Cohort Studies – Part 1

Dr. Borenstein

NO AUDIO ON THIS SLIDE. CLICK NEXT TO CONTINUE >>
Welcome to the next section of Epi 6000. My name is Dr. Borenstein. My office hours are on Fridays from 2-4pm. You can see my phone number and e-mail address here. If you have questions regarding the materials for the class, please contact the TAs first. You may call me during my office hours or come and see me during that time.

A little bit about my background – I have my PhD in Epi from the University of Washington and my MPH from the Univ of Michigan. After I got my MPH, I worked for 1 year at the EPA and another year for a consulting firm that did most of their work with the EPA. I realized that environmental epi wasn’t what I loved, and that I wanted to do work in chronic disease epi. So I left for my PhD and later worked for 9 years at a research institute. In 1996, I came to USF.
I've spent my professional life doing work in the area of neuroepidemiology, focusing mostly on Alzheimer's disease. I've had the opportunity to conduct pretty much all types of observational studies – from case-control studies to prospective cohort studies, and also cross-sectional studies. We are now working on the design of a cohort study in Shanghai, China that will have both an experimental and an observational component. We will be talking more about these designs today.

I also have an interest in multiple sclerosis and am starting some work with the National MS Society, and a new interest in IBD.
Dr Stockwell has already introduced you to different study designs that are used in epi. Remember that study designs can be either descriptive: this is typically done with a series of cases or with ecologic data. Descriptive designs can not test a specific hypothesis, like, I want to see if walking one mile a day has any effect on blood pressure, they can only suggest an hypothesis to be studied in an analytic study. In an analytic study, we can test an hypothesis and either fail to reject the null hypothesis or reject the null hypothesis. If we reject the null hypothesis, we then have to be concerned about bias affecting our results. During the next four weeks, we are going to talk about how to do all these things. Today we are going to focus specifically on two types of analytic study designs: the prospective cohort and the retrospective (or historical) cohort study.

### Study designs

<table>
<thead>
<tr>
<th>Observational</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive:</td>
<td>Clinical trials</td>
</tr>
<tr>
<td>- Case report</td>
<td>Intervention trials</td>
</tr>
<tr>
<td>- Case series</td>
<td>Prevention trials</td>
</tr>
<tr>
<td>- Cross-sectional/</td>
<td></td>
</tr>
<tr>
<td>Correlational/ecologic</td>
<td></td>
</tr>
</tbody>
</table>

| Analytic:                      |                                     |
|--------------------------------|                                     |
| - Prospective cohort           |                                     |
| - Retrospective (or historical) cohort |                     |
| - Case-control                 |                                     |
| - Nested case-control          |                                     |


Before we talk about cohort studies, let's refresh our memories about the definition of an “association”. In an analytic study, we want to test whether there is an association, whether the association is VALID and finally, whether the association is CAUSAL. We are asking the basic question: Is X related to the Probability of Y? Or does the Exposure (E) predict the Probability of the Disease?
Now, associations can be defined as the extent to which things tend to co-occur. This includes whether X is dependent on Y (for example, age and education are dependent in the sense that older people tend to have less education than young people). However, increasing age does not CAUSE low education. In this sense, the association between age and educational is “non-directional”.

In epidemiology, we would like to establish CAUSAL associations, or “uni-directional” ones. We would like to know, for example, whether age predicts the incidence rate of breast cancer. Do older women have a higher risk of breast cancer than younger women? This question addresses the biological phenomena underlying breast cancer, the “etiology” of breast cancer. Some biological mechanism must underlie this association. For example, perhaps as we age, our immune system is less capable of fighting off cancer; or our DNA does not repair itself as easily as when we were young. Once we understand risk and protective factors, then we can study what is going on mechanistically at the cellular level.

It is important to understand that for a risk factor to be causal, the exposure (X) must precede the disease (Y) in time. So exposures MUST occur before the disease begins. The time “before the disease begins” is sometimes difficult to pinpoint, because many diseases take many years to develop. For example, having a myocardial infarct (MI) may happen one day but it takes many years for the pathology to develop before this event. Therefore, looking at exposures in the months or year preceding the MI may not adequately measure exposure history. You may have to go back much further, perhaps even as early as fetal life, because studies have shown that the fetal environment can affect the risk of chronic diseases in adult to late life. Suffice it to say at this point that we must be sure, to establish causal associations, that the exposure precedes the disease in time. How we define disease is also an epidemiologist’s task but it is more complex than our scope is at this point in the course.
Related to this issue of timing of exposure and disease, remember from Dr. Stockwell’s lectures that when we use a case-control approach, the odds ratio really tells us what the ODDS of exposure are in diseased individuals compared to what the ODDS of exposure are in non-diseased individuals, or controls. We do not necessarily know which came first, the disease or the exposure. This problem is fondly referred to as the “chicken-or-egg” dilemma. Of course we try to design our studies and our questionnaires so that we get information only on exposures that occurred before the onset of the disease.

Under some circumstances, we know that the E preceded the Disease in time – for example, when a genetic variant increases the risk of a disease. However, when the disease is present at birth, such as the mutation for Down Syndrome, we could argue about “when the disease begins”. This notwithstanding, it is accepted that a mutation on chromosome 21 is responsible for Down Syndrome.

The c/c approach is used when a cohort study would be too expensive. Say that the incidence rate of a disease is 1 in 100,000. You would have to follow one million people to find 10 cases. This is obviously prohibitive and this is why the c/c design exists.
We could think about all of epidemiology in the context of a cohort. Pretend that the world population is at our disposal. Pretend we could follow every person from the moment they are born (and that we could even measure exposures of the mother) until the moment they die (see the left and right bounds above). During each person’s lifetime, pretend we could document with no error every exposure they had – and document with no error all the diseases they developed (ideally those that were subclinical as well as those that were manifested clinically). This would be the best kind of observational study. There would be no sample, because we are studying every person in the world. We are documenting with NO error, all exposures and diseases. We can look at how any exposure predicts any disease.

Obviously, this fictional fantasy study is impossible to do. Real life is much more messy. We can’t study everyone, we have to pick samples. We can’t document everyone’s exposures from conception; we rely on their memories or we measure them. We have to define the disease we are studying and find people to compare them with. When we talk about cohort studies, though, the ideal cohort study would be one where we follow everyone from birth to death, and document their exposures and outcomes along the way. All of epidemiology is modeled on this concept, although you probably won’t find that statement in any book. C/C studies are designs in which cases and controls are found at one point in time, pulled out of this mythical cohort and cases and controls are questioned about their habits, behaviors, chemical exposures, diet, etc. Of course you already know that there can be a lot of error involved in recalling this information. We will talk more about that in Unit 7. Only in cohort studies are we able to calculate incidence rates. Incidence rates are needed, as stated by Rothman and Greenland, because they document the speed with which a disease develops and we can compare the incidence rates among those exposed to those among those non-exposed individuals in order to calculate the rate ratio, or risk. All of epidemiology is built on this concept.
To complicate matters further, we can not only calculate the association between X and Y. We must also account for other variables that might explain some of that association. These are called CONFOUNDING VARIABLES. For example, let’s say we’re interested in the association between diet (X) and cardiovascular disease (Y). What are some other explanations for this association? People who do not exercise are at higher risk for CVD. People who do not exercise often have poor diets. So we have to measure exercise and treat this as a potential confounding variable. What else do we have to measure? Poverty is associated with higher rates of CVD. Poverty is also associated with a poor diet. So poverty, which may be difficult to measure, must also be treated as a PCF. We will deal with this problem in Unit 8.
Finally, when we consider how E, or X, predicts D, or Y, and we’ve accounted for PCFs, we have to think about the causal pathway. We can draw Directed Acyclic Graphs or DAGs, like the one shown above. We should think when we start a study about how we anticipate each variable to be related to the other variables so that we know what we want to measure in the study and how we plan to analyze the data after we get them. These DAGs are also known as causal pathways, or causal webs, as we will see on the next slide.
This is an example of a DAG for coronary heart disease. The DAG shows “heart disease promoters” or risk factors; and “heart disease inhibitors” or protective factors. The interplay of all of these factors lead to either a heart attack or to the prevention of a heart attack. The epidemiologist’s dream is to eliminate disease. Ideally we would die in our sleep and we would have gone to bed in a healthy state. In the epidemiology of aging, this dream is known as the “compression of morbidity” and “rectangularization of the mortality curve”. See the links for more detail
What is the definition of a “cause”? It is a... This definition implies that we can either PREVENT the disease event from happening all together OR we can DELAY the disease event to occur at a later time. For example, in Alzheimer’s Disease, much of the risk is determined genetically. However the AGE at which someone shows symptoms of AD depend on a person’s habits, behaviors and conditions. Higher levels of education, for example, can delay the age at which an individual presents with the disease. If we can delay the onset of AD by 5 years, we could cut the incidence rate of the disease in HALF! That means we would have HALF as many new cases in one year as we have now. So many variables we study are those that may REDUCE the incidence of the disease. If we are interested in variables that reduce or increase the incidence of disease, we are always concerned that these exposures occurred BEFORE the onset of disease. That is key.
Articulate Quizmaker Quiz Placeholder - review1_lecture5_part1
Let's talk now about cohort studies in more detail.
The essential concept behind any prospective design, be it cohort, historical, or a randomized controlled trial, is that: we identify the exposures first, we follow the population over time, we assess the outcomes later and we then compare the incidence rate of the disease in the exposed and unexposed. E(Bar) here means unexposed group. The only difference between an observational cohort study and an experimental study is that in the latter type of study, the Exposure is manipulated by the Investigator. For example, the investigator has a cohort of 200 people available, say these people are patients who have suffered from an MI. The investigator wants to find out whether a special diet decreases the incidence of 2nd MIs. The investigator designs the study and teaches the patients what to eat and how to keep a food diary. The 200 people are then randomized: one group eats normally (placebo group) and the other group gets the special diet. After, say 2 years, the groups are compared. Did the group who got the special diet have fewer 2nd MIs than the group who didn’t get the diet? We can compare the incidence of 2nd MIs in the two groups to get the incidence rate (in this case we would get the incidence density because we can follow each person and add up the total amount of time they were in the study). An observational study could do the same thing but document the regular diets in the 200 people and then define a “good” diet and a “poor” diet and look at incidence rates of 2nd MIs in these two groups. Note that in the latter example, the investigator did not manipulate anyone’s diet.
We would like to emulate an experimental study design...

- Experimental animal studies:
  - Breed mice so genes are controlled. Control environment in which mice live.
  - Have enough mice so don’t make the mistake of not having sufficient sample size to find something if it exists.
  - Randomize mice to treatment (E) and placebo groups (E)
  - Administer controlled doses of medication by mouth (diet)/injection/topically (compliance is assured)
  - Compare clinical disease in Rx and Rx mice
  - Sacrifice animals, look for pathologic disease in Rx and Rx mice

The experimental study design is the best type of design to test whether a factor is associated in a causal manner with an outcome. This is because when we randomize people, we get people, on average, with similar characteristics in the two groups (exposed and placebo). The larger our sample size, the better the randomization works to equalize the two groups as to other characteristics, such as age, gender, education, genes, etc. Randomization even levels the two groups by variables that are unmeasurable (e.g., all genes in the genome). Laboratory scientists do the best studies: they breed their own mice, so they can control the mice’s genes. They control the environment in which the mice live from conception to death. They breed enough mice because they determine in advance, as epidemiologists do, how many individuals they will need to find a meaningful result. They then randomize the mice to either an exposed group or a placebo group. They then administer the exposure (perhaps it is a medication or a diet or some other intervention) under controlled conditions to the mice. This assures that the mice are complying with the intervention. They then run behavioral studies or other tests on the mice to look at clinical disease, and finally they sacrifice the animals to look for pathology.
In humans, there are three kinds of experimental studies:

- Randomized controlled trials (single- or double-masked)
  - Usually conducted among patients to test drug efficacy,
  - or, among high-risk individuals (e.g., MCI to AD) – secondary/tertiary prevention

- Intervention/prevention trials
  - Conducted among asymptomatic individuals (much larger numbers needed) – primary/secondary prevention

- Community-based trials
  - Unit of measurement is not the individual but a whole community (e.g., fluoride), neighborhoods, schools (classrooms receiving intervention) - primary/secondary prevention
Most of the time, we can’t randomly allocate an intervention in human populations. Many of the exposures we study are risk factors, and we ethically can not randomize people to, say a smoking group and a non-smoking group. Even in drug studies, it may be unethical to have a group on placebo, so most “placebo” groups are people who are being treated with standard of care. When we do an experimental study, we would like to get as close to an experimental study of animals.

What is under our control? The size of the sample – this is very important and we will talk about this in the next unit. If we can do an experimental study, we have control over whether the randomization is done properly. The larger the sample that is randomized, the better the randomization will work. We have to have a measure of compliance as we follow people over time to make sure they are taking the drug, or doing the intervention. This has to be carefully thought out for each situation. Also, we can design the trial to be long enough so that we would observe a difference in incidence rates between the two groups that is clinically meaningful.
Experimental studies vs. observational studies

- Elements of RCTs that are not under investigators’ control:
  - People may not comply with treatment
  - People drop out of the study: death, refusal, move out of study area
  - People are inherently heterogeneous, unlike bred animals
  - People don’t live in controlled environments like experimental animals
  - There may be variations by ethnicity so RCT may only apply to certain subpopulations studied

Some elements of experimental studies that are not under the investigator’s control –
Most times we have to settle for looking at people’s exposures (voluntary exposures, such as their behaviors, work exposures and diet -- and involuntary exposures, such as second hand smoke and chemical exposures in the workplace, infectious agents, environmental agents) and see who develops disease and who doesn’t. This is where the observational cohort study fits in.

The problem with this is that --

The people who happen to be exposed are usually different from the people who are not exposed. For example, people who smoke tend to have different characteristics than people who do not. People who do not smoke many more likely be exercisers, have a more healthy diet, go to regular doctors visits and get screening tests, be of higher socioeconomic status. We can adjust to some extent for the differences found between two groups but sometimes we can not adjust for all of the factors, because there may be too many to measure well, and because some variables are unmeasurable, such as genes.
Sometimes a case-control study is better than a cohort study!

- In 1950’s and 1960’s, diethylstilbestrol (DES), a hormone used to prevent miscarriages, was given to women during pregnancy.
- Case report in 1970 of 7 cases of vaginal clear cell adenocarcinoma (CCAC) in the daughters of mothers who took DES.
- Case-control study in 1971 (Herbst, Ulfelder, Poskanzer, NEJM 284:878-81) of 8 cases of CCAC and 4 matched controls (32 controls matched on birth within 5 days of proposita in same hospital and same type of service (SES)).

All this is not to say that a cohort study is always preferred over another design. Sometimes a case-control study is better than a cohort study. A good example of this is that of DES, which was given to pregnant women in the 1950’s and 1960’s to prevent spontaneous abortions. In 1970 there was a case report of 7 cases of vaginal clear cell adenocarcinoma (CCAC) in the daughters of the mothers who took this drug during their pregnancy. The daughters were young, in their 20’s when the cancer was diagnosed. This case series led to a case-control study conducted in 1971 in which 8 cases (they found another case) were matched to 32 controls (4 controls for each case). Controls selected had to be born in the same hospital as the case within 5 days of the case’s birth, and had to have been born on the same type of service ward, which to some extent controlled for socioeconomic status.
Here is a table from the paper, published in the New England Journal of Medicine in 1971. You can see the case number in the column on the far left. Look at the 6th column: “Estrogen given in this pregnancy”. Here you can see that 7 of the 8 cases’ mothers were given DES during the pregnancy, compared to zero of the 32 controls. If we make a 2x2 table representing this column, we have 8 cases, only 1 who was not exposed to DES. In the controls, 0 of the 32 controls was exposed. We calculate the Odds Ratio and find that it is 224/0 which is equal to infinity. If we change the zero to a 1 so we can do the calculation, it turns out the Odds Ratio is a very large number: 224. The frequency of the exposure in the cases and the rarity of the exposure in non-cases is very telling here. The further away from 1.0 the OR is, the more likely the association is real and important. This is a very large OR. There was almost unanimous acceptance in the field that DES caused CCAC. But there was one problem: this was a case-control study and maybe the cases and the controls were a biased sample? So a prospective study was done later:
In the exposed group, there were 900 girls whose mother took DES when they were in gestation. They were compared to 800 unexposed girls who had not been exposed. Even in this very large cohort, there was not even 1 case of CCAC found in the exposed girls. It turns out that the incidence rate of CCAC is 1 in 1,000 among DES-exposed fetuses. Just by chance, the cohort assembled of 900 exposed girls did not yield even 1 case of CCAC. The disease is just too rare. It can only be studied using a case-control design. So, if a prospective cohort study had been conducted first, this association would not have been discovered. So you have to think very carefully about which design to use and you have to know a good deal about the exposures and the outcomes you are studying.
There are three kinds of prospective studies. The thing that distinguishes between them is where the investigator is when the study is done. The thing that all three studies share is that exposures are always being identified first, and the cohorts are always being followed through time to ascertain the outcome later on. Because of this, all subjects MUST be free of the disease of interest at the baseline, or the beginning of the study period. This is true of all prospective studies. In a case-control study, cases and controls are identified and their exposure histories compared. There is no element of time, as we have discussed already. In a cohort study, people are being followed through time. The difference is whether they have already been followed (historical cohort study) or whether they are currently, in real time, being followed (prospective cohort study). In the example on this slide, we are talking about a prospective cohort study. The investigator, who here is your third professor in this course, Dr. Sanchez-Anguiano, is on board at the beginning of the study. She designs the study, decides which population to sample, what will be measured. Then we all grow old together – the subjects and the investigator. The cohort has to be kept track of, and loss to follow-up has to be minimized. This is Dr. Sanchez-Anguiano’s job. She has a team of trained interviewers and study personnel to help her track people over time and retain them in the study. Depending on how long the follow-up period is, the study could last 1 year or 50 years. It depends on what the research question is.
Historical cohort study (retrospective cohort)

- Exposure status (data) is documented for the beginning of the observation period in a disease-free cohort.
- Incidence of disease is ascertained over the follow-up period.
- Both exposure and disease have already occurred, but the sample is selected by exposure status (not real time).

Dr. Stockwell is an occupational epidemiologist, and she is used to doing historical cohort studies. This design is commonly used when we want to study worker populations and their exposures. We still start with exposure data in a disease-free cohort. But now we will start with an exposed worker population, follow them through time, and wait for the disease of interest to occur. We will then count the number of incident cases in the exposed cohort. We will concurrently find a second worker population, or sometimes we will use the general population, as a comparison, or non-exposed cohort. They too will be free of the disease of interest at the baseline (beginning of the study). Often, industries keep records of their workers' job titles and sometimes even of their health status over time. An example I've dealt with is looking at employees working at an aluminum smelter in Washington state. We were interested in whether workers who were exposed to aluminum in the pot room had worse neuropsychological outcomes over a 20-year period than workers who worked in another part of the factory. The factory kept records of who worked where and for how long. They also conducted neuropsychological testing on all the workers every year. We were able to look at their records and do our analysis in one year. All the data were already collected. In this situation, the investigator has to deal with the data that are available. In a prospective cohort study, the investigator gets to design which data will be available. Obviously, the prospective cohort study gives the investigator more control over the content of the study. The down side is that in a prospective cohort study, we are following the population for the outcomes, and we have to do the study in real time. This is obviously expensive and time-consuming. If a data set has all the data you need to answer a question, then it is cheaper and faster to analyze the data instead of getting them all yourself.
The third type of cohort study is the ambi-directional study. It has both prospective and retrospective (historical) components. Again, the incidence of disease is ascertained over a set follow-up period in a disease-free cohort. We can measure exposures many times during the follow-up period (this applies to all cohort designs). In the ambi-directional study, the disease outcome(s) is measured through real time by some investigator. Then you as the investigator step in, and help to maintain the cohort through real time. An example is the Honolulu Heart Program and the Honolulu Asia Aging Study. The study began in 1965 and they examined over 8,000 Japanese American men who lived on Oahu, Hawaii. From 1965 till 1991, the study was primarily directed at looking at risk factors for Coronary Heart Disease and stroke, and was funded by the NHLBI. In 1991, the NIA took over the study and it became a study of aging (HAAS). The investigators for the HAAS were interested in dementia outcomes. They still documented heart disease events and stroke, but they added on a component to look at memory and dementia. This study continues today. Obviously, a study that began in 1965 has to have different investigators when we are talking about decades of follow-up. There are only a handful of studies like this in the U.S. – and they have produced very interesting results, such as the role of hypertension in midlife and its relation to incidence of Alzheimer’s disease many decades later. The ambi-directional study sample is selected by exposure status, as in all prospective studies.
Example of a prospective cohort vs. historical cohort study

1. Prospective: investigator accompanies Ss through time

2007 → 2017-2022 → 2062-2067
Identify elementary school students (measure baseline characteristics @ ages 5-10) → Measure smoking status at ages 15-20 → Measure lung diseases at ages 55-60

Investigator is dead

2. Historical cohort: investigator can do entire study in 1 year

1942 → 1960 → 2007
Identify elementary school students → Survey of smoking habits → Trace outcomes using Medicare files

Investigator is still young and has a long career ahead of him

Here is an example of a prospective cohort vs. a historical cohort study. In #1, we start today to identify a cohort of representative elementary school students. We measure all the characteristics about them that we wish to measure, and we follow them through time. In 10 years, we measure their smoking status when they are 15-20 years old. Then we wait until 2062 and measure their lung diseases. Obviously, this approach will not work because the investigator will long be dead. We can not conduct prospective cohort studies that are 60 years long. However, we could still use a prospective approach if we could go BACK in time. For example, say that there already exists a cohort of elementary school students who were 5-10 years old in 1942. Say also that someone had done a study linking this cohort to their smoking habits in 1960. Imagine further that we could trace the outcomes of these individuals today, in 2007. We can still start with the cohort and separate them into smokers and non-smokers and look at their outcomes until 2007. We can still calculate incidence rates. We get all the advantages of a cohort study without having to wait. The investigator can complete the study in a couple of years and has a long career ahead of him. The problem? Someone else has to have already conducted the study! And the variables that you need may not all be there. And you may not know how accurate the measurements of the variables are. Either way, someone needs to do the prospective part of the study. The directionality of the study (prospective vs. historical) is defined here as where you as the investigator, are, at the time data collection. If you are at the beginning, and you are following subjects in real time, the design is a prospective cohort. If the data have already been collected and you are at the end, then the design is a historical cohort study.
Every prospective study becomes a historical cohort study if you wait long enough...

- When data collection is over and the study is terminated (usually because funding has ceased), a prospective cohort study becomes a retrospective cohort study, i.e., all the data have been collected.

So in essence, this is a moot point because every prospective cohort study becomes a historical cohort study if you wait long enough; and every historical cohort study was a prospective cohort study at some time in the past. The distinction is important in a methodologic sense, because you have to recognize a historical cohort study as such and you should know right away that you can calculate incidence rates from such a study.
Now, we have talked about a prospective cohort study but we need to emphasize that in order to study the incidence, or the rate, or speed of developing a newly diagnosed illness, we cannot include in the denominator individuals who have already achieved that end point. In other words, people with prevalent disease must be removed from the cohort before we can start time – healthy-person time – ticking. Remember that prevalence = Incidence x Duration. If we leave prevalent cases in the cohort, we will mistake them as incident cases the first time we see the cohort again and we will grossly overestimate the incidence rate because we will be calling prevalent cases incident ones. Also, we cannot study the rate of new disease development among people who already have that disease. It doesn’t make sense. So it is very important to use a disease-FREE cohort when beginning an incidence study. This is not as easy as it may seem. It involves screening for the disease of interest in the baseline cohort (everyone at Time Zero), finding and diagnosing the prevalent cases, and being sure that only disease-free individuals enter the follow-up time clock. This means, as shown in this slide, that the cohort at Wave 1 (the blue circle on the right) will be a little smaller than the cohort that we started with (on the LEFT) if and only if we have a fixed cohort, which means that we are not adding new people to the cohort. Remember too, that unlike mice, not every person will complete their time until Wave 1. People move, die, and refuse to continue in the study. Time in health stops when a person develops the disease of interest, or is lost to follow-up.
Choose a cohort design when:

- The outcome is relatively common
- The exposure is relatively rare
- You want to study multiple outcomes of an exposure
- You are able (or try hard) to minimize loss to follow-up
- The interval between E and D is relatively short or when you think you can follow the cohort for a long enough time to account for the latent period (induction period) of the disease
- You want to be confident about temporal sequence (that D follows E in time)
- You want to study a particular cohort (e.g., The Nun Study, The Kame Project)

So under what circumstances do I want to choose a cohort approach? You have already learned about under what circumstances to choose a case-control approach. We do so when the disease is rare and the exposure is relatively common. The opposite is true in cohort studies. A cohort design is most appropriate when the outcome is relatively common. This makes intuitive sense: if I follow 1,000 people for 1 year and 300 people get the disease, that is a pretty good number. If I follow 1,000 people for 1 year and 10 people get the disease, that is not as good. It will take me many years before I have enough cases to be able to find anything meaningful. This is the concept of Power that we will tackle in the next unit. The exposure can be relatively rare, as we shall see later on. This is true especially when we use an internal exposure group and an external unexposed comparison group (later on in this lecture). In a cohort study, you can study multiple outcomes. Remember, you are waiting for the outcomes to occur, so you can measure the outcomes you are interested in. You may even be interested in intermediate outcomes, like the development of diabetes or hypertension in a study of MI. You must try hard to minimize loss to follow-up, because this can really affect the success of the study. Also, you need to think about how long you have to follow people to allow a sufficient amount of time to pass for your outcome to occur. If you are following a cohort that boarded a cruise and you’re interested in Norwalk Virus, this may only be a matter of days. More commonly, we would wait years to document the development of Alzheimer’s disease or breast cancer. These studies are very expensive and difficult to conduct, so we have to make sure we have a sufficient sample size and we account for the latent or induction period of the disease we are interested in measuring. This will make us more confident about the temporal sequence, that the disease is a result of the exposure (and not the other way around). Lastly, we will use a cohort design if we are interested in a particular cohort, such as the two studies I will discuss next.
Here is an example of a cohort study that is one of the gems in the field of Alzheimer’s disease. You may recognize the Co-Principal Investigator’s name here, if you’ve taken Biostat I in the College. This study began in the early 1990s. The PI, Dr. Snowdon, went to the convents of the potential participants, the School Sisters of Notre Dame. The investigators invited Nuns living in US convents who were born before 1917 to participate in an annual examination. They asked for access to archival and medical records and to donate brain upon death for neuropathologic studies. Of 1,030 Sisters eligible and invited, 678 (66%) agreed to participate. Nuns were age 75-102 at baseline.

Here is an example of a cohort study that is one of the gems in the field of Alzheimer’s disease. You may recognize the Co-Principal Investigator’s name here, if you’ve taken Biostat I in the College. This study began in the early 1990s. The PI, Dr. Snowdon, went to the convents of the potential participants, the School Sisters of Notre Dame. The investigators invited Nuns living in US convents who were born before 1917 to participate in an annual examination. They asked for access to archival and medical records and to donate brain upon death for neuropathologic studies. This was the first study in Alzheimer’s disease where a whole cohort of people could be examined clinically and then neuropathologically for AD. Because the “definitive” dx of AD can only be given when the postmortem tissue is examined, this provided unique data that were not available in any other population.

It is always important to look at the participation rate in any epidemiologic study. In this study, 1,030 Sisters aged 75 to 102 were eligible at baseline and invited to participate and 66% agreed. This is quite high, considering the requirement of brain donation at death. In the 1970’s and 1980’s, participation rates in the community at large were pretty high: about 80%. Today, participation rates are much lower (about 50%). This is definitely bad for research (we will discuss why in Unit 7).
Some of the Sisters are also siblings. Here is a picture of the Gore girls in midlife.
In the panel on the top is a normal brain from the Nun Study. In the bottom panel is a picture of an Alzheimer brain. Notice how the spaces (sulci) increase while the gyri decrease. This is called atrophy. The brain is actually shrinking. This atrophy is thought to lead to the dementia, or loss of intellectual capacity, seen in Alzheimer’s disease. The Nun Study was the first study to be able to quantify all four cells seen on the right side of the slide. Because in the real world people get autopsied for a reason, only a proportion of individuals with dementia will ever get an autopsy. So in the real world, we would be happy if we could quantify accurately the yellow cells. In this study, they were able to quantify accurately not only the yellow cells but also the red hatched cells. We could calculate the Relative Risk (using cumulative incidence, since we do not have person years in this table) of the association between being demented and having AD pathology. That is 190/223 over 81/455. Also, a very interesting fact that this table tells us is that of 271 Sisters who met neuropathologic criteria for AD, 81 showed NO DEMENTIA during life. If we generalize this finding to all people, we can conclude that about 30% of people whose brains are riddled with the pathology of AD have no cognitive symptoms during life. Since the neuropathologic dx is the “gold standard”, we can ask the question: What characteristics about these women buffered them from showing the clinical disease when in fact they were walking around with the real disease in their brains? This finding has led us to many of our current studies.
I will be talking about this study too as we move along in these next units. This is a study that I did from 1990-2001. We were interested in studying rates of AD and VaD in Japanese populations in Japan, Hawaii and Seattle. Our hypothesis was that as persons of Japanese descent moved toward the West, their rates for AD would go up and their rates for VaD would go down. This is because the lifestyle and diet in the West predisposes more to cardiovascular disease and AD, and the lifestyle and diet in the East predisposes more to stroke, the major cause of VaD. In the three locations of the study, Hiroshima, Oahu and Seattle, the research teams met together many times to standardize case-finding and diagnosis in the three cohorts. We also tried to standardize our questionnaires and how we asked the questions. Our study in Seattle was funded by the National Institute on Aging, part of the NIH, from 1991-2001. I wrote the grants that were funded and was responsible for the whole study – hiring, training personnel, implementing the study in the field, data collection and analysis.
This slide shows how we found individuals eligible to invite to participate. In 1991, we started a study census. Most of the individuals were found through the telephone directories of King County. Community volunteers went through a stack of directories that literally were 1 ½ feet tall. They combed through every name looking for a Japanese-sounding first or last name. We also found people from a variety of other sources, including Japanese-American organizations in the community. In total, we identified 3,196 people, 1,993 of whom agreed to participate. This is a participation rate of 62%, which is ok, but not great. The U.S. Census Bureau uses 60% as the participation rate that is required for a study to be valid. We’re above that, but just by a little. 96% of those found in the study census had 2 parents of Japanese descent. Some of the subjects were born in Japan, but most were born in Seattle.
I just want to summarize the methods of this study. The terms should be familiar to you now. We first conducted a prevalence study in the baseline cohort of 1,985 individuals. We found 149 prevalent cases of dementia. We then followed the NON-DEMENTED cohort for up to 8 years, with 4 follow-up waves, each 2 years apart. These are called “biennial” visits. The study was stopped in Dec 2001. At each wave, surviving members of the cohort were given a cognitive screening test, which was scored out of 100 points. If they scored less than 87 points in any wave, whereas they had scored above 87 in a previous wave, they were invited to come in for a clinical evaluation, which included neuropsychologic testing, blood work, and in some cases a CT scan. We found a total of 173 INCIDENT cases of dementia. Stop and calculate: What is the crude prevalence rate? 149/1985=7.51%. What is the cumulative incidence rate? 173/1836=9.42%, but that’s over 8 years, so to get the annual cumulative incidence rate, we have to divide by 8, so that comes to 1.18% per year. I have not given you person-years information, so you are not able to calculate incidence density rates.
Translation...

This lays out the design in a chronologic format. We start on the left and move to the right...
At end of study period

- 149 prevalent cases of dementia found at baseline
- 172 incident cases of dementia found over eight-year follow-up period among non-demented cohort
- Over 10,000 person-years accumulated
- Analyses examine how risk factors from baseline predict incident cases (best to study etiology since \( P = I \times D \)).
  - Prevalence does not measure risk.

So, at the end of the study period, we had found: See slide
Articulate Quizmaker Quiz Placeholder - review2_lecture5_part1
This is a reminder from the first third of the course. Remember that there are two ways to calculate incidence rates: the cumulative incidence rate... see slide.
Here is an example of fictitious data to calculate the two types of Relative Risk. The first one in the upper panel shows the number of people who became cases (54) among those who were exposed (486). This is the Ie. We have to compare this to something. We compare this to the number of people who became cases (75) among those who were unexposed (1,489). No person-time is given to you in this example. Therefore, in this example, we calculate the RR using Cumulative Incidence rates. In this fictitious study, there are 1,489 individuals in the denominator. We can STRATIFY these 1,489 people by EXPOSURE STATUS to get the exposed cohort and the unexposed cohort. In this cohort, 486/1975 or 24.6% of the cohort is exposed. You can see intuitively that if the exposure rate was very low, say 5%, we would only have 11 cases in the upper left-hand cell. This may be too small to analyze. So even though we say that low exposure rates are suited for cohort studies, that is a relative statement. We still have to have sufficient number of people exposed who develop the disease to have meaningful results.

In the lower panel, I am showing you the data you would need to calculate incidence density rates, because I am now giving you the number of person-years that accrued in the cohort, stratified by exposure status. Here you can see that the RR using incidence density rates is the same as in the first panel, and you can also see where the word “relative” for the “relative risk” comes in (see red font in 2nd RR calculation).
A note about the book: to calculate person-time, do not learn Oleckno’s way of doing it. We feel this is unnecessarily complicated. Do it the way we have taught you in the class. Calculate the denominator by adding up the person-time of health contributed by each individual being followed. Calculate the numerator always as the number of events. Express the rate per 100 or 1,000 or 10,000, or 100,000. Remember that incidence and prevalence rates always carry an element of time, and that this unit should always be included when an incidence rate is calculated.
Person-time

- Person-time accumulates when we observe a group of individuals over a period of time to ascertain the development of an event.
- Although we would like to follow everyone in the sample indefinitely, (cumulative incidence), the actual time each individual is observed will most likely vary, since:
  - Subjects may be recruited at different times
  - Subjects migrate (move)
  - Subjects choose to leave study (dropout)
  - Subjects die
  - Subjects develop the disease of interest
  - The study ends

You have already learned about person-time, but I’d like to emphasize a few points about it in this lecture about cohort studies. SEE SLIDE
Also, recall what we are interested when we calculate a relative risk: We compare the incidence rate in the exposed to the incidence rate in the unexposed. The value of 1.0 is considered the “null value”. This is the value of NO ASSOCIATION. If the incidence rate in the exposed is higher than that in the unexposed, then we call the variable a RISK FACTOR. If the incidence rate in the exposed is LOWER than that in the unexposed, we call the variable a protective factor, or we say that there is an INVERSE association. Remember also that the OR, RR, and HR (the hazard ratio, or the comparison of two incidence density rates) are on log scales. That means that an inverse association of 0.10 could also be expressed as an OR or RR of 10, if one were to reverse the reference groups. For example, say we are studying the association between antioxidants in the diet and cardiovascular disease. Say that individuals who consume a high anti-oxidant diet (we’d have to define what this means) are 10 times less likely to develop an MI compared with people who had a low anti-oxidant diet. We are using the “low” anti-oxidant diet group as the reference, or comparison group. If we were to reverse it, we could say that the low anti-oxidant diet group were 10 times MORE likely to develop an MI than the high antioxidant diet group. Both statements are correct, it just depends who you use as the “reference” or “comparison” group.

Also notice that the log scale means that an association >1 can go from 1.0 to infinity. On the left hand side of 1.0, we are bounded by the zero.
Now remember also that we do not usually calculate ORs when we have rates. Under some circumstances, which are topics in more advanced epi courses, you might want to, but for the purpose of this class: whenever you have rates you will NOT want to calculate an OR. The OR is used to approximate the RR. If you have rates, use them.

OR vs. RR

- Do not usually calculate ORs in cohort studies because we have rates. OR is usually used to approximate the RR.
- If we have rates, we use them.
Articulate Quizmaker Quiz Placeholder - review3_lecture5_part1
Please continue to...
Cohort Studies - Part 2