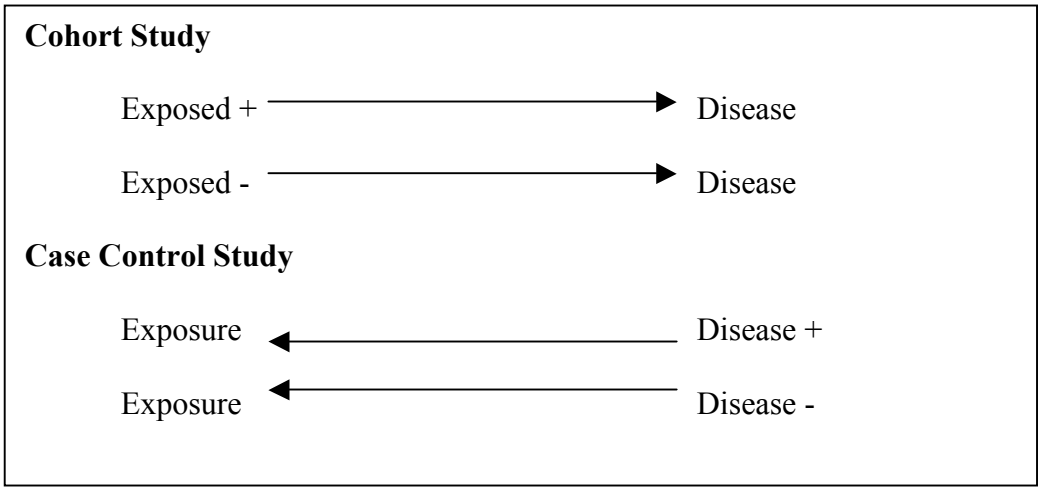
Case Control Studies

A Prospective Cohort Study is not efficient for investigating a rare disease outcome because of the large number of study subjects and/or the long period of follow-up that are needed to obtain a sufficient number of cases of disease. In this situation the Case Control Study is a more efficient alternative design to consider.

Alvin Feinstein, a clinical epidemiologist, proposed the phrase "trohoc" to describe a Case Control Study. The word "trohoc is the reversed spelling of the word "cohort", reflecting the timing relationship between a Case Control Study and a Cohort Study.

The classical description of a Case Control Study is a study that compares previous exposure histories among a group of study subjects who have the disease in question (cases) and a group of subjects who do not have the disease (controls). This description and its relationship to a Cohort Study are depicted in the following figure:



Two important questions related to this description of the Case Control Study design are:

1. What criteria should be used to select the controls?

2. Does comparing exposure history among cases and control result in a measure of association that estimates the causal effect of the exposure on the disease, as in a Cohort Study?

To address the first question, suppose an investigator wished to examine the effect of oral contraceptive use on the risk of breast cancer with a Case Control Study. According to the above description the investigation would enroll cases (women with breast cancer) and controls (women without breast cancer). Suppose the cases were women diagnosed with breast cancer at local hospitals and, for convenience, the

investigator wanted to enroll controls from the same hospital. Would any group of women without breast cancer be appropriate controls? For example, would newborn baby girls in the nursery be an appropriate control group? They are free of breast cancer but almost everyone would agree that they are not appropriate for examining the effect of oral contraceptive use on the risk of developing breast cancer. Therefore, some other characteristic is needed to define an appropriate control group.

The answer to what is an appropriate control group lies with the link between Case Control Studies and Cohort Studies. This link is formed by considering the cases in a Case Control Study to be the outcomes from a corresponding Cohort Study. Sometimes this is true by definition, if the cases were taken from a registry of outcomes in a previously documented Cohort Study. However, in many situations the cases are selected from a hospital or health care plan and were not part of a previous performed Cohort Study. Nevertheless we can still entertain the notion that a cohort of subjects existed in the past and if it were followed over time, then its outcomes would be the cases in our Case Control study. Under this assumption, the role of the controls in a Case Control Study is to provide in estimate of the prevalence of exposure in that Cohort Study. If this holds, then comparing previous exposure history among cases and controls yields estimates of measures of association that were previously discussed for Cohort Study.

The demonstrate the link between the controls in a Case Control Study and a corresponding Cohort Study, consider the following table, displaying data from a Case Control Study:

	Case	Control
Exposure +	a	b
Exposure -	c	d
Total	M_1	M_0

The usual measure of association from a Case Control Study is the Exposure Odds Ratio, comparing the odds of exposure among the cases (a/c) to that of the controls (b/d).

$$\text{Exposure Odds Ratio} = \text{EOR} = (a/c) / (c/b)$$

Mathematically, many of the measures of association from Cohort Studies can also be expressed as ratio of exposure odds. For example, the following table displays the results from a closed Cohort Study and the formulas for calculating two common measures of association: the Risk Ratio and the Disease Odds Ratio.

	Disease		
	+	-	Total
Exposure +	A	B	N_1
Exposure -	C	D	N_0

Risk Ratio = RR = $(A/N1)/(C/N_0)$
= $\mathbf{(A/C)/(N_1/N_0)}$

Disease Odds Ratio= DOR = $(A/B) / (C/D)$
= **(A/C) / (B/D)**

The second equation for the Risk Ratio (RR) demonstrates that is can also be calculated by dividing the odds of exposure among the cases of disease (A/C) by the odds of exposure in the source population (closed cohort) (N_1/N_0).

Also, by the symmetry of the odds ratio, the Disease Odds Ratio (DOR) is equal to the ratio of the odds of exposure among the cases (A/C) divided by the odds of exposure among the subjects who did not develop the disease (B/D).

Similarly the following table displays the results from an open cohort and the calculation for the Rate Ratio

	Disease	Person-Time
Exposure +	A	K_1
Exposure -	C	K_0

RR = $(A/K_1)/(C/K_0)$
= $\mathbf{(A/C)/(K_1/K_0)}$

The second equation for the Rate Ratio demonstrates that is can also be calculated by dividing the odds of exposure among the cases of disease (A/C) by the ratio of exposed to non-exposed person-time in source population.

Control Selection

The role of the controls in a Case Control Study is to estimate the prevalence of exposure in the Cohort Study whose outcomes would be the cases at hand. From the above formulas, this allows the Exposure Odds Ratio from a Case Control study to estimate common measures of association from the corresponding Cohort Study. This cohort could be closed or open.

Corresponding Cohort Study: Closed Cohort

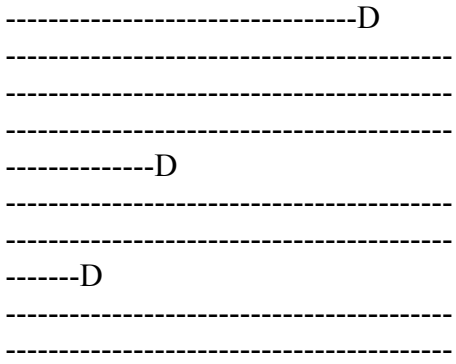
If the corresponding cohort is closed, then the selection of controls is often referred to as cumulative incidence sample and there are two options for selecting controls:

1. Selecting controls from subjects in the cohort who did not develop the outcome during the period of follow-up (this is method used in the Classic Nested Case Control Study)

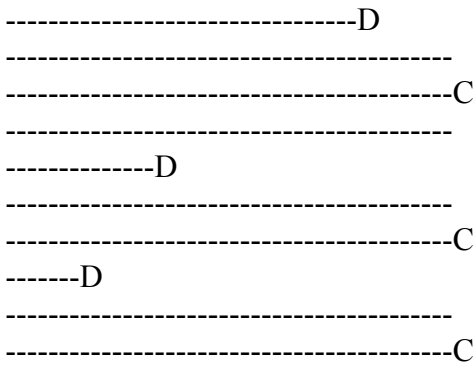
2. Select controls from everyone in the cohort at the beginning of the follow-up (this is an example of a Case Cohort Study).

These options are describes in the following figure:

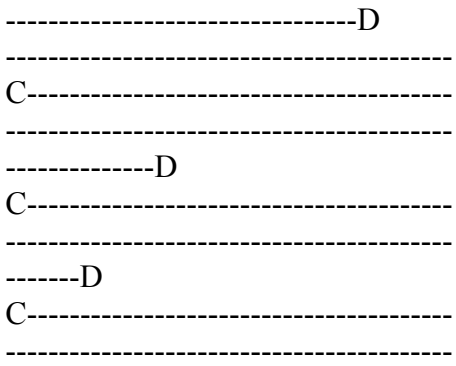
Closed Cohort (D indicates the development of disease)



Nested Case Control Study (C indicates Controls selected from non-disease group)



Case Cohort Study (C indicates Controls selected from full cohort)



Classic Nested Case Control Study

The following example describes the classic Nested Case Control Study (Willett. Lancet 1983 Jul 16;2(8342):130-4). The study involves data from the Hypertension Detection and Follow-up Program (HDFP). This was a previously performed randomized clinical trail that investigated different treatments for hypertension. However, 4480 participants in this RCT provided blood sample, which were frozen for future use. The RCT also created a registry that recorded the names of study subjects who developed cancer from 4480 participants.

111 of these subjects developed cancer and were chosen as the cases in a Case Control Study to examine the relationship between selenium levels and cancer. Controls should be chosen to reflect the prevalence of exposure in the in the corresponding cohort of 4480 subjects. However, in this classic Nested Case Control Study controls were selected from the members of the cohort who did not develop the disease (4480 – 111 = 4369 subjects). 210 controls were selected form these 4369 study subjects. The investigator measured selenium levels from the frozen blood samples of the 111 cases and the 210 controls. 57 of the cases and 84 of the controls had low levels of selenium. The results of the case control Study are displayed in the following table:

	Case	Control
Low Selenium	57	84
High Selenium	54	126
Total	111	210

EOR = [57/54] / [84/126] = 1.6

The following tables displays the results of the Cohort Study had the investigator measured the blood specimens for all 4480 study subjects.

	Cancer		
	Yes	No	Total
Low selenium	57	B	N_1
High selenium	54	D	N_0
Total	111	4369	4480

$RR = [(57/N_1)]/[(54/N_0)] = [57/54]/[N_1/N_0] = ?$

DOR= [57/B]/[54/D] = [57/54]/[B/D] = ?

The values for the Risk Ratio (RR) and Disease Odds Ratio (DOR) would require analyzing the blood specimen on the remaining 4369 subjects. If the selenium distribution of the 210 controls reflects the distribution of all 4369 potential controls then

the Exposure Odds Ratio (EOR) from the Nested Case Control Study will estimate the Disease Odds Ratio from the Cohort Study as demonstrated in the following calculation:

$$\begin{aligned} \text{DOR} &= [(57/B)]/[(54/D)] \\ &= [57/54] / [B/D)] \\ &\approx [57/54] / [84/126] = 1.6 = \text{EOR} \end{aligned}$$

Furthermore, since the disease is rare, the number of potential controls (4369) is almost the same as the number of subjects in the cohort (4480). Therefore, the odds of exposure among the 4369 potential controls (B/D) should be similar to the odds of exposure in the full cohort (N_1 / N_0). Under this **rare disease assumption,** it follows that the Exposure Odds Ratio approximates the Risk Ratio from this Cohort Study.

$$\begin{aligned} \text{EOR} &= [57/54] / [84/126] = 1.6 \\ &\approx [57/54] / [B/D] = \text{DOR} \\ &\approx [57/54] / [N_1 / N_0\} = \text{RR} \end{aligned}$$

Case Cohort Study

In the classic Nested Case Control Study controls are chosen from subjects who did not develop the disease in the corresponding closed cohort (4,369 subjects in the previous example). The Exposure Odds Ratio from the Case Control Study estimates the Disease Odds Ratio from the corresponding Cohort Study, and under the rare disease assumption also estimates the Risk Ratio from the Cohort Study.

An alternative option is to select controls from the 4,480 members of the original cohort. The resulting Case Control Study is usually referred to as a Case Cohort Study. The exposure odds among the selected controls (b/d) should estimate the exposure odds in the full cohort (N_1/N_0). Furthermore, the Exposure Odds Ratio from the Case Cohort Study estimates the Risk Ratio from the Cohort Study without any assumption about the rarity of the disease.

Since the outcomes in a Cohort Study at part of the at-risk subjects at the start of a study, it is possible disease case might also be selected as a control in a Case Cohort Study. This presents some problems in performing tests of significance and confidence interval estimation, but does not invalidate the Exposure Odds Ratio from the Case Cohort Study estimating the Risk Ratio from the Cohort Study.

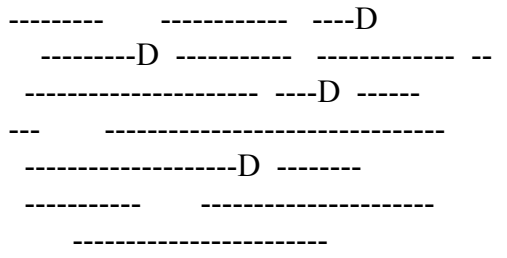
Corresponding Cohort Study: Open Cohort

Returning to a previous example, suppose that an investigator plans a Case Control Study examining the relationship between oral contraceptive use and the risk of developing breast cancer. Furthermore, suppose that the cases are women diagnosed with

breast cancer at a local hospital in the past two years. Since the purpose of the controls is to reflect the prevalence of exposure (oral contraceptive use) in the cohort study that gave rise to the cases, the challenge to the epidemiology if to formulate this cohort study and determine an appropriate control to describe the prevalence of exposure in that cohort.

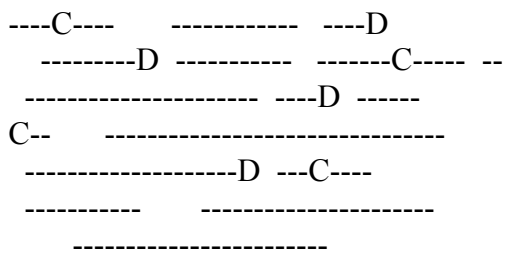
Since the cases are chosen from a single hospital, the corresponding cohort would be the population living in the catchment area of that hospital. Membership in this population may be defined by residential area and also by other factors such as a women's primary care physician and health plan that might influence her being referred to that hospital for testing and ultimately for the diagnosis of breast cancer. Unfortunately a list of women living in the population would not exist. However, if it did exist then it would probably be a dynamic population (open cohort) with women moving in and out of this population. A description of this open cohort is given in the following figure:

Open Cohort (D represent the development of disease)



Since the cohort is open, the appropriate measure of disease incidence is the Incidence Rate. Another term for an Incidence Rate is the Incidence Density. (proposed by Olli Miettinen), and the corresponding Case Control Study is called a Density Type Case Control Study. The controls are chosen so that their odds of exposure will reflect the ratio of the amount of person time in the open cohort that was contributed by oral contraceptive users to the amount of person time in the open cohort that was contributed by non-oral contraceptive users. Therefore controls should be selected from the person-years of the cohort study. The following figure displays this type of density type sampling for controls:

Density Type Sampling of Controls (C represents a selected control)



The measure of association in an open cohort study is the Rate Ratio (RR) as described in the following table

	Cases of Disease	Person-Time
Exposed	A	K_1
Non-Exposed	C	K_0

$$RR = (A/K_1) / (C/K_0)$$
$$= (A/C) / (K_1/K_0)$$

The display of data from the corresponding Density Type Case Control Study is

	Case	Control
Exposure +	A	B
Exposure -	C	D
Total	M_1	M_0

$$EOR = (A/C) / (B/D)$$

If the exposure odds among the controls (B/D) estimates the amount of person time in the open cohort that was contributed by exposed subjects divided by the amount of person time in the open cohort that was contributed by non-exposed subjects (K_1/K_0), then is follows that the Exposure Odds Ratio from the Density Type Case Control Study estimates the Rate Ratio from the corresponding Open Cohort Study

$$EOR = (A/C) / (B/D)$$
$$\approx (A/C) / (K_1 / K_0) = RR$$

Sources of Controls

If the cases in a Density Type Case Control Study are a list of all cases that develop in a geographical population (e.g. state of Massachusetts) then the corresponding open cohort is a census of individuals living in that population in the past. Such cases are referred to as **Population-Based Cases** and the selected are referred to as **Population-Based Controls.**

If the cases are chosen from one (or more) hospitals with a specified diagnosis, then they are referred to as **Hospital Based Cases**. Controls are typically patients selected from the same hospital but with a different diagnosis. The reason for this choice is the assumption that the cases and controls from the same hospital come from the same catchment area. Therefore the controls can be considered a sample from the catchment area. However, for the prevalence of exposure among the controls to reflect the prevalence of exposure in the catchment area, the diagnosis for the controls should not be

one that is caused or prevented by the exposure. For example, if the cases are women who are diagnosed with breast cancer and the exposure of interest is the use of oral contraceptive, then an inappropriate control group would be women diagnosed with venous thrombosis since this might be caused by oral contraceptive use. Controls selected from the same hospital as the cases are referred to as **Hospital Based Controls.**

Example: Density Type Case Control Study

The following results are from a hospital-based Density Case Control Study measuring the association between a series of potential risk factors and the development of Aortic Stenosis (Hoagland. Am J Med 1986;80(6):1041-50). Aortic Stenosis "is a disease of the heart valves in which the opening of the aortic valve is narrowed. The aortic value is the valve between the left ventricle of the heart and the aorta, which is the largest artery in the body" (http://en.wikipedia.org/wiki/Aortic_valve_stenosis).

The cases for this study were 105 subjects with Aortic Stenosis documented by cardiac catheterization (gold standard test). The suspected risk factors of interest (exposures) included smoking, diabetes, hypertension, cholesterol, and a family history of CHD. Three Control Groups were considered for this study:

1. **Group 1**: Patients who underwent cardiac catheterization, which showed no Aortic Stenosis but did show another type of valvular heart disease (n=110)

2. **Group 2**: Patients who underwent cardiac catheterization, which showed no Aortic Stenosis and no other type of valvular heart disease (n=170)

3. **Group 3**: Surgical patients whose reason for surgery was not known to be associated with risk factors of interest (n=269)

All data was obtained from medical record reviews. If no mention of a risk factor was indicated in the medical record, then it was assumed to be absent (i.e. non-exposed). This may results in a large potential for a misclassification bias.

The following table shows the relationship between hypertension and Aortic Stenosis using control Group 3.

	Case	Control
Hypertension	43	91
No Hypertension	62	178
Total	105	269

$$\text{EOR} = (43/62)/(91/178)$$
$$= 1.4$$

One limitation of a Case Control Study is that it does not allow for the estimation of exposure-specific risks or rates for developing the outcome. Since the investigator usually determines the relative sizes of the case and control groups, it follows that the overall prevalence of disease in the data does not reflect the incidence of the disease in the corresponding cohort study. For example, the prevalence of disease, P(D), among the exposed and non-exposed groups from the previous table is

P(Aortic Stenosis| History of Hypertension) = 43/134 = .32
P(Aortic Stenosis| No History of hypertension) = 62/240 = .26

These proportions are somewhat arbitrary and do not reflect the risk of developing hypertension. To demonstrate this, suppose that the investigator selected twice as many controls for the study. The expected results from this study are shown in the following table:

	Case	Control
Hypertension	43	182
No Hypertension	62	356
Total	105	538

EOR = (43/62)/ (91/178)
= 1.4

The value for the Exposure Odds Ratio does not change but the prevalence of Aortic Stenosis in each group changes to

P(Aortic Stenosis| History of Hypertension) = 43/225 = .19
P(Aortic Stenosis| No History of Hypertension) = 62/418= .15

Case Control Studies are sometimes referred to as “quick and dirty” studies. They are labeled as “quick” compared to prospective Cohort Studies in that the follow-up period for the study subjects has happened in the past. On the other hand, they are labeled as “dirty” in part because their potential for selection bias, due to the use of an incorrect control group. This may hold true for the first two control groups considered for this study.

Control Group 1 included patients who underwent cardiac catheterization, which showed no Aortic Stenosis, but did show another type of valvular heart disease. It is very possible that the same risk factors, which cause Aortic Stenosis, may also cause these other types of valvular heart disease. Therefore the exposure history in Control Group 1 may over estimate that for the source population

Control Group 2 included patients who underwent cardiac catheterization, which showed no Aortic Stenosis and no other type of valvular heart disease. It is very possible that the risk factors being considered as exposures in this study may have influenced the

decision for cardiac catheterization for Control Group 2. If so, then the exposure history in Control Group 2 may over estimate that for the source population. This is demonstrated by the suggestion of a protective effect of hypertension in the following table that uses Control Group 2.

	Case	Control
Hypertension	43	89
No Hypertension	62	81
Total	105	170

$$\text{EOR} = (43/62) / (89/81) = .63$$

Measurement bias is a potential in any study but may be a particular problem in the study at hand. All exposure information was recorded from medical records. Control Group 3 included surgical patients whose reason for surgery was not known to be associated with the risk factors of interest. Exposure information on the Cases, members of Control Groups 1, and members of Control Group 2 where obtained from interview by cardiology fellows at the time of cardiac catheterization, which would include detailed questions on Coronary Heart Disease (CHD) risk factors. On the other hand, subjects in Control Group 3 were interviewed by different hospital staff prior to surgery and may have had less detailed questions on CHD risk factors. For, example, it may be that subjects in Control Group 3 were not asked detailed questions about family history of heart disease or such information was not completely recorded in their medical records. This might explain the possible protective effect of this factor that is shown in the following table.

	Case	Control
Family History	42	53
No Family History	63	216
Total	105	269

$$\text{EOR} = (42/63) / (53/216) = 2.72$$

Risk Set Sampling

A Case-Cohort Study is also an option when the corresponding when the cases are considered to be the outcome of an open Cohort Study. This would mean that cases has the potential for being selected as controls when the latter are selected to reflect the amounts of person-time from the exposed and non-exposed groups in the open cohort. For example, if Control Group 3 in the previous example were appropriate to reflect this information, then it is possible that this group may contain some cases of Aortic Stenosis since its members did not undergo cardiac catheterization.

Risk Set Sampling is another option for selection controls, in which the selected controls are matched the follow-up times of cases. The risk-set for a case is the members of the cohort study who were also at risk for developing the disease at the time a case developed the disease. Risk-set sampling involves selecting one of more members of that set as controls. The resulting matched analysis is similar to a survival analysis that could be performed on the full cohort. Risk-set sampling is depicted in the following figure.

Risk-Set Sampling of Controls (C_i represents a potential control for D_i)