STATISTICS IN MEDICINE
Statist. Med. 2005; **0**:1–31 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

TUTORIAL IN BIOSTATISTICS: The self-controlled case series method

Heather J. Whitaker[1], C. Paddy Farrington[1], Bart Spiessens[2] and Patrick Musonda[1]

[1] *Department of Statistics, The Open University, Milton Keynes, MK7 6AA, UK.*
[2] *GlaxoSmithKline Biologicals, Rue de l'Institut 89, B-1330 Rixensart, Belgium.*

SUMMARY

The self-controlled case series method was developed to investigate associations between acute outcomes and transient exposures, using only data on cases, that is, on individuals who have experienced the outcome of interest. Inference is within individuals, and hence fixed covariates effects are implicitly controlled for within a proportional incidence framework. We describe the origins, assumptions, limitations, and uses of the method. The rationale for the model and the derivation of the likelihood are explained in detail using a worked example on vaccine safety. Code for fitting the model in the statistical package STATA is described. Two further vaccine safety data sets are used to illustrate a range of modelling issues and extensions of the basic model. Some brief pointers on the design of case series studies are provided. The data sets, STATA code, and further implementation details in SAS, GENSTAT and GLIM are available from an associated website.

key words: case series; conditional likelihood; control; epidemiology; modelling; proportional incidence

1. Introduction

The self-controlled case series method, or case series method for short, provides an alternative to more established cohort or case-control methods for investigating the association between a time-varying exposure and an outcome event. It was developed to investigate associations between vaccination and acute potential adverse events [1], [2]. Subsequently the method was applied in other settings, for example to investigate non-acute events such as autism, and has been used more widely in pharmaco-epidemiology and other areas of epidemiology.

*Correspondence to: C.P. Farrington. Department of Statistics, The Open University, Milton Keynes, MK7 6AA, UK. E-mail:c.p.farrington@open.ac.uk

Contract/grant sponsor: HJW and CPF were supported by a grant from the Wellcome Trust; contract/grant number: 070346

Contract/grant sponsor: PM was supported by an EPSRC CASE studentship with funding from GlaxoSmithKline Biologicals; contract/grant number: 0307

Accepted 25th April 2005

This tutorial has three main aims. First, we provide an account of the background to the method, and explain how it works by means of a worked example. Second, we describe how the method may be implemented in commercial software packages. The software and several of the data sets are available online from http://statistics.open.ac.uk/sccs. Third, we describe in detail how the method is applied, the assumptions it requires and its limitations, through further practical examples. We do not aim to present the most general version of the model, which can readily be extended to handle time-varying continuous exposures, and to non-parametric modelling of the baseline incidence; pointers are given where appropriate to these more general versions. Instead we concentrate on transient exposures and parametric modelling.

The paper is organised as follows. In section 2 we describe the background of the method, briefly discuss alternative approaches, set out the main advantages and limitations of the case series method, and describe how it has been used. In section 3, we work through an example, small enough to allow all calculations to be done by hand, and set out the more general theory, including the likelihood and the associated regression model. In section 4 we describe how the model is fitted using STATA. In section 5 we present two other data sets, which we use for illustrative purposes in section 6, where we discuss at some length the assumptions and limitations of the method, as well as issues in the analysis of case series studies. In section 7 we briefly discuss the design of case series studies. Section 8 is a short discussion, including areas for further research.

2. The self-controlled case series method

2.1. Background

The method described in this tutorial was developed to analyse record linkage data on MMR vaccination and aseptic meningitis [3], [4]. Data on episodes of aseptic meningitis arising in children aged 1-2 years over a defined calendar time period (the observation period) were obtained from laboratory and hospital records. Throughout the paper, the term 'case' refers to an individual who has experienced one or more events of interest over the observation period. These cases were linked to vaccination records: the resulting data set thus consisted of cases and their exposures. No precise denominators were available, nor was it wholly clear from what population the cases arose, since the catchment areas of the hospitals from which the cases were obtained were not clearly defined. Thus population-based cohort and case-control studies would have required some ingenuity to avoid confounding, as vaccine coverage is not uniform. Furthermore, answers about a possible association with MMR vaccine were required quickly. The case series method was developed in response to these requirements. A positive association between vaccination with the Urabe mumps strain and aseptic meningitis in the period 15-25 days post-vaccination was confirmed, and the composition of MMR vaccines used in the UK was changed.

2.2. Relation to other methods

In a case series analysis, only cases are sampled. To take account of this, the likelihood is conditional on an event having occurred during the observation period. The case series likelihood is thus based on the probability density that an event occurred when it did in relation

to the individual's time of exposure, given that the event occurred during the observation period. Note that this derives from cohort logic: event times are regarded as random, whereas exposure times are regarded as fixed. The focus of estimation, as in a cohort study, is the relative incidence, or relative hazard of an event. This is the ratio of the rate (or hazard) of events in a given post-exposure period, to the rate of events in the absence of the exposure.

An early approach to analysing case-only data was proposed by Aalen et al. [5]. This was later generalised by Prentice et al. [6], who suggested applying Cox regression to the cohort of cases, effectively ignoring the fact that only cases are sampled. This provides a valid test of the null hypothesis of no association, but does not yield a useful effect estimate. This method was used by Miller et al. [3] to analyse the MMR vaccination and aseptic meningitis data described above, since the case-series method had not yet been published at this time.

Maclure [7] proposed a case-control method involving only cases, known as the case-crossover method. In this method, exposures in the period immediately preceding the event time are compared to exposures at earlier 'control' times in the history of the case. There are several variants of this method, reviewed by Greenland [8], and the case-crossover approach has been used in many settings [9]. However, a necessary condition for the case-crossover method to apply is that the exposure distribution is stationary. Indeed, a stronger assumption is required in many settings, namely that the exposure distribution in successive time periods is exchangeable [10]. Without this condition, which mimics the tacit assumption in case-control studies that the ordering of the controls is immaterial, multiple logistic regression applied to case-crossover data with several control periods does not produce consistent estimates. The case series method does not require such an assumption. In particular, age or time effects can be allowed for in much the same way as in a cohort study.

Feldmann [11] [12] proposed an approach similar to the case series method insofar as it is also based on cohort logic, conditioning on the probability that one or more events occur within each individual's observation period. For rare events, Feldmann's method coincides approximately with the case series model without adjustment for age. However, in general, the method does not implicitly control for fixed covariates.

The case series method incorporates features of these other methods: like Prentice's, it controls for age; like Maclure's, it controls for fixed confounders; and like Feldmann's, it allows for multiple events.

2.3. Advantages and limitations

The main advantages of the case series method over other methods of analysis can be summarised as follows.

1. It is based only on cases, and provides consistent estimates of the relative incidence.
2. It controls implicitly for all fixed confounders, that is to say, confounders that do not vary with time over the observation period, such as variables relating to genetics, location, socio-economic status, gender, individual frailty, severity of underlying disease, etc.
3. Age or temporal variation in the baseline incidence can be allowed for in the model.
4. Under certain circumstances it can have high efficiency relative to the retrospective cohort method from which it is derived by conditioning.

The main limitations of the case series method are as follows.

1. It requires that the probability of exposure is not affected by the occurrence of an outcome event.
2. For non-recurrent events the method works only when the event risk is small over the observation period.
3. It does not produce estimates of absolute incidence, only estimates of relative incidence.
4. It requires variability in the time or age of the event: if all events were to happen at exactly the same age, then the method would fail.

These advantages and limitations will be reviewed in subsequent sections. The single most important limitation is the requirement that events do not affect subsequent exposures. In some circumstances, this assumption can be circumvented, as will be described later.

2.4. Applications

Table I documents the published applications of the case series method to transient exposures. Most applications so far are to vaccine safety; Andrew [13] and Farrington [14] review applications in this field using the case series and other methods. The case series method has also been used in other areas of epidemiology. For example, Becker et al. [15] independently used the method to investigate the association between long-haul flights and thromboembolism. Farrington et al. [2] and Andrews [13] include direct comparisons between the case series method and conventional methods of analysis.

Navidi [16] also independently proposed what is essentially a case series method with time-varying exposures, for application in studies of air pollution. This method is described as a bi-directional (or ambidirectional) case-crossover method. The case series version of this method is that in which the entire observation period is used as controls. This model has also been discussed by Lumley and Levy [17]. In the present paper, only transient qualitative exposures, such as receipt of a drug, will be considered. A more general treatment of the case series model, including an application to the analysis of air pollution data, is given by Farrington and Whitaker [18].

3. The case series likelihood

The case series likelihood is first motivated by working through an example which, we hope, will help clarify the key points of the method. Then the more general model is described.

3.1. A worked example

The example is based on hospital data on MMR vaccination and viral meningitis from Oxford, originally analyzed by Miller et al. [3]. Specifically, we investigate the hypothesis that vaccination with a particular type of live mumps vaccine (the Urabe strain) is associated with an increase in the risk of viral meningitis.

In this study, cases of viral meningitis diagnosed in the second year of life between October 1988 and December 1991 were obtained and linked to vaccination records. The observation period for each child is the time during which, if an event arose, it would be sampled. Thus, in this study, the observation period includes all time between 366 and 730 days of age and

Table I. Studies using the SCCS method

Exposure	Outcome	Reference
DTP vaccine*	Febrile convulsion	Farrington et al. [4]
MMR vaccine†	Febrile convulsion	Farrington et al. [4]
MMR vaccine†	Idiopathic thrombo-cytopenic purpura	Farrington et al. [4] Miller et al. [19]
MMR vaccine†	Aseptic meningitis	Farrington et al. [4] Dourado et al. [20]
MMR vaccine†	Autism	Taylor et al. [21] Farrington et al. [22]
MMR vaccine†	Invasive bacterial infection	Miller et al. [23]
MMR vaccine†	Gait disturbance	Miller et al. [24]
Influenza vaccine	Asthma	Kramarz et al. [25] Tata et al. [26]
Influenza vaccine	Bell's palsy	Mutsch et al. [27]
Oral polio vaccine	Intussusception	Andrews et al. [28] Galindo Sardinas [29]
Oral rotavirus vaccine	Intussusception	Murphy et al. [30]
DTP*, MMR†, HBV HIB, OPV vaccines‡	Wheezing	Mullooly et al. [31]
Antidepressants	Hip fracture	Hubbard et al. [32]
Antidepressants	Myocardial infarction	Tata et al. [33]
Long-haul air travel	Venous thromboembolism	Becker et al. [15]
Influenza vaccine	Any medical visits	France et al. [34]
Common vaccines and infections	Myocardial infarction and stroke	Smeeth et al. [35]

*DTP = diphtheria, tetanus, pertussis.
†MMR = measles, mumps, rubella.
‡HBV = hepatitis B vaccine. HIB = haemophilius influenzae type B.
OPV = oral polio vaccine.

between the 1st October 1988 and 31st December 1991 inclusive: note that it is important to be rigorously precise about ages and timings.

Based on evidence from previous studies, and on knowledge of the time taken by the mumps virus to replicate, we choose the risk period to contain days 15 to 35 inclusive after receipt of MMR. Using the information obtained from vaccination records, the exposure histories of all cases during the observation period is documented. Note that vaccination histories between ages 331 and 715 days are required so as to classify all days in the observation period as exposed or unexposed.

Ten viral meningitis diagnoses, in ten children, met the selection criteria. Table II shows the data from these ten children. The data include an identifier for each individual (called *individual* in table II), age on the day before the start of the observation period (*pre-start*), age on the day before the first day at risk (*pre-risk*), age on the last day at risk (*risk end*), age on the day of diagnosis (*diagnosis*) and age on the last day of the observation period (*end day*). The beginning of a period is always indicated by listing the last day of the preceding

period, as this is required by the software described later. In this case each child had a single event, so there is a single record per child. For all ten children, the observation period was 366 to 730 days (but note that this need not have been so). Child 10 had zero days in the post-exposure risk period; the other 9 all had 21 days 'at risk'. Of the ten events, 5 occurred within the post-exposure risk period, and 5 occurred outside it.

Table II. Oxford data

individual	pre-start	pre-risk	risk end	diagnosis	end day
1	365	472	493	398	730
2	365	406	427	413	730
3	365	443	464	449	730
4	365	447	468	455	730
5	365	446	467	472	730
6	365	409	430	474	730
7	365	484	505	485	730
8	365	510	531	524	730
9	365	442	463	700	730
10	365	-	-	399	730

The baseline risk (that is, the risk in the absence of exposure to MMR vaccine) of viral meningitis varies with age. To keep matters simple, in this example only two age groups will be used: ages 366 to 547 days (roughly 1 year to 1 year 6 months) and ages 548 to 730 days. The individual data are then expanded as follows. Each individual's observation time is split up into successive time intervals determined by the changes in age group, exposure status, and by the start and end of the observation period, as shown in Figure 1 for the first 3 children. In this figure, the symbol] indicates that the corresponding day is included in the interval to the left.

The data are then expanded, each line of the expanded data corresponding to a single time interval for one individual with a combination of the indices i, j, k, where $i = 1, ..., 10$ ranges over individuals, $j = 0, 1$ corresponds to age groups, and $k = 0, 1$ corresponds to risk periods (1 if at risk, 0 if not). The length of each interval is denoted e_{ijk} (in days), and the number of events occurring within each interval for individual i is denoted n_{ijk}. Table III shows how the data were expanded for the first three individuals. The expanded data are in a suitable form for log-linear modelling, but first we explain the rationale for the model.

We assume that viral diagnoses arise in a non-homogeneous Poisson process. Denote by e^{ϕ_i} the baseline incidence of viral meningitis for individual i in age group j, and e^{α_1} and e^{β_1} the relative incidence associated with age group $j = 1$ and exposure $k = 1$, respectively, relative to age group $j = 0$ and non-exposure $k = 0$. The baseline incidence for individual i, e^{ϕ_i}, may depend on the characteristics of the individual. With other methods these characteristics can be modelled explicitly; this is not necessary with the case series method.

Take the first child, $i = 1$. This child's observation period was split into four intervals. During the first interval, from day 366 to 472, child 1 was in age group $j = 0$ and was not at risk, so $k = 0$, thus e_{100} =107 days. The Poisson incidence rate during this first interval is $\lambda_1 = 107e^{\phi_1}$. The second interval, including days 473 to 493, was 15-35 days after MMR

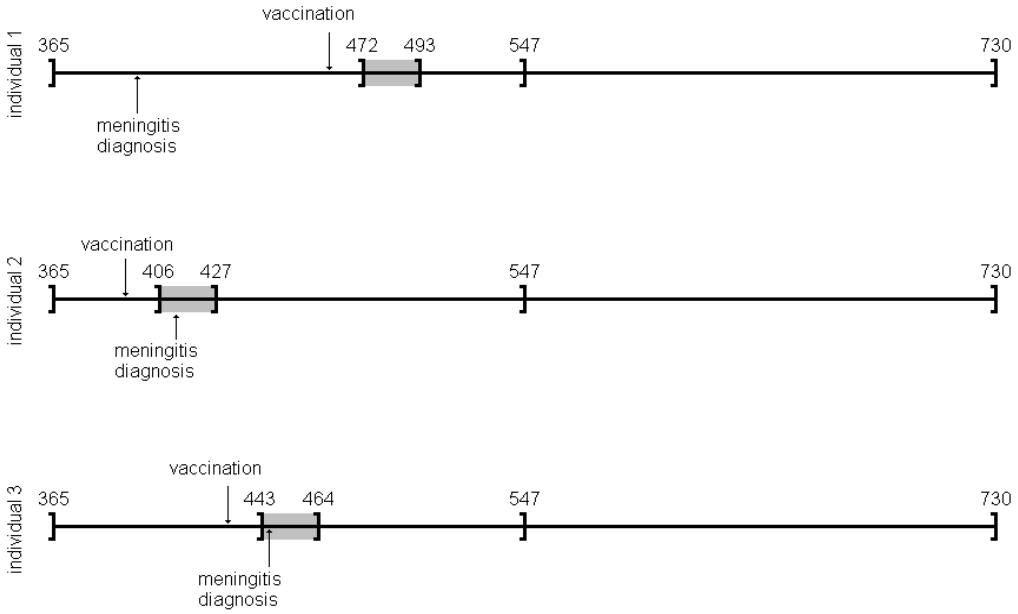

Figure 1. Diagram of the observation period for the first 3 individuals. Vaccine risk periods are shaded.

Table III. MMR and meningitis data, expanded for the first three children.

individual i	age group j	vaccine risk factor k	interval length e_{ijk}	number of events n_{ijk}
1	0	0	107	1
1	0	1	21	0
1	0	0	54	0
1	1	0	183	0
2	0	0	41	0
2	0	1	21	1
2	0	0	120	0
2	1	0	183	0
3	0	0	78	0
3	0	1	21	1
3	0	0	83	0
3	1	0	183	0

vaccination, thus child 1 was at risk, so $k = 1$, but was still in age group $j = 0$. After this, child 1's exposure status returned to $k = 0$, and from day 494 to 547 remained in age group $j = 0$. In the fourth interval, spanning days 548 to day 730, child 1 was in age group $j = 1$, and not at risk, $k = 0$. The Poisson incidence rates for the second, third and fourth intervals respectively are $\lambda_2 = 21e^{\phi_1+\beta_1}$, $\lambda_3 = 54e^{\phi_1}$ and $\lambda_4 = 183e^{\phi_1+\alpha_1}$. Hence the Poisson rate for child 1's observation period as a whole is

$$\Lambda = 107e^{\phi_1} + 21e^{\phi_1}e^{\beta_1} + 54e^{\phi_1} + 183e^{\phi_1}e^{\alpha_1}.$$

If this were a cohort study, child 1's contribution to the Poisson likelihood would be

$$\left(\lambda_1 e^{-\lambda_1}\right) \times e^{-\lambda_2} \times e^{-\lambda_3} \times e^{-\lambda_4} = \lambda_1 e^{-\Lambda}$$

because one event occurred in the first interval and zero events occurred in intervals 2, 3 and 4. However, child 1 was sampled *because* one or more episodes of viral meningitis were diagnosed during the observation period. The probability of this occurring is $\Lambda e^{-\Lambda}$. To take account of the fact that we have sampled only cases, we must condition on this event. Thus the conditional likelihood is

$$\begin{aligned}
\frac{\lambda_1 e^{-\Lambda}}{\Lambda e^{-\Lambda}} &= \frac{\lambda_1}{\Lambda} \\
&= \frac{107e^{\phi_1}}{107e^{\phi_1} + 21e^{\phi_1}e^{\beta_1} + 54e^{\phi_1} + 183e^{\phi_1}e^{\alpha_1}} \\
&= \frac{107}{107 + 21e^{\beta_1} + 54 + 183e^{\alpha_1}}.
\end{aligned}$$

This conditional likelihood is the case series likelihood for child 1. Equivalently, the likelihood for child 1 can be thought of as based on the multinomial probability that the event occurred in the first interval, given that it occurred in one of intervals 1, 2, 3 and 4. The general model described in the next section is based on such a multinomial formulation. For child 1, the multinomial likelihood is:

$$\begin{aligned}
l(\alpha_1, \beta_1) = &\left(\frac{107}{107 + 21e^{\beta_1} + 54 + 183e^{\alpha_1}}\right)^1 \times \left(\frac{21e^{\beta_1}}{107 + 21e^{\beta_1} + 54 + 183e^{\alpha_1}}\right)^0 \times \\
&\left(\frac{54}{107 + 21e^{\beta_1} + 54 + 183e^{\alpha_1}}\right)^0 \times \left(\frac{183e^{\alpha_1}}{107 + 21e^{\beta_1} + 54 + 183e^{\alpha_1}}\right)^0
\end{aligned}$$

The likelihood for all 10 children is obtained in the same way as for child 1. Since children are

independent, the overall log-likelihood is:

$$\begin{aligned} l(\alpha_1,\beta_1) &= \log\left(\frac{107}{161+21e^{\beta_1}+183e^{\alpha_1}}\right)+\log\left(\frac{21e^{\beta_1}}{161+21e^{\beta_1}+183e^{\alpha_1}}\right)+ \\ &\quad \log\left(\frac{21e^{\beta_1}}{161+21e^{\beta_1}+183e^{\alpha_1}}\right)+\log\left(\frac{21e^{\beta_1}}{161+21e^{\beta_1}+183e^{\alpha_1}}\right)+ \\ &\quad \log\left(\frac{80}{161+21e^{\beta_1}+183e^{\alpha_1}}\right)+\log\left(\frac{117}{161+21e^{\beta_1}+183e^{\alpha_1}}\right)+ \\ &\quad \log\left(\frac{21e^{\beta_1}}{161+21e^{\beta_1}+183e^{\alpha_1}}\right)+\log\left(\frac{21e^{\beta_1}}{161+21e^{\beta_1}+183e^{\alpha_1}}\right)+ \\ &\quad \log\left(\frac{183e^{\alpha_1}}{161+21e^{\beta_1}+183e^{\alpha_1}}\right)+\log\left(\frac{182}{182+183e^{\alpha_1}}\right) \\ &= constant+5\beta_1+\alpha_1-9\log(161+21e^{\beta_1}+183e^{\alpha_1})-\log(182+183e^{\alpha_1}). \end{aligned}$$

Maximizing this log-likelihood gives estimates $\alpha_1 = -1.491$ and $\beta_1 = 2.488$, so the relative incidence in the second age group is $e^{\alpha_1} = 0.225$ and the relative incidence associated with the exposure is $e^{\beta_1} = 12.037$.

Note that all the individual effects ϕ_i cancel out. The implications of this are discussed in the next subsection.

3.2. The general case

The general form of the case series likelihood for qualitative exposures is now derived. Assume that events arise within individuals as a non-homogeneous, age-dependent Poisson process. A random selection of n events in N individuals $i = 1, 2, ..., N$ each with at least one event are sampled; let n_i denote the number of events observed for individual i: $n_i \geq 1$ and $n = \sum_i n_i$. Note that more than one event can occur for each individual, in line with the Poisson assumption: the special case where events are necessarily unique, and other departures from the Poisson assumption, will be discussed later, along with other assumptions of the method.

The observation period for individual i is the time period during which an event could be sampled. Observation periods may (and often do) vary between individuals. Individual i's observation period is split into intervals for age groups, indexed by j, and risk periods, indexed by k. The reference age group and risk period are denoted 0. Note that in general we allow several risk periods: for example, in pharmacoepidemiology, an individual is unexposed ($k = 0$) during the period prior to exposure to a drug. For a period after exposure to the drug, the individual might be considered at high risk, coded $k = 1$. This first risk period might then be followed by a second, in which the risk might be intermediate, coded $k = 2$. After this the individual might return to the reference risk level ($k = 0$).

Let e_{ijk} denote the time spent by individual i in age group j and risk period k. The incidence, denoted λ_{ijk}, is assumed to be constant within each interval. We assume a multiplicative model for the incidence function:

$$\lambda_{ijk} = \exp(\phi_i + \alpha_j + \beta_k)$$

where ϕ_i represents an effect for each individual i, α_j represents an age group effect, and β_k represents an effect for risk group k, with $\alpha_0 = 0$ and $\beta_0 = 0$. The incidence function during the baseline period is simply $\lambda_{i00} = \exp(\phi_i)$.

Conditioning on the number of events n_i observed for individual i during the observation period, the log likelihood is multinomial:

$$l(\alpha, \beta) = \sum_{ijk} n_{ijk} \log \left[\frac{\exp(\alpha_j + \beta_k) e_{ijk}}{\sum_{rs} \exp(\alpha_r + \beta_s) e_{irs}} \right]$$

Note that all the individual effects ϕ_i cancel out. This always occurs, because incidence rates are contrasted within the same individual's person-time: in this sense, the case series method is self-controlled. Provided that the model is correct, inferences from a case series analysis cannot be confounded by individual effects, such as genetic factors, location, socio-economic status, sex, underlying health status, individual frailty, and so on. Such covariates can of course modify the effect of exposure: this can be modeled by including suitable interaction terms, as will be illustrated later. Finally, it is important to emphasize that self-control applies to fixed covariates, not age or time-dependent covariates.

4. Fitting the model

The multinomial model can be fitted in standard software as an associated Poisson model with log link function. The response variable is the number of events in each interval, n_{ijk}, and log of the time spent in the interval $ln(e_{ijk})$ is included as an offset. As well as factors for the risk group k and age group j, a factor is included for each individual i to ensure that the fitted individual totals equal the observed values. Thus the associated Poisson main effects model is

$$n_{ijk} \sim Poisson(\lambda_{ijk} e_{ijk})$$
$$\log(\lambda_{ijk}) = \phi_i + \alpha_j + \beta_k.$$

In the remainder of this section it is described how to fit the model with the statistical package STATA [36]. Implementation in SAS [37], GENSTAT [38] and GLIM [39] are also available from the case series website; details are given in section 4.5.

4.1. Example: Oxford data

When the data are in the correct format, as in Table III, fitting the model in STATA or any statistical package is simple using a Poisson regression model. In STATA we have named our variables as follows: `indiv` is the individual identifier (i), `exgr` the exposure group (k) (which takes the value 1 for the period during which the individual is 'at risk' from the exposure of interest and 0 otherwise), `agegr` the age group factor (j), `loginterval` the log of the time spent in the interval (in days) ($\ln(e_{ijk})$), and finally `nevents` the number of events the individual experiences within the interval (n_{ijk}). Then the model can be fitted using the `glm` command:

```
xi: glm nevents i.exgr i.agegr i.indiv,
        offset(loginterval) family(poisson) eform,
```

or using the `poisson` command.

The standard output includes the deviance (20.18), Pearson χ^2 (25.09), residual degrees of freedom (26), Bayesian information criterion (BIC) (-74.40) and a table of the exponentiated coefficient estimates as shown below (or just the coefficient estimates if the `eform` option is not included). The relative incidence is 12.04, with 95% confidence interval (3.00, 48.26).

```
------------------------------------------------------------------------------
             |               EIM
    nevents  |      IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iexgr_1  | 12.03687   8.527978     3.51   0.000     3.002255    48.25915
  _Iagegr_1  |  .225243   .2518756    -1.33   0.183     .0251654    2.016037
  _Iindiv_2  |         1   1.414214    -0.00   1.000     .0625488    15.98751
       ⋮
loginterval  | (offset)
------------------------------------------------------------------------------
```

4.2. Pre-processing the data in STATA

For all but the smallest problems, expanding the data by hand is impractical. One advantage of STATA is that it has commands to expand, collapse and reshape data that makes reformatting the data relatively simple and fast.

In STATA the individual identifier (i) is named `indiv` and the age when the adverse event occurred (meningitis diagnosis) `eventday`. To avoid ambiguity about whether a cut point is included in the interval to the left or the right, all cut points are included in the interval to the left. For example, if the start of the observation period is 366 days of age, then the cut point defining it is 365 days. The do file requires that all the cut points i.e. the day before the start, the end of the observation period, and the age and exposure group cut points, are given names beginning with the same string and ending in a number. We've used the string `cutp` and numbered the variables as follows: `cutp1` and `cutp2` are the day before the start of, and the end of the observation period for each individual respectively, `cutp3` is the cut point between the two age groups (common to all individuals) and finally `cutp4` and `cutp5` are the exposure group cut points, 14 and 35 days after administration of the MMR vaccine, for each individual. If throughout the observation period an individual was unexposed i.e. did not receive the MMR vaccine, such as child $i = 10$ in this example, the data were amended to state that a dose was given after the end of the observation period. A `generate` command is given to create a new variable, and a `rename` command is used to rename an existing variable. The ordering of the cut points should always be kept the same, i.e. cut points for the day before the start of and end of the observation period should always be followed by the age group cut points, and these should always be followed by the exposure group cut points. The number of age group cut points, which is 1 in this example, is listed in a local macro named `nage`.

```
local nage = 1
```

Now we describe how the data are reformatted. We start by using a `reshape` command so that the cut points, `cutp`, are listed by each adverse event (or individual) with a variable, `type`, listing the number corresponding the cut point (e.g. 2 for end of observation period, 3 for age group cut point). The data are sorted so that the cut points are listed in increasing order by each individual and adverse event.

```
sort indiv eventday
reshape long cutp, i(indiv eventday) j(type)
```


Prepared using simauth.cls

```
sort indiv eventday cutp type
```

A new variable, `nevents` (n_{ijk}), is generated which is 1 if an adverse event occurred between neighbouring cut points for each individual. A `collapse` command is given to sum events over each individual, in this example this just enters 0's where no events occurred.

```
by indiv: generate int nevents = 1 if
          eventday > cutp[_n-1] + 0.5 & eventday <= cutp[_n] + 0.5
collapse (sum) nevents, by(indiv cutp type)
```

The difference between consecutive cut points, `interval`, gives the length of each interval (e_{ijk}).

```
by indiv: generate int interval = cutp[_n] - cutp[_n-1]
```

The age groups (j), `agegr`, are generated by partitioning the cut points by a list of all the age group cut points using the `irecode` function. In this example there is only 1 age group cut point at age 547 days. Age groups are labeled 0 and 1.

```
by indiv: generate int agegr = irecode(cutp, 547)
```

The exposure groups (k) are generated in `exgr` by picking out the two exposure risk group cut points, which we know from the `type` variable, and filling in the risk group factors between these cut points. In this example there is one risk period ($k = 1$), and $k = 0$ time not at risk.

```
generate exgr = type-‘nage’-3 if type>‘nage’+2
count if exgr>=.
local nmiss = r(N)
local nchange = 1
while ‘nchange’ > 0{
              by indiv: replace exgr = exgr[_n+1] if exgr >= .
              count if exgr>=.
              local nchange = ‘nmiss’-r(N)
              local nmiss = r(N)
              }
replace exgr = 0 if exgr >= .
```

At this point the data held in STATA for the first three individuals are as in table IV. Variables that are not needed to fit the model and intervals of length 0 are dropped. The first record for each individual (when `interval` is a '.', STATA's missing value character) are also dropped, we are only interested in what is happening between consecutive cut points so there is one too many records for each individual.

```
drop cutp* type
drop if interval == 0 | interval >= .
```

Now the data are in the same format as in table III. Once the log of the intervals has been taken,

```
generate loginterval = log(interval),
```

Table IV. Example of how the MMR and meningitis data are expanded by STATA

indiv	type	cutp	nevents	interval	agegr	exgr
1	1	365	0	.	0	0
1	4	472	1	107	0	0
1	5	493	0	21	0	1
1	3	547	0	54	0	0
1	2	730	0	183	1	0
2	1	365	0	.	0	0
2	4	406	0	41	0	0
2	5	427	1	21	0	1
2	3	547	0	120	0	0
2	2	730	0	183	1	0
3	1	365	0	.	0	0
3	4	443	0	78	0	0
3	5	464	1	21	0	1
3	3	547	0	83	0	0
3	2	730	0	183	1	0

the data are ready to fit the Poisson regression model as described in the previous subsection. We also recommend saving the expanded data.

To run the MMR and meningitis in Oxford example in STATA using the files on our web page start by saving the data file listing one record per event (as in table II) as a STATA data file 'oxford.dta' and run the STATA do file 'oxford.do'.

4.3. Fitting larger models in STATA

The effects for each individual are needed only to fit the multinomial model as a Poisson model. They are nuisance parameters which have the potential to make model fitting very slow if many individuals are included in the analysis. Models can be fitted much faster using an *absorbing factor* for the individual effects. This absorbing factor can be used to partition model terms so that the size of the matrix to be inverted is smaller [40]. The parameters corresponding to the absorbing factor are then not calculated explicitly.

There is no standard command to fit a generalised linear model with absorbing factors in STATA, though linear regression models with absorbing factors can be fitted using the command `areg`. The `glm` command can be amended to fit absorbing factors with the iterative re-weighted least squares `irls` option: this makes a call to `regress` which fits standard linear regression models, and can be replaced with a call to `areg`. Only a few lines need to be changed in the file 'glm.ado' to do this; instructions on the changes to make are given on our website. Alternatively a ready-amended file 'aglm.ado' can be downloaded from the self controlled case series method website (we have renamed the command `aglm`). The models can then be fitted using:

```
xi: aglm nevents i.exgr i.agegr,
            offset(loginterval) family(poisson) irls eform.
```

This gives the following output:

```
------------------------------------------------------------------------------
             |               EIM
    nevents  |      IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Iexgr_1  | 12.03687   8.527977     3.51   0.000     3.002256    48.25914
  _Iagegr_1  |  .225243   .2518756    -1.33   0.183     .0251654    2.016037
loginterval  | (offset)
------------------------------------------------------------------------------
```

Estimates differ only very slightly from those given by the `glm` command. Note that the individual effects are not included in this table.

4.4. Other software

Most statistical packages can fit GLMs, though to our knowledge absorption methods for fitting GLMs with many nuisance parameters are only available as standard in GLIM and GENSTAT, and in STATA and SAS, provided by us. In GLIM, this is the `eliminate` command, and in GENSTAT this is via the `groups` option in the `model` directive. Different packages use different algorithms, occasionally resulting in slightly different parameter estimates. Such differences are of no material importance.

4.5. Website

The case series website may be found at http://statistics.open.ac.uk/sccs. This website contains a brief description of the method and a list of references of published work on and using the method. It also contains STATA, SAS, GLIM and GENSTAT macros to download, including files and data sets to run the examples given in this tutorial.

5. Two further examples

In this section we introduce two further examples, which we shall use in the next section to illustrate modelling issues.

5.1. ITP and MMR vaccine

Miller et al. [19] studied the association between measles, mumps, rubella (MMR) vaccine and idiopathic thrombocytopenic purpura (ITP) (abnormal bleeding into the skin due to low blood platelet count) in children aged 12-23 months during the period from October 1991 to September 1994 within 42 days of receiving the vaccine. The data have since been updated, so the results differ slightly from those published. Six of the 35 children in our data set were admitted to hospital more than once during the observation period, five children were admitted twice ($i = 5, 9, 16, 17, 34$) and one child was admitted five times ($i = 23$). The data include a total of 44 admissions. The observation periods ran from age 366 to 730 days for all but two children whose observation periods started after day 366 ($i = 1, 7$). A further feature of this analysis is that it involves several age groups. Six age groups were used: 366-426 days, 427-487 days, 488-548 days, 549-609 days, 610-670 days and 671-730 days of age. Three two-week long risk groups were defined: 0-14 days after vaccination, 15-28 days after vaccination and

29-42 days after vaccination. Miller et al. [19] found a significant association between hospital admission for ITP and the risk period after receiving the MMR vaccine, though the risk is considerably less than that after natural measles or rubella.

5.2. Intussusception and oral polio vaccine

Andrews et al. [28] studied the association between oral polio vaccine (OPV) and intussusception (a dangerous condition where the bowel folds in on itself causing an obstruction of the intestine) in infants aged 28 to 365 days. Three data sets on intussusception were used: (1) hospital episode statistics (HES) data from January 1991 to March 1997, (2) HES data from March April 1997 to June 1999 and (3) all cases in the general practice research database data (GPRD), which started in 1987. In our examples we use only the first data set, in which there were a total of 218 episodes with 11 infants having more than one episode. Various combinations of two-week risk periods after taking the OPV were considered, including 0-13, 14-27 and 28-41 days after, and the two week period prior to taking the vaccine. The choice of risk periods will be discussed in subsequent sections. Eleven roughly one month long age groups were used for all analyses.

6. Modelling with the case series method

In this section we discuss various modelling issues, assumptions and limitations of the case series method. These are demonstrated using the three examples already introduced: MMR and viral meningitis, MMR and ITP, and OPV and intussusception. We also give details of how to fit the models in STATA.

6.1. Multiple risk periods

It is common to use more than one risk period. This might be to represent different biological effects, or a varying time since exposure effect. For example, Farrington et al. [4] used the two post-MMR risk periods 6-11 days and 15-35 days in their analysis of febrile convulsions. The rationale for this choice was that the 6-11 day risk period corresponded to increased risk from the measles component of the vaccine, as documented in many studies, while the 15-35 day period corresponded to increased risk from some types of mumps components. Multiple risk periods are also commonly used to differentiate between acute, non-acute and washout phases. For example Hubbard et al. [32] studied the risk of hip fracture after exposure to specific antidepressants. Two risk factors $k = 1$ and $k = 2$ were assigned during the 'high risk' period 0-14 and 15-42 days after the start of treatment. Then a risk period $k = 3$ was included for the remaining treatment period from day 43 of treatment until 31 days after the last antidepressant prescription (the exact date when treatment ended was not known). Two further risk periods $k = 4$ and $k = 5$ were also included corresponding to two 3-month wash-out periods.

We illustrate the use of multiple risk periods with the ITP and MMR data. As already mentioned, three two-week long risk groups were defined: day of vaccination to 14 days after vaccination ($k = 1$), and 15-28 ($k = 2$) and 29-42 ($k = 3$) days after vaccination. This is shown for the first three individuals in the data in figure 2 (age groups are not included in this diagram and individual 2 was vaccinated at age 868 days which is outside the observation period).

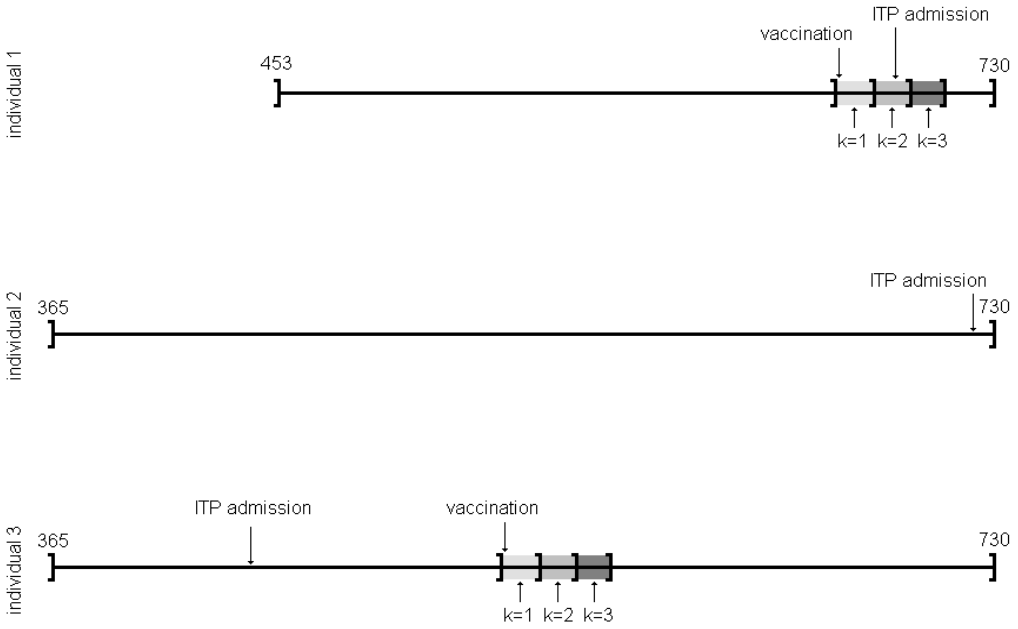

Figure 2. Diagram of the observation period for the first 3 individuals. Vaccine risk periods are shaded.

In STATA, after generating the cut points for the start (`cutp1`) and end (`cutp2`) of the observation period, and the five age group cut points (`cutp3-7`), four risk group cut points corresponding to the day before a new risk period starts or the day a risk period ends were generated (the day of vaccination is named `mmr`):

```
cutp8 = mmr - 1
cutp9 = mmr + 14
cutp10 = mmr + 28
cutp11 = mmr + 42.
```

The remaining STATA code does not need to be amended from that given for the study of MMR and meningitis in Oxford in section 4.2. The exposure group factors, `exgr`, are automatically generated as a list of increasing integers (starting at 0), except for the last interval which returns to level $k = 0$ (so that after the risk periods have ended the risk returns to the same 'unexposed' level as before exposure), for example 0,1,2,3,4,...,0.

Compared to the control period which included all days in the observation period before, and 43+ days after MMR vaccination, the relative incidences (RI) for the post-vaccination risk periods are given in table V. The risk of hospital admission for ITP was significantly raised 15 to 28 days after MMR vaccination.

Prepared using simauth.cls

Table V. Relative incidences for analyses of ITP and MMR

risk period: days after MMR	RI (95% CI)
0-14	1.31 (0.30, 5.73)
15-28	5.95 (2.52, 14.07)
29-42	2.60 (0.75, 9.07)

6.2. Modelling multiple events

This subsection describes how to analyse case series data in which one individual can experience more than one event during their observation period. This analysis assumes that events are independent within individuals: in other words, occurrence of one event does not alter the rate at which subsequent events might occur (the next subsection describes how to analyse data in which this assumption fails).

The ITP and MMR data were analysed under this within-individual independence assumption. There were several children who were admitted to hospital for ITP more than once during the observation period. Again data are analysed in STATA exactly as in the MMR and meningitis example. Now it should be clear why the data need to be reshaped listing all the `cutp` by each adverse event as well as by each individual (with only one event per person it is sufficient to list the `cutp` by individual alone). `nevents` is generated as before: 1 if an event occurred between the `cutp` listed and the `cutp` before (and is left as a missing data point otherwise). The `collapse` command now sums all the 1's in `nevents` over each individual and cut point `cutp` to give a total number of events between each cut point.

6.3. Unique and non-independent events

There are two common instances in which the Poisson assumption, under which the model was derived, fails. The first is when events are necessarily unique. However, the case series method is applicable for non-recurrent events in the limit as $\varphi_i \to -\infty$, where e^{φ_i} is the baseline rate for individual i (see Farrington [1] for details). Thus the case series method remains valid for rare non-recurrent events. Informally, the reason for this is that the times of first occurrence of a rare potentially recurrent event and the times of occurrence of a rare unique event cannot in practice be distinguished.

A second common failure of the Poisson assumption is when events are recurrent, but occurrence of one event increases the probability of subsequent events. There are two strategies for dealing with this. If events cluster in episodes, but episodes can be assumed independent, then the methods described in the previous subsection can be applied to the first event in each episode. For example, Farrington et al. (1995) grouped repeat hospital admissions within 72 hours for febrile convulsions into single episodes. If this is not appropriate but the initiating event is rare, then only the first event should be used, and subsequent events ignored.

For example, in the ITP data set, one individual had 5 events. This might suggest that events cluster within individuals. If this is the case, the Poisson assumption is inappropriate. An alternative analysis is to use only the 35 first events: this is valid since ITP is a relatively

uncommon condition. In STATA, second and subsequent events can be dropped by giving the following commands at the beginning of the do file:

```
sort indiv eventday
by indiv:drop if eventday[_n] > eventday[_n-1]
```

With reference to the baseline period which included all days before vaccination and 43+ days after vaccination the relative incidences for the risk periods are given in table VI. These relative incidences are greater than in the original analysis.

Table VI. Relative incidences for analyses of ITP and MMR

risk period: days after MMR	RI (95% CI)
0-14	1.59 (0.36, 7.15)
15-28	7.19 (2.92, 17.73)
29-42	3.22 (0.89, 11.59)

6.4. Event-dependent exposures

The most important, and possibly restrictive, assumption of the case series method is that occurrence of an event does not alter the probability of subsequent exposure. This assumption is required for the conditioning argument by which the case series likelihood is derived. The most extreme setting in which this assumption fails is when the event of interest is death: clearly, individuals cannot be exposed after death. Case series analyses of deaths are discussed in section 6.12. A less extreme example is myocardial infarction and nicotine patches: the pattern of use of nicotine patches is likely to be altered by the occurrence of a myocardial infarction.

Developing a general approach to the analysis of case series for event-dependent exposures is a topic of ongoing research. In this subsection we discuss two modifications of the standard method, which may be used in certain circumstances with event-dependent exposures.

Suppose first that the post-exposure risk period is not indefinite. For each individual, redefine the observation period as starting with the age at exposure, and ending as before at the end of observation. Then apply the standard case series method with these new observation periods. It follows that only individuals exposed prior to the event are included in the analysis, and that the within-individual comparison is between the immediate post-exposure period and later post-exposure periods.

A second approach applies to situations in which the dependency is short-lived, that is, post-event exposures are affected for a defined time period, but then return to normal. For example, a child who has suffered from intussusception is unlikely to be given an oral polio vaccine until he or she has fully recovered [28, 29]. Thus the exposure probability is reduced for a time after the event, and then can be assumed to return to normal.

In this second setting, it follows that events are unlikely to occur in the immediate pre-exposure period. Ignoring the effect would deplete the baseline incidence and hence exaggerate

the relative incidence. A simple way to correct for this is to use a pre-exposure risk period, of duration ϵ say, to remove this time from the baseline. This has been shown to be valid when exposures arise in a non-homogeneous Poisson process and the probability that two or more events occur in a period of length ϵ is negligible.

In general, there may be some doubt as to whether, and how, exposures are affected by prior events. In this situation it is sensible to undertake several contrasting analyses as described here. We illustrate this procedure with the data on oral polio vaccination and intussusception. Here we use only the first data set: hospital admissions data from January 1991 to March 1997 (this was an exploratory study: further analyses did not confirm the association reported here). After finding no increased risk in the period 0-13 days after analysis, Andrews et al. [28] restricted their analyses to the periods 14-27 and 28-41 days after receiving the oral polio vaccine (OPV), thus we use these two risk periods here. Also, for simplicity, we restrict the analysis to the 3rd dose only. Three analyses were carried out:

Analysis 1 Standard analysis with no allowance for dependence of exposures on previous events. To recap, the observation period runs from age 28 to 365 days inclusive (unless the record starts after or ends before these respective ages), there are 11 roughly 1-month age groups and three exposure groups including the baseline period and the two risk periods, 14-27 and 28-41 days after the 3rd vaccine dose.

Analysis 2 As analysis 1, but the observation period starts on the day dose 3 was given.

Analysis 3 As analysis 1, but including a 14-day long pre-3rd dose period.

The observation period for the first individual in each analysis is illustrated in figure 3. This individual received their 3rd vaccine dose at age 114 days and was admitted to hospital at age 156 days.

Relative incidences (RI) (with 95% confidence intervals (CI)) of risk periods and the 14-day pre-dose period compared with the baseline period are given in table VII.

Table VII. Relative incidences for analyses 1, 2 and 3 of intussusception and OPV.

risk period	analysis 1 RI (95% CI)	analysis 2 RI (95% CI)	analysis 3 RI (95% CI)
14-day pre-OPV period	-	-	0.18 (0.05, 0.76)
14-27 days after OPV	2.14 (1.32, 3.48)	2.24 (1.26, 4.00)	1.91 (1.17, 3.12)
28-41 days after OPV	1.24 (0.69, 2.25)	1.29 (0.69, 2.43)	1.16 (0.64, 2.09)

Analysis 2 gives similar results to analysis 1, though the confidence intervals are wider because cases with events prior to vaccination are excluded, thus reducing study power. In analysis 3 the 14-day pre-vaccination period has a significantly lower incidence of hospital admission for intussusception than the baseline period, suggesting that children are less likely to be given the polio vaccine until they are fully recovered. Ignoring this effect, as in analysis 1, biases the baseline risk downward and hence inflates the relative incidence. Analysis 3 corrects this, thus reducing the relative incidence in the risk periods after OPV.

Pre-exposure risk periods are coded in exactly the same way as other risk intervals, as described in sections 6.1 and 6.6.

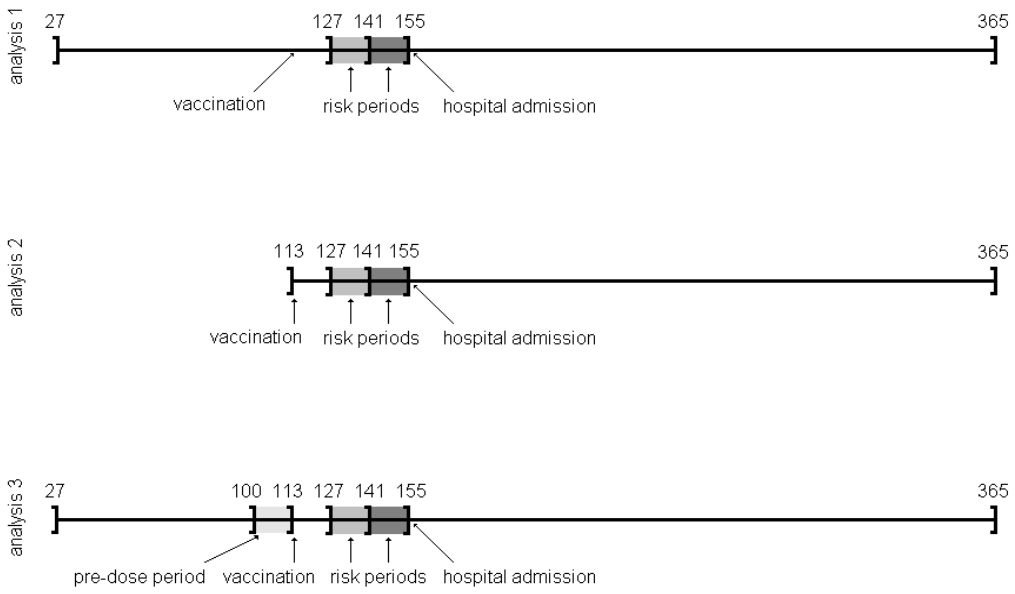

Figure 3. Diagram of the observation period for the first individuals for each analysis.

6.5. Covariates and interactions

A major attraction of the case series method is that it is self-controlled: estimation is within individuals, and hence the (multiplicative) effects of any fixed covariates cancel out. Thus individual-specific characteristics need not be included as main effects in the analysis (if they are, they are aliased as they are redundant). However, fixed covariates may act as effect modifiers: for example, the association between ITP and MMR vaccination might be sex-dependent. To investigate such effects, interactions between covariates, such as sex, and the exposure may be included in the analysis.

For example, to investigate the effect of sex on the association between OPV and intussusception, the STATA code is amended as follows. For simplicity we use analysis 1 as described in section 6.4 (event-dependent exposures). The variable for sex is called `sex`, and the sexes are coded 1 for males and 2 for females. The commands which need to be amended to incorporate an interaction with sex into an analysis are:

```
reshape long cutp, i(indiv eventday sex) j(type)
collapse (sum) nevents, by(indiv cutp type sex)
xi i.exgr*i.sex i.agegr
aglm nevents _Iexgr_* _IexgXsex_* _Iagegr_*,
```

```
offset(loginterval) family(poisson) irls eform.
```

The `xi` command creates four sets of indicator variables (sex `_Isex_1-2`, exposure group `_Iexgr_0-3`, sex × exposure group `_IsexXexg_1-2_0-3` and age group `_Iagegr_0-10`). We do this separately here because we do not want to fit the main effect for sex, only the interaction with the exposure groups. The star `*` is short for all variables beginning with the text given.

Relative incidences (RI) (with 95% confidence intervals (CI)) for analysis 1 and the same analysis including an interaction with sex are given in table VIII.

Table VIII. Relative incidences for analyses of intussusception and OPV with and without an interaction with sex.

risk period days after OPV	effect	analysis 1 (no interaction) RI (95% CI)	effect	analysis 1 with interaction RI (95% CI)
14-27	main	2.14 (1.32, 3.48)	main	2.45 (1.41, 4.26)
28-41	main	1.24 (0.69, 2.25)	main	0.69 (0.27, 1.74)
14-27	-	-	interaction	0.65 (0.24, 1.73)
28-41	-	-	interaction	3.24 (1.04, 10.12)
deviance (d.f.)		1038.36 (2582)		1032.85 (2580)

We performed a likelihood ratio test to test the hypothesis that no interaction term should be included in the model. The log likelihood ratio is equal to the difference in deviances, these are shown in table VIII, together with the degrees of freedom. The interaction between sex and the period 28-41 days after OPV is only marginally significant. The log likelihood ratio $\simeq$ 5.51 on 2 degrees of freedom, p=0.06. The Bayesian Information Criterion (BIC) reported by STATA favours the model without the interaction.

6.6. Repeat exposures

Individuals may experience repeat exposures, for example repeat prescriptions of the same drug. If the risk is expected to be the same for each exposure, the risk periods after each exposure can be given the same factor level. If the risk is expected to differ each time, different factor levels should be assigned to each risk period. The two approaches yield nested models and hence can be compared in the usual way, for example to evaluate the evidence for a cumulative dose effect.

When dealing with repeat exposures, care must be taken to avoid overlapping risk periods, else the software will not divide the data up correctly. A simple convention is that later exposures take precedence over earlier ones. An alternative is to model the first, second etc... exposures as distinct, as described in the next subsection, and investigate possible interactions in overlapping periods.

For example in our OPV and intussusception data 3 doses of OPV were administered. We define two risk periods 14-27 days and 28-41 days after each dose. There are two possible ways to analyse our data:

Analysis 4 Does not allow for a dose effect. We define three exposure factors: $k = 0$ the

baseline period, $k = 1$ 14-27 days after any OPV dose and $k = 2$ 28-41 days after any of the three OPV doses.

Analysis 5 Allows for a dose effect. We define seven separate exposure periods each with their own factor level: $k = 0$ baseline period, $k = 1$ 14-27 days after the first dose, $k = 2$ 28-41 days after the first dose, $k = 3$ 14-27 days after the second dose, $k = 4$ 28-41 days after the second dose, $k = 5$ 14-27 days after the third dose and finally $k = 6$ 28-41 days after the third dose.

Analyses 4 and 5 are illustrated for individual $i = 4$ in figure 4. This child received OPV doses 1, 2, and 3 at ages 98, 132 and 160 days respectively and was admitted to hospital twice at ages 107 and 197 days. If the data records show that a child did not receive a vaccine dose,

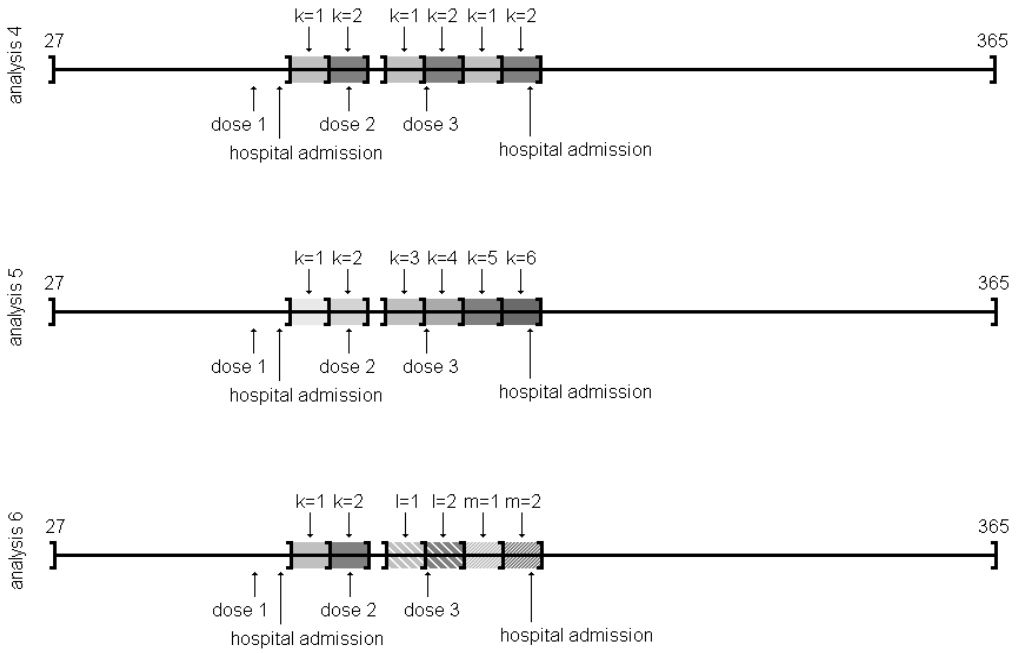

Figure 4. Diagram of the observation period for individual $i = 4$ for analyses 4, 5 and 6. Exposure risk periods are shaded.

the data should be amended to state that a dose was given at some time after the end of their observation period. To fit either of these models in STATA we begin by generating nine exposure group cut points (numbered `cutp13` to `cutp21`): 13, 27 and 41 days after each dose. To ensure risk periods do not overlap we include the following code after the cut points have been generated:

```
foreach i of numlist 13/20{
```


Prepared using simauth.cls

```
local j = `i'+1
replace cutp`i' = cutp`j' if cutp`i' > cutp`j'
},
```

which replaces a cut point with the next cut point if that cut point is greater than the next one (this assumes that later exposures take precedence over earlier ones). Generating nine exposure group cut points means that our basic STATA file gives us nine exposure periods: the periods between the risk periods for each dose are assigned a new factor level when these periods should belong to the baseline exposure group. The factor levels can be corrected by issuing a `recode` command after the exposure groups have been generated. For analysis 4:

```
recode exgr (0=0) (1=1) (2=2) (3=0) (4=1) (5=2) (6=0) (7=1) (8=2),
```

and for analysis 5:

```
recode exgr (0=0) (1=1) (2=2) (3=0) (4=3) (5=4) (6=0) (7=5) (8=6).
```

The two periods between the risk periods for the three doses were coded 3 and 6, and these are changed to 0, the factor for the control period.

The results of these two analyses are shown in Table IX. There was only one significant relative incidence 14-27 days after dose 3 in analysis 5. We can use a likelihood ratio test to

Table IX. Relative incidences for analyses of intussusception and OPV

dose	risk period: days after OPV	analysis 4 RI (95% CI)	analysis 5 RI (95% CI)
all	14-27	1.30 (0.84, 2.01)	-
all	28-41	1.01 (0.64, 1.58)	-
1	14-27	-	0.86 (0.38, 1.94)
1	28-41	-	0.40 (0.14, 1.10)
2	14-27	-	0.56 (0.24, 1.30)
2	28-41	-	1.20 (0.63, 2.28)
3	14-27	-	2.09 (1.25, 3.51)
3	28-41	-	1.23 (0.68, 2.25)
	deviance (d.f.)	1130.83 (3613)	1118.59 (3609)

test the hypothesis that there is no dose effect. The log likelihood ratio is approximately 12.24, which we compare to a χ^2 distribution with 4 degrees of freedom. The test rejects the hypothesis that there is no dose effect $p = 0.02$, thus favouring analysis 5. The Bayesian information criterion (BIC) also supports analysis 5 over analysis 4 (BIC analysis 4 = -28680.55, BIC analysis 5 = -28659.78).

Other methods for modelling repeat exposures could also be envisaged, including parametric dose-response models with number of previous doses or time since previous dose as a covariate. Such methods are not considered here.

6.7. Multiple exposures

The association between an event and several distinct exposures, for example, two different vaccines, both of which may cause the same type of reaction, can be investigated within a single case series analysis. The presence of an interaction between the two exposures can also be investigated. Suppose for example there are two exposures, A and B. To each exposure is associated a factor, possibly with several levels. The model is then

$$\log(\lambda_{ijkl}) = \phi_i + \alpha_j + \beta_k + \gamma_l$$

where β_k and γ_l represent the two exposure factors. The interaction model is

$$\log(\lambda_{ijkl}) = \phi_i + \alpha_j + \beta_k + \gamma_l + \delta_{kl}.$$

To fit this model, the observation periods must be segmented according to the two exposures.

As an example of multiple exposures we use the OPV and intussusception data, but this time we regard each of the three doses of OPV as a separate exposure, and call this analysis 6. We model

$$\log(\lambda_{ijkl}) = \phi_i + \alpha_j + \beta_k + \gamma_l + \delta_m$$

where β_k are the exposure factors for dose 1, γ_l for dose 2 and δ_m for dose 3. For the control periods $k = 0$, $l = 0$ and $m = 0$. Two risk factors are defined for each dose: 14-27 days after receiving OPV ($k = 1$ after dose 1, $l = 1$ after dose 2 and $m = 1$ after dose 3), and 28-41 days after OPV ($k = 2$ after dose 1, $l = 2$ after dose 2 and $m = 2$ after dose 3). Analysis 6 is illustrated alongside the two repeated exposure analyses (analyses 4 and 5) in figure 4. Note that, in this analysis, it no longer matters if the risk periods corresponding to different exposures (i.e. different doses) overlap.

In STATA the exposure group cut points are generated exactly as before, but it is important to note the numbers `type` associated with each `cutp`. The exposure group cut points for dose 1 are numbered `cutp13-15`, for dose 2 `cutp16-18` and for dose 3 `cutp19-21`. The process of generating the exposure groups `exgr` needs to be repeated 3 times, and separate names need to be given to the exposure groups for each separate dose. We have chosen `exgr1`, `exgr2` and `exgr3` for doses 1, 2 and 3 respectively. In the part of the STATA code that generates the exposure groups, other than the names for the exposure groups, only the first line needs to be amended for each dose. Change `type-‘nage’-3` to `type-`(the first `type` number for the relevant exposure group). Change `if type>‘nage’+2` to `if type>=`(the first `type` number for the exposure group) `& type<=`(the last `type` number for the exposure group). For this example:

```
generate exgr1 = type-13 if type>=13 & type<=15
generate exgr2 = type-16 if type>=16 & type<=18
generate exgr3 = type-19 if type>=19 & type<=21.
```

Then change the model to include the three exposure groups:

```
xi: aglm nevents i.exgr1 i.exgr2 i.exgr3 i.agegr,
       offset(loginterval) family(poisson) irls eform.
```

The results of this analysis are shown in table X. There is very little change from the results for the repeat exposures. This is not surprising, since in this example there is very little overlap between the exposure periods as the doses of OPV are generally spaced 28 days or more apart.

Table X. Relative incidences for analysis 6 of intussusception and OPV

dose	risk period	RI (95% CI)
1	14-27	0.86 (0.38, 1.93)
1	28-41	0.39 (0.14, 1.09)
2	14-27	0.56 (0.24, 1.30)
2	28-41	1.18 (0.62, 2.25)
3	14-27	2.09 (1.25, 3.49)
3	28-41	1.23 (0.67, 2.24)

6.8. Confounding between exposure and age

The inclusion of age effects in the model corrects for confounding by age. However, if all exposures occur at roughly the same age, such adjustments may not be sufficient: the confounding is inherent in the data. In this situation, no method of analysis can disentangle the separate effects of age and exposure unless a sub-group of cases are unexposed. Unexposed cases (that is, cases who do not experience exposure in the observation period) contribute information only on the age-specific relative incidence, and hence enable the exposure-specific relative incidence to be estimated. It is generally worth including unexposed cases in the analysis, provided that lack of exposure is correctly recorded, and is not simply the result of missing exposure data.

In one setting the case series method fails completely: this is when the age at event is determinate, so that there is no within-individual variation. If events were to happen at exactly the same age, then the method would fail since the conditional likelihood would trivially reduce to 1. The conditional likelihood is degenerate, all information on the association between exposure and event residing in the margins. Such a pathological situation, however, is unlikely ever to arise in practice.

6.9. Long and indefinite risk periods

The case series method works best for acute events and short risk periods (relative to the observation period). However, it can be used with non-acute events that may occur long after the exposure. For example, the case series method was used to investigate the risk of MMR with respect to autism, possibly long after vaccination [21], [22]. In these analyses longer and indefinite risk periods were used. This use of the case series method is much more prone to confounding between age and exposure effects than when risk periods are short. The confounding is eliminated by including unexposed individuals in the analysis, a result confirmed by extensive simulation studies.

Experience suggests that the results of such analyses can be sensitive to the choice of age groups. Recently, a semi-parametric case series method has been developed [18]. In this semi-parametric model, the age effect is modeled non-parametrically. When ages are recorded in discrete time units, the semi-parametric model is equivalent to a parametric model as described here, with age groups of one time unit. For example, if the data are recorded to the nearest day, the semi-parametric model is equivalent to a parametric model with a separate parameter for each day. A more economical fitting method than that requiring a separate age group for

each day is available from the self-controlled case series website.

6.10. Modelling age effects

The choice of age groups depends on the age-dependence of the outcome event of interest, and should aim to capture this baseline dependence. Within this general framework, the choice of age groups in a parametric case series analysis is to some extent arbitrary. It is generally a good idea to try narrower age groups to check that there is not substantial sensitivity to the choice of age boundaries.

Where it is reasonable to assume that age effects are constant throughout each individual's observation period it may not be necessary to split up the observation time into different age groups (but see the comment at the end of section 6.11). Such a situation is more likely to occur in studies including only adults when each individual's observation period is sufficiently short that age-related changes in baseline incidence can be ignored. Age at start of observation can then be treated as a fixed covariate, and interactions with the exposure can be investigated as described in section 6.5.

In our example of the meningitis cases in Oxford only two 6-month age groups were used. Nine of the ten cases occurred within the baseline age group ($j = 0$), and so the risk associated with the baseline group is high. The relative risk for the vaccine risk group ($k = 1$) versus the baseline group ($k = 0$ and $j = 0$) is 12.04 (3.00, 48.25) when there are 2 age groups, this increases to 12.32 (3.04,49.90) when there are 4 age groups, and respectively, 13.90 (3.17,59.09), 14.25 (2.83,71.16), 16.00 (2.83,90.48) when there are 8, 16 and 32 roughly evenly spaced age groups.

In the semi-parametric model, the choice of age groups is based entirely on the ages at which events occur. For the meningitis and MMR in Oxford data, the semi-parametric estimate is 40.29 (2.42, 669.63), the confidence interval around this estimate is very wide. The conclusion from the semi-parametric model is that there is a strong positive association, but there is insufficient information to estimate its magnitude with any precision.

6.11. Temporal effects

All analyses described so far take age as the underlying time line. In some settings, calendar time is the appropriate time line. For example, Kramarz et al. [25] and Tata et al. [26] studied the incidence of hospital visits for asthma after influenza vaccination. They used calendar time rather than age as the underlying time line, since influenza is highly seasonal. Note that if all exposures occur at the same calendar time, exposure and time effects will be confounded, as described for age effects in section 6.8.

In other settings, it is necessary to allow for both age and temporal effects. For example, in their analysis of oral polio vaccination and intussusception in Cuba, Galindo Sardiñas et al. [29] used age as the underlying time line, but also fitted a seasonal factor to remove the confounding effect of season, since both intussusception and polio vaccination are seasonal in Cuba. The model they fitted used an additional segmentation of the data by month of the year, thus taking the form

$$\log(\lambda_{ijkl}) = \phi_i + \alpha_j + \sigma_k + \beta_l$$

where σ_k is a seasonal factor with 12 levels. The interval lengths are of the form e_{ijkl}, indexed by season as well as age group.

Other time-varying covariates may be included in this way in the model. Generally, the case series method is insensitive to exponentially varying temporal covariates. In the semi-parametric model, such effects are adjusted for implicitly [18].

Failure to take account of temporal effects, whether related to age or calendar time, when such effects are present will generally produce biased estimates if exposure and event are age or calendar time dependent. We recommend that age effects are always included in a first model, if only to demonstrate they are not required. Inclusion of other temporal effects in addition to age is only necessary if these are non-exponential, as is the case with seasonal effects for example.

6.12. Case series analyses of deaths

Case series analyses of deaths present two challenges: exposures must usually precede death (there are some exceptions, such as environmental exposures [17]), and the observation period is censored. Both may be accommodated by taking the observation period as the time from exposure to the end of the *planned* observation period. Provided that the deaths of interest are rare, the case series method can then be applied. For multiple exposures the analysis is more tricky and will not be covered here.

7. Designing a case series study

7.1. Ascertainment of exposures and outcomes

As with any epidemiological design, it is essential that the ascertainment of events and exposures are independent. Thus haphazardly assembled case series are not suitable for case series analysis: cases must be sampled by some objective mechanism, within a defined sampling frame. Commonly used sampling frames include hospital admission lists, and databases such as the General Practice Research Database (GPRD) or The Health Improvement Network (THIN). Exposure information might be included on such databases, or might be obtained by record linkage of case data and exposure data.

7.2. Choice of observation period

The observation period is usually defined by the age and calendar time criteria for case selection. For example, 'all hospital admissions for (the condition) in persons aged (age group) between (date) and (date) inclusive' implicitly defines the observation period for all cases. It is important to be rigorous in the definition of the age and time boundaries.

Choosing the age and time boundaries depends on the condition and the exposure to be analyzed. Generally, the boundaries should be chosen so as to maximise the chance that a case experiences both risk and control periods. For example, in an analysis of acute reactions to MMR vaccination, it makes sense to use the second year of life (defined precisely as days 366 to 730 of life inclusive), since the recommended age for MMR vaccination is 15 months but is occasionally delayed a little. In studies based on pre-existing databases, the observation period may simply depend on how long the patient record lasts. The length and timing of the observation periods are likely to impact on the choice of age groups.

7.3. Choice of risk period

Risk periods should be chosen based on prior hypotheses, previous studies, and presumed biological mechanisms. Caution is required in specifying the risk period, since if it is too long, too short or is placed so that it does not cover the true risk period then the relative incidence estimate may be biased toward the null.

To avoid this it is good practice, especially when there is uncertainty about the true risk period, to use several contiguous periods. On the other hand, use of multiple periods can increase the type I error, and can introduce bias if the total duration of the risk periods becomes long compared to the observation period (see comments above on long risk periods).

7.4. Covariates

A major advantage of the case series method is that the analysis adjusts for fixed covariates. Thus collecting data on covariates is much less important for case series analyses than other epidemiological designs. However, if it is anticipated that some covariates might modify the association, then information on these should be collected and included in the model as interactions with the exposure effect.

7.5. Relative efficiency

Conditioning on the numbers of events results in some loss of efficiency of the case series method compared to the retrospective cohort method from which it is derived: specifically, the marginal information is lost. In some circumstances, however, this marginal information is negligible and hence the case series method retains high relative efficiency. For example if all individuals are exposed, or alternatively if the proportion of cases attributable to the exposure is small, which arises when the risk period is short in relation to the observation period, then the case series method retains high efficiency relative to the cohort method. An explicit expression for the asymptotic relative efficiency was derived by Farrington and Whitaker [18].

7.6. Sample size

The number of events required to detect a given relative incidence in general depends on age effects and the age-distribution of the exposure. The following expression for the sample size applies when age effects can be ignored.

Let β denote the true log relative incidence associated with exposure, r the ratio of the risk period to the observation period, and write

$$\pi = \frac{re^{\beta}}{re^{\beta} + 1 - r}.$$

Assume that a proportion p of individuals in the population are exposed during the observation period (note that p relates to the total population, not just cases), and let α denote the significance level and $1-\gamma$ the power required, $z_{\alpha/2}$ and z_{γ} the $100(1-\frac{\alpha}{2})$ and $100(1-\gamma)$ percentiles of the standard normal distribution. Then the number of events required to achieve the stated power to reject the null hypothesis $\beta = 0$ in a 2-tailed test at the stated significance level is

$$n = \frac{C}{A}\left(z_{\alpha/2} + z_{\gamma}\sqrt{B}\right)^2$$

where

$$\begin{aligned} A &= 2\left\{\pi\beta - \log(re^{\beta} + 1 - r)\right\}, \\ B &= \frac{\beta^2\pi(1-\pi)}{A}, \\ C &= 1 + \frac{1-p}{p(re^{\beta} + 1 - r)} \end{aligned}$$

This expression is based on the signed root likelihood ratio statistic, and can be generalised to allow for age effects. Musonda et al. [41] provide further details and a comparative evaluation of this and other sample size formulae.

8. Concluding comments

The case series method is a relatively recent addition to the panoply of study designs available to epidemiologists. We hope that this tutorial will help clarify its assumptions and limitations, indicate how to implement it in general purpose software packages, and illustrate the types of analyses that can be undertaken.

There are several areas of ongoing research into the case series method which we have barely touched upon in this tutorial: the analysis of event-dependent exposures, inference for small samples, power calculations in the presence of age-related variation, to name but three. There are plenty of other topics worthy of investigation, and we hope that this tutorial will encourage users to develop the method further.

Acknowledgements

We would like to thank Bobby Gutierrez, director of statistics at STATA corp, for showing us how to amend STATA's generalised linear model `glm` command to fit absorbing factors. Also Bill Gould at STATA corp., Stephen Evans for motivating us to fit the model in STATA, Mick Green for information about GLIM and absorbing factors, Liz Miller and Nick Andrews of the HPA for the data sets and two anonymous referees whose comments helped improve the paper.

REFERENCES

[1] . Farrington CP. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics* 1995; **51**:228-235.
[2] . Farrington CP, Nash J and Miller E. Case series analysis of adverse reactions to vaccines: a comparative evaluation. *American Journal of Epidemiology* 1996; **143**:1165-1173. Erratum 1998; **147**:93.
[3] . Miller E, Goldacre M, Pugh S, Colville A, Farrington CP, Flower A, Nash J, MacFarlane L and Tettmar R. Risk of aseptic meningitis after measles, mumps and rubellavaccine in UK children. *The Lancet* 1993; **341**:979-982.
[4] . Farrington CP, Pugh S, Colville A, Flower A, Nash J, Morgan-Capner P, Rush M and Miller E. A new method for active surveillance of adverse events from diphtheria tetanus pertussis and measles mumps rubella vaccines. *Lancet* 1995; **345**:567-569.

Prepared using simauth.cls

[5] . Aalen. OO, Borgan O, Keiding N and Thorman J. Interaction between life history events, Nonparametric analysis for prospective and retrospective data in the presence of censoring. *Scandinavian Journal of Statistics* 1980; **7**:161-171.
[6] . Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; **73**:1-11.
[7] . Maclure M. The case-crossover design: A method for studying transient effects on the risk of acute events. *American Journal of Epidemiology* 1991; **133**(2):144-153.
[8] . Greenland S. A unified approach to the analysis if case-distribution (case-only) studies. *Statistics in Medicine* 1996; **18**:1-15.
[9] . Maclure M and Mittleman MA. Should we use a case-crossover design? *Annual Review of Public Health* 2000; **21**:193-221.
[10]. Vines SK and Farrington CP. Within-subject exposure dependency in case-crossover studies. *Statistics in Medicine* 2001; **20**:3039-3049.
[11]. Feldmann U. Epidemiologic assessment of risks of adverse reactions associated with intermittent exposure. *Biometrics* 1993; **49**:419-428.
[12]. Feldmann U. Design and analysis of drug safety studies, with special reference to sporadic drug use and acute adverse reactions. *Journal of Clinical Epidemiology* 1993; **46**:237-244.
[13]. Andrews NJ. Statistical assessment of the association between vaccination and rare adverse events post licensure. *Vaccine* 2002; **20**:S49-S53.
[14]. Farrington CP. Control without separate controls: Evaluation of vaccine safety using case-only methods. *Vaccine* 2004; **22**:2064-2070.
[15]. Becker NG, Li Z and Kelman CW. The effect of transient exposures on the risk of an acute illness with low hazard rate. *Biostatistics* 2004; **5**(2):239-248.
[16]. Navidi W. Bidirectional case-crossover designs for exposures with time trends. *Biometrics* 1998; **54**(2):596-605.
[17]. Lumley T and Levy D. Bias in the case-crossover design: implications for studies of air pollution. *Environmetrics* 2000; **11**:689-704.
[18]. Farrington CP and Whitaker HJ. Semi-parametric analysis of case series data. Submitted November 2004.
[19]. 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.
[20]. Dourado I, Cunha S, Teixeira M de G, Farrington CP, Melo A, Lucena R, Barreto ML. An outbreak of aseptic meningitis associated with a Urabe-containing MMR mass vaccination campaign: implications for immunization programs. *American Journal of Epidemiology* 2000; **151**:524-530.
[21]. Taylor B, Miller E, Farrington CP, Petropoulos M-C, Favot-Mayaud I, Li J, Waight PA. Autism and measles, mumps and rubella vaccine: no epidemiological evidence for a causal association. *Lancet* 1999; **353**:2026-2029.
[22]. Farrington CP, Miller E and Taylor B. MMR and autism: further evidence against a causal association. *Vaccine* 2001; **19**:3632-3635.
[23]. Miller E, Andrews N, Waight P, Taylor B. Bacterial infections, immune overload, and MMR vaccine. *Archives of Disease in Childhood* 2003; **88**:222-223.
[24]. Miller E, Andrews N, Grant A, Stowe J and Toylor B. No evidence of an association between MMR vaccine and gait disturbance. *Archives of Disease in Childhood* 2005; **90**:292-296.
[25]. Kramarz P, DeStefano F, Garguillo PM, Davis RL, Chen RT, Mullooly JP, Black SB, Shinefield MD, Bohlke K, Ward JI and Marcy MS. Does influenza vaccination exacerbate asthma? *Archives of Family Medicine* 2000; **9**:617-623.
[26]. Tata LJ, West J, Harrison T, Farrington P, Smith C, Hubbard R. Does influenza vaccination increase consultations, corticosteroid prescriptions or exacerbations in people with asthma or chronic obstructive pulmonary disease? *Thorax* 2003; **58**:835-839.
[27]. Mutsch M, Zhou W, Rhodes P, Bopp M, Chen RT, Linder T, Spyr C and Steffen R. Use of the inactivated intranasal influenza vaccine and the risk of bell's palsy in Switzerland. *New England journal of medicine* 2004; **350**:896-903.
[28]. Andrews N, Miller E, Waight P, Farrington CP, Crowcroft N, Stowe J and Taylor B. Does oral polio vaccine cause intussusception in infants? Evidence from a sequence of three self-controlled case series studies in the United Kingdom. *European Journal of Epidemiology* 2001; **17**:701-706.
[29]. Galindo Sardiñas MA, Zambrano Cárdenas A, Coutin Marie G, Santin Peña M, Aliño Santiago M, Valcárcel Sanchez M, and Farrington CP. Lack of association between intussusception and oral polio vaccine in Cuban children. *European Journal of Epidemiology* 2001; **17**:783-787.
[30]. Murphy TV, Garguillo PM, Massoudi MS, Nelson DB, Jumaan AO, Okoro CA, Zanardi LR, Setia S, Fair E, LeBaron CW, Wharton M and Livingood JR. Intussusception among infants given an oral rotavirus vaccine. *New England Journal of Medicine* 2001; **344**:564-572.
[31]. Mullooly JP, Pearson J, Drew L, Schuler R, Maher J, Gargiullo P, De Stefano F, Chen R and the vaccine

safety datalink working group. Wheezing lower respiratory disease and vaccination of full-term infants. *Pharmacoepidemiology and Drug Safety* 2002; **11**:21-30.
[32]. Hubbard R, Farrington P, Smith C, Smeeth L and Tattersfield A. Exposure to tricyclic and selective serotonin reuptake inhibitor antidepressants and the risk of hip fracture. *American Journal of Epidemiology* 2003; **158**:77-84.
[33]. Tata LJ, West J, Smith C, Farrington P, Card T, Smeeth L, Hubbard R. General population based study of the impact of tricyclic and selective serotonin reuptake inhibitor antodepressants on the risk of acture myocardial infarction. *Heart* 2005; **91**:465-471.
[34]. France EK, Glanz JM, Xu S, Davis RL, Black SB, Shinefield HR, Zangwill KM, Marcy SM, Mullooly JP, Jackson LA and Chen R. Safety of the trivalent inactivated influenza vaccine among children - a population based study. *Archives of Pediatrics and Adolecent Medicine* 2004; **158**:1031-1036.
[35]. Smeeth L, Thomas SL, Hall AJ, Hubbard R, Farrington P and Vallance P. Risk of myocardial infarction and stroke after acute infection or vaccination. New England Journal of Medicine 2004; **351**:2611-2618.
[36]. StataCorp. Stata Statistical Software: Release 8.0. College Station, TX: Stata Corporation. 2003.
[37]. SAS 8. SAS Institute Inc. Cary, NC, USA. 1999.
[38]. GenStat release 7.1. VSN international Ltd. Oxford, UK. 2003.
[39]. GLIM release 4. Royal Statistical Society, London, UK. 1992.
[40]. Payne RW (ed). *The guide to GenStat release 7.1. Part 2: Statistics.* VSN International: Oxford, UK, 2003; 555-556.
[41]. Musonda P, Farrington CP and Whitaker HJ. Sample size formulae for the self-controlled case series method. In preparation.