SENTINEL METHODS REPORT

POSTLICENSURE MEDICAL PRODUCT SAFETY DATA-MINING: POWER CALCULATIONS FOR BERNOULLI DATA

Prepared by: Judith C. Maro, PhD, MS,[1] Inna Dashevsky, MS,[1] Martin Kulldorff, PhD,[2]

Author Affiliations: 1. Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, 2. Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA,

December 22, 2017

The Sentinel System is sponsored by the U.S. Food and Drug Administration (FDA) to proactively monitor the safety of FDA-regulated medical products and complements other existing FDA safety surveillance capabilities. The Sentinel System is one piece of FDA's Sentinel Initiative, a long-term, multi-faceted effort to develop a national electronic system. Sentinel Collaborators include Data and Academic Partners that provide access to healthcare data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I. This project was funded by the FDA through HHS Mini-Sentinel contract number HHSF223200910006I.

Sentinel Methods Report

Postlicensure Medical Product Safety Data-Mining: Power Calculations For Bernoulli Data

Table of Contents

I. INTRODUCTION

In observational electronic healthcare databases, data-mining is a technique for simultaneous monitoring of many exposure-outcome pairs. Data-mining analyses have traditionally been performed using spontaneous reporting databases, which lack denominator data and are subject to persistent underreporting.[1] However, the longitudinal nature of administrative claims data, such as in the U.S. Food and Drug Administration's (FDA's) Sentinel System[2], enables systematic evaluation of thousands of outcomes, ensuring that rate and count data are collected and analyzed routinely.

Here, we focus on one data-mining method that leverages these longitudinal data: the tree-based scan statistic as operationalized in TreeScan™ freeware (http://www.treescan.org).[3] This method has previously been used in postmarket medical product safety settings[4–6], and is planned to monitor nine-valent human papillomavirus vaccine exposure.[7] Like most data-mining methods, the tree-based scan statistic is hypothesis-generating, in that it produces an early warning with respect to potential associations. 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.

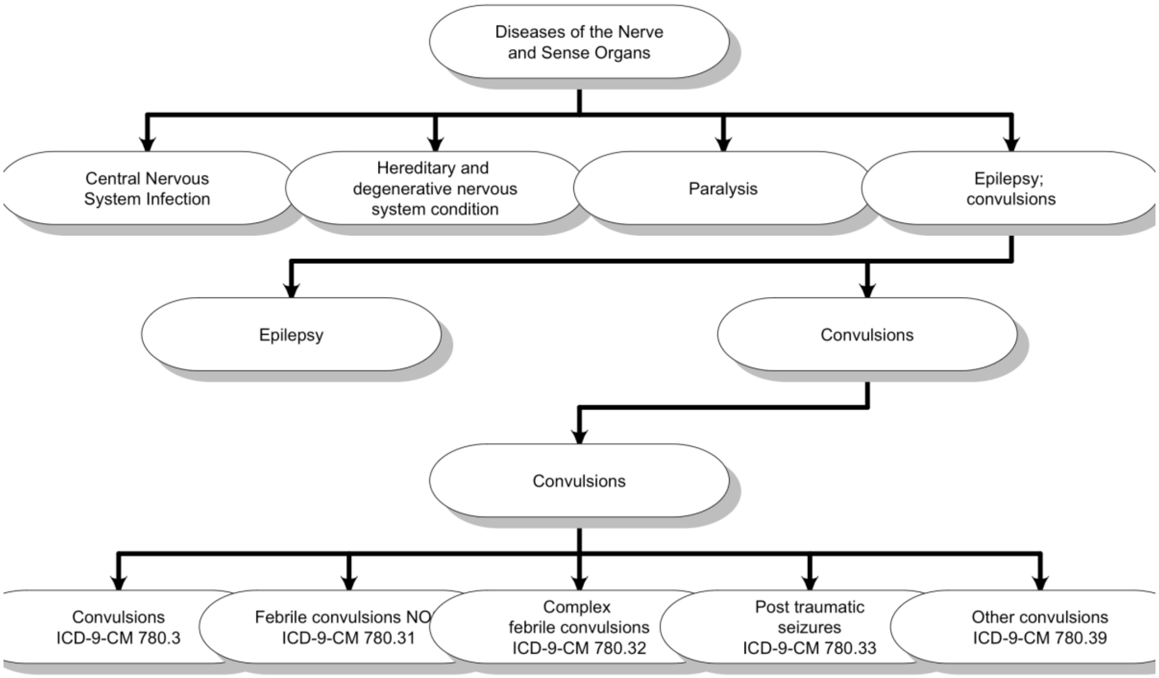
The steps in using TreeScan™ are as follows: first, analytic datasets containing rate or count data for many outcomes are assembled using familiar epidemiologic designs that control for confounding including restriction, stratification, or matching. These design-based confounding strategies (as opposed to analysis-based confounding strategies) are necessary in TreeScan™ because of the many outcomes being evaluated. 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 such as convulsions. **Figure 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. The null hypothesis is that there is no elevated risk for any one of these thousands of outcomes.

Moore *et al.* have expressed concern regarding the potential for missed safety signals in automated data.[8] Here, we demonstrate the procedure to assess the statistical power of the tree-based scan statistic when analytic data is structured according to a Bernoulli distribution. Designs that can be used to create such data include self-controlled designs[9] and fixed-ratio matched designs.[10]

Our work is part of a larger literature that studies the statistical power of other types of scan statistics.[11–20] In general, 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 of patients, and the frequency of the underlying outcome rate. We simulated data using a new user self-controlled risk interval design, which followed patients exposed to a particular medical product for a pre-defined period post-exposure known as the observation window. Within a patient's observation window, the time is divided up between a risk window period when the patient is assumed to be at higher risk for experiencing medical product-associated outcomes and a comparison window period. These periods are then compared. We created known alternative hypotheses that generated clusters of excess risk in the tree structure that happened in particular time periods. We then used the tree-based scan statistics that are compatible with Bernoulli-structured data to analyze these simulated study data.

Figure 1. Example branch of the Multi-Level Clinical Classification System. The branch is not shown in full. The upper level is the root and the lowest level is the leaf level.


Diseases of the Nerve and Sense Organs
Central Nervous System Infection
Hereditary and degenerative nervous system condition
Paralysis
Epilepsy; convulsions
Epilepsy
Convulsions
Convulsions
Convulsions ICD-9-CM 780.3
Febrile convulsions NOS ICD-9-CM 780.31
Complex febrile convulsions ICD-9-CM 780.32
Post traumatic seizures ICD-9-CM 780.33
Other convulsions ICD-9-CM 780.39

II. METHODS

A. HIERARCHICAL TREE

The tree-based scan statistic detects elevated frequencies of outcomes in electronic health data that have been grouped into hierarchical tree structures. This approach takes advantage of the hierarchical nature of clinical concepts, including clinical outcomes and medical product exposures. Here, 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. Figure 1 shows an example branch.

We curated the full 2014 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 6,551 ICD-9-CM codes which all represent individual leaves on the tree. Overall, there are 7,306 nodes on the tree. The curated tree is available upon request.

B. TREE-BASED SCAN STATISTIC FOR BERNOULLI DATA

The null hypothesis assumes outcomes are uniformly distributed across the observation window following the incident medical product exposure. Under the alternative hypothesis, there is at least one outcome, or group of related outcomes, that occurs in excess of what would be expected in a particular risk window. In the conditional forms of the tree-based scan statistic, the outcomes are standardized by the frequency with which they appear in the overall dataset on any given day within the observation window. Conditioning is a mechanism to control for situations when there is an across-the-board increase in healthcare 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.

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 data set is among the 5 percent highest of all the maxima, the null hypothesis is rejected. The fact 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 seven thousand nodes.[21] Specific details of this procedure are included in Appendix A.

C. 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.

We extracted background rates for all the outcomes in the curated MLCCS tree from databases that participate in the Post-licensure Immunization and Safety Monitoring (PRISM) system.[22] We extracted a cohort of 9-26 year olds from June 2006 to December 2014. 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 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 the background rates are listed in **Table 1**.

Table 1. Key Metrics of the Source Data[a] used to Capture the Background Rates of Outcomes of Interest

Key Metrics	
Total person-years followed	34,607,477
Total events	5,552,734
Total persons	1,903,697
Total exposed person-years	147,432
Total expected events	19,498
Total observed events in the risk window	27,714

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

Outcome events were defined by ICD-9-CM codes and visit location or encounter 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 relative to the potential incident outcome 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 **Figure 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.

These background rates are used to simulate outcome counts in the time following medical product exposure (i.e., the observation window) that are used by the Bernoulli tree scan statistic for comparison of the outcomes in the risk window to the outcomes in the comparison window. These counts are simulated for each of the 6,551 nodes on the tree. Only the first dose of the vaccination was simulated.

D. 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 three 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. That is, there was a pre-specified number of excess cases on particular leaves of the tree. The choices of outcomes were incidental, but were required to be differing orders of magnitude in their base frequency in the dataset.

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 healthcare 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. Under these circumstances, we compared the ability of the conditional and unconditional tree-based scan statistics to control for type I error.

E. MIS-SPECIFICATION OF THE RISK WINDOW

In the initial scenarios we tested, 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.[23]

First, we considered the circumstance when the true risk window was longer, but encompassed the observed risk window. We refer to this circumstance as a too-short risk window. For example, the true risk window could occur 1-28 days post-vaccination whereas the observed risk window could occur 1-7 days post-vaccination. That is, outcomes in the 15-28 days following vaccination would be included in the control window. With outcomes that are occurring in the true risk window misclassified as control window outcomes, a too-short risk window biases effect estimates toward the null. 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 because there are no excess cases that occur during portions of the risk window. Again, the net effect is effect estimates that are biased toward the null.

F. POWER EVALUATIONS

All analyses were performed using the power evaluation feature in the free TreeScan™ tool (www.treescan.org, v1.1.4), 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 is always possible to determine which nodes individually alert.

We used Monte Carlo simulation to create multiple alternative datasets under both the null and known alternative hypotheses as described above in sections C and D. For power evaluations related to Poisson-based tree scan statistics, the input dataset has no dependencies.[20] That is, the input dataset is based only on expected rates and it is therefore possible to perform the power evaluation with a single input dataset. For Bernoulli tree scan statistics, the power evaluation depends on the total observation window outcomes for each node. Further, the total observation window outcome count is based on three expected values: expected outcome counts in the risk window under the null hypothesis, expected outcome counts in the comparison window under the null hypothesis, and excess cases in the risk window under the alternative hypothesis. Because Bernoulli data require integer-valued counts, each of these expected rates serves as the input to a Poisson random draw. Therefore, a single input dataset has dependencies as it represents a single instantiation of a Poisson process. Appropriate power evaluations require multiple input datasets (i.e., thousands) and consequent executions of the power evaluation feature in TreeScan™. This is computationally quite expensive.

Through a preliminary phase of this work, we were able to determine that the power range generated by the various input datasets was highly correlated with the total outcomes observed in the node affected by the alternative hypothesis. Full details of this first phase are in Appendix B. Therefore, to eliminate computation time, we report the power evaluations for the median value of outcomes in the affected node. 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.

III. RESULTS

A. STATISTICAL POWER

Figure 2a and 2b shows the statistical power to detect various attributable risks. We vary the total sample size among three outcomes of interest with underlying event rates that vary by orders of magnitude. Among these three outcomes in the population of interest, syncope (ICD-9-CM 780.2) occurs most frequently with 311 expected outcomes in the 56-day observation window for every 1 million doses administered whereas systemic lupus erythematosus (ICD-9-CM 710.0) occurs least frequently with 1.4 expected outcomes in the 56-day observation window for every 1 million doses. These expected outcome totals do not include a value for expected excess cases due to an increased risk.

Figure 2a. Statistical power to detect various attributable risks, accounting for different background event rates, and sample sizes.

		Incidence Rate Difference of Interest (Events per Million doses)									
Total Expected Outcomes	Vaccinees	0	1	2	5	10	20	50	100	200	500
Syncope (ICD-9-CM 780.2), Unconditional Bernoulli Analysis											
31.1	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.07	0.23	0.99
62.2	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.13	0.72	1.00
155.5	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.07	0.43	1.00	1.00
311.0	1 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.17	0.90	1.00	1.00
622.0	2 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.49	1.00	1.00	1.00
1554.8	5 M	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.12	0.98	1.00	1.00	1.00
Syncope (ICD-9-CM 780.2), Conditional Bernoulli Analysis											
31.1	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.07	0.22	0.99
62.2	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.13	0.68	1.00
155.5	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.07	0.43	1.00	1.00
311.0	1 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.16	0.86	1.00	1.00
622.0	2 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.48	1.00	1.00	1.00
1554.8	5 M	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.12	0.98	1.00	1.00	1.00
Syncope (ICD-9-CM 780.2), Unconditional Tree-Temporal Analysis											
31.1	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.10	0.93
62.2	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.35	1.00
155.5	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.17	0.96	1.00
311.0	1 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.07	0.59	1.00	1.00
622.0	2 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.20	0.99	1.00	1.00
1554.8	5 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.84	1.00	1.00	1.00
Thrombocytopenia (ICD-9-CM 275.0), Unconditional Bernoulli Analysis											
0.82	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.09	0.67	1.00	1.00
1.6	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.46	0.99	1.00	1.00
4.1	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.22	0.96	1.00	1.00	1.00
8.2	1 M	≤0.05	≤0.05	≤0.05	0.06	0.13	0.67	1.00	1.00	1.00	1.00
16.4	2 M	≤0.05	≤0.05	≤0.05	0.08	0.32	0.99	1.00	1.00	1.00	1.00
41.0	5 M	≤0.05	≤0.05	0.06	0.26	0.95	1.00	1.00	1.00	1.00	1.00
Thrombocytopenia (ICD-9-CM 275.0), Conditional Bernoulli Analysis											
0.82	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.09	0.67	1.00	1.00
1.6	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.43	0.99	1.00	1.00
4.1	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.22	0.97	1.00	1.00	1.00
8.2	1 M	≤0.05	≤0.05	≤0.05	0.06	0.13	0.67	1.00	1.00	1.00	1.00
16.4	2 M	≤0.05	≤0.05	≤0.05	0.07	0.34	0.99	1.00	1.00	1.00	1.00
41.0	5 M	≤0.05	≤0.05	≤0.05	0.27	0.96	1.00	1.00	1.00	1.00	1.00
Thrombocytopenia (ICD-9-CM 275.0), Unconditional Tree-Temporal Analysis											
0.82	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.19	0.92	1.00
1.6	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.14	0.77	1.00	1.00
4.1	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.09	0.79	1.00	1.00	1.00
8.2	1 M	≤0.05	≤0.05	≤0.05	0.05	0.08	0.28	1.00	1.00	1.00	1.00
16.4	2 M	≤0.05	≤0.05	≤0.05	0.06	0.13	0.87	1.00	1.00	1.00	1.00
41.0	5 M	≤0.05	≤0.05	≤0.05	0.08	0.72	1.00	1.00	1.00	1.00	1.00

Notes: All simulations were performed with 99,999 iterations under the null hypothesis, 10,000 iterations under the known alternative hypothesis. Critical values were set at a signaling threshold of p=0.05.

Figure 2b. Statistical power to detect various attributable risks, accounting for different background event rates, and sample sizes.

		Incidence Rate Difference of Interest (Events per Million doses)									
Total Expected Outcomes	Vaccinees	0	1	2	5	10	20	50	100	200	500
Systemic Lupus Erythematosus (ICD-9-CM 710.0), Unconditional Bernoulli Analysis											
0.14	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.94	1.00	1.00
0.28	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.88	1.00	1.00	1.00
0.69	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.21	1.00	1.00	1.00	1.00
1.4	1 M	≤0.05	≤0.05	≤0.05	≤0.05	0.53	0.97	1.00	1.00	1.00	1.00
2.8	2 M	≤0.05	≤0.05	≤0.05	0.29	0.87	1.00	1.00	1.00	1.00	1.00
6.9	5 M	≤0.05	0.07	0.16	0.89	1.00	1.00	1.00	1.00	1.00	1.00
Systemic Lupus Erythematosus (ICD-9-CM 710.0), Conditional Bernoulli Analysis											
0.14	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.94	1.00	1.00
0.28	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.88	1.00	1.00	1.00
0.69	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.21	1.00	1.00	1.00	1.00
1.4	1 M	≤0.05	≤0.05	≤0.05	≤0.05	0.53	0.97	1.00	1.00	1.00	1.00
2.8	2 M	≤0.05	≤0.05	≤0.05	0.29	0.87	1.00	1.00	1.00	1.00	1.00
6.9	5 M	≤0.05	0.07	0.16	0.88	1.00	1.00	1.00	1.00	1.00	1.00
Systemic Lupus Erythematosus (ICD-9-CM 710.0), Unconditional Tree-Temporal Analysis											
0.14	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.06	0.11	0.97	1.00
0.28	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.09	0.93	1.00	1.00
0.69	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.08	0.97	1.00	1.00	1.00
1.38	1 M	≤0.05	≤0.05	≤0.05	≤0.05	0.08	0.68	1.00	1.00	1.00	1.00
2.76	2 M	≤0.05	≤0.05	≤0.05	0.09	0.53	1.00	1.00	1.00	1.00	1.00
6.90	5 M	≤0.05	≤0.05	0.06	0.48	1.00	1.00	1.00	1.00	1.00	1.00

Notes: All simulations were performed with 99,999 iterations under the null hypothesis, 10,000 iterations under the known alternative hypothesis. Critical values were set at a signaling threshold of p=0.05.

As seen in prior power studies of the tree-based scan statistics, 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 outcomes occurring in the risk window and comparison window time periods.[20] To illustrate, five excess events in the risk window amounts to statistical noise in a more commonly occurring outcome such as syncope when over three hundred outcomes are expected. However, with rare events such as systemic lupus erythematosus, five additional outcomes in the risk window when only a few are expected generates meaningful separation between the two time periods, thereby generating higher statistical power to rule out the same attributable risk. As expected, it is easier to detect the same risk differences with larger sample sizes.

The statistical power of the unconditional and conditional Bernoulli tree scan statistics are quite similar when applied to the same dataset. The unconditional tree-temporal scan statistic has less statistical power for the same fixed risk difference when compared to its Bernoulli counterpart. This occurs because there is an increased level of multiple hypothesis testing when using the tree-temporal scan statistic. In addition to evaluating the test statistic on many nodes across the tree, the tree-temporal scan also evaluates many potential risk windows within the designated observation window for each node. In **Figure 2**a and 2b, both the true risk window and the observed risk window are the 28 days following vaccination. Therefore, the risk window is perfectly specified. In such situations, the Bernoulli test statistic is more efficient than the tree-temporal because there is no unnecessary hypothesis testing coupled with testing multiple risk windows of interest.

B. TYPE I ERROR

Table 2 demonstrates the ability of the conditional v. unconditional Bernoulli tree scan statistic to properly control for across-the-board elevations in healthcare 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 Bernoulli tree scan statistic inflates type I error when general utilization is increased by as little as 2%. 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 1**. That is, when comparing the total observed events in the risk window to the total expected events (27,714 v. 19,498), there is a ~40% increase in across-the-board utilization. However, the conditional Bernoulli tree scan statistic continues to hold type I error to the allowable level even when across-the-board healthcare utilization increases by 200%.

Table 2. Observed Type I Error in the Conditional and Unconditional Bernoulli Tree Scan Statistic under Conditions of Across-the-board Elevations in Healthcare Utilization[a]

Bernoulli Scan Statistic	Vaccinees	Increases in Across-the-board Elevations in Healthcare Utilization								
		0%	**1%**	**2%**	**5%**	**10%**	**20%**	**50%**	**200%**	**500%**
Unconditional	0.1M	0.05	0.05	0.06	0.05	0.11	0.24	0.92	1.00	1.00
	0.2M	0.05	0.06	0.06	0.08	0.13	0.39	1.00	1.00	1.00
	0.5M	0.05	0.04	0.06	0.07	0.17	0.79	1.00	1.00	1.00
	1M	0.05	0.06	0.07	0.11	0.41	1.00	1.00	1.00	1.00
	2M	0.05	0.06	0.07	0.17	0.77	1.00	1.00	1.00	1.00
	5M	0.05	0.07	0.08	0.42	1.00	1.00	1.00	1.00	1.00
Conditional	0.1M	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05
	0.2M	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06
	0.5M	0.05	0.03	0.05	0.05	0.05	0.05	0.05	0.06	0.06
	1M	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.06	0.06
	2M	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.07
	5M	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.10

[a]All simulations were performed with 99,999 iterations under the null hypothesis.

C. MIS-SPECIFICATION OF THE RISK WINDOW

Figure 3 demonstrates the effect on statistical power when a too-short risk window has been specified using the Bernoulli scan statistic. In this case, the true risk window is Days 1-28 and the comparison window is Days 29-56. As seen in **Figure 2**, when the true risk window is correctly specified (as in the upper third of the figure), the tree-temporal scan statistic has less statistical power than the Bernoulli scan statistic because of the additional hypothesis testing accounting for multiple risk windows per node. However, in the lower third of Figure 3, when the true risk window is Days 1-28 and the risk window is specified to be Days 1-7 (i.e., too short), then additional excess cases are misclassified as control window outcomes. The tree-temporal scan statistic has higher statistical power than the Bernoulli scan statistic. The losses in statistical power occur because of the bias toward the null that occurs with misclassification.

Figure 3. Statistical power to detect various attributable risks and sample sizes, while comparing a correctly-specified risk window to an incorrectly-specified risk window (i.e., too short).

Incidence Rate Difference of Interest (Events per Million doses)											
Total Expected Outcomes	Vaccinees	0	1	2	5	10	20	50	100	200	500
Syncope (ICD-9-CM 780.2), Unconditional Bernoulli Analysis with a 28-day risk window (TRUE)											
31.1	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.07	0.23	0.99
62.2	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.13	0.72	1.00
155.5	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.07	0.43	1.00	1.00
311.0	1 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.17	0.90	1.00	1.00
622.0	2 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.49	1.00	1.00	1.00
1554.8	5 M	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.12	0.98	1.00	1.00	1.00
Syncope (ICD-9-CM 780.2), Unconditional Tree-Temporal Analysis											
31.1	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.10	0.93
62.2	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.35	1.00
155.5	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.17	0.96	1.00
311.0	1 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.07	0.59	1.00	1.00
622.0	2 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.20	0.99	1.00	1.00
1554.8	5 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.84	1.00	1.00	1.00
Syncope (ICD-9-CM 780.2), Unconditional Bernoulli Analysis with a 7-day risk window (TOO SHORT)											
31.1	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.20
62.2	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.52
155.5	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.12	0.99
311.0	1 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.32	1.00
622.0	2 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.09	0.73	1.00
1554.8	5 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.34	1.00	1.00

Notes: All simulations were performed with 99,999 iterations under the null hypothesis, 10,000 iterations under the known alternative hypothesis. Critical values were set at a signaling threshold of p=0.05.

Figure 4 demonstrates the effect on statistical power when a too-long risk window has been specified using the Bernoulli scan statistic. Here, the true risk window is Days 1-7 and the comparison window is Days 8-56. As before, when the true risk window is correctly specified (as in the upper third of **Figure 4**), the tree-temporal scan statistic has slightly less statistical power than the Bernoulli scan statistic because of the additional hypothesis testing accounting for multiple risk windows per node. However, in the lower third of **Figure 4**, when the true risk window is Days 1-7 and the risk window is specified to be Days 1-28 (i.e., too-long), the tree-temporal scan statistic has higher statistical power than the Bernoulli scan statistic. The losses in statistical power occur because of the “washing out” of the signal. That is, there are a smaller number of excess outcomes that are spread out over a longer time period.

Figure 4. Statistical power to detect various attributable risks and sample sizes, while comparing a correctly-specified risk window to an incorrectly-specified risk window.

Incidence Rate Difference of Interest (Events per Million doses)											
Total Expected Outcomes	Vaccinees	0	1	2	5	10	20	50	100	200	500
Syncope (ICD-9-CM 780.2), Unconditional Bernoulli Analysis with a 7-day risk window											
31.1	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.13	0.56	0.98	1.00
62.2	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.26	0.90	1.00	1.00
155.5	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.08	0.75	1.00	1.00	1.00
311.0	1 M	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.18	0.99	1.00	1.00	1.00
622.0	2 M	≤0.05	≤0.05	≤0.05	≤0.05	0.08	0.48	1.00	1.00	1.00	1.00
1554.8	5 M	≤0.05	≤0.05	≤0.05	≤0.05	0.22	0.98	1.00	1.00	1.00	1.00
Syncope (ICD-9-CM 780.2), Unconditional Tree-Temporal Analysis											
31.1	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.22	0.91	1.00
62.2	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.11	0.63	1.00	1.00
155.5	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.43	1.00	1.00	1.00
311.0	1 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.08	0.91	1.00	1.00	1.00
622.0	2 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.22	1.00	1.00	1.00	1.00
1554.8	5 M	≤0.05	≤0.05	≤0.05	≤0.05	0.08	0.87	1.00	1.00	1.00	1.00
Syncope (ICD-9-CM 780.2), Unconditional Bernoulli Analysis with a 28-day risk window											
31.1	0.1M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.05	0.31	1.00
62.2	0.2M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.15	0.73	1.00
155.5	0.5M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.08	0.44	1.00	1.00
311.0	1 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.05	0.19	0.89	1.00	1.00
622.0	2 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.06	0.45	1.00	1.00	1.00
1554.8	5 M	≤0.05	≤0.05	≤0.05	≤0.05	≤0.05	0.12	0.98	1.00	1.00	1.00

Notes: All simulations were performed with 99,999 iterations under the null hypothesis, 10,000 iterations under the known alternative hypothesis. Critical values were set at a signaling threshold of p=0.05.

IV. DISCUSSION

We performed numerous simulations to examine the statistical power of both the unconditional and conditional Bernoulli tree scan statistic as well as the unconditional tree-temporal scan statistic. These scan statistics support data that are collected as part of a self-controlled risk interval design or a fixed-ratio matched design. The unconditional and conditional Bernoulli tree-based scan statistics have nearly the same statistical power. The unconditional tree scan statistic inflated type I error even in the presence of low general increases in healthcare utilization following exposure whereas the conditional tree scan statistic controlled type I error well in the same circumstance. When the risk window is known, Bernoulli scan statistics are preferred to the tree-temporal scan statistic because they are power-preserving. However, when the risk window is unknown or uncertain, tree-temporal scan statistics are preferred to potentially mis-specified Bernoulli scan statistics. We also observed reductions in statistical power resulting from specifying either a too-short or a too-long 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. Then, according to the simulation procedure specified in more detail in Appendix B, input datasets and alternative hypothesis files were created and evaluated. The statistical power concepts demonstrated with this example (i.e., a tree-temporal scan statistic is preferred when the risk window is unknown due to statistical power losses that occur with a mis-specified Bernoulli scan statistic) should apply to all problems regardless of the source data.

When performing data-mining in TreeScan™, the same risk window specification for any Bernoulli tree scan statistic will apply to all 6000+ outcomes in the entire tree. Therefore, it is unlikely to correctly specify the identical risk window for all outcomes. Consequently, the tree-temporal scan statistic will likely be preferred in most data-mining exercises (i.e., when the total nodes evaluated is in the hundreds or thousands). While we did not perform power evaluations for the conditional tree-temporal scan statistic (i.e., such evaluations are not built into TreeScan™ because the perturbation process is so computationally intensive), if there is any expectation of increased levels of across-the-board healthcare utilization, then a conditional tree-temporal scan statistic will be preferred to an unconditional tree-temporal scan statistic. The power is expected to be minimally different between these two scan statistics however type I error is expected to be more poorly controlled with the unconditional tree-temporal scan statistic. The PRISM background data used here points to a 40% increase in general healthcare 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

Our preparatory-to-surveillance simulation demonstrates how to estimate 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 with various sample sizes. 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.[24]

There were limitations of this evaluation, which are either inherent to secondary-use observational data, the nature of data-mining, or limitations in computational speed and efficiency.

First, electronic healthcare databases have key advantages including representativeness of routine clinical practice and efficient capture of the healthcare experiences of a large patient population. However, there are fundamental limitations to using administrative claims data for safety surveillance.[25]

Second, data-mining analyses are dependent on design-based confounding control. Self-controlled risk interval designs that compare time periods post-exposure are particularly sensitive to risk window specifications when using either the unconditional or conditional Bernoulli scan statistics. Risk windows have a universal specification for all outcomes being evaluated.

Third, power evaluations for the conditional tree-temporal scan statistic are not developed at this time as a result of computational complexity. However, there has been little difference in statistical power between the conditional and unconditional versions of the Bernoulli tree scan statistics.

V. CONCLUSIONS

Signal identification has traditionally been strongly driven by spontaneous reports, which lack population data to provide context. 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.

VI. REFERENCES

1. Dal Pan, G. J., Lindquist, M. & Gelperin, K. in *Pharmacoepidemiology* (eds. Strom, B. L., Kimmel, S. E. & Hennessy, S.) 137–157 (John Wiley & Sons, 2011).

2. Platt, R. *et al.* The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol. Drug Saf.* **21 Suppl 1,** 1–8 (2012).

3. Kulldorff, M., Fang, Z. & Walsh, S. J. A tree-based scan statistic for database disease surveillance. *Biometrics* **59,** 323–331 (2003).

4. Kulldorff, M. *et al.* Drug safety data mining with a tree-based scan statistic. *Pharmacoepidemiol. Drug Saf.* **22,** 517–523 (2013).

5. Brown, J. S. *et al.* Drug Adverse Event Detection in Health Plan Data Using the Gamma Poisson Shrinker and Comparison to the Tree-based Scan Statistic. *Pharmaceutics* **5,** 179–200 (2013).

6. Yih, W. K. *et al. Pilot of Self-Controlled Tree-Temporal Scan Analysis for Gardasil Vaccine*. (U.S. Food and Drug Administration, 2016).

7. Yih, W. K. *et al. Evaluation of HPV9 (Gardasil9) Vaccine Safety Surveillance Using the TreeScan Data Mining Method Surveillance Protocol*. (U.S. Food and Drug Administration, 2016).

8. Moore, T. J. & Furberg, C. D. Electronic Health Data for Postmarket Surveillance: A Vision Not Realized. *Drug Saf.* **38,** 601–610 (2015).

9. Gault, N. *et al.* Self-controlled designs in pharmacoepidemiology involving electronic healthcare databases: a systematic review. *BMC Med. Res. Methodol.* **17,** (2017).

10. Rothman, K. J., Greenland, S. & Lash, T. L. *Modern epidemiology*. (Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008).

11. Sahu, S. K., Bendel, R. B. & Sison, C. P. Effect of relative risk and cluster configuration on the power of the one-dimensional scan statistic. *Stat. Med.* **12,** 1853–1865 (1993).

12. Jung, I. & Lee, H. Spatial cluster detection for ordinal outcome data. *Stat. Med.* **31,** 4040–4048 (2012).

13. Neill, D. B. An empirical comparison of spatial scan statistics for outbreak detection. *Int. J. Health Geogr.* **8,** 20 (2009).

14. Huang, L., Pickle, L. W. & Das, B. Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases. *Stat. Med.* **27,** 5111–5142 (2008).

15. Huang, L., Kulldorff, M. & Gregorio, D. A spatial scan statistic for survival data. *Biometrics* **63,** 109–118 (2007).

16. Kulldorff, M. *et al.* Benchmark data and power calculations for evaluating disease outbreak detection methods. *MMWR Morb. Mortal. Wkly. Rep.* **53 Suppl,** 144–151 (2004).

17. Kulldorff, M. *et al.* Multivariate scan statistics for disease surveillance. *Stat. Med.* **26,** 1824–1833 (2007).

18. Waller, L. A., Hill, E. G. & Rudd, R. A. The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. *Stat. Med.* **25,** 853–865 (2006).

19. Song, C. & Kulldorff, M. Power evaluation of disease clustering tests. *Int. J. Health Geogr.* **2,** 9 (2003).

20. Maro, J. C., Nguyen, M. D., Dashevsky, I., Baker, M. A. & Kulldorff, M. Statistical Power for Postlicensure Medical Product Safety Data-Mining. *eGEMS.* In press.

21. Dwass, M. Modified Randomization Tests for Nonparametric Hypotheses. *Ann. Math. Stat.* **28,** 181–187 (1957).

22. Nguyen, M., Ball, R., Midthun, K. & Lieu, T. A. The Food and Drug Administration's Post-Licensure Rapid Immunization Safety Monitoring program: strengthening the federal vaccine safety enterprise. *Pharmacoepidemiol. Drug Saf.* **21 Suppl 1,** 291–297 (2012).

23. Rowhani-Rahbar, A. *et al.* Biologically plausible and evidence-based risk intervals in immunization safety research. *Vaccine* **31,** 271–277 (2012).

24. Toh, S. *et al.* Re-using Mini-Sentinel data following rapid assessments of potential safety signals via modular analytic programs. *Pharmacoepidemiol. Drug Saf.* **22,** 1036–1045 (2013).

25. Schneeweiss, S. & Avorn, J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J. Clin. Epidemiol.* **58,** 323–337 (2005).

VII. ACKNOWLEDGEMENTS

We acknowledge the contributions of our other workgroup members including Carolyn Balsbaugh and Megan Reidy.

VIII. APPENDIX A - BRIEF DESCRIPTION OF THE METHODS

A. UNCONDITIONAL BERNOULLI SCAN STATISTIC WITH FIXED RISK WINDOW

All outcomes are first classified into a hierarchical tree structure described in Section II A above. For each leaf *i* of the tree (i.e., finest granularity) which represents a unique outcome or ICD-9-CM code of interest, we note the observed number c_i of outcomes in the risk window and the observed number n_i of outcomes in the comparison window.

The next step is to define nodes on the tree. Each node *G* defines either an outcome (if at the leaf level) or a group of related outcomes, i.e., a branch on the tree. The sums of the observed number of outcomes in this node in the risk and comparison window are denoted as c_G and n_G respectively. Note that a single leaf is one potential node, but a node could also be an entire branch of the tree.

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

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

where:

> p is the length of the risk window divided by the sum of the lengths of the risk and comparison windows. This represents the Bernoulli probability under the null hypothesis that the outcome occurs in proportion to the length of the window.
>
> $I()$ is the indication function, which is 1 when there are more outcomes in the risk window than would be expected by chance. It is included to ensure that we are looking for an excess risk of having the adverse event rather than a protective decreased risk.

Log likelihood ratios are computed for computational convenience and results from them are equivalent 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, and so inference is conducted using Monte Carlo hypothesis testing.[21] 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 length of the risk window relative to the observation window. *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 exposures. 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≤0.05.

B. CONDITIONAL BERNOULLI SCAN STATISTIC WITH FIXED RISK WINDOW

When using the unconditional Bernoulli tree-based scan statistic described above, the null hypothesis is that any outcome is likely to occur in proportion to the length of the risk and comparison windows. In the conditional version, the lengths of the two windows are ignored, and instead the null hypothesis is based on the proportion of the sum of outcomes in the risk window of a particular node as compared to the total number of outcomes in the risk window observed in the whole tree.

Thus, we calculate the total number of outcomes in the risk window $C = \sum_i c_i$ observed in the whole tree and the total number of outcomes in the comparison window $N = \sum_i n_i$ observed in the whole tree.

So, when comparing the unconditional to the conditional, the probability *p* used above is now replaced by $\left(\frac{C}{C+N}\right)$.

The LLR for the conditional Bernoulli tree-based scan statistic is

$$LLR = \ln\left(\frac{\left(\frac{c_G}{c_G+n_G}\right)^{c_G}\left(\frac{n_G}{c_G+n_G}\right)^{n_G}}{\left(\frac{C}{C+N}\right)^{c_G}\left(\frac{N}{C+N}\right)^{n_G}}\right) I\left(\frac{c_G}{c_G+n_G} > \frac{C}{C+N}\right)$$

$I()$ is the indication function, which is 1 when there are more outcomes in the risk window 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* as the real data*,* so that the total number of outcomes in the risk window and control windows are the same in both the real and all the random data sets. The rest of the procedure is the same as described above.

C. UNCONDITIONAL TREE-TEMPORAL SCAN STATISTIC WITH VARYING RISK WINDOW

The unconditional tree-temporal scan statistic – also called the tree-temporal scan – adds a temporal dimension to the data. Now, in addition to the multiple hypotheses tested based on the tree structure as in the fixed risk window studies, each node itself contributes multiple temporal hypotheses related to the length of the risk and comparison windows.

$$LLR = \ln\left(\frac{\left(\frac{c_G}{c_G+n_G}\right)^{c_G}\left(\frac{n_G}{c_G+n_G}\right)^{n_G}}{\left(\frac{w}{O}\right)^{c_G}\left(\frac{O-w}{O}\right)^{n_G}}\right) I\left(\frac{c_G}{c_G+n_G} > \frac{w}{O}\right)$$

where:

c_G is the number of outcomes in the node *G* of interest that are also in the variable risk window

n_G is the number of outcomes in the node that are NOT in the variable risk window

w is the length of the variable risk window

O is the length of the total observation window.

$I()$ is the indication function, which is 1 when there are more outcomes in the risk window than expected under the null, and 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. Note that *O* is a constant that is the same for every node and every potential risk window (i.e., time interval of interest).

Similar to the unconditional fixed risk window analysis described above, the null hypothesis is again that the outcome occurs in proportion to the length of the risk window relative to the total observation window.

As before, 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 Monte Carlo simulation step occurs similarly as described before.

IX. APPENDIX B – FULL SCALE SIMULATION STEPS

A. INITIAL INVESTIGATIONS WITH MEAN STATISTICAL POWER ACROSS MANY SCENARIOS

For both Bernoulli and tree-temporal scan statistics, the power evaluation feature of TreeScan™ uses the input dataset (*.cas file) and its total number of observed outcomes in the node to conduct a perturbation process under both the null and alternative hypotheses. For example, if there were ten total outcomes observed in the incoming dataset in a particular node, and the risk window was equal in length to the control window, then p0 is 0.5. Perturbation of these outcomes under the null hypothesis are repeated binomial random draws with n=total outcomes and p=p0. Consequently, the statistical power of a particular dataset is dependent on the total number of observed outcomes in each node.

To test an alternative hypothesis on this node, one specifies the p1 in the alternative hypothesis file, which is applied to the affected node only, and uses the same technique: a binomial random draw on with n=total outcomes in that node and p=p1. However, to get the correct statistical power, one must seed the total number of cases in the affected node to include the excess cases planned under p=p1. So, the input dataset is also unique to the alternative hypothesis being evaluated.

All of the outcome counts on each node are a particular instantiation or realization of a Poisson process. Because the Bernoulli and tree-temporal scan statistics depend on the input dataset, to calculate the statistical power of an injected elevated risk accurately requires multiple input datasets per scenario (i.e., specific combination of input dataset, alternative hypothesis file, and pre-set parameters). This is computationally quite expensive.

For one outcome – headache (ICD-9-CM 784.0) – we created 1000 input datasets for each scenario. Each input dataset is a realization (or random draw) on the same Poisson process. Each cell in **Figure 5** shows the statistical power based on the mean power across 1000 scenarios. Overall, this simulation process is computationally intensive, and we sought a simpler approximation in order to create a procedure for routine use.

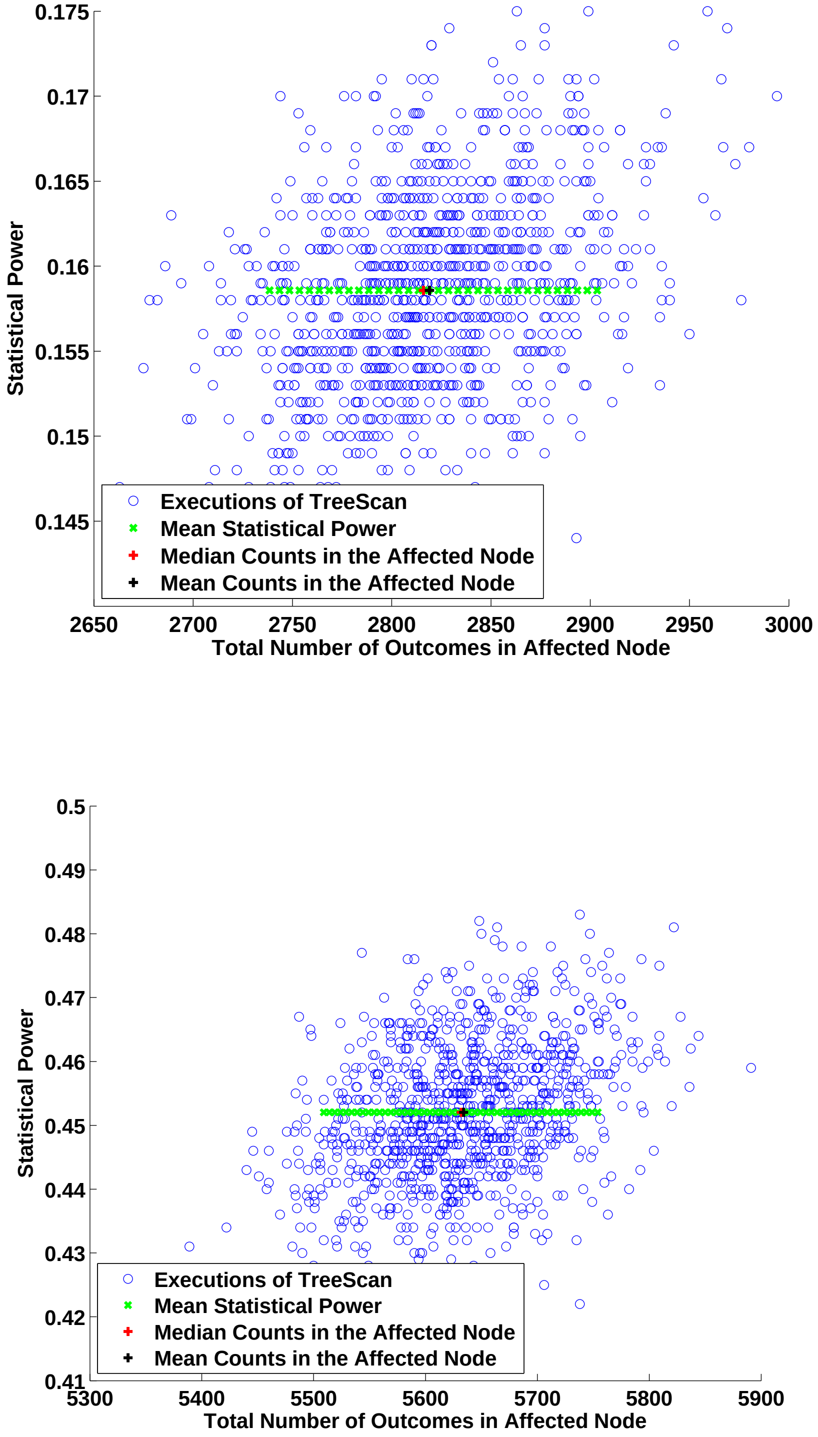
Figure 5. Statistical power to detect various attributable risks and sample sizes.

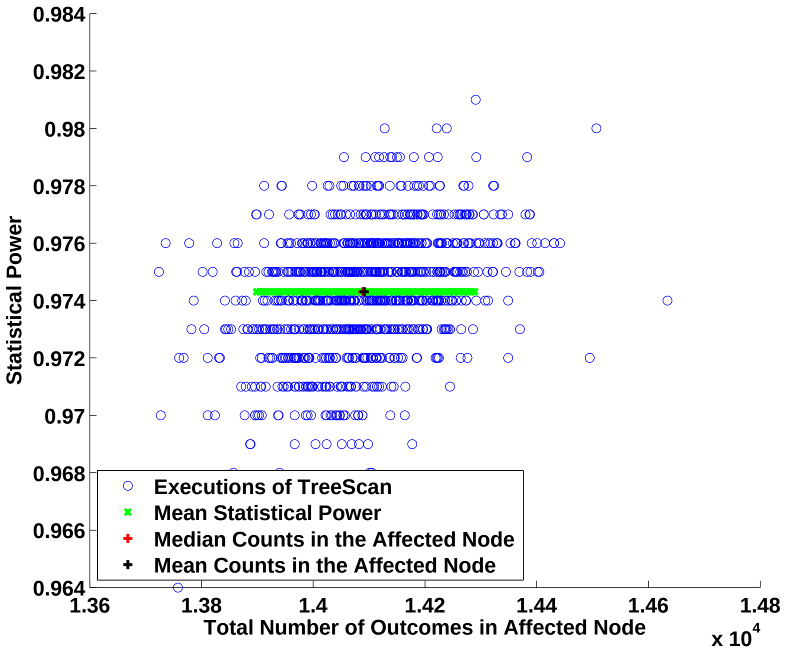
<table>
<tr><th colspan="7">Incidence Rate Difference of Interest (Events per Million doses)</th></tr>
<tr><th>Total Expected Outcomes</th><th>Vaccinees</th><th>0</th><th>100</th><th>200</th><th>500</th><th>1000</th></tr>
<tr><th colspan="7">Headache (ICD-9-CM 784.0), Unconditional Bernoulli Analysis with a 28-day risk window</th></tr>
<tr><td>44.4</td><td>0.1M</td><td>≤0.05</td><td>≤0.05</td><td>0.05</td><td>0.10</td><td>0.64</td></tr>
<tr><td>88.7</td><td>0.2M</td><td>≤0.05</td><td>≤0.05</td><td>0.06</td><td>0.24</td><td>0.97</td></tr>
<tr><td>221.8</td><td>0.5M</td><td>≤0.05</td><td>≤0.05</td><td>0.07</td><td>0.79</td><td>1.00</td></tr>
<tr><td>443.6</td><td>1 M</td><td>≤0.05</td><td>0.06</td><td>0.16</td><td>1.00</td><td>1.00</td></tr>
<tr><td>887.3</td><td>2 M</td><td>≤0.05</td><td>0.07</td><td>0.45</td><td>1.00</td><td>1.00</td></tr>
<tr><td>2218.2</td><td>5 M</td><td>≤0.05</td><td>0.20</td><td>0.97</td><td>1.00</td><td>1.00</td></tr>
</table>

Notes: All simulations were performed with 99,999 iterations under the null hypothesis, 10,000 iterations under the known alternative hypothesis. Critical values were set at a signaling threshold of p=0.05.

Figure 6 demonstrates the correlation between the statistical power and the total number of outcomes in the affected node over 1000 input datasets. The mean power reported in **Figure 5** as a single value is the mean statistical power for the scenario across the 1000 input datasets. While the range of statistical power is still dependent on the particular input dataset received by TreeScan™, the mean statistical power occurs coincident with the mean outcomes in the affected node. Because the mean outcomes in the affected node over 1000 input datasets are not integer-valued, we use the median outcomes to report statistical power.

Figure 6a, 6b, and 6c. Scatter Plot of Statistical Power to detect 200 excess headache outcomes (ICD-9-CM 784.0) in a 1 million, 2 million, and 5 million vaccinee sample (respectively) over 1000 input datasets with each dataset drawn from a Poisson process.





B. PROCEDURE TO CREATE INPUT DATASETS AND ALTERNATIVE HYPOTHESES

1. For every pre-defined risk window, sample size population, and attributable risk difference of interest:

a. Use background rates derived from background population to determine expected counts in the risk window and expected counts in the control window for all nodes on the tree.

b. Generate Bernoulli analytic dataset using Poisson random draws with means defined in (a) above for all nodes EXCEPT the node of interest. By definition, the probability of a case occurring in the risk window under the null hypothesis is equal to the length of the risk window / length of the observation window.

c. For the node of interest, the mean total number of outcomes will be the sum of expected counts in the risk window, expected counts in the control window and the user-defined number of excess cases per the attribute risk difference of interest. Draw 1000 Poisson samples of this sum. Assign the median value of total outcomes to the affected node of interest (i.e., note **not the mean** as the mean will not be integer-valued) in the Bernoulli analytic dataset created in (b).

d. Generate alternative hypothesis files on the affected node based on p1 defined as:

(expected cases in risk window + excess cases)/(expected cases in risk window + excess cases + expected cases in control window)

2. From Bernoulli analytic dataset, create tree-temporal scan dataset by assigning each case in the risk and control windows to specific days according to a random uniform discrete sampling procedure. Adapt Bernoulli alternative hypothesis file to create tree-temporal alternative hypothesis file.