MINI-SENTINEL METHODS

EVALUATION OF SCAN STATISTICS FOR ASSESSING VACCINE SAFETY IN PREGNANCY

Prepared by: Lingling Li, PhD,[1] Stanley Xu, PhD,[2] Lihan Yan, PhD,[3] Alison Tse Kawai, ScD,[1] Gabriela Vazquez Benitez, PhD, MSC,[4] Wei Hua, MD, PhD[3]

Author Affiliations: 1. Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, 2. Institute for Health Research, Kaiser Permanente Colorado, Denver, CO; School of Public Health, University of Colorado, Aurora, CO, 3. Center for Biologics Evaluation and Research, FDA, Silver Spring, MD, 4. HMORN: HealthPartners Research Foundation, Bloomington, MN

April 27, 2016

Mini-Sentinel is a pilot project sponsored by the U.S. Food and Drug Administration (FDA) to inform and facilitate development of a fully operational active surveillance system, the Sentinel System, for monitoring the safety of FDA-regulated medical products. Mini-Sentinel is one piece of the Sentinel Initiative, a multi-faceted effort by the FDA to develop a national electronic system that will complement existing methods of safety surveillance. Mini-Sentinel Collaborators include Data and Academic Partners that provide access to health care data and ongoing scientific, technical, methodological, and organizational expertise. The Mini-Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223200910006I.

Mini-Sentinel Methods

Evaluation of Scan Statistics for Assessing Vaccine Safety in Pregnancy

Table of Contents

I. EXECUTIVE SUMMARY

Pregnancy vaccine safety is particularly challenging to study because the pathophysiology of maternal pregnancy outcomes is not well understood, making it difficult to determine which risk intervals should be used. Pregnancy outcomes may be affected by timing since vaccination, gestational age at vaccination, or both. While scan statistics have been used in a wide array of application areas to identify unusual clusters of events, they have had limited use in the context of vaccine safety during pregnancy. Furthermore, their performance in this setting has not been systematically evaluated. To address this knowledge gap, we conducted a comprehensive simulation study to examine the performance of the one-dimensional temporal scan statistic and a two-dimensional time-time scan statistic in the context of the MS-PRISM Influenza Vaccines and Pregnancy Outcomes study (the empirical study) with a case-time-control design. The one-dimensional temporal scan statistic can be viewed as a special case of the spatial scan statistic or a special case of the space-time scan statistic. The two-dimensional time-time scan statistic is a novel application of the space-time scan statistic.

The simulation study was designed to examine the association between influenza vaccination and an adverse pregnancy outcome spontaneous abortion (SAB) while mimicking the MS-PRISM Influenza Vaccines and Pregnancy Outcomes study. With a case-time-control design, we use the scan statistics to identify unusual clusters of exposed cases compared to the expected counts estimated based on the matched controls. In this study, T_1 denotes the gestational age at vaccination defined in terms of the number of days away from the last menstrual period (LMP, day 0) and T_2 denotes the number of days between vaccination time and the index date which is defined as the case onset date. We considered seven risk regions, i.e., $R_1 \equiv \{-28 \leq T_1 \leq 35\}$, $R_2 \equiv \{15 \leq T_1 \leq 49\}$, $R_3 \equiv \{43 \leq T_1 \leq 84\}$, $R_4 \equiv \{1 \leq T_2 \leq 28\}$, $R_5 \equiv \{-28 \leq T_1 \leq 35 \;\&\; 1 \leq T_2 \leq 28\}$, $R_6 \equiv \{15 \leq T_1 \leq 49 \;\&\; 1 \leq T_2 \leq 28\}$, and $R_7 \equiv \{43 \leq T_1 \leq 84 \;\&\; 1 \leq T_2 \leq 28\}$. For each risk region, we considered 25 settings with varying sample size and effect size. We considered multiple performance metrics including sensitivity, specificity, and bias and 95% CIs coverage probability and width of the effect measure estimator. We also proposed a new performance metric to assess the distance between the empirical distribution of D which is defined as the sum of sensitivity and specificity, and its "null distribution".

We have successfully demonstrated the use and performance of the scan statistics in examining the association of vaccination and adverse pregnancy outcomes. Performance always improves with sample size and/or effect size. In general, the one-dimensional temporal scan statistic has better or similar performance in detecting the T_2-based risk region R_4 than the three T_1-based risk regions. Among R_1, R_2, and R_3, R_2 tends to perform better than R_1 and R_3 whose performance comparison varies across settings and performance metrics. Among the three two-dimensional risk regions, R_6 and R_7 have comparable performance and are better than R_5. It seems that other than the key factors of sample size and effect size, the scan statistics perform better when the true risk region overlaps greatly with the plausible region and also has high baseline rates inside the risk region. In other words, a larger proportion of exposure-attributed cases and a bigger contrast between cases and controls would allow the scan statistics to better detect the true risk region.

Based on our findings, we conclude that these statistics could potentially be of great use for detecting increased risks of pregnancy adverse events for which the pathophysiology is not well understood, making it difficult to appropriately define risk intervals. The scan statistics do require a reasonably large sample size to be able to detect the risk region with good accuracy. For instance, with an odds ratio of 3, to achieve an 80% sensitivity, we need between 50 and 100 matched pairs for the one-dimensional risk regions and between 200 and 500 matched pairs for the two-dimensional risk regions R_6 and R_7. An even larger sample size is required for the risk region R_5. The number of SAB cases chart reviewed in the PRISM SAB protocol is limited, so the power of the scan statistics may be limited in that setting. However, the temporal scan could theoretically be better powered in future analysis of larger sample sizes—for example one that uses the temporal scan initially on automated data only to screen for potential risk intervals, and proceeds to chart review of outcomes and gestational age based on those data; or alternatively another independent study to confirm these results. Overall, this project is useful in i) demonstrating the feasibility of the two-dimensional time-time scan statistic as a novel application of the space-time scan statistic in the setting of vaccine safety during pregnancy, and ii) systematically examining the performance of the one-dimensional temporal scan statistic and the two-dimensional time-time scan statistic in various settings to guide future implementation.

II. INTRODUCTION

The use of a scan statistic to identify unusual clusters of events dates back to the 1960s when Joseph Naus[1] first published on the problem. Martin Kulldorff extended it to multi-dimensional settings and varying scanning window sizes in his 1997 seminal paper[2]. Since then, Martin Kulldorff and others have developed different variants of scan statistics to handle various types of outcomes and data settings[2-7]. The scan statistics have been implemented in numerous medical and public health studies. Please refer to the SaTScan website (http://www.satscan.org/) and manual[8] for an updated list of SaTScan Bibliography.

The scan statistics share a common basic idea that is to scan across space and/or time with a scanning window of varying shape and/or size to find a region with excess of observed events based on a pre-specified test statistic (e.g., the maximum likelihood ratio between the alternative hypothesis and the null hypothesis). Analytic formulas typically do not exist for the variance of the test statistics. Instead, a Monte Carlo simulation approach is used to obtain the significance level[2].

The one-dimensional temporal scan statistic can be viewed as a special case of the spatial scan statistic or a special case of the space-time scan statistic[9]. The two-dimensional time-time scan statistic is a novel application of the space-time scan statistic. The one-dimensional temporal scan has been used in various Vaccine Safety DataLink (VSD) and Mini-Sentinel Post Licensure Rapid Immunization Safety Monitoring Program (MS-PRISM) projects to identify unusual adverse event clusters following vaccination. To our knowledge, the two-dimensional temporal scan has not been used previously in post-marketing safety surveillance studies.

One area for which there has been limited application of scan statistics, whether one-dimensional or two-dimensional, is in the area of vaccine safety during pregnancy. Pregnancy adverse events are particularly challenging to study because their pathophysiology is not well understood in regards to time periods relative to exposure that are relevant for vaccine-related adverse events. Specifically, the risk of

specific pregnancy adverse events could hypothetically be elevated following vaccination received at specific gestational periods, specific time periods following vaccination, or both. Historically, many studies examining risk of adverse pregnancy outcomes following prenatal vaccination have not assumed that the risk is limited to specific time periods with respect to vaccination, and have thus included all person-time following vaccination during which the event could occur (e.g., 20 weeks gestation for SAB). Others have assumed that the effect of vaccines is limited to a specific period following vaccination; or alternatively that it is limited to periods following vaccines received at a specific time window defined by gestational age[10,11]. In such observational studies, the risk interval (i.e., period of potentially increased risk) must be defined before conducting the analysis. However, mis-specifying the risk interval, either with respect to the length or placement, could hypothetically cause relative risk estimates to be biased by washing out an effect if it exists. Yet forgoing defining risk intervals (e.g., examining all person-time following vaccination before 20 weeks gestation, in the instance of SAB) could wash out an effect if the risk is truly limited to a specific time period. Thus, temporal scan techniques may be useful in identifying appropriate risk intervals for the study of pregnancy adverse events following vaccination during pregnancy. In particular, we plan to use both one-dimensional and two-dimensional temporal scans in the MS-PRISM Influenza Vaccines and Pregnancy Outcomes study. While the one-dimensional temporal scan has been used in various VSD and MS-PRISM post-marketing safety studies, no study has used the two-dimensional temporal scan, and a systematic evaluation of their performance was not conducted previously. To fill in this knowledge gap, we conducted this simulation study to assess the performance of the scan statistics in plausible settings that we may encounter in vaccine-related safety studies for pregnancy outcomes such as SAB.

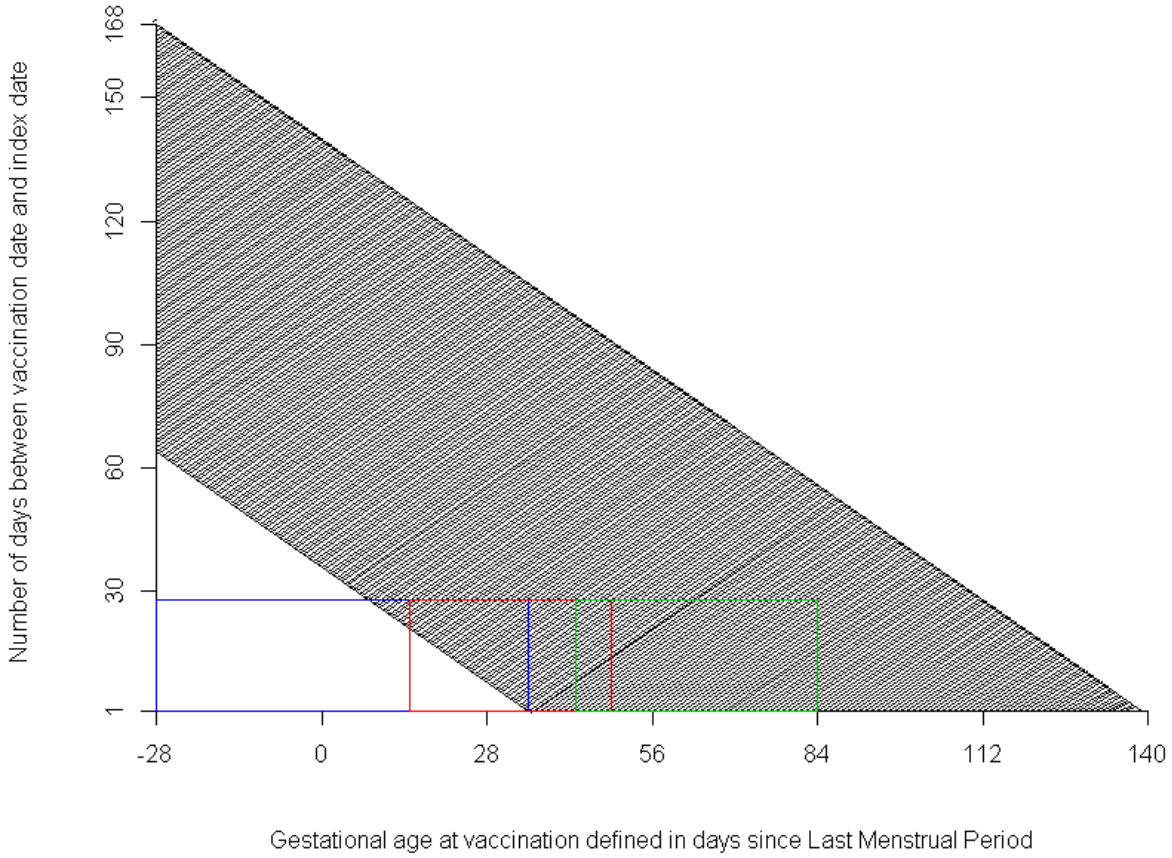
We designed our simulation study mimicking the study design and data structure in the MS-PRISM Influenza Vaccines and Pregnancy Outcomes study (referred to as the empirical study in subsequent sections). The main objective of that proof-of-concept protocol is to develop the data infrastructure, study design framework, and analytic framework to study pregnancy outcome following vaccination. As described in the protocol, the project aims to achieve these objectives using a single vaccine-outcome pair as an example, influenza vaccines and SAB. The empirical study uses the case-time-control design[12,13], which is a variant of the case-crossover study and is well-suited to measuring transient effects of exposures on immediate risks of illness with abrupt onset. We introduce the case-time-control design in details in the Methods section. Intuitively, the design uses vaccinated cases and controls and compares their likelihood of being exposed during a particular period of interest; if the likelihood is much higher among cases than among controls, then it indicates a possible association between the exposure and the adverse event. Therefore in this simulation study, a risk region is defined in terms of exposure time and refers to a particular period of interest with an excess of exposed cases compared to the expected counts under the null hypothesis, for instance, 1 to 28 days prior to the case onset date or 6 to 11 weeks gestation. This is different from the conventional definition of a risk interval which typically applies to a cohort design among all exposed subjects, is defined in terms of adverse event onset time, and refers to a particular time period with an excess of observed events (e.g., 1 to 28 days following exposure).

In this simulation study, we consider three sets of settings which, respectively, correspond to three scenarios with elevated SAB risks following influenza vaccination received at specific gestational periods, specific time periods following vaccination, or both. Specifically, let T_1 denote the gestational age at vaccination defined in terms of the number of days away from the last menstrual period (LMP, day 0) and let T_2 denote the number of days between vaccination time and the index date which is defined as

the case onset date. As illustrated in Figure 1 below, we use the x-axis to denote T_1 and y-axis to denote T_2, then the seven considered risk regions can be denoted as $R_1 \equiv \{-28 \leq T_1 \leq 35\}$, $R_2 \equiv \{15 \leq T_1 \leq 49\}$, $R_3 \equiv \{43 \leq T_1 \leq 84\}$, $R_4 \equiv \{1 \leq T_2 \leq 28\}$, $R_5 \equiv \{-28 \leq T_1 \leq 35 \,\&\, 1 \leq T_2 \leq 28\}$, $R_6 \equiv \{15 \leq T_1 \leq 49 \,\&\, 1 \leq T_2 \leq 28\}$, and $R_7 \equiv \{43 \leq T_1 \leq 84 \,\&\, 1 \leq T_2 \leq 28\}$. For instance, the risk region R_1 means an excess of SAB cases being vaccinated during this gestational period of -28 to 35 days; the risk region R_4 means an excess of SAB cases being vaccinated within the prior 28 days before index date; the two-dimensional risk region R_5 means an excess of SAB cases being vaccinated during the gestational period of -28 to 35 days and within the prior 28 days before index date because the SAB risks are elevated during the first 28 days following vaccination if and only if vaccination was received during this particular gestational period of -28 to 35 days.

The shaded area in Figure 1 is the plausible region $\{(T_1, T_2): 36 \leq T_1 + T_2 \leq 140\}$. We specify such a plausible region to mimic the empirical study. Note that the sum $T_1 + T_2$ denotes the gestational age at SAB onset date. The lower bound is imposed to mimic the exclusion criteria in the empirical study excluding SABs with a very early onset date due to concerns on incomplete capture of earlier cases in the MS-PRISM Distributed Databases and the role of genetic anomalies in early SAB cases. The upper bound is imposed to reflect that SAB occurs before 20 weeks gestation by definition.

Figure 1. Plausible range for (T_1, T_2) and three potential risk regions R_5, R_6, and R_7 in blue, red, and green respectively



In the simulation study, we apply the scan statistics to detect the most likely cluster of vaccination time among cases compared to controls. To assess the performance of the scan statistics, we examine the overlap between the detected cluster and the true risk region. Moreover, we use the detected cluster and conditional logistic regression to calculate an odds ratio (OR) estimator. The OR indicates the magnitude of the association between the exposure and the adverse event. We examine the performance on bias, variance, and the coverage probability of the 95% confidence intervals (CIs).

III. METHODS

In the following subsections, we first introduce the case-time-control design, then we introduce the one-dimensional temporal scan statistic and the two-dimensional time-time scan statistic.

A. THE CASE-TIME-CONTROL STUDY DESIGN

The case-time-control design is an extension of the case-crossover study[14] which is a self-controlled analogue of the case control design. Specifically, in a case-crossover study, the study population consists of patients who met the inclusion and exclusion criteria and experienced an eligible adverse event of interest. A patient serves as his/her own control comparing exposure status in a risk region and a control region. If the likelihood of being exposed in the risk region is significantly higher than in the control region accounting for region size, then it suggests a possibly elevated adverse event risk due to the exposure. A risk region is typically immediately preceding case onset date and a control region is usually before the risk region with or without a wash-out period.

The case-crossover design implicitly adjusts for time-invariant factors such as patient characteristics and genetic factors, but it does not adjust for time-varying factors that could affect the timing of exposure (e.g., seasonality). The case-time-control design was proposed to address this issue by using matched controls. Specifically, for each case, one or several controls are selected based on pre-specified matching criteria (e.g., LMP, age). Each control also contributes a risk region and a control region. The odds of being exposed inside vs. outside the risk region among the matched control(s) represent the baseline odds under the null hypothesis of no exposure-outcome association. If the odds of being exposed inside vs. outside the risk region among cases are systematically higher than the odds among their matched controls, it suggests a possibly elevated risk of adverse event due to the exposure. It is worth noting that the case-time-control design, unlike the standard case control design, does not require cases and controls to have similar outcome risk profiles. Instead, it only requires the cases and controls to have similar exposure trends to account for time-varying factors.

In the empirical study, a case is a pregnant woman whose pregnancy ended with a SAB and was vaccinated between -4 weeks gestation and SAB onset. The matched controls are live births and were vaccinated between -4 weeks gestation and the index date which is defined as the matched case's SAB onset date. We matched controls to cases on LMP, Data Partner, and age within +/- 6 months. Potential risk regions are -4 to 4 weeks gestation, 2 to 6 weeks gestation, 6 to 11 weeks gestation, and 1 to 28 days prior to index date. The corresponding risk regions defined in the simulation study are $R_1, R_2, R_3,$ and R_4 respectively. In the first set of analyses, we examine each of the four potential risk regions as pre-specified risk regions. Specifically, we fit a conditional logistic regression to the matched cases and controls. The dependent variable is whether the woman was vaccinated inside vs. outside the specified risk region. The independent variable is the case status (1 for case and 0 for control). We condition on matched sets to account for non-uniform exposure trend across different matched sets. The exponent of the coefficient for the case status is the estimated OR for being vaccinated inside vs. outside the pre-specified risk region between cases and controls.

As described earlier, mis-specifying the risk region, which occurs because the pathophysiology of many pregnancy adverse events is not well understood, may cause bias. Thus in the empirical study, in addition to the conditional logistic regression analyses with pre-specified risk regions, we will also conduct exploratory analyses by applying the one-dimensional and two-dimensional scan statistics to the data on matched sets to identify potential clusters. The findings from this project help us better understand the utility and power of the scan statistics in the empirical study.

B. THE TEMPORAL SCAN STATISTIC

In this section, we describe the statistical framework behind the one- and two-dimensional scan statistics to detect potential clusters of vaccination with respect to the number of days before SAB, gestational age at vaccination, or both. The one-dimensional temporal scan statistic applies to either Poisson or Bernoulli models. Here we consider the temporal scan with the Bernoulli model as we use matched controls to adjust for non-uniform baseline trend of vaccination with gestational age and seasonality. It can be viewed as a special case of the spatial scan statistic with the Bernoulli model. Next, we use the risk region of 1 to 28 days prior to index date as an example to illustrate the temporal scan method. It applies to other one-dimensional risk regions too.

Suppose there are N cases and M controls matched on LMP and other characteristics and all vaccinated between -4 weeks gestation and index date. We use a scanning window $[s, l]$ with varying starting point and length along the observation period 1-168 days. The upper bound 168 is determined by the maximum possible length between vaccination date and SAB onset date. For a scanning window$[s, l]$, suppose there are n cases and m controls within the scanning window, we define a function $\lambda(s, l)$ such that

$$\lambda(s,l) \equiv \left(\frac{n}{m+n}\right)^{n}\left(\frac{m}{m+n}\right)^{m}\left(\frac{N-n}{N+M-m-n}\right)^{N-n}\left(\frac{M-m}{N+M-m-n}\right)^{M-m}$$

if $\frac{n}{m+n} > \frac{N-n}{N+M-m-n}$, and $\lambda(s, l) = 1$ otherwise. The function $\lambda(s, l)$ is derived based on the maximized likelihood ratio between the alternative hypothesis of $OR > 1$ and the null hypothesis of $OR = 1$. The statistic $\lambda(s, l)$ is evaluated for all possible scanning windows, the window $[s^*, l^*]$ with the maximum $\lambda(s, l)$ value is the detected mostly likely cluster.

C. THE TIME-TIME SCAN STATISTIC

The two-dimensional time-time scan statistic is a novel application of the space-time scan statistic. In the space-time scan statistic, a potential risk region is a cylinder with the base indicating a circular geographic area and the height indicating a time period. For instance, an unusually high number of influenza-alike cases occurred within a 5-mile radius from a particular hospital in November 2015. In this application, we use the latitude in spatial scale to indicate gestational age at vaccination (T_1) while setting the longitude to a constant, and use the time scale in the space-time scan statistic to indicate the time between vaccination date and index date (T_2). In consequence, a potential region reduces from a cylinder to a rectangle to accommodate the two-dimensional time-time nature of a potential risk region in this particular application. For instance, if the risk of SAB is elevated during 1 to 28 days post influenza vaccination among those who received vaccination during 2 to 6 weeks gestation, the corresponding risk region is a rectangle of 2 to 6 weeks gestation times 1 to 28 days prior to index date. The likelihood ratio-based test statistic to identify the most likely cluster is the same as in the one-dimensional temporal scan statistic. The restrictions $36 \leq T_1 + T_2 \leq 140$ do not affect the use of the scan statistic as the non-uniform baseline rates of vaccination are accounted for with the matched controls. However, they need to be accounted for in the calculations of sensitivity, specificity and other performance measures which we explain in details later on.

IV. SIMULATION STUDY

A. DATA GENERATING MECHANISM

We designed our simulation study mimicking the empirical study. The study design parameters were imposed based on literature and preliminary data from the empirical study.

Specifically, for each pregnant woman

- The vaccination time (V_i), measured in number of days after LMP, is simulated from a multinomial distribution between -28 and 139 with the daily probability being 0.005509176 for $t \leq 84$ and 0.006862966 for $84 \leq t < 140$. Only SAB cases that occurred after vaccination were considered as in the empirical study because by the case-time-control design all eligible cases need to be exposed prior to case onset date.
- A woman is at risk for SAB following vaccination or beginning of week 6, whichever is later, until 20 weeks gestation, i.e., on day $t = \max[V_i + 1, 36]$, …, 140. Baseline SAB rates $r_0(t)$ were derived from the published weekly rates in the Li study [15](eTable 1). For simplicity, let's set $r_0(t)$ to 0 for $t < 36$. As mentioned earlier, in real life, SAB cases occur during the early gestational period prior to 6 weeks gestation. We do not consider them in the simulation study to mimic the exclusion of early SAB cases in the empirical study due to concerns on incomplete data and the role of genetic anomalies in early SAB cases.
- If day t is inside a pre-specified risk region, then the SAB rate is elevated by a pre-specified RR to $RR \times r_0(t)$; otherwise, it remains at the baseline risk level.
    - Suppose the risk region is $R_2 = \{15 \leq T_1 \leq 49\}$ and $RR = 2$. If a woman was vaccinated on day $V_i = 60$, then her daily SAB rate remains at $r_0(t)$ for $t = 61, \ldots 140$. If a woman was vaccinated on day $V_i = 25$, then her daily SAB rate is elevated to $2r_0(t)$ for $t = 26, \ldots, 140$.
    - Suppose the risk region is $R_4 = \{1 \leq T_2 \leq 28\}$ and $RR = 2$, then the daily SAB rate is elevated to $2r_0(t)$ for $t = V_i + 1, \ldots V_i + 28$ and comes back to $r_0(t)$ for $t = V_i + 29, \ldots 140$.
    - Suppose the risk region is $R_7 = \{43 \leq T_1 \leq 84 \;\&\; 1 \leq T_2 \leq 28\}$ and $RR = 2$. If a woman was vaccinated on day $V_i = 35$, her daily SAB rate remains at the baseline rate $r_0(t)$ for $36 \leq t \leq 140$ as her vaccination time is outside the risk region. If she was vaccinated on day $V_i = 45$, then her daily SAB rate is elevated to $2r_0(t)$ for $t = 46, \ldots, 73$ (1 to 28 days post vaccination) and then comes back to the baseline rate of $r_0(t)$ for $74 \leq t \leq 140$. Of note, this scenario is generalizable to all scenarios that consider multiplicative effect of two risk factors.
- For $t = V_i + 1$, …, 140, a woman's SAB status (1 for SAB onset and 0 for no SAB) is simulated from a Bernoulli distribution with the probability being the daily rate determined in the previous step, until either SAB occurs or the maximum $t = 140$ is reached.
- If a SAB occurred, this woman is a case; otherwise this woman is placed into the pool of controls.
- For each case with an index date of W, a matched control is selected among those who did not have SAB and were vaccinated before gestational age W (the case's index date). For simplicity, in the simulation study, we did not consider population heterogeneity across Data Partners or with maternal age, thus matching on Data Partner and maternal age was not necessary.

It is worth noting that under this data generating mechanism, the odds ratio of being vaccinated inside vs. outside the risk region between a case and her matched control varies with the index date and may differ from the true RR value unless it is under the null hypothesis of $RR = 1$. The earlier the index date W, the bigger the difference. This is because our controls are pregnant women who did not have SAB throughout 20 weeks gestation. Under an alternative hypothesis of $RR > 1$, a pregnant woman who did not develop SAB throughout 20 weeks gestation has a smaller likelihood of being vaccinated inside the risk region compared to another pregnant woman who did not have SAB by gestational age W but may have SAB later on. This simulation study is designed mimicking the empirical study in which we did not consider an alternative strategy of selecting a control from those who did not have SAB by gestational age W but may end up being a non-live birth later on (SAB or still birth), as this alternative strategy is very difficult to operationalize using electronic health care data such as the Mini-Sentinel Distributed Database. Specifically, it is not possible to ascertain pregnancy start dates or gestational age for non-live births using ICD-9 coded data. Thus, for pregnancies ending in a spontaneous abortion, it is not possible to ascertain LMP date without chart review. If we allow non-live birth in the pool of potential controls and select controls based on the electronically-derived pregnancy start dates, it is possible that after chart review, a fair number of matched pairs would need to be excluded from subsequent analyses due to pregnancy start matching criteria no longer being met. We usually have limited resources to conduct chart review for a fixed number of individuals. Therefore we want to maximize the chance of retaining as many matched sets as possible by restricting to live births as potential controls. Please refer to the protocol[16] for a detailed explanation on why we used a multi-phase procedure to select and chart review cases and matched controls (Figure 1 in the protocol). However, theoretically speaking, a two-phase chart review approach may not be necessary, depending on changes in coding with the ICD-10.

B. CLUSTER DETECTION AND EFFECT ESTIMATION

In each considered setting, we conduct 2000 simulation replications to assess the performance of the scan statistics. For each replication, we generate a dataset with $2N$ cases and $2N$ matched controls. We then randomly and equally split the $2N$ matched pairs to two datasets, apply the scan statistic to one dataset to detect a most likely cluster, and then apply the conditional logistic regression with the dependent variable defined based on the detected cluster to the other dataset to estimate the association between exposure and outcome. Let $\hat{\theta}$ denote the estimated coefficient for the indicator variable for case status. Let θ^* denote the value $\hat{\theta}$ would converge to with a perfectly detected true cluster and $N \to \infty$. The exponent of θ^* refers to the ratio between the odds of being vaccinated inside vs. outside the risk region between cases and controls. Due to the use of live-births as potential controls, the OR differs from the RR in the data generating algorithm unless it is under the null hypothesis such that $RR = OR = 1.0$.

In the simulation study we consider seven risk regions: $R_1 = \{-28 \le T_1 \le 35\}$, $R_2 = \{15 \le T_1 \le 49\}$, $R_3 = \{43 \le T_1 \le 84\}$, $R_4 = \{1 \le T_2 \le 28\}$, $R_5 = \{-28 \le T_1 \le 35 \;\&\; 1 \le T_2 \le 28\}$, $R_6 = \{15 \le T_1 \le 49 \;\&\; 1 \le T_2 \le 28\}$, and $R_7 = \{43 \le T_1 \le 84 \;\&\; 1 \le T_2 \le 28\}$. Any time prior to the index date but outside the risk region is considered the control region. For each risk region, we vary the number of matched pairs between 25 and 500, and then $\exp(\theta^*)$ between 1.0 and 10.0.

When we use the one-dimensional temporal scan to detect the risk regions R_1, R_2, and R_3, we set the observation period to [-28,140]. We consider all possible scanning windows with varying starting point and length less than half of the observation period. When we use the one-dimensional temporal scan to

detect the risk region R_4, we set the observation period to [1,168]. Similarly, we consider all possible scanning windows with varying starting point and length less than half of the observation period. When we use the two-dimensional time-time scan statistic to detect the risk regions R_5, R_6, and R_7, we use the Bernoulli model-based spatial-time scan statistic module of the SaTScan software. We use the latitude of the spatial cluster to denote gestational age at vaccination (T_1), set the longitude to a constant, and use the time scale to denote the number of days between vaccination time and SAB onset date (T_2). The maximum spatial cluster size is 50% of the study population. The maximum time length is half of the observation period [1,168]. All analyses were done with SaTScan version 9.3 (www.satscan.org) and SAS 9.3 (SAS Institute, Cary, NC).

C. PERFORMANCE METRICS

We measured the performance of the scan statistics from two aspects, cluster detection (i.e., how much the detected cluster overlaps with the true risk region) and performance assessment on effect estimation (e.g., bias, coverage probability of the 95% CI). The latter does not directly measure the performance of the scan statistics but measures the performance of the detected risk region on risk estimation, the ultimate goal of the empirical study.

To assess the performance on cluster detection, whenever applicable, we examine the sensitivity and specificity in the two-dimensional time-time scale, the sensitivity and specificity for each of the two time scales, as well as a newly proposed performance metric p_G which we define below. The one-dimensional sensitivity for either the T_1 or T_2 scale is defined as the proportion of days in the true risk interval that is included in the detected cluster. Similarly, the one-dimensional specificity for either the T_1 or T_2 scale is defined as the proportion of days outside the risk interval that is outside the detected cluster. For instance, suppose the true risk region is [43, 84]x[1, 28] and the detected cluster is [41, 90]x[3, 30], then the sensitivity and specificity for the T_1 scale are 42/42=1 and (168-42-8)/(168-42)=0.94 respectively, while the sensitivity and specificity for the T_2 scale are 26/28=0.93 and (168-28-2)/(168-28)=0.99 respectively. In the two-dimensional time-time scale, sensitivity is defined as the proportion of (T_1, T_2) coordinates inside the intersection of the shaded area and the true risk region ("true positives") that are included in the detected cluster, while specificity is defined as the proportion of (T_1, T_2) coordinates inside the shaded area but outside the true risk region ("true negatives") that are not included in the detected cluster. With the same example of the true risk region being [43, 84]x[1, 28] and the detected cluster being [41, 90]x[3, 30], the sensitivity and specificity equal $(42 \times (28-2))/(42 \times 28) = 0.93$ and $\left(1 - \frac{(90-40)*(30-28)+2*(28-2)+(90-84)*(28-2)}{168*169/2+63*64/2}\right) = 0.98.$

We let G denote the sum of sensitivity and specificity. Thus, $0 < G \leq 2$. When the detected cluster overlaps perfectly with the true risk region, $G = 2$, otherwise $G < 2$. If the detected cluster overlaps poorly with the true risk region and takes in most of the control region instead, then G may be close to 0. In this simulation study, because we restrict the length of the scanning window to less than half of the observation period, then G is bounded away from 0. In each setting, we obtain 2000 G values with 2000 simulated datasets, and create a histogram by partitioning the interval (0,2] into 100 equally sized bins and calculate the proportion of G values falling into each bin. We also create a "null distribution" histogram by calculating the G values for all possible clusters with the restrictions on maximum cluster size and calculating the proportions of G values falling into each of the 100 bins among the G values from all possible clusters. Then the performance metric p_G is defined as the sum of the minimums of the two bin-specific proportions across all bins. In other words, p_G denotes the overlap between the setting-

specific histogram and "null distribution" histogram. The smaller the p_G value is, the lesser the overlap, and the better the performance of the scan statistic. We proposed p_G as a summary measure of both sensitivity and specificity. Moreover, p_G measures the overlap between the empirical distribution of G and its "null distribution" in an effort to reduce the dependency of this risk metric on specific settings. In our simulation study, as the considered risk regions are small relative to the entire plausible region, specificity tends to be close to 1, in consequence, p_G tends to be mainly affected by sensitivity. We expect p_G to have a greater utility in other settings with medium to large variations in both sensitivity and specificity.

V. RESULTS

We present the simulation results in Tables 1-14 for seven risk regions: $R_1 = \{-28 \leq T_1 \leq 35\}$, $R_2 = \{15 \leq T_1 \leq 49\}$, $R_3 = \{43 \leq T_1 \leq 84\}$, $R_4 = \{1 \leq T_2 \leq 28\}$, $R_5 = \{-28 \leq T_1 \leq 35 \;\&\; 1 \leq T_2 \leq 28\}$, $R_6 = \{15 \leq T_1 \leq 49 \;\&\; 1 \leq T_2 \leq 28\}$, and $R_7 = \{43 \leq T_1 \leq 84 \;\&\; 1 \leq T_2 \leq 28\}$. For each risk region, we present in two tables the simulation results in 25 settings with varying sample size, measured in the number of matched pairs N, and varying effect size, measured in θ^*, the log odds ratio of being vaccinated inside vs. outside the risk region between cases and controls. The first table provides results on performance assessment on cluster detection, while the second table provides results on performance assessment on effect estimation. Specifically, in Tables 1, 3, 5, and 7, we present medians and robust standard errors for both sensitivity and specificity and the p_G value for the 4 one-dimensional risk regions. The robust standard error is a robust analogue of the conventional standard error and is defined as the inter-quartile range divided by 1.345, the inter-quartile range for a standard normal distribution. The robust standard error equals the conventional standard error if the statistic is normally distributed[17] but is less sensitive to outliers. In Tables 9, 11, and 13, we present medians and robust standard errors for sensitivity and specificity in the two-dimensional scale, as well as the sensitivity and specificity for T_1 and T_2 scales respectively. In Tables 2, 4, 6, 8, 10, 12, and 14, we present medians and robust standard errors for the bias of $\hat{\theta}$. We also present the relative bias which is defined as the median bias divided by θ^* when $\theta^* > 1$. In addition, we present coverage probabilities and median width for the 95% CIs of $\hat{\theta}$.

In all settings, as expected, performance improves with larger N and/or larger θ^*. Specificity is reasonably close to 1 in most settings. This is likely because the risk region is small relative to the control region and we restrict the scanning window size by less than half of the observation period. In contrast, the performance on sensitivity is less satisfactory. Consider the four settings with $N = 25 \;\&\; \theta^* = \log(10)$, $N = 50 \;\&\; \theta^* = \log(5)$, $N = 100 \;\&\; \theta^* = \log(3)$, and $N = 200 \;\&\; \theta^* = \log(2)$, the sensitivity for detecting the risk region R_1 equals 81.3%, 84.4%, 82.8%, and 65.6% respectively; the sensitivity for detecting the risk region R_2 equals 88.6%, 91.4%, 94.3%, and 88.6% respectively; the sensitivity for detecting the risk region R_3 equals 78.6%, 78.6%, 69.0%, and 54.8% respectively; the sensitivity for detecting the risk region R_4 equals 89.3%, 89.3%, 89.3%, and 82.2% respectively. Among the four one-dimensional risk regions, it appears the one-dimensional temporal scan statistic has better performance on sensitivity, bias and 95% CIs coverage in detecting risk regions R_2 and R_4. They have comparable performance on bias and CIs coverage, R_4 has a slightly better performance on specificity while R_2 has a slightly better performance on sensitivity. The risk regions R_1 and R_3 have worse performance compared to R_2 and R_4. For R_1, it is likely because we exclude early onset SAB cases such that the impact on SAB risk is exactly the same if a woman was vaccinated on day -28 versus on day 35. For R_3, it

is likely because the baseline SAB risk decreases dramatically after 13 weeks gestation, and thus when the total number of cases is fixed, a later risk region means a smaller proportion of cases are attributable to the exposure of interest. In consequence, it is more difficult to accurately detect the true cluster as the effective sample size is smaller. The scan statistic is expected to have better performance when the contrast between cases and controls is bigger. Therefore it makes sense that R_2 has either better or comparable performance compared to the other three T_1-based risk regions with SAB risks elevated for all person-time following vaccination before 20 weeks gestation. Between R_1 and R_3, R_3 in general has better performance on bias and CIs coverage but the performance comparison on sensitivity varies across settings. For the performance metric p_G, among the three T_1-based risk regions, R_1 has a slightly better performance than R_2 which in turn has better performance than R_3. Here R_1 has a slightly better performance than R_2 despite a worse performance on sensitivity likely because it has better performance on specificity. The T_2-based risk region R_4 has better performance than the T_1-based risk regions.

The two-dimensional time-time statistic does not perform as well as the one-dimensional temporal scan statistic which is not surprising because the number of scanning windows increases dramatically with the addition of a second time scale. Moreover, it is more difficult to accurately detect a two-dimensional cluster as it has an additional degree of freedom compared to a one-dimensional cluster. With the two-dimensional time-time statistic, we observe a similar trend that performance improves with larger N and/or larger θ^* although the improvement is smaller than what we observed with the one-dimensional temporal scan statistic. For instance, consider the same four settings with $N = 25$ & $\theta^* = \log(10)$, $N = 50$ & $\theta^* = \log(5)$, $N = 100$ & $\theta^* = \log(3)$, and $N = 200$ & $\theta^* = \log(2)$, the two-dimensional sensitivity equals 51.2%, 38.9%, 19.2% and 0.0% for detecting the risk region R_5, equals 59.4%, 46.3%, 28.7%, and 13.2% for detecting the risk region R_6, and equals 57.1%, 47.6%, 32.7%, and 15.5% for detecting the risk region R_7. We fix $\theta^* = \log(5)$ and increase N from 25 to 500, the sensitivity for detecting the three risks regions equals (35.2%, 38.9%, 41.4%, 64.5%, 88.9%) for R_5, (39.0%, 46.3%, 72.3%, 88.1%, 96.0%) for R_6, and (37.2%, 47.6%, 72.3%, 88.1%, 97.6%) for R_7; if we fix the $\theta^* = \log(10)$, the sensitivity equals (51.2%, 61.2%, 74.4%, 84.2%, 96.3%) for R_5, (59.4%, 78.6%, 89.9%, 95.5%, 99.0%) for R_6, and (57.1%, 76.7%, 91.8%, 96.4%, 100.0%) for R_7. Among the three two-dimensional risk regions, all performance metrics provide consistent inference. Specifically, the two-dimensional time-time scan statistic has the worst performance in detecting R_5 as it has little overlap with the plausible region due to the design of excluding SAB cases with early onset. Between R_6 and R_7, R_6 has a slightly better performance with small to medium sample size ($N < 100$) while R_7 has a slightly better performance when there are 100 or more matched pairs. This is different from what we observed with the one-dimensional scan statistic that the risk region $R_3 = \{43 \leq T_1 \leq 84\}$, which corresponds to $R_7 = \{43 \leq T_1 \leq 84 \ \& \ 1 \leq T_2 \leq 28\}$, tends to have worse performance compared to other one-dimensional risk regions. It is likely because with the two-dimensional risk regions, pregnant women have increased SAB risks only during the first 28 days if vaccinated during the corresponding gestational period. Therefore, despite higher baseline SAB rates in early gestational weeks, the smaller overlap with the plausible region decreases the performance of the risk regions R_5 and R_6. Across all three two-dimensional risk regions, the two-dimensional time-time scan statistic has better performance in the T_2 scale than in the T_1 scale, which is consistent with the performance of the one-dimensional temporal scan statistic. This means the scan statistics are better at detecting a cluster defined in terms of the number of days between vaccination time and event date than detecting a cluster defined in terms of gestational age at vaccination.

The 95% CIs have coverage probabilities below nominal level unless it is under the null hypothesis of $\theta^* = 0$. This is because we used the detected cluster to define the exposure variable for cases and controls. Unless the detected cluster overlaps perfectly with the true risk region, the estimator $\hat{\theta}$ under-estimates θ^*. Moreover, the 95% CIs were constructed using the standard errors from the conditional logistic regression without taking into consideration the variation in cluster detection.

VI. DISCUSSION

We conducted a comprehensive simulation study to examine the performance of the one-dimensional temporal scan statistic and the two-dimensional time-time scan statistic in the context of the MS-PRISM Influenza Vaccines and Pregnancy Outcomes study with a case-time-control design. The simulation study was designed to examine the association between influenza vaccination and an adverse pregnancy outcome SAB while mimicking the empirical study. We considered seven risk regions: $R_1 = \{-28 \leq T_1 \leq 35\}$, $R_2 = \{15 \leq T_1 \leq 49\}$, $R_3 = \{43 \leq T_1 \leq 84\}$, $R_4 = \{1 \leq T_2 \leq 28\}$, $R_5 = \{-28 \leq T_1 \leq 35 \ \& \ 1 \leq T_2 \leq 28\}$, $R_6 = \{15 \leq T_1 \leq 49 \ \& \ 1 \leq T_2 \leq 28\}$, and $R_7 = \{43 \leq T_1 \leq 84 \ \& \ 1 \leq T_2 \leq 28\}$. For each risk region, we considered 25 settings with varying sample size and effect size.

We considered multiple performance metrics including sensitivity, specificity, p_G, bias and 95% CIs coverage probability and width of the effect measure estimator $\hat{\theta}$. Specificity and 95% CIs width do not differ much between the scan statistics and across risk regions. Both scan statistics have better performance with larger N and/or larger θ^*. In general, the one-dimensional temporal scan statistic has better or similar performance in detecting the T_2-based risk region R_4 than the three T_1-based risk regions. Among R_1, R_2, and R_3, R_2 tends to perform better than R_1 and R_3 whose performance comparison varies across settings and performance metrics. Among the three two-dimensional risk regions, R_6 and R_7 have comparable performance and are better than R_5. It seems that other than the key factors of sample size and effect size, the scan statistics perform better when the true risk region overlaps greatly with the plausible region and also overlaps greatly with the gestational period with high baseline incidence of SAB. In other words, a larger proportion of exposure-attributed cases and a bigger contrast between cases and controls would allow the scan statistics to better detect the true risk region. The overlap with the gestational period with high baseline incidence of SAB may seem counter-intuitive as higher baseline risk typically leads to less contrast. In our simulation experiments, we fix the total number of cases and the OR, therefore a higher baseline incidence means a larger number of exposure-attributed cases (i.e., a larger effective sample size).

In summary, this project has i) successfully demonstrated the feasibility of the two-dimensional time-time scan statistic as a novel application of the space-time scan statistic, and ii) systematically examined the performance of the one-dimensional temporal scan statistic and the two-dimensional time-time scan statistic in various settings to guide future implementation. These statistics could potentially be of great use for detecting increased risks of pregnancy adverse events for which the pathophysiology is not well understood, making it difficult to appropriately define risk intervals. Studying vaccine safety during pregnancy is particularly complex because the risk of a vaccine-related event could be affected by both temporal proximity to vaccination and gestational age at vaccination. The scan statistics do require a reasonably large sample size to be able to detect the risk region with good accuracy. For instance, with an $\theta^* = \log(3)$, to achieve an 80% sensitivity, we need between 50 and 100 matched pairs for the one-dimensional risk regions and between 200 and 500 matched pairs for the two-dimensional risk regions

R_6 and R_7. An even larger sample size is required for the risk region R_5. The number of SAB cases chart reviewed in the PRISM SAB protocol is limited, so the power of the scan statistics may be limited in that setting. However, the temporal scan could theoretically be better powered in future analysis of larger sample sizes—for example one that uses the temporal scan initially on automated data only to screen for potential risk intervals, and proceeds to chart review of outcomes and gestational age based on those data; or alternatively another independent study to confirm these results. We do not recommend relying solely on scan statistics to detect the risk region. Instead, the detected cluster from applying the scan statistics to automated data shall be used as supplementary information to existing evidence and clinical knowledge to help refine the specification of the risk region.

The simulation study has several limitations. First of all, we did not assess the impact of the study design (i.e., the use of live-births as potential controls) on the performance of scan statistics. Secondly, we did not assess the impact of control to case ratio. In a subset of settings we considered higher matching ratios such as 4 controls per case and observed limited gains in performance improvement with more controls. However, due to time constraints we did not systematically examine its impact across all settings. Thirdly, we did not develop a rule to calibrate the value of p_G, the smaller p_G the better, but how small is small enough? Fourthly, we did not consider population heterogeneity across Data Partners and with gestational age and other risk factors, and did not assess their impact on the performance of the scan statistics. Finally, we did not assess the robustness of the scan statistics when there is residual confounding such as cases and controls do not have the same baseline vaccination rates. These are all very interesting questions that this workgroup did not get to study due to time constraints. These are all very interesting questions that are not in the scope of this study. These are great topics for future methodology work.

VII. TABLES AND FIGURES

Table 1. Performance of the one-dimensional temporal scan statistic on cluster detection, with the true risk region $\boldsymbol{R_1} = \{-28 \leq T_1 \leq 35\}$

N	$\boldsymbol{\theta^*, \exp(\theta^*)}$	Sensitivity		Specificity		$\boldsymbol{p_G}$
		Median	Robust se	Median	Robust se	
25	0,1	0.109	0.116	1	0.057	0.544
25	0.69,2	0.156	0.232	1	0	0.435
25	1.10,3	0.219	0.429	1	0	0.399
25	1.61,5	0.406	0.533	1	0	0.374
25	2.30,10	0.813	0.44	1	0	0.346
50	0,1	0.063	0.081	1	0.043	0.487
50	0.69,2	0.125	0.278	1	0	0.42
50	1.10,3	0.305	0.591	1	0	0.356
50	1.61,5	0.844	0.486	1	0	0.332
50	2.30,10	0.938	0.116	1	0	0.2
100	0,1	0.047	0.069	1	0.036	0.444
100	0.69,2	0.141	0.533	1	0	0.386
100	1.10,3	0.828	0.568	1	0	0.32
100	1.61,5	0.961	0.093	1	0	0.172
100	2.30,10	0.984	0.035	1	0	0.054
200	0,1	0.016	0.046	1	0.029	0.432
200	0.69,2	0.656	0.637	1	0	0.372
200	1.10,3	0.969	0.081	1	0	0.165
200	1.61,5	0.984	0.035	1	0	0.049
200	2.30,10	1	0.012	1	0	0.018
500	0,1	0.016	0.035	0.99	0.029	0.408
500	0.69,2	0.969	0.093	1	0	0.174
500	1.10,3	0.984	0.023	1	0	0.033
500	1.61,5	1	0.012	1	0	0.008
500	2.30,10	1	0	1	0	0.004

- N: number of matched pairs in each of the two datasets for cluster detection and effect estimation respectively
- Robust se: robust standard error which is defined as inter-quartile range (IQR)/1.345

Table 2. Performance of the one-dimensional temporal scan statistic on effect estimation, with the true risk region $R_1 = \{-28 \leq T_1 \leq 35\}$

N	$\theta^*, \exp(\theta^*)$	#sim with valid effect estimates	Bias of $\widehat{\theta}$			95% CI	
			Median	Median/θ^*	Robust se	Coverage probability	Median width
25	0,1	1413	0	.	0.83	0.992	3.578
25	0.69,2	1554	-0.539	-0.778	0.727	0.945	3.28
25	1.10,3	1623	-0.693	-0.631	0.679	0.87	3.201
25	1.61,5	1714	-0.916	-0.569	0.929	0.73	3.099
25	2.30,10	1671	-1.12	-0.488	0.893	0.651	3.036
50	0,1	1623	0	.	0.679	0.988	3.201
50	0.69,2	1763	-0.47	-0.678	0.679	0.911	2.63
50	1.10,3	1855	-0.588	-0.535	0.727	0.809	2.263
50	1.61,5	1919	-0.629	-0.391	0.694	0.708	2.167
50	2.30,10	1903	-0.598	-0.26	0.787	0.71	2.413
100	0,1	1769	0	.	0.601	0.979	2.772
100	0.69,2	1893	-0.379	-0.547	0.469	0.863	1.761
100	1.10,3	1962	-0.405	-0.369	0.514	0.752	1.426
100	1.61,5	1992	-0.31	-0.193	0.532	0.758	1.536
100	2.30,10	1980	-0.248	-0.108	0.587	0.82	1.857
200	0,1	1861	0	.	0.514	0.978	2.4
200	0.69,2	1959	-0.279	-0.402	0.378	0.789	0.978
200	1.10,3	1991	-0.191	-0.174	0.337	0.788	1
200	1.61,5	2000	-0.148	-0.092	0.342	0.834	1.125
200	2.30,10	1999	-0.118	-0.051	0.419	0.89	1.378
500	0,1	1927	0	.	0.401	0.967	1.859
500	0.69,2	1999	-0.117	-0.169	0.198	0.794	0.592
500	1.10,3	2000	-0.069	-0.063	0.189	0.881	0.648
500	1.61,5	2000	-0.049	-0.03	0.202	0.912	0.738
500	2.30,10	2000	-0.024	-0.01	0.244	0.934	0.901

- N: number of matched pairs in each of the two datasets for cluster detection and effect estimation respectively
- Robust se: robust standard error which is defined as inter-quartile range (IQR)/1.345

Table 3. Performance of the one-dimensional temporal scan statistic on cluster detection, with the true risk region $R_2 = \{15 \leq T_1 \leq 49\}$

N	$\theta^*, \exp(\theta^*)$	Sensitivity		Specificity		p_G
		Median	Robust se	Median	Robust se	
25	0,1	0	0.191	0.947	0.067	0.617
25	0.69,2	0.286	0.445	0.977	0.067	0.612
25	1.10,3	0.486	0.508	0.985	0.072	0.523
25	1.61,5	0.771	0.424	0.992	0.061	0.422
25	2.30,10	0.886	0.191	1	0.042	0.33
50	0,1	0	0.106	0.962	0.05	0.514
50	0.69,2	0.286	0.561	0.985	0.061	0.554
50	1.10,3	0.743	0.53	0.992	0.05	0.456
50	1.61,5	0.914	0.212	0.992	0.045	0.364
50	2.30,10	0.971	0.085	1	0.022	0.197
100	0,1	0	0.085	0.977	0.039	0.453
100	0.69,2	0.543	0.614	0.992	0.045	0.499
100	1.10,3	0.943	0.212	0.992	0.045	0.366
100	1.61,5	0.971	0.064	0.992	0.022	0.189
100	2.30,10	1	0.042	1	0.011	0.085
200	0,1	0	0.042	0.985	0.033	0.425
200	0.69,2	0.886	0.381	0.992	0.045	0.418
200	1.10,3	0.971	0.064	0.992	0.028	0.209
200	1.61,5	1	0.021	1	0.011	0.08
200	2.30,10	1	0.021	1	0.006	0.033
500	0,1	0	0.021	0.985	0.028	0.411
500	0.69,2	0.971	0.064	0.992	0.028	0.197
500	1.10,3	1	0.021	1	0.011	0.059
500	1.61,5	1	0	1	0.006	0.022
500	2.30,10	1	0	1	0	0.006

- N: number of matched pairs in each of the two datasets for cluster detection and effect estimation respectively
- Robust se: robust standard error which is defined as inter-quartile range (IQR)/1.345

Table 4: Performance of the one-dimensional temporal scan statistic on effect estimation, with the true risk region $R_2 = \{15 \leq T_1 \leq 49\}$

N	$\theta^*, \exp(\theta^*)$	#sim with valid effect estimates	Bias of $\widehat{\theta}$			95% CI	
			Median	Median/θ^*	Robust se	Coverage probability	Median width
25	0,1	1447	0	.	0.757	0.991	3.578
25	0.69,2	1571	-0.405	-0.585	0.679	0.95	3.201
25	1.10,3	1702	-0.405	-0.369	0.814	0.91	3.064
25	1.61,5	1737	-0.511	-0.317	0.679	0.873	2.994
25	2.30,10	1659	-0.511	-0.222	0.814	0.843	3.036
50	0,1	1619	0	.	0.601	0.981	3.201
50	0.69,2	1788	-0.288	-0.415	0.592	0.931	2.295
50	1.10,3	1909	-0.336	-0.306	0.554	0.887	1.982
50	1.61,5	1957	-0.31	-0.193	0.584	0.861	1.982
50	2.30,10	1922	-0.266	-0.115	0.643	0.888	2.381
100	0,1	1769	0	.	0.601	0.979	2.772
100	0.69,2	1915	-0.208	-0.3	0.43	0.92	1.41
100	1.10,3	1987	-0.194	-0.177	0.366	0.885	1.299
100	1.61,5	2000	-0.157	-0.098	0.401	0.881	1.397
100	2.30,10	1995	-0.124	-0.054	0.465	0.92	1.688
200	0,1	1845	0	.	0.573	0.975	2.457
200	0.69,2	1983	-0.147	-0.212	0.275	0.883	0.878
200	1.10,3	1999	-0.113	-0.103	0.254	0.885	0.895
200	1.61,5	2000	-0.073	-0.045	0.275	0.921	0.992
200	2.30,10	2000	-0.051	-0.022	0.329	0.928	1.237
500	0,1	1931	0	.	0.419	0.972	1.905
500	0.69,2	2000	-0.074	-0.107	0.152	0.884	0.536
500	1.10,3	2000	-0.05	-0.045	0.159	0.922	0.563
500	1.61,5	2000	-0.026	-0.016	0.164	0.943	0.633
500	2.30,10	2000	-0.004	-0.002	0.209	0.947	0.789

- N: number of matched pairs in each of the two datasets for cluster detection and effect estimation respectively
- Robust se: robust standard error which is defined as inter-quartile range (IQR)/1.345

Table 5: Performance of the one-dimensional temporal scan statistic on cluster detection, with the true risk region $R_3 = \{43 \leq T_1 \leq 84\}$

N	$\theta^*, \exp(\theta^*)$	Sensitivity		Specificity		p_G
		Median	Robust se	Median	Robust se	
25	0,1	0	0.106	0.937	0.059	0.62
25	0.69,2	0.143	0.318	0.96	0.071	0.695
25	1.10,3	0.31	0.424	0.984	0.059	0.655
25	1.61,5	0.524	0.424	1	0.047	0.553
25	2.30,10	0.786	0.3	1	0.035	0.415
50	0,1	0	0.071	0.96	0.041	0.546
50	0.69,2	0.167	0.335	0.976	0.041	0.619
50	1.10,3	0.405	0.477	0.992	0.041	0.602
50	1.61,5	0.786	0.371	1	0.035	0.458
50	2.30,10	0.905	0.141	1	0.024	0.289
100	0,1	0	0.071	0.976	0.035	0.481
100	0.69,2	0.214	0.477	0.992	0.029	0.59
100	1.10,3	0.69	0.477	1	0.035	0.509
100	1.61,5	0.929	0.159	0.992	0.029	0.299
100	2.30,10	0.976	0.071	1	0.018	0.18
200	0,1	0	0.053	0.984	0.029	0.445
200	0.69,2	0.548	0.591	0.992	0.029	0.553
200	1.10,3	0.905	0.221	0.992	0.029	0.343
200	1.61,5	0.976	0.071	0.992	0.018	0.174
200	2.30,10	1	0.035	1	0.012	0.089
500	0,1	0	0.035	0.984	0.035	0.44
500	0.69,2	0.905	0.229	0.992	0.029	0.362
500	1.10,3	0.976	0.071	0.992	0.018	0.174
500	1.61,5	1	0.035	1	0.006	0.083
500	2.30,10	1	0.018	1	0.006	0.03

- N: number of matched pairs in each of the two datasets for cluster detection and effect estimation respectively
- Robust se: robust standard error which is defined as inter-quartile range (IQR)/1.345

Table 6. Performance of the one-dimensional temporal scan statistic on effect estimation, with the true risk region $R_3 = \{43 \leq T_1 \leq 84\}$

N	$\theta^*, \exp(\theta^*)$	#sim with valid effect estimates	Bias of $\widehat{\theta}$			95% CI	
			Median	Median/θ^*	Robust se	Coverage probability	Median width
25	0,1	1432	0	.	0.715	0.995	3.578
25	0.69,2	1450	-0.693	-1	0.814	0.932	3.395
25	1.10,3	1514	-0.588	-0.535	0.814	0.892	3.395
25	1.61,5	1585	-0.511	-0.317	0.893	0.844	3.201
25	2.30,10	1478	-0.613	-0.266	0.814	0.852	3.201
50	0,1	1616	0	.	0.649	0.989	3.201
50	0.69,2	1695	-0.431	-0.621	0.649	0.919	2.994
50	1.10,3	1837	-0.405	-0.369	0.649	0.874	2.479
50	1.61,5	1891	-0.288	-0.179	0.617	0.864	2.263
50	2.30,10	1866	-0.266	-0.115	0.62	0.896	2.435
100	0,1	1767	0	.	0.601	0.975	2.772
100	0.69,2	1875	-0.266	-0.383	0.578	0.898	2.087
100	1.10,3	1939	-0.216	-0.197	0.459	0.87	1.532
100	1.61,5	1995	-0.143	-0.089	0.418	0.886	1.532
100	2.30,10	1991	-0.128	-0.056	0.485	0.923	1.837
200	0,1	1845	0	.	0.573	0.975	2.457
200	0.69,2	1970	-0.182	-0.263	0.375	0.869	1.097
200	1.10,3	1998	-0.143	-0.13	0.293	0.884	1.003
200	1.61,5	2000	-0.086	-0.053	0.294	0.927	1.079
200	2.30,10	2000	-0.056	-0.024	0.348	0.941	1.303
500	0,1	1912	0	.	0.412	0.965	1.762
500	0.69,2	1999	-0.078	-0.112	0.181	0.889	0.618
500	1.10,3	2000	-0.057	-0.052	0.17	0.917	0.628
500	1.61,5	2000	-0.034	-0.021	0.191	0.937	0.686
500	2.30,10	2000	-0.02	-0.009	0.206	0.947	0.825

- N: number of matched pairs in each of the two datasets for cluster detection and effect estimation respectively
- Robust se: robust standard error which is defined as inter-quartile range (IQR)/1.345

Table 7. Performance of the one-dimensional temporal scan statistic on cluster detection, with the true risk region $\boldsymbol{R_4} = \{1 \leq T_2 \leq 28\}$

N	$\boldsymbol{\theta^*, \exp(\theta^*)}$	Sensitivity		Specificity		$\boldsymbol{p_G}$
		Median	Robust se	Median	Robust se	
25	0,1	0	0.159	0.95	0.085	0.536
25	0.69,2	0.25	0.344	1	0.053	0.396
25	1.10,3	0.393	0.45	1	0.026	0.255
25	1.61,5	0.75	0.424	1	0.016	0.16
25	2.30,10	0.893	0.212	1	0.005	0.118
50	0,1	0	0.106	0.964	0.058	0.472
50	0.69,2	0.179	0.397	1	0.032	0.331
50	1.10,3	0.679	0.53	1	0.016	0.195
50	1.61,5	0.893	0.185	1	0.005	0.126
50	2.30,10	0.964	0.079	1	0.005	0.063
100	0,1	0	0.053	0.971	0.04	0.433
100	0.69,2	0.393	0.556	1	0.016	0.263
100	1.10,3	0.893	0.265	1	0.011	0.142
100	1.61,5	0.964	0.079	1	0.005	0.063
100	2.30,10	1	0.026	1	0.005	0.023
200	0,1	0	0.026	0.979	0.032	0.391
200	0.69,2	0.821	0.424	1	0.011	0.183
200	1.10,3	0.964	0.079	1	0.005	0.077
200	1.61,5	1	0.026	1	0.005	0.022
200	2.30,10	1	0	1	0	0.011
500	0,1	0	0.026	0.986	0.032	0.397
500	0.69,2	0.964	0.079	1	0.005	0.08
500	1.10,3	1	0.026	1	0	0.02
500	1.61,5	1	0	1	0	0.006
500	2.30,10	1	0	1	0	0.003

- N: number of matched pairs in each of the two datasets for cluster detection and effect estimation respectively
- Robust se: robust standard error which is defined as inter-quartile range (IQR)/1.345

Table 8. Performance of the one-dimensional temporal scan statistic on effect estimation, with the true risk region $R_4 = \{1 \leq T_2 \leq 28\}$

N	$\theta^*, \exp(\theta^*)$	#sim with valid effect estimates	Bias of $\widehat{\theta}$			95% CI	
			Median	Median/θ^*	Robust se	Coverage probability	Median width
25	0,1	1437	0	.	0.794	0.993	3.578
25	0.69,2	1529	-0.405	-0.585	0.679	0.944	3.28
25	1.10,3	1625	-0.405	-0.369	0.814	0.903	3.143
25	1.61,5	1726	-0.431	-0.268	0.75	0.879	3.013
25	2.30,10	1623	-0.511	-0.222	0.814	0.863	3.013
50	0,1	1658	0	.	0.679	0.991	3.201
50	0.69,2	1739	-0.288	-0.415	0.628	0.931	2.705
50	1.10,3	1880	-0.288	-0.262	0.592	0.894	2.024
50	1.61,5	1955	-0.223	-0.139	0.565	0.889	1.994
50	2.30,10	1910	-0.143	-0.062	0.626	0.914	2.381
100	0,1	1795	0	.	0.601	0.986	2.863
100	0.69,2	1908	-0.223	-0.322	0.467	0.923	1.561
100	1.10,3	1986	-0.15	-0.136	0.368	0.903	1.315
100	1.61,5	1999	-0.118	-0.073	0.374	0.912	1.376
100	2.30,10	1994	-0.069	-0.03	0.471	0.935	1.683
200	0,1	1862	0	.	0.573	0.97	2.4
200	0.69,2	1979	-0.111	-0.16	0.282	0.907	0.894
200	1.10,3	1999	-0.084	-0.076	0.238	0.919	0.89
200	1.61,5	2000	-0.048	-0.03	0.256	0.94	0.98
200	2.30,10	2000	-0.033	-0.014	0.307	0.95	1.193
500	0,1	1930	0	.	0.381	0.968	1.762
500	0.69,2	2000	-0.048	-0.07	0.149	0.911	0.539
500	1.10,3	2000	-0.026	-0.023	0.149	0.934	0.561
500	1.61,5	2000	-0.013	-0.008	0.171	0.95	0.623
500	2.30,10	2000	0	0	0.196	0.953	0.765

- N: number of matched pairs in each of the two datasets for cluster detection and effect estimation respectively
- Robust se: robust standard error which is defined as inter-quartile range (IQR)/1.345

Table 9. Performance of the two-dimensional time-time scan statistic on cluster detection, with the true risk region $R_5 = \{-28 \leq T_1 \leq 35 \ \& \ 1 \leq T_2 \leq 28\}$

N	$\theta^*, \exp(\theta^*)$	T_1		T_2		Sensitivity		Specificity		p_G
		Sens	Spec	Sens	Spec	Median	Robust se	Median	Robust se	
25	0,1	0.238	1	0	0.814	0	0.089	0.948	0.039	0.53
25	0.69,2	0.222	1	0.214	0.843	0	0.232	0.951	0.04	0.518
25	1.10,3	0.222	0.981	0.357	0.871	0.197	0.329	0.957	0.038	0.453
25	1.61,5	0.238	0.971	0.5	0.9	0.352	0.365	0.963	0.041	0.322
25	2.30,10	0.254	0.971	0.607	0.929	0.512	0.318	0.968	0.044	0.184
50	0,1	0.206	1	0	0.85	0	0.022	0.962	0.028	0.552
50	0.69,2	0.206	1	0.214	0.871	0	0.219	0.965	0.03	0.497
50	1.10,3	0.206	0.99	0.357	0.886	0.192	0.325	0.967	0.032	0.436
50	1.61,5	0.238	0.981	0.5	0.943	0.389	0.31	0.973	0.035	0.262
50	2.30,10	0.27	0.981	0.643	0.975	0.612	0.299	0.98	0.038	0.106
100	0,1	0.159	1	0	0.879	0	0	0.975	0.019	0.638
100	0.69,2	0.175	0.99	0.214	0.907	0	0.183	0.976	0.022	0.507
100	1.10,3	0.206	1	0.357	0.929	0.192	0.305	0.98	0.023	0.407
100	1.61,5	0.238	1	0.536	0.971	0.414	0.345	0.985	0.031	0.207
100	2.30,10	0.302	1	0.75	0.993	0.744	0.239	0.99	0.031	0.063
200	0,1	0.127	1	0	0.907	0	0	0.984	0.014	0.811
200	0.69,2	0.159	1	0.179	0.929	0	0.168	0.986	0.016	0.572
200	1.10,3	0.19	1	0.393	0.975	0.219	0.372	0.988	0.019	0.365
200	1.61,5	0.286	1	0.679	0.993	0.645	0.332	0.995	0.018	0.106
200	2.30,10	0.349	1	0.857	1	0.842	0.124	0.998	0.011	0.031
500	0,1	0.079	1	0	0.936	0	0	0.992	0.008	0.737
500	0.69,2	0.159	1	0.25	0.971	0.099	0.278	0.992	0.012	0.493
500	1.10,3	0.286	1	0.679	0.993	0.666	0.356	0.996	0.013	0.155
500	1.61,5	0.365	1	0.857	1	0.889	0.121	0.998	0.005	0.031
500	2.30,10	0.397	1	0.964	1	0.963	0.055	1	0.002	0.007

- N: number of matched pairs in each of the two datasets for cluster detection and effect estimation respectively
- Robust se: robust standard error which is defined as inter-quartile range (IQR)/1.345

Table 10. Performance of the two-dimensional time-time scan statistic on effect estimation , with the true risk region $R_5 = \{-28 \leq T_1 \leq 35 \ \& \ 1 \leq T_2 \leq 28\}$

N	$\theta^*, \exp(\theta^*)$	#sim with valid effect estimates	Bias of $\widehat{\theta}$			95% CI	
			Median	Median/θ^*	Robust se	Coverage probability	Median width
25	0,1	1451	0	.	1.028	0.998	3.92
25	0.69,2	1437	-0.693	-1	0.893	0.96	3.92
25	1.10,3	1493	-1.1	-1	0.814	0.849	3.92
25	1.61,5	1489	-1.2	-0.748	0.679	0.708	3.578
25	2.30,10	1498	-1.49	-0.648	0.862	0.598	3.578
50	0,1	1770	0	.	0.794	0.992	3.201
50	0.69,2	1742	-0.693	-1	0.814	0.909	3.201
50	1.10,3	1729	-0.847	-0.771	0.679	0.761	2.994
50	1.61,5	1792	-0.916	-0.569	0.814	0.642	2.863
50	2.30,10	1786	-1.1	-0.477	0.75	0.587	2.654
100	0,1	1853	0	.	0.636	0.979	2.772
100	0.69,2	1889	-0.588	-0.848	0.66	0.854	2.705
100	1.10,3	1880	-0.762	-0.694	0.641	0.718	2.53
100	1.61,5	1927	-0.836	-0.52	0.645	0.59	2.235
100	2.30,10	1953	-0.815	-0.354	0.685	0.572	1.975
200	0,1	1950	0	.	0.601	0.971	2.4
200	0.69,2	1953	-0.575	-0.83	0.592	0.799	2.295
200	1.10,3	1964	-0.657	-0.598	0.628	0.665	2.029
200	1.61,5	1993	-0.545	-0.338	0.544	0.619	1.518
200	2.30,10	1999	-0.464	-0.202	0.532	0.659	1.504
500	0,1	1978	0	.	0.514	0.972	2.181
500	0.69,2	1992	-0.431	-0.621	0.414	0.737	1.761
500	1.10,3	1998	-0.329	-0.299	0.378	0.659	0.953
500	1.61,5	2000	-0.233	-0.145	0.299	0.746	0.915
500	2.30,10	2000	-0.159	-0.069	0.306	0.839	1.017

- N: number of matched pairs in each of the two datasets for cluster detection and effect estimation respectively
- Robust se: robust standard error which is defined as inter-quartile range (IQR)/1.345

Table 11. Performance of the two-dimensional time-time scan statistic on cluster detection, with the true risk region $R_6 = \{15 \leq T_1 \leq 49 \ \& \ 1 \leq T_2 \leq 28\}$

N		T_1		T_2		Sensitivity		Specificity		p_G
	$\theta^*, \exp(\theta^*)$	Sens	Spec	Sens	Spec	Median	Robust se	Median	Robust se	
25	0,1	0.229	0.895	0	0.829	0	0.143	0.952	0.039	0.523
25	0.69,2	0.371	0.917	0.393	0.871	0.173	0.299	0.96	0.039	0.453
25	1.10,3	0.457	0.94	0.5	0.914	0.281	0.302	0.967	0.04	0.338
25	1.61,5	0.543	0.977	0.643	0.964	0.39	0.287	0.975	0.042	0.188
25	2.30,10	0.686	0.992	0.786	0.986	0.594	0.31	0.978	0.047	0.103
50	0,1	0.2	0.91	0	0.85	0	0.096	0.963	0.029	0.508
50	0.69,2	0.343	0.94	0.393	0.9	0.177	0.258	0.972	0.034	0.418
50	1.10,3	0.486	0.955	0.536	0.95	0.281	0.268	0.975	0.034	0.28
50	1.61,5	0.629	0.985	0.714	0.986	0.463	0.339	0.982	0.039	0.131
50	2.30,10	0.8	0.992	0.893	1	0.786	0.252	0.982	0.046	0.062
100	0,1	0.143	0.925	0	0.879	0	0.051	0.975	0.021	0.576
100	0.69,2	0.343	0.962	0.357	0.943	0.152	0.22	0.981	0.027	0.39
100	1.10,3	0.514	0.985	0.607	0.986	0.287	0.352	0.987	0.029	0.209
100	1.61,5	0.771	0.985	0.857	0.993	0.723	0.368	0.986	0.036	0.094
100	2.30,10	0.886	0.992	0.964	1	0.899	0.114	0.987	0.039	0.038
200	0,1	0.086	0.94	0	0.907	0	0.027	0.984	0.014	0.721
200	0.69,2	0.357	0.977	0.357	0.979	0.132	0.245	0.989	0.021	0.363
200	1.10,3	0.686	0.992	0.786	0.993	0.581	0.474	0.99	0.028	0.145
200	1.61,5	0.886	0.992	0.964	1	0.881	0.139	0.989	0.031	0.057
200	2.30,10	0.943	1	1	1	0.955	0.066	0.996	0.014	0.012
500	0,1	0.057	0.955	0	0.936	0	0	0.991	0.01	0.741
500	0.69,2	0.6	0.992	0.679	0.993	0.4	0.53	0.993	0.019	0.233
500	1.10,3	0.886	0.992	0.964	1	0.881	0.139	0.993	0.02	0.057
500	1.61,5	0.971	1	1	1	0.96	0.056	0.998	0.006	0.011
500	2.30,10	1	1	1	1	0.99	0.022	1	0.002	0.003

- N: number of matched pairs in each of the two datasets for cluster detection and effect estimation respectively
- Robust se: robust standard error which is defined as inter-quartile range (IQR)/1.345

Table 12. Performance of the two-dimensional time-time scan statistic on OR estimation, with the true risk region $R_6 = \{15 \leq T_1 \leq 49 \ \& \ 1 \leq T_2 \leq 28\}$

N	$\theta^*, \exp(\theta^*)$	#sim with valid effect estimates	Bias of $\widehat{\theta}$			95% CI	
			Median	Median/θ^*	Robust se	Coverage probability	Median width
25	0,1	1424	0	.	0.679	0.998	3.92
25	0.69,2	1485	-0.693	-1	0.814	0.949	3.578
25	1.10,3	1488	-0.762	-0.694	0.845	0.876	3.578
25	1.61,5	1550	-0.916	-0.569	0.814	0.785	3.395
25	2.30,10	1522	-1.05	-0.456	0.814	0.73	3.28
50	0,1	1745	0	.	0.757	0.989	3.28
50	0.69,2	1768	-0.511	-0.737	0.727	0.907	2.994
50	1.10,3	1805	-0.693	-0.631	0.679	0.824	2.772
50	1.61,5	1845	-0.654	-0.406	0.649	0.763	2.53
50	2.30,10	1889	-0.734	-0.319	0.679	0.711	2.435
100	0,1	1891	0	.	0.641	0.979	2.772
100	0.69,2	1922	-0.47	-0.678	0.679	0.856	2.479
100	1.10,3	1939	-0.511	-0.465	0.592	0.775	2.147
100	1.61,5	1973	-0.511	-0.317	0.514	0.715	1.676
100	2.30,10	1992	-0.487	-0.212	0.526	0.722	1.722
200	0,1	1950	0	.	0.601	0.978	2.4
200	0.69,2	1950	-0.405	-0.585	0.459	0.826	2.095
200	1.10,3	1990	-0.405	-0.369	0.416	0.722	1.233
200	1.61,5	2000	-0.364	-0.226	0.365	0.686	1.123
200	2.30,10	2000	-0.274	-0.119	0.424	0.759	1.309
500	0,1	1978	0	.	0.514	0.967	2.18
500	0.69,2	1994	-0.269	-0.388	0.319	0.745	0.884
500	1.10,3	2000	-0.22	-0.2	0.255	0.675	0.7
500	1.61,5	2000	-0.126	-0.079	0.241	0.805	0.746
500	2.30,10	2000	-0.074	-0.032	0.25	0.898	0.879

- N: number of matched pairs in each of the two datasets for cluster detection and effect estimation respectively
- Robust se: robust standard error which is defined as inter-quartile range (IQR)/1.345

Table 13. Performance of the two-dimensional time-time scan statistic on cluster detection, with the true risk region $R_7 = \{43 \leq T_1 \leq 84 \ \& \ 1 \leq T_2 \leq 28\}$

N		T_1		T_2		Sensitivity		Specificity		p_G
	$\theta^*, \exp(\theta^*)$	Sens	Spec	Sens	Spec	Median	Robust se	Median	Robust se	
25	0,1	0	0.865	0	0.829	0	0.096	0.951	0.04	0.534
25	0.69,2	0.286	0.897	0.393	0.907	0.143	0.265	0.964	0.04	0.462
25	1.10,3	0.452	0.929	0.571	0.971	0.255	0.344	0.974	0.039	0.365
25	1.61,5	0.619	0.976	0.679	1	0.372	0.292	0.987	0.032	0.2
25	2.30,10	0.81	0.992	0.821	1	0.571	0.318	0.993	0.025	0.096
50	0,1	0	0.881	0	0.85	0	0.048	0.962	0.028	0.518
50	0.69,2	0.31	0.921	0.429	0.929	0.162	0.252	0.974	0.03	0.4
50	1.10,3	0.5	0.968	0.589	1	0.276	0.261	0.985	0.028	0.277
50	1.61,5	0.762	0.984	0.786	1	0.476	0.365	0.99	0.027	0.116
50	2.30,10	0.905	0.984	0.929	1	0.767	0.279	0.99	0.029	0.063
100	0,1	0	0.897	0	0.879	0	0.038	0.975	0.019	0.578
100	0.69,2	0.357	0.944	0.393	0.979	0.155	0.228	0.984	0.022	0.354
100	1.10,3	0.667	0.976	0.679	1	0.327	0.385	0.989	0.025	0.164
100	1.61,5	0.929	0.976	0.929	1	0.786	0.318	0.985	0.034	0.075
100	2.30,10	0.976	0.984	0.964	1	0.918	0.141	0.987	0.03	0.034
200	0,1	0	0.921	0	0.907	0	0.022	0.984	0.013	0.744
200	0.69,2	0.405	0.968	0.464	1	0.155	0.269	0.991	0.018	0.29
200	1.10,3	0.881	0.976	0.857	1	0.702	0.402	0.988	0.03	0.1
200	1.61,5	0.976	0.976	0.964	1	0.929	0.141	0.987	0.031	0.047
200	2.30,10	1	0.992	1	1	0.964	0.069	0.992	0.021	0.012
500	0,1	0	0.944	0	0.943	0	0.009	0.992	0.008	0.716
500	0.69,2	0.833	0.976	0.821	1	0.582	0.541	0.991	0.03	0.149
500	1.10,3	0.976	0.976	0.964	1	0.929	0.111	0.987	0.03	0.035
500	1.61,5	1	0.992	1	1	0.976	0.053	0.993	0.014	0.009
500	2.30,10	1	0.992	1	1	1	0.026	0.996	0.008	0.003

- N: number of matched pairs in each of the two datasets for cluster detection and effect estimation respectively
- Robust se: robust standard error which is defined as inter-quartile range (IQR)/1.345

Table 14. Performance of the two-dimensional time-time scan statistic on effect estimation, with the true risk region $R_7 = \{43 \leq T_1 \leq 84 \ \& \ 1 \leq T_2 \leq 28\}$

N	$\theta^*, \exp(\theta^*)$	#sim with valid effect estimates	Bias of $\widehat{\theta}$			95% CI	
			Median	Median/θ^*	Robust se	Coverage probability	Median width
25	0,1	1471	0	.	0.893	0.999	3.92
25	0.69,2	1438	-0.693	-1	0.814	0.949	3.92
25	1.10,3	1444	-0.693	-0.631	0.814	0.885	3.92
25	1.61,5	1459	-0.916	-0.569	1.028	0.833	3.578
25	2.30,10	1349	-0.799	-0.347	0.679	0.835	3.92
50	0,1	1734	0	.	0.757	0.99	3.201
50	0.69,2	1751	-0.511	-0.737	0.763	0.922	3.201
50	1.10,3	1768	-0.405	-0.369	0.814	0.868	2.977
50	1.61,5	1836	-0.511	-0.317	0.679	0.835	2.553
50	2.30,10	1829	-0.405	-0.176	0.645	0.861	2.466
100	0,1	1893	0	.	0.649	0.984	2.772
100	0.69,2	1909	-0.388	-0.559	0.613	0.894	2.553
100	1.10,3	1934	-0.351	-0.32	0.55	0.852	2.063
100	1.61,5	1979	-0.28	-0.174	0.446	0.872	1.636
100	2.30,10	1996	-0.203	-0.088	0.499	0.88	1.846
200	0,1	1947	0	.	0.601	0.976	2.479
200	0.69,2	1969	-0.288	-0.415	0.534	0.858	2.024
200	1.10,3	1994	-0.216	-0.197	0.353	0.845	1.141
200	1.61,5	2000	-0.184	-0.114	0.308	0.862	1.109
200	2.30,10	2000	-0.147	-0.064	0.366	0.875	1.3
500	0,1	1971	0	.	0.527	0.966	2.181
500	0.69,2	1997	-0.158	-0.227	0.249	0.857	0.748
500	1.10,3	2000	-0.115	-0.105	0.194	0.855	0.663
500	1.61,5	2000	-0.082	-0.051	0.193	0.881	0.707
500	2.30,10	2000	-0.059	-0.026	0.223	0.915	0.839

- N: number of matched pairs in each of the two datasets for cluster detection and effect estimation respectively
- Robust se: robust standard error which is defined as inter-quartile range (IQR)/1.345

VIII. REFERENCES

1. Naus JI. The distribution of the size of the maximum cluster of points on a line. Journal of the American Statistical Association 1965;60:532-8.
2. Kulldorff M. A spatial scan statistic. Communications in Statistics-Theory and methods 1997;26:1481-96.
3. Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2001;164:61-72.
4. Huang L, Kulldorff M, Gregorio D. A spatial scan statistic for survival data. Biometrics 2007;63:109-18.
5. Jung I, Kulldorff M, Klassen AC. A spatial scan statistic for ordinal data. Stat Med 2007;26:1594-607.
6. Jung I, Kulldorff M, Richard OJ. A spatial scan statistic for multinomial data. Stat Med 2010;29:1910-8.
7. Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. Stat Med 2006;25:3929-43.
8. Kulldorff M. SaTScan User Guide version9.4. http://www.satscan.org/cgi-bin/satscan/register.pl/SaTScan_Users_Guide.pdf?todo=process_userguide_download.
9. Kulldorff M. Temporal Scan Statistics for Pharmacoepidemiology. Unpublished paper. 2016.
10. Irving SA, Kieke BA, Donahue JG, et al. Trivalent inactivated influenza vaccine and spontaneous abortion. Obstet Gynecol 2013;121:159-65.
11. Baril L, Rosillon D, Willame C, et al. Risk of spontaneous abortion and other pregnancy outcomes in 15-25 year old women exposed to human papillomavirus-16/18 AS04-adjuvanted vaccine in the United Kingdom. Vaccine 2015;33:6884-91.
12. Suissa S. The case-time-control design. Epidemiology 1995;6:248-53.
13. Suissa S. The case-time-control design: further assumptions and conditions. Epidemiology 1998:441-5.
14. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. American journal of epidemiology 1991;133:144-53.
15. Li DK, Odouli R, Wi S, et al. A population-based prospective cohort study of personal exposure to magnetic fields during pregnancy and the risk of miscarriage. Epidemiology 2002;13:9-20.
16. Kawai AT, Li L, Kulldorff M, et al. Mini-sentinel CBER/PRISM surveillance protocol - influenza vaccines and pregnancy outcomes. http://www.mini-sentinel.org/work_products/PRISM/Mini-Sentinel_PRISM_Influenza-Vaccines-and-Pregnancy-Outcomes-Protocol.pdf.
17. Casella G, Berger RL. Statistical Inference: Duxbury Pacific Grove, CA; 2002.

IX. APPENDIX

eTable 1. Weekly SAB rates per 1000 woman-week

Week	Days post LMP	Weekly rate (per 1000 woman-weeks)*
6	36-42	21.05
7	43-49	28.59
8	50-56	34.62
9	57-63	28.55
10	64-70	21.88
11	71-77	23.96
12	78-84	21.76
13	85-91	11.55
14	92-98	8.15
15	99-105	7.05
16	105-112	1.18
17	113-119	1.18
18	120-126	1.18
19	127-133	2.38
20	132-140	1.2