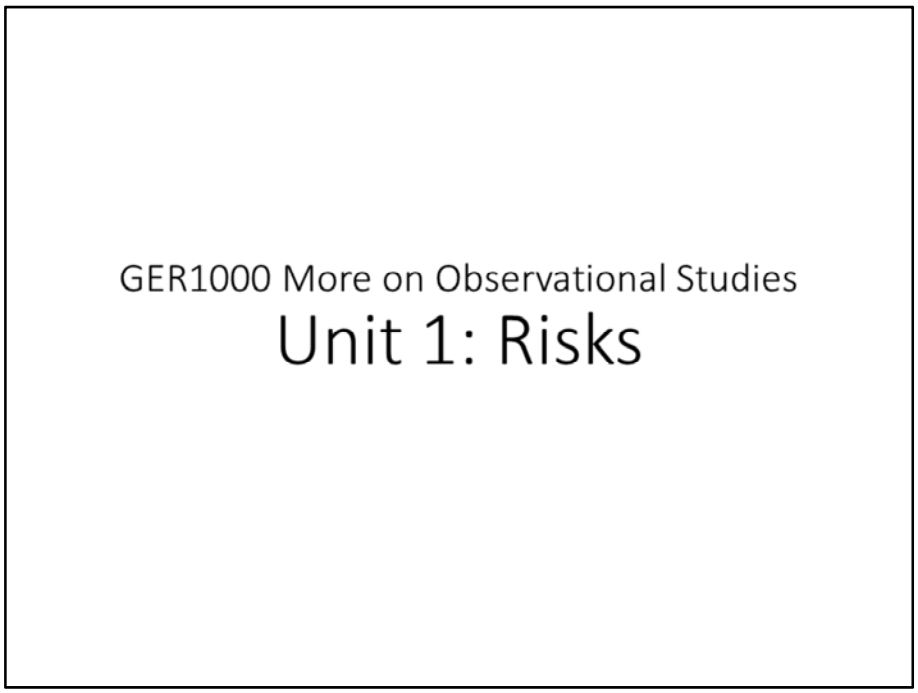
GER1000
More on Observational Studies

Previous chapters discuss controlled experiments and observational studies, association and correlation, as well as sampling methods. The relevant steps in the QR framework are Collect, Analyse and Communicate, though Frame and Specify are also important. Now we will look at two kinds of observational studies that feature prominently in studies of risk of all kinds of undesirable outcomes, be it disease or financial crisis. Two keys concepts of this chapter are risks and odds.



[1] Human survival has always been a struggle, though the adversaries change with time. Malnutrition and communicable diseases used to kill many, but now, lifestyle diseases seem to pose a greater risk, at least to the more affluent societies. Consider a population where every individual initially did not have a condition, but after some time, some of them will acquire that condition, in a manner which is not predictable. We will define the risk of such an uncertain condition as its rate in the population. So a risk is a number between 0 and 1, or between 0% and 100%. Not all rates are risks. For instance, the rate of female in any adult population is around 0.5, but it is odd to talk about the risk of being female, since the outcome has been fixed.

Contingency Table

	Diabetic	Healthy	Row total
Female	72,000	144,000	216,000
Male	52,000	156,000	208,000
Column total	124,000	300,000	424,000

- rate(diabetes) = $\frac{124{,}000}{424{,}000} = 0.29$, or 29%.
- Definition: risk(diabetes) = 0.29 or 29%.
- Diabetes risk by country:
  https://www.indexmundi.com/facts/indicators/SH.STA.DIAB.ZS/rankings

[2] The table a categorises a hypothetical adult population of 424,000 people, by sex and by diabetes status. Such a table is called a contingency table. The variables are called “categorical”, since their values are categories, unlike a numerical variable such as height. The rate of diabetes is 124,000/424,000 = 0.29, or 29%. We shall call the number 0.29 the risk of diabetes for the population. It is useful to compare risks across populations from different countries. According to the website, the diabetes risk in Mauritius is about twice as high as in Singapore, while Argentinians are half as at risk as Singaporeans. In the website, the figures are prevalence, which is the same as our rate. The population risk of an uncertain outcome is defined as the population rate.

Simple Random Sampling

	Diabetic	Healthy	Row total
Female	72,000	144,000	216,000
Male	52,000	156,000	208,000
Column total	124,000	300,000	424,000

- Take SRS of 1,000, calculate sample diabetes rate. It fluctuates
  - unpredictably.
  - around population rate.
- Sample diabetes rate estimates population rate, or risk.

[3] Suppose we take a simple random sample of size 1,000 from this population. What can we say about the diabetes rate in the sample? There are two things. Firstly, it is a random quantity, meaning if we repeat the sampling many times, calculating the resulting sample rate each time, these numbers will fluctuate, and it is impossible to predict exactly what will be obtained. Secondly, the sample rates will fluctuate around the population rate, which is 0.29 in this case. These conclusions follow from the chapter on Sampling.

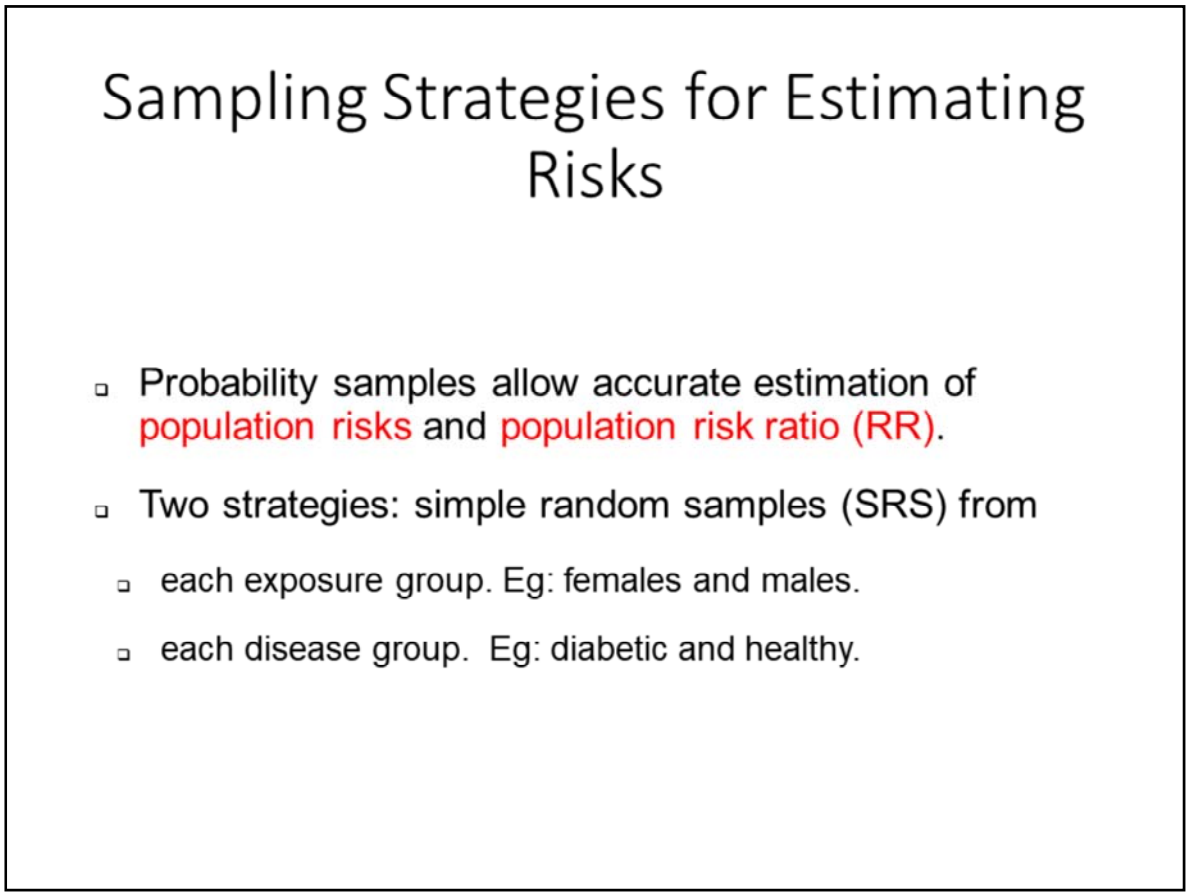
To know the diabetes rate, one has to test everyone of the 424,000 people. The work is too time-consuming and costly, even though the information is very important. A more practical solution is to take a simple random sample, and to use the sample rate to estimate the population rate. Although we expect some error from this approach, it can be made quite small by taking a larger sample, while still incurring only a fraction of work required for studying the whole population.

Risk Ratio (RR)

	Diabetic (D)	Healthy	Row total
Female (F)	72,000	144,000	216,000
Male (M)	52,000	156,000	208,000
Column total	124,000	300,000	424,000

- $\text{risk(D|F)} = \frac{72{,}000}{216{,}000} \approx 0.33 > 0.25 = \frac{52{,}000}{208{,}000} = \text{risk(D|M)}$
- Risk ratio $= \frac{\text{risk(D|F)}}{\text{risk(D|M)}} \approx \frac{0.33}{0.25} \approx 1.33$. Also called relative risk.
- RR = 1: no association.

[4] Within a country, the population risk is too crude, since it averages over people of various risks. For diabetes, diet and other lifestyle factors matter a lot, so it is important to look at risks of different subpopulations. In this example, we will look at sex as a risk factor. Sex and age are often considered risk factors in many studies of diseases, although they cannot be changed easily. According to the table, the risk of diabetes among females is about 0.33, while it is 0.25 for males. The notation for the risks are consistent with that for rates, which you have seen in the first chapter. So females have higher risk than males, and this fact can be summarised by a risk ratio. The risk ratio of diabetes between females and males is 0.33/0.25 = 1.33. This quantity is also called a relative risk. In general, the risk in the first group is greater or smaller than the second group according to whether the RR is more or less than 1. If the RR is 1, then the two groups have the same risk, meaning there is no association between the disease and the exposure that defines the groups.



[5] Just like the population risk, for accurate estimation of the risks of exposure groups, probability samples are practical, due to considerable savings in resources. There will be uncertainty in the estimates obtained, though it can be made smaller by taking larger samples. Towards this goal, we attach the word “population” to those quantities that are to be estimated by the samples. In the example, 0.33 and 0.25 are population risks for females and males respectively, and 1.33 is the population RR between females and males.
We will consider two sampling strategies. The first takes SRS separately from the exposure groups, which in this example are females and males. The second takes SRS separately from disease groups: the healthy and the diabetic. We will look at the pros and cons of the two strategies.

Study 1: the sample table

	Diabetic	Healthy	Row total
Female	378	702	1,080*
Male	53	155	208**
Column total	431	857	1,288

*: 5 in 1000 females sampled. **: 1 in 1000 males sampled.

- Expected no. of diabetic females = $1{,}080 \times \frac{1}{3} = 360$. Actual = 378.
- Expected number of diabetic males = $208 \times \frac{1}{4} = 52$. Actual = 53.
- Numbers in black fluctuate with repeated samplings.

[6] The table shows the result of sampling 5 females out of 1,000, and sampling 1 male out of 1,000, both at random, using a computer. This means that the sample will have 216,000 x 5/1,000 = 1,080 females and 208,000 x 1/1,000 = 208 males, as shown in the last column. Since in the population 1/3 of women are diabetic, we expect about 1/3 of the 1,080 sample women to be diabetic, which is 1,080 x 1/3 = 360. It turns out that 378 of the sample women are diabetic; the discrepancy with 360 is because of random fluctuation. So there are 702 healthy sample women. Similarly, we expect 208 x 1/4 = 52 of the sample men to be diabetic; actually 53 are. In the combined sample of 1,288 people, 431 turn out to be diabetic, and 857 healthy. The row totals are fixed by the sampling method, but all other numbers will fluctuate around their expected numbers if the sampling process were repeated.

Study 1: estimation

	Diabetic	Healthy	Row total
Female*	378	702	1,080
Male**	53	155	208
Column total	431	857	1,288

*: 5 in 1000 females sampled. **: 1 in 1000 males sampled.

- Estimated risk(D | F) = $\frac{378}{1{,}080} \approx 0.35$. 0.33 = population risk(D|F)
- Estimated risk(D | M) = $\frac{53}{208} \approx 0.25$. 0.25 = population risk(D|M)
- Estimated RR ≈ $\frac{0.35}{0.25} \approx 1.37$. 1.33 = population RR

[7] We can estimate the population risks by doing the same calculations on the sample table. The estimated risk of diabetes for females is 378/1,080 = 0.35, and the estimated risk for males is 53/208 = 0.25. The population RR is estimated as 0.35/0.25 = 1.37. The estimates are very close to the respective population quantities. This is not due to good luck. If the sampling were repeated, so that the table looks different, the estimates will also be different, but they will still be quite close to their targets. We conclude that simple random samples from exposure groups allows accurate estimation of population risks and RR, even if the fractions sampled are different among the exposure groups.

Study 2

	Diabetic	Healthy	Row total
Female	364	142	506
Male	256	158	414
Column total	620*	300**	920

*: 5 in 1000 diabetic sampled. **: 1 in 1000 healthy sampled.

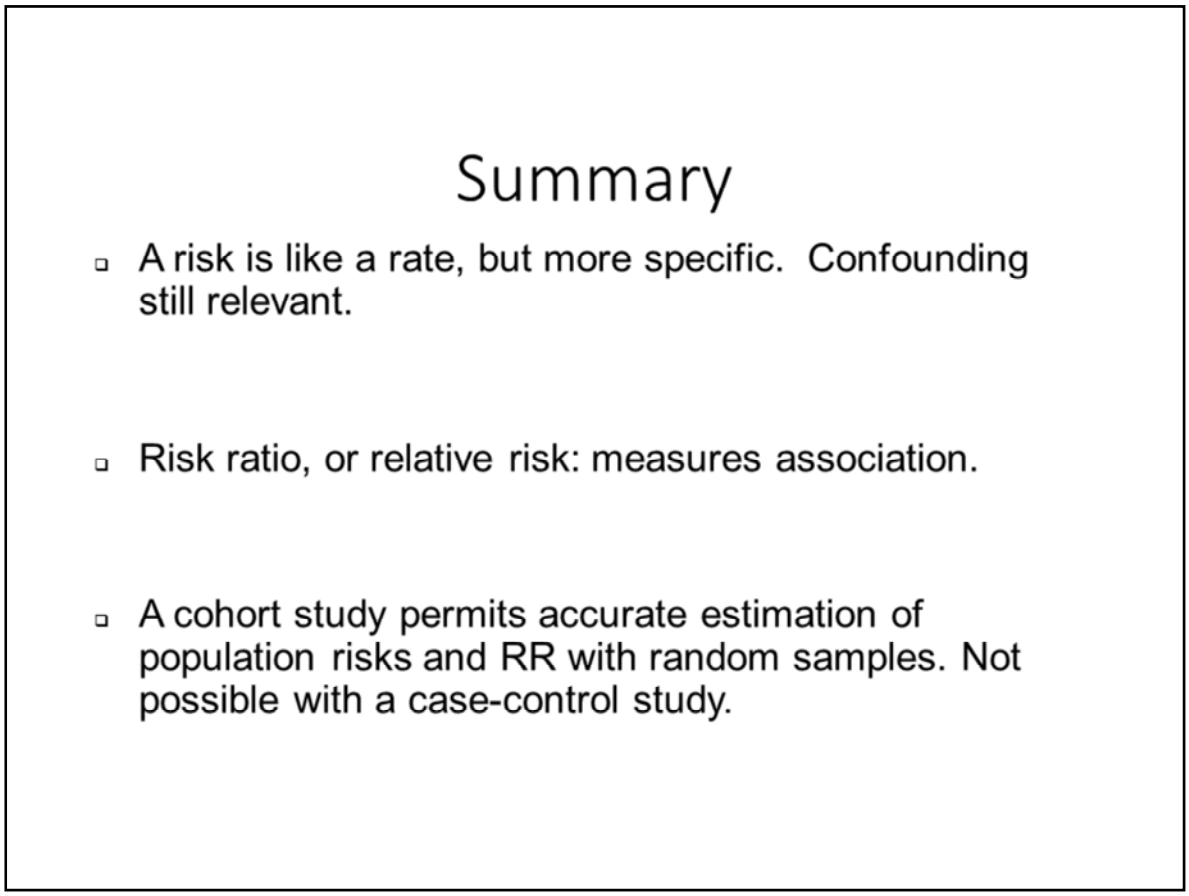
- $\frac{364}{506} \approx 0.72 \neq 0.33$. $\frac{256}{414} \approx 0.62 \neq 0.25$
- $\frac{0.72}{0.62} \approx 1.16 \neq 1.33$

[8] The table shows the results of another study, which randomly samples 5 out of 1,000 diabetic patients, and 1 of 1,000 healthy persons. The situation is similar to Study 1, but flipped. Now the column totals are fixed, and all other numbers fluctuate. For example, since in the population about 0.42 of diabetic patients are male, we expect the number of sample males to be 620 x 0.42 = 260; the actual number is 256, because of random fluctuation. If we try to estimate the risks, we get, for females, 364/506 = 0.72, and for males, 256/414 = 0.62. These numbers are not close to the population risks of 0.33 and 0.25. In fact, they are too high, because the sample has a higher rate of diabetic patients than the population. We can push on and see that 0.72/0.62 = 1.16, so the RR is also not estimated accurately. This is not bad luck. If the sampling were repeated, the calculations will still produce risk estimates that are too high. Unlike the previous case, the current sample table cannot be used in the same way to yield good estimates of population risks and RR.

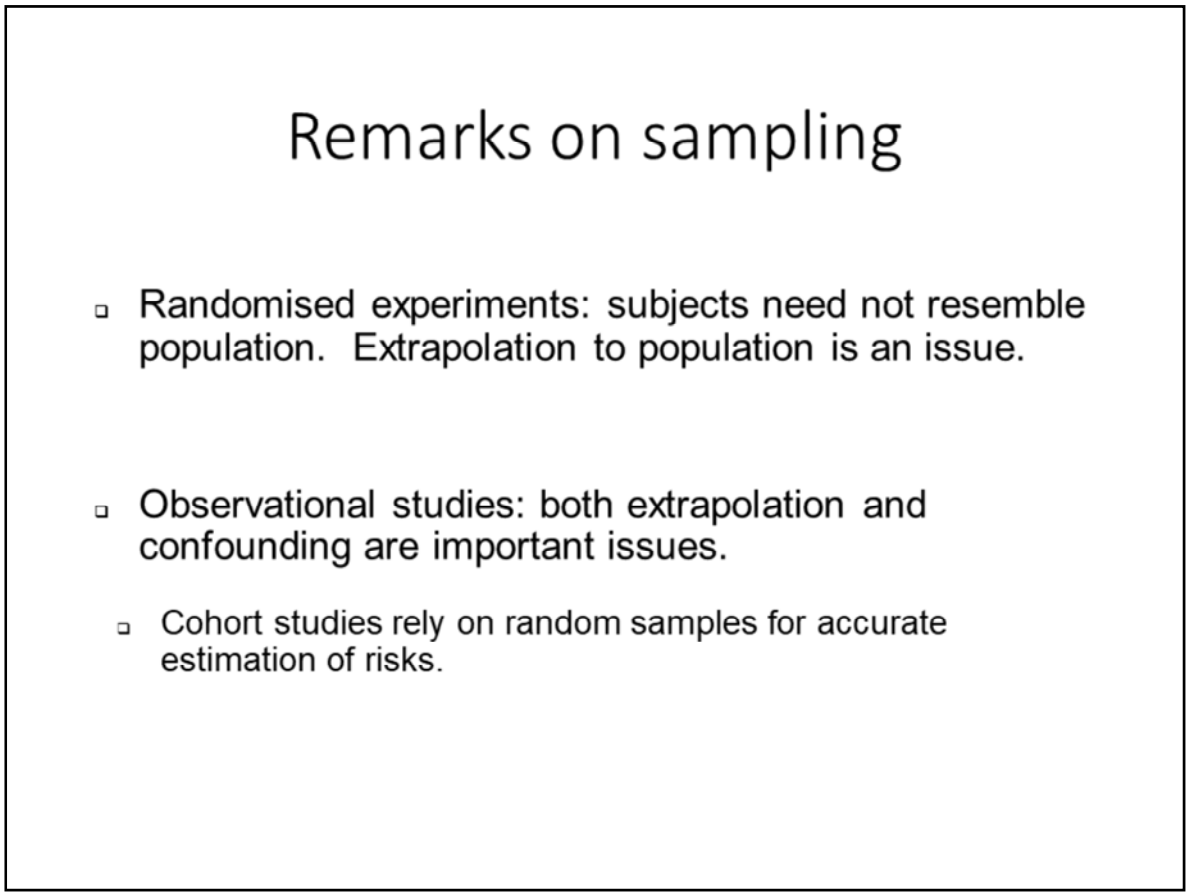
Cohort and Case-control Studies

Study	Samples from	Advantage
Cohort (study 1)	Exposure groups	Risks and RR can be estimated from sample table
Case-control (study 2)	Disease groups	Good for rare diseases

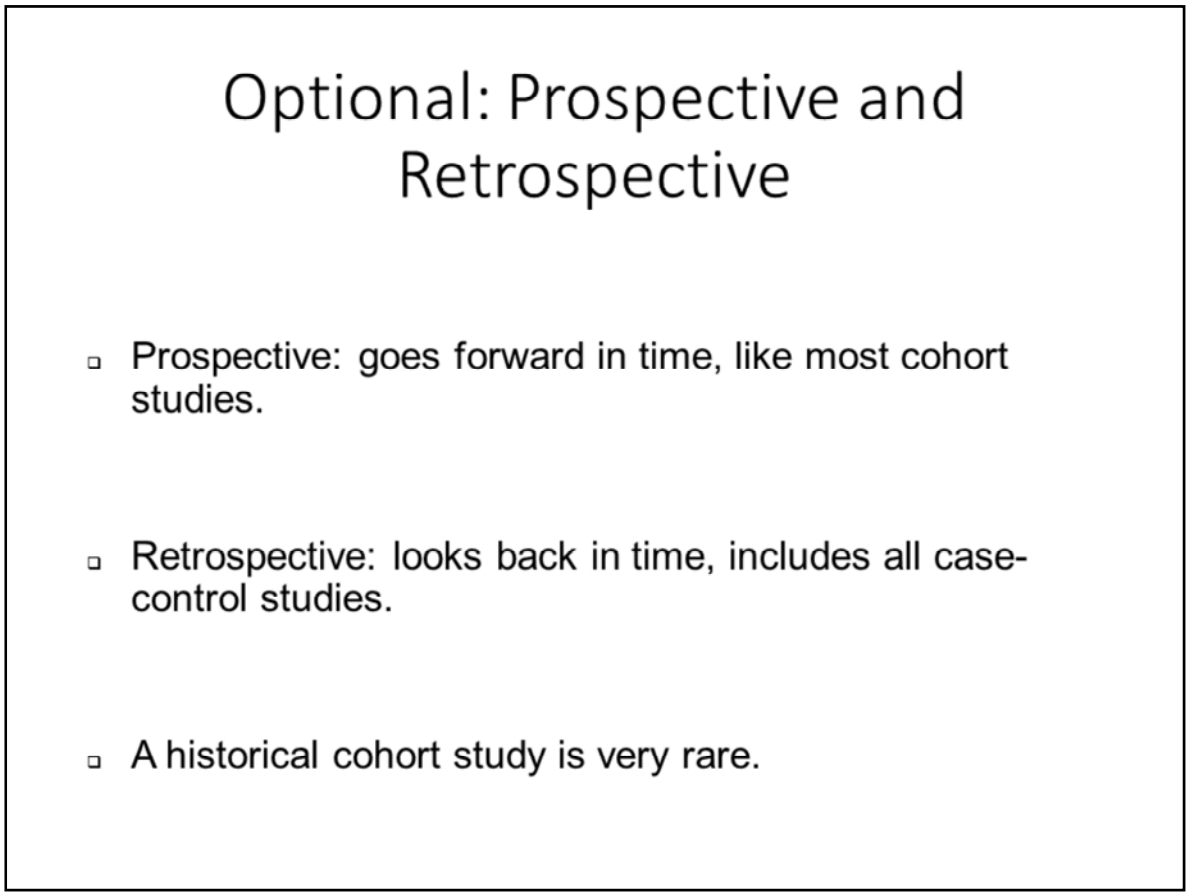
[9] Study 1 is an example of a cohort study. A cohort study typically enrols subjects from different exposure groups, who are then monitored for a period of time. At the end of the study, the investigator will know how many subjects in each exposure group have developed diabetes. Thus population risks and RR can be estimated, provided the samples are randomly selected. Study 2 is an example of a case-control study. A case-control study enrols subjects from different disease groups, so is always conducted at the end of a time period. Generally, the sample data cannot be used to estimate the population risks and RR. An exception is if the same fractions are sampled from the disease groups; the detailed reason is out of the scope of this module. Case-control studies are valuable for studying rare diseases that can only be observed in very large samples of exposed individuals.



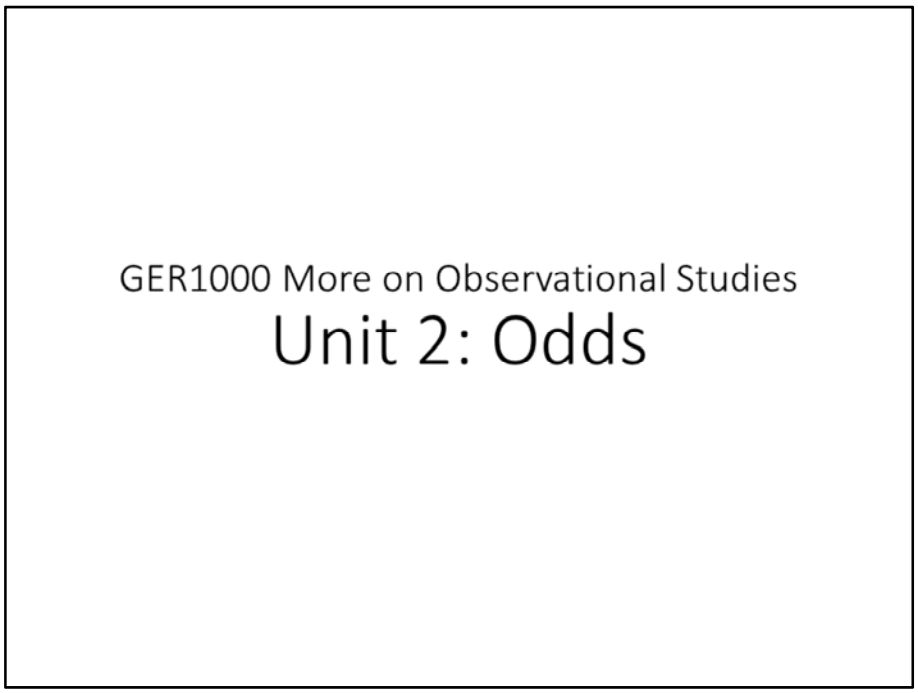
[10] A risk is a rate, but more specific. For example, when we speak of the risks of diabetes for females and males, we know that sex, or genetics, plays some role in diabetes, and not the other way. Still, confounding is an issue: there can be a third variable associated with both sex and diabetes, which may implicate other causal factors besides genetics.
The risk ratio, or relative risk, is a useful measure of association. It is greater or less than 1 according to whether the first group has higher or lower risk than the second group.
With a cohort study, population risks and population risk ratio can be accurately estimated from the sample table, if samples are selected at random. The estimation is generally not possible with a case-control study, even with random samples.



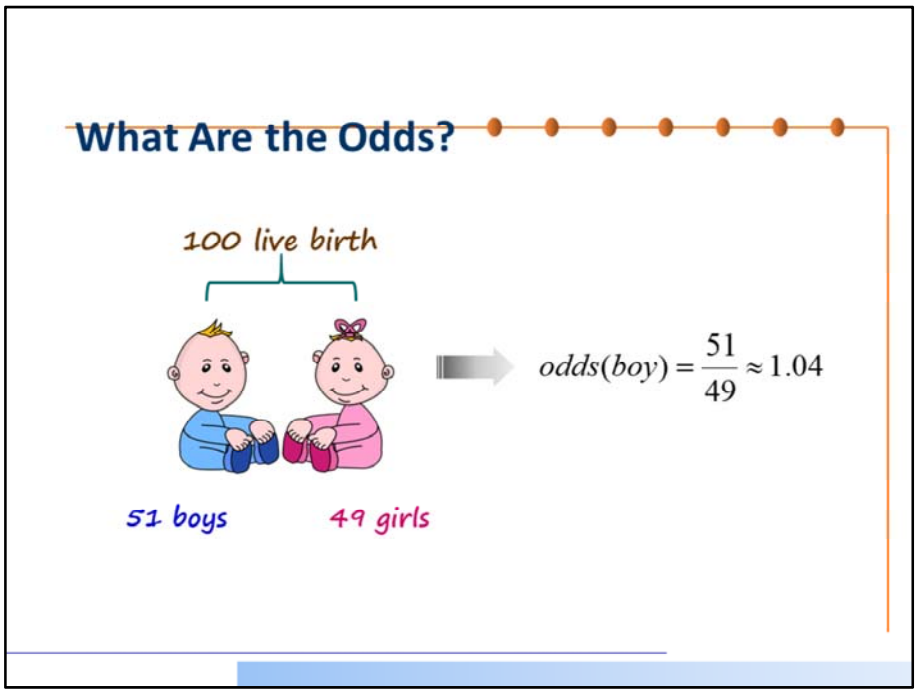
[11] We have seen confounding as the main contrast between a randomised experiment and an observational study. Sampling issues are relevant to both designs. Subjects in a randomised experiment usually do not resemble the population of interest. For example, the Salk randomised experiment on a polio vaccine involved only children from richer families, and such families are different from other families in many important ways. This can pose some challenge in extrapolating the results to the larger population. Extrapolation is also often an issue for observational studies, since taking a random sample is a lot of hard work. More specifically, for a cohort study to accurately estimate risks, the samples should be chosen at random from the population. In studies with non-random samples, it cannot be taken for granted that the population risks can be estimated accurately.



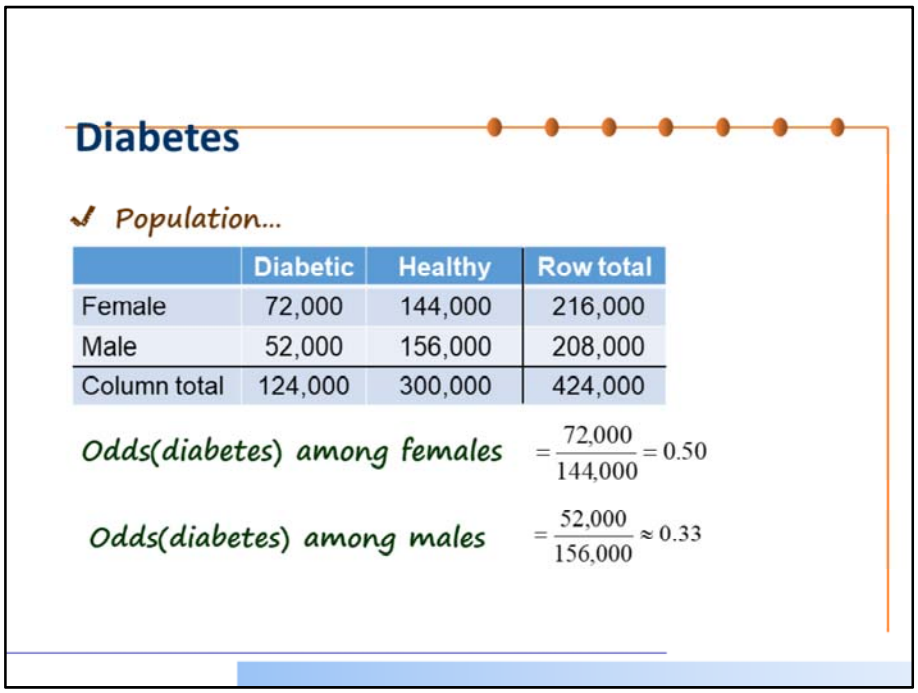
[12] This optional slide touches on two terms that may be encountered in reports on observational studies. Most cohort studies are prospective, namely, the exposure groups are enrolled, then followed through a fixed time period. All case-control studies must be retrospective, since disease status can only be determined at the end. The exposure state is then measured by looking back in time. Very rarely, we come across a historical cohort study, which samples from exposure groups at the end of the study period.



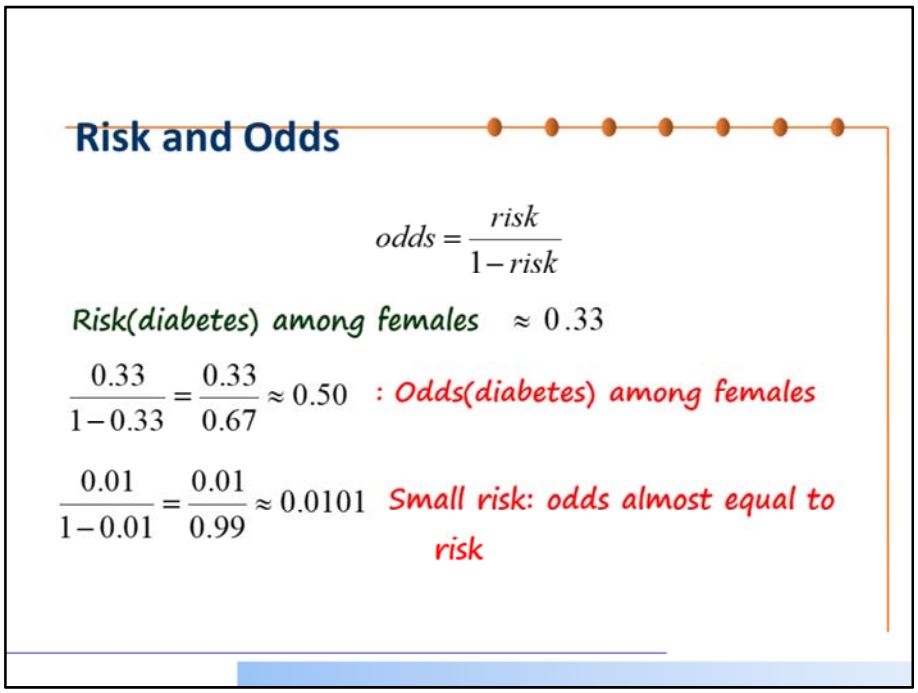
[1] As described in the previous unit, the risk ratio, RR, can be estimated from the sample table of a cohort study. However, this is generally not the case for a case-control study. In this unit we will introduce the population odds ratio, which can be estimated from both kinds of studies.



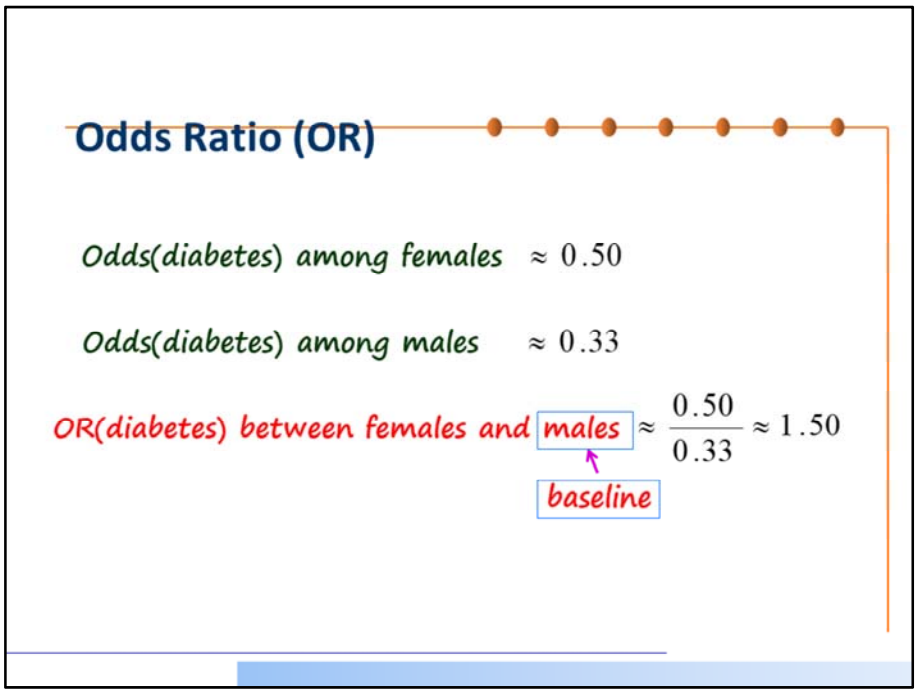
[2] Suppose that in 100 live births, there are 51 boys and 49 girls. Then the odds for boy is the ratio of 51 to 49, or about 1.04, meaning a live birth is slightly more likely to be a boy than a girl. If an event is very likely, the numerator is much larger than the denominator, so its odds is large. If it is very unlikely, then its odds is very close to 0.



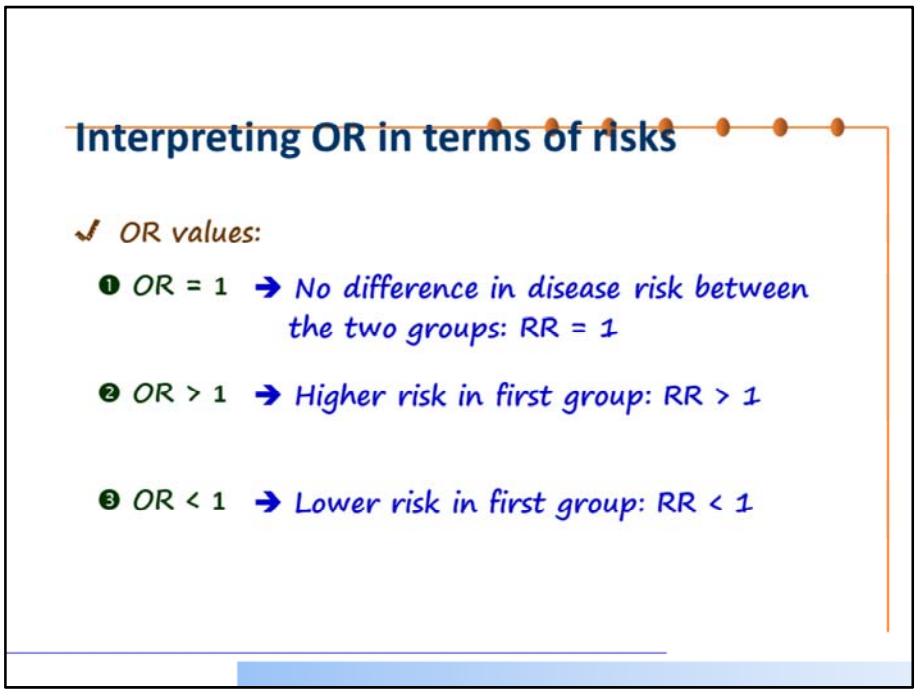
[3] Let us apply the concept of odds on the diabetes population introduced in the previous chapter. The odds of diabetes among females is 72,000/144,000 = 0.50. So for every diabetic woman, there are two who are not. Similarly, the odds of diabetes among males is 52,000/156,000 ~ 0.33. Thus, males have lower odds than females.



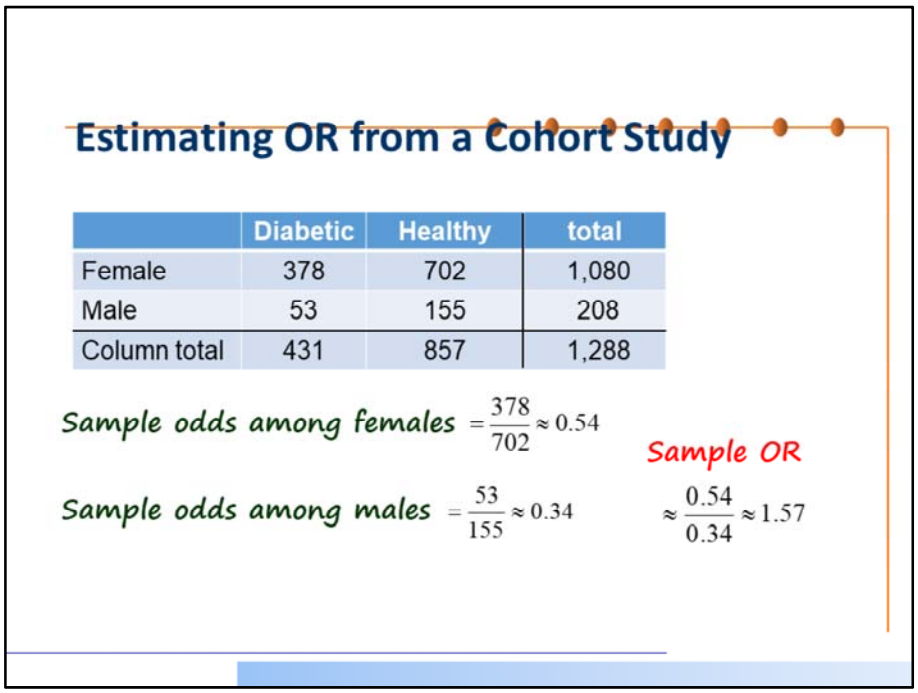
[4] The odds for an event is determined from its risk via this formula. For example, the risk of diabetes among females is 0.33. Applying the formula gives 0.50, which is the odds we calculated in the previous slide. You may want to verify this relationship by doing a similar calculation for the males. The formula relating odds to risk tells us two things. First, the value of odds is always larger than the risk. For example, the odds 0.50 is larger than the risk 0.33. This is because the risk is divided by (1 – risk), which is less than 1. Second, if the risk is very small, then the odds has a similar value. For example, if risk is 0.01, then the odds is practically also 0.01, as shown by the last calculation.



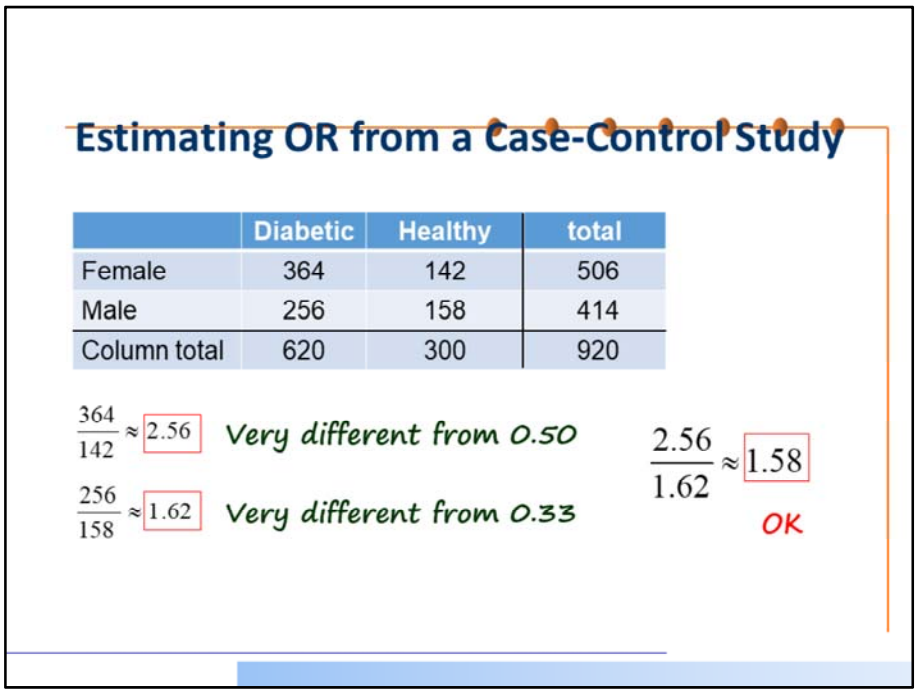
[5] We introduce the odds ratio, more specifically, the odds ratio for an event between one group to another group. In the diabetes population, the odds for diabetes are 0.50 and 0.33 respectively among the females and the males. So the OR works out to be 1.50. Note that this does not mean that the females' risk is 1.50 times higher than the males' risk, since odds is different from risk. In fact, if someone just tells us the OR is 1.50, we will not be able to calculate the RR without additional information. However, the OR being more than 1 does mean that the RR is also more than 1. In the odds ratio here, males are the so-called "baseline" group. There is nothing special about the choice of baseline. If we choose the females, then you can check that the odds ratio for diabetes between males and females is 0.33/0.50, about 0.67.



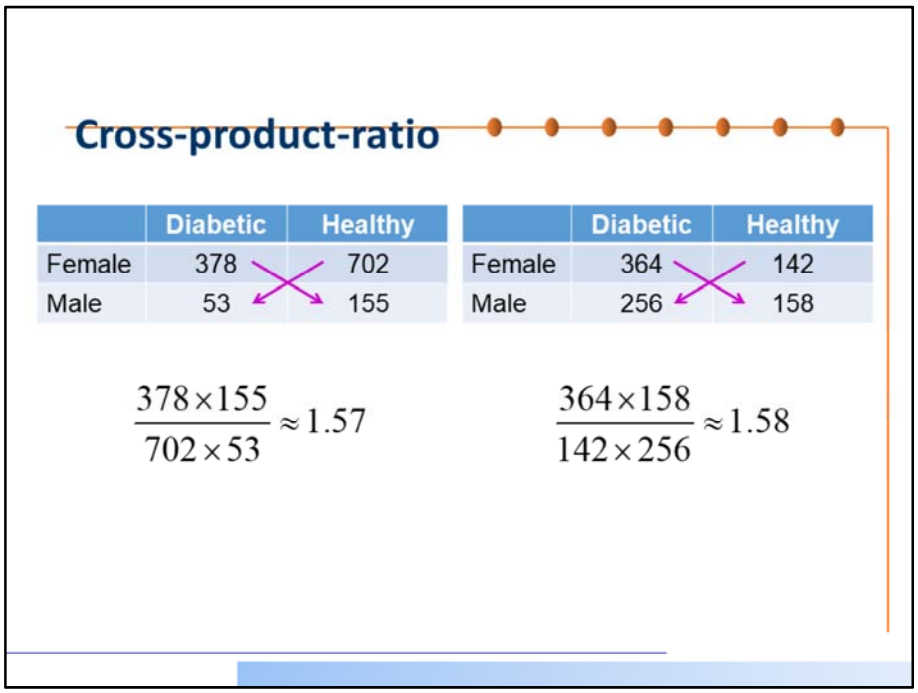
[6] This slide interprets the OR in terms of risks. If the OR for a disease between two groups has value 1, then the odds are equal, the risks are equal, so RR = 1 also. This is the only situation where a definite RR value can be deduced from a given OR value. If OR is larger than 1, the first group has higher odds than the second group, hence higher risk, so RR > 1. If OR is smaller than 1, the first group has smaller odds, hence lower risk, so RR < 1.



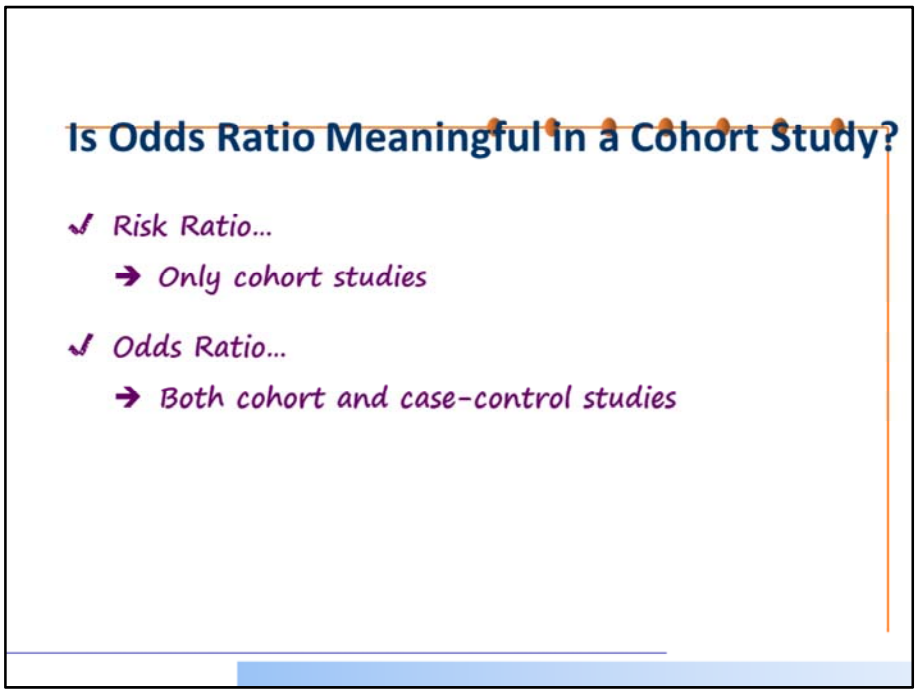
[7] So far, odds and odds ratios are population-level quantities. Now we turn to the issue of estimating these quantities from the cohort study introduced in the previous unit. Recall that from the female population of 216,000, 5 out of every 1000 were randomly chosen, and that from the male population of 208,000, 1 out of every 1000 were randomly chosen, to get the table shown here. We have seen before that the sample RR of 1.37 is quite close to the population RR of 1.33. The sample odds among females is 378/702 ~ 0.54, which is quite close to the population odds 0.50. Similarly, the sample odds among males is 53/155 ~ 0.34, quite close to the population odds 0.33. Hence the sample OR for diabetes between females and males is 0.54/0.34 ~ 1.57. This is very close to the population OR 1.50. So just like the RR, the OR can be accurately estimated from a cohort study.



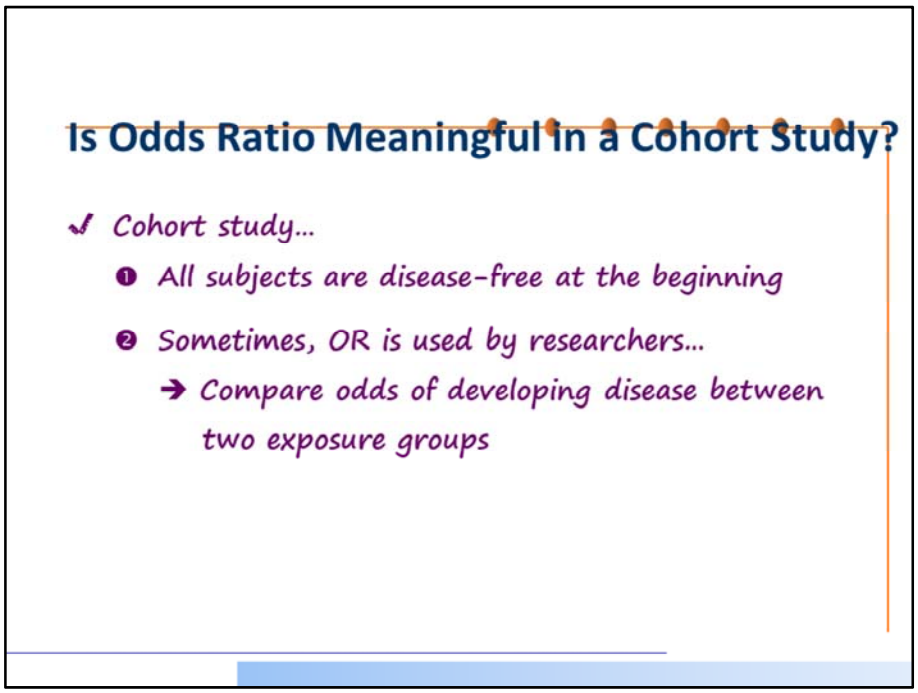
[8] Next, we look at a case-control study presented before, where from the diabetic population of 124,000, 5 out of every 1000 were randomly chosen, and from the healthy population of 300,000, 1 out of every 1000 were randomly chosen. We have seen in the previous unit that the population risks for females and for males cannot be accurately estimated from the sample. Not surprisingly, the same holds for odds. Going through the same mechanics as in the previous slide, 364/142 ~ 2.56 is very different from 0.50, the population female odds; 256/158 ~ 1.62 is also very different from 0.33, the population male odds. However, if we continue on the seemingly hopeless path, we get 2.56/1.62 ~ 1.58, which is remarkably close to the population OR 1.50. This is not a lucky break; we will likely see a similar thing if the case-control study were to be repeated, giving a new sample. So, unlike the population RR, the population OR can be estimated accurately from a case-control study, provided enough subjects are chosen. The detailed reason is beyond the scope here. Very briefly, the errors in the two numbers 2.56 and 1.62 cancel each other when taking ratio.



[9] There is a simple way to calculate the estimated OR, called the cross-product-ratio, which works the same way regardless of the study design. The table on the left is from the cohort study. As indicated by the arrows, we take the product of 378 and 155, then divide by the product of 702 and 53, to get 1.57, the same as before. The table on the right, from the case-control study, works the same way to give 1.58, also seen before. In fact this method works for the whole population as well; you may wish to practice on the table on slide 3. The most important thing to note is that the table must be set up correctly. The event of interest, diabetes in this case, should be in the first column. The first group, females in this case, should in the first row. Then the cross-product-ratio will give either an estimated, or the population, OR for the event between the first and second groups.



[10] It is crucial to identify the study design used to collect data, so that an appropriate measure, either a risk ratio or an odds ratio, can be used to investigate association. Population RR can only be accurately estimated from the sample table of a cohort study, but not from a case-control study in general. Population OR can be accurately estimated from both kinds of studies.

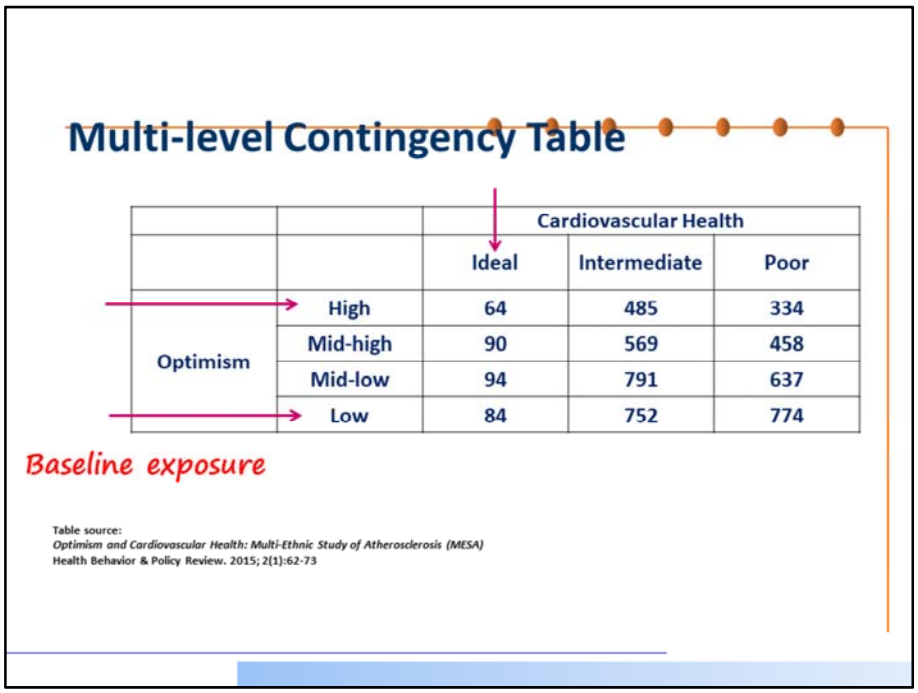


[11] In a cohort study, all subjects, by design, are disease-free at the beginning of the study. We segregate the subjects by exposure groups, for example, exposed versus unexposed groups. Therefore, what need to be compared at the endpoint of the study would be the odds of developing the disease between the two exposure groups if we apply odds ratio to a cohort study.

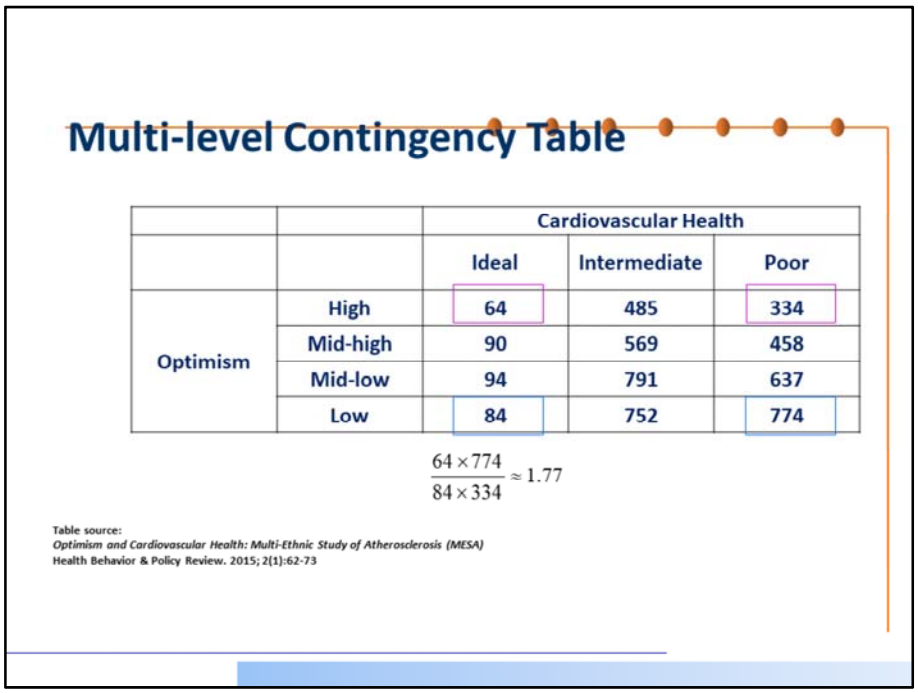
2 X 2 Contingency Table

	Diabetes (Case)	Non-diabetes (Control)	total
Female	364	142	506
Male	256	158	414
Total	620	300	920

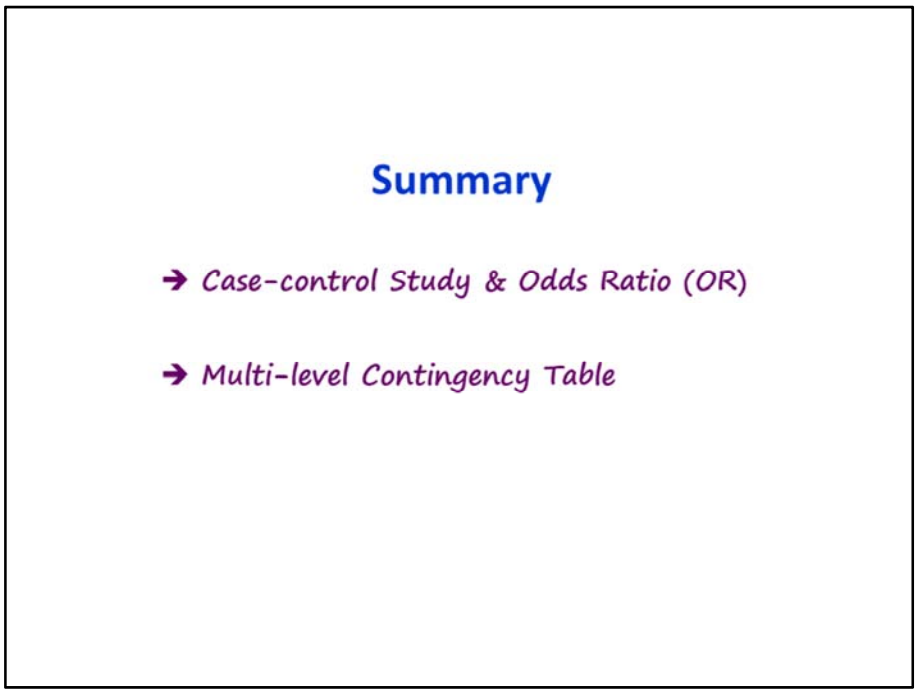
[12] So far, we have presented calculations and interpretations of odds ratio when the data are organized in a two by two table. That is, we only have two levels of exposure and two levels of disease status. What if we have three or more levels of exposure and disease status? Can we meaningfully apply odds ratio? The issue is similar for risk ratio.



[13] Let's look at a study on the relationship between optimism and cardiovascular health, or CVH, treating optimism as exposure and CVH as disease. Without going into details on how the actual measurements work, we see that optimism has four levels: low, mid-low, mid-high and high, and CVH has three levels: poor, intermediate and ideal. The essential idea is to select a 2x2 table out of the 4x3 table, according to a question of interest. For example: What is the odds ratio for ideal CVH between high and low optimism? Thus, the low optimism group is the baseline exposure group. The remaining task is to choose a baseline disease group, to compare with ideal CVH.



[14] Either intermediate or poor CVH is fine as the baseline disease group. We will choose poor CVH. Then the four highlighted numbers form the appropriate 2x2 table. Using the cross-product-formula, the odds ratio for ideal CVH (relative to poor CVH) between high and low optimism is 1.77. This confirms our expectation that the high optimism group has a lower risk of heart problems, compared to the low optimism group. 1.77 is a good estimate of the population OR provided random samples were taken from the exposure groups (a cohort study) or from the disease groups (a case-control study). Even then, this is an observational study, so on its own is not definitive proof that optimism causes better heart health. Unlike the 2x2 case, where it is often clear which groups are the baseline, in larger tables, it is important to state these groups clearly.



[15] In this unit, we have discussed the following:
In a cohort study, we compare the risks of outcomes between two exposure levels by taking the ratio of the risks to obtain a risk ratio. This measure is easily understood, however, it is NOT workable for a case-control study.
In a general case-control design, generally the sample table cannot give accurate estimate of the population RR, the exception being the case when the same fractions have been sampled from the disease groups. But the population OR can be estimated. Though OR measures the association between exposure and disease in an unfamiliar way, it is widely used in medical studies.
We also extend the odds ratio to a multilevel contingency table.
Regardless of whether the data come from a cohort or a case-control study design, it is an observational study. We need to keep in mind that association does not imply causation.