Self-controlled case series analyses: small sample performance

Patrick Musonda[1], Mounia N. Hocine[1,2], Heather J. Whitaker[1]
and C. Paddy Farrington[1]*
[1]The Open University, Milton Keynes, MK7 6AA, UK
[2]INSERM U780 ; Univ Paris-Sud, Villejuif F-94807, France

Abstract

We derive second-order expressions for the asymptotic bias and variance of the log relative incidence estimator for the self-controlled case series method in a simplified scenario, and study in qualitative terms how bias and variance depend on factors such as the relative incidence and ratio of risk to observation period. Small-sample performance of the estimator in realistic scenarios is investigated using simulations. We find that in scenarios likely to arise in practice, asymptotic methods are valid for numbers of cases in excess of 20 – 50 depending on the ratio of the risk period to the observation period and on the relative incidence. The application of Monte Carlo methods to self-controlled case series analyses is also discussed.

Keywords: Asymptotic bias; Asymptotic variance; Bootstrap; Randomization test; Self-controlled case series method; Simulation; Small-sample performance

* Corresponding author.
Department of Statistics
Faculty of Mathematics and Computing
The Open University
Walton Hall
Milton Keynes MK7 6AA
Tel. : +44 (0) 1908 654 840
Fax.: +44 (0) 1908 652 140
Email: c.p.farrington@open.ac.uk

This research was supported by EPSRC (CASE0307), GlaxoSmithKline Biologicals, and Wellcome Trust project grant 070346

1 Introduction

The self-controlled case series method, or case series method for short, is a conditional cohort method for estimating the strength of association between the incidence of specified events and a time-varying exposure using data only on cases. The method was originally developed to investigate associations between vaccination and acute potential adverse events [3]. Other applications, along with a detailed account of the theory and its implementation in standard statistical packages are described in Whitaker et al [11]. A semi-parametric version of the method has also been developed [7].

While the maximum likelihood estimator of the relative incidence is guaranteed good asymptotic properties for both parametric and semi-parametric models, in practice samples are often small, especially for rare conditions. Limited small-sample simulations for the semi-parametric model suggest that it performs well in samples of moderate size [7]. However, no systematic evaluation of the statistical properties of the method has been undertaken. Some comparative evaluations have been done, comparing the case series method with case-control, cohort and other case only methods [1, 4, 6].

Our aim in this paper is to investigate in more detail the factors that influence the magnitude of the bias and variance of the relative incidence estimator, or more precisely the estimator of the log relative incidence. For simplicity, we confine our investigations to the parametric self-controlled case series model and to the risks associated with exogenous point exposures [2].

The paper is organised as follows. In section 2 we introduce the case series model. Explicit expressions for the asymptotic bias, variance and mean square error in a simplified but relevant scenario are derived and studied in section 3. Section 4 describes a simulation study to evaluate bias and variance in small samples under more realistic scenarios. The results from this simulation study are presented in section 5. In section 6, we discuss the application of Monte Carlo methods to self-controlled case series analyses, including bootstrap estimation and randomization tests. Finally in section 7 we discuss our findings and make some recommendations.

2 The self-controlled case series model

The self-controlled case series model is derived from an underlying Poisson cohort model. Thus, we consider a cohort of individuals, individual *i* being observed in the interval $(a_i, b_i]$. This interval is the observation period for individual *i*; we shall use age as the underlying time line, but other choices are possible, notably calendar time.

The observation period for individual *i* is partitioned into intervals indexed by $j = 0,1,...,J$ (for age groups) and $k = 0,1,...,K$ (for risk periods). The age groups are pre-determined, as are the durations of the post-exposure risk periods. Risk periods $k = 1,...,K$ correspond to increased or reduced risks relative to the baseline control period, which is coded $k = 0$. The age groups are typically of the form $(0, A_0], (A_0, A_1], ..., (A_{J-2}, A_{J-1}], (A_{J-1}, \infty)$. Post-exposure risk periods are typically of the form $(E_i + B_{k-1}, E_i + B_k]$ where E_i is the age at exposure of individual *i* and $B_0 < ... < B_K$, the remainder of the observation time being allocated to the control period.

Let e_{ijk} denote the duration of time that individual *i* spends in age group *j* and in risk period *k* during the course of his or her observation period. Conditioning on the exposure history over the entire observation period $(a_i, b_i]$, we assume that events of interest for individual *i* arise as a non-homogeneous Poisson process with rate λ_{ijk}. If n_{ijk} denotes the number of events arising for individual *i* in age group *j* and risk period *k*, then

$$n_{ijk} \sim \text{Poisson}(\lambda_{ijk} e_{ijk}).$$

Conditioning on the total number of events $n_i = \Sigma_{j,k} n_{ijk}$ arising in $(a_i, b_i]$, which is possible by virtue of the assumption that the exposure is an exogenous variable [2, 7], the log-likelihood contribution of individual *i* is multinomial with kernel

$$l_i = \sum_{j,k} n_{ijk} \log\left(\frac{\lambda_{ijk} e_{ijk}}{\sum_{r,s} \lambda_{irs} e_{irs}} \right). \quad (1)$$

We assume a log-linear model for the Poisson rate of the form

$$\log(\lambda_{ijk}) = \varphi_i + \alpha_j + \beta_k \quad (2)$$

where φ_i is an individual effect, α_j is the age effect associated with age group *j*, and β_k is the exposure effect associated with risk group *k*, with $\alpha_0 = \beta_0 = 0$. The parameters α_j and β_k are thus log relative incidences.

Substituting (2) in (1), and summing over individuals, we obtain a product multinomial log-likelihood kernel:

$$l(\alpha, \beta) = \sum_i \sum_{j,k} n_{ijk} \log\left(\frac{\exp(\alpha_j + \beta_k) e_{ijk}}{\sum_{r,s} \exp(\alpha_r + \beta_s) e_{irs}} \right). \quad (3)$$

This is the self-controlled case series log-likelihood. The model is *self-controlled* because the individual effects φ_i cancel out. Thus multiplicative confounders that do not vary over the individual's observation period – which might include, for example, genetic effects, socio-economic status, location, underlying state of health, individual frailties – are necessarily adjusted in the analysis. It is a *case series* model because only individuals who have experienced one or more events, that is individuals for whom $n_i \geq 1$, contribute non-trivially to the log-likelihood (3). Thus, only cases need to be sampled. These features make the self-controlled case series method an attractive alternative to other methods in some settings.

The efficiency of the case series model relative to the underlying cohort model, and the assumptions required, in particular the important assumption that the exposure variable is exogenous, are discussed in [7].

3 Asymptotic bias, variance and mean square error

In this section we study in greater detail the asymptotic properties of the estimators of the log relative incidence.

3.1 A simplified scenario

Our aim is to obtain qualitative insight into the factors which affect bias and variance. So as to obtain simple explicit expressions, we make the following assumptions.

- All cases have the same observation period $(a_i, b_i] = (a, b]$.
- There are no underlying age effects, that is, $\alpha_j = 0$ for all j.
- There is at most one post-exposure risk period, that is, $K = 1$.
- All cases experience an exposure risk period of common duration e_1 and a control period of common duration e_0, with $e_0 + e_1 = b - a$.

The age parameters may thus be dropped from the model. We denote $\beta = \beta_1$. Under these assumptions, the log-likelihood (3) for n events reduces to the expression

$$l(\beta) \quad = \quad x\beta - n\log\left(e_1 e^{\beta} + e_0\right) \tag{4}$$

where x is the number of events occurring in the exposure risk period. The maximum likelihood estimator of β is

$$\hat{\beta} \quad = \quad \log\left(\frac{x}{n-x}\right) - \log\left(\frac{r}{1-r}\right)$$

where $r = e_1/(e_0 + e_1)$ is the ratio of the length of the risk period to the observation period.

Expanding $\hat{\beta}$ as a function of x by Taylor series to fourth order, we obtain the following expressions for the asymptotic bias and variance, to second order.

$$\begin{aligned}\text{bias}(\hat{\beta}) &= E(\hat{\beta})-\beta \\ &= \frac{1}{2n}\left(re^{\beta}-(1-r)\right)\left(\frac{1}{re^{\beta}}+\frac{1}{1-r}\right) \\ &\times\left[1+\frac{5(re^{\beta})^2+4re^{\beta}(1-r)+5(1-r)^2}{6nre^{\beta}(1-r)}\right]+O(n^{-3})\end{aligned} \qquad (5)$$

$$\text{var}(\hat{\beta}) = \frac{1}{n}\frac{\left(re^{\beta}+(1-r)\right)^2}{re^{\beta}(1-r)}\times\left[1+\frac{3(re^{\beta})^2-2re^{\beta}(1-r)+3(1-r)^2}{2nre^{\beta}(1-r)}\right]+O(n^{-3}). \qquad (6)$$

Combining expressions (5) and (6), we obtain the asymptotic mean squared error:

$$\text{AMSE}(\hat{\beta}) = \frac{1}{n}\frac{\left(re^{\beta}+(1-r)\right)^2}{re^{\beta}(1-r)}\times\left[1+\frac{7(re^{\beta})^2-6re^{\beta}(1-r)+7(1-r)^2}{4nre^{\beta}(1-r)}\right]+O(n^{-3}). \qquad (7)$$

3.2 Asymptotic properties

Consider first the asymptotic bias. The expression in square brackets in (5) is always greater than 1, so that

$$\text{sgn}\left(\text{bias}(\hat{\beta})\right) = \text{sgn}\left(re^{\beta}-(1-r)\right)$$

and the second-order bias is always greater in magnitude than the first-order bias.

The asymptotic bias is zero when $re^{\beta}=1-r$, which occurs when the expected number of cases in the risk period equals the expected number of events in the control period. The asymptotic bias is negative (respectively, positive) when the expected number of events in the risk period is less (respectively, greater) than that in the control period. In practice, the risk period is determined by the scientific question of interest, and the observation period is determined both by the age range at which exposures occur and by the practicalities of data collection. For a given value of *r*, the asymptotic bias is minimized when

$$e^{\beta} = \frac{1-r}{r}.$$

For a fixed value of β, the asymptotic bias increases in magnitude as r tends to 0 or 1. Similarly, for a fixed value of *r*, the asymptotic bias increases in magnitude as β tends to $\pm\infty$. Figure 1 shows the value of the second-order asymptotic bias for $n = 50$, for different values of *r* and e^{β}. The asymptotic bias is negligible unless the ratio of the risk period to observation time is very close to 0 or 1, but increases sharply in these regions for smaller sample sizes.

Turning now to the asymptotic variance, its value to second-order is always greater than to first order. Regarding expression (6) as a function of *r*, its minimum is attained when $re^{\beta} = 1-r$. Thus, the asymptotic variance is smallest when the expected number of events in the risk period equals the expected number in the control period. Figure 2 shows $\mathrm{var}(\hat{\beta})$ for $n = 50$, for different values of *r* and e^{β}. As for the bias, the asymptotic variance increases as r tends to 0 or 1 and as β tends to $\pm\infty$.

The second-order asymptotic mean squared error (7) is close to the second-order variance. It is minimized when $re^{\beta} = 1-r$, but is typically very flat for values *r* in the range (0.1, 0.9) and $|\beta| < \log(10)$.

4. Simulation study

In this section we study the properties of the maximum likelihood estimator $\hat{\beta}$ by simulation, in more realistic scenarios than that described in section 3. In particular, we no longer assume that there is no effect of age, or that all individuals have the same exposure risk period. Our aim is to investigate the limits of validity of asymptotic theory in finite samples.

Because $\hat{\beta}$ is the logarithm of a ratio estimator, it takes values $\pm\infty$ with positive probability in finite samples. Thus, rather than the bias *per se*, which is undefined, we investigate the median $m_n(\hat{\beta})$ of the estimator in samples of size *n*. This provides an

appropriate measure of central tendency of the estimator in finite samples. Note that $\lim_{n\to\infty} m_n(\hat{\beta}) = E(\hat{\beta})$ since $\hat{\beta}$ is asymptotically normally distributed. From now on, the term 'bias' refers to $m_n(\hat{\beta}) - \beta$. We also investigate the coverage probability of the Wald 95% confidence interval calculated from $\hat{\beta} \pm 1.96 \times \text{se}(\hat{\beta})$ where $\text{se}(\hat{\beta})$ is the asymptotic standard error (for unbounded estimates the confidence interval is in effect $(-\infty, +\infty)$).

The simulations were set up to mimic those scenarios that typically occur in studies of paediatric vaccines. The simulation experiments are described in the following sections.

4.1. Structure of the simulation study

Each simulation required the following parameters to be specified.

- Observation period, always taken to be 500 days for all individuals.
- Length of the risk period following exposure (days): 1, 5, 10, 25, 50, 100, 200, indefinite (described in section 4.4).
- True relative incidence RI = e^{β} = 0.5, 1, 1.5, 2, 5, 10.
- Distribution for age at exposure E_i (section 4.3).
- Age groups and age-specific relative incidences (section 4.2, Figure 4).
- Baseline rate, always taken to be $\varphi_i = 2 \times 10^{-7}$ per day, or one per hundred thousand over 500-day observation period. Thus the event is assumed to be rare, and with high probability a case has only a single event.
- Sample size n = 10, 20, 50, 100, 200, 500, 1000 cases.

Figure 3 shows the structure of the simulation study in graphical form. For a given set of parameters (listed above) and random seed, a set of n exposure times were generated, together with n marginal total number of events per individual. These marginal totals were generated using a truncated Poisson distribution (excluding zero), conditionally on the exposure history.

The exposures and marginal totals were resampled between runs. however, in each run of 10,000 simulations, the exposures and marginal totals were kept fixed. This is to mimic the fact that the case series method is conditional on exposures and marginal totals.

Within a run, the events for each individual were randomly reallocated 10,000 times to the age and risk categories within each individual's person time. This was done based on the case series model, using a multinomial distribution.

The run size of 10,000 ensures that the coverage probability for a 95% confidence interval is estimated with Monte Carlo standard error of about 0.0022, and hence is accurate to within about 0.005 (or 0.5% when expressed as a percentage).

4.2 Age effects

In most self-controlled case series analyses, it is necessary to control for age. We varied the effect of age on the event incidence according to four practically realistic scenarios. These four types of age effect are defined as follows; in each case the age groups are given, along with the associated age-specific relative incidences e^{α_j} (in brackets)

- Weak symmetric age effect: 1-100 (1), 101-200 (1.2), 201-300 (1.5), 301-400 (1.2), and 401-500 (1).
- Strong symmetric age effect: 1-50 (1), 51-100 (2), 101-150 (3), 151-200 (4), 201-250 (5), 251-300 (5), 301-350 (4), 351-400 (3), 401-450 (2), and 451-500 (1).
- Weak monotone increasing age effect: 1-100 (1), 101-200 (1.1), 201-300 (1.2), 301-400 (1.3), 401-500 (1.4)
- Strong monotone increasing age effect: 1-50 (1), 51-100 (1.5), 101-150 (2), 151-200 (2.5), 201-250 (3), 251-300 (3.5), 301-350 (4), 351-400 (4.5), 401-450 (5), and 451-500 (5.5).

Figure 4 shows bar charts representing each of the above four choices of age groups and age-specific relative incidences.

4.3 Exposure distribution

The precision of the relative incidence estimator depends on the extent of between-individual variation in age at exposure. We used the following four beta distributions on $[0,500]$ to generate age at exposure.

- Mean age 250 days and standard deviation 100 days.
- Mean age 250 days and standard deviation 50 days.
- Mean age 125 days and standard deviation 100 days.
- Mean age 125 days and standard deviation 50 days.

These distributions are shown in Figure 5.

For some simulations, much more highly peaked distributions of age at exposure were also considered, with mean age of 125 days and standard deviation of 10, 20, 30, and 40 days.

4.4 Risk periods

Before carrying out a self-controlled case series analysis, a major issue to consider is how to define the risk periods. Generally speaking the risk periods are elicited from experts. Different studies need different risk periods. These range from very short (a few days) to very long (several months), and occasionally may be indefinite.

We simulated data with risk periods of 1, 5, 10, 25, 50, 100 and 200 days. We also investigated indefinite risk periods. Owing to potentially strong confounding between age and exposure effects with indefinite risk periods, we considered these separately and varied the proportion of cases exposed (in other simulations we assumed all cases were exposed).

5 Results of the simulation study

The presentation of results is organised in five subsections. In subsection 5.1 we present results for our ‘standard scenario’. In subsection 5.2 we vary the risk period.

In subsection 5.3 we vary the age effect. In subsection 5.4 we vary the age at exposure. Finally, in subsection 5.5 we consider indefinite risk periods.

5.1 The standard scenario

For our standard scenario the risk period was 25 days, all cases experienced the exposure, the age effect was weak symmetric (see Figure 4) and the distribution of age at exposure has mean age 250 days and standard deviation 100 days (see Figure 5).

Table 1 shows the results for the standard scenario. For very small samples ($n \leq 20$) and low relative incidences (RI ≤ 1), there is considerable bias: effectively, in most samples there were zero events within a risk period, yielding unbounded estimates of β. For relative incidences greater than 1, the bias is moderate even for sample sizes as small as 10. For sample sizes in excess of 20, the bias is small for most values of the relative incidence (the exception being RI = 0.5).

The bias tends to be negative for low relative incidences, and positive for large relative incidences. This reflects the asymptotic results obtained in section 3, namely that, in the absence of age effects, the asymptotic bias is negative when $e^{\beta} < (1-r)/r$ and positive when $e^{\beta} > (1-r)/r$. Here, $r = 25/500 = 0.05$. Thus, asymptotically, and provided that age effects are not too strong, one might expect zero bias at $e^{\beta} \approx 20$. In finite samples, this point appears to be reached for lower relative incidences: for example, with $n = 50$, it is reached at $e^{\beta} \approx 5$ in the standard scenario.

Finally, note from Table 1 that the coverage probabilities of the Wald 95% confidence intervals are close to their nominal values for all combinations of sample size and relative incidence, though tend to be conservative especially for low sample sizes. Similar results (not shown) were obtained for 90% and 99% confidence intervals.

5.2 Risk period of fixed length

The fixed-length risk periods were: 1, 5, 10, 50, 100 and 200 days. Table 2 shows the results (with *n* = 20, 100 and 500) for the short risk periods of 1 and 5 days, and Table 3 shows the results for longer risk periods of 50 and 100 days.

As expected from the asymptotic calculations, the bias increases in absolute value as r, the ratio of the risk period to the observation period (500 days), tends towards zero. With a 1-day risk period, the bias is considerable in small or moderate samples, unless the relative incidence is high: it is possible to estimate β with little bias for a 1-day risk period with sample sizes of 100 cases provided that the relative incidence is in excess of 5. A slight increase in the length of the risk period has a big effect: there is little bias with sample sizes as small as 20 for relative incidences in excess of 5 when the risk period is 5 days.

For longer risk periods (50 and 100 days), Table 3 shows that there is little bias even for sample sizes as small as 20, when the relative risk is greater than 1. The results for the 10 day risk period were broadly similar to those for 25 days (the standard scenario), while the results for the 200 day risk period were similar to those for the 100 day risk period (not shown).

5.3 Age at event

In this section, we summarize the results we obtained by varying the underlying age effect. We investigated sample sizes 20, 100 and 500 and risk periods of 10, 25 and 50 days, with relative incidences of 1, 2 and 5. The distribution of age at exposure was as in the standard scenario, namely mean 250 days and standard deviation 100 days. Table 4 gives the results for sample size 100 with risk period 25 days. Varying the age effect has little influence on the magnitude of the bias or on the coverage probabilities, for any of the risk intervals considered here. Similar results were obtained for other sample sizes (not shown).

5.4 Age at exposure

In the standard scenario, the distribution of age at exposure was a symmetrical beta distribution with 250 days and standard deviation 100 days. Here we evaluate the performance of the model when we vary the mean and standard deviation. In view of possible confounding between age and exposure effects, interest focuses particularly on the bias when a positively skewed distribution of age at exposure is combined with a strong monotone increasing age at event effect.

Table 5 presents the results for samples of 100 cases, risk periods 25 and 50 days, relative incidences of 1 and 5, and both the weak symmetric and the strong monotone age effects. There is little evidence that the mean or standard deviation of the age at exposure have any discernible impact on the bias or coverage probabilities. Similar results were obtained for the 10 day risk period, and for RI = 2 (not shown).

5.5 Indefinite risk periods

The self-controlled case series method can be used even when the risk period following an exposure is indefinite [5, 11]. However, exposure and age effects may be confounded. This can be controlled by including unexposed cases, which contribute exclusively to the estimates of the age effects.

For age at event, we used the weak symmetric, and the strong monotone increasing age distributions. We investigated six beta distributions of age at exposure: mean 250 days and standard deviation 100 days, mean 125 days and standard deviation 50 days, and four more peaked distributions with mean 125 days and standard deviations 40, 30, 20 and 10 days. We studied relative risks of 1, 2 and 5.

We used samples of 100 exposed cases, augmented by 0%, 20%, 50% and 100% unexposed cases. For example, the sample augmented by 20% unexposed cases contained 100 exposed cases and 20 unexposed cases. Table 6 shows the results for the strong symmetric age effect and distributions of age at exposure with mean 125 days and standard deviations 10, 30 and 50 days.

When the relative incidence is 1, β is estimated without substantial bias even with no unexposed cases. The greater the relative incidence and the more peaked the distribution of age at exposure, the greater the bias: when the relative incidence is 5, the estimate is swamped by bias. However, inclusion of just 20 unexposed cases is sufficient to greatly reduce the bias. Interestingly, inclusion of more than 20 unexposed cases has little further beneficial effect. The coverage probabilities of the 95% confidence intervals are unaffected.

When the distribution of age at exposure is more evenly spread over the observation period (mean 250 and standard deviation 100), there is little bias even when only exposed cases were included (not shown).

6 Monte Carlo methods

In this section we describe the application of Monte Carlo methods to the self-controlled case series method, with reference to two example data sets relating to measles, mumps and rubella (MMR) vaccine.

6.1 The data

In the first data set the outcome is aseptic meningitis, which is occasionally associated with receipt of MMR vaccines containing the Urabe mumps strain. There are 10 events in 10 children observed from ages 366 to 730 days of age inclusive. The analysis uses two age groups (366 to 547 days, and 548 to 730 days) and a single risk period 15 – 35 days post-MMR. There were 5 events in the risk period. For further details, see [9, 11].

In the second data set, the outcome is idiopathic thrombocytopenic purpura (ITP), an uncommon bleeding disorder occasionally associated with MMR vaccination. The observation period is 366 to 730 days of age. There are 35 children with 44 ITP events. The analysis uses three age groups (366 – 487, 488 – 609, and 610 – 730 days of age) and three risk periods: 0 – 14 days, 15 – 28 and 29 – 42 days post-MMR. There were 2 events in the 0 – 14 day, 8 in the 15 – 28 day, and 3 in the 29 – 42 day risk periods. For further details see [10, 11].

In both data sets, the small number of events in the risk periods calls into question the validity of the asymptotic theory underpinning the calculation of confidence intervals and p values.

6.2 Bootstrap

The most readily applicable bootstrap method for self-controlled case series studies is the non-parametric method based on resampling of cases. This is preferred to resampling of residuals, since it is far from clear what an appropriate residual, or set of residuals, would be in this context. Note that the units to be resampled are the

cases, not the events (an individual who has experienced several events constitutes one case).

As previously noted, the bias of $\hat{\beta}$ is undefined in finite samples. We thus investigate the median $m_B(\hat{\beta})$ of the bootstrap samples; it is desirable that $\hat{\beta}$ should lie close to this value. We also obtain percentile and bias-corrected percentile confidence intervals [8]. All results are based on 4999 bootstrap samples. The results are shown in Table 7. Figure 6 shows the centres of the distributions of the bootstrap replicates for the two data sets; unbounded estimates have been excluded from the figure.

The point estimates are close to the median bootstrap values, suggesting that the bias is mild, but there are substantial discrepancies between asymptotic and bootstrap 95% confidence intervals. With the possible exception of Figure 6(c), the bootstrap distributions display marked evidence of non-normality. The multiple modes correspond to estimates based on distinct numbers of events within the risk period.

6.2 A randomization test

Throughout this paper the emphasis has been on point and interval estimation. In some circumstances, however, it is required to test the null hypothesis of no association between the exposure and event of interest. For this purpose, the likelihood ratio test is readily applicable when the sample size is sufficiently large that asymptotic theory can be relied upon. When this is not the case, however, other methods may be required. We describe a suitable randomization test, implemented by Monte Carlo methods.

Under the null hypothesis of no association, exposure histories and event histories are independent. A randomization test may thus be obtained by randomly pairing event times and exposures. More specifically, consider a sample of *n* cases, case *i* having n_i events at times $t_{i1},...,t_{in_i}$ and exposure history E_i. We then permute the exposure histories from $\{E_1,...,E_n\}$ and allocate the permuted values $E_{\sigma(i)}$ to obtain new data of the form $\{(a_i,b_i];t_{i1},...,t_{in_i};E_{\sigma(i)}\}$. These data are then analysed using the self-controlled case series method to produce a value of the log-likelihood ratio statistic

D_σ. The distribution of the D_σ over all permutations (which thus includes the observed value D_0, say) constitutes the null distribution, from which the *p* value may be calculated from $\#\{D_\sigma : D_\sigma \geq D_0\}$. In practice, it is usually not feasible to obtain all permutations, in which case a random sample is used, augmented by D_0.

This randomization test is standard [8]. The only special point to note is that the test requires that exposure histories are collected in the range $(\min\{a_i\}, \max\{b_i\}]$ to ensure that reallocated histories are relevant to all the observation periods $(a_i, b_i]$.

For the aseptic meningitis data, none of 999 randomly sampled values of D_σ exceeded D_0 = 11.51. Thus, the estimated *p*-value is (0+1) / (999+1) = 0.001. The *p* value based on the asymptotic $\chi^2(1)$ distribution is 0.0007. For the ITP data, 9 values of D_σ out of 999 exceeded D_0 = 13.43. Thus the estimated *p* value is (9+1) / (999+1) = 0.010. The *p* value based on the asymptotic $\chi^2(3)$ distribution is 0.0038. Figure 7 shows the randomization and asymptotic distributions under the null hypothesis. There is a substantial difference between the randomization and asymptotic distributions in each case, though the randomization and asymptotic tests lead to identical conclusions in these examples.

7 Discussion

The aim of this paper was to study the bias and variance of the maximum likelihood estimator of the relative incidence in self-controlled case series studies. We were particularly interested in two aspects: determining which factors most substantially affect the bias and the variance, and the performance of the estimators in small to medium samples.

The asymptotic expressions we obtained in a simple scenario suggest that the bias in β is small unless (a) the risk period is short in relation to the observation period and the relative risk is low, and (b) the risk period is long in relation to the observation period and the relative risk is high. Specifically, the direction and magnitude of the bias is

governed by the quantity $re^{\beta} - (1-r)$, where r is the ratio of the risk period to the observation period and e^{β} is the relative incidence.

This qualitative conclusion was confirmed in simulations. Thus, we found that the bias is small when there are 50 or more cases, the relative incidence is not less than 1, and r is at least 0.05. For sample sizes of 20, the bias is large when the relative incidence is less than 2 or r is less than 0.05. Variation in age at exposure and age at event have only marginal effect on the bias for finite risk periods. For indefinite risk periods, confounding between exposure and age effects may be controlled by inclusion of about 20% of unexposed cases. The asymptotic Wald confidence intervals are generally slightly conservative, but perform well whatever the sample size. When the estimate of the log relative incidence is unbounded, a confidence interval obtained by profile likelihood methods [8] is preferable.

When there is doubt about the validity of asymptotics, simulation inference methods may be used. These include non-parametric bootstrap methods based on resampling complete cases (that is, individuals rather than events), and randomization tests. Note, however, that the use of randomization tests requires that exposures over the entire period $(\min\{a_i\}, \max\{b_i\}]$ are obtained.

The scenarios we chose to investigate relate to those that are likely to arise in studies of vaccine safety with a single post-vaccination risk period. For simplicity, we did not consider multiple exposures, long but fixed risk periods (with r close to 1), semi-parametric estimation of the age effect, between-individual variation in observation periods, and continuous exposures. Most of these more general scenarios can nevertheless be related to those used here. Thus, distinct risk periods can be considered separately, using a value of r calculated as the ratio of the risk period of interest to the sum of the risk period and control period; long fixed risk periods will yield results similar to those obtained with indefinite risk periods; between-individual variation in observation periods may be accommodated by taking r to be the ratio of the risk period to the average observation period; and age effects were shown to have only moderate impact, though of course semi-parametric estimation will necessarily yield less precise estimates.

Our findings are thus broadly relevant to case series studies of point exposures. For continuous time-varying exposures, further investigations in small samples are required. In such settings, the notion of risk period is no longer relevant, and the within-individual standard deviation of the exposure variable must be considered instead. To date, the only application of self-controlled case series methods with continuous exposure variables of which we are aware is to environmental time series. We have argued elsewhere that time series methods are generally more appropriate than case series methods for the analysis of such data [12], and in any case the sample sizes used in such studies are usually large.

References

[1] Andrews, N.J., 2002. Statistical assessment of the association between vaccination and rare adverse events post licensure. *Vaccine* **20** S49-S53.

[2] Diggle, P.J., Heagerty, P., Liang, S.L. and Zeger, S.L., 2002. Analysis of Longitudinal Data, 2nd edition. Oxford University Press, New York.

[3] Farrington, C.P., 1995. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics*, **51** 228-235.

[4] Farrington, C.P., Nash, J., and Miller, E., 1996. Case series analysis of adverse reactions to vaccines: a comparative evaluation. *American Journal of Epidemiology*, **143** 1165-1173 (Erratum 1998 **147** 93).

[5] Farrington, C.P., Miller, E. and Taylor, B., 2001. MMR and autism: further evidence against a causal association. *Vaccine*, **19** 3632-3635.

[6] Farrington, C.P., 2004. Control without separate controls: Evaluation of vaccine safety using case-only methods. *Vaccine*, **22** 2064-2070.

[7] Farrington, C.P. and Whitaker, H.J., 2006. Semiparametric analysis of case series data (with Discussion). *Journal of Royal Statistical Society, Series C*, In Press.

[8] Garthwaite, P.H., Jolliffe, I.T. and Jones, B., 2002. Statistical Inference, 2nd edition. Oxford University Press, New York.

[9] Miller, E., Goldacre, M., Pugh, S., Colville, A., Farrington, P., Flower, A., Nash, J., MacFarlane, L. and Tettmar, R., 1993. Risk of aseptic meningitis after measles, mumps and rubella vaccine in UK children. *The Lancet*, **341** 979-982

[10] Miller, E., Waight, P., Farrington, P., Stowe, J. and Taylor, B., 2001. Idiopathic thrombocytopenic purpura and MMR vaccine. *Archives of Disease in Childhood*, **84** 227-229.

[11] Whitaker, H.J., Farrington, C.P., Spiessens, B. and Musonda, P., 2006. Tutorial in Biostatistics: The self-controlled case series method. *Statistics in Medicine*, **25** 1768-1797.

[12] Whitaker, H.J., Hocine, M.N. and Farrington, C.P., 2006. On case-crossover methods for environmental time series data. *Environmetrics*, in press.

Table 1 Standard scenario. First row: median estimate of $\beta = \log(RI)$. Second row: percentage coverage of 95% confidence interval.

True value		$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
RI	β							
0.5	**-0.693**	-∞	-∞	-0.973	-0.676	-0.752	-0.703	-0.701
		97	96	96	97	97	96	95
1	**0.000**	-∞	-0.120	-0.006	-0.005	-0.011	-0.004	-0.004
		96	97	96	97	95	95	95
1.5	**0.405**	0.541	0.347	0.391	0.380	0.400	0.401	0.404
		97	97	97	0.96	95	95	96
2	**0.693**	0.695	0.646	0.676	0.681	0.689	0.693	0.691
		96	96	97	96	95	95	95
5	**1.609**	1.584	1.617	1.611	1.612	1.612	1.610	1.610
		98	97	96	95	95	95	95
10	**2.303**	2.415	2.367	2.325	2.315	2.306	2.304	2.305
		99	96	95	95	95	95	95

Table 2 Short risk periods. First row: median estimate of $\beta = \log(RI)$. Second row: percentage coverage of 95% confidence interval.

		1 day risk period			**5 day risk period**		
True value		$n = 20$	$n = 100$	$n = 500$	$n = 20$	$n = 100$	$n = 500$
RI	β						
0.5	**-0.693**	-∞	-∞	-∞	-∞	-∞	-0.634
		98	98	98	97	98	96
1	**0.000**	-∞	-∞	-0.074	-∞	-0.108	-0.058
		96	98	99	98	97	97
1.5	**0.405**	-∞	-∞	-0.035	-∞	0.018	0.390
		94	96	98	96	97	96
2	**0.693**	-∞	-∞	0.623	-∞	0.625	0.646
		94	96	97	94	97	96
5	**1.609**	-∞	1.534	1.554	1.557	1.577	1.607
		98	97	96	96	97	95
10	**2.303**	-∞	2.367	2.269	2.299	2.291	2.299
		95	96	96	97	96	95

Table 3 Longer risk periods. First row: median estimate of $\beta = \log(RI)$. Second row: percentage coverage of 95% confidence interval.

True value		50 day risk period			100 day risk period		
RI	β	***n* = 20**	***n* = 100**	***n* = 500**	***n* = 20**	***n* = 100**	***n* = 500**
0.5	**-0.693**	-0.813 97	-0.712 97	-0.699 95	-0.790 97	-0.712 95	-0.697 95
1	**0.000**	-0.058 97	-0.015 97	-0.003 95	-0.016 97	-0.015 97	0.003 95
1.5	**0.405**	0.401 97	0.398 96	0.402 95	0.412 97	0.409 95	0.405 95
2	**0.693**	0.675 97	0.690 95	0.693 95	0.709 96	0.700 95	0.694 95
5	**1.609**	1.637 97	1.616 95	1.611 95	1.706 95	1.623 95	1.611 95
10	**2.303**	2.410 96	2.322 95	2.306 95	2.437 96	2.335 95	2.308 95

Table 4 Effect of age at event for samples of size 100. First row: median estimate of $\beta = \log(RI)$. Second row: percentage coverage of 95% confidence interval.

Risk period (days)	True value		Weak symmetric age effect	Strong symmetric age effect	Weak monotone increasing age effect	Strong monotone increasing age effect
	RI	β				
10	**1**	**0.000**	-0.054 97	-0.021 96	-0.026 95	-0.032 97
	2	**0.693**	0.641 97	0.683 97	0.679 97	0.686 96
	5	**1.609**	1.596 96	1.605 95	1.592 95	1.601 96
25	**1**	**0.000**	-0.005 97	-0.033 96	-0.017 97	-0.029 97
	2	**0.693**	0.681 96	0.685 97	0.683 95	0.684 96
	5	**1.609**	1.612 95	1.619 95	1.615 95	1.616 96
50	**1**	**0.000**	-0.015 97	-0.009 96	-0.016 97	-0.011 97
	2	**0.693**	0.690 95	0.698 97	0.689 95	0.693 96
	5	**1.609**	1.616 95	1.627 95	1.615 95	1.628 96

Table 5 Effect of age at exposure for samples of size 100. First row: median estimate of $\beta = \log(RI)$. Second row: percentage coverage of 95% confidence interval.

Exposure distribution		True value		25 day risk period		50 day risk period	
Mean	SD	RI	β	Weak symmetric age effect	Strong monotone increasing age effect	Weak symmetric age effect	Strong monotone increasing age effect
250	**100**	**1**	**0.000**	-0.005 97	-0.029 97	-0.015 97	-0.011 97
		5	**1.609**	1.612 95	1.616 96	1.616 95	1.628 96
250	**50**	**1**	**0.000**	-0.027 97	-0.030 97	-0.093 96	-0.019 95
		5	**1.609**	1.616 95	1.613 95	1.609 95	1.626 95
125	**100**	**1**	**0.000**	-0.020 97	-0.052 97	-0.014 96	-0.026 96
		5	**1.609**	1.611 95	1.620 95	1.620 95	1.628 95
125	**50**	**1**	**0.000**	-0.030 97	-0.039 96	-0.014 96	-0.017 97
		5	**1.609**	1.608 95	1.622 95	1.618 95	1.629 95

Table 6 Indefinite risk periods. First row: median estimate of $\beta = \log(RI)$. Second row: percentage coverage of 95% confidence interval.

Exposure distribution		True value		100 exposed cases	100 exposed cases and 20 unexposed	100 exposed cases and 50 unexposed	100 exposed cases and 100 unexposed
Mean	SD	RI	β				
125	**50**	**1**	**0.000**	0.004 95	0.004 96	0.001 95	0.005 95
		2	**0.693**	0.715 96	0.713 96	0.712 96	0.713 97
		5	**1.609**	1.696 96	1.669 97	1.684 97	1.670 97
125	**30**	**1**	**0.000**	0.003 95	0.002 95	0.001 95	0.002 96
		2	**0.693**	0.731 96	0.716 97	0.722 97	0.712 97
		5	**1.609**	1.733 96	1.652 97	1.682 97	1.666 97
125	**10**	**1**	**0.000**	-0.090 96	-0.012 96	0.009 96	0.003 97
		2	**0.693**	0.731 98	0.684 98	0.716 96	0.728 97
		5	**1.609**	2.824 98	1.679 98	1.668 98	1.662 97

Table 7 Asymptotic and bootstrap results for aseptic meningitis and ITP data

Data set	Risk period (days)	Asymptotic		Bootstrap		
		Estimate	95% CI	Median $m_B(\hat{\beta})$	Percentile 95% CI	Bias corrected 95% CI
Meningitis	15 – 25	2.488	1.099, 3.876	2.488	0.938, 4.116	1.075, 4.116
ITP	0 – 14	0.269	-1.206, 1.745	0.221	$-\infty$, 1.494	$-\infty$, 1.571
	15 – 28	1.784	0.924, 2.644	1.798	0.702, 2.741	0.647, 2.718
	29 – 42	0.995	-0.294, 2.205	0.932	$-\infty$, 2.092	$-\infty$, 2.130

Figure 1 bias($\hat{\beta}$) for $n = 50$ against r, the ratio of the risk period to the observation period, for different values of the relative incidence (RI).

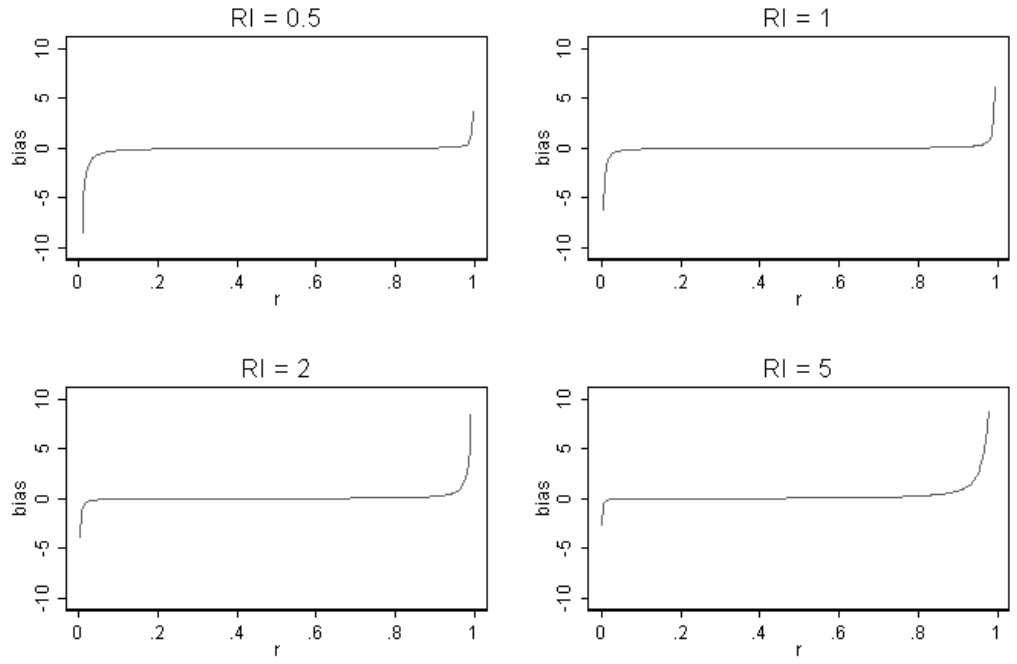

Figure 2 var($\hat{\beta}$) for $n = 50$ against r, the ratio of the risk period to the observation period, for different values of the relative incidence (RI).

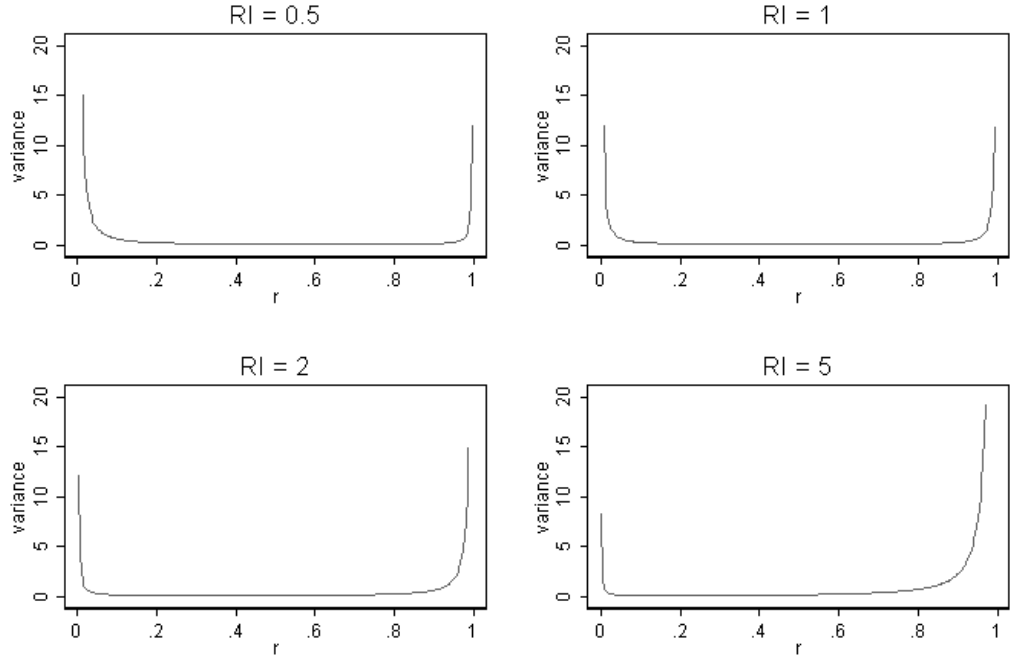

Figure 3 Structure of the simulation study.

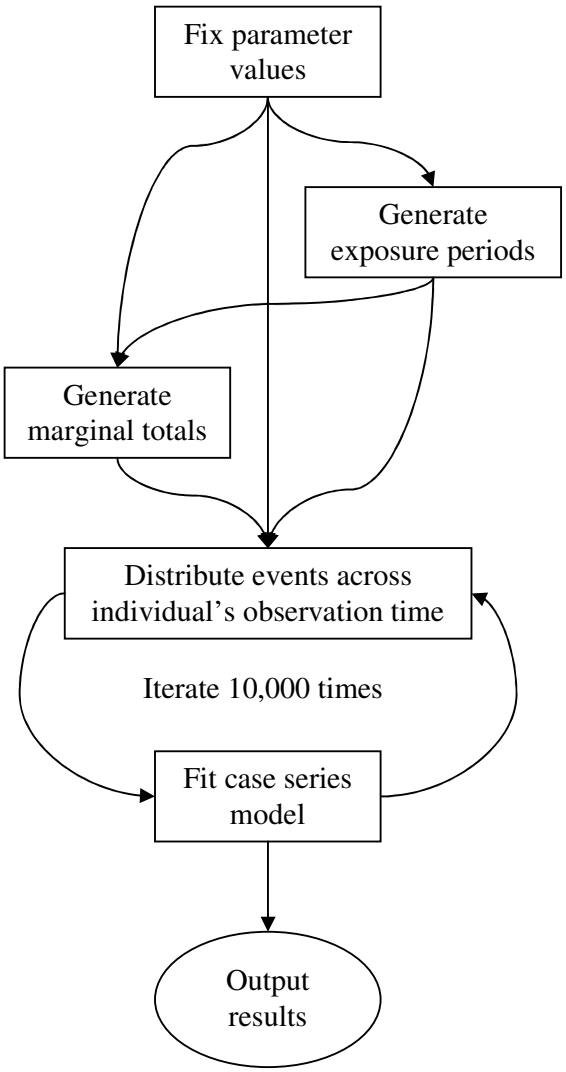

Figure 4 The four effects of age at event used in the simulations

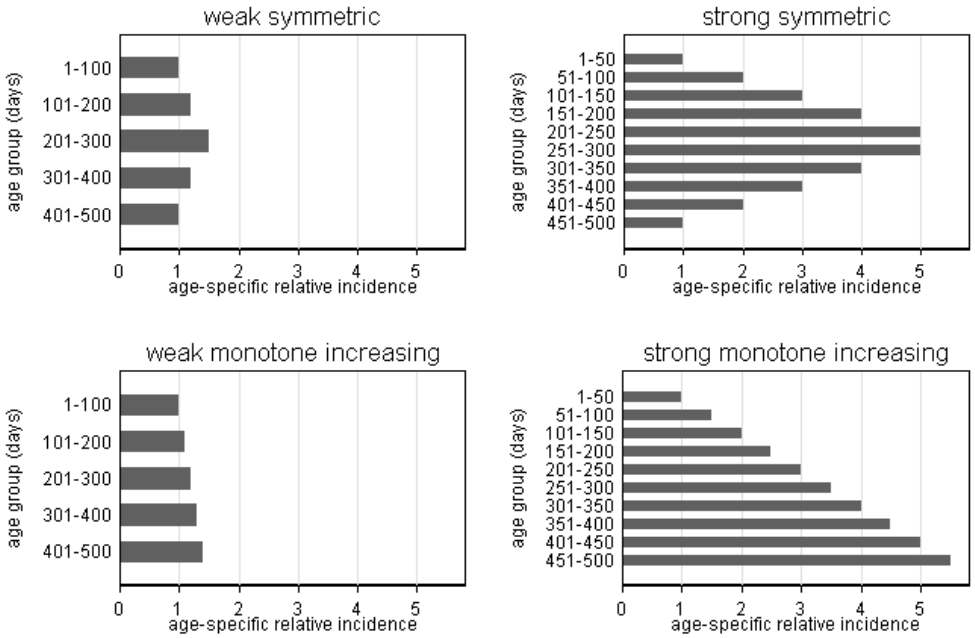

Figure 5 Four distributions of age at exposure used in the simulations

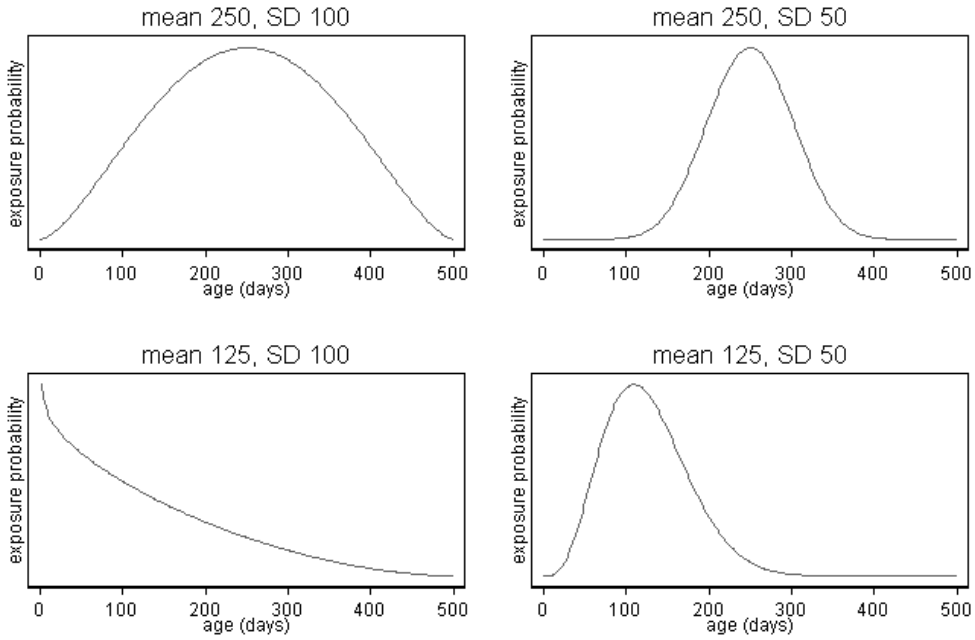

Figure 6 Bootstrap distribution of relative incidence for aseptic meningitis and ITP data, by risk period.

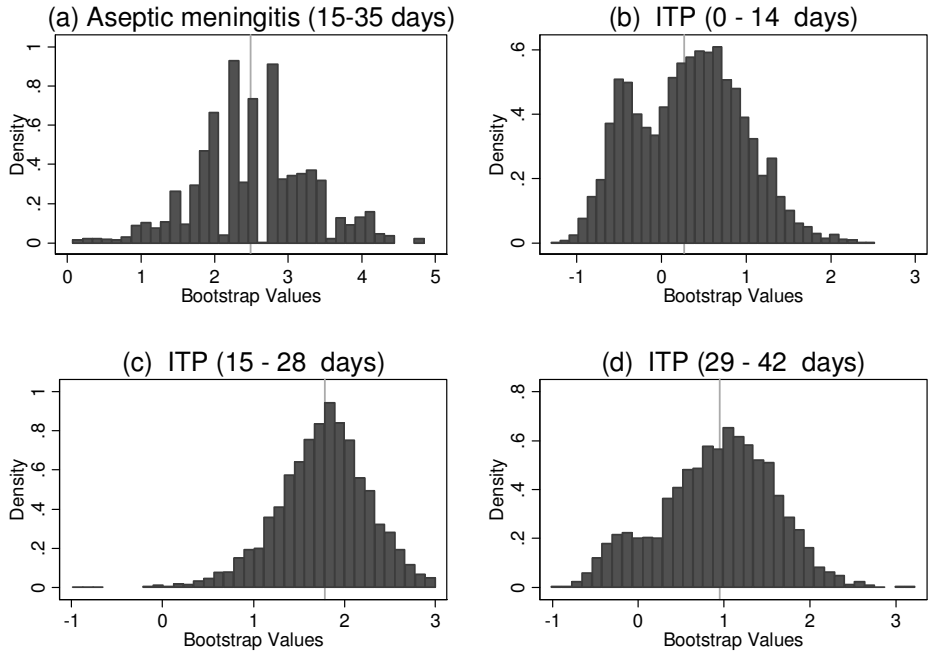

Figure 7 Randomization and asymptotic distributions of the likelihood ratio statistic