Published by Oxford University Press on behalf of the International Epidemiological Association *International Journal of Epidemiology* 2012;**41**:1480–1489

doi:10.1093/ije/dys147

Case–control studies: basic concepts

Jan P Vandenbroucke[1]* and Neil Pearce[2,3]

[1]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands, [2]Faculty of Epidemiology and Population Health, Departments of Medical Statistics and Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK and [3]Centre for Public Health Research, Massey University, Wellington, New Zealand

*Corresponding author. Department of Clinical Epidemiology, Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, The Netherlands. E-mail: J.P.Vandenbroucke@lumc.nl

Accepted 7 August 2012

The purpose of this article is to present in elementary mathematical and statistical terms a simple way to quickly and effectively teach and understand case–control studies, as they are commonly done in dynamic populations—without using the rare disease assumption. Our focus is on case–control studies of disease incidence ('incident case–control studies'); we will not consider the situation of case–control studies of prevalent disease, which are published much less frequently.

Keywords Education corner, methods, case–control studies, teaching, rare disease assumption, dynamic population

Introduction

Readers of the medical literature were once taught that case–control studies are 'cohort studies in reverse', in which persons who developed disease during follow-up are compared with persons who did not. In addition, they were told that the odds ratio calculated from case–control studies is an approximation of the risk ratio or rate ratio, but only if the disease is 'rare' (say, if <5% of the population develops disease). These notions are no longer compatible with present-day epidemiological theory of case–control studies which is based on 'density sampling'. Moreover, a recent survey found that the large majority of case–control studies do not sample cases and control subjects from a cohort with fixed membership; rather, they sample from dynamic populations with variable membership.[1] Of all case–control studies involving incident cases, 82% sampled from a dynamic population; only 18% of studies sampled from a cohort, and only some of these may need the 'rare disease assumption' (depending on how the control subjects were sampled). Thus, the 'rare disease assumption' is not needed for the large majority of published case–control studies. In addition, different assumptions are needed for case–control studies in dynamic populations and those in cohorts to ensure that the odds ratios are estimates of ratios of incidence rates.

The underlying theory for case–control studies in dynamic populations has been developed in epidemiological and statistical journals and textbooks over several decades,[2–19] and its history has been described.[20] Still, the theory is not well known or well understood outside professional epidemiological and statistical circles. Introductory textbooks of epidemiology often fall back on methods of control sampling, which involve the 'rare disease assumption' as it was proposed by Cornfield in 1951,[3] because it seems easier to explain.[1] Moreover, several advanced textbooks or articles depict the different ways of sampling cases and control subjects from the point of view of a cohort with fixed membership.[13,18] This reinforces the view of case–control studies as constructed within a cohort, even though this applies to only a small minority of published case–control studies.

The purpose of this article is to present in elementary mathematical and statistical terms a simple way to quickly and effectively teach and understand case–control studies as they are commonly done in dynamic populations—without using the rare disease assumption. Our focus is on case–control studies of disease incidence ('incident case–control studies'); we will not consider the situation of case–control studies of prevalent disease, which are published much less frequently,[1] except in certain situations as discussed by Pearce[21] (e.g. for diseases such as asthma in which it is difficult to identify incident cases).

The theory of case–control studies in dynamic populations cannot be explained before first going back to the calculation of incidence rates and risks in

dynamic populations. In a previous article, we have reviewed the demographic concepts that underpin these calculations.[22] In the current article, these concepts will first be applied to case–control studies involving sampling from dynamic populations. Second, we discuss how to teach the theory in the situation of sampling from a cohort. In the third part, it is explained how these two distinct ways of sampling cases and control subjects can be unified conceptually in the proportional hazards model (Cox regression). Finally, we discuss the consequences of this way of teaching case–control studies for understanding the assumptions behind these studies, and for appropriately designing studies. We propose that the explanation of case–control studies within dynamic populations should become the basis for teaching case–control studies, in both introductory and more advanced courses.

Case–control studies in dynamic populations

Basic teaching

To understand the application of the basic concepts of incidence rate calculations to case–control studies, we start with the demographic perspective of a dynamic population in which we calculate and compare incidence rates of disease.[22]

Suppose that investigators are interested in the effect of oral contraceptive use on the incidence of myocardial infarction among women of reproductive age. They might investigate this in a large town in a particular calendar year (we base this example loosely on one of the first case–control studies that investigated this association[23]). The time-population structure of the study is depicted in Figure 1.

In Figure 1, for the sake of simplicity, imagine that, on average, 120 000 young women of reproductive age (between ages 15 and 45 years) who have never had coronary heart disease (CHD), are living in the town, on each day during the calendar year of investigation. This is a dynamic population: each day, new young women will become 15 years old, others will turn 46, some will leave town and others will come to live in the town, some will develop CHD and be replaced by others who do not have the disease and so forth. Such a population can be safely regarded as being 'in steady state'. The demographic principle of a steady-state population was explained in our previous article;[22] in brief, it assumes that over a small period, e.g. a calendar year, the number of people in a population is approximately constant from day to day because the population is constantly depleted and replenished at about the same rate. It was also explained why this assumption holds, even if the population is not perfectly in a steady state.[22] Thus, we take it that each day of the year, ~120 000 women of reproductive age, free of clinically recognized CHD,

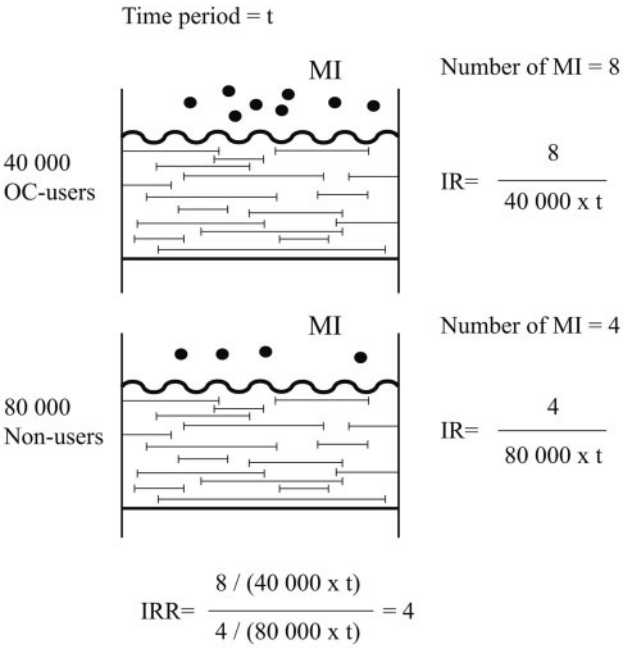

Figure 1 The underlying dynamic 'source' population of a study of myocardial infarction (MI) and oral contraceptive use. The bold undulating lines show the fluctuating number of users and non-users of oral contraceptives in a population that is in a steady state. The finer lines below it depict individuals who enter and leave the populations of users and non-users. Closed circles indicate cases of MI emanating from the population. For users and non-users separately, an incidence rate (IR) of MI can be calculated. The incidence rate ratio (IRR) can be used to compare the incidence of MI between users and non-users. In the description of the example in the text, the time t was set to one calendar year. Figure adapted from Miettinen[9]

live in the town. Suppose that, on average, 40 000 women use oral contraceptives and 80 000 do not. Again, these are two dynamic subpopulations that can be regarded as being in a steady state. Women start and stop using oral contraceptives for various reasons and switch from use to non-use and back again. As such, in one calendar year, we have 40 000 woman-years of pill use and 80 000 woman-years of non-use, free of CHD.

Suppose that a group of investigators surveys all coronary care units in the town each week to identify all women, aged 15–45 years, admitted with acute myocardial infarction during that period. When a young woman is admitted, the investigators enquire whether she was on the pill—and whether she had previously had a coronary event (if she had, she is excluded from the study). Suppose that, in total, 12 women were admitted for first myocardial infarction during the year of study: eight pill users and four non-users. That produces an incidence rate of 8/40 000 woman-years among pill users and 4/80 000 woman-years among non-users. The ratio of these incidence rates becomes (8/40 000 woman-years)/(4/80 000

woman-years), which is a rate ratio of 4, indicating that women on the pill have an incidence rate of myocardial infarction that is four times that of those not on the pill.

Transformation to a case–control study

In total, 12 cases arise from the population: eight users and four non-users. Those are the potential cases for a case–control study in which the investigators would survey all coronary care units each week of the year. Suppose that the investigators, as their next step, would take a random sample of 600 control subjects from the total source population of the cases (the total of 120 000), by asking 600 women aged 15–45 years, without previous CHD, whether they are 'on the pill' at the time the question is asked. Then, on whatever day of the year, this sample of control subjects will include, on average, 200 users and 400 non-users of oral contraceptives. These numbers represent the underlying distribution of woman-years of users and non-users. Together with the cases, this is the complete case–control study (see Table 1).

From Table 1, an odds ratio can be calculated as $(8 \times 400)/(4 \times 200)$. This exactly equals the ratio of the incidence rates in the underlying population. Algebraically: the incidence rate ratio from the complete dynamic population, which we calculated earlier, can be easily rewritten as (8/4)/(40 000 woman-years/80 000 woman-years). Between parentheses in the numerator of this formula is the number of pill users divided by the number of non-users among all women newly admitted with CHD (= cases in the case–control study). In the denominator, we find the proportion of woman-years on the pill divided by the proportion of woman-years of non-use. It is immediately obvious that—if the steady-state assumption holds—we can estimate the latter proportion directly from the sample of 600 women (= control subjects in case–control study). Among the 600 control subjects, the ratio of exposed to unexposed is expected to be the same as the ratio of the woman-years—except for sampling fluctuations. Thus, what we do in a case–control study is to replace the denominator ratio (40 000 woman-years /80 000 woman-years) by a sample (200/400). We still obtain, on average, the same rate ratio of 4. It follows that to estimate the rate ratio, we do not have to measure, nor to estimate, all the person-years of pill-using and non-using women in town; we can simply determine the ratio of those woman-years by asking a representative sample of women free of CHD from the population from which the cases arise, about their pill use. The complete dynamic population is called the 'source population' from which we identify the cases and the sample of control subjects, and the period over which cases and control subjects are identified is the 'time window' of observation, also called the 'risk period'.

The 'odds ratio' which is calculated from Table 1 is technically also known as the 'exposure odds ratio', as it is the 'odds of exposure' in the cases divided by the 'odds of exposure' in the controls: $(8/4)/(200/400) = 4$, the same as the ratio of incidence rates in the whole source population. The great advantage of case–control studies is that we can calculate relative incidences of disease in a population, by collecting all the data for the numerator (by collecting cases in hospitals or registries where they naturally come together), and sampling control subjects from the denominator, i.e. sampling 'control subjects' to estimate the relative proportions (exposed vs non-exposed) of the person-years of the exposure of interest in the source population. Thus, one achieves the same result as in a comprehensive population follow-up, at much less expense of time and money. Just imagine the effort of having to do a follow-up study of all 120 000 women of reproductive age in town, also keeping track of when they move in and out of town and constantly updating their oral contraceptive use in a particular calendar year!

Table 1 Layout of case-control data sampled from dynamic population: study of occurrence of myocardial infarction in users vs non-users of oral contraceptives, corresponding to Figure 1

	Myocardial infarction	Control subjects
Oral contraceptive use		
Yes	8	200
No	4	400
Odds ratio	4	

Advanced teaching

Cohorts vs dynamic populations

For researchers who are used to think in terms of clinical cohorts, it can be difficult to understand that populations are not depleted: is it not true that the people with a particular risk factor will develop some disease more often, and thus in the course of time, there will be less of them who are still candidates for developing the disease? That will be true in cohorts because their membership is fixed, but not in dynamic populations. One way to understand this is to think of genetic exposures. People with blood group O develop clotting disorders more frequently, whereas people with blood group A develop more often gastric cancer. However, in a dynamic population, the numbers of people with blood group O or A are not constantly depleted—blood group distribution is fairly constant over time, as new people are born with these blood groups so that an equilibrium is maintained.[22]

Another way to understand this concept is to think about an imaginary town and the cases of myocardial infarction that are enrolled in a study. For the aforementioned discussion, we assumed that we were studying all women living in a town during some time over the course of one calendar year (this could

be the whole year or a few months). The situation would be entirely different if we restricted our study to all women who lived in the town on the 1 January of that year: then we would only count the myocardial infarctions that happened during this year in women who had been living in town on the 1 January; indeed, the number of women on the pill might decline more than the number of women not on the pill because the myocardial infarctions predominantly occur in the users. That situation would be akin to a clinical cohort study, i.e a study with fixed membership defined by a single common event.[22] However, in a dynamic population, a myocardial infarction that happens in a woman who moved into town during the year also counts in the numerator; she and the other women who move into town replenish the denominator because other women move out. By and large, as with blood groups, the population denominator remains constant in terms of its exposure distributions: the woman-years of oral contraceptive use vs non-use. If the population is truly in steady state, it does not matter when the control subjects are sampled—at the beginning, at the end or at the halfway point of the calendar period (the time window or 'risk period').

To refine the concept, the members of a dynamic population do not necessarily have to be present for long periods in the population—as might be surmised from the examples about towns and countries of which one is either an inhabitant or not, and usually for several years. Members of a dynamic population may also switch continuously between being in and out of the population.[22] Take a study on car accidents and mobile phone use by the driver. The risk periods of interest are the periods when people drive. The exposure of interest is phone use. In a case–control study, car accidents are sampled, and it is ascertained (say, via mobile phone operators) whether the driver was phoning at the time of the accident. Control moments might be sampled from the same driver (say, in the previous week) or from other drivers, by sampling other moments of time when they were driving; for each of these control moments, it might be ascertained, via the same mechanism as for the cases, whether they were phoning while driving. These control moments are contrasted with the moment of the accident (the case). If the same driver is used as his or her own control, this type of case–control study is called a 'case–crossover study'.[24] From the example, it can be understood readily that such a case–control study compares the incidence rate of accidents while driving and phoning vs the incidence rate of accidents while driving and not phoning.[25]

What if the exposure distribution of the population is not in steady state?

But what if the exposure distribution in the population is not in steady state? For example, suppose that one wants to investigate in a case–control study whether two different types of oral contraceptives give a different risk of venous thrombosis: 'third-generation oral contraceptives' vs 'second-generation oral contraceptives' (this was once a real and hotly debated question[26]). Suppose further that the newer 'third-generation oral contraceptives' are strongly marketed, and that their market share clearly increases in the course of the calendar year. That situation is depicted in Figure 2.

There are two solutions:

(i) Sample the control subjects in the middle of the period when the cases accrued, and thereby use the additional assumption that the rise (or fall) of the use of a particular brand of pill is roughly linear over the risk period. Then the control subjects will still represent the average proportion of person-years over the risk period. This is depicted in Figure 2 and is the same solution as is used to calculate person-years (i.e. the denominator) when populations are not in steady state [see previous article on the calculation of incidence rates for explanation].[22] Alternatively, if one assumes that the incident cases in the dynamic population are evenly spread over time, one might sample control subjects evenly over time.

(ii) The more sophisticated solution is the one that researchers often use spontaneously: they sample a (number of) control subject(s) each time there is a case, which amounts to 'matching on calendar time'. Then the control subject(s) will reflect the underlying population distribution of exposure at each point in time a case occurs, and any assumption about linearity is not needed. This is the most exact solution and is represented in Figure 3. Matching on calendar time can be done in two ways: (i) invite the control subject(s) around the same calendar date as the case and ask them about their exposure (at that time or at previous times if exposure has a lag time to produce disease); or (ii) if control subjects are invited at a later point in time, present them with an 'index date', which is the date as the event of the matching case, and question them and/or measure their exposures for that index date. If control subjects are matched on calendar time, then it is appropriate to take the time matching (and, of course, any other matching factors) into account in the analysis, or at least to check whether it is necessary to control for them.

Hospital-based case–control studies

In most examples presented earlier, the patients are assumed to be sampled from a defined geographical population (via disease registries or by having access to all hospitals of some region), and control subjects are sampled from the underlying dynamic population

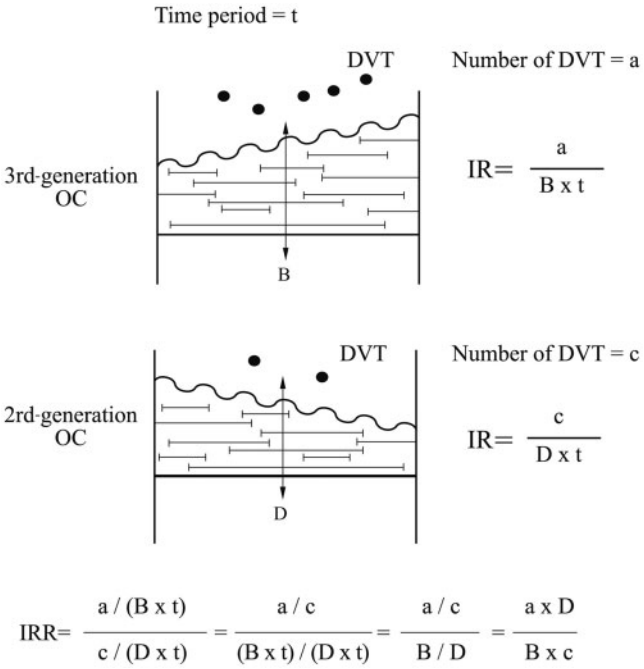

Figure 2 Sampling from the middle of the 'risk period' when the exposure distribution is not in steady state. The bold undulating lines show the increasing use of one type of oral contraceptives and the decreasing use of the other type during the time period (risk period). The finer lines below it depict individuals who enter and leave the populations of users of these types of oral contraceptives. Closed circles indicate cases of deep venous thrombosis (DVT) emanating from the population. B and D represent the numbers of users of one type or the other contraceptive at a cross-section in the middle of the time period. Incidence rates (IRs) of DVT can be calculated for both populations separately, and an incidence rate ratio (IRR) can be used to compare these two incidence rates. In a case–control study, B and D are estimated by 'b' and 'd', the numbers of users of one type or the other type of oral contraceptives in a sample from the source population taken in the middle of the period. The algebraic redrafting of the IRR shows that a ratio of IRs is algebraically equivalent to an 'exposure odds ratio' or the 'cross-product' that is obtained in a case–control study

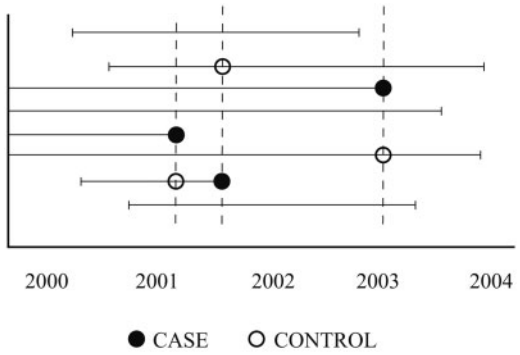

Figure 3 Case–control sampling in dynamic populations when a control is sampled each time a case occurs: matching on calendar time. Persons move in or out of the population by mechanisms such as birth or death, or move in or out from this population to another. Person-time is indicated by horizontal lines. The time axis is calendar time. The sampling of the control subjects is 'matched on calendar time': each time a case occurs, one or more control subjects are sampled. Cases and control subjects can be either exposed or unexposed (not shown here). A person who will become a case can be a control subject earlier, and multiple control subjects or even a variable number of control subjects can be drawn for each case

of this geographical area. If cases from a case–control investigation are sampled from one or more hospitals that do not reflect a well-defined geographic population, still each hospital has a 'catchment population', consisting of the patients who will be admitted to that hospital when they develop a particular disease. Such a catchment population can be seen as a dynamic population, with inflow and outflow depending on patient and referring doctor preferences, religious or insurance affiliations, or on the reputation of a particular hospital for particular diseases and so forth. To obtain control subjects for such cases, the investigator should consider patients who are admitted to the same hospital and come from the same catchment population—meaning that if they had developed the case disease, they would have been admitted to that same hospital. This approach obviously has some risks in that the control disease may be associated with the exposure that one wants to study; that risk can (it is hoped) be minimized by using a mix of control diseases, none of which is known to be associated with the exposure under study.[27] Still, the principle of sampling control subjects from a dynamic population remains the same, whether the controls are population-based or hospital-based.

The early case–control study on oral contraceptives and myocardial infarction, which inspired the example presented earlier, sampled cases from a number of coronary care units that were surveyed in one geographically defined hospital area in the UK; for each case interviewed, three women of the same age who were discharged after some acute or elective medical or surgical condition were similarly interviewed about their use of oral contraceptives.[23] Likewise, the first case–control studies on smoking and lung cancer were hospital-based, and control subjects were non-cancer patients being present in the same wards or the same hospital as the lung cancer patients.[2,28]

Case–control studies within cohorts

Doing a case–control study by sampling from a cohort with fixed membership is relatively rare—a recent survey found that it only occurs in 18% of published

case–control studies.[1] It is mostly done when investigators have data available from a cohort, and when it is too expensive to go back and assess the exposures of everybody in the cohort. For example, in an occupational cohort study, personnel records may be available for all cohort members from date of employment, but it may take a considerable amount of work to assess these work histories and estimate cumulative exposures to particular chemicals, whether by using a job-exposure matrix or by an expert panel assessment.[29]

Another example is the 're-use' of data or samples from a randomized controlled trial (RCT) for a subsequent investigation. For example, the data from the 'Physician's Health Study'[30] were re-used several years after the trial was finished for a new genetic case–control study; baseline blood samples of participants who developed cardiovascular end points in the trial were used, as were blood samples of matched participants in the trial who remained free of those diseases, and the frequency of one genetic factor (Factor V Leiden) was compared between these cases and control subjects. This investigation thereby considered the trial data as a single cohort in which new exposures were assessed, irrespective of the original randomization.

Figure 4 depicts a cohort with fixed membership from time 0. The cases accrue in the course of the follow-up in the exposed and unexposed part of the cohort. The available cohort data may only relate to exposure status at baseline (as in the aforementioned RCT example), but may also indicate changes in exposure over time, for example, if repeated measurements were done in the cohort study, or if time-related exposure information can be assessed from personnel records, prescribing records or other sources (as in the occupational example).

For each case, one or more control subjects are selected from the overall cohort, and the exposure statuses of the case and control subjects are determined at the time they are sampled. There are three options to sample control subjects:[12,13,18]

(i) As in the aforementioned RCT example, investigators often sample control subjects from the people who have still not developed the disease of interest at the end of follow-up (this is termed 'cumulative incidence sampling' or 'exclusive sampling'), and exposure status at beginning of follow-up is used for these cases and controls. As shown algebraically in many textbooks, in that situation, the odds ratio is exactly the same (on average) as the corresponding odds ratio from the full cohort study, and this will approximate the risk ratio or rate ratio (in the full cohort study) only if the disease is rare (say, <5% of exposed and non-exposed develop the disease). This is the 'rare disease assumption', as historically first proposed by Cornfield in 1951.[3] It can be seen from Figure 4 that if the disease is rare, even in the exposed (sub)cohort, the ratio of people with and without exposure among those without disease at the end of the follow-up will remain about the same as at the beginning of the follow-up, which is why the 'rare disease assumption' works.

(ii) An imaginative solution, first proposed by Kupper *et al.*,[8] is to sample control subjects

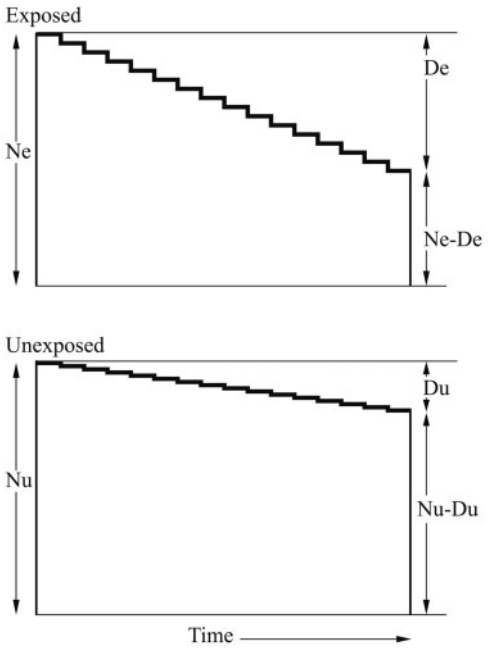

Figure 4 First two methods of case–control sampling from cohorts consisting of a subgroup of exposed and a subgroup of non-exposed persons: 'exclusive' sampling at end of follow-up, and 'inclusive' sampling at beginning of follow-up. The bold lines that go down stepwise represent the number of people who remain without the disease of interest; each step is a case of the disease. Total cohort consists of N = number of persons at beginning of follow-up, the sum of exposed and unexposed subgroups (Ne + Nu). The total number of persons who become diseased in exposed and unexposed subgroups is D, sum of exposed and unexposed diseased persons (De + Du). In a case–control study, cases are sampled from D, which is De and Du together. Control subjects are either sampled 'exclusive' from N − D, which is (Ne − De) and (Nu − Du) together, or 'inclusive' from N, which is Ne and Nu together. This leads to the following measures:

Measure	Definition	Alternative formulation
Odds ratio under exclusive sampling	$\frac{De/(Ne-De)}{Du/(Nu-Du)}$	$\frac{De/Du}{(Ne-De)/(Nu-Du)}$
Risk ratio under inclusive sampling	$\frac{De/Ne}{Du/Nu}$	$\frac{De/Du}{Ne/Nu}$

Figure refers to methods 1 and 2 in text under subheading 'Case–control studies within cohorts', and is adapted from Rodrigues *et al.* [13] and Szklo and Nieto [18]

from all those in the cohort at the beginning of follow-up instead of at the end ('case-cohort' or 'inclusive sampling'). At the beginning of the follow-up, all persons are still disease free (if they are not, then they would not have been included in the cohort). Then, the control subjects reflect the proportion exposed among the source population at the start of follow-up. Some of the control subjects who are sampled at baseline may become cases during follow-up. This seems strange at first sight, but it is not: if in a cohort study or an RCT, the risk is calculated, one uses all persons developing a disease outcome in the numerator, and divides by the denominator, which consists of all people who were present at start of follow-up, including those who will later turn up in the numerator. As can be seen from Figure 4, sampling from the persons present at the beginning of the follow-up makes the odds ratio from the case–control study exactly the same (on average) as the risk ratio from the full cohort study. This can be understood most easily if one imagines taking a control sample of 100%, that is, all persons present at the beginning: then the odds ratio in the case–control study will be exactly the same as the risk ratio from the cohort study. Next, if one imagines taking a 50% sample for the control subjects, the odds ratio will remain the same (on average). One complication with this method is the calculation of the standard error of the odds ratio, as some persons are both cases and control subjects; different solutions exist.[31] A further complication is that, just as with the estimation of risks (which this sampling scheme corresponds to), losses to follow-up for other reasons than developing the disease that is studied are not easily taken into account; such losses to follow-up may produce bias if they are substantial and differ between exposed and unexposed.

(iii) The third option is to sample control subjects longitudinally throughout the risk period (i.e. not just at the beginning or just at the end). Throughout the follow-up of a cohort, the numbers of both exposed and unexposed persons who are free of disease will decrease, and people may be lost to follow-up for other reasons. Moreover, persons may move between exposure categories. The 'royal road' is to sample one or more control subjects at each point in time when a case occurs ('density sampling', 'risk-set sampling' or 'concurrent sampling') and determine the exposure status of cases and control subjects at that point in time. This is depicted in Figure 5. By this sampling approach, the odds ratio from the case–control study will estimate the rate ratio from the cohort study. This is the equivalent of 'matching on time' in dynamic populations. This approach is most correct theoretically, but can only be used for cohorts when one has information about disease status of all persons at regular intervals during follow-up (e.g. when cancer incidence or mortality data are available over time).

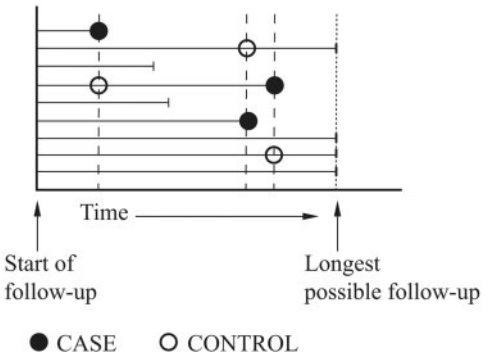

Figure 5 Third method of sampling from a cohort: longitudinal sampling, also called concurrent sampling, density sampling or risk-set sampling. Persons start follow-up at inclusion in the cohort (e.g. date of surgery) and are followed until either end point occurs (person becomes a case), or the last calendar day of the study. Persons are indicated by fine lines from start of follow-up onwards. The time axis is follow-up time from inclusion (time 0). The longest period of follow-up is by persons who enter the cohort on the calendar day that the study starts; persons entering later will have shorter follow-up because they will be withdrawn from the study at the last calendar day of the study. Cases and control subjects can be either exposed or non-exposed (not shown here). A person who will become a case can be a control subject earlier, and multiple control subjects, or even a variable number of control subjects, can be drawn for each case. In text, see method 3, under subheading 'Case–control studies within cohorts'

The first solution corresponds to the original theory proposed by Cornfield,[3] and requires the 'rare disease assumption' if the goal is to estimate rate ratio or risk ratios; it was the most frequently used method in case–control studies within cohorts in the past—and that approach was used in almost all case–control studies based on cohorts that were identified in the review by Knol *et al.*[1] Solution 2 still pertains to cohort thinking, but has an imaginative solution to calculate risk ratios; it is often called a 'case–cohort' study, and is particularly useful in studies in which a single control sample can be used for multiple case–control studies of various outcomes. Solution 3 is the more sophisticated development in case–control theory, in which the case–control odds ratio estimates the rate ratio from the cohort population over the follow-up period without the need for any rare disease assumption.[10,11] However, it is used relatively rarely.[1]

A note about terminology: the term 'nested case–control studies' seems to be mostly used to

denote case–control studies within cohorts which use the third sampling option. However, it is sometimes loosely used to denote all types of case–control sampling within a cohort.

Unity of the concept of density sampling from dynamic populations and sampling from cohorts

The last method of sampling (method 3) immediately points to a conceptual unity of 'incidence density sampling' or 'density sampling' in cohorts and in dynamic populations. This was described by Prentice and Breslow in 1978[10] and expanded by Greenland and Thomas in 1982.[11] It can be grasped intuitively by comparing Figures 3 and 5. The basis of the conceptual unity is that person-years can be calculated from cohorts and from dynamic populations, as was explained in our earlier article.[22]

In a case–control study in a dynamic population, investigators often use matching on calendar time spontaneously (a control is chosen each time a case occurs), which is an ideal way of sampling, as it produces an odds ratio that directly estimates the incidence rate ratio, as in Figure 3. In cohorts, however, one has to use sampling strategy 3, presented earlier, to estimate the incidence rate ratio, as in Figure 5. The latter necessitates advanced insight and is used infrequently. In advanced textbooks, the 'matching on time' in dynamic populations and the 'concurrent sampling' in cohorts are often mentioned together as 'density sampling'. This is theoretically correct, although it obscures the practicalities of the different sampling options.

'Density sampling' or 'risk-set sampling' from a cohort (i.e. the purer form of sampling of aforementioned strategy 3) involves sampling control subjects from the risk sets that are used in the corresponding Cox proportional hazards model.[10,11] A 'hazard' or 'hazard rate' is the name used in statistics for a peculiar form of 'incidence rate', wherein the duration of the follow-up approaches the limit of zero and becomes infinitesimally small; it is also called an 'instantaneous hazard'.[22] When follow-up time is small, there is no numerical difference between risks and incidence rates.[22] Intuitively, a proportional hazards model in a follow-up analysis of a cohort can be understood as comparing the exposure odds of all successive cases at each point in time with those of the non-cases who are still at risk at that point in time (some of whom may become cases later), that is, the 'risk set'. The exposure odds ratio or hazard ratio is then averaged over all of these comparisons, assuming it to be constant. Thus, a Cox proportional hazards model in a cohort becomes conceptually similar to a study that is 'matched' on time with a 'variable control-to-case-ratio' in a dynamic population. The estimation of the proportional hazard in a Cox model can be seen as an average of odds ratios over several risk sets; as the follow-up time in each risk set is small (say, the day of occurrence of the case disease), the odds ratios directly translate to relative risks and incidence rates, for reasons explained in the article on incidence calculations in dynamic populations.[21,22]

Discussion: differences with classic case–control teaching, and consequences

The main difference between the approach we have described in this paper and the classic view of case–control studies as a 'cohort study in reverse' is that the dynamic population view reflects how the large majority of case–control studies are actually done. They are not done within cohorts, neither real nor imaginary. Rather, most case–control studies have an underlying population that is dynamic: for example, the geographically defined source population of a disease registry, the catchment areas of a hospital region or people who are driving.

The first case–control studies on smoking and lung cancer were done using cases and control subjects admitted to hospital from vaguely defined catchment areas.[2,28] Doll and Hill showed in the discussion of their original case–control study on smoking and lung cancer how one might calculate back to the general population,[2] as they assumed that they had sampled from that population—an insight that was far ahead of their time because it did not need the 'rare disease assumption'. Although it originated during the period when Cornfield proposed his 'rare disease assumption', Doll and Hill's solution was largely forgotten. Only occasionally does one read back-calculations from case–control studies to the background or source population, perhaps because such back-calculations have intricacies of their own, for example, in the case of matching.[32]

An important consequence of primarily teaching case–control studies in dynamic populations, without the rare disease assumption, is that the real assumptions that are necessary for the majority of case–control studies become clear: either the exposure distribution should be in steady state in the dynamic population, or sampling of control subjects should be matched on time in a dynamic population (or equivalently, concurrent in the follow-up of a cohort).

An often-heard precept to guide the design of case–control studies is 'Think of an imaginary randomized trial when planning your case–control study'. This gives the impression of automatically assuming a cohort, as all randomized trials are cohorts with a fixed membership. However, randomized trials can be done equally well on dynamic populations—public health interventions are often on dynamic

populations. When the intervention or the exposure is studied in a case–control study with an underlying dynamic population, design features can be construed that are impossible or difficult in cohorts. For example, a dynamic population free of other key risk factors can be proposed: in a case–control study of the risk of oral contraceptives and venous thrombosis, an investigator might stipulate a dynamic population that has neither major surgery nor plaster casts after breaking legs and so forth—thus limiting the study to 'idiopathic cases'. That would be difficult in a cohort; for example, in an imaginary randomized trial on oral contraceptives, wherein the outcome would be venous thrombosis, it would seem strange to truncate follow-up at the time of major surgery or plaster cast. In a dynamic population, however, the population is constantly renewed, and this exclusion comes naturally and may have advantages in attributing causality because other major risk factors for the outcome are excluded.

It should be emphasized that when cases and control subjects are selected from a dynamic population (or by risk-set sampling from a cohort), exposures do not need to be assessed solely at the time cases and control subjects are selected (e.g. 'current use' of oral contraceptives). In many circumstances, investigators need information on the duration of exposure and/or cumulative exposure. For example, in studies of smoking, the effect on lung cancer only becomes clear after several years. In contrast, the cardiovascular adverse effects of hormone replacement therapy may be limited to the first year of use, so recent exposure is most relevant. Recent and historical exposures can be assessed by a variety of methods in case–control studies, ranging from subjective (e.g. questionnaires) to more objective methods (e.g. birth records, pharmacy records and work histories combined with historical exposure monitoring data). The exposure definition can be easily adapted, by defining as many time windows of exposure as is deemed necessary, for recent and for long-term exposure, because there is a continuous turnover between these categories over time in the underlying population.

In summary, case–control studies with incident cases can be conducted in two contexts—dynamic populations and cohorts—of which the first is the most commonly used[1] because it comes naturally to most investigations. This method should become the basis of teaching case–control studies—in both introductory and more advanced courses:

- Case–control studies can be conducted in a dynamic population, and the resulting odds ratio directly estimates the rate ratio from this dynamic population, provided that the control subjects represent the source population's distribution of person-time of exposure over the risk period. This can be achieved either by matching on time or by selecting control subjects more loosely from the same period, if the population is judged to be in steady state for the exposure(s) and other variables of interest.
- Case–control studies can also be conducted within a cohort; in this situation, control subjects can be sampled in three different ways, and the resulting odds ratio can estimate the odds ratio, risk ratio or rate ratio from the corresponding full cohort analysis.[21] Because such case–control studies are a minority, and the need for the rare disease assumption only applies for one method of sampling in such studies, they should not be made central to the basic teaching of case–control studies.

Funding

Jan P Vandenbroucke is an Academy Professor of the Royal Netherlands Academy of Arts and Sciences. The center for Public Health research is supported by a Programme Grant from the Health Research Council of New Zealand.

Conflict of interest: None declared.

References

[1] Knol MJ, Vandenbroucke JP, Scott P, Egger M. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. *Am J Epidemiol* 2008;**168:**1073–81.
[2] Doll R, Hill AB. Smoking and carcinoma of the lung: preliminary report. *Br Med J* 1950;**2:**739–48.
[3] Cornfield J. A method of estimating comparative rates from clinical data. *J Natl Cancer Inst* 1951;**11:**1269–75.
[4] Woolf B. On estimating the relation between blood group and disease. *Ann Hum Genet* 1955;**19:**251–53.
[5] Kraus AS. Comparison of a group with a disease and a control group from the same families, in the search for possible etiologic factors. *Am J Public Health Nations Health* 1960;**50:**303–11.
[6] Sheehe PR. Dynamic risk analysis in retrospective matched pair studies of disease. *Biometrics* 1962;**18:** 323–41.
[7] Thomas DB. The relationship of oral contraceptives to cervical carcinogenesis. *Obstet Gynecol* 1972;**40:**508–18.
[8] Kupper LL, McMichad AJ, Spinal R. A hybrid epidemiologic design useful in estimating relative risk. *J Am Stat Assoc* 1975;**70:**524–28.
[9] Miettinen OS. Estimability and estimation in case-control studies. *Am J Epidemiol* 1976;**103:**226–35.
[10] Prentice RL, Breslow NE. Retrospective studies and failure time models. *Biometrika* 1978;**65:**153–58.
[11] Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. *Am J Epidemiol* 1982;**116:**547–53.
[12] Pearce N. What does the odds ratio estimate in a case-control study? *Int J Epidemiol* 1993;**22:**1189–92.
[13] 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.
[14] MacMahon B, Pugh TF. *Epidemiology: Principles and Methods*. Boston, MA: Little, Brown & Co, 1970.
[15] Breslow NE, Day NE. *Statistical Methods in Cancer Research. Vol. 1–The Analysis of Case-Control Studies*. Lyon, France: IARC, 1980.
[16] Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research*. Belmont, CA: Lifetime Learning Publications, 1982.
[17] Schlesselman JJ. *Case-Control Studies: Design, Conduct, Analysis*. New York, NY: Oxford University Press, 1982.
[18] Szklo M, Nieto FJ. *Epidemiology: Beyond the Basics*. 2nd edn. Sudbury, MA: Jones and Bartlett Publishers, 2007.
[19] Rothman KJ, Greenland S, Lash TL (eds). *Modern Epidemiology*. 3rd edn. Philadelphia, PA: Lippincott Williams & Wilkins, 2008.
[20] Morabia A. *A History of Epidemiologic Methods and Concepts*. Basel, Switzerland: Birkhäuser Verlag, 2004.
[21] Pearce N. Classification of epidemiological study designs. *Int J Epidemiol* 2012;**41:**393–97.
[22] Vandenbroucke JP, Pearce N. Incidence rates in dynamic populations. *Int J Epidemiol* 2012;**41:**1472–49.
[23] Mann JI, Vessey MP, Thorogood M, Doll SR. Myocardial infarction in young women with special reference to oral contraceptive practice. *Br Med J* 1975;**2:**241–45.
[24] McEvoy SP, Stevenson MR, McCartt AT *et al*. Role of mobile phones in motor vehicle crashes resulting in hospital attendance: a case-crossover study. *BMJ* 2005; **331:**428.
[25] Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol* 1991;**133:**144–53.
[26] Vandenbroucke JP, Rosing J, Bloemenkamp KW *et al*. Oral contraceptives and the risk of venous thrombosis. *N Engl J Med* 2001;**344:**1527–35.
[27] Jick H, Vessey MP. Case–control studies in the evaluation of drug-induced illness. *Am J Epidemiol* 1978;**107:**1–7.
[28] Wynder EL. Graham EA. Tobacco smoking as a possible etiological factor in bronchiogenic carcinoma. *JAMA* 1950; **143:**329–36.
[29] Checkoway H, Pearce N, Kriebel D. *Research Methods in Occupational Epidemiology*. 2nd edn. New York, NY: Oxford University Press, 2004.
[30] Ridker PM, Hennekens CH, Lindpaintner K, Stampfer MJ, Eisenberg PR, Miletich JP. Mutation in the gene coding for coagulation factor V and the risk of myocardial infarction, stroke, and venous thrombosis in apparently healthy men. *N Engl J Med* 1995;**332:** 912–17.
[31] Schouten EG, Dekker JM, Kok FJ *et al*. Risk ratio and rate ratio estimation in case-cohort designs: hypertension and cardiovascular mortality. *Stat Med* 1993;**12:**1733–45.
[32] Greenland S. Estimation of exposure-specific rates from sparse case-control data. *J Chronic Dis* 1987;**40:**1087–94. Erratum in: *J Clin Epidemiol* 1988;**41**:423.