The case-cohort design in outbreak investigations

O Le Polain de Waroux (olivier.lepolain@hpa.org.uk)[1,2], H Maguire[1,2], A Moren[3]
1. London Region Epidemiology Unit, Health Protection Agency (HPA), London, United Kingdom
2. European Programme for Intervention Epidemiology Training (EPIET), European Centre for Disease Control and Prevention (ECDC), Stockholm, Sweden
3. EpiConcept, Paris, France

Citation style for this article:
Le Polain de Waroux O, Maguire H, Moren A. The case-cohort design in outbreak investigations. Euro Surveill. 2012;17(25):pii=20202. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20202

Article submitted on 25 November 2011 / published on 21 June 2012

The use of the case–cohort design for outbreak investigations has been limited. Here we discuss its strengths and limitations based on real and fictitious examples. The case–cohort is a case–control study where controls are sampled from the initial population at risk, and may thus include both cases and non-cases. An advantage of the design, compared to traditional case–control studies, is that risk ratios can easily be obtained directly from the cross-product of exposed and unexposed cases and controls (rare disease assumption is not required). We illustrate this in the context of point source gastrointestinal outbreaks and in field studies on vaccine effectiveness. The design is also useful to investigate multiple outcomes with a unique sample of controls or to test hypotheses when different case-definitions (from the most sensitive to the most specific) are used for a particular outcome. Strengths and limitations are presented, and discussed in the context of outbreak investigations.

Introduction

Outbreaks are defined as any excess in the number of cases of disease that would normally be expected in a particular geographic area over a particular period of time [1]. Outbreak investigations differ from standard epidemiological research as they are often conducted under time and resource constraints. Design options mostly depend on the outbreak setting, the size of the outbreak and of the population affected, whether or not the affected population is well defined and its members identifiable, and what measure of association is desired. In addition, the approach may vary depending on the pathogen (or the environmental hazard) and its mode of transmission, as well as time, staff and resource constraints.

The two main study designs generally considered by field epidemiologists in the investigation of outbreaks are the retrospective cohort design and the traditional case–control design. In the retrospective cohort, all members of a defined cohort are included in the study and information on their exposure to different factors is investigated retrospectively [2]. Risk of illness in exposed and unexposed individuals is obtained and the measure of association is the risk ratio (RR). Traditional case–control designs (also called 'cumulative' or 'classic' case–control) offer an efficient alternative when the source population (i.e. the population from which cases arose) is large and/or the outcome rare. Exposures in cases are compared to exposures in a sample of the non-cases (i.e. the controls) drawn from the same at-risk population, and the most common measure of association is the odds ratio (OR).

The case–cohort design is an alternative to the traditional case–control design. In the case-cohort design controls are randomly sampled from the source population, regardless of their disease status.

Although the case–cohort design has gained popularity in large prospective studies [3], its use in outbreak investigations has been limited [4,5]. There is, to the best of our knowledge, no publication that explains, summarises and discusses the use of case–cohort designs in the context of outbreak investigations.

In that context, the aim of this paper is therefore to summarise the theory of the case–cohort design, illustrate its use in four different outbreak scenarios and discuss its strengths and limitations.

Description of the case–cohort design

The foundation of case–cohort design is generally attributed to Prentice who, in 1986, described it as an efficient alternative to a full cohort design in the context of prospective research when the collection and follow-up of covariate information in each cohort member is costly and time-consuming [6]. Similar approaches were suggested by others under the terminology 'hybrid epidemiologic design', 'case–base' design or 'inclusive' case–control design [7-9].

Sampling and sample size

In a case–cohort design, all cases (or a random sample of all cases) and a random sample of the source population (i.e. the controls) are included in the study. The controls may therefore include some of the cases included in the case group [10].

Figure 1 illustrates the sampling of cases and controls in the case–cohort study design, and compares this with the traditional case–control and retrospective cohort designs.

The sampling strategies for controls include the whole range of probabilistic sampling methods used in cross-sectional studies, also including complex sampling designs. There are not many examples, but one is in a case-cohort study during and outbreak in Darfur, Sudan that used complex sampling to recruit controls [4].

Generally a little less statistical power is achieved with a case–cohort study, compared to a traditional case–control study, if both have an equal number of controls, inversely proportional to the primary attack rate (AR). A simple way of estimating the number of controls required for a defined power is to apply sample size calculations used in traditional case–control studies and multiply the number of controls by a weighting factor corresponding to the inverse of the proportion of non-cases in the initial cohort. For example, if the AR is 33% then 50% more controls (as $(1\text{-}0.33)^{-1}=1.5$) will have to be selected than in a traditional case–control study, whereas if the AR is only 5% the number of controls will only need to be increased by 5% (as $(1\text{-}0.05)^{-1}=1.05$). In some situations however, the AR will not be known at the start of the investigation.

Measure of association and analysis

Provided that cases are a random sample of all cases and the controls are sampled randomly from the source population, the cross product of exposed and unexposed cases and controls will yield a true estimate of the crude RR (allowing for sampling error), unlike the traditional case–control study where the OR obtained from the cross product of exposed and unexposed cases and controls will generally overestimate the RR (if true RR>1) or underestimate the true RR (if true RR<1). This inflation – or deflation – of the OR in case–control studies increases as the AR increases – or decreases – and also depends on the magnitude of the true RR (Figure 2).

Standard logistic regression can be used for multivariable analysis, in the same way as in a traditional case–control study, to obtain direct estimates of the adjusted RRs from the model output. This approach, taken in previous case–cohort studies [4,5], is limited however by the lack of precision around the estimates,

Figure 1

Comparing three study designs: case–cohort, case–control and retrospective cohort

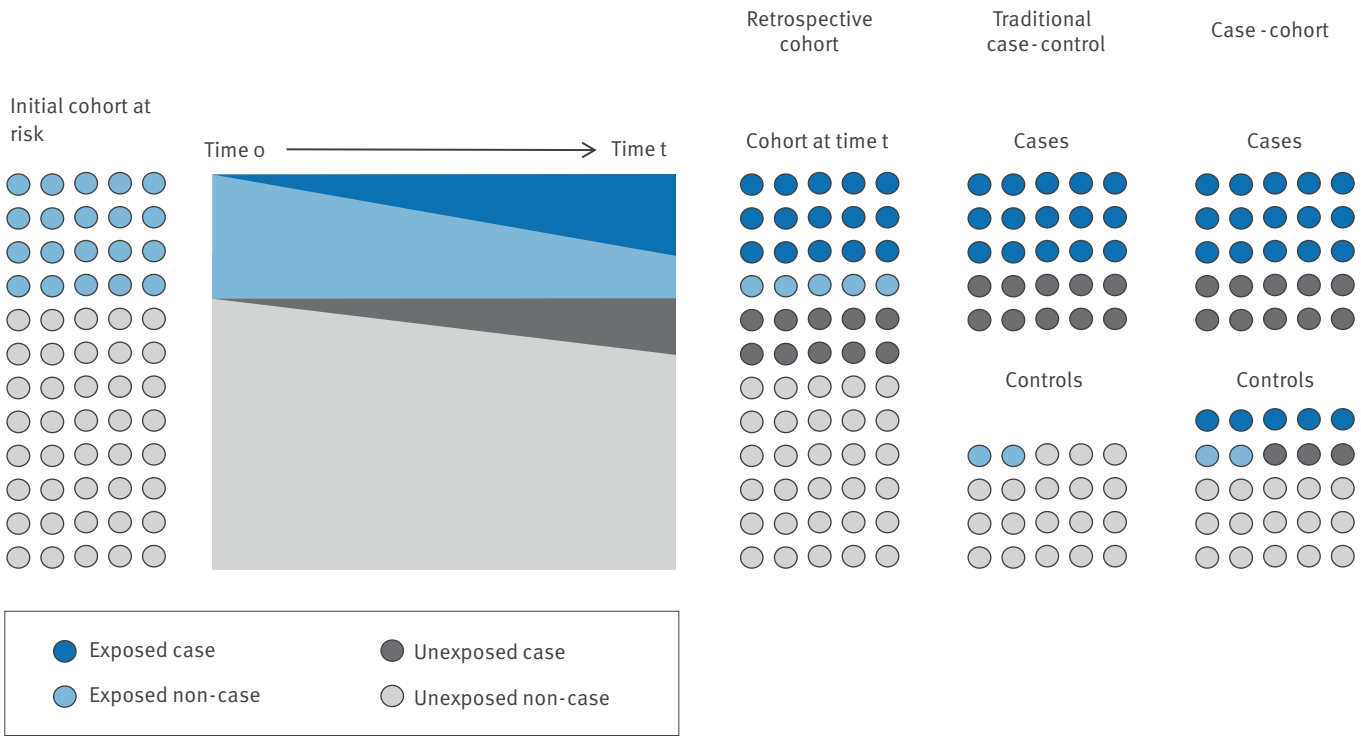
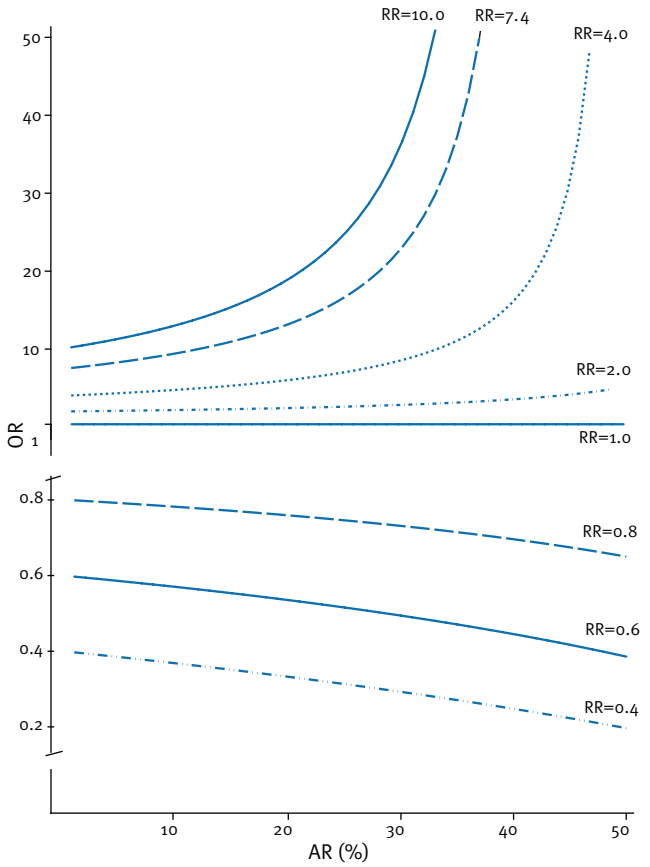


Figure 2

The relationship between odds ratio and risk ratio, for increasing attack rates



AR: attack rate; OR: odds ratio; RR: risk ratio

with standard errors generally being equal or larger than the true standard errors [10]. This may not be a major constraint when a strong association is found for a particular exposure variable; however in situations where weak evidence of an association is found, this should be taken into account in the interpretation of the results. Several solutions have been proposed to deal with this [8,11,12]. Schouten et al. [11] developed pseudo-likelihood risk models using logistic regression with a so called 'sandwich estimator' (or robust variance estimator) derived from the covariate matrix of the model output. Logistic regression is applied in traditional case–control studies, but RRs are obtained directly from the model output. The sandwich estimator adjusts the standard errors of the RR. This approach only requires common statistical software but may be more challenging if software commands are not readily available.

Outbreak scenarios

We will illustrate the case–cohort design through four commonly encountered outbreak scenarios, and discuss its strengths and limitations compared with traditional case–control and retrospective cohort designs. Examples are either based on published outbreak investigations, or, if no such outbreak investigation was published, are fictitious for illustrational purposes. We chose examples which illustrate the design well and cover different types of scenarios where the case–cohort design might be considered.

Scenario 1: A point-source outbreak in a closed setting

Outbreaks occurring in closed settings, such as schools, cruise ships or parties are common. To illustrate design options in this context, let us imagine a Salmonella outbreak following a party attended by 400 people, of whom 100 developed symptoms of diarrhoea within two days following the event and were defined as primary cases (AR: 25%).

If contact details of all participants can be obtained, the first choice would be to conduct a retrospective cohort design. Let us assume that in a retrospective cohort study a particular food item (food x) emerged as the most important risk factor in a univariable analysis, with 80 ill with diarrhoea amongst the 140 exposed (AR: 57.1%) and 20 ill amongst the 260 unexposed (AR: 7.4%), giving a crude risk ratio of 7.4 (95% confidence interval (CI): 4.8–11.4).

Under time and resource constraints of outbreak investigations, there is often a need to collect data on smaller sample sizes, and the use of traditional case–control or case–cohort studies (in this case, nested in the cohort) could be envisaged.

The Table compares the results of the univariable analysis for food x obtained with a retrospective cohort to those obtained in a traditional case–control study and in a case–cohort study, in which the sample size would be half that of the cohort. The true RR is obtained in the case–cohort study, whereas, in the traditional case–control studies, the OR does not approximate the true RR because the overall primary AR is high (25%), as shown in Figure 2.

Table. Comparing measures of association in the retrospective cohort, case–cohort and traditional case–control studies

Attributable risk fractions (the proportion of cases explained by the association=(RR-1)/RR) can also be calculated easily with case–cohort studies.

Arguably, in most outbreaks such as food-borne outbreaks the exact quantification of the risk increase associated with a particular factor may be unimportant

as long as there is evidence that that particular factor is associated with an increased risk, and over- or underestimating the RR may not matter that much.

Scenario 2: A vaccine effectiveness study during an outbreak

Outbreaks provide good opportunities to measure vaccine effectiveness (VE). Traditional case–control studies are often conducted when the source population is too large to conduct a retrospective cohort design [13].

In these studies VE is calculated as 1-OR of being vaccinated, where OR is assumed to approximate the RR. However, in studies of VE it is important to accurately obtain a precise estimate of the true RR. The use of the traditional case–control study in that context should therefore be discouraged given the difficulties in interpreting the OR [14].

The case–cohort design offers a suitable alternative in the context of VE studies during outbreaks, as it allows (i) to obtain true estimates of the RR and (ii) to randomly sample controls from the population without the need to enquire about disease history.

During a mumps outbreak in Switzerland, Richard et al. [15] investigated and compared the VE of two mumps vaccines. Cases were obtained from outbreak and surveillance data and controls were selected from a random systematic sample of GP registers, regardless of children's disease status. Similarly, Carrat et al. [16] used a case–cohort design to investigate influenza VE. Vaccination status in cases of confirmed influenza was compared to the vaccination status in controls randomly selected from GP registers, irrespective of whether or not they suffered from ILI during the influenza epidemic period. The design was particularly useful to obtain true estimates of the RR, and thus the VE, given the high incidence rate of ILI and influenza in the population.

Scenario 3: A food-borne outbreak at a restaurant

Food-borne outbreaks linked to restaurants are common. The use of a retrospective cohort design in restaurant outbreaks is often limited by the lack of identifiable controls, either because the guests' details have not been recorded or because the restaurant management may refuse to release details on their customers [17]. Traditional case–control studies are therefore often seen as the only available option, in which controls are a convenient sample selected from the non-ill meal companions of cases [17-19]. There may be few of these unaffected individuals, or they may not represent the average meal consumption of the customers as they tend to be more similar to the cases with regard to their meal consumption. In a situation where non-ill meal companions were scarce, Giraudon et al. [19] instead used a case–case approach in their investigation of a Salmonella PT1 outbreak linked to a fast-food restaurant in London. They compared consumption in mild cases to that reported by severe cases assuming an exposure dose–response effect.

We suggest that in food-borne outbreaks linked to restaurants, where no customers' list is available, a case–cohort design could be performed, in which meal consumption in the cohort of customers (e.g. based on receipts or any other type of restaurant record) would be compared to meal consumption in cases. Limitations with this approach include the lack of adjustment for the possible confounding effects age and sex, and the assumption that all food and drinks served were consumed. Its advantage is a rapid test of hypotheses, with no need of selection and interviewing controls. This can be particularly useful during ongoing outbreaks where speed is crucial.

Scenario 4: Investigating multiple outcomes

The opportunity to study multiple outcomes is particularly helpful in outbreak situations because, unlike in standard epidemiological research, case definitions are often dynamic. Generally, the case definition is initially broad (sensitive) and is narrowed down (more specific) as more information is gathered (e.g. laboratory confirmation).

With case–cohort studies, hypotheses can be tested with different sets of cases (e.g. from the most sensitive to the most specific case definition) using only one sample of controls.

Moreover, in situations where several outbreaks occur at the same time, especially outbreaks linked to similar

Table

Comparing measures of association in the retrospective cohort, case–cohort and traditional case–control studies

Type of design	Sample size	Number of cases	Type of measure of association	OR or RR (95%CI)
Retrospective cohort	400	100	RR	7.4 (4.8–11.6)
Traditional case–control	200	100	OR	16.0 (8.0–32.0)
Case–cohort	200	100	RR	7.4 (4.2–13.1)[a]

OR: odds ratio; RR: risk ratio.

[a] Variance derived from a first-order Taylor series approximation.

risk factors, the case–cohort design allows for one single control group to be used as reference group to investigate multiple outcomes.

For example, Martin et al. [5] used a case–cohort study to investigate a *Campylobacter* outbreak in the municipality of Söderhamn, Sweden, linked to the consumption of communal water. Although the number of confirmed campylobacteriosis cases was small (n=101) in comparison to the population of Söderhamn (n=27,765), the use of a traditional case–control study was complicated by the fact that another large outbreak of acute gastrointestinal illness (initially thought to affect more than 20% of the residents) occurred simultaneously, possibly including some unconfirmed cases of campylobacteriosis and possibly linked to the same source. A case–cohort study was conducted, and the control group was a simple random sample of the community, thus including some individuals with gastrointestinal illness. The investigation found that consuming communal water increased the risk of both campylobacteriosis and acute gastrointestinal illness, and the risk increased with the amount of water consumed.

Conclusions

We have described the use of the case–cohort design in field epidemiology, and illustrated its strengths and weaknesses through examples.

Among the advantages we identified is that a true estimate of the RR is possible. Although the OR may be good enough in most outbreak situations, there are situations (in particular VE studies) where obtaining a precise estimate of the true RR is important.

Further, the control group represents a random sample of the source population, and detailed disease history is therefore not required. This is particularly advantageous when cases and controls are sampled from different source databases, for instance a surveillance database for cases and a GP practice register for controls.

In addition, the control group can easily be used as a reference group to investigate multiple outcomes.

There are also a few limitations such as reduced statistical power compared with a traditional case–control study and the few analytical challenges, which can be addressed, but need more statistical expertise than a traditional case–control design.

Acknowledgments

The authors are grateful to Marta Valenciano and Ioannis Karagiannis for their comments on an earlier draft. Many thanks also to Sheila O'Malley for her help in retrieving articles.

References

1. Heyman DL, editor. Control of Communicable Disease Manual. 19th ed. Washington, DC: American Public Health Association; 2008.
2. Dwyer DM, Strickler H, Goodman RA, Armenian HK. Use of case-control studies in outbreak investigations. Epidemiol Rev. 1994;16(1):109-23.
3. Kulathinal S, Karvanen J, Saarela O, Kuulasmaa K. Case-cohort design in practice - experiences from the MORGAM Project. Epidemiol Perspect Innov. 2007;4(1):15.
4. Guthmann JP, Klovstad H, Boccia D, Hamid N, Pinoges L, Nizou JY, et al. A large outbreak of hepatitis E among a displaced population in Darfur, Sudan, 2004: the role of water treatment methods. Clin Infect Dis. 2006;42(12):1685-91.
5. Martin S, Penttinen P, Hedin G, Ljungstrom M, Allestam G, Andersson Y, et al. A case-cohort study to investigate concomitant waterborne outbreaks of Campylobacter and gastroenteritis in Soderhamn, Sweden, 2002-3. J Water Health. 2006;4(4):417-24.
6. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986;73:1-11.
7. Kupper LL, McMichael AJ, Spritas R. A hybrid epidemiologic study design useful in estimating relative risk. J Am Stat Assoc. 1975;70(351):524-8.
8. Miettinen O. Design options in epidemiologic research. An update. Scand J Work Environ Health. 1982;8(Suppl 1):7-14.
9. Rodrigues L, Kirkwood BR. Case-control designs in the study of common diseases: updates on the demise of the rare disease assumption and the choice of sampling scheme for controls. Int J Epidemiol. 1990; 19(1):205-13.
10. Rothman KJ, Greenland S, Lash TJ. Case-control Studies. In: Rothman KJ, Greenland S, editors. Modern Epidemiology. Philadelphia: Lippincott Williams and Wilkins; 2008: p. 111-27.
11. Schouten EG, Dekker JM, Kok FJ, Le Cessie S, Van Houwelingen HC, Pool J, et al. Risk ratio and rate ratio estimation in case-cohort designs: hypertension and cardiovascular mortality. Stat Med. 1993;12(18):1733-45.
12. Onland-Moret NC, van der A DL, van der Schouw YT, Buschers W, Elias SG, van Gils CH et al. Analysis of case-cohort data: a comparison of different methods. J Clin Epidemiol. 2007;60(4):350-5.
13. Goodson JL, Perry RT, Mach O, Manyanga D, Luman ET, Kitambi M, et al. Measles outbreak in Tanzania, 2006-2007. Vaccine. 2010;28(37):5979-85.
14. Moulton LH, Wolff MC, Brenneman G, Santosham M. Case-cohort analysis of case-coverage studies of vaccine effectiveness. Am J Epidemiol. 1995;142(9):1000-6.
15. Richard JL, Zwahlen M, Feuz M, Matter HC, Swiss Sentinel Surveillance Network. Comparison of the effectiveness of two mumps vaccines during an outbreak in Switzerland in 1999 and 2000: a case-cohort study. Eur J Epidemiol. 2003;18(6):569-77.
16. Carrat F, Tachet A, Rouzioux C, Housset B, Valleron AJ. Field investigation of influenza vaccine effectiveness on morbidity. Vaccine. 1998;16(9-10):893-8.
17. Baker K, Morris J, McCarthy N, Saldana L, Lowther J, Collinson A, et al. An outbreak of norovirus infection linked to oyster consumption at a UK restaurant, February 2010. J Public Health (Oxf). 2011;33(2):205-11.
18. Barton BC, Mody RK, Jungk J, Gaul L, Redd JT, Chen S, et al. 2008 outbreak of Salmonella Saintpaul infections associated with raw produce. N Engl J Med. 2011;364(10):918-27.
19. Giraudon I, Cathcart S, Blomqvist S, Littleton A, Surman-Lee S, Mifsud A, et al. Large outbreak of salmonella phage type 1 infection with high infection rate and severe illness associated with fast food premises. Public Health. 2009;123(6):444-7.