J Rehabil Med 2009; 41: 217–222

SPECIAL REPORT

WHEN IS A CASE-CONTROL STUDY A CASE-CONTROL STUDY?

Nancy E. Mayo, PhD and Mark S. Goldberg, PhD

From the Department of Medicine, Division of Clinical Epidemiology, McGill University, Montreal, Quebec, Canada

Rehabilitation professionals rarely ask questions about the etiology of health events or outcomes and may not have formal training or relevant experience in the design of studies whose intent is to identify causal factors. The case-control study, which is one design used to answer questions about etiology, is particularly difficult to understand and research has shown that this study design label is often used incorrectly. This paper outlines the main features of case-control studies, with a particular focus on sampling strategies. The goal is to educate clinical rehabilitation colleagues about the fundamental principles of this powerful epidemiologic design. Examples illustrate how the parameters of cumulative incidence, incidence-density, and prevalence are estimated and the effect of sampling strategy on these parameters. Also shown is how sampling strategy affects conclusions drawn about the effects of an exposure on outcome. Even when used appropriately, case-control studies are methodologically complex to design and analyze to ensure an unbiased answer to the research question. The hypothetical and real-life examples given here could be used as course material to educate rehabilitation researchers.

Key words: rehabilitation, methodology, case-control studies, incidence, prevalence, incidence-density, odds ratio.

J Rehabil Med 2009; 41: 217–222

Correspondence address: Nancy Mayo, Division of Clinical Epidemiology, MUHC – Royal Victoria Hospital Site, 687 Pine Ave. West, R4.29 Montreal, Quebec, Canada. E-mail: nancy.mayo@mcgill.ca

Submitted September 13, 2008; accepted January 8, 2009

INTRODUCTION

In the first paper of this dyad on case-control studies (1), compelling evidence was presented that many clinical research disciplines have difficulty in distinguishing a case-control study from other types of designs. For example, of 221 research articles labeled as case-control studies in a sampling of medical research journals, 34% were found to be mislabeled. However, what was more distressing from the perspective of rehabilitation research, 97% of 86 articles from rehabilitation literature were found to be mislabeled. The purpose of this second paper is to outline the main features of case-control studies, with a particular focus on sampling strategies. The overall aim of the exercise is to educate clinical rehabilitation colleagues about the fundamental principles of this powerful epidemiologic design.

Case-control studies demystified

As discussed in the first paper, case-control studies are one of 2 designs to identify factors hypothesized to be causally associated with the outcome. Both cohort and case-control studies use rigorous statistical sampling strategies to ensure that the study subjects are representative of the target population. Fig. 1 is a schematic representation of these 2 designs, which differ only in the manner in which exposure and outcome are ascertained. In cohort studies, the sample is drawn from a population known to be free of the outcome of interest but with different values for the exposure; the outcome is ascertainedor determined after exposure has occurred. In case-control studies, subjects are sampled from the target population according to outcome status: those with the outcome of interest are referred to as cases, those without the outcome are referred to as controls; both cases and controls are sampled. In a case-control study, the exposure is ascertained or evaluated for a time period before the outcome. Implicit in the case-control design is that there is an underlying cohort from which the cases and controls are sampled. From a purely statistical perspective, the 2 designs are equivalent and make the argument that an exposure causes outcome.

Calendar time can be different in both of these designs. In some cohort studies, exposure status is ascertained in the present, and persons with or without exposure are followed into the future to ascertain outcomes. This is referred to as a prospective cohort study. Studies of outcomes that occur a long time after exposure, for example the occurrence of stroke following development of hypertension, may take decades to complete. In the other type of cohort study (optimally termed historical, but commonly, although less ideally, termed retrospective), the cohort is assembled based on exposure to a par-

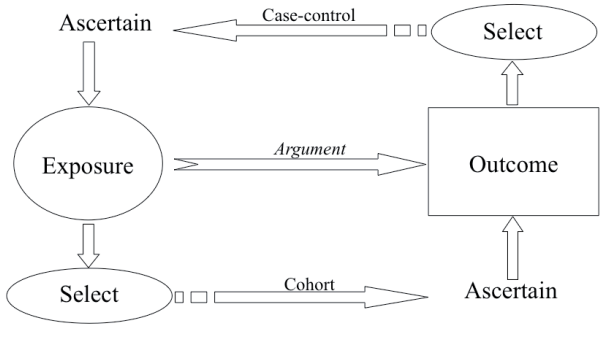

Fig. 1. Cohort and case-control studies

© 2009 The Authors. doi: 10.2340/16501977-0341

Journal Compilation © 2009 Foundation of Rehabilitation Information. ISSN 1650-1977

ticular factor in the (distant) past and the outcomes of interest are ascertained during the follow-up period, which ends at the present time. In case-control studies, ascertainment of exposure is always for a time in the past and this retrospective view can introduce bias, especially if cases and controls are asked to recall past events or exposures.

Untangling features of basic research designs

The mislabeling identified in the first paper of this dyad indicates several areas of misunderstanding of basic research designs. The key concepts that need to be understood to design a study to produce an unbiased answer to a question of etiology are the measurement of outcomes (distinguishing incidence and prevalence) and exposures, and the sampling procedure. These principles are best understood by thinking about how to conduct the "perfect" cohort study and then deciding which sampling methods could be used to reduce costs and increase efficiency.

Fig. 2 sets up a hypothetical cohort of 10 individuals, labeled A to J. Lines ending in points are persons who develop the outcome under study (there are 6 in all, labeled A, D, E, G, H and J); the point indicates when they develop the outcome. Lines ending in circles are persons who do not develop the outcome during the study period (there are 4: B, C, F and I). The length of the line shows how long each person was in view. Table I shows the different parameters that can be estimated from the cohort approach. *Incidence* is the number of new cases ($n=6$) that accumulate over the study period (20 months). The *cumulative incidence proportion* is 6/10 or 60%. The *incidence rate* is the number of new cases divided by the person-time in view. Person-time can be considered as a density. This estimator considers that people are not all followed for the same amount of time and hence should not contribute equally to the denominator. Fig. 2 can be used to calculate the density represented by person and time in months (person-time): A = 10; B = 4; C = 14; D = 8; E = 4; F = 12; G = 4; H = 18; I = 10; and J = 6 for a total of 100 person-months. The *incidence-density rate* is 6 per 100 person-months.

As the questions answered by cohort or case-control studies relate to the effect of an exposure on outcome, the incidence rate in the exposed needs to be compared with the incidence rate in the unexposed. Table I shows these calculations, assuming that the exposed subjects are A, B, C, D and E and the unexposed subjects are F, G, H, I and J. The ratio of the

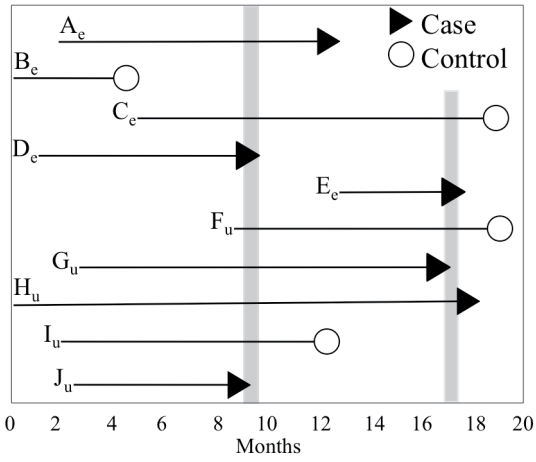

Fig. 2. Cross-sectional sampling within a cohort. Horizontal lines indicate people, labeled A to J; e and u indicate whether the person was exposed or unexposed; 2 cross-sectional samples made at 10 and 18 months are shown in shaded boxes.

2 cumulative incidence proportions (exposed vs unexposed) is 1.0 and the incidence-density rate ratio is 1.5; these are not identical, as the latter estimator is more informative because of the added information on time in view. Note that confidence intervals and other statistical testing details are omitted for the sake of focusing on the understanding of what a case-control study is and is not.

The above shows how these key parameters are estimated in a cohort study. It is informative now to see the effect of changing the sampling strategy from taking all subjects in a cohort to taking a sample at one point in time (i.e. a cross-sectional study). Cross-sectional studies provide estimates of prevalence but cannot be used to estimate incidence explicitly as subjects are not followed-up in time.

During a cross-sectional study at 10 months (shown in grey in Fig. 2), 8 subjects were recruited (A, C, D, F, G, H, I and J); 2 persons were not sampled, B was lost prior to the start of the study and E was not in view at the start of the study. Two persons (D and J) were identified as having the outcome for a prevalence proportion of 2/8 (25%). Three people (A, G and H) developed the outcome after the study ended and 3 never did (C, F and I).

A different estimate of prevalence would be obtained depending on when the sampling started. Imagine if the sampling

Table I. *Calculations of key parameters from hypothetical cohort and cross-sectional studies shown in Fig. 2*

	Cohort study		Cross-sectional studies	
	Cumulative Incidence	Incidence density	Prevalence	Prevalence
Time period (months)	0–20	0–20	10	18
Denominator	Persons ($n=10$)	Person-months ($n=100$)	Flagged at study start ($n=8$)	Flagged at study start ($n=5$)
Cases identified, *n*	6 (incident)	6 (incident)	2 (prevalent)	3 (prevalent)
Rate	6/10 or 60%	6/100 or 6/100 pm	2/8 or 25%	3/5 or 60%
Exposed (A, B, C, D, E)	3 (A, D, E)/5 or 60%	3/40 or 7.5/100 pm	1 (D)/3 or 33%	1 (E)/2 or 50%
Unexposed (F, G, H, I, J)	3 (G, H, J)/5 or 60%	3/60 or 5/100 pm	1 (J)/5 or 20%	2 (G, H)/3 or 67%
Rate ratio	1.0/1.0 = 1.00	7.5/5.0 = 1.50	33% /20% = 1.65	50%/67% = 0.75

pm: person-months; A-J: the 8 persons recruited.

was now done at 18 months (shown as the second grey line in Fig. 2). There are 5 persons in the sample for this time point (C, E, F, G and H); 5 persons (A, B, D, I and J) were lost prior to the study start and were not sampled. Three (E, G and H) of the 5 in view had the outcome for a point prevalence of 60%.

The major flaw with the cross-sectional design is that the effect of exposure on outcome cannot be estimated as the incidence rate ratio is inestimable. Only the prevalence proportion ratio can be estimated, which, at 10 months is 1.65 (33%/20%), larger than the incidence density rate ratio of 1.5, which is an unbiased estimate of the effect of exposure on outcome. In this case, the incidence density ratios and the prevalence proportion both point to a higher risk of the outcome for the exposed in comparison with the unexposed and would identify the exposure variable as a risk factor for the outcome.

The prevalence proportion ratio at month 18 is 0.75, which is the ratio of the rate in the exposed (50%) to the rate in the unexposed (66%). This ratio is less than 1.0, suggesting that exposure is protective against the outcome, which is clearly the wrong answer.

If no explicit cohort has been defined, then a cross-sectional study has basically sampled persons at a time of convenience and subjects may not be representative of the target population. The bottom line is that valid *causal* inferences cannot be made in cross-sectional studies. As shown in the first paper in this dyad (1), the most frequent type of study misclassified as a case-control study in the rehabilitation literature was a cross-sectional study (55 of 86 studies). In our teaching experience of graduate students, both in rehabilitation and in epidemiology, we have observed that students have difficulty distinguishing the salient features of these 2 designs.

With these issues in mind, incidence vs prevalence and sampling, we will now turn to the case-control studies. Case-control and cohort studies are linked, as shown in Fig. 1. In the case-control study, incident (i.e. new) cases (people with the outcome) from a defined location are sampled over a specified period of time, driven by the number of cases that need to be identified to achieve an acceptable level of statistical power. The "exposure profiles" of these cases are compared with the exposure profiles of the control subjects sampled from among those who do not have the outcome and who are representative of the population from which the cases arose. The major advantage of the case-control study is that data collection is much more efficient when the outcome is rare. While data are col lected on all of the cases, data are not collected on all potential controls but only on a random sample of potential controls. Most case-control studies use criteria to define the selection of cases and controls. The aim is to have controls that are as similar as possible to the cases on important variables, except the exposure. This is termed matching; more than one control can be selected per case to increase the power of the study. The following example illustrates the unique features of a case-control study.

Illustration of a case-control study to identify causes of falls

In 1989, Mayo et al. (2) published a study in the *American Journal of Physical Medicine and Rehabilitation* the objective of which was to identify factors predicting falls in a rehabilitation hospital. The target population for this study was persons undergoing in-patient physical rehabilitation and, hence, it was appropriate to sample from within this rehabilitation setting. Sampling was from an explicit cohort – persons who were admitted to this hospital – and all members of the cohort could be identified. This study was undertaken solely to investigate an "outbreak" of falls in one particular rehabilitation hospital. The factors identified and solutions implemented would be relevant only to this setting; other similar institutions could learn from this study but would be advised to identify their own specific factors. The inference from this study was to other years, but the proof of the translation would be a reduction in falls in subsequent years once the fall prevention strategies were implemented. Often, studies performed in rehabilitation settings aim to identify factors associated with outcomes and in many instances the very factor under study is one of the deciding factors for accessing rehabilitation services. If the factor potentially has a negative impact on outcome regardless of services accessed, then to gain entry into rehabilitation the persons would likely need to have other positive factors to balance out the negative. Hence the outcome may be more positive than expected. To arrive at an unbiased estimate of a prognostic factor for functional outcome or recovery would require sampling both admitted and non-admitted persons. Researchers working in rehabilitation settings need to word their research questions carefully to avoid implying the role of the factor in prognosis in general; it would also be important to recognize and address the potential for bias when carrying out research in a referral centre.

In the study on predicting falls (2), over a 2-year period, there were 1805 admissions to the targeted rehabilitation hospital, and 356 people (20%) fell. Thus, the cohort, the 1805 admissions, is explicit in this study, but detailed exposure assessment was made only for the cases and a matched group of controls, making this design efficient in terms of data collection. When people had more than one fall, only the first fall was analyzed.

Persons with falls (cases) were identified from incident reports, required by law to be completed and kept, for every fall in or around the hospital, witnessed or not. The controls had: (*i*) to have been admitted to the hospital within one week on either side of the admission date of the case; (*ii*) to be of the same sex; (*iii*) to have been in the hospital at the time the case became a case; and (*iv*) not to have fallen themselves before the time the case became a case. Thus, cases and controls were matched on sex, time of admission, and the time of the event for the case. In the predicting falls study, only one control was selected for each case because there was sufficient power with 386 cases and 386 controls. If more than one control was eligible, one was selected at random. This type of sampling, known as incidence density sampling, is an unbiased method of selecting cases and controls from the source population.

Fig. 3 illustrates the sampling procedure for case-control pairs for the hypothetical cohort presented in Fig. 2. At the time that each case was identified, marked by a number at the end of the line, potential control subjects are all persons who

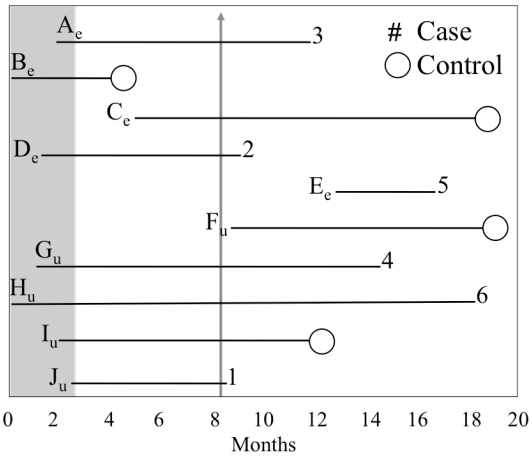

Fig. 3. Selection process of controls for the cases in the hypothetical cohort shown in Fig. 2. Case status is identified by a number in order of time. The study entry time for case 1 is shown in the shaded box**.**

had not yet fallen at that time and were, therefore, at risk of falling. A random selection of controls was made from among those individuals in the cohort who, at the time the case has occurred, had not yet developed the outcome (i.e. those still at risk). The control selection for cases is shown in Table II using the hypothetical cohort shown in Figs. 2 and 3 for illustration. Case 1, person J, had 6 possible control subjects who entered the cohort at a similar time (shown in the shaded box); person C entered later and was not eligible; person B left the cohort (without the outcome) prior to person J becoming a case and was not eligible. Case 2 (person D) entered close to the time of Case 1 and had 4 possible control subjects (A, G, H and I); B and C were not eligible and neither was J because of reaching case status before. Imagine a vertical line when each case became a case. The persons whose time course in the cohort intersects with the time the case became a case are eligible to be controls as long as they met the other criteria (cohort entry around the same time, still being in view, and not being a case previously). Table II shows all possible controls for each case. Also shown are the calculations for the matched odds ratio (OR) quantifying the association between exposure and outcome when one control is selected per case (remember persons A, B, C, D and E were exposed and persons F, G, H, I and J were unexposed). For the matched OR, only the discordant pairs provide any information upon which to judge the relationship between exposure and outcome (case-control status). For control selection 1, the OR is 2.0 and for control selection 2, the OR is 1.5, close to that estimated by the incidence density ration (IDR), which is 1.5 for the scenario shown in Fig. 2. In this example, we illustrate 2 random selections of controls to show that the results will differ depending on which control is selected, but the conclusion about the relationship between exposure and outcome is still the same. Because of the very small sample size, the confidence intervals around these estimates would be very large, but this example is for the purpose of illustration only.

One can take as many controls as necessary, and formal power calculations are used to help select an optimum number. Such a method is used in "nested case-control studies" of defined cohorts. Thus, at each time a case is identified, all individuals still at risk have an equal probability of being selected, and this selection strategy thus assures that the exposure distribution in the controls is representative of the study population (at that time). (It also follows

Table II. *Illustration of case-control sampling and calculation of matched odds ratio*

		Random control 1		Random control 2	
Case in order of time of event (i.e. fall)	Possible controls (matched on time of entry)	Choice	Exposure status Case/control	Choice	Exposure status Case/control
1 J	ADGHI	G	uu	A	ue
2 D	AGHI	A	ee	H	eu
3 A	GHI	I	eu	G	eu
4 G	H	H	uu	H	uu
5 E	F	F	eu	F	eu
6 H	C*	C	ue	C	ue
		Cases			
		Unexposed	Exposed		
	Controls 1				
	Unexposed	2	2		
	Exposed	1	1		
	OR (b/c)	2.0			
	Controls 2				
	Unexposed	1	3		
	Exposed	2	0		
	OR (b/c)	1.5			

*C is not an ideal control for subject 6 because the time of study entry was not close, but it is better to widen the window than exclude a case.
OR for matched analysis is b/c.
e: exposed; u: unexposed.

that whatever eligibility constraints are applied to the case subjects must also be applied to the control subjects.) This also implies that all subjects who were lost-to-follow-up or who died (censored) or developed the outcome before the case became a case would be ineligible to be controls. However, any subject who is not yet a case is eligible to be a control and that includes those subjects who become cases in the future. For example, in Fig. 3 and Table II, A is a control for D and later on A became a case.

This latter point often seems counter-intuitive, but at the time the subject became a case, the future status of the control, if he or she will become a case or not, is not known. This might make more sense if one considers a mortality study where the cases are people who died and controls are persons who are still alive, at the time the case becomes a case by dying. Of course, taking this to the extreme, all controls will eventually die and then become cases. Not stipulating that controls could become cases later on, and therefore eligible to serve as potential controls for a subject who fell, can lead to serious biases in the estimation of risk ratios (3).

Having selected the cases and controls, the next challenge in case-control studies is the accurate measurement of exposures that reflect the etiological time period (occurring before the outcome). In the study on predicting falls (2), the average time to the first fall was more than 3 weeks. Thus, there were many days during which the exposures of interest could have occurred, and ascertaining multiple exposures for these days for the 386 cases and 386 controls would have been a daunting task. Instead, data collection focused on: (*i*) the 24-h period preceding the recorded time of the fall; (*ii*) the 7-day period before the day and recorded time of the fall; and (*iii*) status at admission. A table in the article shows the data collected for each of these 3 time periods, which were identical for the cases and the controls. Having the same time period was important because environmental factors, such as wet floors, have been implicated as causes of falls. To be fair, both cases and controls need to be equally at risk for encountering the wet floor.

At admission, the focus was on diagnosis, vision, hearing, orientation, verbal comprehension and physical and emotional function. The focus for both cases and controls during the 7-day period before the fall was on change in functional status as well as any changes or events occurring in the persons' lives, such as a medical procedure or a birth or death in the family. The focus extended to physiological data on blood pressure, temperature and medications during the 24-h period preceding the recorded time of the fall.

Once the cases and controls have been identified and data collected on the exposures of interest as well as any important confounding variables, case-control studies are analyzed using a technique called conditional logistic regression. The term conditional refers to a matched analysis because the controls are not a random sample of the population but their inclusion is conditional upon who are the cases (because of the matching). Logistic regression is used because the outcome is dichotomous (binary), "case" or "control", and the exposures can be on any measurement scale: continuous for age, blood pressure, and temperature; ordinal for functional status (e.g. walks independently without an aid, walks independently with an aid, walks with help of one person, does not walk); or dichotomous for being on a hypnotic or not, or using eye medication or not. This analysis yields the OR associated with a particular variable; the OR is interpreted as the relative risk in the case of rare outcomes. Table III, recast from the original publication, presents the results of the matched analysis after adjusting for confounding variables.

Table III. *Factors identified as predicting falls*

Variable	Cases	Controls	Adjusted OR
Stroke at admission			
Yes	124	50	3.99
No	232	306	
Incontinence week prior to admission			
Ever	137	56	2.80
Never	219	300	
Anticonvulsant 24 h prior to admission			
Yes	36	19	2.98
No	320	337	
Topical eye preparation 24 h prior to admission			
Yes	33	17	3.39
No	323	339	

Table reworked from reference (2); OR is adjusted for all other variables in the table using conditional logistic regression.
OR: odds ratio.

This example should illustrate that case-control studies are methodologically difficult to design and can be quite complex to analyze, but most importantly, these studies answer questions about potential etiological factors, as they are just cohort studies in disguise. The results of the sudy on predicting falls (2) indicated that the factors increasing the risk of falls (hence, potentially causally related to falls through one or more paths) at the institution under study were: an admission diagnosis of stroke; incontinence in the week prior to the fall; use of anticonvulsants; and use of topical eye preparations during the 24 h period to the fall. A full prospective cohort study including all eligible subjects would have yielded approximately the same answer, but with less statistical uncertainty because of the increased sample of controls. However, the effort of data collection would have been insurmountable.

CONCLUSION

Rehabilitation professionals rarely ask questions about etiology and hence may not have had much exposure to the methodological features of designs that answer questions about etiology. This was apparent in the high degree of mislabeling of case-control studies in the rehabilitation literature sampled in the first paper of this dyad (83 of 86 studies were found to be mislabeled) (1). The hypothetical and real examples given here should help to illustrate the unique features of case-control studies and could be used as course material to educate rehabilitation researchers. There are also many textbooks on research design that provide more in-depth material (4–7).

REFERENCES

1. Mayo NE, Goldberg MS. When is a case-control study not a case-control study? J Rehabil Med 2009; 41: 209–216.
2. Mayo NE, Korner-Bitensky N, Becker R, Georges P. Predicting falls among patients in a rehabilitation hospital. Am J Phys Med Rehabil 1989; 68: 139–146.
3. Lubin JH. Case-control methods in the presence of multiple failure times and competing risks. Biometrics 1985; 41: 49–54.
4. Hulley SB, Cummings SR, Brown WS, Grady DG, Newman T, editors. Designing clinical research: an epidemiologic approach. Baltimore: Wolters Kluwer Lippincott Williams & Wilkins; 2006.
5. Bailar JC, Louis TA, Lavori PW, Polansky M. A classification for biomedical research reports. In: Bailar JC, Mosteller F, editors. Medical uses of statistics. 2nd edn. Boston: Massachusetts Medical Society; 1992, p. 141–156.
6. Schlesselman JJ, editor. Case-control studies: design, conduct, analysis. New York: Oxford University Press; 1982.
7. Rothman KJ, Greenland S, editors. Modern epidemiology. 2nd edn. Philadelphia: Lippincott-Raven Publishers; 1998.