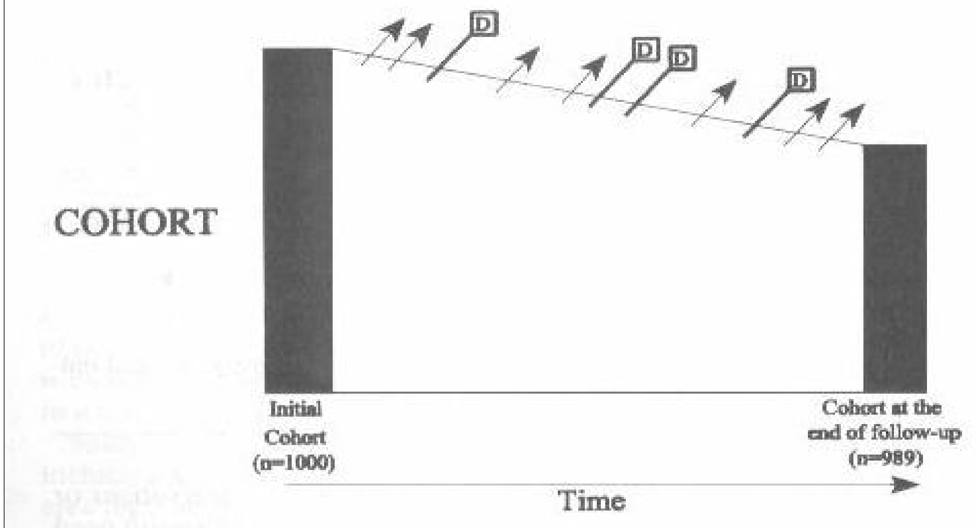
Cohort study design

D = disease occurrence; arrow = losses to follow-up

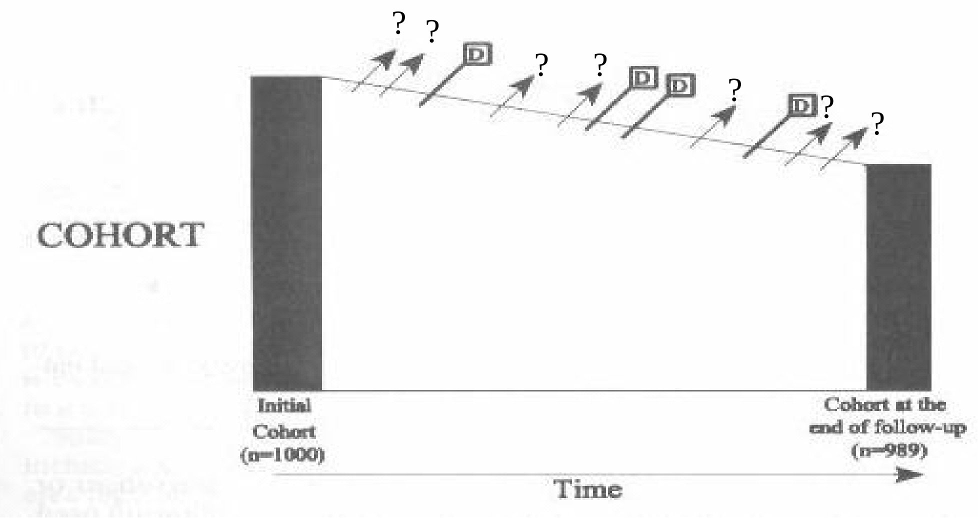


Minimum loss to follow-up (1%)

A cohort study begins with a group of individuals identified at time zero, none of whom have the disease outcome of interest, who are then observed over time for diagnosis of the disease. Some persons develop the outcome of interest, here depicted as D for disease. Others become lost to follow-up, while others stay in the cohort but have not yet developed the outcome of interest by the last date of study observation. In some studies the outcome might be death, but that is not the intention here. Unless death is the outcome, deaths are one of the ways subjects become "lost to follow-up." In other words, in this graphic subjects who died would be among those leaving the cohort, represented by the arrows. Others could be those who decline further follow-up, those who cannot be located for return visit/evaluation by the investigators, and, in some studies, those who experience a disease outcome other than the one under investigation which makes them ineligible for further follow-up.

There is the further assumption in this schematic that the disease diagnosis is a one-time event. Although this is frequently the case in cohort studies either because the event can only occur once (eg, death) or because the focus is on time to the first event (eg, time to first myocardial infarction), repeating events can also be studied (eg, frequency of debilitating back pain). In the case of repeating events a diagnosis does not remove an individual from follow-up as depicted in the above text book schematic.

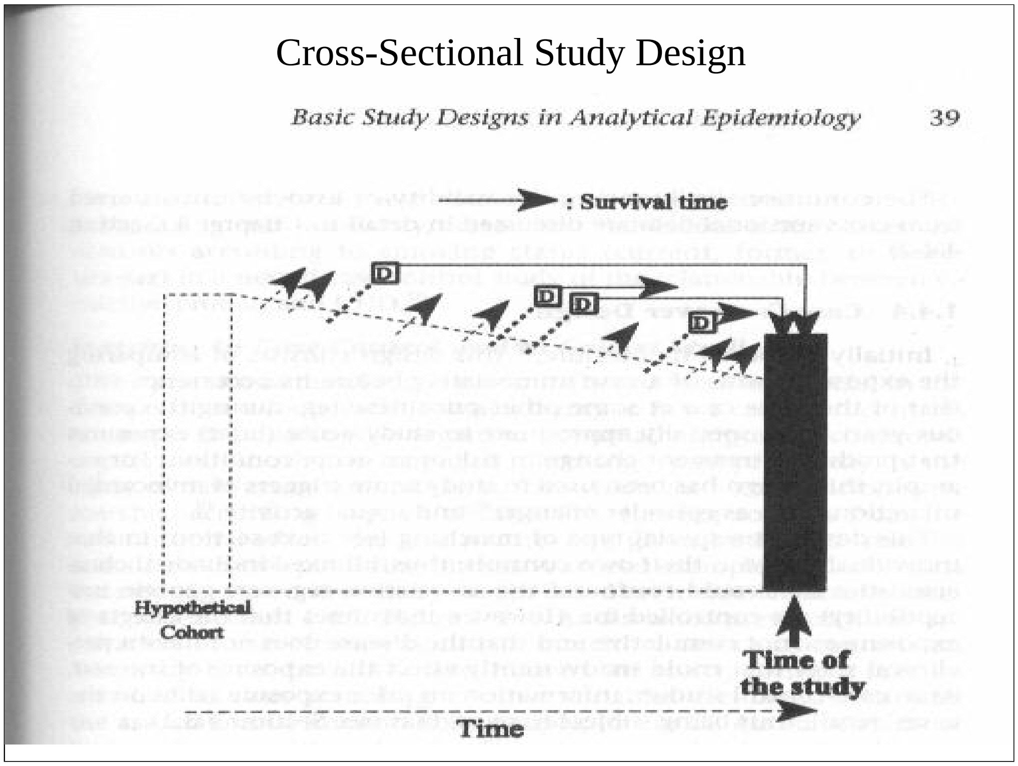
Crucial issue is who is leaving cohort: What bias do the losses to follow-up introduce? Are disease diagnoses being missed? Are those with a risk factor more likely to leave and then be diagnosed?



In practice, it may often not be possible to answer these questions in a cohort study for the simple reason that subjects who are lost or refuse to participate further may not be available to supply the answers.

At a minimum you need to attempt to answer them with whatever information you do have about the characteristics of those leaving the cohort by comparing those characteristics with those you retain.

Another strategy to guard the validity of your result is to do a sensitivity analysis in which you assume a worst case scenario for those you have lost and see how much it could affect your findings. We will discuss this option more later on.

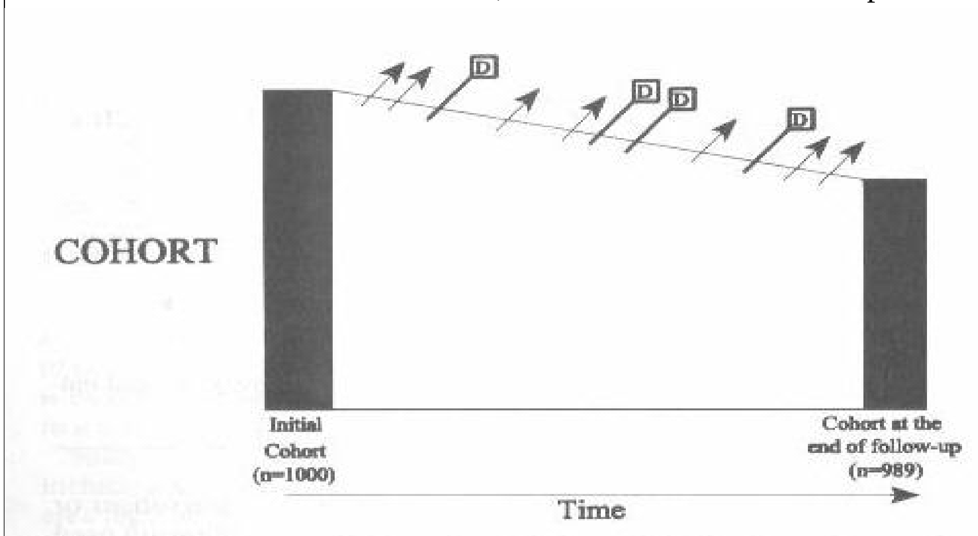


It is easy to describe a cross-sectional study as a sample of a population at one point in time, a cross-section of that population. What is perhaps not so well appreciated is the point illustrated in this graphic from the text by showing the design in the setting of a hypothetical cohort. It demonstrates the prevalent nature of the sample. In other words, only those individuals who were present at the time of the cross-sectional sample have a chance to be included. So, for example, in the illustration there are two members of the cohort who were diagnosed with the disease outcome who did not survive to the time of the sample. Cross-sectional sampling, then, will only capture prevalent cases of disease, which means that the probability of inclusion is related to the length of disease duration or survival. It will over-represent those cases of the disease with longer disease duration or survival times.

Likewise, those without disease are also “prevalent,”meaning that persons with certain characteristics may be more or less likely to be represented in the cross-sectional sample. This would be the case if the individuals who left the population, represented by the arrows in the schematic, differed on characteristics of interest from those who remained to the time of study.

Cohort study design

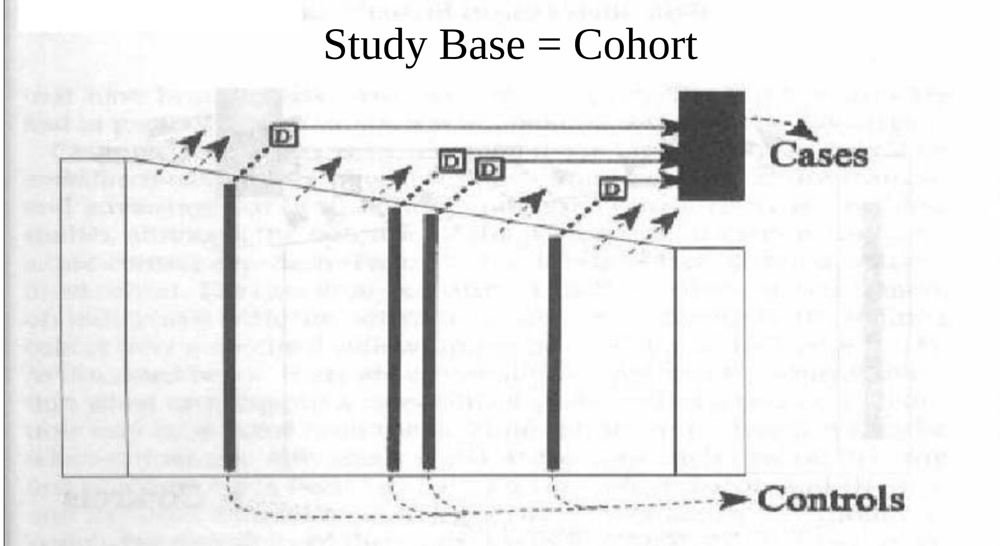
D = disease occurrence; arrow = losses to follow-up



Given that all the cases are diagnosed, how would you sample controls from this cohort for a case-control study?

The main advantage of case-control designs is that it allows you to sample the experience of the study base most efficiently. Stated in other words, case-control designs allow you to make measurements on far fewer subjects than cohort studies but still get the same answer. The reason to do this is to conserve resources, something that is becoming more and more important these days as funding is drying up. A typical example is when expensive testing on stored biological samples are required for an analysis. It is often prohibitively costly to test everyone in the cohort.

Incidence Sampling within a Cohort Study



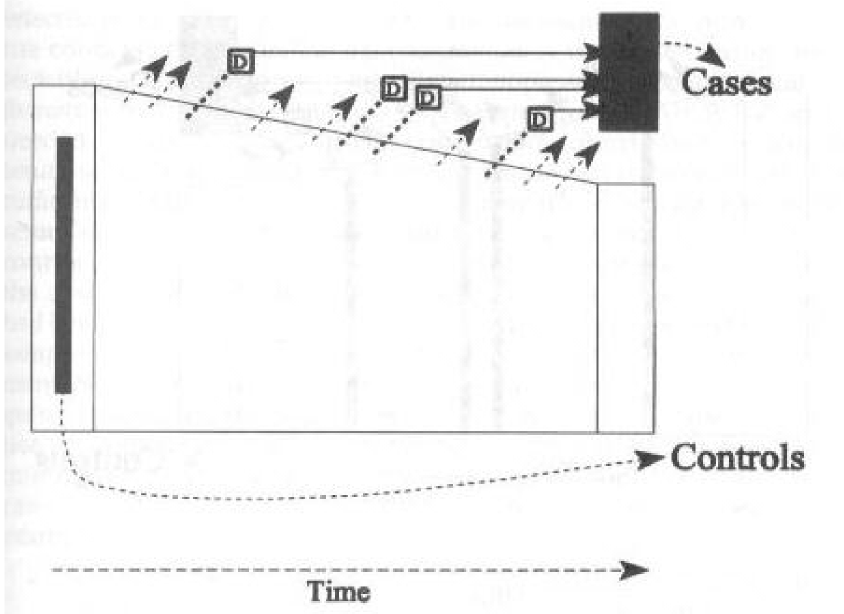
In this example, controls are sampled each time a case is diagnosed.

In incidence density sampling the selection of controls is governed by the diagnoses of cases. Every time a case is diagnosed one or more controls are selected from other members of the cohort who, **at that time**, do not have the diagnosis. The term incidence density comes from the fact that the time of follow-up and the incidence of new disease are involved in determining eligible controls.

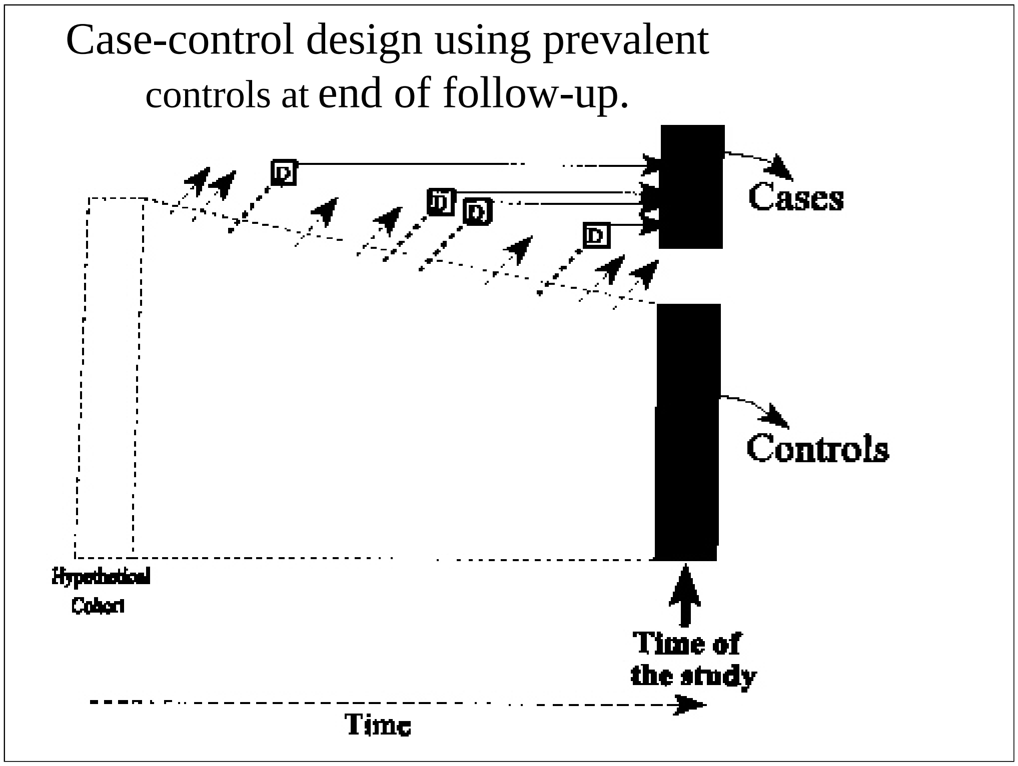
In our example of conserving resources by not testing all of the cohort members, the investigator would test stored biological samples only on those subjects chosen as controls. If the predictor variable were a questionnaire item everyone in the cohort had already answered, there wouldn't be any point in selecting controls as the data is already available on the entire cohort.

The text book (and a number of others) call this design a “nested case-control study,” but nested is a imprecise term. It seems that it should more properly refer to any case control studies selecting controls from within a cohort study. In other words, all three of the sampling methods we are describing can be viewed as “nested” within a cohort.

Case-cohort design: sample baseline of cohort

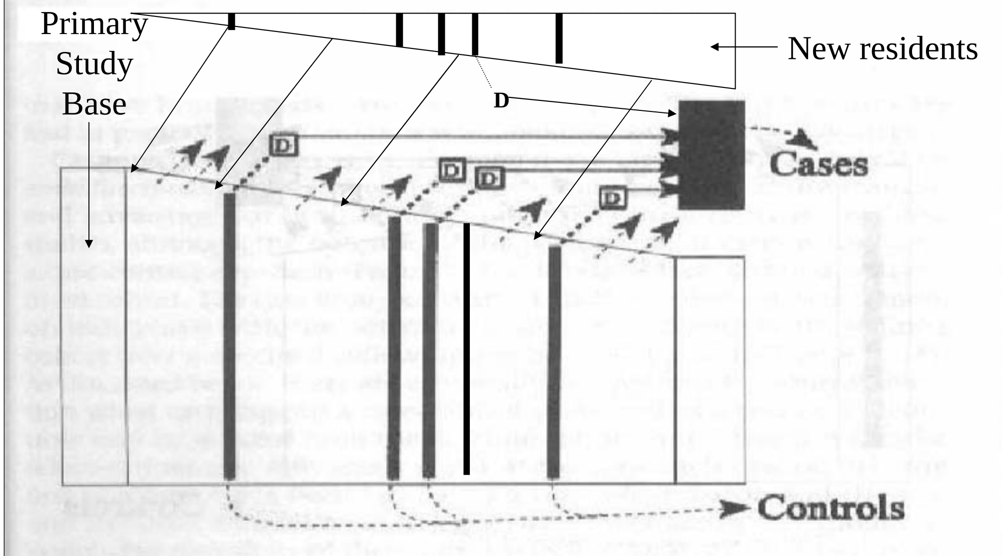


“Case-cohort” is type of design you may not be acquainted with as it is relatively new and still has not been used frequently. It was first described by the statistician Ross Prentice in the 1980’s. It seems odd at first to realize that you will likely be sampling future cases as well as controls when you take a random sample of a cohort at its baseline. This means that a subject may be included both as a case and a control. But this is also true of incidence density sampling since a subject selected as a control at one time point may later become a case. This troubles many new to these sampling designs and results in their thinking that the best design must be to wait until the end of follow-up to select controls so that the investigator can be sure they will not be cases. We will spend some time trying to demonstrate why this is not the right way to think about it. For starters, becoming a case is an artifact of the follow-up period of the cohort. The investigator cannot know whether many of the controls will be diagnosed with the study outcome the day after the study ends. This is made even clearer by the example of the cohort study that uses death as an outcome, as some do. Everyone is eventually a case. In summary, when we are looking for (i.e., sampling) controls, we do not necessarily have to guarantee that these are subjects who will never become cases. All that is needed is to be sure that they are not cases at the time of control sampling.



This is the design that most neophytes are drawn to, as discussed in the notes on the previous slide on case-cohort design. When we discuss the measures of association linked to each of these sampling designs in future lectures, we will show more formally why this is not a good design. For now, it will suffice to note that there is an obvious source of potential bias in waiting until the end of follow-up to select controls because factors that influence loss to follow-up will influence the selection of controls. If those factors are associated with both your predictor variable and the outcome, the measure of association will be biased.

Incidence Density Sampling in a Primary Study Base
(e.g., San Francisco County)



Incidence-density sampling in a specified population with new subjects entering

The same graphic could be shown for a case-control study in a cohort with open enrollment of new study subjects.