Clinical Chemistry Trainee Council

Webcasts

www.traineecouncil.org

TITLE: PRINCIPLES OF STUDY DESIGN AND ANALYSIS: COHORT STUDIES

PRESENTER: Julie E. Buring, ScD

Slide 1: In these series of lectures we are talking about the principles of study design and analysis and today, we are going to be talking about cohort studies.

Slide 2: If you remember when we were talking in our first lecture on case control studies, when we're trying assess where we are in our state of knowledge on the question in a particular point in time, what we do is not look at one particular specific study, but look at the totality of evidence, all data that are available on this question.

Those data and evidence come from basic research studies, laboratory, animal studies, which gives us information about a possible mechanism by which the intervention or the exposure could be related to the outcome.

But the unique evidence that comes in humans comes from epidemiologic studies and the first differentiation on the categorization of epidemiologic studies is whether they are descriptive or analytic.

Descriptive studies are the co-relational studies or ecologic studies, case supports and case series and cross sectional surveys and what they do is raise hypothesis by describing who is getting the outcome, what outcome are they getting at, when are they getting the outcome where are they getting the outcome.

We take those hypotheses that are raised by the descriptive studies and we test them in the analytic studies. We try to evaluate why that exposure affects that outcome.

And the next classification of studies is whether they are observational or intervention.

Slide 3: Observational studies means that we simply record what people chose to do. We record what their exposure is, we record what their outcome is, but we, the investigator, played no role in what exposure is chosen by the participants or what exposure the participants' experience.

So is an observational study the exposures are self selected. Being an intervention study, the exposures are actually allocated by the investigators. So the difference between an observational study and then an intervention study or randomized clinical trial is the role the investigator plays in the study.

Among the observational studies now, there are two types, case control studies, which we discussed last time, and the cohort studies which we will discuss today, and the studies are the same in terms of evaluating the association between the exposure and the outcome, they get both of them get information on whether someone is exposed or not exposed, diseased or not diseased. But the difference between them is the basis on which the participants are selected into the study at the beginning.

Where case control studies initially select on the basis of the outcome whether the person has the outcome cases, or does not, the control, and goes back and looks at their exposure history and in a cohort study the initial selection is upon the basis of the exposure status, they are classified as exposed or non-exposed and then followed forward for the development of the outcome.

Slide 4: So remembering again, case control studies, start at the end of the picture. The end of the story, they start with a group that's diseased non-diseased and then go back and assess their exposure habits, their exposure patterns.

Slide 5: In a cohort study the investigator starts by classifying people as exposed and non-exposed and then follows and forwards to the development to the outcome.

Slide 6: A couple of examples of cohort study. Let's say that the question was, what are the adverse effects associated with being exposed to dioxins or Agent Orange in the Vietnam War?

The exposed individuals would be 1264 Air Force personnel who were involved in the defoliant spraying in Vietnam between 1962 and 1971 what's called the Ranch Hand Project.

The non-exposed individuals are 1264 Air Force personnel who flew a variety of cargo missions in South East Asia during the same time period but were not involved in the defoliant spraying.

So the same Air Force personnel that classified as exposed were non-exposed depending on whether they were in the Ranch Hand Project or not, and the outcomes being looked at, would be dermatologic conditions from the actual defoliant on the skin, adverse pregnancy outcomes in their spouses or partners and then cancer in the actual personnel who were there.

Slide 7: Another example, are there adverse effects associated with the use of oral contraceptive? The cohort that was chosen was the nurses' health study, which is the 122000 female nurses in the United States who have been followed since 1976.

They then classified them as exposed and non-exposed where exposed are those who reported the use of oral contraceptive at baseline and non-exposed are those who reported never using of oral contraceptive at baseline. Those two groups exposed and non-exposed are then followed forward and compared for the development of breast cancer, myocardial infarction and blood clots.

Slide 8: So a cohort study is defined as a group of subjects free of the disease of interest at baseline who are defined or classified by the presence or absence of the exposure of interest, they are then followed over time for the occurrence of the disease or the outcome of interest.

And then another example of that could be something like if we did genetics, women with and without the BRCA1 gene mutation, who are free of cancer and then followed for the development of breast cancer. Synonyms for cohort study that you might hear are longitudinal study, prospective study, follow-up study.

Slide 9: Now the cohorts can be of different forms. One is what's called an open cohort, where members are defined by a characteristic that is changeable. So let's say for example, that we want our cohort to be defined by location, people who are living in the Boston area or it could be by experienced those are in the Vietnam war, those who are in the Iraq war. Another group could be students in college, another group could be an occupational exposure, such as employees and the faculty in a factory.

But the important thing about it is that new subjects can be added or eliminated during the follow-up, it's going to be a dynamic cohort. So as people start college or leave college, move to Boston and leave Boston, start working in a factory and leave a factory, they would come in and come out of a cohort. And the exposure status can actually change over time as people come in and out of what we are going to be calling the exposed cohort.

Slide 10: A fixed cohort is more where members are defined by an irrevocable event, so for example, if we have a disaster, so as exposure to the Japanese earthquake, exposure to Katrina in the New Orleans. It also, could be inhabitants of a specified location at a specified time such as a Framingham heart study.

All of these involve having a common starting point and a defined period of follow-up whether it'd be one year, ten years or until all the cohort dies and the exposure is going to be defined at the start of the follow and no new enrollees will be allowed into the cohort during the follow-up period.

Slide 11: So how do we select this kind of a cohort? Do we do a general cohort or do we do a special exposure cohort? Well, it really depends on the research question that you are asking.

Slide 12: A general cohort means that it is selected on a characteristic that there is nothing special about the exposure related to it. So in other words when we select the Framingham community for the Framingham heart study or college attendants or a professional group of some kind, there is not anything particularly special about that exposure, but what really is, is that we have an enhanced ability to follow-up the participants.

So whether it'd be the Framingham heart study, the nurses heart study, the Harvard alumni health study again, the study was not begun at Harvard because it was thought that Harvard would have a different exposure experience in another college or university, it was because the alumni association kept list of the Harvard alumni and therefore it would enhance our ability to follow them up over time.

So what you can then do is take not only the exposed and the non-exposed but do it internally, so you've got the group, you ask them questions and you yourself can classify them as exposed or non-exposed from the information that they provided.

And this is going to be a very appropriate cohort as long as the prevalence of the exposure is not extremely rare or extremely common.

Slide 13: You could also though taken exposed cohort and in that case the cohort is actually chosen because of a higher prevalence of the exposure and it is especially done when the exposure is rare in the general population, so it might mean workers who are exposed to Man-made Vitreous Fibers or women who have breast implants or groups that have a special lifestyle patterns, such as the Seventh-day Adventists who actually have a higher prevalence than the general population of having lower risk factors for disease.

Slide 14: Regardless of what you do, a non-exposed and the exposed subjects should be as similar as possible with respect to all factors other than the factor under investigation. So basically under the nuller assumption, the null hypothesis that there is no association between exposure and the outcome, then the disease rates of the two population should be essentially the same if that is correct. And what we need to do is collect data when any potential baseline differences that could affect the outcome under study.

Slide 15: So where do we get the non-exposed subjects. Well we can actually take an internal subgroup of the general cohort, just like we said, nurses' health study, divided them into those who at the baseline said that they used oral contraceptive, and those that said they did not use oral contraceptives. So that's an internal subgroup of non-exposed from the general cohort.

It's usually the most comparable group to the exposed population in that cohort, so you stay within the cohort and you internally divide them into exposed and non-exposed. So you could take, for example again in the nurses' health study not just oral contraceptives but you could classify them as baseline as reporting high versus low intake of fat or with or without a BRCA 1 gene.

On the other hand you can get your non-exposed from the general population and the reason that we would do that is the following. Let's say we were interested in looking at the mortality in a specific occupational group, like the rubber workers and who do we compare them to because I don't -- we don't think that we can use an internal cohort.

It isn't that you can take all rubber workers and classify them as exposed or non-exposed to the chemicals of concern, because either they are working with the chemicals, where there is a possibility that they might have the chemicals for example on their clothing and even those who are working in that industry in the office might have exposure to those chemicals.

So then you can say, all right let's not use rubber workers as both the exposed and non-exposed, let's compare to the general US population, but the problem with that is that those who are working in an occupation actually are healthier than the general US population which is formed of those who are working and non-working, those who are retired, those who are too ill to work, different populations are in the general US population than just those who are working in an occupation and that's called the healthy worker effect.

So in that case it would be better to actually get another comparison cohort. So you would look at the mortality in a specific occupational group, compared to the mortality in an another occupational group who was not exposed to that chemical that you are concerned about in the rubber workers for example and that will avoid the healthy worker effect.

Slide 16: Where do we get the information on the exposure? Well we can get them from records, collected independently in the study, so occupational records, medical records, pharmaceutical records, school records. We can also get information from the research staff.

We can do medical exams on the participants, make biological measurements, electronic devices that are worn by the subjects, environmental or workplace measurements and finally we can get information reported by the study subject themselves, such as questionnaires or interviews.

Slide 17: But the aim is to collect data uniformly from the exposed and non-exposed subjects not just on the exposure, but now on the outcome and some options for sources of the outcome information can be reported by the subjects sometimes with validation where when they self report, you get the medical record and you confirm the diagnosis.

You can get it from medical records, from physical examination or from links to other pre-existing databases.

Slide 18: The analysis of a cohort study is pretty straightforward. We setup our data in a 2×2 or in R×C, R number of rows, C number of columns table and we directly calculate the measures of disease frequency.

We calculate something called the cumulative incidence if there is uniform follow-up, meaning we have a number of people that have been followed for the same length of time. So we'll put the number of new cases divided by the number of individuals in our cohort.

On the other hand if there is variable follow-up, let's say it's a dynamic cohort where people can come in and out of the study as they move in and out of a community then we must take that variable length of follow-up of the participants into account.

So it would be new cases of the outcome, but divided now by a composite denominator which is a combination of a number of people in the study, times the amount of time that each of them has followed up as part of the study.

And then we can take these measures of disease frequency to cumulative incidence with uniform follow-up, the incidence rate if we have variable follow-up and we can go ahead and calculate the measures of association, the ratio measures which are to divide the two measures of disease frequency and the difference measures which are to subtract the two.

So the risk ratio is taking our two cumulative incidences with uniform follow-up and just dividing them and the rate ratio is taking our two incidence rates with variable follow-up and dividing them. The risk difference is **subtracting the two cumulative incidences and the rate difference is subtracting the two incidence rates.**

Slide 19: There are a couple of special issues of concern in a cohort study. There is the potential for bias due to loss of follow-up, because we are really now following people up for an extended period of time. And the concern is, if we lose people to follow-up and it is differential. So those who are exposed and develop the outcome are more or less likely to be lost to follow-up than those who are exposed who do not develop the outcome and those who are not exposed who do develop the outcome.

So we have a potential here now for the people that we have in our analysis being there differentially, either because of their exposure status and/or their outcome status, and that's a big problem in any study where you are following people up overtime whether it'd be cohort studies here or randomized trails which will be the next thing we talk about.

And the other issue just to be thinking about in a cohort study is the fact that we have usually a specific kind of cohort that often is chosen, not because their exposure is unique, but because their ability to follow them up is unique. So we would pick nurses, because we would be able to follow them up more easily over time than the general populations, or the Framingham inhabitants or an alumni association.

And the question is whether if you find the relationship between physical activity preventing heart disease in the Harvard alumni, is there any reason to believe that you can not generalize that finding to others who did not go to Harvard and were not in the study.

So we always have to be aware of the group in which we did the cohort study and think about and make a judgment as to whether those findings can be generalized to a broader population.

Slide 20: The advantages of a cohort study are many. We are much more likely to have the correct temporal sequence between the exposure and the outcome, because we get the information on the exposure prior to any evidence of the outcome emerging.

So it also generally involves good information on exposure status, because that's when we are coming into the study, that's when we can measure it most carefully.

Very, very efficient study to do when the exposures are rare because otherwise, we would need a tremendous number of cases and control in a case control study to have enough people who have this exposure for us to be able to compare the two groups and we can study several outcomes associated with a single exposure.

So now we have physical activity as our exposure and we can look at physical activity and ask about many, many things that will happen to people over time and we can follow them for all of those.

If it's a prospective cohort, so I have assessed exposure now and I am following them forward into time, it will minimize the bias in ascertaining the exposure because at the time the people tell us about their exposures, they don't know whether they are going to get the disease or not, they have not evidence of it yet and we can directly measure the incidence of the disease among the exposed and non-exposed groups.

Slide 21: But there are some disadvantages of cohort studies. They are generally inefficient for studying rare diseases because we would need a tremendous amount of exposed and non-exposed people to have enough, having the diseases for us to make their comparison, if it's prospective cohort again, meaning we assess exposure now and we all go along and follow that cohort over time, often it could be decades, is very time consuming to do and we have to worry about losses to follow up if we are going to make sure that we have valid results.

And if we went ahead and did it the other way of retrospective cohort where we actually got records from the past to assess the exposure and then we got records, other records to assess the outcome after that, then that study might be more efficient to do, less time consuming, less expensive, but it requires the availability of this pre-recorded information on exposures and confounders because we get it from the past, we are never going to be able to talk to the participants, everything must be available to us in preexisting record.

Slide 22: So what we are going to do next to say can we get rid of some of the disadvantages of a cohort study. In particular we said, we really have to worry about confounders, are there differences between the groups that we have to take into account. The next study design that we are going to talk about is actually the intervention study which is the structure of a cohort study, exposed and non-exposed, followed forward for the development of the disease but the investigator actually allocates the exposure and it will have some advantages over the cohort study that make it a complimentary technique for us to use to answer a question.

Slide 23: Thank you very much.