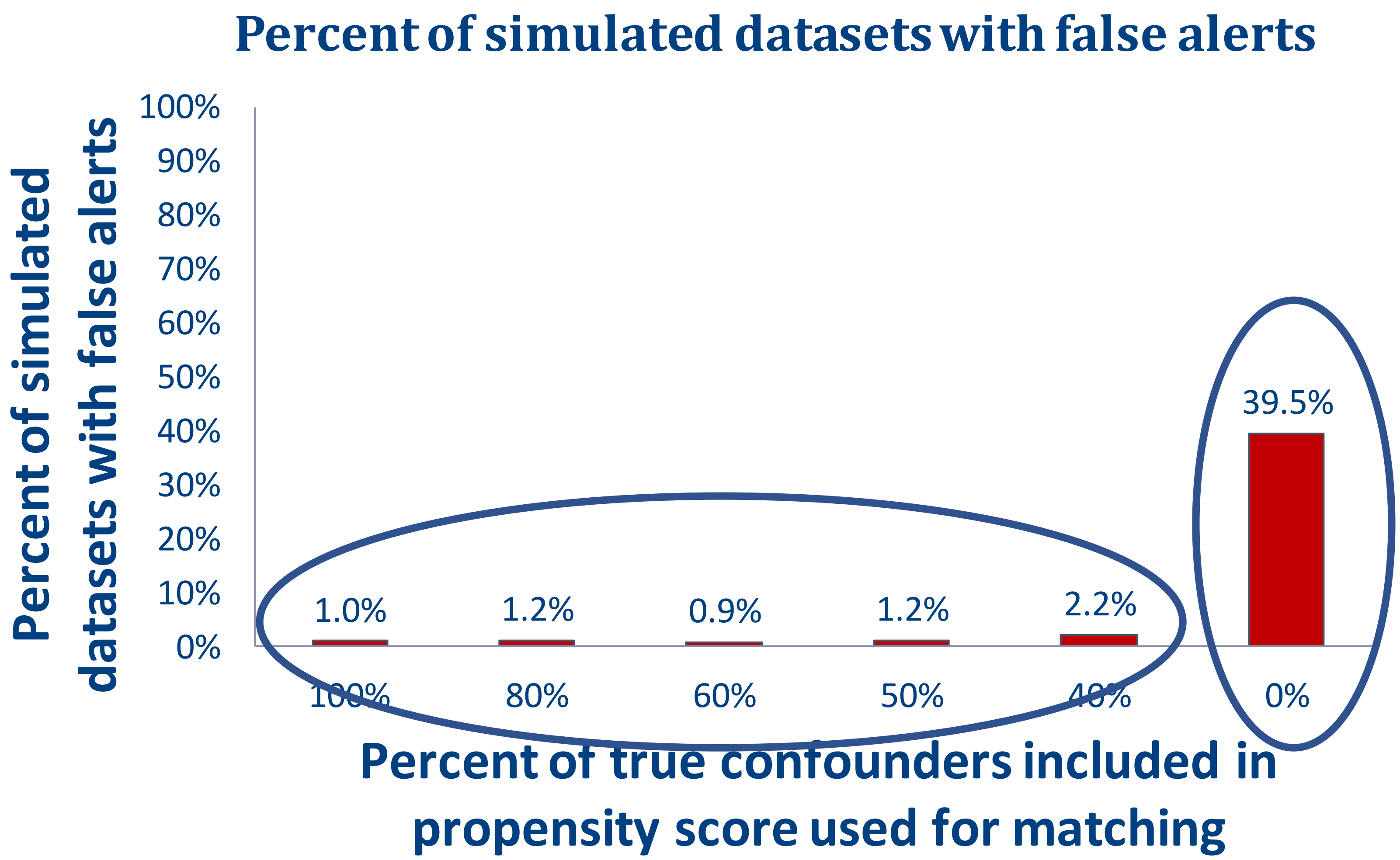
Level 1	Diseases of the genitourinary system
Level 2	Diseases of the urinary system
Level 3	**Acute and unspecified renal failure**
Level 4	Acute renal failure Unspecified renal failure
Leaf	*Numerous diagnosis codes*

Results: Take-home points

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

- 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

Node	Percent	MLCCS Level 3
08.06.01	32.6	Respiratory failure
03.08.01	18.6	Hyposmolality
06.03.01	17.7	Hemiplegia
07.01.02	17.5	Hypertension with complications
03.08.05	13.7	Other fluid and electrolyte disorders
17.01.05	11.0	Shock
10.01.03	10.4	Chronic kidney disease
Other	…	…

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