Inclusion of unexposed subjects improves the precision and power of self-controlled case series method

Xiangmei Ma [a], K. F. Lam [a,b], Yin Bun Cheung [a,c,d,*]

a. Centre for Quantitative Medicine, Duke-NUS Medical School, 8 College Road, Outram Park, Singapore 169857

b. Department of Statistics and Actuarial Science, University of Hong Kong, Pokfulam Road, Hong Kong, China

c. Programme in Health Services & Systems Research, Duke-NUS Medical School, 8 College Road, Outram Park, Singapore 169857

d. Tampere Center for Child, Adolescent and Maternal Health Research, Tampere University, Arvo Ylpön katu 34, Tampere 33520, Finland

* Correspondence:
Professor Yin Bun Cheung, Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, 8 College Road, Outram Park, Singapore 169857
yinbun.cheung@duke-nus.edu.sg
Tel: +65 9858 9470

Abstract

The self-controlled case series is an important method in the studies of the safety of biopharmaceutical products. It uses the conditional Poisson model to make comparison within persons. In models without adjustment for age (or other time-varying covariates), cases who are never exposed to the product do not contribute any information to the estimation. We provide analytic proof and simulation results that the inclusion of unexposed cases in the conditional Poisson model with age adjustment reduces the asymptotic variance of the estimator of the exposure effect and increases power. We re-analysed a vaccine safety dataset to illustrate.

Key words: Asymptotic variance; conditional Poisson model; drug safety; self-controlled case series; vaccine safety

Word count: 3417

1. Introduction

The self-controlled case series is a popular method in the studies of vaccine safety (Whitaker et al. 2006). It has also been recommended as a major alternative to case-control study design in drug safety research using electronic healthcare databases (Schuemie et al. 2019). The estimation of the effect of an exposure to a risk factor is based on within-person comparison. Given that a person has experienced the outcome event, the method evaluates whether the event occurrence was more likely during the person-time exposed or unexposed to the risk factor. Therefore, it has the advantage of being unconfounded by time-constant covariates. The method is based on fitting a conditional Poisson regression model (Whitaker et al. 2006; Weldeselassie et al. 2011; Xu et al. 2011). The model does not require data from persons who have not experienced the outcome event (i.e. non-cases) because they have no contribution to the conditional likelihood. The conditional Poisson model is equivalent to the fixed effects Poisson model as their likelihoods are proportional (Xu et al. 2011). In drug safety research there is also a large group of methods known as signal detection algorithms, safety data mining or proportional reporting analysis (Almenoff et al. 2007; DuMouchel 1999; Harpaz et al. 2012). We only mention them in passing as their study design/data source (spontaneous reporting system) and research purpose (screening for drug-event combinations) are different from that of the self-controlled case series.

Since the estimation of the exposure effect is based on within-person comparison, cases who are never exposed during the study period are supposed to contribute no information to the estimation of the exposure effect. Only exposure variables that vary within persons should be included in the model (Gardiner et al. 2009). However, within-unit comparisons almost always require statistical adjustment for age, which is often related to both disease outcomes and medical exposures (Musonda et al. 2008; Whitaker et al. 2006; Xu et al. 2012), or some other time-related factors. Such adjustment is seen in SCCS studies of people at different age ranges (e.g., Peach et al. 2021; Whitaker et al. 2006). Inclusion of cases who were never exposed during the study period in the conditional Poisson model can help to reduce bias arising from confounding by age or other time-varying covariates (Whitaker et al. 2006). This has been demonstrated by simulation (Musonda et al. 2008). The inclusion of these cases is sensible only if it can be assumed that the underlying pattern of covariate effects is the same for them and the other cases.

We identified and reviewed 44 medical studies published in 2020 that used SCCS to evaluate the safety/effect of interventions/exposures. (Details of the search and review are provided in Online Supplementary Material 1.) Of the 44 studies, 11 did and 22 did not

include subjects who were unexposed, 9 were “hybrid”, and 2 were indeterminable. Each of the 9 “hybrid” studies involved more than one intervention/exposure. Subjects must be exposed to at least one of the interventions/exposures for them to be included in the studies, but they could be included in the analysis of an intervention/exposure without having specifically exposed to it.

Most studies did not provide the rationale of the inclusion or exclusion. However, in their comments on SCCS, three studies that did not include unexposed subjects explicitly made the following statements: “To be included in the SCCS methods, individuals must experience the exposure and the outcome of interest” (Duncan et al. 2020), “patients must have both the outcome and the exposure of interest” (Aspinall et al. 2020), and “Comparisons are made within individuals rather than between individuals. Thus, only those who have experienced both the outcome and the exposure of interest are included” (Forbes et al. 2020). Although the SCCS is increasingly popular in medical research, there is a need to promote proper understanding of its properties and usage.

For brevity, we will refer to cases who were never exposed and cases who had some exposed person-time and some unexposed person-time during their observation periods as “unexposed cases” and “exposed cases”, respectively. Without loss of generality, we will consider age as the time-varying covariate, but there can be other time-varying covariates such as calendar time or season (Moulton et al. 2006; Peach et al. 2021; Whitaker et al. 2006).

While the statistics literature has shown that the inclusion of unexposed cases in the conditional Poisson model is beneficial in terms of reducing bias, little attention has been given to the question of whether the inclusion can improve the precision of the estimate of the exposure effect. Precision is the inverse of the variance of the estimator. In this article, we provide analytic proof that the inclusion can improve the precision, demonstrate the impact by simulation, and illustrate the difference by re-analysis of a vaccine safety study with and without inclusion of unvaccinated cases.

2. Model, likelihood and variance

2.1. No time-varying covariate

We begin with the simplest scenario that there is only one binary exposure variable and there is no time-varying covariate in the model. For the time being, person-time is either exposed or unexposed (i.e. before or after the occurrence of an exposure). Suppose there are N cases.

Each case's observation period is $(a_i, b_i], i = 1,2, \ldots, N$. Suppose that case i experiences n_i events in total within $(a_i, b_i]$. Denote the exposure indicator as $k = 1$ for an exposed period and $k = 0$ for an unexposed period.

With the convention that $0^0 = 1$, the conditional Poisson likelihood is:

$$L \propto \prod_{i=1}^{N} \prod_{k=0}^{1} \left(\frac{\exp(\beta k)\, \tau_{ik}}{\sum_{k'=0}^{1} \exp(\beta k')\, \tau_{ik'}} \right)^{n_{ik}}$$

where τ_{ik} is the duration of the time case i spent in exposure status k, n_{ik} is the number of events that occurred in this period, and $\sum_{k=0}^{1} n_{ik} = n_i$ (Xu et al. 2011, 2012). The parameter β represents the effect of the exposure of interest in terms of log incidence rate ratio, also called log relative incidence in safety research.

Let $l = \log(L)$ be the log-likelihood function:

$$l = \log(L) = \sum_{i=1}^{N} \sum_{k=0}^{1} n_{ik} \left[\log(\tau_{ik}) + \beta k \right] - \sum_{i=1}^{N} n_i \left[\log \left(\sum_{k'=0}^{1} \exp(\beta k')\, \tau_{ik'} \right) \right] + constant.$$

The second derivative of l with respect to β is

$$\frac{\partial^2 l}{\partial \beta^2} = - \sum_{i=1}^{N} \frac{n_i \tau_{i0}\, \tau_{i1} \exp(\beta)}{[\tau_{i0} + \exp(\beta)\, \tau_{i1}]^2}.$$

The inverse of the Fisher information matrix, $I(\beta) = -\frac{\partial^2 l}{\partial \beta^2}$, is an estimator of the asymptotic variance:

$$Var(\tilde{\beta}) = [I(\beta)]^{-1} = \left[\sum_{i=1}^{N} \frac{n_i \tau_{i0}\, \tau_{i1} \exp(\beta)}{(\tau_{i0} + \exp(\beta)\, \tau_{i1})^2} \right]^{-1} . \qquad (1)$$

If case i is unexposed for the entire observation period, $\tau_{i1} = 0$ (no exposed time) and the case contributes no information to the variance estimator. Therefore, in the absence of covariates to adjust for, the inclusion of unexposed cases does not affect the variance of the estimator of the exposure effect, i.e. $Var(\hat{\beta}) = Var(\tilde{\beta})$, where $\hat{\beta}$ and $\tilde{\beta}$ are the estimators of the exposure effect in analysis involving exposed cases only and analysis involving all cases, respectively.

2.2. Simplified scenario with a time-varying covariate

Consider a time-varying covariate such as age, with $J + 1$ age intervals over $(\min(a_i), \max(b_i)]$. Each case's person-time can be partitioned into intervals jointly defined by age and exposure. The likelihood becomes:

$$L \propto \prod_{i=1}^{N} \prod_{j=0}^{J} \prod_{k=0}^{1} \left[\frac{\exp(\alpha_j + \beta k)\tau_{ijk}}{\sum_{j'=0}^{J} \sum_{k'=0}^{1} \exp(\alpha_{j'} + \beta k')\tau_{ij'k'}} \right]^{n_{ijk}}, \qquad (2)$$

where τ_{ijk} is duration of time case i spent in age interval j with exposure status k, n_{ijk} is the number of events occurred in this interval, and $\sum_{j=0}^{J} \sum_{k=0}^{1} n_{ijk} = n_i$. The parameters α_j represent age effects for intervals $j = 1,2,\dots,J$, relative to the reference age interval 0, with $\alpha_0 = 0$.

The log-likelihood function is:

$$l = \log(L) = \sum_{i=1}^{N} \sum_{j=0}^{J} \sum_{k=0}^{1} n_{ijk} \left[\log(\tau_{ijk}) + \alpha_j + \beta k \right] - \sum_{i=1}^{N} n_i \left[\log \left(\sum_{j'=0}^{J} \sum_{k'=0}^{1} \exp(\alpha_{j'} + \beta k')\tau_{ij'k'} \right) \right] + constant.$$

We begin with a simplified scenario that the time-varying covariate (age) has two levels, i.e. $J = 1$. The log-likelihood function is

$$l = \sum_{i=1}^{N} n_{i00}[\log(\tau_{i00})] + n_{i01}[\log(\tau_{i01}) + \beta] + n_{i10}[\log(\tau_{i10}) + \alpha_1] + n_{i11}[\log(\tau_{i11}) + \alpha_1 + \beta] - \sum_{i=1}^{N} n_i[\log(\tau_{i00} + \exp(\beta)\tau_{i01} + \exp(\alpha_1)\tau_{i10} + \exp(\alpha_1 + \beta)\tau_{i11})].$$

The first derivative of l with respect to β is

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{N} \sum_{j=0}^{1} n_{ij1} - \sum_{i=1}^{N} n_i \frac{\exp(\beta)\tau_{i01} + \exp(\alpha_1 + \beta)\tau_{i11}}{\tau_{i00} + \exp(\beta)\tau_{i01} + \exp(\alpha_1)\tau_{i10} + \exp(\alpha_1 + \beta)\tau_{i11}},$$

and the first derivative of l with respect to α_1 is

$$\frac{\partial l}{\partial \alpha_1} = \sum_{i=1}^{N} \sum_{k=0}^{1} n_{i1k} - \sum_{i=1}^{N} n_i \frac{\exp(\alpha_1)\tau_{i10} + \exp(\alpha_1 + \beta)\tau_{i11}}{\tau_{i00} + \exp(\beta)\tau_{i01} + \exp(\alpha_1)\tau_{i10} + \exp(\alpha_1 + \beta)\tau_{i11}}.$$

Without loss of generality, we assume that the first N_0 cases are unexposed during the entire follow-up. Denote l_0 and l_1 as the contributions to the log-likelihood by unexposed and exposed persons who have experienced the outcome event, respectively, with $l = l_0 + l_1$. If case i is unexposed during his/her entire observation period, $n_{ij1} = 0$ and $\tau_{ij1} = 0$ for all j, and $\frac{\partial l_0}{\partial \beta} = 0$. Therefore, given an estimate for α_1, unexposed cases contribute no information to the point estimate for the exposure effect. However,

$$\frac{\partial l_0}{\partial \alpha_1} = \sum_{i=1}^{N_0} n_{i10} - \sum_{i=1}^{N_0} \frac{n_i \exp(\alpha_1)\,\tau_{i10}}{\tau_{i00} + \exp(\alpha_1)\,\tau_{i10}} = \sum_{i=1}^{N_0} \frac{n_{i10}\tau_{i00} - n_{i00}\exp(\alpha_1)\,\tau_{i10}}{\tau_{i00} + \exp(\alpha_1)\,\tau_{i10}}$$

with $n_{i00} + n_{i10} > 0$ and $\tau_{i00} + \tau_{i10} = b_i - a_i > 0$. If an unexposed case only has person-time in one age level, then either $(n_{i10} = 0 \cap \tau_{i10} = 0)$ or $(n_{i00} = 0 \cap \tau_{i00} = 0)$, making no contribution to the point estimate of the age effect, or vice versa.

The second derivative of l with respect to β is

$$B = \frac{\partial^2 l}{\partial \beta^2} = -\sum_{i=1}^{N} n_i \left[\frac{\exp(\beta)\,\tau_{i01} + \exp(\alpha_1 + \beta)\,\tau_{i11}}{\tau_{i00} + \exp(\beta)\,\tau_{i01} + \exp(\alpha_1)\,\tau_{i10} + \exp(\alpha_1 + \beta)\,\tau_{i11}} - \frac{(\exp(\beta)\,\tau_{i01} + \exp(\alpha_1 + \beta)\,\tau_{i11})^2}{(\tau_{i00} + \exp(\beta)\,\tau_{i01} + \exp(\alpha_1)\,\tau_{i10} + \exp(\alpha_1 + \beta)\,\tau_{i11})^2} \right].$$

The second derivative of l with respect to α_1 is

$$A = \frac{\partial^2 l}{\partial \alpha_1^2} = -\sum_{i=1}^{N} n_i \left[\frac{\exp(\alpha_1)\,\tau_{i10} + \exp(\alpha_1 + \beta)\,\tau_{i11}}{\tau_{i00} + \exp(\beta)\,\tau_{i01} + \exp(\alpha_1)\,\tau_{i10} + \exp(\alpha_1 + \beta)\,\tau_{i11}} - \frac{(\exp(\alpha_1)\,\tau_{i10} + \exp(\alpha_1 + \beta)\,\tau_{i11})^2}{(\tau_{i00} + \exp(\beta)\tau_{i01} + \exp(\alpha_1)\,\tau_{i10} + \exp(\alpha_1 + \beta)\,\tau_{i11})^2} \right]$$

and

$$D = \frac{\partial^2 l}{\partial \beta \partial \alpha_1} = \frac{\partial^2 l}{\partial \alpha_1 \partial \beta} = -\sum_{i=1}^{N} n_i \left[\frac{\exp(\alpha_1 + \beta)\,(\tau_{i11}\tau_{i00} - \tau_{i01}\tau_{i10})}{(\tau_{i00} + \exp(\beta)\,\tau_{i01} + \exp(\alpha_1)\,\tau_{i10} + \exp(\alpha_1 + \beta)\,\tau_{i11})^2} \right].$$

Let $B = B_1 + B_0$, $A = A_1 + A_0$, $D = D_1 + D_0$, where B_1, A_1, D_1 are the contribution of exposed cases and B_0, A_0, D_0 are the contribution of unexposed cases. Since $\tau_{i01} = \tau_{i11} = 0$ for the unexposed cases, $B_0 = \frac{\partial^2 l_0}{\partial \beta^2} = 0$ and $D_0 = \frac{\partial^2 l_0}{\partial \beta \partial \alpha_1} = 0$.

$$A_0 = \frac{\partial^2 l_0}{\partial \alpha_1^2} = -\sum_{i=1}^{N_0} n_i \left[\frac{\exp(\alpha_1)\,\tau_{i10}}{\tau_{i00} + \exp(\alpha_1)\,\tau_{i10}} - \frac{(\exp(\alpha_1)\,\tau_{i10})^2}{(\tau_{i00} + \exp(\alpha_1)\,\tau_{i10})^2} \right]$$

$$= \sum_{i=1}^{N_0} n_i \left[\frac{(\exp(\alpha_1)\,\tau_{i10})^2}{(\tau_{i00} + \exp(\alpha_1)\,\tau_{i10})^2} - \frac{\exp(\alpha_1)\,\tau_{i10}}{\tau_{i00} + \exp(\alpha_1)\,\tau_{i10}} \right]$$

$$= \sum_{i=1}^{N_0} n_i (r_i^2 - r_i)$$

where $r_i = \frac{\exp(\alpha_1)\tau_{i10}}{\tau_{i00}+\exp(\alpha_1)\tau_{i10}}$. Since $\tau_{i00} \geq 0$ and $\tau_{i10} \geq 0$, it follows that $0 \leq r_i \leq 1$.

Furthermore, $r_i^2 - r_i = \left(r_i - \frac{1}{2}\right)^2 - \frac{1}{4} \leq 0$.That is $A_0 \leq 0$ and $A = A_1 + A_0 \leq A_1$. If an unexposed case only has person-time in one age level, then either $\tau_{i10} = 0$ or $\tau_{i00} = 0$. That gives either $r_i = 1$ or $r_i = 0$. In this case, $A_0 = 0, A = A_1$.

The Fisher information matrix is

$$I(\beta, \alpha_1) = -\begin{pmatrix} \frac{\partial^2 l}{\partial \beta^2} & \frac{\partial^2 l}{\partial \beta \partial \alpha_1} \\ \frac{\partial^2 l}{\partial \alpha_1 \partial \beta} & \frac{\partial^2 l}{\partial \alpha_1^2} \end{pmatrix} = -\begin{pmatrix} B & D \\ D & A \end{pmatrix}.$$

The variance-covariance matrix is

$$I(\beta, \alpha_1)^{-1} = -\frac{1}{BA - D^2}\begin{pmatrix} A & -D \\ -D & B \end{pmatrix}, \text{ if } BA - D^2 \neq 0.$$

Therefore, $Var(\tilde{\beta}) = -\frac{A}{BA-D^2}$.

Define a function $y(x) = -\frac{x}{bx-d^2}$. Its first derivative $y'(x) = \frac{\partial y(x)}{\partial x} = \frac{d^2}{(bx-d^2)^2} \geq 0$. So $y(x) \leq y(x_0)$ if $x \leq x_0$. Since $A \leq A_1$ and $B_0 = D_0 = 0$, we have $-\frac{A}{BA-D^2} \leq -\frac{A_1}{BA_1-D^2} = -\frac{A_1}{B_1A_1-D_1^2}$. Therefore, $Var(\tilde{\beta}) \leq Var(\hat{\beta})$. Simulation studies have demonstrated that the exposure effect was more accurately estimated by inclusion of the unexposed cases in the adjustment for age effects (Musonda et al. 2008; Whitaker et al. 2006). If $\hat{\beta}$ and $\tilde{\beta}$ differ substantially, one would trust $\tilde{\beta}$ over $\hat{\beta}$ and the comparison of $Var(\hat{\beta})$ and $Var(\tilde{\beta})$ is not meaningful.

Note that the analytic proof does not depend on the size of α_j. It concerns conditional Poisson models that include adjustment for time-varying covariates. The results hold no matter if the time-varying covariates have any effects on the outcome event or not.

2.3. Generalizations

For brevity we have focused on unexposed cases. But the same conclusion can be made for cases who are always exposed during the study period. For example, in equation (1), cases who are always exposed have $\tau_{i0} = 0$ instead of $\tau_{i1} = 0$. But in either situation, the product $\tau_{i0}\tau_{i1} = 0$ and there is no contribution to the precision of the estimate.

The analysis above has assumed that person-time is classified as either before or after the occurrence of an exposure. In other words, the duration of the altered risk level after an exposure is indefinite (Musonda et al. 2008; Weldeselassie et al. 2011). In safety

investigation, it is quite common to assume that the duration is definite, say 28 days in vaccine studies or longer durations in studies of repeated drug prescriptions. Afterwards, the risk level returns to the pre-exposure level. In Section 1 of Online Supplementary Material 2, we show that the conclusions remain unchanged in the situation of the exposure giving rise only to a duration-limited risk interval.

Studies may also consider more than one level of elevated risk after exposure. For example, incidence rate ratios may be estimated for the first X days and second X days after exposure as compared to the pooled person-time before exposure and 2X days after exposure. In Section 2 of Online Supplementary Material 2 we show that the variance of the estimator for the two incidence rate ratios based on all cases are smaller than or equal to that based only on exposed cases.

The above only considered two age intervals and up to two duration-limited risk periods. In Section 3 of Online Supplementary Material 2 we provide the proof for three age intervals. In Section 4 the proof is generalized to multiple age intervals and multiple duration-limited risk periods.

In the next section we use simulation to demonstrate the above property in a variety of settings.

3. Simulation study

3.1. Simulation settings

We carried out a simulation study for demonstration of the above analytic findings, adopting some simulation methods from a previous work on the conditional Poisson model (Musonda et al. 2008). We assumed all subjects were observed for 500 days. We considered sample size of exposed cases n = 100 or 250. We considered the length of the risk period following exposure to be either 200 days or 28 days, similar to situations of repeated drug prescriptions or one dose of vaccination, respectively. We generated age at exposure from the beta distribution on [0, 500] with mean age 150 days and standard deviation 50 days or 100 days. The patterns of age at exposure are shown in Figure 1.

[Figure 1 about here]

The observation period was partitioned into 5 age intervals (100 days each). Four patterns of age effects were considered. The incidence rate ratios, e^{α_j}, from the youngest to oldest age intervals were, respectively:

1) No age effect, i.e., $e^{\alpha_j} = 1$ for all 5 age intervals.
2) Symmetric age effect: 1, 1.5, 2, 1.5 and 1.
3) Monotone increasing age effect: 1, 1.5, 2, 2.5, and 3.
4) Monotone decreasing age effect: 3, 2.5, 2, 1.5, and 1.

The true incidence rate ratio was set to be e^{β}= 1, 2 or 3. Totally there were 96 scenarios: 2 levels of number of exposed cases, 2 levels of variance of age at exposure, 2 levels of risk intervals, 4 patterns of age effect, and 3 levels of incidence rate ratios. Within each scenario, we considered exposed cases only $(r = 0)$ and 3 levels of augmentation by unexposed cases $(r = 0.1, 0.2, 0.5)$. The total sample size is $n(1 + r)$ cases.

The marginal total number of events per subject was generated using a zero truncated Poisson distribution with rate $\sum_{j=1}^{5}\sum_{k=0}^{2}\lambda_{ijk}e_{ijk} = \sum_{j=1}^{5}\sum_{k=0}^{2}\exp(\varphi_i + \alpha_j + \beta k(2-k))e_{ijk}$, conditionally on the exposure history. Here, e_{ijk} is the duration of time subject i spends in age interval j and in risk period k. The baseline rate is set at $\exp(\varphi_i) = 1 \times 10^{-5}$, thus the event is rare, and with no more than 1% of the cases having more than one event during the observation period. Then we used a multinomial distribution to randomly allocate each event with the probability $\frac{\lambda_{ijk}e_{ijk}}{\sum_{j=1}^{5}\sum_{k=0}^{1}\lambda_{ijk}e_{ijk}}$ to the age interval and risk period for each subject (Musonda et al. 2008). For each given setting, we conducted 10,000 replicates of the simulation. In a small number of replicates there was zero event within the risk window (as indicated in table footnotes). This mainly occurred in the scenarios with n = 100, $e^{\beta} = 1$ and 28-day risk period. They were excluded and the simulation continued until there were 10,000 replicates with non-zero events in the risk period.

We present the simulation results on the properties of the maximum conditional likelihood estimator of β, having adjusted for age, in terms of Bias, SE, Root Mean Square Error (RMSE), Power, and CP, where bias is the mean of the 10,000 estimates minus the true β, SE is the mean of 10,000 estimated standard errors, RMSE $=\sqrt{\sum_{i=1}^{10,000}(\hat{\beta}_i - \beta)^2 / 10{,}000}$, Power refers to empirical power (if $\beta \neq 0$) or empirical type 1 error rate (if $\beta = 0$) for rejecting the null hypothesis of $\beta = 0$ at the two-sided significance level of 0.05, and CP is the coverage probability of the 95% confidence interval.

3.2. Simulation results

Table 1 shows the results in scenarios with $\exp(\beta) = 2.0$ ($\beta = 0.693$), standard deviation of age at exposure = 50 days and 200-day risk period. Across both levels of sample size and four patterns of age effect, including the pattern with no age effect $(\alpha_j = 0 \; \forall \; j)$, SE and RMSE monotonically decreased and power monotonically increase with inclusion of r = 10%, 20% or 50% of unexposed cases. As compared to analysis of exposed cases only, the SE declined by about 4-16% with addition of 10-50% of unexposed cases. When n = 100 and n = 250, there was 3-15% and 1-3% increase in statistical power (absolute value), respectively, as compared to analysis of exposed cases only. There was a small reduction in bias in the analysis with inclusion of unexposed cases when n = 100, but practically no difference when n = 250. There was practically no difference in CP between the analyses with and without inclusion of unexposed cases.

[Table 1 about here]

Table 2 shows the results in the same setting except that the risk period was 28 days. Similar pattern of results was observed, but the improvement in power in relation to increasing number of unexposed cases was mild in this set of scenarios, by no more than 4%.

[Table 2 about here]

Further simulation results are available in Tables S1 to S10 in Online Supplementary Material 3. The findings are qualitatively similar to the above. In addition, when $\exp(\beta) = 1.0$, the empirical type 1 error rate was almost identical between the analyses with and without inclusion of unexposed cases.

4. Application

We re-analysed the measles, mumps, rubella (MMR) vaccine and meningitis data that Whitaker et al. (2006) used to illustrate the self-controlled case series method (https://sccs-studies.info/stata.html; accessed 7 August 2020). The study involved 10 cases of meningitis. Based on biological knowledge and previous studies, they defined 15-35 days inclusive post-MMR as the elevated risk period due to the vaccine, as compared to the time before and after the 15-35 days window as the reference time period. Five of the 10 events occurred within

the post-MMR risk period. Furthermore, the observations were partitioned into two age intervals, 366-547 days (reference) and 548 to 730 days (older).

The previous analysis by Whitaker et al. (2006) included all 10 cases, one of whom did not receive MMR during the study period. Here we re-analysed the data and compare the results using all the 10 cases versus the 9 exposed cases only.

Without adjustment for age, the standard error (0.671) of the vaccine effect estimate was identical in the analysis of all 10 cases and analysis of only the 9 exposed cases. In the analysis with adjustment for age using all cases, the standard error for the vaccine effect was 0.708 (Table 3). The SE based only on the exposed cases was larger, at 0.730.

[Table 3 about here]

5. Discussion

The self-controlled case series is an important method in biomedical research. It is commonly used in the studies of the safety of vaccines and drugs. There has been limited guidance on the relevancy of inclusion of unexposed cases. The practice in medical studies is variable, and there is a lack of transparency in the rationale. In the recent literature, we have seen exclusions based on a misunderstanding that unexposed subjects cannot be included (Aspinall et al. 2020; Duncan et al. 2020; Forbes et al. 2020). Note that the inclusion does not necessarily improve precision. The precision gain is realized only if the model adjusts for time-varying covariates. Medical exposure and outcomes are often related to time-varying covariates such as age or season. Therefore, the analysis usually requires covariate adjustment. A previous methodological study demonstrated that inclusion of unexposed cases in covariate-adjusted analysis tended to offer a benefit in terms of bias reduction, but the benefit was obvious mainly when the incidence rate ratio was very large and the standard deviation of the timing of exposure was very small (Musonda et al. 2008). The previous study did not evaluate whether inclusion of unexposed cases might improve precision and power.

In the analytic proof and simulation, we have demonstrated the gain in precision by inclusion of unexposed subjects occur when time-varying covariates are included in the model, regardless of whether the covariates have any effects on the outcome event or not. In the simulation, we have demonstrated the advantage of inclusion of unexposed cases in self-controlled case series with covariate-adjustment in terms of precision, root mean squared error and power in a variety of scenarios, as well as the absence of disadvantage in terms of coverage probability of 95% confidence intervals or type 1 error rate. In simulation settings

with a longer risk period, the SE declined substantially with addition of unexposed subjects. But when the risk period was 28 days, the reduction in SE was much smaller. This reflects that the longer the risk period is, the more likely the older age intervals are dominated by exposed person-time. With such strong correlation between the indicator variables for age and exposure, the inclusion of unexposed subjects has a strong impact, or vice versa. As such, we expect that the inclusion of unexposed subjects is more beneficial in terms of precision and power in situations of drug prescriptions that may continue for a substantial duration (see, e.g. Duncan et al., 2020; Man et al. 2017) than one-off prescription or vaccination.

From a practical point of view, the case-only methodology is an attractive research strategy because it can save the resources for the ascertainment of the exposure status among non-cases. The self-controlled case series methodology begins with identifying the cases before identifying their exposure status. Having identified the cases, the exposure status (and its timing) needs to be ascertained anyway; there is no extra cost to keep unexposed cases in the analysis even if the benefit is small. From the viewpoint of precision and power, it is useful to include the unexposed cases in covariate-adjusted analysis.

Disclosure statement

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work was supported by the National Medical Research Council, Singapore (MOH-000526-00).

ORCID

Yin Bun Cheung http://orcid.org/0000-0003-0517-7625

Data availability statement

We used a publicly available dataset from https://sccs-studies.info for the illustration (accessed 7 August 2020).

References

Almenoff, J. S., E. N. Pattishall, T. G. Gibbs, W. DuMouchel, S. J. W. Evans, and N. Yuen. 2007. Novel statistical tools for monitoring the safety of marketed drugs. *Clinical Pharmacology and Therapeutics* 82 (2): 157-166. doi: 10.1038/sj.clpt.6100258.

Aspinall, S. L., N. P. Sylvain, X. Zhao, R. Zhang, D. Dong, K. Echevarria, et al. 2020. Serious cardiovascular adverse events with fluoroquinolones versus other antibiotics: A self-controlled case series analysis. *Pharmacology Research & Perspectives* 8(6):e00664. doi: 10.1002/prp2.664.

DuMouchel, W. 1999. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The American Statistician* 53 (3): 177-190. doi: 10.1080/00031305.1999.10474456.

Duncan, A. D. S., S. Hapca, N. De Souza, D. Morales, and S. Bell. 2020. Quinine exposure and the risk of acute kidney injury: a population-based observational study of older people. *Age and Ageing* 49 (6): 1042-1047. doi: 10.1093/ageing/afaa079.

Forbes, H., I. Douglas, A. Finn, J. Breuer, K. Bhaskaran, L. Smeeth, et al. 2020. Risk of herpes zoster after exposure to varicella to explore the exogenous boosting hypothesis: self controlled case series study using UK electronic healthcare data. BMJ. 2020 Jan 22;368:l6987. doi: 10.1136/bmj.l6987.

Gardiner, J., Z. Luo, and L. A. Roman. 2009. Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine* 28 (2): 221-239. doi: 10.1002/sim.3478.

Harpaz, R., W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan, and C. Friedman. 2012. Novel data mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology and Therapeutics* 91 (6): 1010-1021. doi: 10.1038/clpt.2012.50.

Man, K. K. C., D. Coghill, E. W. Chan, W. C. Y. Lau, C. Hollis, E. Liddle, et al. 2017. Association of risk of suicide attempts with methylphenidate treatment. *JAMA Psychiatry* 74 (10): 1048-1055. doi: 10.1001/jamapsychiatry.2017.2183.

Moulton, L. H., K. L. O'Brien, R. Reid, R. Weatherholtz, M. Santosham, and G. R. Siber. 2006. Evaluation of the indirect effects of a pneumococcal vaccine in a community-randomized study, *Journal of Biopharmaceutical Statistics* 16 (4): 453-462. doi: 10.1080/10543400600719343.

Musonda, P., M. N. Hocine, H. J. Whitaker, and C. P. Farrington. 2008. Self-controlled case series analyses: Small-sample performance. *Computational Statistics and Data Analysis* 52 (4): 1942-1957. doi: 10.1016/j.csda.2007.06.016.

Peach, E. J., F. A. Pearce, J. Gibson, A. J. Cooper, L. C. Chen, and R. D. Knaggs. 2021. Opiods and the risk of fracture: a self-controlled case series study in the clinical practice research datalink. *American Journal of Epidemiology*, Online ahead of print. Doi: 10.1093/aje/kwab042.

Schuemie, M. J., P. B. Ryan, K. K. C. Man, I. C. K. Wong, M. A. Suchard, and G. Hripcsak. 2019. A plea to stop using the case-control design in retrospective database studies. *Statistics in Medicine* 38 (22): 4199-4208. doi: 10.1002/sim.8215.

Weldeselassie, Y. G., H. J. Whitaker HJ, and C. P. Farrington. 2011. Use of the self-controlled case-series method in vaccine safety studies: Review and recommendations for best practice. *Epidemiology and Infection* 139: 1805-1817. doi: 10.1017/S0950268811001531.

Whitaker, H.J., C. P. Farrington, B. Spiessens, and P. Musonda. 2006. Tutorial in biostatistics: the self-controlled case series method. *Statistics in Medicine* 25 (10): 1768-1797. doi: 10.1002/sim.2302.

Xu, S., Zeng, C., Newcomer, S., Nelson, J., and J. Glanz. 2012. Use of fixed effects models to analyze self-controlled case series data in vaccine safety studies. *Journal of Biometrics and Biostatistics* Suppl 7:006. doi: 10.4172/2155-6180.s7-006.

Xu, S., L. Zhang, J. C. Nelson, C. Zeng, J. Mullooly, D. McClure, and J. Glanz. 2011. Identifying optimal risk windows for self-controlled case series studies of vaccine safety. *Statistics in Medicine* 30 (7): 742-752. doi: 10.1002/sim.4125.

Table 1. Simulation results with $\exp(\beta) = 2.0$ ($\beta = 0.693$), standard deviation of age at exposure = 50 days, 200-day risk period, four patterns of age effect, number of exposed cases (n) = 100 or 250, and number of unexposed cases (r) = 0% to 50% of n.

Age effect	r	n = 100					n = 250				
		Bias	SE	RMSE	Power	CP(%)	Bias	SE	RMSE	Power	CP(%)
No	0%	0.012	0.299	0.303	0.660	95.0	0.007	0.188	0.188	0.969	95.0
	10%	0.009	0.284	0.287	0.704	94.9	0.006	0.178	0.179	0.979	95.0
	20%	0.008	0.272	0.275	0.738	94.9	0.005	0.171	0.172	0.985	95.2
	50%	0.006	0.252	0.254	0.796	95.2	0.004	0.159	0.159	0.994	95.3
Symmetric	0%	0.021	0.297	0.302	0.683	95.1	0.011	0.186	0.187	0.975	95.4
	10%	0.020	0.285	0.289	0.717	94.9	0.011	0.179	0.179	0.982	95.3
	20%	0.019	0.277	0.280	0.742	95.0	0.010	0.174	0.174	0.986	95.3
	50%	0.017	0.261	0.266	0.785	95.3	0.009	0.164	0.164	0.992	94.9
Increasing	0%	0.012	0.304	0.309	0.644	94.8	0.006	0.190	0.193	0.961	95.0
	10%	0.011	0.288	0.291	0.689	94.8	0.005	0.180	0.183	0.975	94.9
	20%	0.010	0.276	0.280	0.724	94.9	0.004	0.173	0.175	0.981	95.0
	50%	0.008	0.255	0.260	0.787	94.7	0.002	0.160	0.162	0.992	95.0
Decreasing	0%	0.017	0.305	0.305	0.645	95.6	0.005	0.191	0.191	0.961	95.0
	10%	0.014	0.289	0.291	0.688	95.3	0.005	0.181	0.181	0.975	95.0
	20%	0.013	0.277	0.278	0.727	95.4	0.005	0.174	0.173	0.982	95.2
	50%	0.011	0.255	0.255	0.791	95.4	0.004	0.161	0.159	0.993	95.1

Table 2. Simulation results with $\exp(\beta) = 2.0$ ($\beta = 0.693$), standard deviation of age at exposure = 50 days, 28-day risk period, four patterns of age effect, number of exposed cases (n) = 100 or 250, and number of unexposed cases (r) = 0% to 50% of n.

Age effect	r	$n = 100$ *					$n = 250$				
		Bias	SE	RMSE	Power	CP(%)	Bias	SE	RMSE	Power	CP(%)
No	0%	-0.036	0.390	0.405	0.443	95.8	-0.013	0.228	0.230	0.827	95.3
	10%	-0.038	0.387	0.402	0.448	95.9	-0.004	0.226	0.229	0.838	95.1
	20%	-0.038	0.384	0.400	0.453	95.8	0.003	0.225	0.228	0.846	94.9
	50%	-0.040	0.378	0.394	0.462	95.8	0.018	0.221	0.225	0.865	94.6
Symmetric	0%	-0.031	0.386	0.396	0.451	95.8	-0.009	0.206	0.208	0.887	95.0
	10%	-0.033	0.382	0.393	0.457	95.8	-0.009	0.205	0.207	0.891	95.1
	20%	-0.034	0.380	0.390	0.461	95.9	-0.009	0.204	0.206	0.893	95.1
	50%	-0.037	0.374	0.385	0.469	95.9	-0.010	0.201	0.203	0.898	95.1
Increasing	0%	-0.045	0.458	0.478	0.365	96.2	-0.017	0.265	0.275	0.713	95.1
	10%	-0.047	0.453	0.472	0.369	96.2	-0.018	0.262	0.272	0.719	95.0
	20%	-0.049	0.449	0.468	0.371	96.3	-0.018	0.260	0.270	0.724	95.0
	50%	-0.052	0.441	0.461	0.384	96.2	-0.020	0.256	0.265	0.734	95.1
Decreasing	0%	-0.025	0.349	0.361	0.513	95.4	-0.012	0.204	0.205	0.892	95.0
	10%	-0.025	0.347	0.358	0.521	95.4	-0.012	0.202	0.204	0.896	95.1
	20%	-0.025	0.345	0.355	0.525	95.4	-0.012	0.201	0.203	0.899	95.1
	50%	-0.026	0.340	0.351	0.533	95.3	-0.013	0.198	0.200	0.901	95.2

* 3 replicates in the scenario with $n = 100$ and increasing age effect that had zero event within the risk window were re-generated.

Table 3. Re-analysis of MMR vaccine and meningitis data.

Independent variables	All cases			Exposed case only		
	β	SE	(95% CI)	β	SE	(95% CI)
MMR	2.488	0.708	(1.099 to 3.877)	2.548	0.730	(1.116 to 3.979)
Older age	-1.491	1.118	(-3.682 to 0.701)	-1.227	1.155	(-3.490 to 1.036)

Legend to Figure 1.

Two distributions of age at exposure in the simulation study.