Published by Oxford University Press on behalf of the International Epidemiological Association

International Journal of Epidemiology 2012;**41**:393–397
doi:10.1093/ije/dys049

EDUCATION CORNER

Classification of epidemiological study designs

Neil Pearce[1,2]

[1]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, UK and [2]Centre for Public Health Research, Massey University, Wellington, New Zealand

Correspondence to: Department of Medical Statistics, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. E-mail: neil.pearce@lshtm.ac.uk

Accepted 2 March 2012

In this article, I present a simple classification scheme for epidemiological study designs, a topic about which there has been considerable debate over several decades. I will argue that when the individual is the unit of analysis and the disease outcome under study is dichotomous, then epidemiological study designs can best be classified according to two criteria: (i) the type of outcome under study (incidence or prevalence) and (ii) whether there is sampling on the basis of the outcome. This classification system has previously been proposed by Greenland and Morgenstern (1988)[1] and Morgenstern and Thomas (1993),[2] all of whom followed previous authors[3,4] in rejecting directionality (i.e. prospective/retrospective or from exposure to outcome vs from outcome to exposure) as a key feature for distinguishing study designs.

Once this two-dimensional classification system has been adopted, then there are only four basic study designs (Table 1):[2,5,6] (i) incidence studies; (ii) incidence case–control studies; (iii) prevalence studies; and (iv) prevalence case–control studies (Rothman *et al*.[7] use the terms 'incident case–control study' and 'prevalent case–control study' where the adjective refers to the incident or prevalent cases[2]).

In this article, I will briefly illustrate these four different study designs for dichotomous outcomes; I then briefly consider the extension of this classification to include studies with continuous exposure or outcome measures and I briefly mention other possible axes of classification.

The four basic study designs

It should first be emphasized that all epidemiological studies are (or should be) based on a particular population (the 'source population') followed over a particular period of time (the 'risk period'). Within this framework, the most fundamental distinction is between studies of disease 'incidence' and studies of disease 'prevalence'. Once this distinction has been drawn, then the different epidemiological study designs differ primarily in the manner in which information is drawn from the source population and risk period.[8]

Incidence studies

Incidence studies ideally measure exposures, confounders and outcome times of all population members. Table 2 shows the findings of a hypothetical incidence study involving 10 000 people who are exposed to a particular risk factor and 10 000 people who are not exposed. When the source population has been formally defined and enumerated (e.g. a group of workers exposed to a particular chemical), then the study may be termed a 'cohort study' or 'follow-up study' and the former terminology will be used here. Incidence studies also include studies where the source population has been defined but a cohort has not been formally enumerated by the investigator, e.g. 'descriptive' studies of national death rates. Furthermore, there is no fundamental distinction between incidence studies based on a broad population (e.g. all workers at a particular factory or all persons living in a particular geographical area) and incidence studies involving sampling on the basis of exposure, since the latter procedure merely redefines the study population (cohort).[4]

Three measures of disease occurrence are commonly used in incidence studies.[9] Perhaps the most common measure is the person–time 'incidence rate'; a second measure is the 'incidence proportion' (average risk), which is the proportion of study subjects who experience the outcome of interest at any time during the follow-up period. A third possible measure is the 'incidence odds', which is the ratio of the number of subjects who experience the outcome to the number of subjects who do not experience the outcome. These three measures of disease occurrence all involve the same numerator: the number of incident cases of disease. They differ in whether their denominators

Table 1 Four basic study types

	Sampling on outcome	
Study outcome	No	Yes
Incidence	Incidence studies	Incidence case–control studies
Prevalence	Prevalence studies	Prevalence case–control studies

Table 2 Findings from a hypothetical cohort study of 20 000 persons followed for 10 years

	Exposed	**Non-exposed**	**Ratio**
Cases	1813 (a)	952 (b)	
Non-cases	8187 (c)	9048 (d)	
Initial population size	10 000 (N_1)	10 000 (N_0)	
Person-years	90 635 (Y_1)	95 163 (Y_0)	
Incidence rate	0.0200 (I_1)	0.0100 (I_0)	2.00
Incidence proportion (average risk)	0.1813 (R_1)	0.0952 (R_0)	1.90
Incidence odds	0.2214 (O_1)	0.1052 (O_0)	2.11

Table 3 Findings from a hypothetical incidence case–control study based on the cohort in Table 1

	Exposed	**Non-exposed**	**OR**
Cases	1813 (a)	952 (b)	
Controls			
From survivors (cumulative sampling)	1313 (c)	1452 (d)	2.11
From source population (case–cohort sampling)	1383 (c)	1383 (d)	1.90
From person-years (density sampling)	1349 (c)	1416 (d)	2.00

represent person–time at risk, persons at risk or survivors.

Corresponding to these three measures of disease occurrence, the three ratio measures of effect used in incidence studies are the 'rate ratio', 'risk ratio' and 'odds ratio'.

Incidence case–control studies

Incidence studies are usually the preferred approach to studying the causes of disease, because they use all of the available information on the source population over the risk period. However, they are often very expensive in terms of time and resources, and the equivalent results may be achieved more efficiently by using an incidence case–control study design.

Table 3 shows the data from a hypothetical incidence case–control study of all 2765 incident cases in the full cohort in Table 2 and a random sample of 2765 controls. Such a study would on an average achieve the same findings as the full cohort study (Table 2), but would be considerably more efficient, since it would involve ascertaining the exposure histories of 5530 people (2765 cases and 2765 controls) rather than 20 000 people. When the outcome under study is rare, an even more remarkable gain in efficiency can be achieved with only a minimal reduction in the precision of the effect estimate.

In incidence case–control studies, the relative risk measure is the 'odds ratio'. The effect measure that the odds ratio (OR) obtained from this case–control study will estimate depends on the manner in which controls are selected. Once again, there are three main options that define three subtypes of incidence case–control studies.[10,11]

One option is to select controls at random from those who do not experience the outcome during the follow-up period, i.e. the 'survivors' (those who did not develop the outcome at any time during the follow-up period). In this instance, a sample of controls chosen by 'cumulative sampling' (or exclusive sampling[11]) will estimate the exposure odds of the survivors, and the OR obtained in the case–control study will therefore estimate the incidence OR in the base population. Early descriptions of the case–control approach were usually of this type.[12]

These descriptions emphasized that the OR was approximately equal to the risk ratio when the disease was rare (in Table 3; this OR = 2.11).

It was later recognized that controls can be sampled at random from the entire 'source population' (those at risk at the beginning of follow-up) rather than just from the survivors (those at risk at the end of follow-up). This approach, which has been reinvented several times since it was first proposed by Thomas,[13] has more recently been termed 'case–cohort sampling'[14] (or inclusive sampling[11]). In this instance, the controls will estimate the exposure odds in the source population at the start of follow-up, and the OR obtained in the case–control study will therefore estimate the risk ratio in the source population (which is 1.90 in Table 3). The method of calculation of the OR is the same as for any other case–control study, but special formulas must be used to compute confidence intervals and P-values.[15]

The third approach is to select controls longitudinally throughout the course of the study, an approach now usually referred to as 'density sampling'[7] (or concurrent sampling[11]); the resulting OR will estimate the rate ratio in the source population (which is 2.00 in Table 3). Most case–control studies involve density sampling (often with matching on a time variable such as calendar time or age), and therefore estimate the incidence rate ratio without the need for any rare disease assumption.[16]

Prevalence studies

Incidence studies are usually the preferred approach to studying the causes of disease, but they often involve lengthy periods of follow-up and large resources.[17] Also, for some diseases (e.g. asthma and diabetes), incidence may be difficult to measure without very intensive follow-up. Thus, it is often more practical to study the 'prevalence' of disease at a particular point in time. This approach has one major potential shortcoming, since disease prevalence may differ between two groups because of differences in age-specific disease incidence, disease duration or other population parameters;[7] thus, it is much more difficult to assess causation (i.e. whether an exposure increases disease incidence) in prevalence studies. Nevertheless, for many common diseases, studying prevalence is often the only practical option and may be an important first step in the research process; furthermore, prevalence may be of interest in itself, e.g. because it measures the population burden of disease. For example, motor neurone disease and multiple sclerosis have similar incidence and mortality rates, but multiple sclerosis represents a greater burden of morbidity for the health services, because survival for motor neurone disease is so short.[18]

Table 4 Findings from a hypothetical prevalence study of 20 000 persons

	Exposed	Non-exposed	Ratio
Cases	909 (a)	476 (b)	
Non-cases	9091 (c)	9524 (d)	
Total population	10 000 (N_1)	10 000 (N_2)	
Prevalence	0.0909 (P_1)	0.0476 (P_0)	1.91
Prevalence odds	0.1000 (O_1)	0.0500 (O_0)	2.00

Table 4 shows data from a prevalence study of 20 000 people (this example has been designed to correspond to the incidence study examples given above, assuming that the exposure has no effect on disease duration and that there is no immigration into or emigration from the prevalence pool, so that no one leaves the pool except by disease onset, death or recovery[7]). The prevalence is 0.0909 in the exposed group and 0.0476 in the non-exposed group, and the prevalence ratio (PR) and prevalence odds ratio (POR) are 1.91 and 2.00, respectively.

Note that this definition of prevalence studies does not involve any specification of the timing of the measurement of exposure. In many prevalence studies, information on exposure will be physically collected by the investigator and at the same time information on disease prevalence is collected. Nonetheless, exposure information may include factors that do not change over time (e.g. gender) or change in a predictable manner (e.g. age), as well as factors that do change over time. The latter may have been measured at the time of data collection [e.g. current levels of airborne asbestos exposure, body mass index (BMI)] or at a previous time (e.g. historical records on past asbestos exposure levels, birthweight recorded in hospital records), or integrated over time (e.g. using a job–exposure matrix and work history records). The sole defining feature of prevalence studies is that they involve studying disease prevalence. There is no restriction on when the exposure information is collected or whether it relates to current and/or historical exposures.

Also note that some prevalence studies may involve sampling on exposure status, just as some incidence studies may involve such sampling. For example, in a study of a group of factory workers, asthma prevalence may be measured in all exposed workers and a sample of non-exposed workers. This sampling scheme does not change the basic study type, rather it redefines the population that is being studied (from the entire group of workers in the factory to the newly defined subgroup).[17]

Prevalence case–control studies

Just as an incidence case–control study can be used to obtain the same findings as a full cohort study, a prevalence case–control study can be used to obtain the same findings as a full prevalence study in a more efficient manner. In particular, if obtaining exposure information is difficult or costly, then it may be more

Table 5 Findings from a hypothetical prevalence case–control study based on the population represented in Table 3

	Exposed	Non-exposed	Ratio
Cases	909 (a)	476 (b)	
Controls	676 (c)	709 (d)	
Prevalence odds	1.34 (O_1)	0.67 (O_0)	2.00

efficient to conduct a prevalence case–control study by obtaining exposure information on some or all of the prevalent cases and a sample of controls selected from the non-cases.

Suppose that a prevalence case–control study is conducted using the source population in Table 4, involving all the 1385 prevalent cases and a group of 1385 controls (Table 5). In this instance, there is one main option for selecting controls, namely to select them from the non-cases. This will enable us to estimate the exposure odds of the non-cases, and the OR obtained in the prevalence case–control study will therefore estimate the POR in the source population (2.00).[17] Alternatively, if the PR is the effect measure of interest, controls can be sampled from the entire source population (i.e. in a manner analogous to case–cohort sampling) and the resulting prevalence case–control ‘OR’ will estimate the PR in the source population.

Extension to continuous exposures or outcomes

The basic study designs presented above can be extended by the inclusion of continuous exposure data and continuous outcome measures. The extension to continuous exposure measures requires minor changes to the data analysis, but it does not alter the 4-fold categorization of study design options presented above. However, the extension to continuous outcome measures does require further discussion.

Continuous outcome measures

Cross-sectional studies

In the presentation of prevalence studies above, the health outcome under study was a ‘state’ (e.g. having or not having hypertension). Studies could involve observing the incidence of the ‘event’ of acquiring the disease state (e.g. the incidence of being diagnosed with hypertension), or the prevalence of the disease state (e.g. the prevalence of hypertension). More generally, the health state under study may have multiple categories (e.g. non-hypertensive, mild hypertension, moderate hypertension and severe hypertension) or may be represented by a continuous measurement (e.g. blood pressure). Since these measurements are taken at a particular point in time, such studies are often referred to as ‘cross-sectional studies’. Prevalence studies are a subgroup of cross-sectional studies in which the disease outcome is dichotomous.

Longitudinal studies

Longitudinal studies (cohort studies) involve repeated observation of study participants over time. They represent the most comprehensive approach since they use all of the available information on the source population over the risk period. Incidence studies are a subgroup of longitudinal study in which the outcome measure is dichotomous. More generally, longitudinal studies may involve repeated assessment of categorical or continuous outcome measures over time (e.g. a series of linked cross-sectional studies in the same population). A simple longitudinal study may involve comparing the disease outcome measure or more usually changes in the measure, over time, between exposed and non-exposed groups. For example, rather than comparing the incidence of hypertension (as in an incidence study) or the prevalence at a particular time (as in a prevalence study), or the mean blood pressure at a particular point in time (as in a cross-sectional study), a longitudinal study might involve measuring baseline blood pressure in exposed and non-exposed persons and then comparing changes in blood pressure (i.e. the change from the baseline measure) over time in the two groups. One special type of longitudinal study is that of ‘time series’ comparisons in which variations in exposure levels and symptom levels are assessed over time with each individual serving as their own comparison.

Other axes of classification

Finally, it should be noted that there are other possible axes of classification or extension of the above classification scheme. These include the timing of collection of exposure information (which is related to classifications based on ‘directionality’), the sources of exposure information (routine records, questionnaires and biomarkers) and the level at which exposure is measured or defined (e.g. population or individual). However, none of these axes is crucial in terms of classifying studies in which the individual is the unit of analysis.

Discussion

There is no definitive approach to classifying types of epidemiological studies, and different classification schemes may be useful for different purposes. A classification scheme will be useful if it helps us to teach and learn fundamental concepts without obscuring other issues, including the many ‘messier’ issues that occur in practice. The scheme presented here involves ‘ideal types’ that are not always followed in practice and mixes can occur along both axes. For example, two-stage designs are not unambiguously cohort or case–control (usually, the second stage involves sampling on outcome and the first stage does not), and studies of malformations are not unambiguously incidence or prevalence. Thus, undoubtedly some readers will find the scheme presented here simplistic. Nonetheless, this 4-fold classification of study types has several advantages over other classification schemes. First, it captures the important distinction between incidence and prevalence studies; in doing so it clarifies the distinctive feature of cross-sectional (prevalence) studies, namely that they involve prevalence data rather than incidence data. Secondly, it captures the important distinction between studies that involve collecting data on all members of a population and studies that involve sampling on outcome (this is the widely accepted distinction between cohort and case–control studies). Finally, it clarifies the range of possibilities and problems of different study designs, particularly by emphasizing that the issues of the timing of data collection are not unique to case–control studies and are not crucial in terms of classification of epidemiological study designs.

Funding

Programme Grant from the Health Research Council of New Zealand (The Centre for Public Health Research).

Conflict of interest: None declared.

References

[1] Greenland S, Morgenstern H. Classification schemes for epidemiologic research designs. *J Clin Epidemiol* 1988;**41:** 715–16.

[2] Morgenstern H, Thomas D. Principles of study design in environmental epidemiology. *Environ Health Perspect* 1993; **101:**23–38.

[3] Miettinen OS. *Theoretical Epidemiology*. New York: John Wiley & Sons, Inc., 1985.

[4] Rothman KJ. *Modern Epidemiology*. Boston: Little, Brown, 1986.

[5] Pearce N, Crane J. Epidemiologic methods. In: Balmes J (ed.). *Occupational and Environmental Respiratory Disease*. St Louis, MI: Mosby, 1995, pp. 13–27.

[6] Pearce N. The four basic epidemiologic study types. *J Epidemiol Biostat* 1998;**3:**171–77.

[7] Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd edn. Philadelphia: Lippincott Williams & Wilkins, 2008.

[8] Checkoway H, Pearce N, Kriebel D. *Research Methods in Occupational Epidemiology*. 2nd edn. New York: Oxford University Press, 2004.

[9] Greenland S, Rothman KJ. Measures of occurrence. In: Rothman KJ, Greenland S, Lash TL (eds). *Modern Epidemiology*. 3rd edn. Philadelphia: Lippincott Williams & Wilkins, 2008.

[10] Pearce N. What does the odds ratio estimate in a case–control study? *Int J Epidemiol* 1993;**22:**1189–92.

[11] 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:**205–13.

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

[13] Thomas DB. Relationship of oral contraceptives to cervical carcinogenesis. *Obstet Gynecol* 1972;**40:**508–18.

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

[15] Greenland S. Adjustment of risk ratios in case-base studies (hybrid epidemiologic designs). *Stat Med* 1986;**5:** 579–84.

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

[17] Pearce N. Effect measures in prevalence studies. *Environ Health Perspect* 2004;**112:**1047–50.

[18] Leigh PN, Abrahams S, Al-Chalabi A, Ampong MA, Goldstein LH, Johnson J *et al*. The management of motor neurone disease. *J Neurol Neurosurg Psychiatry* 2003;**74:**32–47.