

Overview of Analytic Studies

Overview of Analytic Studies

Introduction

We search for the determinants of health outcomes, first, by relying on descriptive epidemiology to generate hypotheses about associations between exposures and outcomes. Analytic studies are then undertaken to test specific hypotheses. Samples of subjects are identified and information about exposure status and outcome is collected. The essence of an analytic study is that groups of subjects are compared in order to estimate the magnitude of association between exposures and outcomes. The two main types of analytic study design are the cohort study and the case-control study, although there are several variations on these general designs.

Learning Objectives

After successfully completing this section, the student will be able to:

- Describe the difference between descriptive and scientific/analytic epidemiologic studies in terms of information/evidence provided for medicine and public health.
- Define and explain the distinguishing features of a cohort study.
- Describe and identify the types of epidemiologic questions that can be addressed by cohort studies.
- Define and distinguish among prospective and retrospective cohort studies using the investigator as the point of reference.
- Define and explain the distinguishing features of a case-control study.
- Explain the distinguishing features of an intervention study.
- Identify the study design when reading an article or abstract.

Preview of Pre-Class Quiz #2

(Overview of Analytic Studies)



Boston University School of Public Health

Cohort Type Studies

A cohort is a "group." In epidemiology a cohort is a group of individuals who are followed over a period of time, primarily to assess what happens to them, i.e., their health outcomes. In cohort type studies one identifies individuals who do not have the outcome of interest initially, and groups them in subsets that differ in their exposure to some factor, e.g., smokers and non-smokers. The different exposure groups are then followed over time in order to compare the incidence of health outcomes, such as lung cancer or heart disease. As an example, the Framingham Heart Study enrolled a cohort of 5,209 residents of Framingham, MA who were between the ages of 30-62 and who did not have cardiovascular disease when they were enrolled. These subjects differed from one another in many ways: whether they smoked, how much they smoked, body mass index, eating habits, exercise habits, gender, family history of heart disease, etc. The researchers assessed these and many other characteristics or "exposures" soon after the subjects had been enrolled and before any of them had developed cardiovascular disease. The many "baseline characteristics" were assessed in a number of ways including questionnaires, physical exams, laboratory tests, and imaging studies (e.g., x-rays). They then began "following" the cohort, meaning that they kept in contact with the subjects by phone, mail, or clinic visits in order to determine if and when any of the subjects developed any of the "outcomes of interest," such as myocardial infarction (heart attack), angina, congestive heart failure, stroke, diabetes and many other cardiovascular outcomes.

Over time some subjects eventually began to develop some of the outcomes of interest. Having followed the cohort in this fashion, it was eventually possible to use the information collected to evaluate many hypotheses about what characteristics were associated with an increased risk of heart disease. For example, if one hypothesized that smoking increased the risk of heart attacks, the subjects in the cohort could be sorted based on their smoking habits, and one could compare the subset of the cohort that smoked to the subset who had never smoked. For each such comparison that one wanted to make the cohort could be grouped according to whether they had a given exposure or not, and one could measure and compare the frequency of heart attacks (i.e., the incidence) between the groups. Incidence provides an estimate of risk, so if the incidence of heart attacks is 3 times greater in smokers compared to non-smokers, it suggests an association between smoking and risk of developing a heart attack. (Various biases might also be an explanation for an apparent association. We will learn about these later in the course.) The hallmark of analytical studies, then, is that they collect information about both exposure status and outcome status, and they compare groups to identify whether there appears to be an association or a link.

The Population "At Risk"

From the discussion above, it should be obvious that one of the basic requirements of a cohort type study is that none of the subjects have the outcome of interest at the beginning of the follow-up period, and time must pass in order to determine the frequency of developing the outcome.

- For example, if one wanted to compare the risk of developing uterine cancer between postmenopausal women receiving hormone-replacement therapy and those not receiving hormones, one would consider certain eligibility criteria for the members prior to the start of the study: 1) they should be female, 2) they should be post-menopausal, and 3) they should have a uterus. Among post-menopausal women there might be a number who had had a hysterectomy already, perhaps for persistent bleeding problems or endometriosis. Since these women no longer have a uterus, one would want to exclude them from the cohort, because they are no longer at risk of developing this particular type of cancer.
- Similarly, if one wanted to compare the risk of developing diabetes among nursing home residents who exercised and those who did not, it would be important to test the subjects for diabetes at the beginning of the follow-up period in order to exclude all subjects who already had diabetes and therefore were not "at risk" of developing diabetes.

Eligible subjects have to meet certain criteria to be included as subjects in a study (inclusion criteria). One of these would be that they did not have any of the diseases or conditions that the investigators want to study, i.e., the subjects must be "at risk," of developing the outcome of interest, and the members of the cohort to be followed are sometimes referred to as "the population at risk."

However, at times decisions about who is "at risk" and eligible get complicated.

Example #1: Suppose the outcome of interest is development of measles. There may be subjects who:

- Already were known to have had clinically apparent measles and are immune to subsequent measles infection
- Had sub-clinical cases of measles that went undetected (but the subject may still be immune)
- Had a measles vaccination that conferred immunity
- Had a measles vaccination that failed to confer immunity

In this case the eligibility criteria would be shaped by the specific scientific questions being asked. One might want to compare subjects known to have had clinically apparent measles to those who had not had clinical measles and had not had a measles vaccination. Or, one could take blood sample from all potential subjects in order to measure their antibody titers (levels) to the measles virus.

Example #2: Suppose you are studying an event that can occur more than once, such as a heart attack. Again, the eligibility criteria should be shaped to fit the scientific questions that are being answered. If one were interested in the risk of a first myocardial infarction, then obviously subjects who had already had a heart attack would not be eligible for study. On the other hand, if one were interested in tertiary prevention of heart attacks, the study cohort would include people who had had heart attacks or other clinical manifestations of heart disease, and the outcome of interest would be subsequent significant cardiac events or death.

Prospective and Retrospective Cohort Studies

Cohort studies can be classified as prospective or retrospective based on when outcomes occurred in relation to the enrollment of the cohort.

Prospective Cohort Study

The Framingham Heart Study, described above, and the Nurses Health Study are good examples of prospective cohort studies.



The **DISTINGUISHING FEATURE** of a prospective cohort study is that at the time that the investigators begin enrolling subjects and

collecting baseline exposure information, none of the subjects has developed any of the outcomes of interest.

In a prospective study like the Nurses Health Study baseline information is collected from all subjects in the same way using exactly the same questions and data collection methods for all subjects. The investigators design the questions and data collection procedures carefully in order to

obtain accurate information about exposures *before* disease develops in any of the subjects. After baseline information is collected, subjects in a prospective cohort study are then followed "longitudinally," i.e. over a period of time, usually for years, to determine if and when they become diseased and whether their exposure status changes. In this way, investigators can eventually use the data to answer many questions about the associations between "risk factors" and disease outcomes. For example, one could identify smokers and non-smokers at baseline and compare their subsequent incidence of developing heart disease. Alternatively, one could group subjects based on their body mass index (BMI) and compare their risk of developing heart disease or cancer.

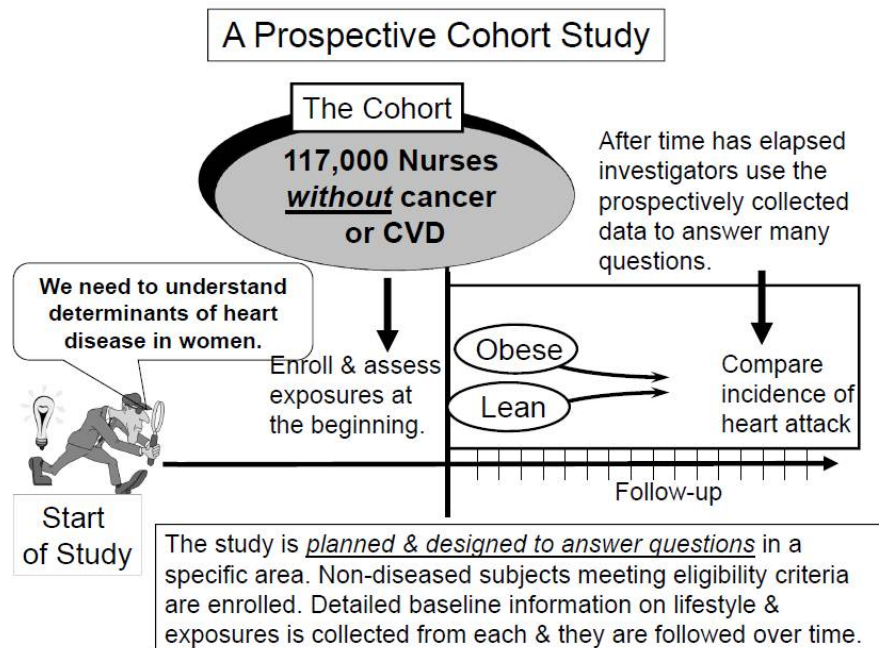


Pitfall: Note that in these prospective cohort studies a comparison of incidence between the groups can only take place after enough time has elapsed so that some subjects developed the outcomes of interest. Since the *data analysis* occurs after some outcomes have occurred, some students mistakenly would call this a retrospective study, but this is incorrect. The analysis always occurs after a certain number of events have taken place. The characteristic that distinguishes a

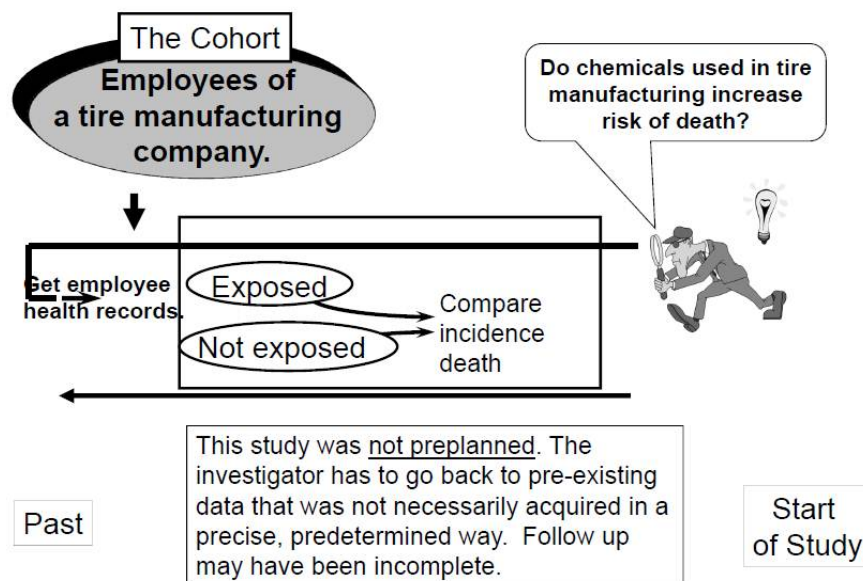
study as prospective is that the subjects were enrolled, and baseline data was collected *before* any subjects developed an outcome of interest.

Retrospective Cohort Study

In contrast, retrospective studies are conceived **AFTER** some people have already developed the outcomes. The investigators find the subjects and begin collecting information about them after outcomes have already occurred. Suppose investigators wanted to test the hypothesis that working with the chemicals involved in



A Retrospective Cohort Study



tire manufacturing increases the risk of death. Since this is a fairly rare exposure, it would be advantageous to use a special exposure cohort such as employees of a large tire manufacturing factory. The employees who actually worked with chemicals used in the manufacturing process would be the exposed group, while clerical workers and management might constitute the "unexposed" group. However, rather than following these subjects for decades, it would be more efficient to use employee health and employment records over the past two or three decades as a source of data. In essence, the investigators are jumping back in time to identify the study cohort. They can classify them as "exposed" or "unexposed" based on their employment records, and they can use a number of sources to determine outcome status, such as death (e.g., using health records, next of kin, National Death Index, etc.).



The **DISTINGUISHING FEATURE** of a retrospective cohort study is that the investigators conceive the study and begin identifying and enrolling subjects *after outcomes have already occurred*.

Retrospective cohort studies like this are very efficient for studying rare or unusual exposures, but there are many potential problems here. Sometimes exposure status is not clear when it is necessary to go back in time and use whatever data is available, especially because the data being used was not designed to answer a health question. Even if it was clear who was exposed to tire manufacturing chemicals based on employee records, it would also be important to take into account (or adjust for) other differences that could have influenced mortality, i.e., confounding factors. For example, it might be important to know whether the subjects smoked, or drank, or what kind of diet they ate. However, it is unlikely that a retrospective cohort study would have accurate information on these many other risk factors.

Intervention Studies (Clinical Trials)

Intervention studies (clinical trials) are experimental research studies that compare the effectiveness of medical treatments, management strategies, prevention strategies, and other medical or public health interventions. Their design is very similar to that of a prospective cohort study. However, in cohort studies exposure status is determined by genetics, self-selection, or life circumstances, and the investigators just observe differences in outcome between those who have a given exposure and those who do not. In clinical trials exposure status (the treatment type) is assigned by the investigators. Ideally, assignment of subjects to one of the comparison groups should be done randomly in order to produce equal distributions of potentially confounding factors. Sometimes a group receiving a new treatment is compared to an untreated group, or a group receiving a placebo or a sham treatment. Sometimes, a new treatment is compared to an untreated group or to a group receiving an established treatment. For more on this topic see the module on Intervention Studies.

In summary, the characteristic that distinguishes a clinical trial from a cohort study is that the investigator assigns the exposure status in a clinical trial, while subjects' genetics, behaviors, and life circumstances determine their exposures in a cohort study.

Key Features of Cohort Type Studies



In general, there are three distinguishing features common to all three types of cohort studies:

1. None of the subjects have the outcome of interest at the beginning of the follow-up period.
2. The groups being compared differ in their exposure status.
3. One measures and compares the incidence of the outcome in order to determine whether there is an association between the exposure and the outcome.

Summarizing Data in a Cohort Study

Investigators often use contingency tables to summarize data. In essence, the table is a matrix that displays the combinations of exposure and outcome status. If one were summarizing the results of a study with two possible exposure categories and two possible outcomes, one would use a "two by two" table in which the numbers in the four cells indicate the number of subjects within each of the 4 possible categories of risk and disease status.

For example, consider data from a retrospective cohort study conducted by the Massachusetts Department of Public Health (MDPH) during an investigation of an outbreak of *Giardia lamblia* in Milton, MA in 2003. The descriptive epidemiology indicated that almost all of the cases belonged to a country club in Milton. The club had an adult swimming pool and a wading pool for toddlers, and the investigators suspected that the outbreak may have occurred when an infected child with a dirty diaper contaminated the water in the kiddy pool. This hypothesis was tested by conducting a retrospective cohort study. The cases of *Giardia lamblia* had already occurred and had been reported to MDPH via the infectious disease surveillance system (for more information on surveillance, see the Surveillance module). The investigation focused on an obvious cohort - 479 members of the country club who agreed to answer the MDPH questionnaire. The questionnaire asked, among many other things, whether the subject had been exposed to the kiddy pool. The incidence of subsequent *Giardia* infection was then compared between subjects who been exposed to the kiddy pool and those who had not.

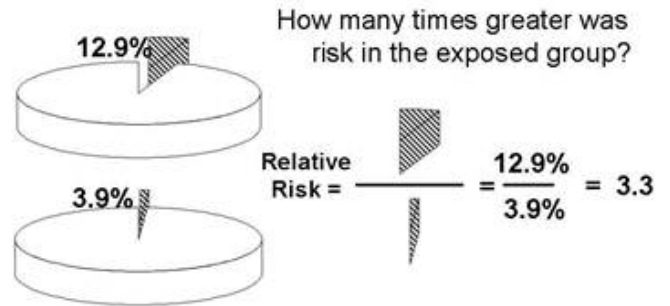
Summarizing Data with a 2 x 2 Table

		Got <i>Giardiasis</i>		Cumulative Incidence
		Yes	No	
Exposed to Kiddy Pool	Yes	16	108	124 12.9%
	No	14	341	355 3.9%
		30	449	479 subjects

The 2 x 2 table to the right summarizes the findings. A total of 479 subjects completed the questionnaire, and 124 of them indicated that they had been exposed to the kiddy pool. Of these, 16 subsequently developed *Giardia* infection, but 108 did not. Among the 355 subjects who denied kiddy pool exposure, 14 developed *Giardia* infection, and the other 341 did not.

Organization of the data this way makes it easy to compute the cumulative incidence in each group (12.9% and 3.9% respectively). The incidence in each group provides an estimate of risk, and the groups can be compared in order to estimate the magnitude of association. (This will be addressed in much greater detail in the module on Measures of Association.) One way of quantifying the association is to calculate the relative risk, i.e., dividing the incidence in the exposed group by the incidence in the unexposed group). In this case, the risk ratio is $(12.9\% / 3.9\%) = 3.3$. This suggest that subjects who swam in the kiddy pool had 3.3 times the risk of getting *Giardia*

Comparing Incidence - Relative Risk



"The risk of giardiasis was 3.3 *times greater* in people who swam in the kiddy pool compared to those who did not."

infections compared to those who did not, suggesting that the kiddy pool was the source.

Unanswered Questions



If the kiddy pool was the source of contamination responsible for this outbreak, why was it that:

1. Only 16 people exposed to the kiddy pool developed the infection?
2. There were 14 Giardia cases among people who denied exposure to the kiddy pool?

Before you look at the answer, think about it and try to come up with a possible explanation.

Likely Explanation

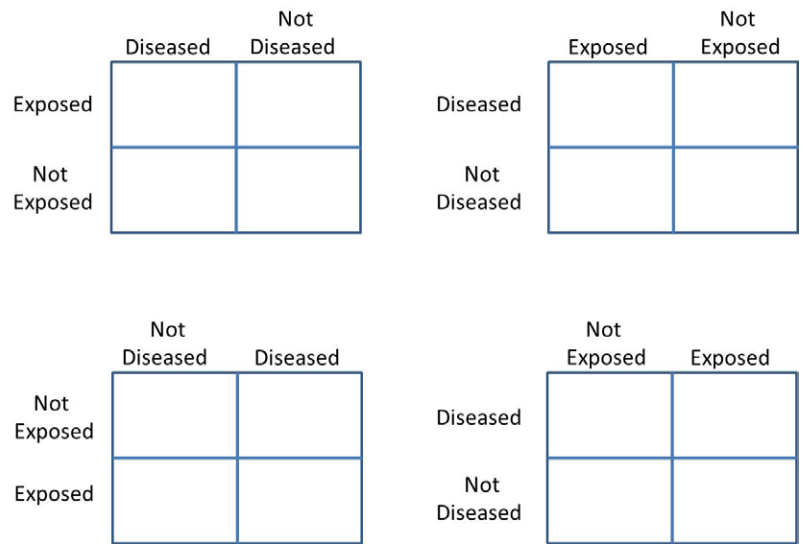
Optional Links of Interest

- The 2003 Giardia outbreak
- CDC: Organizing Data



Pitfall: Contingency tables can be oriented in several ways; some of the possibilities are shown to the right. The two tables in the top row are probably the most common, but you see all kinds. There is no standard rule about which arrangement to use. However, when you begin to use these

tables to calculate incidence estimates and measures of association, the formulas you use will depend on the orientation of the table. This can generate a lot of confusion and incorrect answers. Because of this, I almost always set up my contingency tables the same way, so I don't get confused. I use the orientation shown in the upper left hand cell in the illustration to the right. Specifically, I list outcome status in columns, with the first column designated for those who had the outcome, and I list the exposure status in rows, with the most exposed group on top and the least exposed group on the bottom. I will try to consistently use this format throughout the course in order to minimize confusion. However, be aware that other authors use different orientations.



Case-Control Studies

Cohort studies have an intuitive logic to them, but they can be very problematic when:

1. The outcomes being investigated are rare;
2. There is a long time period between the exposure of interest and the development of the disease; or
3. It is expensive or very difficult to obtain exposure information from a cohort.

In the first case, the rarity of the disease requires enrollment of very large numbers of people. In the second case, the long period of follow-up requires efforts to keep contact with and collect outcome information from individuals. In all three situations, cost and feasibility become an important concern.

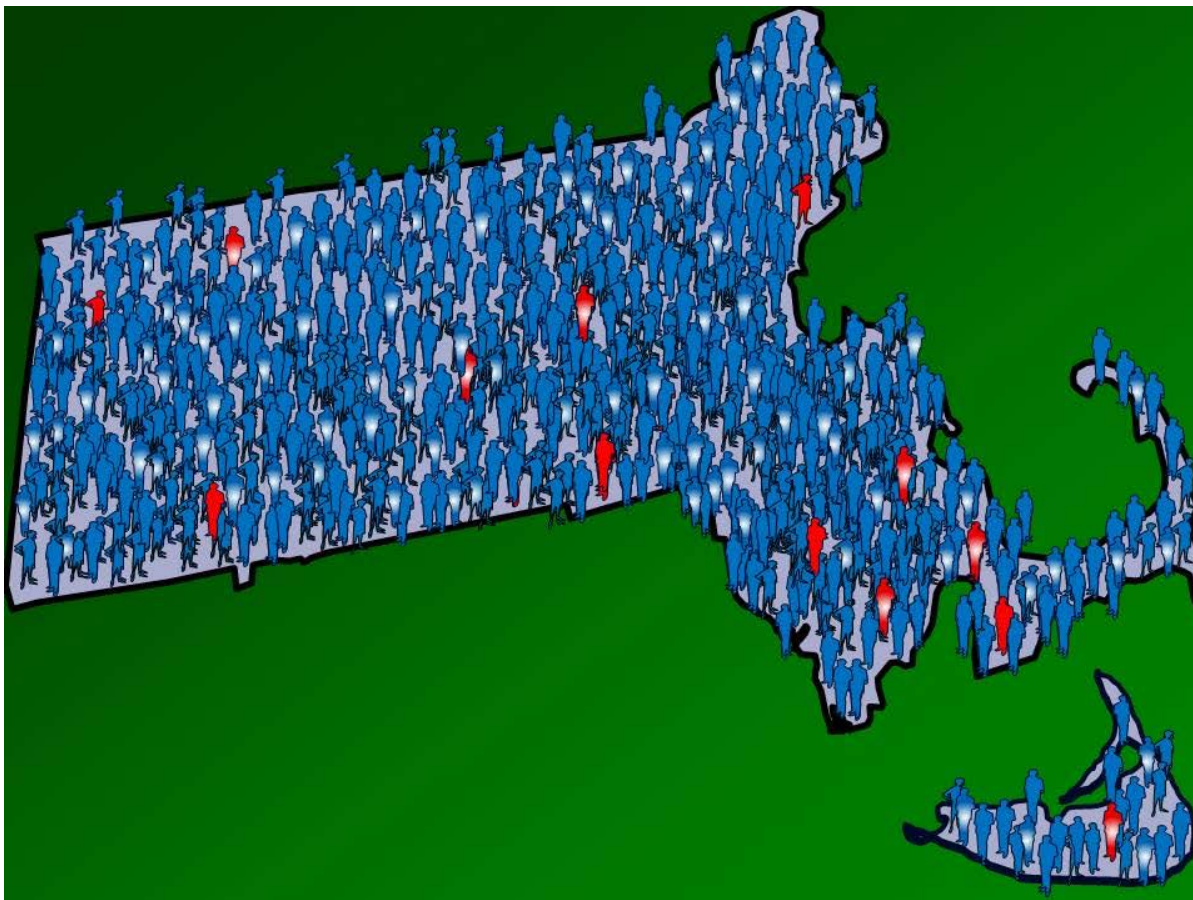
A case-control design offers an alternative that is much more efficient. The goal of a case-control study is the same as that of cohort studies, i.e. to estimate the magnitude of association between an exposure and an outcome. However, case-control studies employ a different sampling strategy that gives them greater efficiency. As with a cohort study, a case-control study attempts to identify all people who have developed the disease of interest in the defined population. This is not because they are inherently more important to estimating an association, but because they are almost always rarer than non-diseased individuals, and one of the requirements of accurate estimation of the association is that there are reasonable numbers of people in both the numerators (cases) and denominators (people or person-time) in the measures of disease frequency for both exposed and reference groups. However, because most of the denominator is made up of people who do not develop disease, the case-control design avoids the need to collect information on the entire population by selecting a sample of the underlying population.

Rothman describes the case-control strategy as follows:

"Case-control studies are best understood by considering as the starting point a *source population*, which represents a hypothetical study population in which a cohort study might have been conducted. The *source population* is the population that gives rise to the cases included in the study. If a cohort study were undertaken, we would define the exposed and unexposed cohorts (or several cohorts) and from these populations obtain denominators for the incidence rates or risks that would be calculated for each cohort. We would then identify the number of cases occurring in each cohort and calculate the risk or incidence rate for each. In a case-control study the same cases are identified and classified as to whether they belong to the exposed or unexposed cohort. Instead of obtaining the denominators for the rates or risks, however, a control group is sampled from the entire source population that gives rise to the cases. Individuals in the control group are then classified into exposed and unexposed categories. The purpose of the control group is to determine the relative size of the exposed and unexposed components of the source population."

To illustrate this consider the following hypothetical scenario in which the source population is the state of Massachusetts. Diseased individuals are red, and non-diseased individuals are blue. Exposed individuals are indicated by a whitish midsection. Note the following aspects of the depicted scenario:

1. The outcome being investigated is rare.
2. There is a fairly large number of exposed individuals in the state, but most of these are not diseased.
3. The proportion of exposed individuals among the disease cases (7/13) is higher than the proportion of exposure among the controls.



If I somehow had exposure and outcome information on all of the subjects in the source population and looked at the association using a cohort design, it might look like this:

	Diseased	Non-diseased	Total
Exposed	7	1,000	1,007
Non-exposed	6	5,634	5,640

Therefore, the incidence in the exposed individuals would be $7/1,007 = 0.70\%$, and the incidence in the non-exposed individuals would be $6/5,640 = 0.11\%$. Consequently, the risk ratio would be $0.70/0.11=6.52$, suggesting that those who had the risk factor (exposure) had 6.5 times the risk of getting the disease compared to those without the risk factor. This is a strong association.

In this hypothetical example, I had data on all 6,647 people in the source population, and I could compute the probability of disease (i.e., the risk or incidence) in both the exposed group and the non-exposed group, because I had the denominators for both the exposed and non-exposed groups.

The problem, of course, is that I usually don't have the resources to get the data on all subjects in the population. If I

took a random sample of even 5-10% of the population, I might not have any diseased people in my sample.

An alternative approach would be to use surveillance databases or administrative databases to find most or all 13 of the cases in the source population and determine their exposure status. However, instead of enrolling all of the other 5,634 residents, suppose I were to just take a sample of the non-diseased population. In fact, suppose I only took a sample of 1% of the non-diseased people and I then determined their exposure status. The data might look something like this:

	Diseased	Non-diseased	Total
Exposed	7	10	unknown
Non-exposed	6	56	unknown

With this sampling approach I can no longer compute the probability of disease in each exposure group, because I no longer have the denominators in the last column. In other words, I don't know the exposure distribution for the entire source population. However, the small control sample of non-diseased subjects gives me a way to estimate the exposure distribution in the source population. So, I can't compute the probability of disease in each exposure group, but I can compute the odds of disease in each group.

The Odds Ratio

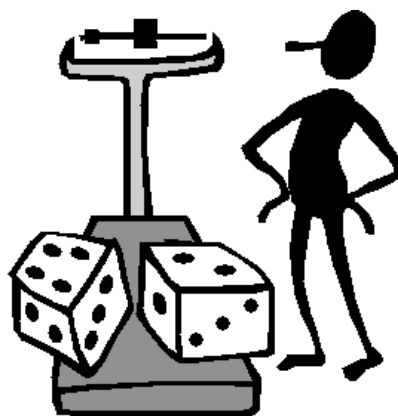
The odds of disease in the exposed group are 7/10, and the odds of disease in the non-exposed group are 6/56. If I compute the odds ratio, I get $(7/10) / (6/56) = 6.56$, very close to the risk ratio that I computed from data for the entire population. We will consider odds ratios and case-control studies in much greater depth in a later module. However, for the time being the key things to remember are that:

1. The sampling strategy for a case-control study is very different from that of cohort studies, despite the fact that both have the goal of estimating the magnitude of association between the exposure and the outcome.
2. In a case-control study there is no "follow-up" period. One starts by identifying diseased subjects and determines their exposure distribution; one then takes a sample of the source population that produced those cases in order to estimate the exposure distribution in the overall source population that produced the cases. [In cohort studies none of the subjects have the outcome at the beginning of the follow-up period.]
3. In a case-control study, you cannot measure incidence, because you start with diseased people and non-diseased people, so you cannot calculate relative risk.
4. The case-control design is very efficient. In the example above the case-control study of only 79 subjects produced an odds ratio (6.56) that was a very close approximation to the risk ratio (6.52) that was obtained from the data in the entire population.
5. Case-control studies are particularly useful when the outcome is rare is uncommon in both exposed and non-exposed people.

The Difference Between "Probability" and "Odds"?

- The probability that an event will occur is the fraction of times you expect to see that event in many trials. Probabilities always range between 0 and 1.
- The odds are defined as the probability that the event will occur divided by the probability that the event will not occur.

If the **probability** of an event occurring is Y , then the probability of the event not occurring is $1-Y$. (Example: If the probability of an event is 0.80 (80%), then the probability that the event will not occur is $1-0.80 = 0.20$, or 20%.



The **odds** of an event represent the ratio of the (probability that the event will occur) / (probability that the event will not occur). This could be expressed as follows:

$$\text{Odds of event} = Y / (1-Y)$$

So, in this example, if the probability of the event occurring = 0.80, then the odds are $0.80 / (1-0.80) = 0.80/0.20 = 4$ (i.e., 4 to 1).

- If a race horse runs 100 races and wins 25 times and loses the other 75 times, the probability of winning is $25/100 = 0.25$ or 25%, but the odds of the horse winning are $25/75 = 0.333$ or 1 win to 3 losses.
- If the horse runs 100 races and wins 5 and loses the other 95 times, the probability of winning is 0.05 or 5%, and the odds of the horse winning are $5/95 = 0.0526$.
- If the horse runs 100 races and wins 50, the probability of winning is $50/100 = 0.50$ or 50%, and the odds of winning are $50/50 = 1$ (even odds).
- If the horse runs 100 races and wins 80, the probability of winning is $80/100 = 0.80$ or 80%, and the odds of winning are $80/20 = 4$ to 1.

NOTE that when the probability is low, the odds and the probability are very similar.



On Sept. 8, 2011 the New York Times ran an article on the economy in which the writer began by saying "If history is a guide, the odds that the American economy is falling into a double-dip recession have risen sharply in recent weeks and may even have reached 50 percent."

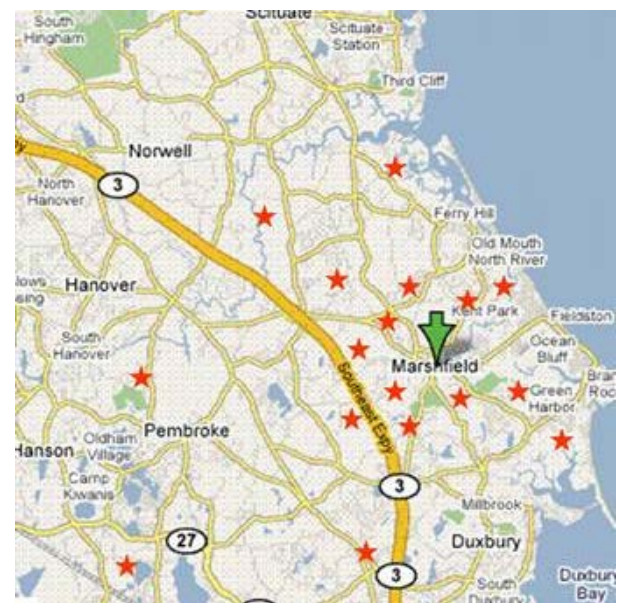
Further down in the article the author quoted the economist who had been interviewed for the story. What the economist had actually said was, "Whether we reach the technical definition [of a double-dip recession] I think is probably close to 50-50."

Question: was the author correct in saying that the "odds" of a double-dip recession may have reached 50 percent?

ANSWER

Hepatitis Outbreak in Marshfield, MA

In 2004 there was an outbreak of hepatitis A on the South Shore of Massachusetts. Over a period of a few weeks there were 20 cases of hepatitis A that were reported to the MDPH, and most of the infected persons were residents of Marshfield, MA. Marshfield's health department requested help in identifying the source from MDPH. The investigators quickly performed descriptive epidemiology. The epidemic curve indicated a point source epidemic, and most of the cases lived in the Marshfield area, although some lived as far away as Boston. They conducted hypothesis-generating interviews, and taken together, the descriptive epidemiology suggested that the source was one of five or six food establishments in the Marshfield area, but it wasn't clear which one. Consequently, the investigators wanted to conduct an analytic study to determine which restaurant was the source.

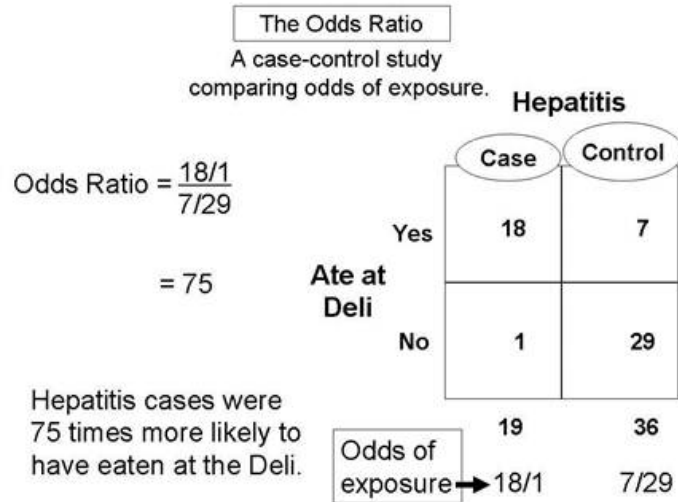




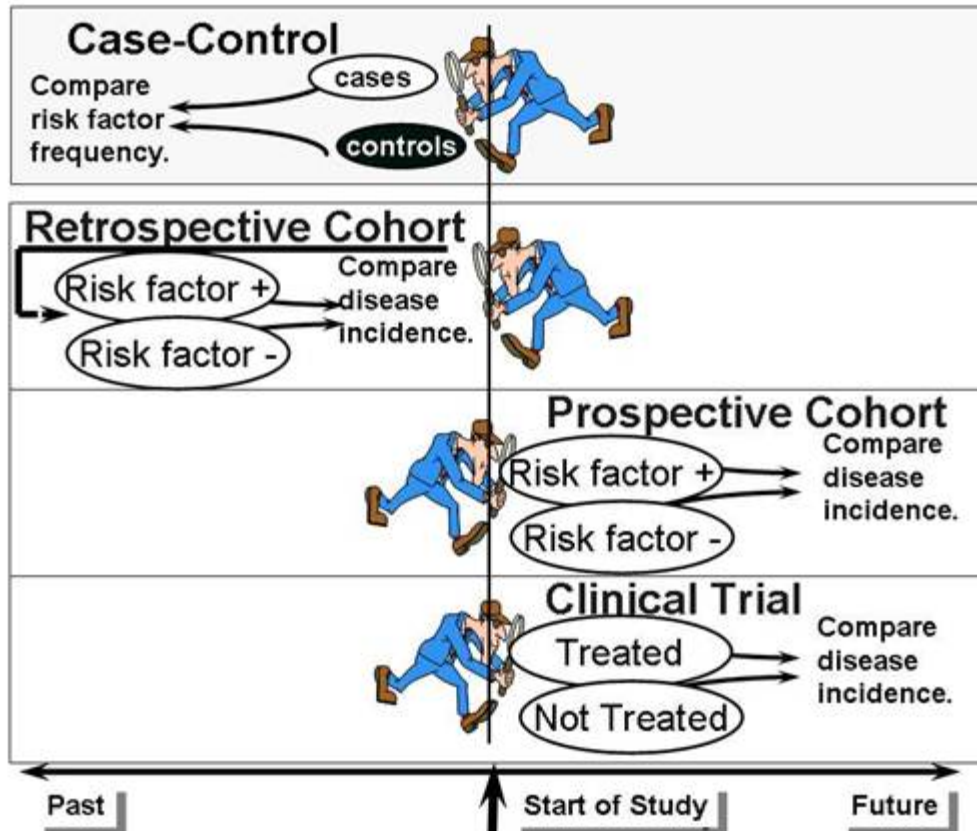
Quiz Me



They invited all 20 cases of hepatitis A to answer questions from a questionnaire designed for this study, and 19 of the cases agreed to complete the survey.



Summary



Note that the lower three study designs (retrospective and prospective cohort studies and clinical trials) are similar in that an initially disease free cohort is divided into groups based on their "exposure" status, i.e., whether or not they have a particular "risk factor," and for all three, the investigator measures and compares the incidence of disease. In contrast, case-control studies identify diseased and non-diseased subjects and then measure and compare their likelihood of having had certain prior exposures.

Which Study Design Is Best?

There are some situations in which more than one study design could be used.

Smoking and Lung Cancer: For example, when investigators first sought to establish whether there was a link between smoking and lung cancer, they did a study by finding hospital subjects who had lung cancer and a comparison group of hospital patients who had diseases other than cancer. They then compared the prior exposure histories with respect to smoking and many other factors. They found that past smoking was much more common in the lung cancer cases, and they concluded that there was an association. The advantages to this approach were that they were able to collect the data they wanted relatively quickly and inexpensively, because they started with people who already had the disease of interest.



Quiz Me



However, there were several limitations to the study they had done. The study design did not allow them to measure the incidence of lung cancer in smokers and non-smokers, so they couldn't measure the absolute risk of smoking. They also didn't know what other diseases smoking might be associated with, and, finally, they were concerned about some of the biases that can creep into this type of study.

As a result, these investigators then initiated another study. They invited all of the male physicians in the United Kingdom to fill out questionnaires regarding their health status and their smoking status. They then focused on the healthy physicians who were willing to participate, and the investigators mailed follow-up questionnaires to them every few years. They also had a way of finding out the cause of death for any subjects who became ill and died. The study continued for about 50 years. Along the way the investigators periodically compared the incidence of death among non-smoking physicians and physicians who smoked small, moderate or heavy amounts of tobacco.



Quiz Me



These studies were useful, because they were able to demonstrate that smokers had an increased risk of over 20 different causes of death. They were also able to measure the incidence of death in different categories, so they knew the absolute risk for each cause of death. Of course, the downside to this approach was that it took a long time, and it was very costly.

So, both a case-control study and a prospective cohort study provided useful information about the association between

smoking and lung cancer and other diseases, but there were distinct advantages and limitations to each approach.

However, there are two very important distinctions to make here.



1. **In a study that is designed and conducted as a case-control study, you cannot calculate incidence.** Therefore, you cannot calculate relative risk. You can only calculate an odds ratio. Consider the first case-control study comparing lung cancer patients to patients without lung cancer. There was no discrete cohort. They found subjects with the disease of interest and a sample of subjects without lung cancer and compared their past use of tobacco (the exposure of interest).

2. In certain situations a case-control study is the only feasible thing to do.

Case-control studies are particularly efficient for rare diseases because they begin by identifying a sufficient number of diseased people (or people have some "outcome" of interest) to enable you to do an analysis that tests associations. Case-control studies can be done in just about any circumstance, but they are particularly useful when you are dealing with rare diseases or disease for which there is a very long latent period, i.e. a long time between the causative exposure and the eventual development of disease.

In general, the decision regarding which study design to use rests on a number of factors that include:

For a video review of the concepts presented in this module you may want to watch the video below. It provides a 26 minute review of the class lecture material.

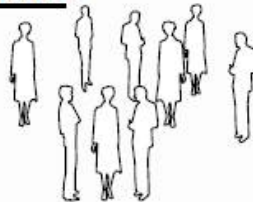
 **Overview of Analytic Studies** (click the camera icon)



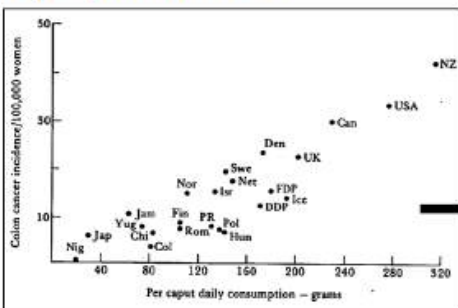
Tips on Identifying a Study Design

Did they collect exposure and outcome information on individual people, or was the data based on average exposure in a group, e.g., per capita expenditures for soft drinks or tobacco? If it is the latter, then you are probably dealing with an ecological study.

Is it based on information about individuals?



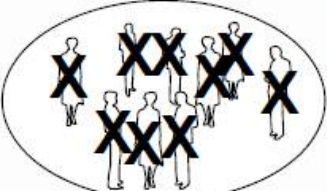
Or averages in populations?



→ **Correlational (Ecologic)**

Was a hypothesis being tested by comparing two or more groups? Or was the focus on a single group of people with a common circumstance. For example, were all of the subjects HIV positive individuals who were all treated with antiretrovirals? And was the report basically an elaborate description of how well it worked? Even if a study like this has a large number of subjects, if there is no comparison group, then it is a case series. Remember that a case series is description of a single group consisting of anywhere from a few to thousands of subjects.

Is there just one group?



8 people with bird flu

Did all subjects have the disease? (**Case Series**)

Did they evaluate presence of disease and risk factors at the same point in time?

(**Cross-sectional Survey**)

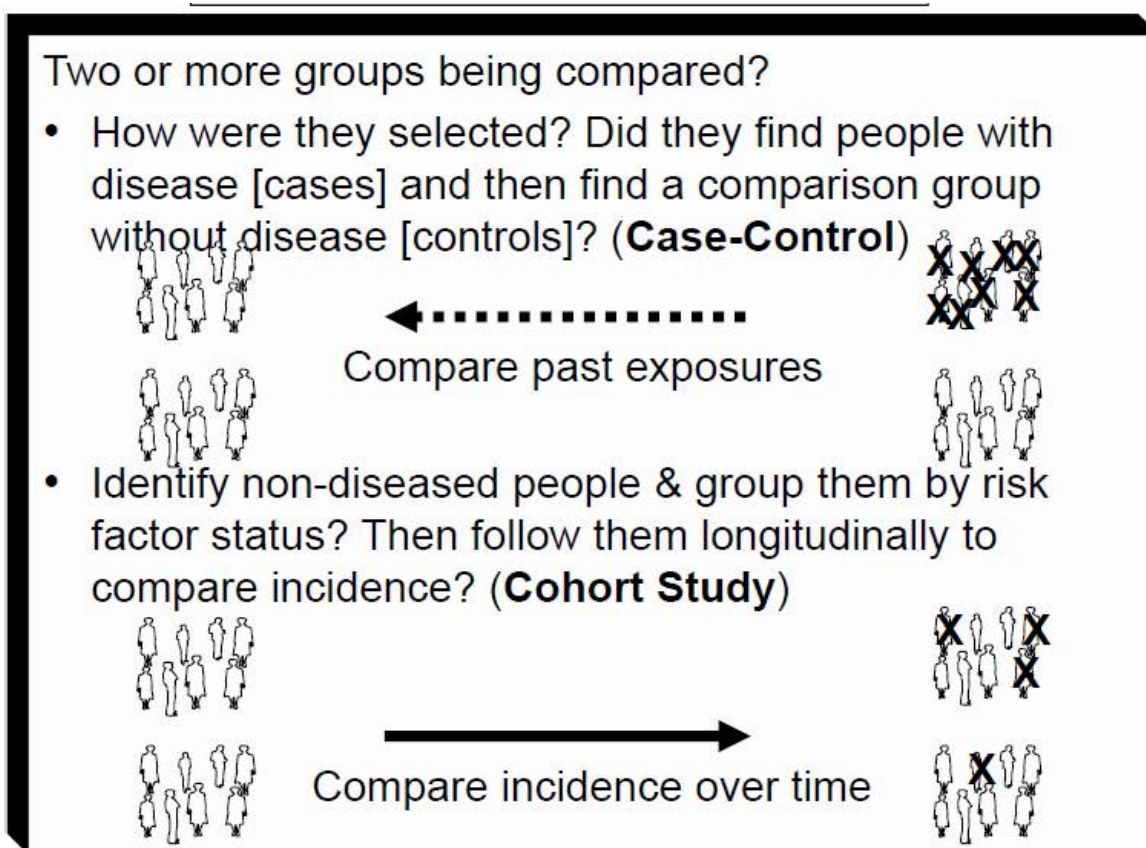
Do you have heart disease?
Are you active?

↓

→

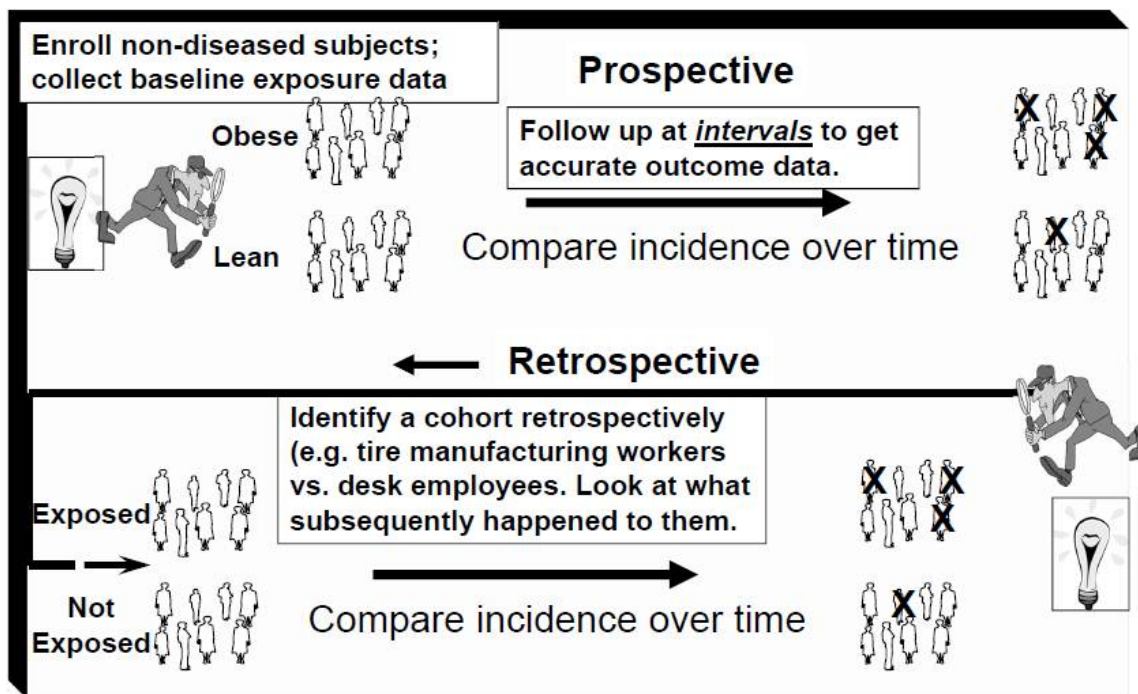
Did the investigators identify a group at risk, assess their exposure status, and then follow them over time (a longitudinal study)? Or did the investigators assess the prevalence of exposures and health outcomes at a single point in time (cross-sectional study)?

Was there a pre-planned comparison of two or more groups? If so, did they enroll subjects based on their disease status and then assess prior exposures (case-control study)? Or did they identify a non-diseased cohort, assess their exposure status, and then following longitudinally in order to compare the frequency of health outcomes?

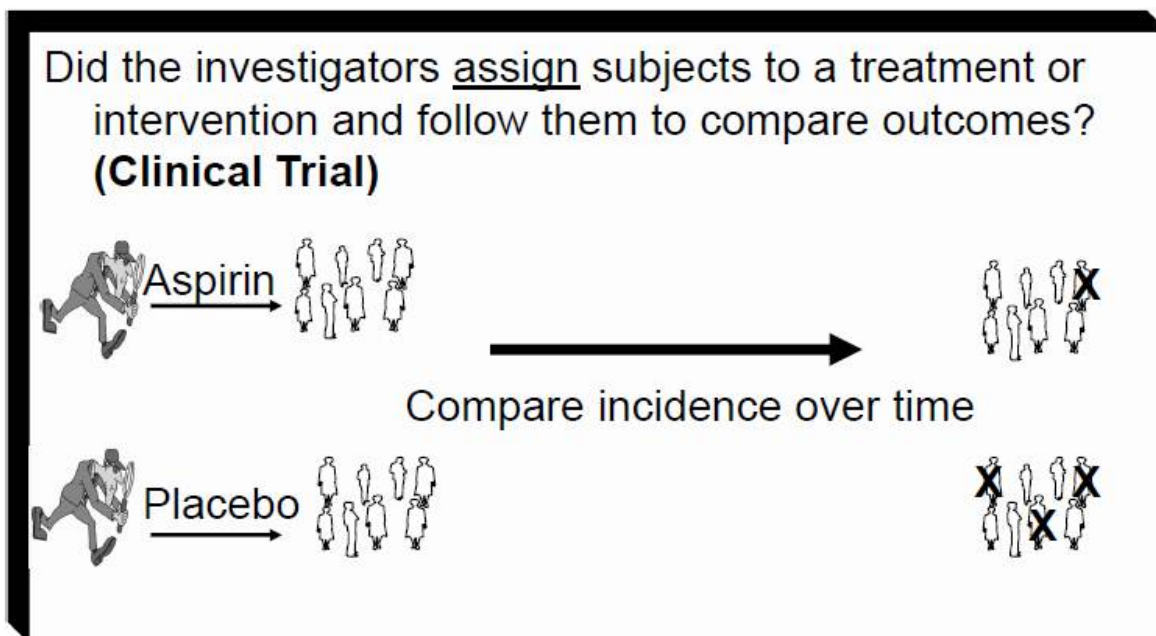


If it was a cohort type of study, was it prospective or retrospective? Was the study designed prospectively, and did they enroll subjects who were "at risk", i.e., did not have the outcome of interest at the time of enrollment (prospective cohort study)? Or were subjects enrolled retrospectively after some had developed the outcome already (retrospective)?

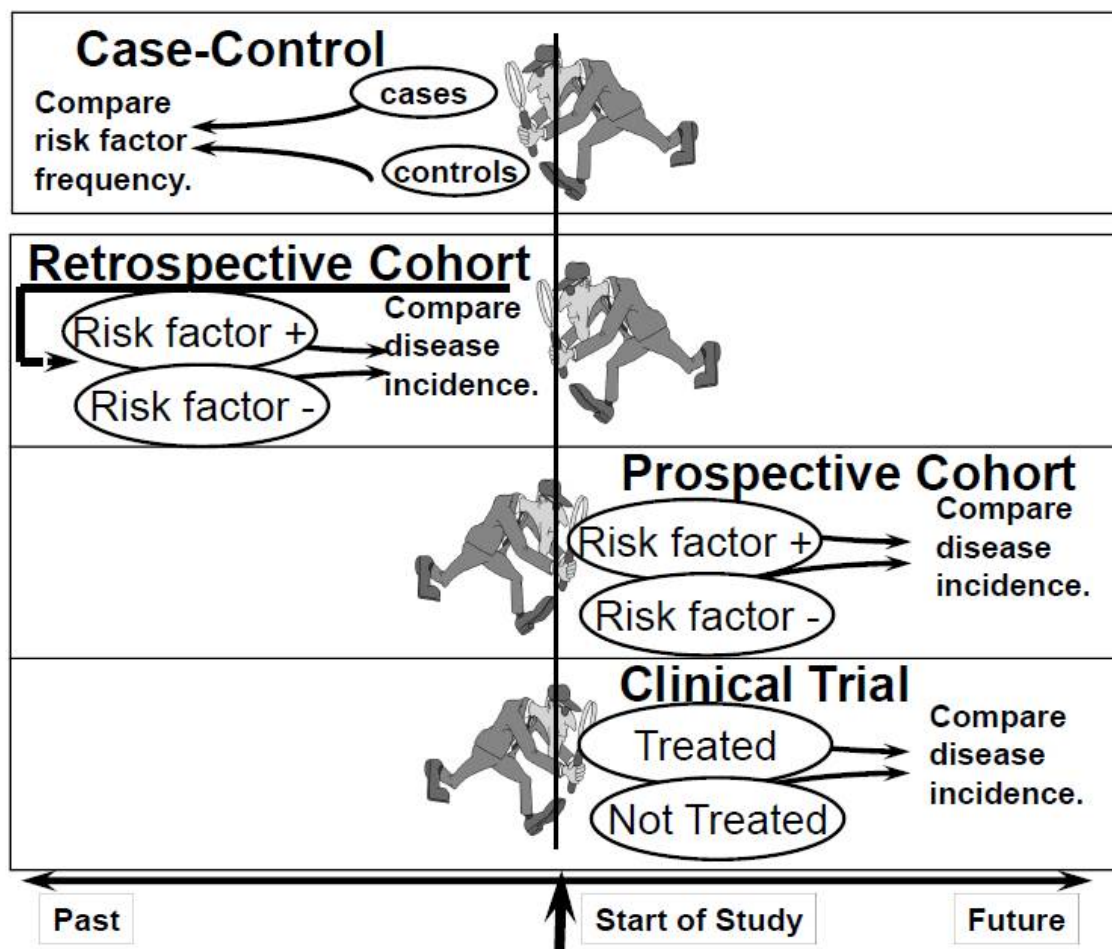
In prospective cohort studies conception, design, & enrollment occur before anyone develops the outcome.



Finally, did the investigators assign subjects to a treatment or program and then follow them prospectively? If so, it was likely a clinical trial.



This last cartoon summarizes the different types of analytical studies.



Preview of the Post-Class Quiz

(Study Designs)

