6-12-2017

Statistical Power for Postlicensure Medical Product Safety Data-Mining

Judith C. Maro
Harvard Medical School and Harvard Pilgrim Health Care Institute, jmaro@mit.edu

Michael D. Nguyen
U.S. Food and Drug Administration

Inna Dashevsky
Harvard Medical School and Harvard Pilgrim Health Care Institute

Meghan A. Baker
Harvard Medical School and Harvard Pilgrim Health Care Institute

See next pages for additional authors

The Electronic Data Methods (EDM) Forum is supported by the Agency for Healthcare Research and Quality (AHRQ), Grant 1U18HS022789-01. eGEMs publications do not reflect the official views of AHRQ or the United States Department of Health and Human Services.

Statistical Power for Postlicensure Medical Product Safety Data-Mining

Abstract

Objective: To perform sample size calculations when using tree-based scan statistics in longitudinal observational databases.

Methods: Tree-based scan statistics enable data-mining on epidemiologic datasets where thousands of disease outcomes are organized into hierarchical tree structures with automatic adjustment for multiple testing. We show how to evaluate the statistical power of the unconditional and conditional Poisson versions. The null hypothesis is that there is no increase in the risk for any of the outcomes. The alternative is that one or more outcomes have an excess risk. We varied the excess risk, total sample size, frequency of the underlying event rate, and the level of across-the-board healthcare utilization. We also quantified the reduction in statistical power resulting from specifying a risk window that was too long or too short.

Results: For 500,000 exposed people, we had at least 98% power to detect an excess risk of 1 event per 10,000 exposed for all outcomes. In the presence of potential temporal confounding due to across-the-board elevations of healthcare utilization in the risk window, the conditional tree-based scan statistic controlled type I error well, while the unconditional version did not.

Discussion: Data-mining analyses using tree-based scan statistics expand the pharmacovigilance toolbox, ensuring adequate monitoring of thousands of outcomes of interest while controlling for multiple hypothesis testing. These power evaluations enable investigators to design and optimize implementation of retrospective data-mining analyses.

Acknowledgements

We gratefully acknowledge comments received from Katherine Yih and Jeffrey S. Brown on this project and manuscript as well as the project management efforts of Carolyn Balsbaugh.

Keywords

data-mining, population health, data analysis method, research networks, pharmacovigilance, signal detection

Disciplines

Clinical Epidemiology | Epidemiology | Other Public Health

Authors

Judith C Maro, *Harvard Medical School and Harvard Pilgrim Health Care Institute*; Michael D Nguyen, *U.S. Food and Drug Administration*; Inna Dashevsky, *Harvard Medical School and Harvard Pilgrim Health Care Institute*; Meghan A Baker, *Harvard Medical School and Harvard Pilgrim Health Care Institute*; Martin Kulldorff, *Harvard Medical School and Brigham and Women's Hospital*.

eGEMs
Generating Evidence & Methods to improve patient outcomes

Statistical Power for Postlicensure Medical Product Safety Data Mining

Judith C. Maro, PhD;[i,ii] Michael D. Nguyen, MD;[iii] Inna Dashevsky, MS;[i,ii] Meghan A. Baker, MD, PhD;[i,ii] Martin Kulldorff, PhD[iv]

ABSTRACT

Objective: To perform sample size calculations when using tree-based scan statistics in longitudinal observational databases.

Methods: Tree-based scan statistics enable data mining on epidemiologic datasets where thousands of disease outcomes are organized into hierarchical tree structures with automatic adjustment for multiple testing. We show how to evaluate the statistical power of the unconditional and conditional Poisson versions. The null hypothesis is that there is no increase in the risk for any of the outcomes. The alternative is that one or more outcomes have an excess risk. We varied the excess risk, total sample size, frequency of the underlying event rate, and the level of across-the-board health care utilization. We also quantified the reduction in statistical power resulting from specifying a risk window that was too long or too short.

Results: For 500,000 exposed people, we had at least 98 percent power to detect an excess risk of 1 event per 10,000 exposed for all outcomes. In the presence of potential temporal confounding due to across-the-board elevations of health care utilization in the risk window, the conditional tree-based scan statistic controlled type I error well, while the unconditional version did not.

Discussion: Data mining analyses using tree-based scan statistics expand the pharmacovigilance toolbox, ensuring adequate monitoring of thousands of outcomes of interest while controlling for multiple hypothesis testing. These power evaluations enable investigators to design and optimize implementation of retrospective data mining analyses.

[i]Harvard Medical School, [ii]Harvard Pilgrim Health Care Institute, [iii]United States Food and Drug Administration, [iv]Brigham and Women's Hospital

Introduction

New methods are emerging that enable data mining for unsuspected drug and vaccine adverse reactions in large longitudinal databases, such as the United States Food and Drug Administration's (FDA's) Sentinel System,[1] a distributed data network of administrative claims databases. Data mining is a technique for simultaneous monitoring of many exposure-outcome pairs without having to pre-specify particular pairs of interest. Data mining analyses have traditionally been performed using spontaneous reporting databases, which lack denominator data. Moreover, spontaneous reports require suspicion by a health care worker or patient that an outcome is potentially the result of exposure to a given medical product, which means that these databases suffer from selective reporting of particular exposure-outcome pairs and persistent underreporting.[2]

The longitudinal nature of administrative claims data provides the ability to systematically evaluate thousands of outcomes as potential adverse reactions. Data mining analyses using longitudinal data can act as a wide-ranging safety net, ensuring that rate and count data are collected and analyzed routinely. Such general safety surveillance can fulfill the congressional mandate to provide access to safety data summaries utilizing its new pharmacovigilance infrastructure[1] including identification of any new risks not previously identified, potential new risks, or known risks reported in unusual number.[3]

Here, we focus on one data mining method that leverages these longitudinal data: the tree-based scan statistic.[4] Previously, it has been shown to perform well in postmarket medical product safety settings,[5–7] and is planned to monitor nine-valent human papillomavirus vaccine exposure in the FDA's Sentinel System.[8] Additionally, the United States Centers for Disease Control and Prevention have indicated that they intend to use the method in their vaccine monitoring system, the Vaccine Safety Datalink.[9,10] First, analytic datasets containing rate or count data for many disease outcome pairs are assembled using familiar epidemiologic designs

Table 1. Example Branch of the Multi-Level Clinical Classifications Software Tree

TREE LEVEL	TREE NODE	TREE NODE NAME
1	06	Diseases of the nervous system and sense organs
2	06.04	Epilepsy; convulsions
3	06.04.02	Convulsions
4	06.04.02.00	Convulsions
5 / Leaf	ICD-9-CM 780.3	Convulsions
5 / Leaf	ICD-9-CM 780.31	Febrile convulsions not otherwise specified
5 / Leaf	ICD-9-CM 780.32	Complex febrile convulsions
5 / Leaf	ICD-9-CM 780.33	Post traumatic seizures
5 / Leaf	ICD-9-CM 780.39	Other convulsions

Volume 5 (2017) Issue Number 1

that control for confounding. Second, data for these outcomes are organized into a hierarchical tree. For example, febrile seizures can be combined with other similar outcomes under a more general heading, e.g., convulsions. Table 1 shows a very small part of an example tree. Then, the tree-based scan statistic is calculated for the entire analytic dataset using maximum likelihood estimation and Monte Carlo hypothesis testing to automatically control for multiplicity among the many outcomes being evaluated.

In the analyses described herein, the hierarchical tree is predefined based on clinical knowledge and used to structure data, and the main results are the expected statistical power. The null hypothesis is that there is no elevated risk for any of the thousands of outcomes. Conceptually, this use of the tree is very different from the tree structures created by classification and regression trees (CART), another data-mining method. In those analyses, the trees are the results of the analyses and that work is aimed at tree generation itself.

The advantage of employing a pre-defined hierarchical tree structure to arrange the analytic dataset is that it allows one to "borrow strength" when a clinical concept may be coded in multiple ways. Therefore, it is unnecessary for the investigator to specify which set of codes is used to identify a particular clinical concept. Additionally, a clinical concept can be experienced somewhat differently by certain individuals in a population, and the tree allows biologically-related reactions to be aggregated.

Nelson et al. have published a comprehensive review paper that describes other data mining techniques in longitudinal data,[11] including logistic regression approaches,[12,13] and disproportionality analyses.[14–20] The former approaches execute multiple logistic regression analyses and then use post hoc techniques to adjust for multiple hypothesis testing. Thus, while the maximum likelihood estimation is similar, the multiplicity control is different and there is not a way to leverage the tree structure to account for variable ways that clinical concept is coded. Disproportionality approaches were originally designed for spontaneous reporting data and then extended to make use of newly available longitudinal data. They use shrinkage estimators to informally control for multiplicity rather than through formal hypothesis testing, and also do not make use of the tree structure. Nelson et al.'s paper does not formally compare methods. Brown et al. compared Poisson tree-based scan statistics to disproportionality analyses and found reasonable concordance with the two approaches.[6] In other words, when the methods are applied to the same empirical dataset, both techniques showed similar detection capability although no formal power studies were performed for this comparison.

The tree-based scan statistic is hypothesis-generating, in that it produces an early warning with respect to potential associations. Because thousands of outcomes are evaluated simultaneously, confounding control is design-based using familiar epidemiologic techniques such as confounder adjustment of expected counts, restriction, stratification, or matching. As with any other data mining method, statistically significant "alerts" generated using the tree-based scan statistic must be carefully evaluated using other pharmacoepidemiologic methods where confounding control is more specifically tailored to the exposure-outcome pair of concern. In addition to generating statistically significant alerts, the method will also produce estimates of relative risk and attributable risk.

Moore et al. have expressed concern regarding the potential for missed safety signals in automated data.[21] Here, we demonstrate how to assess the

statistical power of the tree-based scan statistic allowing regulators to understand its statistical power for different sample sizes and outcome frequencies. Our work is part of a larger literature that studies the statistical power of other types of scan statistics.[22–30] These sample size calculations should be used in the same way that sample size calculations are used for traditional epidemiologic studies: to allow the investigator to decide whether to proceed with a study or to wait for more sample size to accrue based on the desired ability to detect particular effect sizes of interest.

Statistical power varies with the effect size, the sample size, and the frequency of the underlying outcome rate. We simulated data using a new user cohort design, which compared an exposed population to an unexposed population. We created known alternative hypotheses that generated clusters of excess risk in the tree structure. We then used the tree-based scan statistic to analyze these data.

Based on these preparatory-to-surveillance power simulations, regulators can properly frame the aforementioned mandatory safety data summaries at eighteen months postmarket, clearly spelling out what level of risk was detectable. Further, such simulations allow regulators to make key process decisions related to the timing of retrospective data-mining analyses.

Methods

Tree-Based Scan Statistics for Cohort Data

The tree-based scan statistic detects elevated frequencies of outcomes in electronic health data that have been grouped into hierarchical tree structures. In our case, the tree structure is derived from the Agency for Healthcare Research And Quality's Multi-Level Clinical Classifications Software (MLCCS) (http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp). The MLCCS groups outcomes into clinically meaningful categories and arranges them into four grouping levels. The broadest grouping identifies eighteen body systems and the narrowest grouping may contain multiple ICD-9-CM codes, forming a "branch." Each individual ICD-9-CM code is a "leaf." Any particular location on the tree – be it at the leaf or branch level – is referred to as a node. Table 1 shows an example branch.

We curated the full MLCCS tree by excluding ICD-9-CM outcome codes that 1) are unlikely to be caused by medical product exposures such as well care visits and pregnancy; 2) are unlikely to manifest within a few weeks after exposure, such as cancer; and 3) are common and of a less serious or unspecific nature, such as fever or diarrhea. Following the curation of the original thirteen thousand unique ICD-9-CM codes, we evaluated 6,162 ICD-9-CM codes which all represent individual leaves on the tree. Overall, there are 6,861 nodes on the tree. The curated tree is available upon request.

The null hypothesis being tested is that, for all nodes on the tree, an outcome is expected to occur in proportion to the underlying expected count that defined that node, as generated from a Poisson distribution. The alternative hypothesis is that one or more particular nodes on the tree have outcomes occurring with higher probability than the specified expected counts on those nodes.

A log-likelihood ratio was calculated for every node on the tree. The maximum among these log-likelihood ratios from the real data set is the test statistic for the entire analytic dataset. This maximum is compared with the maximum log-likelihood ratios that were calculated in the same way from simulated datasets generated under the null hypothesis. If the test statistic from the real dataset is among the 5 percent highest of all the maxima, the null hypothesis is rejected. The fact

Volume 5 (2017) Issue Number 1

that it is the maxima over the whole tree is what adjusts for the multiple testing. This hypothesis testing method allows one to detect whether any node on the tree had clusters of excess outcomes that were statistically significant while adjusting for multiple testing inherent to evaluating more than six thousand nodes.[31] Specific details of this procedure are included in the eAppendix.

Tree-based scan statistics can be used unconditionally or one can condition on the total number of observed outcomes in the dataset. Mathematical expressions for both versions can be found in the eAppendix. Conditioning is a mechanism to control for situations when there is an across-the-board increase in health care utilization during a particular time period that is unrelated to the exposure of interest. This situation might occur commonly in vaccine safety surveillance when the cohort has follow-up tests or visits in the days immediately following their well-care visit when a vaccine was administered. The conditional tree-based scan statistic attenuates this health care utilization unrelated to the exposure by standardizing all diagnoses by the frequency with which they appear in the dataset.

Simulated Datasets

To create the simulated datasets, we required background rates, and chose the exposure of interest to be quadrivalent human papillomavirus vaccine (Gardasil, Merck and Co. Inc.), identified by CPT code 90649. The choice of exposure is incidental to the power evaluations, but we chose this example to motivate how one might use these power evaluations in decision-making.

We extracted background rates for all the outcomes in the curated MLCCS tree from Florida Medicaid data using a cohort of 9-26 year olds from June 2006 to June 2009. All persons were minimally enrolled for 183 days in the health plan to ascertain chronic medical conditions and then began contributed time to the background rates. Contributed time was censored for any of the following criteria: 1) the last date of the study period, 2) disenrollment, 3) when the first incident outcome occurred with incidence criteria defined next, 4) or when a subsequent identical vaccination occurred. Vaccinated individuals only contributed unexposed time post-vaccination in days after the designated risk window. Never-vaccinated individuals were allowed to contribute time after the 183 day run-in period. Key metrics to describe the source data for

Table 2. Key Metrics of the Source Dataa used to Capture the Background Rates of Outcomes of Interest

KEY METRICS	
Total person-years followed	1,807,325
Total events	256,117
Total persons	24,369
Total exposed person-years	1,664
Total expected events	164.1
Total observed events in exposed time	379

[a]These data are based on 183-day lookback period, with an "exposed" risk window of 1-28 days following vaccination.

the background rates are listed in Table 2. These background rates are used to calculate age-adjusted expected counts that are used by the Poisson-based tree-based scan statistic for comparison with the simulated observed counts in the risk window.

Outcome events were defined by ICD-9-CM codes and visit location/setting. An incident outcome was defined as the chronologically first third-level MLCCS outcome observed in the inpatient or emergency department setting, which was not observed during the prior 183 days in either the emergency department, inpatient or outpatient setting. This means that, even if it was a never before seen ICD-9-CM code, it was not counted if a different ICD-9-CM code belonging to the same third level MLCCS group, i.e. the same branch, was observed during the prior 183 days. For example, as shown in Table 1, a febrile seizure (ICD-9-CM 780.31) and a complex febrile seizure (ICD-9-CM 780.32) are part of the same branch at the third-level node on the MLCCS tree (06.04.02). Therefore, in order for a 780.31 code to be incident, none of those branch-level outcomes could have occurred in the previous 183 days.

Alternative Hypotheses

To understand the statistical power to detect various effect sizes, we pre-defined effect sizes of interest ranging from 5 excess event per million doses to 500 excess events per million doses. We chose four different outcomes that have varying incidence rates and created known alternative hypotheses by injecting the risk at the leaf level (i.e., ICD-9-CM code) on the tree. The choices of outcomes also were incidental, but were required to be differing orders of magnitude in their base frequency in the dataset. We used Monte Carlo simulation to create multiple alternative datasets under both the null and known alternative hypotheses. The incidence rates and the known alternative hypotheses were inputs to stochastic Poisson processes. That is, these values allow us to calculate expected counts that serve as the parameter of interest for Poisson random draws. Using the maximum log-likelihood ratio as the test statistic, we computed the percentage of time an alert is raised when the type I error was set to 0.05. This output was the statistical power.

All analyses were performed using the power evaluation feature in the free TreeScan tool (www.treescan.org, v1.1.3), which calculates pure power of the analytic dataset. That is, when performing a power evaluation, we do not know which particular nodes give rise to the alert, only that an alert was generated. The probability of signaling on the particular node with the injected elevated risk is slightly lower than the pure power since there is an allowance for false positive alerts (i.e., 0.05). For actual analyses of real data (i.e., those that do not use the power evaluation feature), it always possible to determine which nodes individually alert.

We also explored the effect on statistical power of composite alternative hypotheses of risk. A composite alternative hypothesis is one in which the elevated risk is assigned to an outcome that is defined over a grouping of ICD-9-CM codes rather than being assigned to a singular ICD-9-CM code. Such a scenario is more likely to occur when multiple ICD-9-CM codes could be assigned for the same clinical concept. We used both optic neuritis and thrombocytopenia as examples. To illustrate, optic neuritis may be coded as 377.30 or 377.39, and in our source data, the latter was coded 10 times more frequently. When we created a simple injected risk scenario, then the risk was only elevated at one node on the tree, i.e., at the most frequently coded ICD-9-CM code. In contrast, when we created a complex injected risk scenario, then the risk was elevated at all nodes on the tree associated with the concept. Thrombocytopenia can be coded with eight different ICD-9-CM codes. We held effect sizes constant and performed the same statistical power analyses.

Volume 5 (2017) Issue Number 1

We also created artificial elevations in the occurrence of all outcomes uniformly throughout the tree on all nodes, representing an across-the-board increase in health care utilization during the risk window. We used these known alternative hypotheses to evaluate the conditional tree-based scan statistic that is designed to control for such utilization. For this comparison, we held effect sizes constant and compared the probability of rejecting the null hypothesis of the conditional and unconditional tree-based scan statistics.

Mis-Specification of the Risk Window

In the scenarios described above, the risk window was perfectly specified, meaning that the true risk window was coincident with the observed risk window. Data-mining does not involve pre-specification of hypotheses of interest, and therefore there is a universal risk window applied to the 6000+ outcomes. Consequently, we considered circumstances when the specified risk window is either too short or too long, and the consequent effects on statistical power. Appropriate risk window specification has been considered in detail elsewhere.[32]

First, we considered the circumstance when the true risk window was longer, but encompassed the observed risk window. For example, the true risk window could occur 1-28 days post-vaccination whereas the observed risk window could occur 1-14 days post-vaccination. That is, exposed outcomes in the 15-28 days following vaccination would be missed. Losses in sensitivity underestimate the true attributable risk but do not bias the true relative risk when assuming a Poisson likelihood, i.e. the risk is constant over the relevant time period. It has the same effect as reducing the overall sample size. That is, specifying a too-short risk window simply means that one needs more vaccinees to attain the same statistical power.

Then, we considered the circumstance when specifying a too-long risk window, i.e. when the true risk window was shorter and contained within the observed risk window. In these circumstances, the true relative risk is diluted or washed out, but the attributable risk remains unbiased. Therefore, in these scenarios, we calculated the observed effect size and created the known alternative hypotheses accordingly.

Result

Simple and Complex Injected Risk

Figure 1 shows the statistical power to detect various attributable risks. We vary the total sample size among four outcomes of interest with underlying event rates that vary by orders of magnitude. In the population of interest, syncope (ICD-9-CM 780.2) occurs most frequently at 95.6 events per 1 million doses whereas optic neuritis occurs least frequently at 0.30 events per 1 million doses.

When using a fixed risk difference measure, it is more difficult to detect the identical risk difference in a more frequently occurring event because it takes many such events to provide adequate separation of the treatment and comparator group population. To illustrate, five excess events in the treatment group amounts to statistical noise in a commonly occurring outcome such as a headache. With rare events, separation between the two groups is observable even with few events, thereby generating higher statistical power to rule out the same attributable risk. For example, five excess cases are quite meaningful for some autoimmune diseases that are only expected to occur once in a million exposed. As expected, it is easier to detect the same risk differences with larger sample sizes.

In administrative data, clinical concepts may be coded uniquely (i.e., a singular ICD-9-CM code), or as a collection of codes. In Figure 1, we included

Figure 1. Statistical Power to Detect Various Attributable Risks

Incidence Rate Difference of Interest (Events per million doses)									
Expected Counts	Vaccinees	0	5	10	20	50	100	200	500
Syncope (ICD-9-CM 780.2)									
9.6	0.1M	0.05	0.05	0.05	0.05	0.06	0.18	0.83	1.00
19.1	0.2M	0.05	0.05	0.05	0.05	0.08	0.50	1.00	1.00
47.8	0.5M	0.05	0.05	0.05	0.06	0.29	0.98	1.00	1.00
95.6	1.0M	0.05	0.05	0.05	0.07	0.74	1.00	1.00	1.00
191.2	2.0M	0.05	0.05	0.06	0.15	0.99	1.00	1.00	1.00
478.0	5.0M	0.05	0.05	0.09	0.64	1.00	1.00	1.00	1.00
Thrombocytopenia (ICD-9-CM 287.1, 287.3, 287.30, 287.31, 287.33, 287.39, 287.4, 287.5)									
9.5	0.5M	0.05	0.05	0.06	0.18	0.96	1.00	1.00	1.00
Thrombocytopenia (ICD-9-CM 287.5)									
1.4	0.1M	0.05	0.05	0.05	0.06	0.25	0.82	1.00	1.00
2.9	0.2M	0.05	0.05	0.06	0.09	0.58	0.99	1.00	1.00
7.2	0.5M	0.05	0.05	0.07	0.26	0.99	1.00	1.00	1.00
14.3	1.0M	0.05	0.06	0.11	0.65	1.00	1.00	1.00	1.00
28.7	2.0M	0.05	0.07	0.30	0.97	1.00	1.00	1.00	1.00
71.7	5.0M	0.05	0.16	0.88	1.00	1.00	1.00	1.00	1.00
Systemic Lupus Erythematous (ICD-9-CM 710.0)									
0.2	0.1M	0.05	0.05	0.06	0.13	0.62	0.98	1.00	1.00
0.4	0.2M	0.05	0.06	0.09	0.33	0.95	1.00	1.00	1.00
1.0	0.5M	0.05	0.08	0.30	0.87	1.00	1.00	1.00	1.00
2.0	1.0M	0.05	0.15	0.67	1.00	1.00	1.00	1.00	1.00
4.0	2.0M	0.05	0.46	0.98	1.00	1.00	1.00	1.00	1.00
9.9	5.0M	0.05	0.95	1.00	1.00	1.00	1.00	1.00	1.00
Optic Neuritis (ICD-9-CM 377.39 and 377.30)									
0.16	0.5M	0.05	0.30	0.75	0.99	1.00	1.00	1.00	1.00
Optic Neuritis (ICD-9-CM 377.39)									
0.03	0.1M	0.05	0.07	0.14	0.39	0.90	1.00	1.00	1.00
0.06	0.2M	0.05	0.14	0.39	0.80	1.00	1.00	1.00	1.00
0.15	0.5M	0.05	0.32	0.77	0.99	1.00	1.00	1.00	1.00
0.30	1.0M	0.05	0.63	0.98	1.00	1.00	1.00	1.00	1.00
0.59	2.0M	0.05	0.96	1.00	1.00	1.00	1.00	1.00	1.00
1.48	5.0M	0.05	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: This figure accounts for different background event rates, sample sizes, and coding algorithms (i.e., singular or multiple ICD-9-CM codes of interest). All simulations were performed with 99,999 iterations under the null hypothesis that observed counts for all nodes on the tree were expected to occur proportionately to the underlying expected counts; and with 10,000 iterations under the known alternative hypothesis using the unconditional tree-based scan statistic. Critical values were set at a signaling threshold of p=0.05.

Volume 5 (2017) Issue Number 1

examples of injected risk when the risk is spread among a collection of related ICD-9-CM codes (e.g., thrombocytopenia) to observe differences between singular injections of risk. The expansion of the outcome definition to encompass several codes increases the expected counts for the clinical concept. The higher expected counts are equivalent to testing a different outcome with a higher underlying frequency. In other words, when we hold the attributable risk constant, the total expected counts - regardless of whether it was derived using a singular ICD-9-CM code or a collection of codes - corresponds to the statistical power.

Adjusting for Across-the-Board Elevations in Health Care Utilization

Table 3 demonstrates the ability of the conditional versus unconditional tree-based scan statistic to properly control for across-the-board elevations in health care utilization that happen to occur in the risk window but are unrelated to the exposure. We compare actual type I error observed to allowable type I error (i.e., 0.05). The unconditional tree-based scan statistic inflates type I error when general utilization is increased by as little as 3 percent. Utilization increases of this magnitude are not unusual in administrative data and have been observed by the authors in other analyses as well as in the source data as seen in Table 2. However, the conditional tree-based scan statistic continues to hold type I error to the allowable level even when across-the-board health care utilization increases by 500 percent.

Figure 2 is a comparison of the statistical power of the conditional versus unconditional tree-based scan statistic in the absence of general health care utilization increases in the risk window. We compare small sample sizes to illustrate situations when the conditional tree-based scan statistic performs less well than the unconditional tree-based scan statistic. Under these circumstances, the conditional tree-based scan statistic detects the attributable risk less often than the unconditional statistic because the small sample size intensifies the "attenuation effect" that occurs when conditioning on the total number of cases. For example, observe the situation of a 1000 vaccine sample size with a true attributable risk of 5 excess cases per 1000 vaccinees (5000 events per million doses in Figure 2). In this dataset, the caseload is almost entirely due to the risk (5 observed cases that are all excess cases of syncope). Therefore, the conditional tree-based scan statistic mistakes part of the "signal" for noise and subsequently has reduced statistical power to detect it properly. Compare the 89 percent statistical power in the unconditional analysis to the 79 percent power in the conditional. Particular situations that warrant concern include the very early uptake period or low exposure prevalence medical products.

Figure 3 is a comparison of the statistical power of the conditional tree-based scan statistic's

Table 3. Type I Error in the Conditional and Unconditional Tree-Based Scan Statistic under Conditions of Across-the-board Elevations in Health care Utilizationa

	GENERAL INCREASES IN HEALTH CARE UTILIZATION APPLIED TO DATASET										
	0%	1%	2%	3%	5%	8%	10%	20%	50%	200%	500%
Unconditional	0.05	0.06	0.06	0.08	0.24	0.82	0.98	1.00	1.00	1.00	1.00
Conditional	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

[a]All simulations were performed with 99,999 iterations under the null hypothesis that observed counts for all nodes on the tree were expected to occur proportionately to the underlying expected counts with a sample size of 500,000 vaccinees. Allowable type I error set to 0.05.

Figure 2. Statistical Power to Detect Various Attributable Risks with Various Sample Sizes Using Both the Unconditional Versus Conditional Tree-Based Scan Statistic in the Absence of Overall Increases in Health Care Utilization

Incidence Rate Difference of Interest (Events per million doses)									
		0	50	100	200	500	1000	2000	5000
Vaccinees	Syncope (ICD-9-CM 780.2)								
100	Unconditional	0.05	0.05	0.05	0.05	0.05	0.06	0.07	0.16
	Conditional	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.12
200	Unconditional	0.05	0.05	0.05	0.05	0.05	0.07	0.11	0.32
	Conditional	0.05	0.05	0.05	0.05	0.06	0.07	0.11	0.29
500	Unconditional	0.05	0.05	0.05	0.05	0.05	0.07	0.14	0.50
	Conditional	0.05	0.05	0.05	0.05	0.05	0.06	0.10	0.36
1000	Unconditional	0.05	0.05	0.05	0.05	0.07	0.15	0.39	0.89
	Conditional	0.05	0.05	0.05	0.05	0.06	0.12	0.31	0.79
2000	Unconditional	0.05	0.05	0.05	0.06	0.09	0.23	0.63	0.99
	Conditional	0.05	0.05	0.05	0.05	0.08	0.18	0.53	0.97
5000	Unconditional	0.05	0.05	0.05	0.06	0.13	0.50	0.95	1.00
	Conditional	0.05	0.05	0.05	0.06	0.19	0.58	0.96	1.00
10000	Unconditional	0.05	0.05	0.05	0.08	0.42	0.92	1.00	1.00
	Conditional	0.05	0.05	0.05	0.07	0.35	0.88	1.00	1.00
20000	Unconditional	0.05	0.05	0.06	0.13	0.76	1.00	1.00	1.00
	Conditional	0.05	0.05	0.05	0.11	0.70	1.00	1.00	1.00
50000	Unconditional	0.05	0.05	0.07	0.35	1.00	1.00	1.00	1.00
	Conditional	0.05	0.05	0.08	0.37	1.00	1.00	1.00	1.00

Note: All simulations were performed with 99,999 iterations under the null hypothesis that observed counts for all nodes on the tree were expected to occur proportionately to the underlying expected counts; and with 10,000 iterations under the known alternative hypothesis. Critical values were set at a signaling threshold of p=0.05.

performance with varying levels of attributable risk and varying levels of general increased utilization unrelated to the health outcome of interest. As before, when using fixed attributable risks, it is easier to detect signals against very rare background rates. However, the statistical power of the conditional tree-based scan statistic is lower for the same fixed risk difference because of the attenuation effect due. For example, the statistical power of the conditional tree-based scan statistic to detect an excess risk of 100 excess events per million doses in an event that occurs with the frequency of syncope in a 500,000 vaccinee population is 97 percent when there is no background elevation in overall health care utilization. If overall health care utilization is 50 percent higher for reasons unrelated to the exposure, the statistical power to detect the same risk difference drops to 59 percent.

Volume 5 (2017) Issue Number 1

Figure 3. Statistical Power to Detect Various Attributable Risks, Accounting for Different Background Event Rates and Different Levels of Overall Increases in Health Care Utilization

Incidence Rate Difference of Interest (Events per million doses)								
Increase	0	2	5	10	20	50	100	200
Syncope (ICD-9-CM 780.2)								
0%	0.05	0.05	0.05	0.05	0.05	0.27	0.97	1.00
1%	0.05	0.05	0.05	0.05	0.05	0.26	0.97	1.00
2%	0.05	0.05	0.05	0.05	0.05	0.26	0.97	1.00
5%	0.05	0.05	0.05	0.05	0.05	0.23	0.95	1.00
10%	0.05	0.05	0.05	0.05	0.05	0.20	0.93	1.00
20%	0.05	0.05	0.05	0.05	0.05	0.15	0.86	1.00
50%	0.05	0.05	0.05	0.05	0.05	0.08	0.59	1.00
200%	0.05	0.05	0.05	0.05	0.05	0.06	0.27	0.97
500%	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.13
Thrombocytopenia (ICD-9-CM 287.5)								
0%	0.05	0.05	0.05	0.06	0.22	0.98	1.00	1.00
1%	0.05	0.05	0.05	0.06	0.22	0.98	1.00	1.00
2%	0.05	0.05	0.05	0.06	0.21	0.97	1.00	1.00
5%	0.05	0.05	0.05	0.06	0.19	0.96	1.00	1.00
10%	0.05	0.05	0.05	0.06	0.17	0.95	1.00	1.00
20%	0.05	0.05	0.05	0.05	0.14	0.90	1.00	1.00
50%	0.05	0.05	0.05	0.05	0.09	0.70	1.00	1.00
200%	0.05	0.05	0.05	0.05	0.06	0.40	0.98	1.00
500%	0.05	0.05	0.05	0.05	0.05	0.06	0.22	0.88
Systemic Lupus Erythematous (ICD-9-CM 710.0)								
0%	0.05	0.05	0.07	0.27	0.84	1.00	1.00	1.00
1%	0.05	0.05	0.07	0.27	0.83	1.00	1.00	1.00
2%	0.05	0.05	0.07	0.26	0.82	1.00	1.00	1.00
5%	0.05	0.05	0.07	0.24	0.80	1.00	1.00	1.00
10%	0.05	0.05	0.07	0.22	0.77	1.00	1.00	1.00
20%	0.05	0.05	0.06	0.18	0.69	1.00	1.00	1.00
50%	0.05	0.05	0.05	0.11	0.49	1.00	1.00	1.00
200%	0.05	0.05	0.05	0.07	0.27	0.95	1.00	1.00
500%	0.05	0.05	0.05	0.05	0.06	0.27	0.84	1.00
Optic Neuritis (ICD-9-CM 377.39)								
0%	0.05	0.07	0.29	0.75	0.99	1.00	1.00	1.00
1%	0.05	0.07	0.28	0.74	0.99	1.00	1.00	1.00
2%	0.05	0.07	0.28	0.73	0.99	1.00	1.00	1.00
5%	0.05	0.07	0.27	0.71	0.99	1.00	1.00	1.00
10%	0.05	0.07	0.25	0.67	0.98	1.00	1.00	1.00
20%	0.05	0.06	0.22	0.62	0.97	1.00	1.00	1.00
50%	0.05	0.06	0.15	0.46	0.90	1.00	1.00	1.00
200%	0.05	0.05	0.09	0.29	0.75	1.00	1.00	1.00
500%	0.05	0.05	0.05	0.07	0.20	0.75	0.99	1.00

Note: All simulations were performed assuming a sample size of 500,000 vaccinees with 99,999 iterations under the null hypothesis that observed counts for all nodes on the tree were expected to occur proportionately to the underlying expected counts; and with 10,000 iterations under the known alternative hypothesis using a conditional tree-based scan statistic. Critical values were set at a signaling threshold of p=0.05.

Risk Window Mis-Specification

Figure 4 demonstrates the effect on statistical power when a too-long risk window has been specified. The ratio defined in the figure represents the ratio of the too-long observed risk window to the true risk window contained within it. The losses in statistical power occur because of the "washing out" of the signal. For example, compare the circumstance when the specified window is twice as long as the true risk interval (i.e., the ratio is 2) and we are interested in the statistical power to detect 100 events per million doses for an outcome that occurs with the frequency of syncope. In Figure 1, the statistical power is 98 percent as compared to 29 percent in Figure 4. The too-short risk window does not result in loss of statistical power as a consequence of bias. Rather, a too-short risk window is equivalent to having a smaller sample size. Therefore, one can refer to Figure 1 and treat a loss of risk window days as a loss of vaccinees.

Sensitivity Analyses: Pruning the Tree

Alerts at the most aggregated nodes on the tree are typically not actionable because they are so general. For example, an alert raised for quadrivalent human papillomavirus vaccine and "diseases of the circulatory system" is unlikely to be useful information. However, hypothesis testing is performed at these nodes. We tested whether "pruning the tree" to eliminate hypothesis testing at the top two levels of aggregated nodes would result in an increase in statistical power. The results were unaffected by this pruning because of the relatively small number of nodes there, i.e. 18 nodes at the root level as compared to 6000+ nodes at the leaf level.

Discussion

We performed numerous simulations to examine the statistical power of both the unconditional and conditional Poisson tree-based scan statistic for cohort-type data. In studies with small sample sizes, the unconditional tree-based scan statistic had slightly higher statistical power to detect attributable risk than the conditional tree-based scan statistic. However, the unconditional tree-based scan statistic inflated type I error even in the presence of low general increases in health care utilization following exposure. The conditional tree-based scan statistic controlled type I error well when faced with general increases in health care utilization following exposure but experienced slightly decreased power as a consequence of the increasing noise. We observed reductions in statistical power resulting from specifying a too-long risk window, and reductions in sample size from specifying a too-short risk window.

To give our statistical power study context, we considered an example problem of quadrivalent human papillomary virus vaccine, which is administered to 9-26 year olds. We further developed background rates based on their "unexposed time" when we considered exposed time to occur in the first 28 days following vaccination. These background rates were used to compute expected counts for various sample sizes. The statistical power concepts and trends demonstrated with this example should apply to all problems regardless of the source data or the particular tree being utilized. We focus here on demonstrating the process to perform statistical power calculations using the power evaluation feature within the TreeScan software.

We also use these tables to prepare for future vaccine safety monitoring (e.g., nine-valent human papillomary virus vaccine) in a population that is represented by the source data and by using the same tree.[8] First, we estimate the number of vaccinated individuals expected at eighteen months. If the expected sample size of vaccinated individuals is small (as it is in Figure 2) and the overall health care utilization following exposure is not expected to be elevated, then an unconditional analysis is

Volume 5 (2017) Issue Number 1

Figure 4. Statistical Power to Detect Various Attributable Risks When Mis-Specifying the Risk Window

Incidence Rate Difference of Interest (Events per million doses)								
	0	5	10	20	50	100	200	500
Ratio	Syncope (ICD-9-CM 780.2)							
2.0	0.05	0.05	0.05	0.05	0.06	0.29	0.98	1.00
1.7	0.05	0.05	0.05	0.05	0.08	0.49	1.00	1.00
1.4	0.05	0.05	0.05	0.05	0.11	0.70	1.00	1.00
1.3	0.05	0.05	0.05	0.05	0.15	0.85	1.00	1.00
1.1	0.05	0.05	0.05	0.06	0.21	0.94	1.00	1.00
1.0	0.05	0.05	0.05	0.06	0.29	0.98	1.00	1.00
Thrombocytopenia (ICD-9-CM 287.5)								
2.0	0.05	0.05	0.05	0.07	0.45	0.99	1.00	1.00
1.7	0.05	0.05	0.06	0.08	0.66	1.00	1.00	1.00
1.4	0.05	0.05	0.06	0.11	0.81	1.00	1.00	1.00
1.3	0.05	0.05	0.06	0.15	0.91	1.00	1.00	1.00
1.1	0.05	0.05	0.06	0.19	0.96	1.00	1.00	1.00
1.0	0.05	0.05	0.07	0.26	0.99	1.00	1.00	1.00
Systemic Lupus Erythematous (ICD-9-CM 710.0)								
2.0	0.05	0.05	0.08	0.30	0.96	1.00	1.00	1.00
1.7	0.05	0.06	0.10	0.44	0.99	1.00	1.00	1.00
1.4	0.05	0.06	0.14	0.57	1.00	1.00	1.00	1.00
1.3	0.05	0.06	0.18	0.70	1.00	1.00	1.00	1.00
1.1	0.05	0.07	0.24	0.80	1.00	1.00	1.00	1.00
1.0	0.05	0.08	0.30	0.87	1.00	1.00	1.00	1.00
Optic Neuritis (ICD-9-CM 377.39)								
2.0	0.05	0.11	0.32	0.77	1.00	1.00	1.00	1.00
1.7	0.05	0.14	0.42	0.87	1.00	1.00	1.00	1.00
1.4	0.05	0.18	0.53	0.93	1.00	1.00	1.00	1.00
1.3	0.05	0.22	0.62	0.97	1.00	1.00	1.00	1.00
1.1	0.05	0.27	0.71	0.98	1.00	1.00	1.00	1.00
1.0	0.05	0.32	0.77	0.99	1.00	1.00	1.00	1.00

Note: Ratio is the length of the observed/assumed risk window to the length of the true risk window. All simulations were performed assuming a sample size of 500,000 vaccinees with 99,999 iterations under the null hypothesis that observed counts for all nodes on the tree were expected to occur proportionately to the underlying expected counts; and with 10,000 iterations under the known alternative hypothesis using an unconditional tree-based scan statistic. Critical values were set at a signaling threshold of p=0.05.

preferred to a conditional analysis. However, if the expected number of vaccinated individuals is larger, or if overall health care utilization is expected to be elevated in the "designated risk window" for reasons unrelated to the exposure, then the conditional tree-based scan statistic is preferred because it minimizes false positive alerting. One could estimate this increased level of utilization in the source data. For example, in Table 2, comparing the total observed events in exposed time (i.e., 379 events) to the total expected events (i.e., 164.1 events) yields a 2.2x elevation in overall health care utilization. These source data have greater overall health care utilization in the time period immediately following vaccination, which is expected due to follow-up visits that occur closely after well-visits for reasons unrelated to vaccination. Then, an investigator could use Figure 3 to get a sense of the statistical power available for various underlying event frequencies.

Our preparatory-to-surveillance simulation demonstrates what magnitudes of risk can be ruled out or detected based on expected sample size at the time of performance of a TreeScan analysis. Regulators can use these simulations to contextualize what type of safety information can reasonably be available at the congressionally mandated eighteen month/10000 user postlicensure review. Further, if multiple TreeScan analyses are likely to be performed over the course of a medical product's lifetime, these simulations can be used to optimize analyses and limit potential reuse of observational data.[33]

Data mining analyses using tree-based scan statistics expand the safety net of pharmacovigilance, ensuring adequate monitoring of thousands of outcomes of interest while controlling for multiple hypothesis testing. They are an important complement to the existing armamentarium of knowledge generation about the effects of medical products, and we have shown how to estimate statistical power for such analyses.

Acknowledgements

We gratefully acknowledge comments received from Katherine Yih and Jeffrey S. Brown on this project and manuscript as well as the project management efforts of Carolyn Balsbaugh.

References

1. Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH, Hennessy S, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. PharmacoepidemiolDrug Saf. 2012;21 Suppl 1:1–8.
2. Dal Pan GJ, Lindquist M, Gelperin K. Postmarketing Spontaneous Pharmacovigilance Reporting Systems. In: Strom BL, Kimmel SE, Hennessy S, editors. Pharmacoepidemiology. Fifth. John Wiley & Sons; 2011. p. 137–57.
3. Food and Drug Administration Amendments Act of 2007, Public Law 110-85. 110–85 2007 p. 823–978.
4. Kulldorff M, Fang Z, Walsh SJ. A tree-based scan statistic for database disease surveillance. Biometrics. 2003 Jun;59(2):323–31.
5. Kulldorff M, Dashevsky I, Avery TR, Chan AK, Davis RL, Graham D, et al. Drug safety data mining with a tree-based scan statistic. Pharmacoepidemiol Drug Saf. 2013 May;22(5):517–23.
6. Brown JS, Petronis KR, Bate A, Zhang F, Dashevsky I, Kulldorff M, et al. Drug Adverse Event Detection in Health Plan Data Using the Gamma Poisson Shrinker and Comparison to the Tree-based Scan Statistic. Pharmaceutics. 2013;5(1):179–200.
7. Yih WK, Maro JC, Nguyen MD, Baker MA, Balsbaugh C, Cole DV, et al. Pilot of Self-Controlled Tree-Temporal Scan Analysis for Gardasil Vaccine [Internet]. Silver Spring, MD: U.S. Food and Drug Administration; 2016 Sep [cited 2016 Oct 24]. Available from: https://www.sentinelsystem.org/sites/default/files/Methods/Mini-Sentinel_PRISM_Pilot-Self-Controlled-Tree-Temporal-Scan-Analysis-Gardasil-Vaccine-Report.pdf
8. Yih WK, Maro JC, Dashevsky I, Anderson S, Baker MA, Mba-Jonas A, et al. Evaluation of HPV9 (Gardasil9) Vaccine Safety Surveillance Using the TreeScan Data Mining Method Surveillance Protocol [Internet]. Silver Spring, MD: U.S. Food and Drug Administration; 2016 Jun [cited 2016 Nov 7]. Available from: https://www.sentinelsystem.org/vaccines-blood-biologics/assessments/evaluation-hpv9-gardasil9-vaccine-safety-surveillance-using
9. Baggs J, Gee J, Lewis E, Fowler G, Benson P, Lieu T, et al. The Vaccine Safety Datalink: a model for monitoring immunization safety. Pediatrics. 2011;127 Suppl 1:S45-53.
10. Remarks made during the Post-Licensure Rapid Immunization Safety Monitoring (PRISM) Public Workshop. Bethesda, MD. December 7, 2016 [cited 2017 Mar 22]. Available from: https://www.fda.gov/downloads/BiologicsBloodVaccines/NewsEvents/WorkshopsMeetingsConferences/UCM544856.pdf
11. Nelson JC, Shortreed SM, Yu O, Peterson D, Baxter R, Fireman B, et al. Integrating database knowledge and epidemiological design to improve the implementation of data mining methods that evaluate vaccine safety in large healthcare databases. Statistical Analy Data Mining. 2014 Oct 1;7(5):337–51.

Volume 5 (2017) Issue Number 1

12. Chao C, Klein NP, Velicer CM, Sy LS, Slezak JM, Takhar H, et al. Surveillance of autoimmune conditions following routine use of quadrivalent human papillomavirus vaccine. J Intern Med. 2012 Feb;271(2):193–203.
13. Klein NP, Hansen J, Chao C, Velicer C, Emery M, Slezak J, et al. Safety of quadrivalent human papillomavirus vaccine administered routinely to females. Arch Pediatr Adolesc Med. 2012 Dec;166(12):1140–8.
14. DuMouchel W, Ryan PB, Schuemie MJ, Madigan D. Evaluation of disproportionality safety signaling applied to healthcare databases. Drug Saf. 2013 Oct;36 Suppl 1:S123-132.
15. Zorych I, Madigan D, Ryan P, Bate A. Disproportionality methods for pharmacovigilance in longitudinal observational databases. Stat Methods Med Res. 2013 Feb;22(1):39–56.
16. Schuemie MJ. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. Pharmacoepidemiol Drug Saf. 2011 Mar;20(3):292–9.
17. Norén GN, Hopstadius J, Bate A, Edwards IR. Safety surveillance of longitudinal databases: methodological considerations. Pharmacoepidemiol Drug Saf. 2011 Jul;20(7):714–7.
18. Svanström H, Callréus T, Hviid A. Temporal data mining for adverse events following immunization in nationwide Danish healthcare databases. Drug Saf. 2010 Nov 1;33(11):1015–25.
19. Walker AM. Signal detection for vaccine side effects that have not been specified in advance. PharmacoepidemiolDrug Saf. 2010;19(3):311–7.
20. Curtis JR, Cheng H, Delzell E, Fram D, Kilgore M, Saag K, et al. Adaptation of Bayesian data mining algorithms to longitudinal claims data: coxib safety as an example. Med Care. 2008 Sep;46(9):969–75.
21. Moore TJ, Furberg CD. Electronic Health Data for Postmarket Surveillance: A Vision Not Realized. Drug Saf. 2015 Jul;38(7):601–10.
22. Sahu SK, Bendel RB, Sison CP. Effect of relative risk and cluster configuration on the power of the one-dimensional scan statistic. Statist Med. 1993;12(19–20):1853–65.
23. Jung I, Lee H. Spatial cluster detection for ordinal outcome data. Stat Med. 2012 Dec 20;31(29):4040–8.
24. Neill DB. An empirical comparison of spatial scan statistics for outbreak detection. Int J Health Geogr. 2009;8:20.
25. Huang L, Pickle LW, Das B. Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases. Stat Med. 2008 Nov 10;27(25):5111–42.
26. Huang L, Kulldorff M, Gregorio D. A spatial scan statistic for survival data. Biometrics. 2007 Mar;63(1):109–18.
27. Kulldorff M, Zhang Z, Hartman J, Heffernan R, Huang L, Mostashari F. Benchmark data and power calculations for evaluating disease outbreak detection methods. MMWR Morb Mortal Wkly Rep. 2004 Sep 24;53 Suppl:144–51.
28. Kulldorff M, Mostashari F, Duczmal L, Katherine Yih W, Kleinman K, Platt R. Multivariate scan statistics for disease surveillance. Stat Med. 2007 Apr 15;26(8):1824–33.
29. Waller LA, Hill EG, Rudd RA. The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. Stat Med. 2006 Mar 15;25(5):853–65.
30. Song C, Kulldorff M. Power evaluation of disease clustering tests. Int J Health Geogr. 2003 Dec 19;2(1):9.
31. Dwass M. Modified Randomization Tests for Nonparametric Hypotheses. Ann Math Statist. 1957 Mar;28(1):181–7.
32. Rowhani-Rahbar A, Klein NP, Dekker CL, Edwards KM, Marchant CD, Vellozzi C, et al. Biologically plausible and evidence-based risk intervals in immunization safety research. Vaccine. 2012 Dec 17;31(1):271–7.
33. Toh S, Avorn J, D'Agostino RB, Gurwitz JH, Psaty BM, Rothman KJ, et al. Re-using Mini-Sentinel data following rapid assessments of potential safety signals via modular analytic programs. Pharmacoepidemiol Drug Saf. 2013 Oct;22(10):1036–45.

APPENDIX

Unconditional Tree-based Scan Statistic for Cohort Data

All outcomes are first classified into a hierarchical tree structure described in the main paper. For each leaf i of the tree (i.e., finest granularity) which represents a unique outcome of interest, we note the observed number c_i of outcomes in the risk window and the expected number n_i of outcomes based on the background rate and sample size.

The next step is to define nodes on the tree. Each node G defines either an outcome (if at the leaf level) or a cluster of related outcomes (if at the branch level). The sums of the observed and expected number of outcomes in this node are denoted as c_G and n_G respectively. Again, note that a single leaf is one potential node, but a node could also be a branch of the tree.

The log likelihood ratio is derived from a Poisson-based maximum likelihood estimator and is:

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

where:

$I()$ is the indication function, which is 1 when there are more observed outcomes than would be expected by chance. It is included to ensure that we are looking for an excess risk of the having the adverse event rather than a protective decreased risk.

Log likelihood ratios are computed for computational convenience, and results from them are identical to results based on likelihood ratios). The order in which the nodes are evaluated does not impact the results. The node G with the maximum LLR is the most likely cluster of unexplained outcomes in the risk window and its log likelihood ratio is the test statistic:

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

The distribution of T is not known analytically, so inference is conducted using Monte Carlo hypothesis testing (Dwass, 1957). First, a user-defined number of random data sets (e.g., 99,999) are generated under the null hypothesis that the observed number of outcomes in the risk window should be proportional to the expected number of outcomes for that same period. T is calculated for the 99,999 random data sets and the 1 real data set.

If the T in the real data is among the 5% highest of all the maxima from the real and 99,999 random data sets generated under the null hypothesis, then that node constitutes a signal at the $alpha=0.05$ statistical significance level. The Monte Carlo based p-value is calculated as $p=R/(99999+1)$, where R is the rank of the T in the real data set in relation to the T in the random data sets. That way the method formally adjusts the p-values for the multiple testing generated by the many overlapping groupings of outcomes. This means that, when the null hypothesis is true, there is a 95% probability that all p-values are greater than 0.05, or in other words, that there is not a single exposure-outcome pair or grouping with $p \leq 0.05$.

Volume 5 (2017) Issue Number 1

Conditional Tree-based Scan Statistic for Cohort Data

When using the unconditional tree-based scan statistic described above, the null hypothesis is that any outcome is equally likely to occur in proportion to underlying background rate of the event as given by the expected counts. In the conditional version, the null hypothesis is based on the relative magnitude of the expected counts rather than the expected counts themselves, and the analysis is conditioned on the total number of outcomes in the whole tree. Thus, the statistical model is a multinomial distribution. Full derivation of the equations is in the paper by Kulldorff, 2003.

Thus, we calculate the total number of outcomes in the risk window $C = \sum_i c_i$ and the total number of expected outcomes $N = \sum_i n_i$, summed over all the leaves on the tree.

$$LLR(G) = \left[c_G \ln\left(\frac{c_G}{n_G}\right) + (C - c_G)\ln\left(\frac{C - c_G}{N - n_G}\right)\right] I\left(\frac{c_G}{n_G} > \frac{C - c_G}{N - n_G}\right)$$

$I()$ is the indication function, which is 1 when there are more observed outcomes than would be expected by chance. It is included to ensure that we are looking for an excess risk of the having the adverse event rather than a protective decreased risk.

Again, log likelihood ratios are used for computational convenience as opposed to likelihood ratios. The order in which the nodes are evaluated does not impact the results. The node G with the maximum LLR is the most likely cluster of unexplained outcomes in the risk window and its log likelihood ratio is the test statistic:

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

The other difference occurs in the Monte Carlo simulation step. Now, every random data set has to have the same C and N, so that the total number of observed outcomes and the total number of expected outcomes are the same in both the real and all the random data sets. The rest of the procedure is the same as described above.

References:

Dwass M. Modified Randomization Tests for Nonparametric Hypotheses. Ann Math Statist. 1957 Mar;28(1):181–7.
Kulldorff M, Fang Z, Walsh SJ. A tree-based scan statistic for database disease surveillance. Biometrics. 2003 Jun;59(2):323–31.