Research Article

Statistics in Medicine

Received 29 June 2009, Accepted 22 September 2010 Published online 30 November 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4125

Identifying optimal risk windows for self-controlled case series studies of vaccine safety

Stanley Xu,[a*†] Lijing Zhang,[a] Jennifer C. Nelson,[b,c] Chan Zeng,[a] John Mullooly,[d] David McClure[a] and Jason Glanz[a]

In vaccine safety studies, subjects are considered at increased risk for adverse events for a period of time after vaccination known as risk window. To our knowledge, risk windows for vaccine safety studies have tended to be pre-defined and not to use information from the current study. Inaccurate specification of the risk window can result in either including the true control period in the risk window or including some of the risk window in the control period, which can introduce bias. We propose a data-based approach for identifying the optimal risk windows for self-controlled case series studies of vaccine safety. The approach involves fitting conditional Poisson regression models to obtain incidence rate ratio estimates for different risk window lengths. For a specified risk window length (L), the average time at risk, $T(L)$, is calculated. When the specified risk window is shorter than the true, the incidence rate ratio decreases with $1/T(L)$ increasing but there is no explicit relationship. When the specified risk window is longer than the true, the incidence rate ratio increases linearly with $1/T(L)$ increasing. Theoretically, the risk window with the maximum incidence ratio is the optimal risk window. Because of sparse data problem, we recommend using both the maximum incidence rate ratio and the linear relationship when the specified risk window is longer than the true to identify the optimal risk windows. Both simulation studies and vaccine safety data applications show that our proposed approach is effective in identifying medium and long-risk windows. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: optimal risk window; self-controlled case series; incidence rate ratio; change point; adverse events after immunization

1. Introduction

Observational vaccine safety studies are important for determining whether particular adverse events following immunization (AEFI) are associated with a specific vaccine [1–8]. Large linked clinical databases have been successfully used to study associations between vaccinations and adverse events. As an example, the Vaccine Safety Datalink (VSD) project combines eight managed care organizations in the United States to collect vaccination and other medical data on more than 8.8 million people each year. Information about AEFIs is derived from billing ICD-9 codes and electronic medical records of patients. The available electronic data on this large population make it possible to study not only common AEFIs (e.g. fever, soreness), but also rare AEFIs (e.g. death, seizures, idiopathic thrombocytopenic purpura (ITP)). In addition, hypotheses can be generated from screening studies of multiple AEFIs using these data.

One of the challenges in vaccine safety studies is that the vaccinated population can be very different from the unvaccinated. To address this selection bias, a novel study design called the

[a] *Institute for Health Research, Kaiser Permanente Colorado, Denver, CO, U.S.A.*
[b] *Group Health Research Institute, Seattle, WA, U.S.A.*
[c] *Department of Biostatistics, University of Washington, WA, U.S.A.*
[d] *Centre for Health Research, Northwest Kaiser Permanente, Portland, OR, U.S.A.*
[*] *Correspondence to: Stanley Xu, Institute for Health Research, Kaiser Permanente Colorado, Denver, CO, U.S.A.*
[†] *E-mail: stan.xu@kp.org*

self-controlled case series (SCCS) design was developed to account for measured and unmeasured confounders [9]. In a typical SCCS vaccine safety study, only subjects with adverse events are included in the analysis. Subjects are considered at increased risk for adverse events for a period of time after vaccination called the risk window. To our knowledge, risk windows for most AEFIs in vaccine safety studies have been pre-defined, partially informed by prior studies, or by prior hypotheses based on biological understanding of how the vaccines work, but not based on preliminary exploration of data from the current study for the specific AEFIs. This approach can lead to bias if there is little or no prior information available to characterize the risk window. In particular, a pre-specified risk window that includes the true control period will underestimate the incidence rate ratio. Conversely, including a portion of the true risk window as the control period will also underestimate the true risk due to the inflation of incidence rate in the control period. For these reasons, it is important to define the risk window for vaccine safety studies as accurately as possible.

This problem can be illustrated using simulated data with a known risk window (Figure 1). We simulated vaccination and adverse event data with known risk windows (14 and 42 days) and a known incidence rate ratio (2.0) as in Glanz *et al.* [10]. Then, we used conditional Poisson regression models to obtain a set of incidence rate ratios for different risk window lengths. Figure 1 shows that misspecification of the risk windows leads to biased estimation of the true association between a vaccine and an adverse event. Thus, given the bias associated with risk window misspecification, it would be useful to have a data-based statistical approach to aid in identifying the optimal risk window, one that uses the information about the change point in data like those shown in Figure 1.

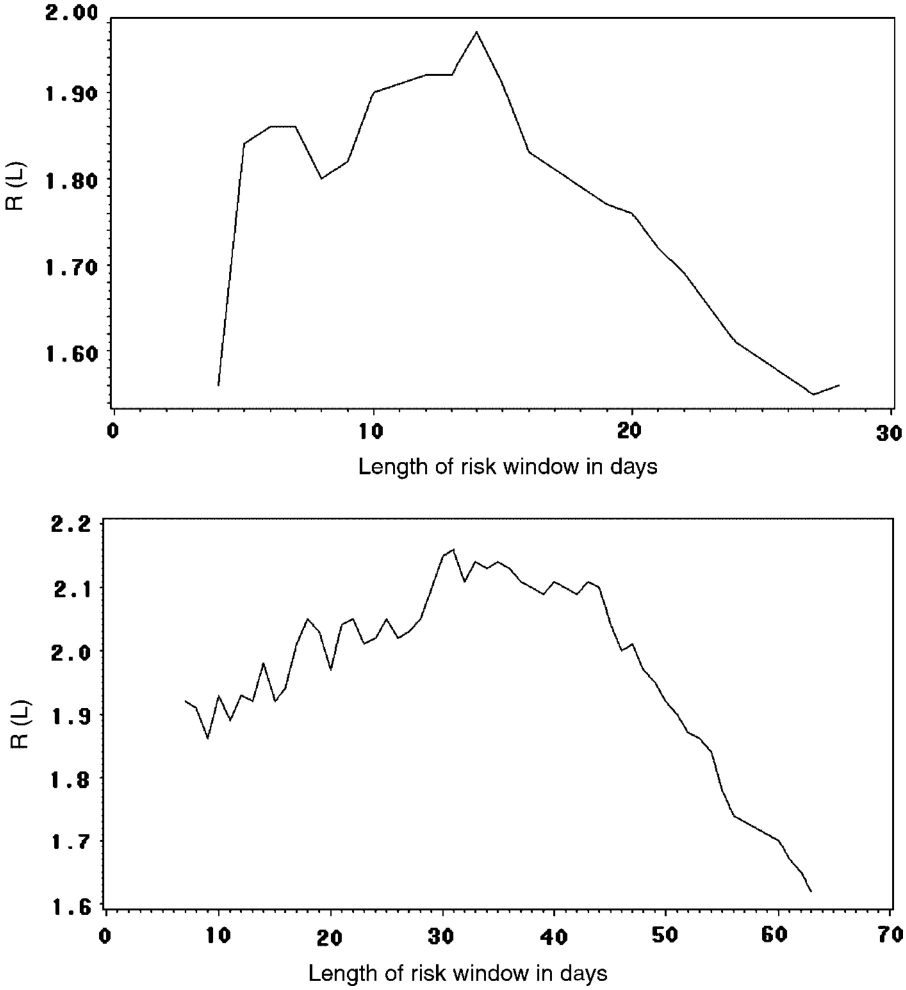

Figure 1. Estimated incidence rate ratios for different specified risk window lengths (in days) using simulated data with a known risk window of 14 (upper panel) and 42 days (lower panel) and a true incidence rate ratio of 2.0.

There is much work in the literature on estimating an unknown change point as in Figure 1. Methods range from simple detection of a change in the mean of stationary, Gaussian dependent variables, to more general approaches covering changes in conditional distributions of non-stationary, non-Gaussian, and non-linear processes [11–15]. In addition, there is considerable literature on estimation and testing of change points in hazard rate models [16–22]. However, none of these methods is applicable in SCCS vaccine safety studies. First, data structure in an SCCS study is unique since only subjects who experienced adverse events (i.e. cases) are included in the study, whereas in traditional survival analyses the entire population (with or without adverse events) is included. Second, the underlying distribution of adverse events in an SCCS study is non-Gaussian, hence the methods used for Gaussian variables cannot be applied here.

In this paper we propose a data-based statistical method to identify the optimal risk window for an SCCS study. To determine the optimal risk window for a particular vaccine we use conditional Poisson regression models to estimate a set of incidence rate ratios by varying the specified length of the risk window as in Figure 1. We then plot the estimated incidence rate ratios against the reciprocals of the average time at risk to identify the length of the optimal risk window empirically. The approach is described in detail in Section 2.

2. Statistical methods

2.1. Self-controlled case series and conditional Poisson regression

The SCCS design is a case-only method in which a subject's follow-up period is partitioned into mutually exclusive risk and control periods or windows. Further partitions may be needed to adjust for time-varying covariates such as age and season. The adverse event incidence rate in the risk period following vaccination is compared to that in the control period so that each case acts as its own control [9, 23]. Incidence rate ratios for risk versus control periods in an SCCS study are estimated using conditional Poisson regression models, which accounts for the within-subject dependence. Conditioning is used to obtain a likelihood kernel in which only cases need to be sampled. As a result, intercept parameters characterizing each subject's baseline risk for the adverse event of interest are not present in the likelihood kernel. This results in a *de facto* adjustment for all time-invariant subject-level risk factors and confounders (measured or not measured), allowing only within-subject comparisons of incidence rates between the risk and control periods. This is similar to a stratified analysis with each subject as a unique stratum. SCCS is particularly useful for controlling for confounding by indication, where the probability of vaccination is related to the subject-level risk of the adverse event.

Suppose we have a sample of n subjects with adverse events and that the baseline incidence rate (I_0) of an adverse event is constant over the follow-up period. For simplicity we also assume that the incidence rate ratio is constant. Let $R=\exp(\beta)$ denote the incidence rate ratio for the vaccination effect where β is the coefficient of vaccine effect. According to Farrington [9], the conditional Poisson likelihood kernel is the product of the likelihood kernel across subjects which is of the following form for the ith subject

$$L_i(\beta,\alpha_1,\ldots,\alpha_J)=\prod_{j=1}^{J}\prod_{k=0,1}\left\{\frac{t_{ijk}\exp(\alpha_j+\beta k)}{\sum_{c=1}^{J}\sum_{d=0,1}t_{icd}\exp(\alpha_c+\beta d)}\right\}^{y_{ijk}} \tag{1}$$

where α_j, $j=1,\ldots,J$, are age effects, t_{ijk} is the person-time (in days) for subject i in age group j and risk period k, $k=0$ for control period and equal to 1 for risk period, and y_{ijk} is the corresponding number of adverse events which will be binary when the adverse event is rare. Model (1) can be easily modified to allow for multiple risk levels in the risk window [24].

Given a pre-defined risk window, the maximum likelihood estimates of the αs and β can be obtained using the SAS NLMIXED procedure [25] by defining the log conditional Poisson likelihood and using the *general*(*ll*) statement. In the model (1), other time-varying covariates (e.g. seasonality) can be included in the same way as the age effects. Recently, a semi-parametric model was developed to fit the conditional Poisson model adjusting for age at event [26–28].

2.2. *Relationship between estimated incidence rate ratios (R(L)) and length of specified risk windows (L)*

We now derive the analytic relationship between the length of a specified risk window and the incidence rate ratio in order to obtain an accurate estimate of the optimal risk window. With only the vaccine effect, the maximum likelihood estimator for R is the solution of equation (2) [29, p. 36]

$$n_1 = \sum_{i=1}^{n} \frac{Rt_{i1}}{Rt_{i1}+t_{i0}} \tag{2}$$

where n_1 is the total number of adverse events in the risk period, t_{i1} is the person-time in risk period, and t_{i0} is the person-time in control period. Note that t_{i1} and t_{i0} can be different across subjects due to different lengths of the subjects' follow-up periods. Let L_0 be the true risk window length in days. In addition, let L denote a specified length of the risk window from a series of possible risk window lengths, and t'_1 be the person-time at risk in days based on L. There are three possible scenarios for the specified risk window, L: (1) $L<L_0$; (2) $L>L_0$; and (3) $L=L_0$ meaning that the specified risk window is the same as the true risk window. There is no explicit relation between $R(L)$ and L when $L<L_0$. However, when $L>L_0$, the estimator of R based on L, called it $R(L)$, is the solution to (3)

$$n'_1 = \sum_{i=1}^{n} \frac{R(L)t'_{i1}}{R(L)t'_{i1}+t'_{i0}} \tag{3}$$

For ith subject

$$I_0 R(L)t'_{i1} = I_0 Rt_{i1} + I_0(t'_{i1}-t_{i1}) \tag{4}$$

Summing over the subjects gives

$$R(L)\sum_{i=1}^{n} t'_{i1} = R\sum_{i=1}^{n} t_{i1} + \sum_{i=1}^{n}(t'_{i1}-t_{i1}) \tag{5}$$

Further

$$R(L) = 1+(R-1)T(L_0)\frac{1}{T(L)} \tag{6}$$

where $T(L_0)=(\sum_{i=1}^{n} t_{i1})/n$ is the average time at risk given the true risk window, L_0, and $T(L)=(\sum_{i=1}^{n} t'_{i1})/n$ is the average time at risk when the risk window is specified as L. Thus the relation between $R(L)$ and $1/T(L)$ is positive and linear when $L>L_0$.

With both a vaccine effect and age effects α_j, $j=1,\ldots,J$, the maximum likelihood estimator for R is the solution of equation (7) when there is one adverse event per subject [29, p. 236]

$$n_1 = \sum_{i=1}^{n} \frac{R\sum_{j=1}^{J} t_{ij1}\exp(\alpha_j)}{\sum_{j=1}^{J}(Rt_{ij1}+t_{ij0})\exp(\alpha_j)} \tag{7}$$

Given a specified length of the risk window, L, for subject i and age j,

$$\exp(\alpha_j)I_0 R(L)t'_{ij1} = \exp(\alpha_j)I_0 Rt_{ij1} + \exp(\alpha_j)I_0(t'_{ij1}-t_{ij1}) \tag{8}$$

Summing over subjects and age groups gives

$$R(L)\sum_{i=1}^{n}\sum_{j=1}^{J} t'_{ij1} = R\sum_{i=1}^{n}\sum_{j=1}^{J} t_{ij1} + \sum_{i=1}^{n}\sum_{j=1}^{J}(t'_{ij1}-t_{ij1}) \tag{9}$$

Further,

$$R(L) = 1+(R-1)T(L_0)\frac{1}{T(L)} \tag{10}$$

where $T(L_0)=(\sum_{i=1}^{n}\sum_{j=1}^{J} t_{ij1})/n$ is the average time at risk given the true risk window and $T(L)=(\sum_{i=1}^{n}\sum_{j=1}^{J} t'_{ij1})/n$ is the average time at risk when the risk window is L. Note that equations (10) and (6) reveal the same relation between $R(L)$ and L (or $1/T(L)$).

2.3. Identification of the optimal risk window

Although there is no explicit relation between $R(L)$ and $1/T(L)$ when $L<L_0$, $R(L)$ decreases with $1/T(L)$ increasing because the control period will contain part of the risk period when $L<L_0$. Theoretically, the optimal risk window is the selected risk window at which the estimated $R(L)$ is the maximum. However, this will not be always true when data are sparse as often seen in SCCS design of vaccine safety studies. Thus, we propose to use both $R(L)$ when $L>L_0$ and R_M, the maximum incidence rate ratio, to identify the optimal risk window. We do not use $R(L)$ when $L<L_0$ because there is no explicit relationship between $R(L)$ and $1/T(L)$. Additionally, there is more volatility in $R(L)$ estimation when $L<L_0$ because fewer cases in true risk window are included as supposed in the analysis.

First, fit a series of conditional Poisson regression models with different specified risk window lengths (L). Time-varying covariates such as age effects can be included. A set of $R(L)$s is thus obtained in this step. Usually, one would choose a sufficient number of Ls both shorter and longer than L_0 (the unknown true risk window), so that an accurate optimal risk window length can be identified. Then, plot $R(L)$ against $1/T(L)$ and visualize their relation. Next identify the risk window (L_M) with the maximum incidence rate ratio. If the plot shows an approximate linear relationship between $R(L)$ and $1/T(L)$ when $L>L_M$, then L_M is the optimal empirical risk window. If not, then identify the region where the estimated incidence rate ratios $R(L)$ and $1/T(L)$ have a linear relationship and find the decreasing change point. The value of L at this change point represents the optimal risk window length. One can then use this identified risk window length to calculate the corresponding R. This change point may not be one with R_M due to the volatility of data, especially when $L<L_0$. Note that $T(L)$ is the average time at risk. It is the same as L only when censoring occurs after the specified risk window for all subjects.

3. Simulation study and results

3.1. Simulation

The purpose of our simulation study was to evaluate the accuracy and precision of our proposed approach for identifying the optimal risk window under a variety of plausible vaccine safety study settings. We simulated data with true risk windows of 14 and 42 days representing medium (e.g. for gastritis/duodenitis in influenza vaccine safety studies) and long (e.g. for ITP) post-vaccination risk windows, respectively. In particular, each subject has a follow-up period of 365 days, which consists of both risk and control periods. Typically, there are three periods for each subject: the control period before vaccination, the risk period after vaccination, and the control period after the risk period. With age effects, the follow-up period will also be partitioned into intervals based on age groups. The probability of having a case in each of these intervals was calculated based on the probability model (11) similar to that described by Glanz *et al.* [10]

$$P_{ijk}=\frac{t_{ijk}}{1+\exp(-(\beta_0+\alpha_j+\beta k))} \tag{11}$$

where P_{ijk} is the probability of having a case, β_0 is the intercept of the model which is the estimator of baseline incidence rate (I_0), α_j is the age effect, β is the vaccine effect and main parameter of interest, and $\exp(\beta)$ is the corresponding incidence rate ratio, k is an indicator equal to 1 for the risk period and 0 for the control periods, and t_{ijk} represents the amount of person-time during the interval. The realization of an adverse event was carried out under a binary distribution assumption with the probabilities specified in (11). The model (11) implies that the chance that an adverse event occurs in an interval is proportional to the person-time in that interval. When β_0 is small, P_{ijk} is much smaller than 1.0 although multiplied by t_{ijk}, and thus the chance for a subject to have more than one adverse event during the entire follow-up period is extremely low.

To assess the potential application range of our approach, β_0 was chosen to be -13, -14, and -15, which results in the average number of cases ranging from seventy to several hundreds; α was chosen to be zero for days 1–90 (reference group), 0.3 for days 91–270, and 0.1 for days 271–365; β was chosen to be 0.69, 1.39, and 1.79, which represents incidence rate ratios of 2, 4, and 6, respectively. Vaccination times were assumed to follow a normal distribution. However, we believe that the results apply to the situations where the vaccination in-take may have a different distribution. Adverse event dates were simulated by assuming a uniform distribution within the interval that the adverse event falls.

One thousand data sets with a population of 5 00 000 subjects for each data set were simulated for a variety of settings. Only those subjects with adverse events are used in SCCS analyses to identify the optimal risk window.

3.2. *Evaluation measures*

For each simulated data set, we estimated a series of incidence rate ratios using conditional Poisson regression model as in (1) by varying the length of the risk windows. We then applied the proposed approach to identify the optimal risk window. In each simulated data set, we call the specified risk window with the maximum $R(L)$, L_M, the optimal risk window because it was not feasible to examine 1000 plots for each combination of simulation parameters. Although this is not as ideal as examining each plot, we found the average estimated risk windows to be close to the truth in simulations. We evaluated our proposed statistical method by computing type I error rates, the mean of the estimated optimal risk windows and its standard deviation from 1000 data sets for each setting, and the percent bias from the true risk window.

3.2.1. Type I error rates. Under the null (i.e. with no risk window and $\beta=0$), the type I error rate is defined as the proportion of simulated data sets in which we identify (1) a positive linear relation between $R(L)$ and $1/T(L)$ when $L \geqslant L_M$, where L_M is the specified length of risk window with maximum $R(L)$; and (2) an apparent point at which visually the increasing trend of $R(L)$ changes by examining the plots of $R(L)$ versus $1/T(L)$ after a positive linear relation is identified between $R(L)$ and $1/T(L)$ when $L \geqslant L_M$.

3.2.2. Mean and standard deviation of the estimated optimal risk windows. The mean of the estimated optimal risk windows is calculated as the average of the 1000 estimates of the optimal risk windows. The standard deviation of the optimal risk windows is the standard deviation of the 1000 estimates of the optimal risk windows. The standard deviation was used to evaluate the precision of the approach.

3.2.3. Percent bias. The percent bias from the truth was used to evaluate the accuracy of our approach and was calculated as follows:

$$\text{Percent bias} = \frac{\text{mean estimate} - \text{true}}{\text{true}} \times 100$$

A negative percent bias implies underestimation of the true value and a positive percent bias represents overestimation. We also provided analogous measures for the incidence rate ratio estimate e^{β}, the main quantity of interest of a vaccine safety study after an optimal risk window is identified.

3.3. *Simulation results*

The type I error rates for falsely detecting a risk window when there truly is no elevated risk during the follow-up period are 4.3, 5.5, and 6.0 per cent for $\beta_0 = -13$, -14, and -15, respectively. The average number of cases in our simulations ranges from 70 to 817 (Table I). The total number of cases is primarily influenced by the value of β_0. Table I also shows the estimated risk window lengths and incidence rate ratios (and their standard deviations) for each simulation setting. For the medium risk window length (14 days), our approach produced mean percent bias of -5.7 to 5.0 per cent compared with the true risk window lengths, which represents a one day difference at most. Using the identified optimal risk window, the mean percent bias of R ranged from -2.5 to 15.0 per cent compared with the true incidence rate ratios. For the longer risk window length (42 days), our approach produced mean percent bias of -2.1 to 3.3 per cent, and -5.0 to 10 per cent, respectively, for the risk window lengths and the incidence rate ratios. For risk windows 14 and 42 days, the standard deviations of the estimated risk window lengths and relative incidence ratios increased with baseline incidence rates decreasing. The standard deviations of the risk window length decreased with true relative incidence ratios increasing.

Table I. Average number of cases ($\bar{n}$), mean optimal risk window length (standard deviation), and incidence rate ratios (standard deviation) by different true risk window lengths, intercept (β_0), and incidence rate ratios from 1000 simulated data sets.

True L_0 (in days)	True R	β_0	$\bar{n}$	Estimated L_0 (std)	Estimated R (std)
14	2.0	−13	517	13.5 (5.5)	2.1 (0.3)
		−14	189	13.2 (7.8)	2.2 (0.6)
		−15	70	13.4 (8.3)	2.3 (0.9)
	4.0	−13	560	13.2 (3.4)	3.9 (0.5)
		−14	206	13.9 (4.2)	3.9 (0.8)
		−15	72	14.7 (6.2)	4.1 (1.3)
	6.0	−13	606	13.3 (2.4)	6.0 (0.6)
		−14	223	13.4 (3.3)	5.9 (1.0)
		−15	81	14.5 (4.6)	6.0 (1.6)
42	2.0	−13	559	45.4 (8.2)	2.0 (0.2)
		−14	205	45.3 (9.5)	2.0 (0.3)
		−15	75	43.9 (11.9)	2.2 (0.6)
	4.0	−13	688	45.1 (1.6)	3.8 (0.3)
		−14	252	44.9 (2.7)	3.9 (0.5)
		−15	92	43.1 (5.4)	4.1 (0.9)
	6.0	−13	817	42.0 (0.3)	6.0 (0.5)
		−14	300	41.8 (1.2)	6.1 (0.8)
		−15	110	41.1 (3.4)	6.3 (1.3)

4. Examples

As an example, our proposed approach was applied to identify the optimal risk window length for a study of the association between measles–mumps–rubella (MMR) vaccination and ITP among children aged 12–23 months in the United Kingdom (U.K.) by Miller *et al.* [30]. The data were later updated in an SCCS tutorial by Whitaker *et al.* [26]. Thirty-five children were admitted to the hospital for ITP at least once and six were admitted more than once during the follow-up period. The risk period was defined to start immediately after MMR vaccination. The control period was defined as the time period before receipt of MMR plus the period after the specified risk window. In the tutorial, six age groups were used for a parametric model: 366–426 days, 427–487 days, 488–548 days, 549–609 days, 610–670 days, and 671–730 days. We used the same age groups and increased the length of the risk window by seven-day increment to obtain a set of incidence rate ratios. The upper panel of Figure 2 shows how the estimated incidence rate ratio for ITP changes in the presence of age effects when different risk window lengths are specified. The maximum incidence rate ratio is 3.28 (p-value=0.002) when $L=77$ days. There is an approximate linear relationship between $R(L)$ and $1/T(L)$ when $L>L_M$. Based on our method, it therefore appears that 77 days after vaccination is the empirically optimal risk window.

Because misspecification of the age groups can produce biased estimation of the association of ITP with MMR, we also used a semi-parametric model [26–28], in which the age effects were left unspecified, to obtain a set of incidence rate ratios. A corresponding plot $R(L)$ versus $1/T(L)$ is displayed in the lower panel of Figure 2 . Again, a risk window of 77 days after vaccination is suggested. The corresponding incidence rate ratio is 3.03 (p-value$=$0.006).

France *et al.* [31] also examined the association between MMR vaccination and ITP. Among a U.S. pediatric population and using a 42-day risk window after MMR vaccination, they found an incidence rate ratio of 7.06 for those vaccinated between 366 and 690 days. They excluded the 42-day healthy period immediately preceding MMR vaccination. We applied our approach to these U.S. MMR–ITP data but included all subjects with follow-up 366–730 days regardless of vaccination status and without excluding the healthy period before MMR vaccination. We used the same parametric and semi-parametric conditional Poisson regression models as in the prior example to estimate incidence rate ratios for a set of specified risk window lengths. For the parametric method, the maximum incidence rate ratio is 7.47 (p-value$<$0.0001) when $L=35$ days. However, the incidence rate ratio only changed slightly (7.41, p-value$<$0.0001) when $L=42$ days. From a vaccine safety point view, a risk window of 42 days is preferred as used by France *et al.* [31] (Figure 3). For the semi-parametric method, the maximum incidence rate ratio is 7.16 (p-value$<$0.0001) when $L=42$ days and there is an approximate

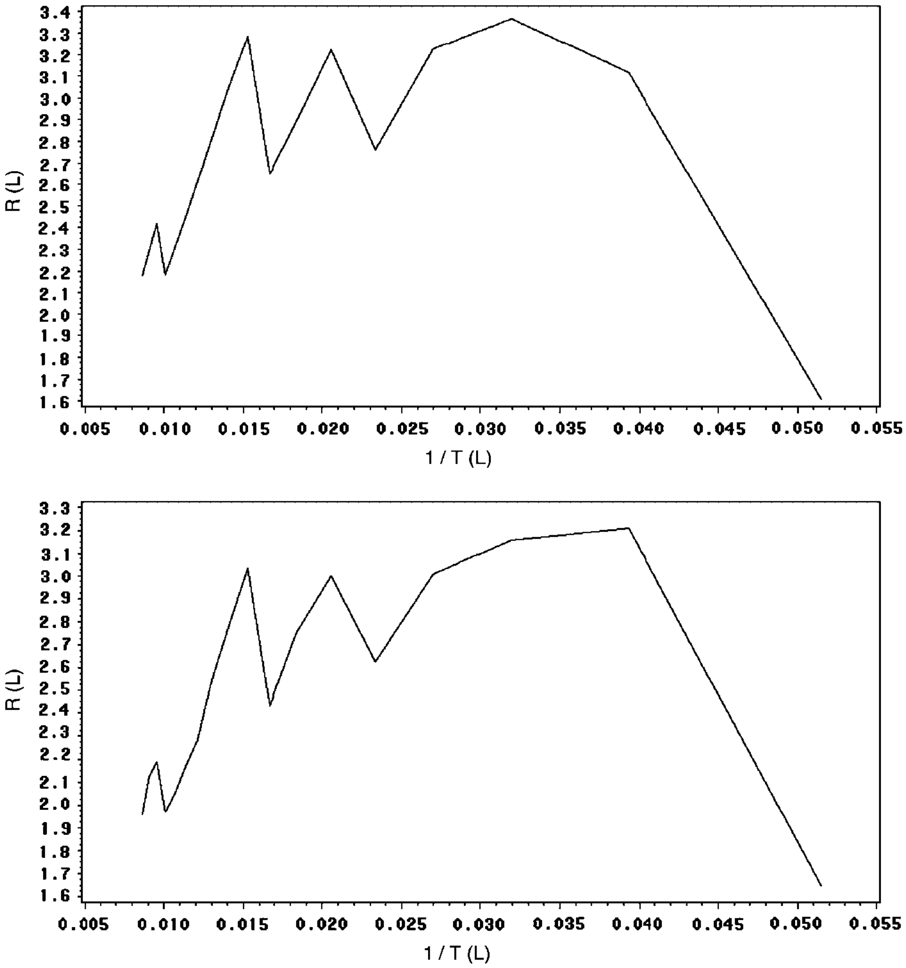

Figure 2. Estimated incidence rate ratios ($(R(L))$) for different values of $1/T(L)$ in the U.K. MMR–ITP data where $R(L)$s are estimated using a parametric (upper panel) and semi-parametric (lower panel) model.

linear relationship between $R(L)$ and $1/T(L)$ when $L>L_M$. We conclude that the optimal risk window for the U.S. MMR–ITP data is 42 days after MMR vaccination.

5. Discussion

Accurately defining the risk window length is an important step when conducting vaccine safety studies. Traditionally, risk windows are defined *a priori* and sometimes using arbitrary criteria. However, misspecifying the risk window length can lead to bias of vaccine risk estimates. Our data-based method can effectively identify the optimal lengths for medium and long risk windows. Additionally, time-varying covariates can be accommodated simply by including them in conditional Poisson regression models. Our theoretical work and data application also showed that this approach is robust to misspecification of age effects. Because the proposed method relies on choosing $L>L_0$, it does require some advance knowledge of the order of magnitude of L_0, which may be available from previous studies or biological understanding of how the vaccines work (e.g. 14 days as the risk window for influenza vaccine studies [6, 7]). In addition to vaccine safety studies, our approach can also be used to identify the optimal risk window in vaccine effectiveness studies (i.e. when the vaccine effects are protective). The linear relation between $R(L)$ and $1/T(L)$ will be similar when $L>L_0$ except that it will be negative due to the protective effect.

Although theoretically it is true, the risk window with the maximum incidence ratio may not be optimal due to sparse data problem. We recommend using both the maximum incidence rate ratio and the linear relationship between the estimated incidence rate ratios and risk window lengths when

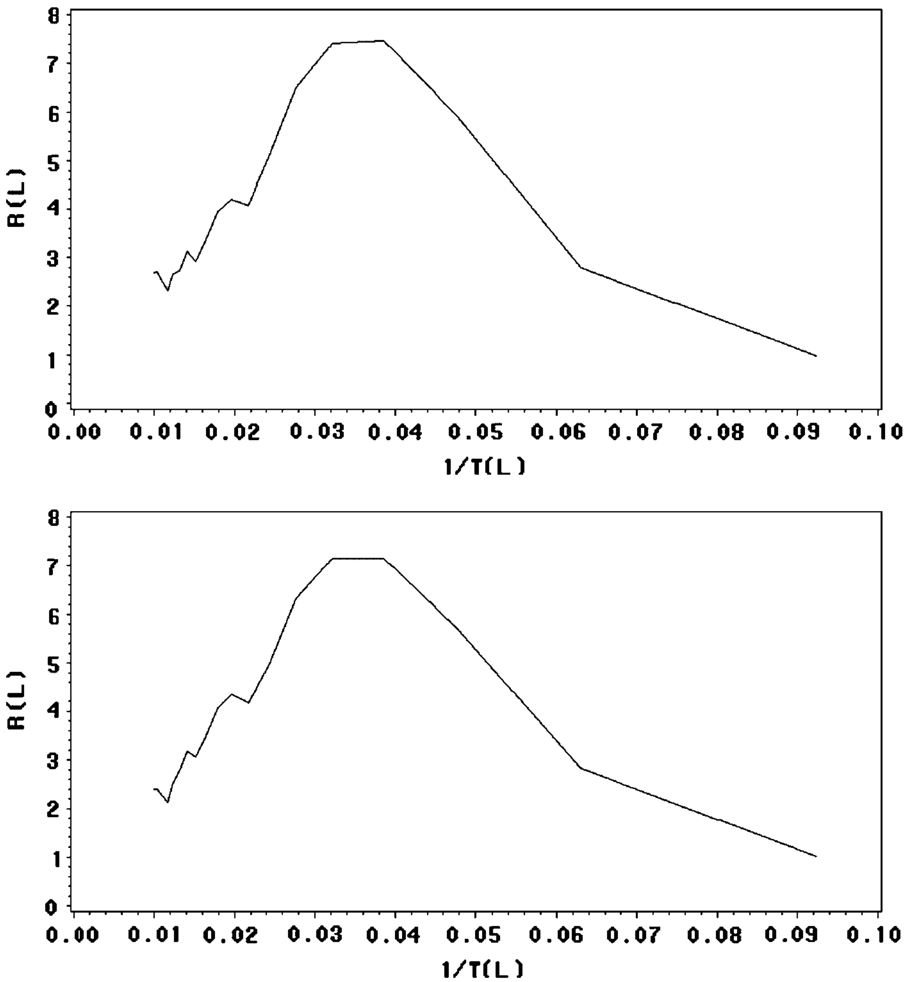

Figure 3. Estimated incidence rate ratios ($R(L)$) for different values of $1/T(L)$ in the U.S. MMR–ITP data where $R(L)$s are estimated using a parametric (upper panel) and semi-parametric (lower panel) model.

they are longer than the true risk window to identify the optimal empirical risk window length in the following way. First, plot the estimated incidence rate ratios ($R(L)$, either from parametric or from non-parametric models) versus $1/T(L)$. Then use the simple method of identifying the risk window (L_M) with the maximum incidence rate ratio, R_M, which our simulations show performs quite well in most circumstances. If the plot shows an approximate linear relationship when $L > L_M$, then L_M is the optimal empirical risk window. If not, then use the change-point method (i.e. identify the region where the estimated incidence rate ratios $R(L)$ and $1/T(L)$ have a linear relationship and find the decreasing change point). We recognize that this method provides helpful guidance but may not find the risk window exactly. In case the two methods suggest different risk windows, one should consider both risk windows along with biological plausibility and prior knowledge to define the risk window.

In this paper, we assumed that the risk window began immediately after vaccination. Although most vaccine safety studies are conducted under that assumption, in reality this may not be true. Our proposed method can be modified to accommodate risk windows that do not start immediately after vaccination by taking a two-step approach. First, the end of risk window can be identified under the assumption that the risk window starts immediately after vaccination. Then by varying the start time, a series of incidence rate ratios can be obtained. Similarly, the start time may be identified by plotting the $R(L)$ and $1/T(L)$.

One may apply smoothing techniques when there is the sawtooth variation as in Figure 2. One commonly used method involves non-parametric estimation of regression surfaces [32, 33] which is implemented in the SAS LOESS procedure [25]. The proper level of smoothness can be determined by assessing the match between the predicted and observed values and examining the scatter plots of the fitted residuals versus $1/T(L)$. However, the optimal risk window length only changed slightly after we smoothed the plots for the U.K. MMR–ITP example.

Because our approach is data-driven, the identified optimal risk window lengths may be dependent on the given data. Thus, cross validation with a training and test data set or using a different population (e.g. different MMR–ITP data sets) is important. However, the approach using a training and test data set is feasible only when the number of cases is relatively large, which may not be the case for many vaccine safety studies that involve rare adverse events. The two MMR–ITP data sets that we evaluated resulted in two very different risk windows (77 and 42 days) and different incidence rate ratios (3.28 and 7.16 for the semi-parametric method). These differences may be due to the underlying difference between these two populations or sampling variability. In the latter case one may choose to combine the two populations to ensure that the optimal risk window length and the incidence rate ratio are more broadly represented.

One limitation of our approach is that it is not suitable for short risk windows (less than seven days) because the identified risk window lengths tend to be longer than the true risk window (data not shown). Another limitation is that our approach gives a single incidence rate ratio for an adverse event for the entire risk period and is not designed to identify multiple risk windows after vaccination. In simulation studies, we only considered constant incidence rate ratio and did not study the robustness of this approach to multiple levels of risks. However, the MMR–ITP data applications show that the estimated optimal risk window lengths are very consistent with the visual perception. Although Miller *et al.* [30] showed that the incidence rate ratios differed for 0–14, 15–28, and 29–42 days after MMR vaccination in the U.K. data, it is also desirable to have an overall incidence rate ratio estimate and our approach permits this. The future research should focus on how to examine the existence of multiple risk levels within the identified optimal risk window and how to discretize and estimate the risk levels within the identified optimal risk window.

In summary, despite some limitations, our proposed method can provide vaccine safety researchers a useful tool to define appropriate risk windows while considering biological understanding of how the vaccines work or information from prior studies. In addition to vaccine, our method also has the potential to be used in safety studies of other medical products [34].

Acknowledgements

This study was supported by the Centers for Disease Control and Prevention via contract 200-2002-00732 (the Vaccine Safety Datalink Project) with America's Health Insurance Plans. The authors are very thankful to the two reviewers for their comments that helped to improve the manuscript.

References

1. Chen RT, Glasser JW, Rhodes PH, Davis RL, Barlow WE, Thompson RS, Mullooly JP, Black SB, Shinefield HR, Vadheim CM, Marcy SM, Ward JI, Wise RP, Wassilak SG, Hadler SC. The vaccine safety datalink project: a new tool for improving vaccine safety monitoring in the United States. *Pediatrics* 1997; **99**:765–73. DOI: 10.1542/peds.99.6.765.
2. Fine PM, Chen RT. Confounding in studies of adverse reactions to vaccines. *American Journal of Epidemiology* 1992; **136**:121–135.
3. DeStefano F, Gu D, Kramarz P, Truman BI, Iademarco BF, Mullooly JP, Jackson LA, Davis RL, Black SB, Shinefield HR, Marcy SM, Ward JI, Chen RT. Childhood vaccinations and risk of asthma. *Pediatric Infectious Diseases Journal* 2002; **21**(6):498–504. DOI: 10.1097/00006454-200206000-00004.
4. Chen RT, Davis RL, Sheedy KM. Safety of immunizations. In *Vaccines*, Chapter 61 (4th edn), Plotkin SA, Orenstein WA (eds). WB Saunders Company: Philadelphia, PA, 2004.
5. Eriksen EM, Perlman JA, Miller A, Marcy SM, Lee H, Vadheim C, Zangwill KM, Chen RT, DeStefano F, Lewis E, Black S, Shinefield H, Ward JI. Lack of association between hepatitis B birth immunization and neonatal death: a population-based study from the Vaccine Safety Datalink Project. *The Pediatric Infectious Disease Journal* 2004; **23**(7):656–662. DOI: 10.1097/01.inf.0000130953.08946.d0.
6. France EK, Glanz JM, Xu S, Davis RL, Black SB, Shinefield HR, Zangwill KM, Marcy SM, Mullooly JP, Jackson LA, Chen R. Safety of the trivalent inactivated influenza vaccine among children: a population-based study. *Archives of Pediatrics and Adolescent Medicine* 2004; **158**:1031–36. DOI: 10.1001/archpedi.158.11.1031.
7. Hambidge SJ, Glanz JM, France EK, McClure D, Xu S, Yamasaki K, Jackson L, Mullooly JP, Zangwill KM, Marcy SM, Black SB, Lewis EM, Shinefield HR, Belongia E, Nordin J, Chen RT, Shay DK, Davis RL, DeStefano F, Vaccine Safety Datalink Team. Safety of trivalent inactivated influenza vaccine in children 6 to 23 months old. *Journal of the American Medical Association* 2006; **296**(16):1990–1997. DOI: 10.1001/jama.296.16.1990.
8. France EK, Glanz J, Xu S, Hambidge S, Yamasaki K, Black SB, Marcy M, Mullooly JP, Jackson LA, Nordin J, Belongia EA, Hohman K, Chen RT, Davis R, Vaccine Safety Datalink Team. Risk of immune thrombocytopenic purpura after measles–mumps–rubella immunization in children. *Pediatrics* 2008; **121**(3)e:687–92. DOI: 10.1542/peds.2007-1578.

9. Farrington CP. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics* 1995; **51**:228–235. DOI: 10.2307/2533328.
10. Glanz JM, McClure DL, Xu S, Hambidge SJ, Lee M, Kolczak MS, Kleinman K, Mullooly JP, France EK. Four different study designs to evaluate vaccine safety were equally validated with contrasting limitations. *Journal of Clinical Epidemiology* 2006; **59**:808–818. DOI: 10.1016/j.jclinepi.2005.11.012.
11. Lerman PM. Fitting segmented regression models by grid search. *Applied Statistics* 1980; **29**:77–84. DOI: 10.2307/2346413.
12. Kim HJ, Siegmund D. The likelihood ratio test for a change-point in simple linear regression. *Biometrika* 1989; **76**:409–423. DOI: 10.1093/biomet/76.3.409.
13. Bai J. Estimating multiple breaks one at a time. *Econometric Theory* 1997; **13**:315–352. DOI: 10.1017/S0266466600005831.
14. Bai J. Likelihood ratio tests for multiple structural changes. *Journal of Econometrics* 1999; **91**:299–323. DOI: 10.1016/S0304-4076(98)00079-7.
15. Bai J, Perron P. Estimating and testing linear models with multiple structural change. *Econometrica* 2003; **66**:47–48. DOI: 10.2307/2998540.
16. Matthews DE, Farewell VT. On testing for a constant hazard against a change-point alternative (Corr: V41 p1103). *Biometrics* 1982; **38**:463–468. DOI: 10.2307/2530460.
17. Nguyen GS, Rogers G, Walker EA. Estimation in change-point hazard rate models. *Biometrics* 1984; **71**:299–304. DOI: 10.1093/biomet/71.2.299.
18. Luo X. The asymptotic distribution of MLE of treatment lag threshold. *Journal of Statistical Planning and Inference* 1996; **53**:33–61. DOI: 10.1016/0378-3758(95)00142-5.
19. Luo X, Turnbull B, Clark L. Likelihood ratio tests for a changepoint with survival data. *Biometrika* 1997; **84**:555–565. DOI: 10.1093/biomet/84.3.555.
20. Pons O. Estimation in a Cox regression model with a change-point at an unknown time. *Statistics* 2002; **36**:101–124. DOI: 10.1080/02331880212043.
21. Gijbels I, Gurler U. Estimation of a change point in a hazard function based on censored data. *Lifetime Data Analysis* 2003; **9**:395–411. DOI: 10.1023/B:LIDA.0000012424.71723.9d.
22. Liu M, Lu W, Shao Y. A Monte Carlo approach for change-point detection in the Cox proportional hazards model. *Statistics in Medicine* 2008; **27**(19):3894–909. DOI: 10.1002/sim.3214.
23. Farrington CP, Nash J, Miller E. Case series analysis of adverse reactions to vaccines: a comparative evaluation. *American Journal of Epidemiology* 1996; **143**:1165–73.
24. Whitaker HJ, Hocine MN, Farrington CP. The methodology of self-controlled case series studies. *Statistical Methods in Medical Research* 2009; **18**:7–26. DOI: 10.1177/0962280208092342.
25. SAS 9.1 SAS Institute Inc., Cary, NC, U.S.A., 2002–2003.
26. Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. *Statistics in Medicine* 2006; **25**(10):1768–1797. DOI: 10.1002/sim.2302.
27. Farrington CP, Whitaker HJ. Semiparametric analysis of case series data (with Discussion). *Applied Statistics* 2006; **155**:553–594. DOI: 10.1111/j.1467-9876.2006.00554.x.
28. Xu S, Gargiullo P, Mullooly J, McClure D, Hambidge S, Glanz J. Fitting parametric and semi-parametric conditional Poisson regression models with Cox's partial likelihood in self-controlled case series and matched cohort studies. *Journal of Data Science* 2010; **8**:349–360.
29. Musonda P. The self-controlled case series method: performance and design in studies of vaccine safety. *Ph.D. Thesis*, The Open University, December 2006.
30. Miller E, Waight P, Farrington P, Stowe J, Taylor B. Idiopathic thrombocytopenic purpura and MMR vaccine. *Archives of Disease in Childhood* 2001; **84**:227–229. DOI: 10.1136/adc.84.3.227.
31. France EK, Glanz JM, Xu S, Hambidge S, Yamasaki K, Black SB, Marcy M, Mullooly J, Jackson L, Nordin J, Belongia E, Hohman K, Chen RT, Davis R. Risk of immune thrombocytopenic purpura after measles–mumps–rubella immunization in children. *Pediatrics* 2008; **121**:687–692. DOI: 10.1542/peds.2007-1578.
32. Cleveland WS, Devlin SJ, Grosse E. Regression by local fitting methods, properties, and computational algorithms. *Journal of Econometrics* 1988; **37**:87–114. DOI: 10.1016/0304-4076(88)90077-2.
33. Cleveland WS, Grosse E. Computational methods for local regression. *Statistics and Computing* 1991; **1**:47–62. DOI: 10.1007/BF01890836.
34. Platt R, Wilson M, Chan A, Benner J, Marchibroda J, McClellan M. The new sentinel network: improving the evidence of medical-product safety. *The New England Journal of Medicine* 2009; **361**:645–647. DOI: 10.1056/NEJMp0905338.