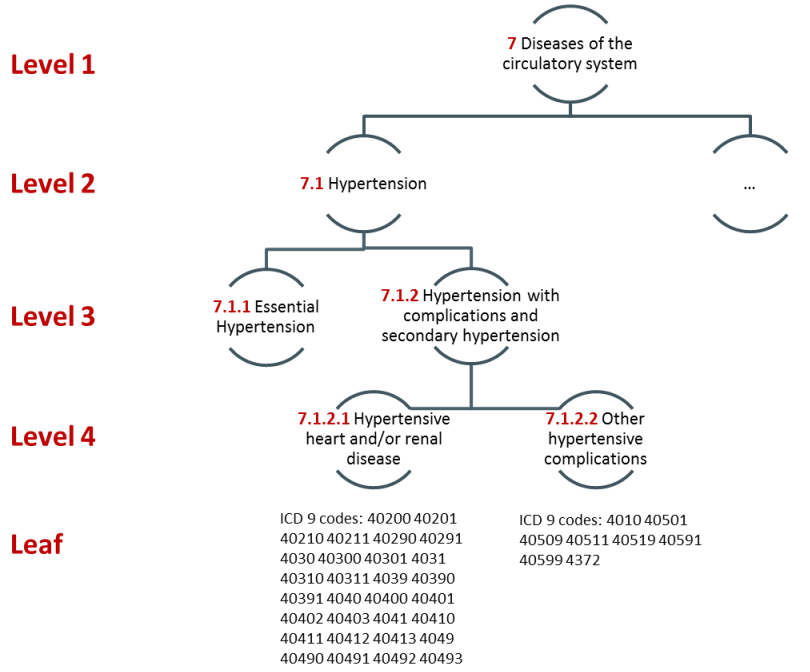
Figure 1. Example from Multi-Level Clinical Classification



Decimal points in ICD-9-CM codes at the leaf level have been removed.

The MLCCS based tree that we will use is a curated tree based on the 2015 version of ICD-9-CM codes. The tree has been independently curated by 2 members of FDA (**Appendix A**). Disagreements were adjudicated after discussion between the curators. Curation of the tree involved removal of conditions that were 1) congenital/hereditary, 2) unlikely to be caused by drugs (e.g. pregnancy, flu, well-visits) as well as 3) conditions with long induction times such as cancer (details on curation in the appendix). The diagnostic codes and their classification into different levels are not based on validated algorithms and could misclassify outcomes. Nevertheless, MLCCS classification of ICD-9-CM codes into clinical concepts can be useful as part of a screening tool for potential adverse events, followed by more rigorous and targeted protocol-based investigations. While an ICD-10-CM version of the MLCCS tree is available, our case studies will be conducted using data prior to widespread use of ICD-10-CM codes in the United States. We may elect to use trees that have been further curated for specific examples.

B. DEFINING INCIDENT OUTCOMES

We will define incident outcomes based on level 3 nodes across the MLCCS tree hierarchy. Incident outcomes will be defined by the first diagnosis from the node that occurs in the emergency department (ED) or inpatient (IP) setting; without any diagnoses in the same MLCCS level 3 node in the prior 183 days in any care settings. Multiple incident outcomes may be contributed by each patient as long as they meet the incidence criteria at MLCCS level 3 nodes.

Each patient will be allowed to enter the cohort only one time, after the first qualifying incident use of either the exposure or comparator of interest. Patients will be censored at death, disenrollment, or maximum days follow up for the example. If one member of a 1:1 propensity score matched set is censored, the other member will also be censored at the same time. Incident outcomes occurring during the patient's follow up after treatment initiation will be included in outcome counts for TreeScan.

C. UNCONDITIONAL BINOMIAL TREE SCAN STATISTIC

We will use the unconditional Bernoulli version of the tree-based scan statistic. This statistic conditions on the number of cases in the node but does not fix the total number of cases across the tree for each exposure group to be the same in the observed and randomly permuted data. The threshold for alerting in both Aims 1 and 2 will be $p \leq 0.01$ (1-sided).

The distribution of the test statistic *T* below is unknown. However, a Monte Carlo based p-value can be obtained by generating random datasets under the null hypothesis that every outcome occurs, independently of other outcomes, with the same probability among in the treatment group versus the comparator group.[4]

The log likelihood ratio (LLR) based test statistic *T* can be calculated as:

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

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

Where: T = unconditional Bernoulli tree scan statistic

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

Random datasets can be generated under the null hypothesis by creating replicates of the original data where each node contains the same number of events as observed in the original data, however the events within each node are assigned to exposure based on a binomial draw with the expected proportion based on the null hypothesis. In our 1:1 matched setting, this proportion is 0.5. For these Monte Carlo generated data sets, the outcomes in each node are assigned randomly. If 9,999 random replicates are generated, and ranked according to *T,* then the Monte Carlo based p-value = Rank of the observed data/(9999+1). When the type 1 error alpha for alerting is set to a threshold of 0.01, then only nodes with rank within the top 1% of the real and random replicates will constitute a statistical alert. If the null hypothesis is true, the probability that all p-values are larger than 0.01 is 99%.

Hypothesis tests will be performed at level 3 and all more finely specified nodes at numerically higher levels of the MLCCS, including level 4 and the leaf level (specific ICD-9-CM codes). A LLR will be computed at every node where a hypothesis test is performed.

Some of the major strengths of tree-based scan statistics include:

1. They were developed based on scan statistical theory
2. They use a hierarchical diagnosis tree to simultaneously evaluate outcomes at different levels of granularity (including specific diagnoses and groups of related diagnoses)
3. They use a frequentist method to formally adjust for the multiple testing inherent in evaluation of thousands of potential adverse events that accounts for correlation between tests of related hypotheses (unlike traditional frequentist methods which are too conservative)
4. They can be useful when screening for unanticipated safety signals where there is no informative prior

48. Layton D, Hughes K, Harris S, Shakir SA. Comparison of the incidence rates of selected gastrointestinal events reported for patients prescribed celecoxib and meloxicam in general practice in England using prescription-event monitoring (PEM) data. Rheumatology 2003;42:1332-41.

49. Asghar W, Jamali F. The effect of COX-2-selective meloxicam on the myocardial, vascular and renal risks: a systematic review. Inflammopharmacology 2015;23:1-16.

50. Layton D, Hughes K, Harris S, Shakir SAW. Comparison of the incidence rates of thromboembolic events reported for patients prescribed celecoxib and meloxicam in general practice in England using Prescription-Event Monitoring (PEM) data. Rheumatology 2003;42:1354-64.

51. Valproic acid and sodium valproate approved for use in epilepsy. FDA drug bulletin 1978;8:14-5.

52. Goldenberg MM. Overview of Drugs Used For Epilepsy and Seizures: Etiology, Diagnosis, and Treatment. Pharmacy and Therapeutics 2010;35:392-415.

53. Highlights of Prescribing Information. (Accessed July 6, 2018, at https://www.accessdata.fda.gov/drugsatfda_docs/label/2009/022251,020764s029,020241s036lbl.pdf.)

54. Karceski SG, P; Dashe, JF. Initial treatment of epilepsy in adults. UpToDate. Jun 2018 ed.

55. Depakote Tablets (divalproex sodium). (Accessed July 6, 2018, at https://www.accessdata.fda.gov/drugsatfda_docs/label/2009/018723s039lbl.pdf.)

56. Lamictal (lamotrigine): Drug Safety Communication - Serious Immune System Reaction. (Accessed 7/6/2018, at https://www.fda.gov/Safety/MedWatch/SafetyInformation/SafetyAlertsforHumanMedicalProducts/ucm605628.htm.)

VIII. APPENDICES

A. CURATION OF THE MLCCS TREE

Two members of the FDA independently reviewed the MLCCS tree based on 2015 ICD-9-CM codes to remove codes that were unlikely to be caused by drug exposures within a short follow up window. Each reviewer flagged codes that were:

1. of known etiology (e.g. pregnancy, congenital condition),
2. unlikely to be an adverse reaction caused by drugs (e.g. gingival recession, recurrent dislocation of shoulder),
3. not an incident diagnosis (e.g. alcoholism in family, social maladjustment),
4. for conditions with long latency/induction periods (e.g. cancer, osteoporosis)

After adjudication of disagreement, 7,078 of 15,075 (47%) ICD-9-CM diagnostic codes were excluded.

Source:

https://www.sentinelinitiative.org/sentinel/surveillance-tools/software-toolkits/treeextraction-documentation (supporting tree file)

B. DETAILED PROTOCOL SPECIFICATIONS FOR EXAMPLES

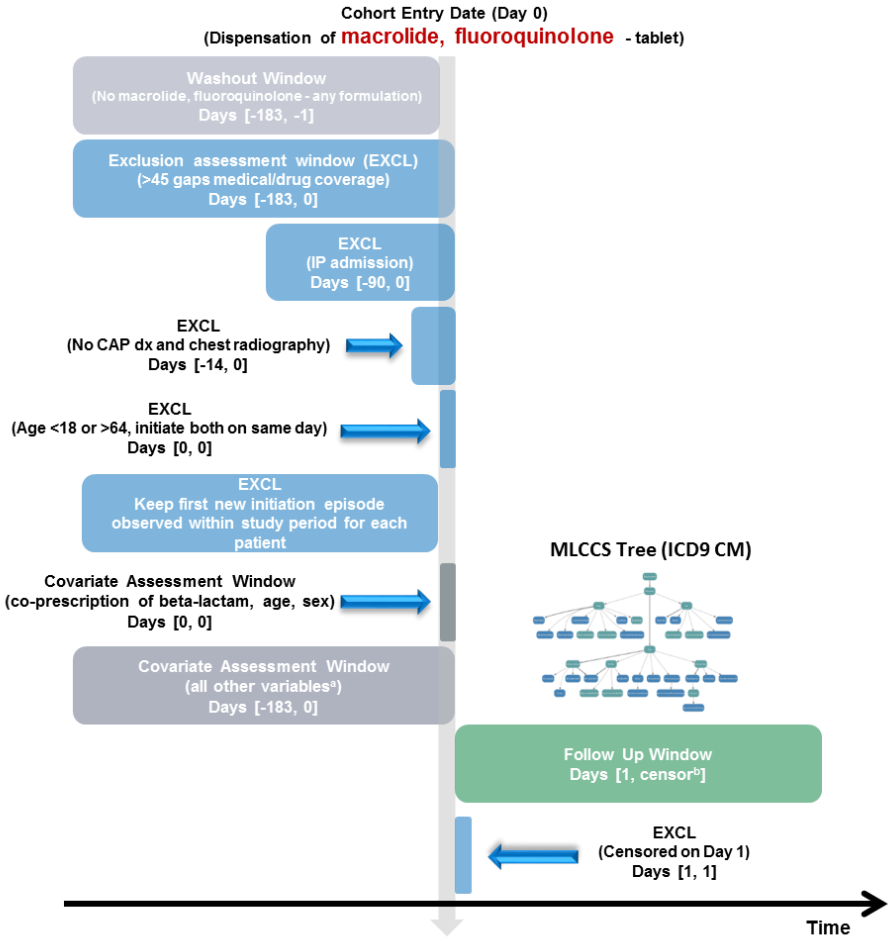
Example 1

[a]All other variables considered in candidate global propensity scores

- Age (continuous)
- Gender
- Metastatic cancer
- Tumor
- Arrhythmia
- Congestive heart failure
- Dementia
- Renal failure
- Weight loss
- Hemiplegia
- Alcohol abuse
- Pulmonary disease
- Coagulopathy
- Complicated diabetes
- Anemia
- Fluid and electrolyte disorder
- Liver disease
- Peripheral vascular disorder
- Psychosis
- Pulmonary circulation disorders
- HIV/AIDS
- Hypertension
- Degenerative disease of central nervous system
- Durable medical equipment
- Vaccine administration
- Screening examinations and disease management training
- Pap smear
- HPV DNA test
- Mammogram
- Fecal occult blood test
- Colonoscopy
- PSA test
- Number of inpatient hospitalizations
- Number of outpatient visits
- Number of emergency department visits
- Number of unique generics
- Prior prescription of penicillins
- Prior prescription of cephalosporins
- Prior prescription of sulfonamides
- Prior prescription of tetracyclines
- Prior prescription of aminoglycosides
- Co-prescription of beta-lactam
- Pregnancy at time of initiation
- Empirically selected

[b]Censoring

- 183 days
- Sep 30, 2015
- Discharged dead
- Disenroll medical or drug (45 day gaps allowed)



Example 2

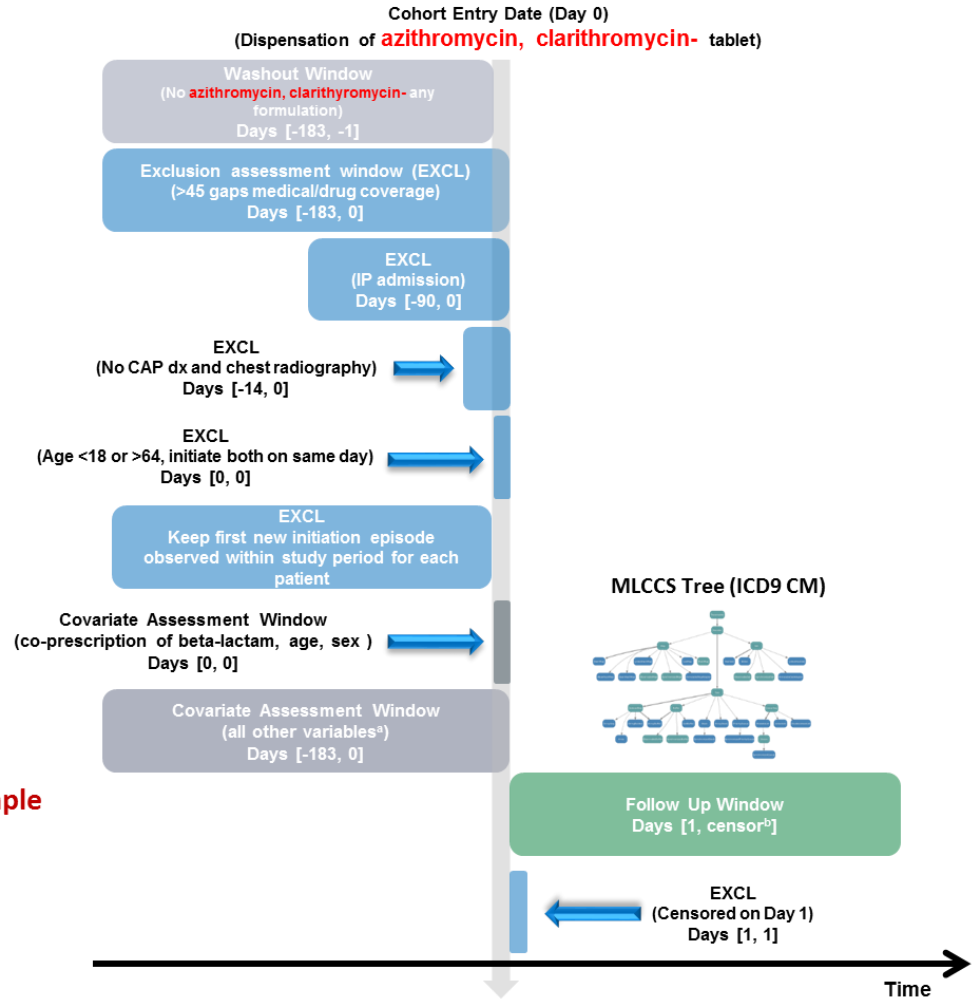
[a]All other variables considered in candidate global propensity scores

- Age (continuous)
- Gender
- Metastatic cancer
- Tumor
- Arrhythmia
- Congestive heart failure
- Dementia
- Renal failure
- Weight loss
- Hemiplegia
- Alcohol abuse
- Pulmonary disease
- Coagulopathy
- Complicated diabetes
- Anemia
- Fluid and electrolyte disorder
- Liver disease
- Peripheral vascular disorder
- Psychosis
- Pulmonary circulation disorders
- HIV/AIDS
- Hypertension
- Degenerative disease of central nervous system
- Durable medical equipment
- Vaccine administration
- Screening examinations and disease management training
- Pap smear
- HPV DNA test
- Mammogram
- Fecal occult blood test
- Colonoscopy
- PSA test
- Number of inpatient hospitalizations
- Number of outpatient visits
- Number of emergency department visits
- Number of unique generics
- Prior prescription of penicillins
- Prior prescription of cephalosporins
- Prior prescription of sulfonamides
- Prior prescription of tetracyclines
- Prior prescription of aminoglycosides
- Prior prescription of other macrolides
- Prior prescription of fluoroquinolones
- Co-prescription of beta-lactam
- Pregnancy at time of initiation
- Empirically selected

Tailored to example

[b]Censoring

- 183 days
- Sep 30, 2015
- Discharged dead
- Disenroll medical or drug (45 day gaps allowed)



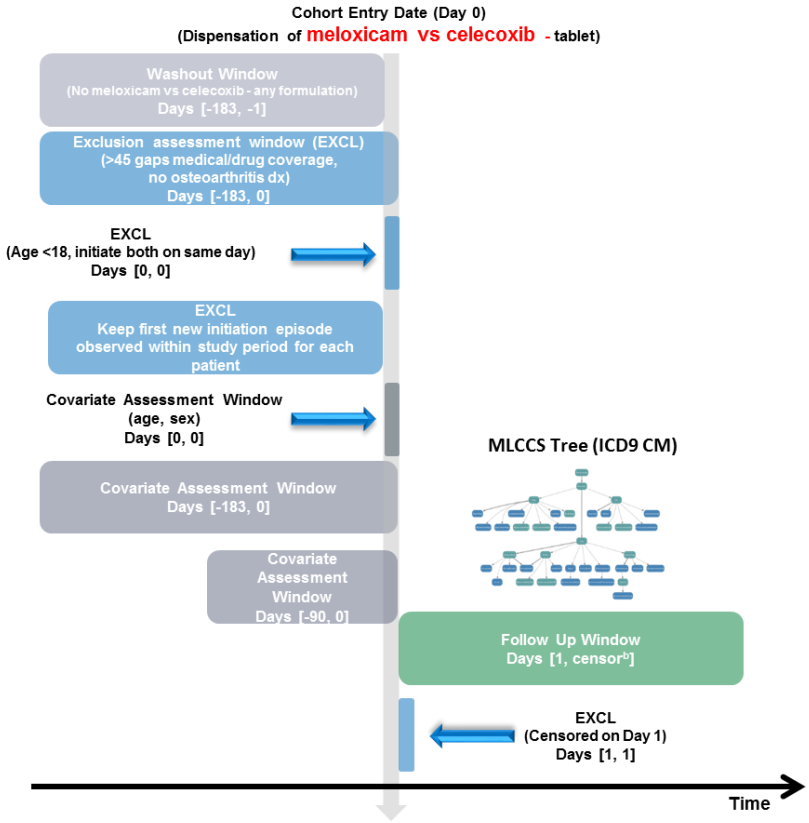
Example 3

Predefined Global Covariates Days:
Empirical Covariates Days:
Tailored Covariates Days:

- Upper GI events
- Lower GI events
- Myocardial infarction
- Cerebrovascular events (stroke, TIA)
- Renal failure, acute kidney injury, hyperkalemia
- Obesity
- Smoking
- Angina
- Coronary Revascularization
- Pregnancy at the time of initiation
- Fibromyalgia
- Rheumatoid arthritis
- Hormone Replacement Therapy
- Statins
- Opioids
- Non-selective NSAIDS
- Selective NSAIDS

Tailored Covariates:

- Anticoagulants
- Antiplatelets
- Antidepressants
- Fluconazole
- Lithium
- Antihypertensives
- Cyclosporine
- Methotrexate
- Steroids
- PPI
- H2B



[b]**Censoring**

- 183 days
- Sep 30, 2015
- Discharged dead
- Switch
- Disenroll medical or drug (45 day gaps allowed)

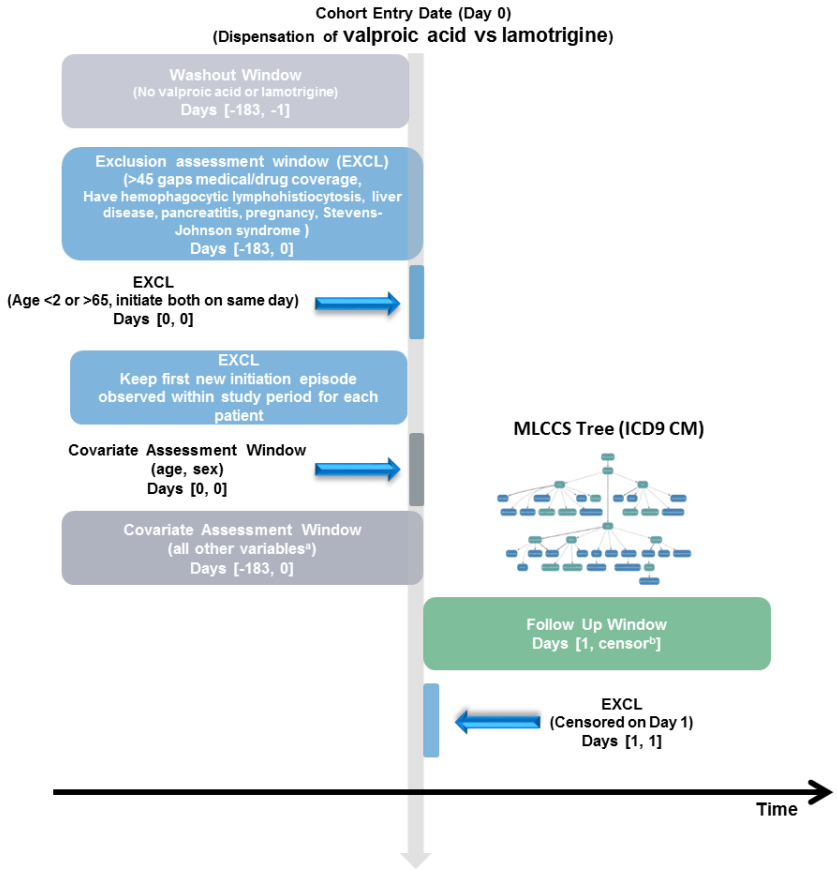
Example 4

[a] **Covariates**
Predefined Global Covariates
Empirical Covariates
Tailored Covariates:

- HIV infection treatments
- Other viral infections (mumps, flu, herpes, Coxsackie, Epstein-Barr, cytomegalovirus, parvovirus B19, pneumocystosis and histoplasmosis)
- Bacterial infection
- Organ transplant
- Autoimmune diseases
- Chemotherapy
- Cancer
- Acute b-lymphoblastic leukemia
- Medicines for gout (e.g. NSAIDS, corticosteroids, other)
- Sulfa antibiotics (e.g. Bactrim, septra),
- Alcohol use disorders
- Gallstones
- Cystic fibrosis
- Kawasaki disease
- Reye's syndrome
- Hemolytic uremic syndrome (HUS)
- Thrombotic thrombocytopenic purpura (TTP)
- Hyperparathyroidism
- Migraine
- Bipolar disorder
- Epilepsy/Convulsions
- Depression
- Schizophrenia
- Other anticonvulsants
- Antidepressants: SSRI, SNRI, TCA, MOA, Atypical, Other
- Atypical antipsychotics, typical antipsychotics
- Short, long acting BZD
- Stroke
- TBI

[b] **Censoring**

- 183 days
- Sep 30, 2015
- Discharged dead
- Switch
- Disenroll medical or drug (45 day gaps allowed)



C. RESPONSES TO PUBLIC COMMENT

Comments on "Development and Evaluation of a Global Propensity Score for Data Mining with Tree-Based Scan Statistics"

1. Given the large number of available empirical approaches for model selection, it could be helpful to provide motivation for why hdps was chosen for this evaluation as opposed to other options. For instance, Karim et al. (Epidemiology 2018 Mar; 29(2): 191-198) recently showed that a machine learning with hdps hybrid often outperforms hdps alone.

   This paper found that machine learning based approaches such as LASSO and ElasticNet in combination with hdPS performed marginally better than hdPS alone in the context of selection based on potential for bias for a single outcome. The machine learning component of the hybrid empirical variable selection methods worked to further reduce the dimensionality of variables identified with hdPS.

   In our context, we are scanning across thousands of potential outcomes. It would not be feasible to apply a hybrid approach which selects variables based on association with outcome. Furthermore, it may be helpful in our scanning context to include a slightly broader base of variables to provide proxy adjustment for confounders on a wider range of outcomes.

   We will include this citation and a brief explanation as above in the background.

2. Similarly, it could be useful to motivate why the TreeScan methodology was selected as opposed to other scan statistics (or, minimally, to provide its major advantages and limitations in this specific setting).

   We will list some of the major strengths and limitations of TreeScan in the protocol. A comparison of different signal detection methods is currently underway in another task order (signalx3).

   Strengths:

   i. Developed based on scan statistical theory
   ii. Use a hierarchical diagnosis tree to simultaneously evaluate outcomes at different levels of granularity (including specific diagnoses and groups of related diagnoses)
   iii. Use a frequentist method to formally adjust for the multiple testing inherent in evaluation of thousands of potential adverse events that accounts for correlation between tests of related hypotheses (unlike traditional frequentist methods which are too conservative)
   iv. Useful when screening for unanticipated safety signals where there is no informative prior

Weaknesses:

v. Bias is adjusted by design, not inherent in the scan statistic

vi. Hierarchical classification system for outcomes are not based on validated algorithms

vii. Adjusting for multiplicity when scanning across outcomes will decrease power compared to evaluating a single pre-specified hypothesis

3. What is the rationale for NOT focusing on data in the ICD-10 era, since future safety studies will this new system and your ICD-9 based results may not (?) be seamlessly generalizable. Minimally, it could be helpful to comment on why an ICD-9 based evaluation is proposed and why you don't have major concerns that your conclusion won't be limited by this feature.

   Although different hierarchical trees may have different properties, the method of TreeScan with PS-matching is not tied to a particular coding system and should be extensible.

   We deliberately chose older examples with known safety profiles. Given the delays in data refreshes, we have limited years of data available after Oct 2015. Although an ICD10 based tree is available, if we focused only on the ICD-10 era, we would have lower power to detect known signals in our examples. Doing hdPS in a mixed ICD9-10 era would require incorporation of mapping and is beyond the scope of this project.

4. It could be helpful to explain why you chose to focus your Aim 1 evaluation on a subset of outcomes (feasibility?) and whether/how this may limit the generalizability of the evaluation of method performance. For instance, why would we expect performance metrics for the selected set of outcomes to carry-over to the other 100's of outcomes one might evaluate? Are there any performance metrics that could be feasibly evaluated for 'all' outcomes to avoid issues of selection and uncertain generalizability?

   The general performance metric of balance on predefined and empirically selected covariates used in the PS will apply to all outcomes. However, the covariates in the PS will not necessarily be relevant or optimal for all outcomes in all drug comparison scenarios. It would be infeasible to identify the best set of covariates for each outcome and evaluate balance on each.

   That is why we will be focusing our deeper dive on balance for known risk factors on a subset of outcomes with and without alerts in a variety of examples where we have prior knowledge of where true signals may or may not be present. These examples are intended to be diverse with respect to study populations as well as types of outcomes with true signals. That said, the performance in these examples will not necessarily be generalizable to all contexts.

   Realistically, most of the potential outcomes scanned will be unrelated to the evaluated drugs. By focusing on areas where there are known signals or unanticipated alerts, we target high yield areas for learning about the method and its performance.

   We will include discussion to that effect in the protocol.

5. How will you ensure that you have identified example scenarios with adequate power to find signals of interest?

   As in a real surveillance activity, we may not necessarily have adequate power to find signals of interest at stringent pre-specified alpha levels. In each of our example scenarios, we have known signals that were previously identified. We will be looking at patterns of alerting in these examples to observe how signal detection using the method could play out in a real scenario. Outcomes that don't alert at the pre-specified threshold may still have relatively low likelihood under the null. The method can play an important part in screening and prioritization even if there is not sufficient power to alert at a pre-specified threshold by painting a clinical picture of the pattern of outcomes that are unlikely to be observed if there was no relationship with exposure.

6. Is there an existing 'standard' data mining method with which you can compare your new global PS-based options? This may help give some evidence about the level of improvement your methods may provide beyond what a more basic approach that a researcher might do typically in practice now.

   There does not appear to be a 'standard' cohort based approach. The signalx3 workgroup will be comparing three signal detection methods that use a self-controlled design, which are more frequently used in pharmacovigilance activities.

7. You might consider referencing an article from our group that motivates the importance of work in this research area. Although this review was more narrowly written in the context of vaccine safety, its conclusions apply to drug safety applications as well.

   This is very relevant. We will cite this review.

   Nelson JC, Shortreed S, Yu O, et al. on behalf of the Vaccine Safety Datalink project. Integrating database knowledge and epidemiological design to improve the implementation of data mining methods to evaluate vaccine safety in large healthcare databases. *Stat Analysis Data Mining* 2014 Oct;7(5):337-351. doi:10.1002/sam.11232

D. FOLLOW UP ANALYSES FOR EXAMPLES 1-3

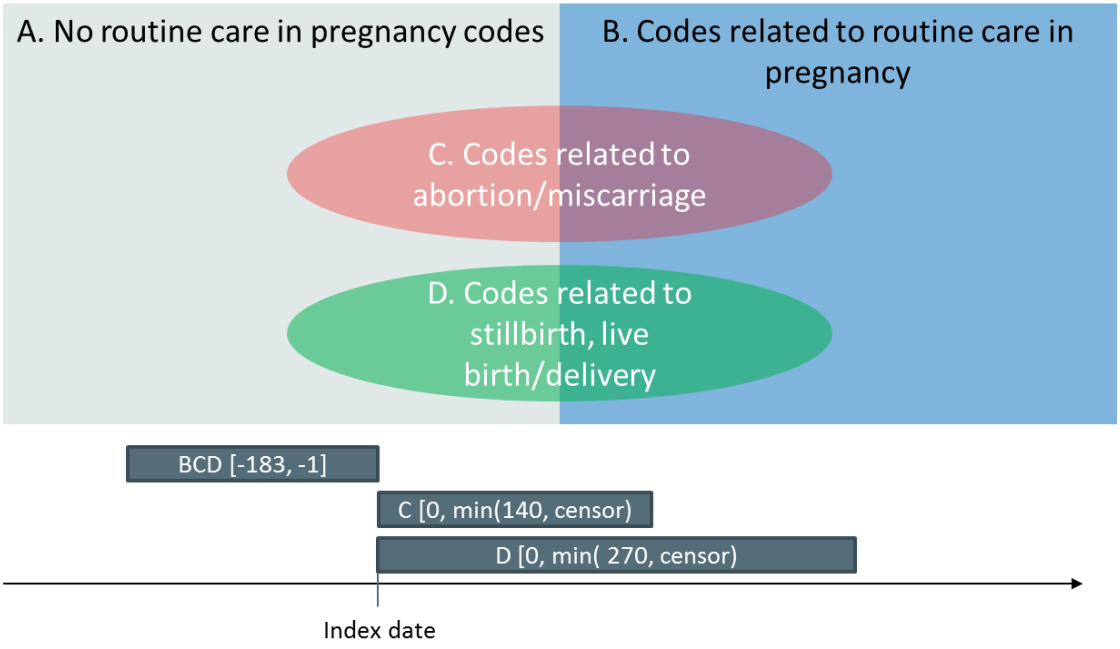
After observing the results for examples 1-3, we decided to initiate post-hoc follow up analyses that would allow us to 1) dig deeper and better understand how baseline pregnancy status affected the results observed after implementing the pre-specified protocol for examples 1 and 2, and 2) evaluate how explicitly balancing on nodes that were highly ranked by LLR would affect subsequent alerting patterns for example 3.

Example 1:

Pregnancy was originally included as a tailored confounder in the pre-specified protocol because we anticipated channeling due to different patterns of antibiotic use in pregnancy. The baseline pregnancy algorithm that we used in the pre-specified analyses may not be ideal. The follow up analyses are exploratory, digging deeper into how we measure pregnancy - how well are we capturing it with different algorithms and how does using different definitions of pregnancy as an exclusion criterion affect results. Screening for outcomes associated with drug exposure in pregnancy will be evaluated in another project. Follow up analyses include first re-analyzing the data after exclusion of pregnant women using different code algorithms and assessment windows, some of which case a broader net to

remove patients that were pregnant during the assessment window, second, describing the impact of these different definitions of pregnancy on the study population and results, and third, generating tables to show balance on empirically identified covariates for matched populations that may not have been matched on a PS that included those covariates (additional detail below).

1. Reanalyze after removing pregnant women (ever during the assessment window)
    - Codes indicative of
        - Routine care during pregnancy
        - Live birth/delivery
        - Other end of pregnancy (stillbirth, miscarriage, termination)
    - Compare results with different assessment windows for pregnancy
        - [-183, 1] any codes +
          [0, min(140, censor)] for miscarriage, abortion codes +
          [0, min(270, censor)] for stillbirth or delivery codes
2. Describe frequency of exclusion with different codes and assessment windows



3. Add tables showing balance on empirically identified covariates when they are versus are not included in the PS

Example 2:

Same follow up analyses as example 1.

Example 3:

The follow up analyses include restriction of data to time prior to a generic version of celecoxib entering the market in case the availability of a generic affected prescribing patterns for patients at different risk of adverse outcomes. Follow up analyses also will involve reanalysis after inclusion of nodes that alerted at the pre-specified threshold or were highly ranked in the propensity score to allow evaluation of scanning results after balancing on those nodes at baseline.

1. Restrict years of study to before May 2014 (before generic celecoxib came on the market)
2. Re-analyze after adding prior history of top nodes from LLR ranking to PS (e.g. stroke, headache, cerebral infarct) and mental health related covariates (e.g. depression, anxiety, antidepressants, anti-anxiety).
3. Add tables showing balance on empirically identified covariates when they are versus are not included in the PS

E. ADDITIONAL JUSTIFICATION FOR DESIGN DECISIONS

Justification for allowing tailored covariate assessment window to differ from fixed predefined and empirical covariate assessment windows:

Predefined covariates and empirically identified covariates can be generically applied to every exposure-comparator evaluation without customization for different exposure-comparator evaluations.

Tailored covariates represent thoughtful consideration of what investigators think is relevant for the exposure-comparator pair and at least one potential adverse event. As such, there are no limitations on which or how tailored covariates are defined. The tailored covariates can be highly customized – including not just different variables but different assessment windows, requirement multiple diagnoses within certain time frames or more complicated algorithms.

Thus, the contrast between PSs that include tailored covariates versus those that do not is between a base case scenario of generically applied covariates using default assessment windows versus the addition of investigator selected covariates defined using whatever covariate definitions the investigators think are the most relevant to the exposure-comparator evaluation. These tailored covariates can and will vary in terms of which variables are included and complexity of definition across different evaluations.

Justification for allowing ascertainment of pregnancy at the time of drug initiation using codes recorded after index date for drug initiation

In pharmacoepidemiology studies, we generally avoid using future information to make decisions about whether patients are eligible for cohort entry, and for good reason. For example, to determine whether patients should enter the cohort at treatment initiation, we generally would not want to require that they have a full year of follow-up after treatment initiation during which they remain alive. If the outcome(s) of interest could sometimes be fatal or were correlated with mortality in other ways, this can create immortal time bias, a form of selection bias. The bias arises when the factor we are conditioning on (mortality, in this example) can be affected by the exposure(s) of interest (or some common cause of exposure and the factor).

In contrast, when using information during follow-up to identify patients who have birth outcomes, the objective is to exclude those that had to have been pregnant prior to the start of follow-up. In this

situation, the pregnancies that began prior to treatment initiation could not have been affected by the treatment. There are other (and perhaps rare) situations in which it would also be valid to use future information. For example, imagine that we wanted to conduct an analysis among patients who were 65 years of age or older and we did not know patients' ages at the time of cohort entry, but we did know their ages one year later (or at the time of death for anyone who may have died within that year). In this case, it is safe to use the future information about age to infer patients' ages at the time of cohort entry because the exposure(s) could not have affected patients' age. This is analogous to the pregnancy situation.

There are other reasons to avoiding using future information to define study variables – such as looking into the future to assign exposure status at the start of follow-up, which is classical immortal time bias. However, these are slightly different situations than the pregnancy scenario.