Clinical Chemistry Trainee Council Webcasts

www.traineecouncil.org

TITLE: PRINCIPLES OF STUDY DESIGN AND ANALYSIS: CASE-CONTROL STUDIES

PRESENTER: Julie E. Buring, ScD

Slide 1: In this next section of talk I am going to be talking about the principles of study design and analysis for case-control studies, cohort studies, and randomized clinical trials.

Slide 2: It's very important for us to know that the different kinds of study designs are relevant to us in terms of choosing what to do ourselves, but also interpreting everybody else’s findings. It's important for us to remember that any research question can be addressed using a number of different research strategies. And we are going to make a choice between them based on our scientific question and our resources; how much time we have, how much money we have; what we can literally do in terms of doing the study.

But each study design has its strengths and limitations that must be clearly understood. So we may choose one study design because it's quicker and it's cheaper, but it will usually have some more problems associated with it that we have to keep in mind.

It's also important to always remember that no one epidemiologic study can answer a research question definitively. So if we are sitting here today and we are trying to figure out our current state of knowledge about a question that’s important to you, we need to consider not just one study, but the totality of evidence that is available on that particular question.

Slide 3: And our totality of evidence includes both basic research and epidemiologic studies. The basic research is very important for us to understand how an exposure could impact an outcome, what is the possible mechanism by which it would be working. And so we work with our laboratory and our animal colleagues to be able to get that evidence.

But the unique feature of epidemiologic studies is that they are done in humans, so they are complementary to the basic research and really provides information we would not get from basic research.

And one of the first classifications for epidemiologic studies is that they are either descriptive or analytic. Descriptive studies are the ones that basically tell us who is getting this outcome? What’s the definition of what they are getting? Does it matter if they are young or old, if they are smokers or nonsmokers? If we are talking about 1950 versus 2013, if we are talking about a developed country or a developing country, just describe the outcome that we are interested in looking at.

All of those are going to raise hypothesis for us. The correlation or ecological studies, the case reports and the case series, the cross-sectional studies or surveys are going to raise hypotheses which we will then test in the analytic studies, the studies that are going to allow us to figure out why this exposure is related to this outcome.

And the next differentiation is between whether those analytic studies are observational or intervention. Observational studies mean we simply watch what people do. We record whether they have a high fat diet or not, take a particular drug or not, what their genetic profile is or not, and we look at what outcome they get. But we don't intervene in any way, we just record.

In an intervention study, randomized clinical trial, that will be the closest to a basic research study that either we can do in epidemiology, because we are actually going to allocate people to either get the intervention or not.

Slide 4: So again, the main difference is between descriptive and analytic. We are now going to be talking about the three kinds of analytic studies; the search for factors associated with or predictive of the outcome; there will be observational studies, either case-control or cohort; and then we will talk about intervention studies or the randomized clinical trial.

Slide 5: And the key distinction between the observational studies and the intervention studies is the role the investigator plays. So in an observational study the exposures are self-selected and we just record what happens, and in an intervention study the exposures are actually allocated by the investigators.

And first then we are going to look at the two types of observational studies; case-control and cohort. And it's going to turn out that in both cases we need to know about the exposure status and the non-exposure status, and whether they get the outcome or the disease or not. The only difference between a case-control and a cohort study is how we bring people into the study.

And in a case-control study our initial classification of our two groups is on the basis of the outcome, whether they got the disease or not. And in a cohort study the initial selection or classification is going to be on the basis of the exposure status, whether they are exposed or not, which is coming in at different ends of the exposure disease relationship.

Slide 6: So a case-control study is a type of observational study in which the subjects are selected on the basis of whether they do or do not have a particular disease or outcome under study. If they do have the outcome they are called cases; if they do not have the outcome they are called control. And then we take those two groups and we compare them with respect to the proportion having a history of an exposure or a characteristic of interest.

Slide 7: And looking at it as a diagram, as we said before, for every single study, case-control cohort or trials, we have to know the exposure status and we have to know the disease status by the time the study is over.

In a case-control study we are going to start at the end, classify people on the basis of whether they are diseased or not, and then go back and assess their exposure status between the two groups.

Slide 8: Let me just give you a few examples of a case-control study. Let's say our question is, is there an association between a baby's sleeping position on the stomach or on the back and risk of sudden infant death syndrome?

So the cases have to be deaths from SIDS, so all deaths from SIDS in infants between the ages of 7 and 364 days, so within the first year; in two regions of the United Kingdom, between February 1993 and January 1994; reported through a communication network of professional organizations who report all sudden unexpected deaths. The controls need to be similar children who did not die of sudden infant death syndrome.

So what they did is go back to the hospital, the same hospital in which the index baby had been born and identify the next two younger and the next two older babies who are born in the same hospital, within two weeks of the age of the index baby but they did not die of SIDS. So cases were death from SIDS and controls were babies born at the same time, in the same hospital, who did not die of SIDS, case-control.

Slide 9: Another example, the question is, do young women, so women less than 40, who regularly participate in physical activity, during their reproductive years have a reduced risk of subsequent breast cancer? So the cases have to be young women with breast cancer.

So they got 545 women aged 40 or younger, diagnosis, newly diagnosed with in situ or invasive breast cancer, between July '83 to January '89, identified by the Population Based Cancer Registry for Los Angeles County.

Controls need to be women of the same age who didn't get breast cancer, but we would like them to be as alike as possible with respect to everything else besides physical activity. So what they did was go to the house where the case lived at the time of diagnosis and basically walked the neighborhood to find a neighborhood control, individually matched to each case by date of birth within three years, race, and parity based on a predefined walk pattern for the neighborhood where the case lived at the time of diagnosis.

Slide 10: So what are the strengths of a case-control study? Well, this approach actually began relatively recently compared to a cohort study or a trial as the diseases of interest to us shifted from acute to chronic public health problems. So as chronic diseases have become more prevalent as causes of death, not only in developed countries, but in developing countries, then we had to figure out a way to deal with that.

And the problem that we had to deal with is that there was a long latent period between when the person was exposed and when they got the outcome. We all know that it's not what we eat today that's going to increase our risk of developing cancer tomorrow, that what our pattern of diet is will affect our risk of developing cancer 5, 10, or 15 years from now.

So there are long latent periods in chronic diseases between adequate exposure and the development of the outcome. And by picking in a case-control study people who already have the disease, then in fact it becomes a very efficient design with respect to time and money, because you don't have to start with people eating a particular diet and then follow them for long periods of time until they develop the outcome; you are starting with the outcome, they already have it, so that’s going to save time and money.

Slide 11: And since we are going to select on the basis of the disease status, we can identify adequate numbers of diseased and non-diseased people. So it's going to be the ideal design when the outcome is rare. We are not going to need to follow large numbers of people in order to get sufficient numbers who subsequently develop a particular outcome.

Slide 12: It also allows for the evaluation of multiple exposures or risk factors for a single disease. So I myself for case-control studies do it one of two times.

Number one, because my disease is so rare that I don't want to have a massively big sample size to get enough people with mesothelioma or a particular other outcome that’s of interest to me.

And the other time I use it is when I just don't know much about this disease. I really don't have a good body of information about its risk factors. So when I want to understand a lot about the risk factors for the disease, then I pick on the disease; I take people with it and people without it, and I can ask them many, many questions to try to elucidate the risk factors for that particular outcome.

So when I am interested in the disease and everything that might be related to it, then I pick the disease, which would be a case-control study.

And it can be used to test hypotheses, but it also can be used in the example I just gave, where we really don't know much about this outcome, to explore a wide range of exposures. What’s sometimes called a fishing expedition, but that doesn’t make it bad, it's not a pejorative term. It just means that I am going to ask loads of questions and see what sort of comes out, so that I advance my knowledge and then can test that in another study. So it's particularly useful in the early stages of knowledge about a disease or an outcome.

Slide 13: Now, the limitations are though that at the time you start to study and you ask people about what their exposure has been in the past, they already know that they have the disease. So if the exposure status, when assessing the exposure status you already know the disease status, you already know it, the participant already knows it, then there's a real potential for bias.

In that, people being brought into the study are selected differently based on whether they were exposed or not. Or they answer questions differentially, because they know that they are diseased or not.

So there's the potential for selection bias, which is the selection of cases or controls on the basis of whether they were exposed or not, whether they were smokers or not, as well as observation bias, where the people report their smoking, either more precisely, or they deny that they do it, one way or the other, but they do report it differently than people who know that they do not have lung cancer.

Slide 14: And the other thing that’s a little bit tough about a case-control study is that the temporal sequence of the exposure and the outcome may be difficult to establish. The logic is fine. Some people say it seems a little artificial, you are going backward, so you have the outcome and now you are trying to figure out what the exposure must have been, but that’s not an unusual logic in everyday life to look at something that happens and go back and try to figure out what must have caused it.

But what's most of a concern is, I don't actually know when I am supposed to be getting your exposure history. Let’s say you now have colon cancer, am I supposed to be asking about what you ate last month, last year, five years ago, when you were a teenager? I don't know when the biologic period is that exposure is supposed to make a difference.

And the farther back I go, let's say, I do want your dietary history when you were a teenager, are you going to be able to give that to me accurately? So I worry about the ability to both get accurate past information and to know what time period I am supposed to be getting it for.

And finally, because there is no rate of development of disease in the case-control study, you either have it, you are a case, or you don't have it, you are a control, nobody is developing it during the study, then you actually can't calculate an incidence rate. You can't do a rate of development if there is no development.

So you cannot calculate incidence rates in a case-control study, but you can estimate the relative risk, the attributable risk by an alternate measure of odds ratio, which is a valid measure in a case-control study. So we get to the measurement we want but we will never be able to tell people the rate of development of the disease in the exposed and the non-exposed group. If you need that, you have to be doing a cohort study or a trial.

Slide 15: So the bottom line, the strengths and limitations, or especially the limitations, are not reasons to not do a case-control study, they are just reasons to design it very carefully.

And the bottom line is case-control studies work. It's actually the most common analytic epidemiologic study design in the medical literature. It's often the first study design that’s used in analytic epidemiology. And it’s absolutely the optimal approach when we have a new condition that we are interested in, or when we need to do the study a little bit cheaper and a little bit faster, then a cohort study or a trial could be done.

Slide 16: Well, now we need to go through where do you get your cases, where do you get your controls, where do you get your exposure, where do you get your outcome information? Selection of cases is not usually the difficult part; all we need to do is figure out what outcome you are interested in and define it in a way that can be reproducible by other people.

And we really do want as homogeneous a disease entity as possible. You don’t want to combine all congenital malformations, because no teratogen causes all congenital malformations.

And you are tempted to do it because the sample size will be higher, but you don't want to do it because if it's only related to one for example of all congenital malformations, one type, you are going to harm your ability to find that relationship with one particular congenital malformation type, because you have merged everything together and it's going to mask the relationship that you want to see.

The same for combining uterine cancer and cervical cancer, again, the sample size will be bigger, but uterine cancer and cervical cancer have very, very different etiologies and you will harm yourself in terms of finding the relationship that you want to find.

So you will need strict diagnostic criteria for your outcome, which are reproducible. So you will need to cite that you used the WHO definition for myocardial infarction or the TOAST criteria for stroke.

And then also what you might want to do is stratify your outcome, your cases and controls, by definite, probable, and possible, so that you really know, did I have all the criteria or just some of the criteria for the disease or the outcome under study.

Slide 17: You can get your cases from many different places. You can get them from the hospital, a hospital-based case-control study. People who were treated at medical facilities; a hospital, health maintenance organization, a private practice during a specified period of time, and that’s relatively easy to do and relatively inexpensive.

Slide 18: You can also do it from the population. You basically enumerate all cases of this disease in a general population or a random, and then take either all of them or a random sample of them. But that’s a little bit more logistically difficult, because it involves locating and obtaining data from all affected individuals or from a sample of them from a defined population.

It's very good in that it avoids bias from selected factors that led someone to use a particular hospital. And if it's complete ascertainment, then you actually can calculate rates, which you can't usually do in a case-control study, because now you have a description of the entire picture in the community. But really big logistic and cost consideration, so it is not frequently done.

Slide 19: You can get your cases from disease registries or from special surveys, any sort of other source that would allow you to delineate cases of your disease.

Couple of things we need to think about; are you going to do incident cases, which are new cases of the diagnosis of the disease, or prevalent cases, people who are existing with this disease?

And you will be tempted to say let me just use prevalent cases, let me just go out there and take everybody in my community who has this outcome or disease. But just remember that incident cases are allowing you to only look at risk factors for the development of the disease; prevalent cases, you are looking at risk factors for not only the development of the disease, but for the prognosis, for the survival from the disease once you get it.

It will be very hard to figure out whether you are getting risk factors for prognosis or survival or risk factors for getting it in the first place. So if you really want etiology or development, if you can, it would be better to use incident cases, not prevalent cases.

And finally, someone might say, well, don't I want representative cases? I want a random sample of all cases in the population rather than just using the five biggest hospitals in my area. Absolutely will increase generalizability, much more representative of all cases with the disease, but it will really increase your logistic difficulties.

So if it is more valid, you will have a better chance of getting all the cases from five hospitals than from everybody in your city, your state, or your country, then go ahead and go for validity, do it well in your five biggest hospitals, and then discuss whether it can be generalized to your entire community.

Validity comes first. You cannot generalize an invalid result. So do this study well, do it so there's a valid or true relationship between the exposure and the outcome in whoever you pick to do the study among, and then sit and judge whether it can be generalized or not to a broader population.

Slide 20: Where do you get your controls? Well, that actually is the most difficult and critical issue for the design and the validity of a case-control study. It is necessary though to have good controls to allow the evaluation of whether what is observed in your cases is different than what would have been expected to be seen based on comparable people who just don't have the disease.

There is no control group that’s going to be optimal for all situations, so you really have to think it through, beginning with where did you get your cases. So depending on the source from which the cases were chosen, the controls must be selected to represent not the entire non-diseased population, but the population of individuals who would have been identified and included as cases if they had developed the disease.

So it is a comparable population to the population which gave rise to the cases. So again, like the cases, they may not be representative of the general population, but the crucial requirement is that they be comparable to the source population that gave rise to the cases, and any exclusions or restrictions that you made in the identification of the cases would have to also be made to the controls. So the cases and controls have to have the same inclusion and exclusion criteria.

Slide 21: So if the cases came from the hospital, then you should consider the controls coming from the hospital. The advantages are that it’s very convenient. They are as easily identified as the cases are. They are readily available in sufficient numbers. It's relatively inexpensive to do. And it takes minimal effort because you are already in the hospital anyhow.

The cases and controls are likely to be similar in their accuracy of recall, because they are sort of both sick. So they are both thinking about what led them to be in the hospital, so it's going to minimize recall bias.

And they will have the same selection factors of coming to the same hospital. And there’s generally a high level of cooperation, because compared to healthy individuals, they really are sitting thinking about their health, they are at that moment sort of a captive audience, and they are going to be very willing to talk to you, so it's going to minimize non-response bias.

Slide 22: The only disadvantages are actually big disadvantages or potentially big disadvantages, we have to figure out whether they are or not. These controls are now ill, so they differ from healthy individuals in ways that’s associated with illness or hospitalization in general.

They actually may not represent the exposure distribution in the population from which the case is derived. And worse than that, it may be that that controls are hospitalized for diseases that are also related to the risk factors under study.

We all would know that if you are looking at smoking and lung cancer, you would never take as your control group bronchitis or chronic obstructive pulmonary disease, because those are also related to smoking, those are sort of obvious.

But the more that you are doing a lifestyle factor, like an alcohol and coronary heart disease, can you use accidents as your comparison group? You might be concerned about doing that because car or sports accidents might in fact be related to alcohol consumption. So how can you really be confident that the diagnostic groups that you have chosen as your control group are truly unrelated to the factors under study?

And there are also issues about selection factors leading to hospitalization. A hospital may for some diseases be specialized. So it's known that you would come to this hospital for cancer, or for gunshot wounds, a particular outcome that you are interested in. There is a referral pattern to go there, it's a tertiary hospital. People know to go to that hospital if this is the condition that they have.

But for the comparison group, the comparison group might in fact be just coming from the local community, because this hospital treats that kind of a problem just as well as any other hospital does.

So the cases and controls may be coming to that particular hospital for a particular disease for different reasons and that could cause a selection bias.

Slide 23: So when the cases come from the hospital, that the hospital controls are of concern to you, because they may not be scientifically adequate, or when the cases actually are coming from the general population, then you might want to consider taking the controls from the general population.

The advantages of that is that generally that ensures comparability; everybody is coming from the same source population. But the disadvantages are just logistics. It is often difficult to enumerate all members of the population and we have to do that to be able to select individuals.

Some places like Massachusetts do have town lists, or you can get all the names of people living in a community, but most places don't have that. So how are you going to enumerate the members of your population or identify in some reproducible way where your controls are going to come from, from the population?

It's also more difficult to gain cooperation. These are healthy people. They may be working multiple jobs, they don't have time, they are not as motivated to do this, and non-response is always greater than for hospitalized cases, which is a threat to the validity.

So all in all, it's going to be more expensive, it's going to be more time-consuming, and it may not have as good quality of information, because these healthy people who are living their lives in the community may not recall the exposures as well or may not be motivated to recall the exposures with the same degree of accuracy as the cases do.

Slide 24: Now, there are some special methods to get the controls. There is random digit dialing. We have all received phone calls where we have been selected and we are talking to people after our telephone number has been at random selected. And you can match on area code, you can match on dialing prefix, so you can actually get somebody from the same restricted neighborhood as you are in. There are a lot of problems with this one.

We are by definition excluding people that don't have a phone. And that could cause a problem of socioeconomic status as a variable that's of interest to you.

We are going to exclude people without landlines, which may mean we exclude the young, which are less likely now to have landlines than older people do.

We have to deal with the probability of being home, so that means we are going to have to call during the day and the evening, as well as weekends to try to reach people, because the probability of home may be related to some exposures; anything that might have to do with income or socioeconomic status or exercise patterns, anything.

We now have answering machines, and people are screening their calls. So you will have less people who are willing to talk to you, because if they look when the phone rings, they look at their machine, they don't recognize the phone number, they may not answer. So our response rates are going to be lower now than ever, and increasingly low.

And finally, the increasing cellphone use is a problem, because there are no lists on cellphone numbers.

So for all of those reasons a random digit dialing is still considered, but it is certainly less popular than it was in the past.

Slide 25: And finally, something called snowballing, where you take your case and you ask your case to help you find a control. Identify a friend, a relative, a spouse, a sibling, a neighbor, who would be willing to possibly be in the study. And actually those people are much more likely to be cooperative than the general population.

And if you needed to control for certain factors, so if you need to control for the current environment, you can get a neighbor; early environment, you can get a sib; household environment, you can get a spouse; genetics, you can get a relative. So it does give a degree of control of important confounding factors that are related to ethnic background, socioeconomic status, current, or early environment.

Slide 26: But the problem is if the study factor is one for which family members and friends are likely to be similar to the cases, then in fact it's going to make the cases and controls artificially alike with respect to the exposure and underestimate the true effect of the exposure.

People tend to eat alike, smoke alike, exercise alike. If one person is big on owning pets, other family members may be big on owning pets; social interaction.

So if the case identifies the control, they actually may identify someone specifically, because they might say, well, I have the disease, but my friend has the same risk factors I do and they don’t, so pick them.

Or my friend has absolutely no risk factors at all so let's see what it would have been if we did that, or has lots of risk factors. So they may elect to choose the control based on their exposure habits. That’s going to lead to overmatching and absolutely require a matched analysis to take that into account.

Slide 27: So bottom line, the cases must be selected independently of the exposures and the controls are a direct random sample ideally of the reference population from which the cases originated, but just like the cases, they must be sampled independently of the exposure.

Slide 28: So how many control groups do you get, one, two, more? Ideally you just want a single control group that is most comparable to the cases. But if there is situation where it is believed that one selected control group really could have a specific deficiency, and if you took another control group, included a second control group, it might be better, then you can use multiple control groups.

So again, if you are doing a lifestyle factor and you really are concerned that there is nobody in the hospital that is not in some way related to this lifestyle factor, then do your hospital control group, but also get a population control group.

So multiple control groups when there is a specific deficiency that could be overcome by the inclusion of another control group.

Slide 29: And how many controls per case? What's your control to case ratio, one to one, two to one, a 100 controls per case, what do you pick? Well, when the number of available cases and controls is large and the cost of getting information from a case and a control is comparable, then the optimal control to case ratio is one to one.

The exceptions are the following. If your sample size of cases is very limited, for example, you have a very rare outcome that even if you work for two years to get the cases and multiple sites of hospitals, you can only still get a limited number of them.

Or when the cost of getting the information is greater for the cases than the controls, then the control to case ratio should be altered and you should increase the number of control per case. And as that control to case ratio will increase, your power will increase. But that increased power really levels off at four to one. So one to one is ideal, but you can go up to four to one and you will increase your power of your study.

The only exception to the four to one rule is if you are working in a situation where the data are what are called free. So you are identifying your cases or your controls by just going to a health maintenance organization dataset, or a preexisting dataset of some kind. Basically you are just putting a request into the computer, give me all people who are matched on these characteristics, who were seen in the organization the same time the case was, those data are sort of free, it doesn’t cost you much to get those, then you can actually increase your control to case ratio beyond four to one, if the data are free.

And sometimes you will see that in a study where they say, we got ten controls per case. Every time I see that I know that the control data were just free and they just picked more people.

So again, control to case ratio, maximal one-to-one; control to case ratio can go up to four to one and you will see an increase in power. After that there will be no increase in power unless the data are free, in which case the control to case ratio can go up as high as you want.

Slide 30: Now, how do you get the disease and the exposure? Any potential source of information has to be assessed in terms of its ability to provide accurate and comparable information for the study groups.

The disease sources can be from death records, case registries, office records, hospital admission or discharges, pathology logs. The exposure sources can be the study subjects themselves, by interview or a mail questionnaire, from a surrogate or a proxy, like a mother for a child, a spouse for a dead patient, from information that is recorded in medical records.

Slide 31: So the important thing is that the procedures that are used to obtain the information must be as similar as possible for the cases and the controls. If you are interviewing cases and controls, make the place that you go to interview them and the circumstances the same.

Blind your abstractors so they don't know whether the person is diseased or not diseased.

The interviewers and the patients should be unaware of the specific hypothesis being studied, so you can blind them to the purpose of the study.

And try to train your interviewers the exact same way, with role modeling, so they don't use more probing questions for cases, for example, than the controls.

And if you can obtain records that are completed before the occurrence of the outcome, that's going to be especially valuable in a case-control study.

So use birth certificates for birth weight and gestational age. Use prenatal x-rays coming from obstetrics records for congenital malformation. Try to use information that could not be biased by the fact that we all know now that a child has been born with a congenital malformation or a child has been born with low birth weight for example.

Slide 32: And then though the hard part is actually to decide the basis on which a given individual is going to be considered exposed. What part of that person's exposure history is relevant? And to do that you need to understand the mechanisms of the disease process, as well as the likely latent period.

So if you are doing smoking and lung cancer, for lung cancer we know that current smoking is not the main variable to be looking at, it's pack years, it's total duration of smoking and how much you have smoked at each one of those periods of time. But if you are doing smoking and myocardial infarction, then in fact current smoking is the most relevant.

So depending on the outcome you have to understand the disease process enough to know what exposure history you need to get and at what time period.

If the time period is too wide, you use ever use when current use is what is relevant, then you are just going to harm your ability to find the association you are looking for because of random misclassification which will bias towards the null.

What you might need to do if you don't know everything biologically yet about this exposure disease relationship is evaluate data from different time windows of exposure. So you can begin to look at the time periods and see what seems to be the one that's most relevant with the most increased risk.

Slide 33: The analysis; we set up our data in a 2x2 or a rxc table, where r is the number of rows, c is the number of columns. As we said before, we can't directly calculate the incident measures of disease frequency unless we are doing a population-based case-control study, because there is no development of the disease in a standard case-control study.

We then estimate the measures of association by the odds ratio, set up our 2x2 table, do the cross-product ratio. Odds ratio is ad over bc, (ad)/(bc). And we can do the risk difference as a percentage. The attributable risk among the exposed group percentage is the odds ratio minus one divided by the odds ratio, and we can do it among the exposed, attributable risk with a small e next to it, so among the exposed. Or we can do a PAR, which is an attributable risk among the entire population percent; either one.

Slide 34: A couple of special things then about the interpretation or the analysis of a case-control study, and that's back to the role of chance, because many times we are going to do a fishing expedition, where we actually derive hypotheses from the data in our study.

So most case-control studies will test a small number of specific hypotheses, but while you are doing that you also collect data on a multitude of other risk factors and you conduct many comparisons besides the ones you are doing as a hypotheses testing situation.

So it's very important to distinguish between test of hypotheses, where those hypotheses were specified in advance, a priori hypotheses, and fishing expeditions, in which you just analyze your data and you see what pops out, what associations emerge when the data are analyzed, called data derived hypothesis or a posteriori hypothesis; a priori or a posteriori hypothesis, they are different. They are different in the way that we must consider the interpretation of the findings.

Slide 35: And data derived hypotheses have to be interpreted with caution. Remember the meaning of the p-value. If you do 20 comparisons and you just see what comes out as statistically significant, remember 1 out of 20 of them will be statistically significant by chance alone, even if the null hypothesis is true.

Even if statistically significant, if you are in a hypothesis formulating situation, you have to interpret very cautiously, because any finding that was not specified in advance as hypothesis testing could be due to chance.

So these formulated hypotheses can then be tested in other studies specifically designed to do so, but you must say we are raising hypotheses in the study that could be due to chance and need to be looked at in another study.

There is also a special kind of case-control study that's very worth knowing about, a nested case-control study within a cohort.

Slide 36: So in that design the population is actually defined first, a cohort study; the Framingham Heart Study, the Mayo Clinic Study, a study of a particular exposure, a study in a particular geographical area, whatever your cohort is. And as you are following that cohort over time, cases are going to develop and they are going to be identified, and you could use those then as your cases in a nested case-control study, where your controls are a random sample of non-cases that are going to be selected as your control subjects.

Slide 37: So let me give you an example of that. Between 1987 and 1992 there were 10,786 women between the ages of 35 and 69 that were recruited into a prospective study on breast cancer in Italy. At the recruitment they collected urine from everybody and stored it. After an average of 5.5 years of follow up, there were a 144 breast cancer cases that were identified.

And they decided that what they wanted to look at is whether estrogen metabolism affected breast cancer risk. So they took 144 breast cancer patients, they selected about four times, 576 controls from the cohorts who had not developed breast cancer. They matched them on age, they matched them on other variables, and they conducted the study on just these individuals.

On the 144 plus the 576, they actually did the assays on estrogen metabolism just among that group, nested case-control study.

Slide 38: And why did they do it that way? Completely because of efficiency in terms of time and money. Think about it. If an assay costs $25 and you needed to do the entire cohort, then it would cost about $269,650, but if I just did the assay on the 144 cases, and between one and four controls, however many you decided to use, it would cost between $7,200 and $18,000. So much more cost and time efficient way to do it.

It also is going to turn out that it's going to reduce the bias that's inherent in a case-control study, because every piece of information, the baseline questionnaire and the blood sample, was actually obtained prior to the development of the breast cancer, because it's a cohort study.

So at the time that everybody came into the cohort, nobody had developed breast cancer. So nobody knew about whether they would get breast cancer or not when they filled out the questionnaire. So there was a reduction in bias in giving us the exposure information, because at that time the outcome had not yet occurred.

So reduces bias; it's much more efficient in terms of time and money. So this is very commonly done, especially in these tight financial times. It's an ancillary case-control study nested into or grafted onto an existing parent cohort study, where that cohort study becomes a resource for other people to use for nested case-control studies.

Slide 39: So summary for case-control studies, the strengths are they are ideal for rare outcomes. They are not very good for rare exposures, because you would need a tremendous number of cases and controls to have enough who have had a rare exposure for you to be able to make that comparison. So if your exposure is rare, you turn to a cohort study; if your outcome is rare, you turn to a case-control study.

If you want to evaluate multiple exposures for a single outcome, so multiple risk factors for a single outcome, you select on your outcome, so you do a case-control study. And if you wanted to do multiple outcomes for a single exposure, then you are going to select on the exposure and you are going to do a cohort study.

Case-control studies are very efficient in terms of time and money. It cannot calculate incidence rates, but you can estimate the relative measures.

You have the potential for selection and observation bias, but that will be minimized if you do a nested case-control study. And you will always have a difficulty in knowing the appropriate time window for assessing exposures and getting accurate past exposure information. But get it all to the best of your ability and then you can look at the relationship for different time periods and for different qualities of information and learn from that in terms of the exposure disease relationship.

Slide 40: Thank you!