International Journal of Epidemiology
© International Epidemiological Association 1990

Vol. 19, No. 1
Printed in Great Britain

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

LAURA RODRIGUES AND BETTY R KIRKWOOD

Rodrigues L (Department of Epidemiology and Population Sciences, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK) and 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. *International Journal of Epidemiology* 1990, **19**: 205–213.
In recent years the use of case-control designs has been extended to the study of common diseases. It has been shown that the rare disease assumption is not necessary, and that by a suitable choice of sampling scheme for controls, it is possible to obtain direct estimates of relative risk and relative rate, instead of relying on the odds ratio as an indirect estimate. The majority of papers addressing these issues are theoretical, and the arguments have been couched in mathematical terms. As such they are not readily accessible to many practising epidemiologists. This paper summarizes the discussion in a simplified manner. It describes the three different measures of relative incidence, namely the relative risk, the relative rate and the odds ratio, together with their corresponding case-control designs.

The discussion is extended to show that the choice of the appropriate measure of relative incidence depends on the mode of action of the risk factor, as well as on characteristics of disease. We propose a classification scheme comprising five different categories of situation, and make recommendations regarding study designs for each.

Case-control methodology was originally developed in the context of studying the causality of non-infectious diseases. This design is particularly efficient for 'rare' diseases, since it requires only a fraction of those escaping disease to be studied, and for diseases with long latent or incubation periods, since it is retrospective in nature. Case-control methods have also been applied in the assessment of the efficacy of health actions when, for ethical or logistic reasons, randomized controlled trials are not possible. Examples include the assessment of efficacy of vaccines during outbreaks,[1,2] of Pap's smears in preventing invasive cancer of the cervix through early detection,[3] of post-menopausal oestrogen in preventing fractures of the hip,[4,5] of aspirin in reducing the risk of myocardial infarction,[6] of the risk of brain damage related to different interventions during delivery,[7] of the risk of neurological disease as a complication of pertussis vaccine,[8] and of the efficacy of BCG vaccination against leprosy[9] and against tuberculosis.[10]

More recently, the use of case-control methods has been extended to the field of common diseases. In this situation the case-control design is no longer clearly more efficient as regards sample size. It does, however, (along with retrospective cohort studies) avoid many of the ethical issues inherent in longitudinal and interventional studies, since the disease status of the study individuals is already determined. In addition a case-control design has the logistic benefits that it may be based entirely in health facilities and that it is relatively quick and easy to carry out. For these reasons, Smith *et al*[11] explored the use of case-control methodology for the measurement of the efficacy of measles immunization, and Briscoe *et al*[12] advocated its use in the evaluation of the health impact of improved water supply and sanitation facilities on diarrhoeal diseases.

Initially, case-control studies were analysed by testing for significant differences between the proportions exposed in cases and controls (e.g. 'Do more lung

Department of Epidemiology and Population Sciences, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK.

cancer patients have a history of smoking than controls?'), without attempting to estimate the size of the risk associated with exposure (e.g. 'By how many times does smoking increase the risk of lung cancer?'). In 1951, Cornfield[13] pointed out for the first time that it is possible to estimate the relative risk associated with exposure, using the ratio of the odds of exposure of cases to the odds of exposure of controls, provided that the disease in question is rare.

The necessity for this 'rare disease assumption' remained unchallenged until 1976 when Miettinen[14] argued that the rationale was only applicable to case-control studies 'in which the subjects are ascertained after the end of the entire risk period of interest' and not for the 'ordinary type of case-control study in chronic disease epidemiology', namely one in which incident cases and controls are recruited concurrently. He showed that in the latter it is possible to obtain an estimate of the relative rate (alternatively called the incidence density ratio), and that no rarity assumption is needed for this. These ideas were explored more fully by Greenland and Thomas[15] in 1982, and in 1984 Smith *et al*[11] showed that it is also possible to design a case-control study to yield a direct estimate of relative risk, and again that no rarity assumption is needed for this.

The majority of papers in this area are theoretical and the arguments have been couched in mathematical terms. As such they are not readily accessible to many practising epidemiologists. Our aim is to summarize and present the discussion in a simplified manner, with a particular focus on the use of case-control methods in the study of common diseases. We will explain the concepts of relative risk (alternatively called cumulative incidence ratio), relative rate (alternatively called incidence density ratio) and odds ratio, and show that each of these three measures can be estimated directly in a case-control study, by a suitable choice of sampling scheme for controls, even when the disease in question is common. We will include a discussion of situations in which subjects selected as controls can/should also be selected as cases, and *vice versa.* To clarify this, we suggest a classification into five types of situation, based on the characteristics of the disease and of the risk factor; and describe the appropriate sampling scheme for each. The classification is based on whether the disease is rare or common; whether or not there is recovery; whether recovery is accompanied by lifelong immunity; and on the mode of operation of the risk factor. In this paper we assume throughout a fixed population and do not explore the impact of a dynamic population, with influx of new people and emigration or death, of change of an individual's exposure status over time, of duration of the latent period or of differing duration of disease. These will be the subject of a later review.

THE DIFFERENT MEASURES OF RELATIVE INCIDENCE

We will start by describing the measurement of relative incidence in terms of a cohort study, and then discuss how this applies in the context of case-control designs. A typical cohort study is illustrated in Figure 1. This shows the gradual accumulation of cases during the study period (time 0 to T) among people initially disease-free, in an exposed population (E) and in an unexposed population (U).

The size of the association between exposure to the risk factor and the subsequent development of disease is measured by the ratio of the incidence among those exposed to that and among those who are not exposed. There are two conceptually different ways of defining incidence; it may be measured either as a risk or as a rate. Although the distinction between the two was clearly described by William Farr in the early nineteenth century, it is only in recent years that it has again attracted attention[16,17] and current terminology remains muddled.

The *incidence risk* (alternatively called cumulative incidence) is the probability that a person initially free from the disease develops it at some time during the period of observation. It is calculated as the number of cases divided by the population initially at risk and is usually expressd as a percentage or, if small, as per 1000 people. The incidence risk therefore equals C_E/N_E for the exposed group and C_U/N_U for the unexposed. The ratio of the two is called the relative risk (or cumulative incidence ratio), and is shown in Table 1.

The *incidence rate* (alternatively called incidence density), on the other hand, is the rate of contracting the disease among those still at risk; when a person contracts the disease they are no longer at risk. The number of new cases is related not to the number initially at risk but to the sum of the lengths of time each person stayed at risk during the period of observation. This sum is called the number of *person years at risk* (pyar) and is equivalent to the *average* number at risk during the period, multiplied by the length of the observation period. The shaded areas in Figure 1 represent the number of person years at risk for exposed and unexposed populations. These areas are equal in size to the rectangles defined by the average numbers at risk in the two groups and the length of the observation period. The incidence rate is often multiplied by 1000 and expressed per 1000 person years at

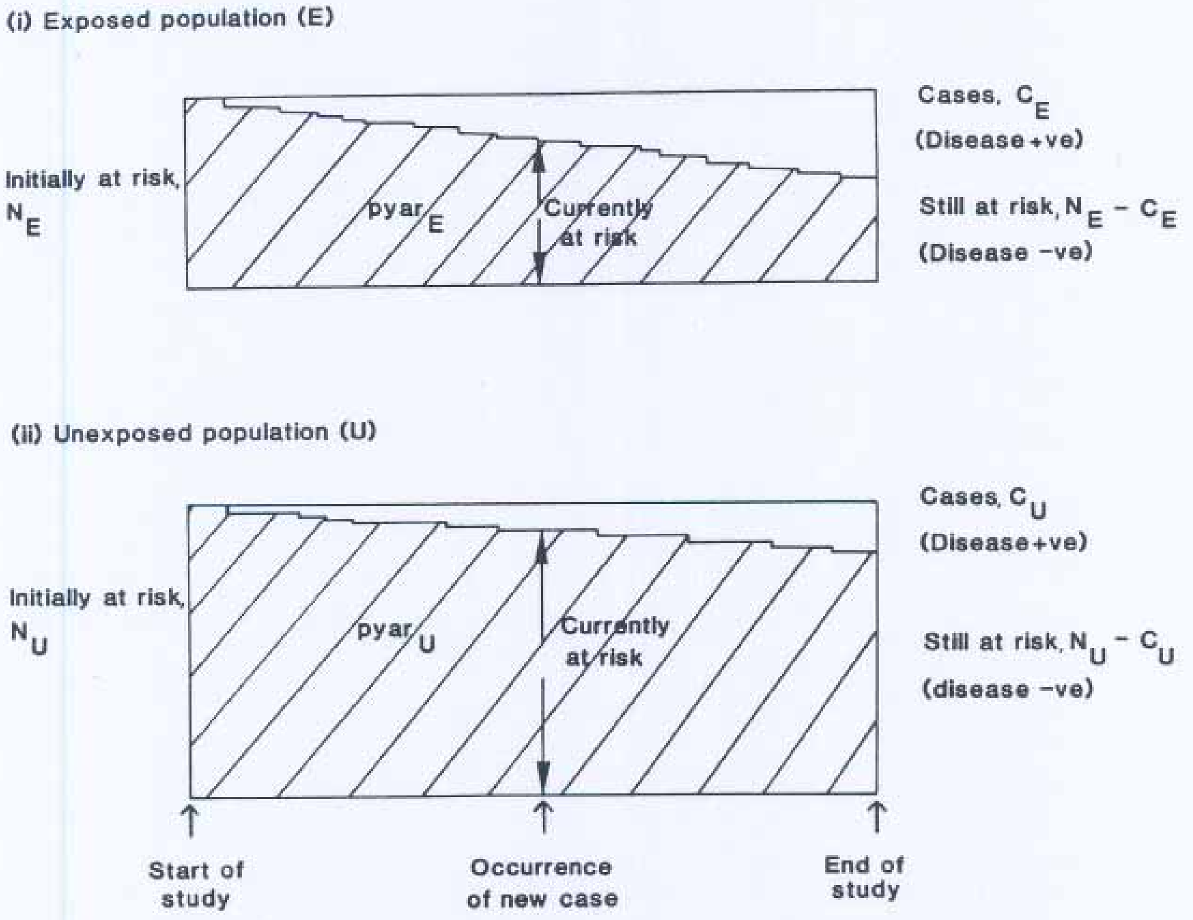

FIGURE 1 *A graphical representation of a cohort study in a fixed population to illustrate the relationships between the three different measures of relative incidence defined in Table 1. The shaded areas represent the number of person years at risk (pyar) in the exposed and unexposed groups.*

risk.* The ratio of the incidence rate in the exposed group ($C_E/pyar_E$) to that in the unexposed group ($C_U/pyar_U$) is called the relative rate (or incidence density ratio), and is shown in Table 1.

A third measure of incidence is the *odds* of disease to non-disease, which equals the total number of cases (C) divided by the number of people still at risk at the end of the study (N-C). This measure has no clear conceptual meaning (outside of horse racing circles). Its importance arises from the fact that the ratio of the odds of disease in the exposed group (C_E/N_E-C_E) to that in the unexposed group (C_U/N_U-C_U) can be expressed as the ratio of the odds of exposure to non-exposure in the cases to the odds of exposure in the non-cases, as shown in Table 1. In other words the disease odds ratio is equivalent to the exposure odds ratio. For this reason the *odds ratio* has played a central role in case-control studies, as an estimate of relative risk or rate.

For a *rare* disease, such as many cancers, the number at risk will approximately equal the total population at all times. The lines in Figure 1 showing the accumulation of cases will be almost indistinguishable from the top lines representing the total population under study. The number of people initially at risk, the average number at risk during the study and the number still at risk at the end will all approximately equal the total population, and all three measures of incidence will therefore be numerically equal. For a *common* disease, however, the three measures are different, and when used to assess an association between a risk factor and disease, will lead to different estimates of

* A distinction has also been made between average incidence rate and instantaneous incidence rate, also commonly called force of morbidity[14,18] or, particularly in the context of survival analysis, hazard. Instantaneous rate is a mathematical concept measuring the rate of change of the at-risk population at a single point in time. In practice we seldom obtain instantaneous incidence rates and estimate, instead, the average of the instantaneous rates over a period of time. We have called this average simply the incidence rate; others such as Kleinbaum *et al.*[19] prefer to use the term incidence density for the average, as suggested by Miettinen,[14] and to restrict incidence rate to mean the instantaneous rate.

TABLE 1 *Measuring relative incidence in a case-control study, using the notation defined in Figure 1.*

Measure	Definition	Alternative formulation	Design	Controls sampled from
Relative risk/Cumulative incidence ratio	$\frac{C_E/N_E}{C_U/N_U}$	$\frac{C_E/C_U}{N_E/N_U}$	Inclusive	Total study population regardless of past/future disease status
Relative ~~risk~~ rate/Incidence density ratio	$\frac{C_E/pyar_E}{C_U/pyar_U}$	$\frac{C_E/C_U}{pyar_E/pyar_U}$	Concurrent	People currently at risk
Odds ratio	$\frac{C_E/(N_E - C_E)}{C_U/(N_U - C_U)}$	$\frac{C_E/C_U}{(N_E - C_E)/(N_U - C_U)}$	Traditional (Exclusive)	People disease-free thoughout study period

relative incidence. Moreover these measures will have different properties depending on the particular circumstance.

The use of the odds ratio has been almost entirely confined to case-control studies, where it has been used as an approximation to the relative risk or relative rate. However, it is now known that it is also possible to obtain direct estimates of these two measures in a case-control design, by employing a suitable choice of sampling scheme for the controls. This will be described in the following section. We will then discuss the factors that determine the choice of appropriate measure, and categorize the situations relevant to each. The role of the odds ratio in this context will be re-explored.

SAMPLING SCHEMES FOR CONTROLS IN CASE-CONTROL DESIGNS

Although a case-control study does not directly estimate incidence among the exposed and non-exposed populations, it does yield a measure of relative incidence by comparing the odds of exposure among the cases and the controls. The formulae for all three measures of relative incidence can be written in a form where the numerator is the ratio of exposed to non-exposed cases in the population, namely C_E/C_U, as shown in Table 1.

This ratio is readily estimated from the study cases. The corresponding denominators of the three measures vary. They are the ratios of exposed to non-exposed of (i) people at risk at the start of the study (N_E/N_U), (ii) person years at risk for the duration of the study ($pyar_E/pyar_C$), and (iii) people still disease free at the end of the study ($N_E–C_E/N_U–C_U$)). By a suitable choice of sampling scheme, each of these three types of denominators may be estimated by the ratio of exposed to non-exposed controls, as summarized in Table 1.

Traditional or (Exclusive) Design

Traditionally controls were sampled from the population still at risk at the end of the study period. In this case the odds ratio of exposure of cases to controls is equivalent to the odds ratio of disease of exposed to non-exposed.

Concurrent Design

In this design controls are selected concurrently from those still at risk when a new case is diagnosed.[14] A person originally selected as a control can therefore, at a later date, be ascertained as a case. The opposite can not happen, since once a person has acquired the disease they are no longer at risk, and therefore not eligible for selection as a control. (Diseases which do not kill or induce immunity are exceptions to this and will be considered separately below). People selected as both control and case are included in both groups.

In this design the control group is representing the person years at risk experience, and an analysis *matched* on time of selection will yield an unbiased estimate of relative rate, provided that this is constant over the study period. (This assumption is the same as that known as 'proportional hazards' in the context of survival analysis). This is justified in the appendix, where it can be seen that an *unmatched* analysis will also yield an unbiased estimate of relative rate provided that the rates of acquiring disease are constant over time among both the exposed and non-exposed populations and that the total numbers at risk remain constant in both populations.

Note that including controls who later become cases as cases only and not as both will turn this concurrent design into an equivalent traditional one, and lead to an estimate of the odds ratio, rather than the relative rate.

'Inclusive' Design

In the third type of sampling scheme, controls are chosen from among all individuals in the population, regardless of whether or not they have already had the disease under study. The role of the control group is to estimate the proportion of the total population that is exposed. In a fixed population, it represents those identified as being at risk at the start of the study. The odds ratio of exposure of cases to controls therefore yields an estimate of relative risk.

This design has been variously called 'hybrid retrospective',[20] 'case-base',[23] 'case-exposure'[22] and 'case-cohort',[23]. We propose instead the term 'inclusive'. Since the control group reflects the total population, a person ascertained as a case may also be selected as a control, and *vice versa* and are included in the study as both cases and controls. A major advantage of this ampling scheme is that it is not necessary to obtain disease histories of selected controls since it is not required to exclude past cases.

Not that it is also possible to obtain estimates of the relative rate and odds ratio from this design by simply excluding at the analysis stage people who do not satisfy the conditions of respectively the concurrent and traditional sampling schemes. If this is planned, the overall sample size should be adjusted accordingly to allow for these exclusions, and disease histories of controls should be collected.

CHOICE OF MEASURE

For a rare disease, the at-risk population will be approximately equivalent to the total population throughout the study, and the three case-control designs will yield almost identical results. It will be a rare event to select a control that has already been or will later become a case, and so for practical purposes this possibility can be ignored. For a common disease, however, the three sampling schemes will lead to different measures of relative incidence, with different properties depending on the particular circumstance. For example, in some situations (see type B below), the relative rate will be *invariant* over time, while the value of the relative risk will depend on the length of follow-up. In other situations (see type C below), the properties will be reversed.

The choice of design will be determined by which of the three possible measures of relative incidence best meets the study objectives. If the main objective is to estimate the size of the increase in incidence, then the measure of choice is the one that is constant over time (invariant) and unbiased. In the majority of situations this will be the relative rate, although in some (see type C below) it will be the relative risk. The exception is

TABLE 2 *Choice of design according to type of situation*

Situation	A	B	C	D	E
Type of disease	Rare disease	Non-recurrent common disease	Non-recurrent common disease	Non-recurrent common disease	Recurrent common diseases (short duration, low case fatality)
Type of risk factor	All risk/protective factors	Risk/protective factor, which affects all exposed equally	Protective factors, which does not affect all exposed equally	Risk factor, which does not affect all exposed equally	All risk/protective factors
Example(s)	Most cancers/ Congenital disorders/Accidents	Crowding as a risk factor for measles/ Vaccines which give partial protection to those vaccinated	Vaccines giving 'all or nothing' protection, such as measles and hepatitis vaccines	Factor 8 transfusion as risk factor for HIV infection in haemophiliacs	Diarrhoea/Acute respiratory infections
Cases return to population at risk	No	No	No	No	Yes—Possibility of multiple episodes
Proportion exposed to risk factor constant in population at risk	Yes, because number of cases is negligible	No	No	No	Yes
Exposed group at uniform risk of disease	Yes/No	Yes	No	No	Yes/No
Invariant measure	Relative risk/ Relative rate/Odds ratio	Relative rate	Relative risk	None (Relative rate changes most slowly)	Relative rate
Appropriate design	Inclusive/ Concurrent/ Traditional	Concurrent	Inclusive	None (? Concurrent)	Concurrent

where the focus is on increased risk to the individual over a specific time period, such as in an investigation of risk factors for death during infancy. In this case the relative risk is the summary measure of choice, regardless of whether or not it is invariant over the first year of life.

On the other hand, if the primary objective is to identify an association, but not accurately to quantify it, then the odds ratio is the best choice, since this tends to overestimate the real effect and is therefore more sensitive. In other words the odds ratio will be numerically larger than both the relative risk and the relative rate. Thus, for studies of similar sizes, the traditional sampling scheme leading to an estimate of odds rati will have the greatest power of finding a significant result. An extension of this idea applying to the special case of diseases that do not cause immunity, and where individuals may suffer multiple episodes of disease, is described in the particular context below (see situation type E).

CHOICE OF DESIGN

In Table 2 we classify situations into one of five different types according to the characteristics of the disease

TABLE 3 *Comparison of relative rate and relative risk over time in Situation B—An example of a non-recurrent common disease with a risk factor affecting all the exposed equally. λ_E and λ_U represent the instantaneous disease rates in the exposed and unexposed populations, with $\lambda_E/\lambda_U = 2$. The disease rates are assumed constant over time. The exponential survival model therefore applies. The number of cases in a year equals ne^{λ} and the person years at risk ne^{λ}/λ, where n is the number at risk at the start of the year.*

	Year 1	Year 2	Year 3	Year 4
Exposed population ($\lambda_E = 0.2$)				
No. at risk at start of period	1000.0	818.7	670.3	548.8
No. of cases during period	181.3	148.4	121.5	99.5
Person years at risk	906.3	742.0	607.5	497.4
Unexposed population ($\lambda_U = 0.1$)				
No. at risk at start of period	1000.0	904.8	818.7	740.8
No. of cases during period	95.2	86.1	77.9	70.5
Person years at risk	951.6	861.0	779.1	705.0
Relative rate*	2.00	2.00	2.00	2.00
Relative risk*	1.90	1.82	1.74	1.67
Odds ratio*	2.10	2.22	2.35	2.49

*Calculated using the cumulative number of cases accrued from the beginning of year 1.

TABLE 4 *Comparison of relative rate and relative risk over time in Situation C—An example of a non-recurrent common disease with a protective factor conferring complete protection to a subgroup of the exposed population, leaving the rest of the exposed population at the same risk as the unexposed. λ_P and λ_U represent the instantaneous disease rates in the protected and unexposed populations, with 80% of the exposed population protected, giving an overall protective efficacy of 80%.*

	Year 1	Year 2	Year 3	Year 4
Exposed population				
Protected subgroup ($\lambda_P = 0$)				
No. at risk at start of period	800.0	800.0	800.0	800.0
No. of cases during period	0.0	0.0	0.0	0.0
Person years at risk	800.0	800.0	800.0	800.0
Unprotected subgroup ($\lambda_U = 0.2$)				
No. at risk at start of period	200.0	163.7	134.0	109.7
No. of cases during period	36.3	29.7	24.3	19.9
Person years at risk	181.3	148.4	121.5	99.4
Overall				
No. at risk at start of period	1000.0	963.7	934.0	909.7
No. of cases during period	36.3	29.7	24.3	19.9
Person years at risk	981.3	948.4	921.5	899.4
Unexposed population ($\lambda_U = 0.2$)				
No. at risk at start of period	1000.0	818.7	670.3	548.8
No. of cases during period	181.3	148.4	121.5	99.5
Person years at risk	906.3	742.0	607.5	497.4
Relative rate*	0.18	0.17	0.16	0.15
Relative risk*	0.20	0.20	0.20	0.20
Odds ratio	0.17	0.14	0.12	0.10

*Calculated using the cumulative number of cases accrued from the beginning of year 1.

TABLE 5 *Comparison of relative rate and relative risk over time in Situation C—An example of a non-recurrent common disease with a risk factor that does not affect all exposed individuals equally. The exposed group consists of two distinct subgroups, one of which is at increased risk of disease (instantaneous disease rate, λ_E). The other subgroup is at the same risk as the unexposed population (instantaneous disease rate, λ_U). Of the exposed population 50% is assumed to be at increased risk, with $\lambda_E/\lambda_U = 2$.*

	Year 1	Year 2	Year 3	Year 4
Exposed population				
Subgroup at increased risk ($\lambda_E = 0.2$)				
No. at risk at start of period	500.0	409.4	335.2	274.4
No. of cases during period	90.6	74.2	60.8	49.7
Person years at risk	453.1	371.1	303.8	248.7
Subgroup at normal risk ($\lambda_E = 0.1$)				
No. at risk at start of period	500.0	452.4	409.3	370.3
No. of cases during period	47.6	43.1	39.0	35.2
Person years at risk	475.8	430.5	389.5	352.4
Overall				
No. at risk at start of period	1000.0	861.8	744.5	644.7
No. of cases during period	138.2	117.3	99.8	84.9
Person years at risk	928.9	801.6	693.3	601.1
Unexposed population ($\lambda_U = 0.1$)				
No. at risk at start of period	1000.0	904.8	818.7	740.8
No. of cases during period	95.2	86.1	77.9	70.5
Person years at risk	951.6	861.0	779.1	705.0
Relative rate*	1.49	1.48	1.47	1.46
Relative risk*	1.45	1.41	1.37	1.34
Odds ratio*	1.52	1.55	1.58	1.60

*Calculated using the cumulative number of cases accrued from the beginning of year 1.

TABLE 6 *Comparison of relative rate and relative risk over time in Situation E—An example of a recurrent common disease of short duration and low case fatality. λ_E and λ_U represent the instantaneous disease rates in the exposed and unexposed populations, with $\lambda_E/\lambda_U = 2$.*

	Year 1	Year 2	Year 3	Year 4
Exposed population ($\lambda_E = 0.2$)				
No. at risk at start of period	1000.0	1000.0	1000.0	1000.0
No. of cases during period	200.0	200.0	200.0	200.0
Person years at risk	1000.0	1000.0	1000.0	1000.0
Unexposed population ($\lambda_U = 0.1$)				
No. at risk at start of period	1000.0	1000.0	1000.0	1000.0
No. of cases during period	100.0	100.0	100.0	100.0
Person years at risk	1000.0	1000.0	1000.0	1000.0
Relative rate*	2.00	2.00	2.00	2.00
Relative risk (undefined)	—	—	—	—

*Calculated using the cumulative number of cases accrued from the beginning of year 1.

and risk factor under study, and their implications for the choice of appropriate measure of relative incidence and therefore of the sampling scheme for controls in a case-control study. Type A is the classic case-control situation of the study of rare diseases. All other categories refer to the study of common diseases. Types B, C and D are non-recurrent diseases, subclassified according to the mode of action of the risk factor, namely whether it is thought to affect all individuals equally (B) or differentially (C and D). Type E refers to recurrent common diseases, which we assume to be of short duration and low case fatality.

Situation A

The first category covers the study of rare disorders, such as most cancers, congenital malformations, accidents and diabetes. There is no precise definition, but it may be considered to include any disease with an incidence risk below 5% over the study period. This is the classic case-control situation. Because the disease is rare, cases constitute a negligible fraction of the population, and the proportion of the population exposed to the risk factor remains constant over time, being essentially unaffected by any accumulation of individuals into the case group, whether or not this is related to exposure. All three sampling schemes for controls are essentially the same and the three associated measures of relative incidence numerically equivalent. This applies whatever the other characteristics of the disease and the risk factor.

Situation B

Situations type B, C and D refer to non-recurrent common diseases, that is diseases that either confer immunity from further attacks or lead to death. The difference is the assumed mode of action of the risk factor. We start by considering risk factors that are thought to affect all exposed people equally. Examples are the study of crowding as a risk factor for measles, of poor sanitation for hepatitis, and of vaccines that confer partial protection to those immunized.

In this situation, the relative rate is invariant over time, as shown in Table 3, yielding an unbiased estimate of the ratio of the instantaneous disease rates in the exposed and unexposed populations. The relative risk is found to decrease with increasing length of follow-up, and to underestimate the true ratio even during the first time period. In contrast the odds ratio increases with time and consistently overestimates the effect; this is true in all situations.

The estimation of a relative rate with the concurrent design is therefore the appropriate procedure, unless the aim is specifically to estimate the relative risk over a particular time period, such as infancy. As explained earlier, in most circumstances it will be necessary to carry out a matched analysis, matching for time of selection. This will certainly be the case for any disease exhibiting seasonality or time trends, such as measles, since the disease rates will not be constant over time. For diseases that occur with constant rates, it is necessary to match if the sizes of the exposed and/or non-exposed populations change over time, but not if they remain constant. For common diseases the latter is unlikely, unless the population under consideration is dynamic and the influx of new at-risk people balances the occurrence of cases.

Situation C

Situations C and D also relate to the study of non-recurrent common diseases, and are characterized by the mode of action of the risk factor, which does not affect all exposed individuals equally. In situation C, the factor is protective; in situation D the factor is associated with increased risk of disease. An example of situation C would be the assessment of the efficacy of measles immunization in an endemic area, if the vaccine had an 'all or nothing' effect, that is if it worked by conferring complete protection to some individuals and none to others, rather than partial protection to everyone.

In such a case, the exposed group consists of two distinct sub-populations, one of which is totally protected against disease, and the other of which is at the same risk as the non-exposed population. This is illustrated in Table 4, which shows that the relative rate is not constant over time, but increasingly overestimates the protective effect of the factor. The relative risk is, however, invariant, and measures the proportion of the exposed population that was totally protected,[11] with one minus the relative risk yielding an estimate of protective efficacy. The inclusive design is therefore the appropriate choice for a case-control study. Since the relative risk remains constant over time, the case-control study may be based on cases diagnosed during any period following the point of definition of exposure. It is not necessary to cumulate cases continuously from this point. Thus, in the example given in Table 4, the case-control study could be based on cases diagnosed during year 3 only, with controls selected from the total population.

Situation D

Situation D relates to the study of non-recurrent common diseases where the risk factor does not affect all exposed individuals equally. Instead, the exposed group consists of two distinct sub-populations, one of which is at increased risk of disease, and the other of which is at the same risk as the non-exposed population. An example would be the study of Factor 8 transfusion as a risk factor for becoming HIV positive, since a transfused haemophiliac will only be at increased risk if the blood was infected.

There are two parameters to be estimated; the proportion of the exposed population that is affected and the increase in rate. These parameters are not directly estimated by either the relative risk or the relative rate, as can be seen in Table 5. Both measures decrease with time, reflecting a decreasing proportion of affected individuals in the at-risk exposed population, due to their differentially higher rate of acquiring disease. Neither measure is completely appropriate, although the relative rate decreases more slowly over time than the relative risk. Furthermore, the odds ratio appears to increase with time. The lack of invariance of all three measures affects the ability of all studies, longitudinal as well as case-control, to examine changes in risk over time, due to risk factors that affect only a proportion of the population classified as exposed. This area needs more theoretical attention.

Situation E

Finally, we turn to recurrent common diseases, such as diarrhoea and acute respiratory infections, which an individual may experience more than once. Cases therefore return to the population at risk after recovery. For diseases with a short duration and low case fatality, this means that the proportion of the at-risk population that is exposed remains constant over time, unless there is differential influx of new people (migrants or newborns). This applies irrespective of whether or not the risk factor affects all exposed equally. In all situations, the relative rate is the meaningful measure as this directly takes account of the possible multiplicity of episodes, estimating, for example, the ratio of the number of episodes of diarrhoea per child per year among the exposed to that among the unexposed. The concurrent design is therefore the appropriate choice for case-control studies. It should be noted that, in this type of situation, cases can become controls as well as *vice versa*, because of th recurrent nature of the disease.

An alternative approach is to focus on individuals rather than episodes, and define cases as those who have suffered from at least one episode of disease. If the definition is changed in this way, cases no longer return to the at-risk pool and the situation is identical to type B. This also applies if attention is restricted to risk factors for repeated episodes of disease, when cases might be defined, for example, as those who suffered at least three times, representing particularly susceptible individuals.

It is also possible to define different categories of cases according to the number of episodes of disease experienced, for example three categories such as 1–2, 3–4 and 5+ episodes, and to compare each group of cases with a control group who remained disease-free throughout the period of study (and with the other groups of cases). The resulting measures would be a type of 'modified' odds ratio, and the appropriate statistical procedure would be an analysis of trend.

All the suggestions can be applied at the analysis stage of the concurrent episode-based design, as well as being options for the overall design of the study. In some circumstances, it may be desirable in the design to minimize sample sizes by selecting only individuals from the top category as cases and comparing these with disease-free controls, as this 'modified' odds ratio comparison would have the greatest chance of detecting effects. The disadvantages are that the information obtained will not provide as complete a picture and that no estimate can be made for the relative rate.

ACKNOWLEDGEMENTS

We would like to thank the following colleagues for reading earlier versions of this paper and making many helpful comments: Sandy Cairncross, Matthias Egger, Richard Feachem, Paul Fine, Peter Sandiford and Cesar Victora.

REFERENCES

[1] Linnemman C C Jr, Dine M S, Bloom J E. Measles antibodies in previously immunized children. *Am J Dis Child* 1972; **124:** 53–7.

[2] Marks J S, Halpin T J, Orenstein W A. Measles vaccine efficacy in children previously vaccinated at 12 months of age. *Pediatrics* 1978; **62:** 955–60.

[3] Clarke E A, Anderson T W. Does screening for 'Pap' smears help prevent cancer? *Lancet* 1979; **ii:** 1–4.

[4] Hutchinson T A, Polansky S M, Feinstein A R. Post-menopausal oestrogens protect against fractures of the hip and distal radius: A case control study. *Lancet* 1979; **ii:** 705–9.

[5] Weiss E S, Kendrick P L. The effectiveness of pertussis vaccine : An application of Sargent and Merrel's method of measurement. *Am J Hygiene* 1943; **38:** 306–9.

[6] Mustard J F, Kinlough-Rathbone R L, Packham M A. Aspirin in the treatment of cardiovascular disease: A review. *Am J Med* 1983; **74:** 43–9.

[7] Niswander K, Henson G, Elbourne D, Chalmers I, Redman G, MacFarlane A, Tizard P. Adverse outcome of pregnancy and the quality of obstetric care. *Lancet* 1984; **ii:** 827–31.

[8] Miller D L, Rose E M, Aldersdale R, Bellman M H, Rawson N S. Pertussis immunization and serious, acute neurological illness in children. *Br Med J* 1981; **282:** 1595–9.

[9] Fine P E, Ponnighaus J M, Maine N, Clarkson J A, Bliss L. Protective efficacy of BCG against leprosy in Northern Malawi. *Lancet* 1986; **ii:** 449–502.

[10] Smith P G. Retrospective assessment of the effectiveness of BCG vaccination against tuberculosis using the case-control method. *Tubercle* 1982; **62:** 23–35.

[11] Smith P G, Rodrigues L C, Fine P E M. Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies. *Int J Epidemiol* 1984; **13:** 87–93.

[12] Briscoe J, Feachem R G, Rahaman M M. *Evaluating Health Impact: Water Supply, Sanitation, and Hygiene Education.* International Development Research Centre 1986; Ottawa, Canada.

[13] Cornfield J. A method of estimating comparative rates from clinical data: Applications to cancer of the lung, breast and cervix. *J Nat Cancer Inst* 1951; **11:** 1269–75.

[14] Miettinen O S. Estimability and estimation in case-referrent studies. *Am J Epidemiol* 1976; **103:** 226–35.

[15] Greenland S, Thomas D C. On the need for the rare disease assumption in case-control studies. *Am J Epidemiol* 1982; **116:** 547–53.

[16] F[illegible]dt-Johnson R C. Definition of rates: Some remarks on their use and misuse. *Am J Epidemiol* 1975; **102:** 267–71.

[17] Vandenbroucke J P. On the rediscovery of a distinction. *Am J Epidemiol* 1985; **121:** 627–8.

[18] Gehan E A. Estimating survival functions from the life table. *J Chronic Dis* 1969; **21:** 629–44.

[19] Kleinbaum D G, Kupper L L, Morgenstern H. *Epidemiological Research: Principles and Quantitative Methods.* Van Nostrand Reinhold Company 1982; New York.

[20] Kupper L L, McMichael A S, Spirtas R. A hybrid epidemiological study design useful in estimating relative risk. *J Amer Statist Ass* 1975; **70:** 524–8.

[21] Miettinen O. Design options in epidemiological research. An update. *Scand J Work Environ Health* 1982; **8 (Suppl 1):** 7–14.

[22] Hogue C J R, Gaylor D W, Schulz K F. The case exposure study—a further explication and response to a critique. *Am J Epidemiol* 1986; **124:** 877–83.

[23] Prentice R L. A case cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; **73:** 1–11.

(Revised version received May 1989)

APPENDIX

Estimation of the relative rate from a concurrent case-control study

In a concurrent design, a control is selected when each new case occurs. At any point in time (t), the probability that a new case occurs in the exposed group will be proportional to the number of people still at risk in the exposed group N_E (t), multiplied by the instantaneous disease rate associated with exposure, $\lambda_E(t)$. Similarly the probability that a new case occurs in the non-exposed group will be proportional to $N_U(t)$, multiplied by $\lambda_U(t)$. Therefore, if a new case does occur at a particular time t, the likelihood that it will be from the exposed or non-exposed groups will be in proportion to these two probabilities. When the corresponding control is selected, the likelihood of exposed and non-exposed status will be directly in proportion to the numbers at risk in the two groups, namely $N_E(t)$ and $N_U(t)$. Therefore, on average, the following composition is expected to be found for a case-control pair occurring at time t.

	Case	Control
Exposed	$\frac{N_E(t)\lambda_E(t)}{N_E(t)\lambda_E(t) + N_U(t)\lambda_U(t)}$	$\frac{N_E(t)}{N_E(t) + N_U(t)}$
Non-exposed	$\frac{N_E(t)\lambda_E(t)}{N_E(t)\lambda_E(t) + N_U(t)\lambda_U(t)}$	$\frac{N_E(t)}{N_E(t) + N_U(t)}$
Total	1	1

The cross-product ratio of this table equals $\lambda_E(t)/\lambda_U(t)$, which is the relative rate at time t. If this relative rate is constant over time, a *matched* analysis will yield an unbiased estimate of it.

An *unmatched* analysis, with just one summary 2×2 table showing the case-control comparison, is only appropriate if all the denominators in the separate pair tables are constant over time. This requires that the numbers at risk in the exposed and non-exposed groups, $N_E(t)$ and $N_U(t)$, and the incidence rates in the two groups, $\lambda_E(t)$ and $\lambda_U(t)$, do not change over the study period.