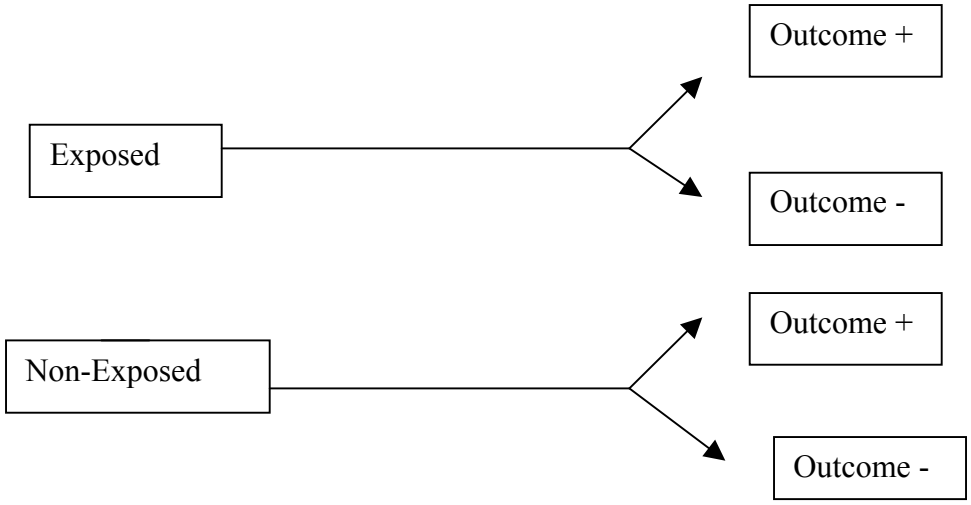
Cohort Study Design

The basic components of a Cohort Study design include:

1. Enrolling subjects at-risk for developing the outcome
2. Measuring exposure status on study participants,
3. Following subjects over time, and
4. Recording outcomes.

This describes the same general features of randomized controlled trial (RCT) in the last series of lectures. The RCT is a special case of a cohort study with the defining feature that the investigator assigns exposure status, usually with a randomization device, for the purpose of evaluating the effect of the exposure on an outcome. In the general Cohort Study design, exposure is typically not determined otherwise by the investigator and not for the primary purpose of examining the effect of the exposure on some outcome. The general features of a cohort study are described by the following figure.



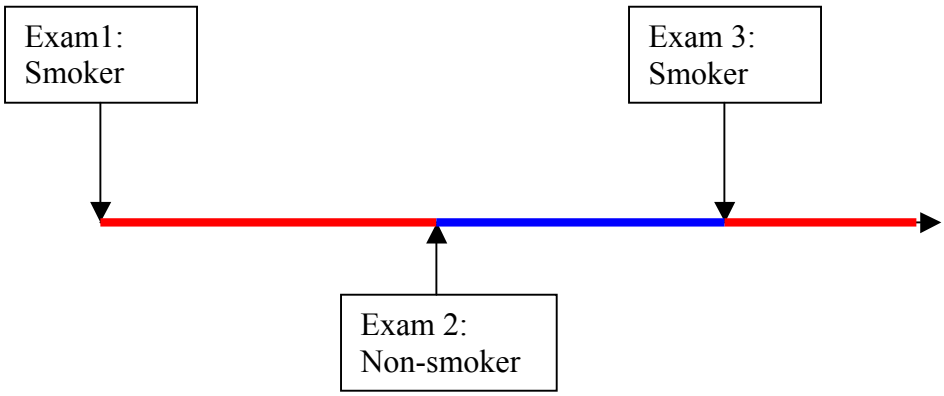
Suppose an investigator wished to study the association between smoking (exposure) and the incidence of a first CHD event. This would require enrolling subjects who are free of CHD and at risk for a first CHD event. There are numerous options for measuring the exposure of interest, smoking. Some are contained in the following list:

1. Smoking versus not smoking at baseline
2. Extent of smoking at baseline (non-smoker vs. light smoker vs. heavy smoker
3. Number of cigarettes smoked per day at baseline
4. Current smoker vs. never smoker vs. ex-smoker
5. Duration of smoking (# years smoked)
6. Pack-years of smoking (combination of duration and extent of smoking)
7. Any of the above measured at various points in time during the follow-up period

One option is to classify subjects according to their baseline smoking status (static measurement) as smokers or non-smokers. This classification does not take into account any changes in smoking status that occurs during a follow-up period, meaning that some of the subjects classified as smokers at baseline might later become non-smokers and some of the non-smokers may become smokers. Baseline smoking also does not reflect the duration and extent of past smoking. Some of the smokers at baseline might have recently started smoking, while others might have a long history of past smoking. Similarly some of the non-smokers at baseline may have recently quit smoking, while others may have never smoked. Finally, classifying subjects and smokers and non-smokers does not take into account the extent of smoking among the smokers. Some of the smokers may be light smokers, while other may be heavy smokers.

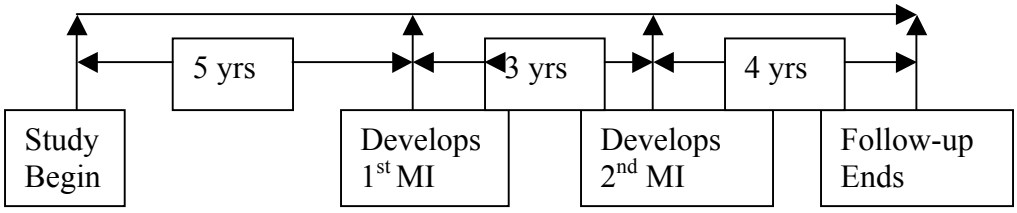
Having enrolled at-risk subjects and determined the appropriate method for measuring exposure, the investigator then needs to decide how to follow subjects over time and record outcomes. Follow-up is recorded in the Framingham Heart Study by having each subject return to an examination center every two years (biennial exams). In the Nurses Health Study follow-up questionnaires are sent by mail every two years. Alternatively, if the subjects are members of an insurance plan then follow-up information may be available through medical records or billing records. A variation of this is the SEER-Medicare data bases that captures information form cancer registries and from Medicare billing records.

If a person changes smoking status during a follow-up period then he/she can contribute person-years to the calculations of the CHD Incidence Rate for both the smoking and non-smoking group. In addition, if this person developed CHD during the follow-up period, then he/she will also contribute an outcome case to the calculation of one of these Incidence rates. For example, the following figure describes the person-years of follow-up for an individual in the FHS teaching data set who was a smoker at the first exam, a non-smoker at the second exam, and returned to be a smoker at the third exam. This person contributes person-time to both the smoking (red lines) and non-smoking group (blue lines).



Finally the investigator needs to determine which outcomes to record during the follow-up period. For non-fatal events, one question to address to how to handle repeated events. For example, a subject may develop and survive multiple myocardial infarctions (MI, heart attacks) during the follow-up period. A common practice is to focus on only the first of such multiple events and stop the follow-up period at the time of the first myocardial infarction. Once a person develops a myocardial infarction he/she is no longer at risk for developing a first myocardial infarction.

Alternatively, multiple events could be taken into account by dividing each subject's follow-up time into periods when he/she is at risk for different events. For example, the following figure describes the follow-up experience for a subject, who developed a first MI after 5 years of follow-up, second MI 3 later, and is followed for another 4 years until the study ends. This person would contribute 5 person-years of follow-up to the denominator and 1 case to the numerator for the calculation of the incidence rate for developing a first MI. He would contribute 3 person-years of follow-up to the denominator and 1 case to the numerator for the calculation the incidence rate for developing a second MI. He would also contribute 4 person-years of follow-up and nothing to the numerator for the calculation of the incidence rate of developing a third MI.



The incidence of developing an outcome in a cohort study potentially can be measured in two ways:

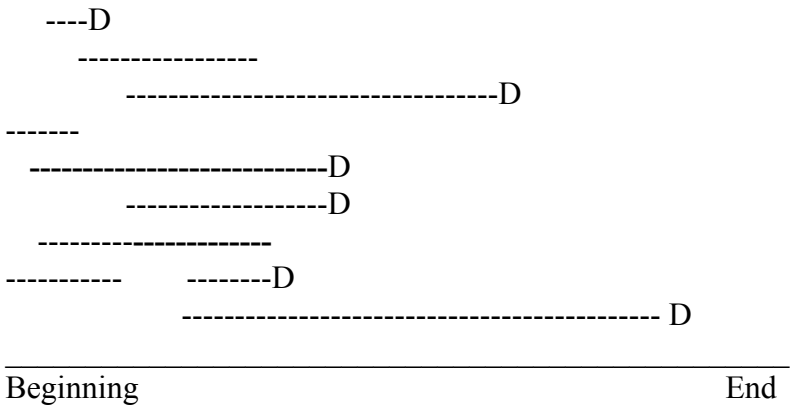
1. Cumulative Incidence
2. Incidence Rate

The Cumulative Incidence requires that all subjects be followed for a fixed period of time and that there be no losses-to-follow-up or losses due to no longer being at risk during the follow-up period because of a competing event (for example dying from another cause). The Incidence Rate requires that the reason for terminating follow-up not be related to the risk of developing the outcome.

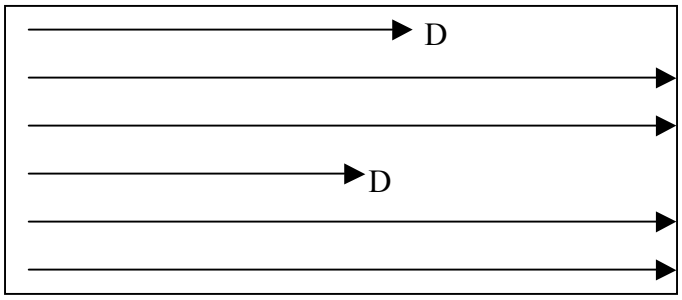
Open vs. Closed Cohort

An Open Cohort is a dynamic population with migration into and out of the cohort occurring during the follow-up period. Exposure status may change over time so that the same subjects can contribute person-time to different exposure groups. The

outcome measure in an open cohort is the Incidence Rate. The following figure shows the general features of an open cohort study.



A Closed Cohort has a common starting point and fixed potential period of follow-up for all subjects. For example, the Framingham Cohort Study started in 1948. It enrolled 5,209 subjects in 1948 with the plan to follow all subjects for 20 years. Exposure is defined at the start of the follow-up and does not change over time. There are no losses-to-follow-up. The outcome measure for such cohorts is either the Cumulative Incidence or the Incidence Rate. The following figure describes the general features of a Closed Cohort.

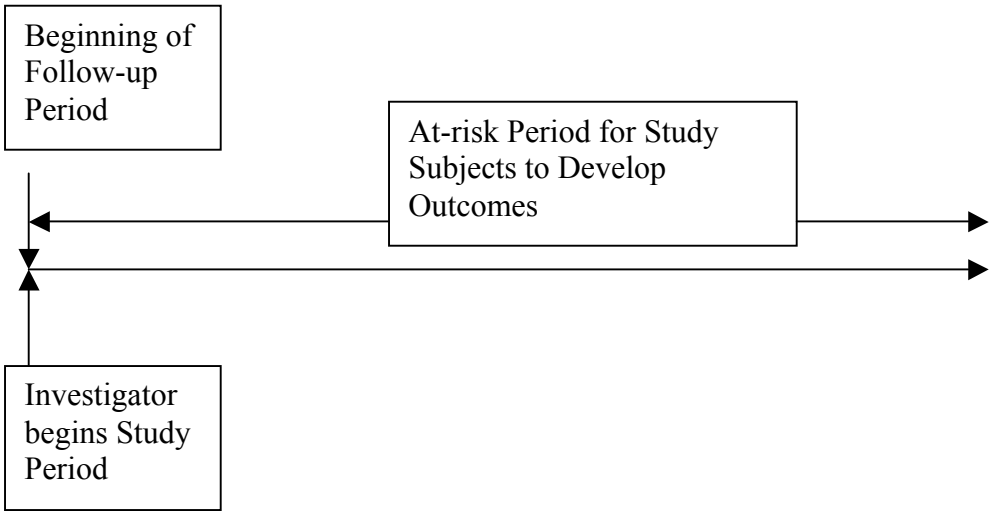


A Fixed Cohort is similar to a closed cohort with the exception that there are some subjects who are lost-to-follow-up. The Framingham Heart Study with the outcome of all-cause mortality and sex as an exposure is an example of a Closed Cohort as there is complete ascertainment of this outcome on all subjects. On the other hand, if the outcome is the development of CHD then it should be considered as a Fixed Cohort as some subjects are lost-to-follow-up and their subsequent development of CHD is unknown. Also, there are other subjects who die from non-CHD causes (competing risk) and therefore are no longer at risk for developing CHD.

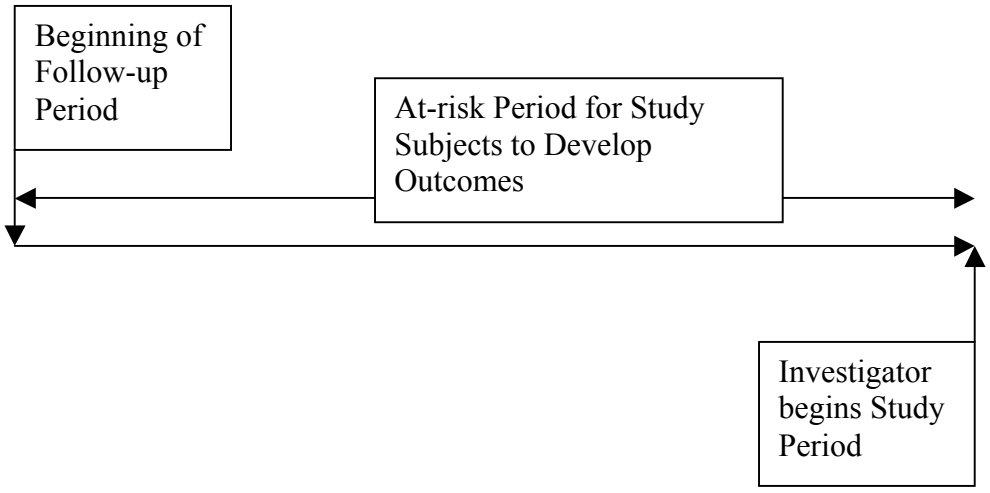
Prospective versus Retrospective Cohort Studies

The distinction between a Prospective and Retrospective Cohort Study concerns the two time periods involved in a cohort study: the time period spent by the investigator to perform the study and the time period spent by the study subjects when they were followed and at-risk for developing the outcome. The **Study Period** is the time period spent by the investigator to perform the study. The **Follow-up Period** is the time period in which subjects at risk and followed to ascertain outcome of interest.

For example, suppose an investigator received grant funding and begins working on a study on 1/1/12, the start of the Study Period. If the investigator enrolls at risk subjects on that date and follows them for the next 5 years to record outcomes, then the Follow-up Period also begins on that date. This is an example of a **Prospective Cohort Study.** Under this design all outcome cases occur **after** the beginning of the Study Period. This design is depicted by the following figure:



Suppose that, at the start of the Follow-up Period (1/1/12), the investigator began reviewing medical records to establish an at-risk cohort that existed 10 years ago and then followed that cohort through the medical records to see which of these subjects had developed outcome in the past 10 years. This is an example of a **Retrospective Cohort Study**. Under this design all outcome cases occur **before** the beginning of the Study Period. This design is depicted by the following figure:



A Retrospective Cohort Study uses data to create a historical cohort of at-risk subjects that existed in the past. It requires data collected in the past by others for a reason other the research goal of the investigator. Candidate data sets for creating historical cohorts are medical records and billing records for insurance plans.

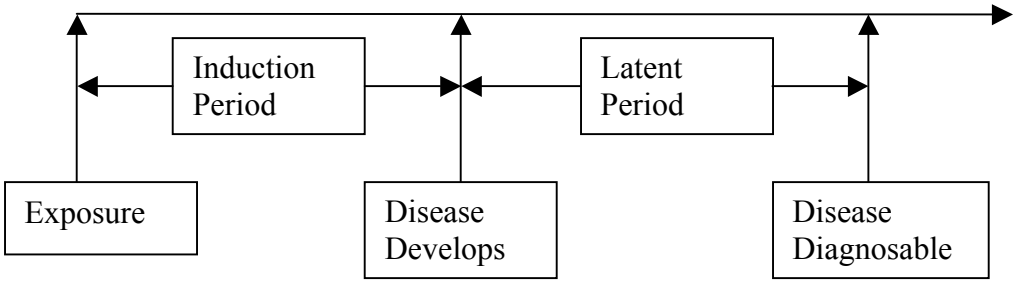
Given the growth of electronic medical records, the potential for creating historical cohorts and performing Retrospective Cohort Studies has increased. However, a limitation of a Retrospective Cohort Study is the quantity and quality of the existing data. On the other hand, the quality and quantity of the data in a Prospective Cohort Study may be superior because it is under the control of the investigator.

Retrospective Cohort Studies tend to be less expensive, in time and money, than Prospective Cohort Studies. In a Retrospective Cohort Study the at-risk follow-up period for the study subjects has already happened. Their outcomes have also happened. On the other hand, In a Prospective Cohort Study the investigator must wait, perhaps many years, for the at-risk follow-up period to end and for outcomes to develop. The Study Period is usually much longer for a Prospective Cohort Study than for a Retrospective Cohort Study. Also, in a Prospective Cohort Study the investigator must develop a system to monitor subjects during the at-risk follow-up period and record outcomes. This might involve having subjects come to an examination center on a regular schedule (Framingham Heart Study) or sending questionnaires in the mail on a regular schedule (Nurses Health Study). On the other hand, the data already exists for a Retrospective Cohort Study, so the financial cost is much less than for a Prospective Cohort Study.

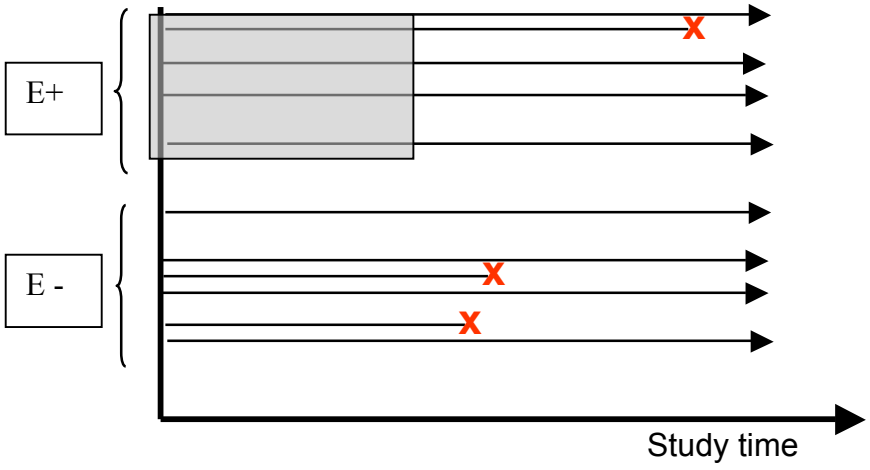
Induction Period

The **Induction Period** is the time between the exposure to a risk factor and development of disease. For example, a woman might be exposed to radiation on a specific data and 3years later might develop the first evidence of leukemia. However, it may take additional time for the disease to reach a state where it can be diagnosed by

current technology. The **Latent Period** is time period between disease development and the ability to detect it. Often these periods are combined and referred to as the **Empirical Latent Period**. These time periods are depicted in the following figure:



Suppose that Induction Period for a particular exposure to cause a disease is 3 years and the Latent Period to be able to detect this disease is another 2 years. Any disease outcome that occurs within 5 years of the exposure could not be caused by the exposure. This implies that person-time and outcomes that are observed during the Induction Period and Latent Period among exposed subjects should not contribute to the calculation of the Incidence Rate for the exposed group. The shaded area in the following graph (provided by Heather Baer – EPI208, 2012) pertains to person-time for the exposed group during the Induction and Latent Periods. This experience should not be included in the calculation of the Incidence Rate for the exposed group. It can either be eliminated from the analysis or folded into the calculation of the Incidence Rate for the non-exposed group.



In practice the investigator will probably not know the length of the Induction and Latent Periods. However a secondary analysis might be useful whereby the investigator

examines the robustness of the main conclusions from the primary analysis by accounting for reasonable estimates for the length of these periods.

Bias

Measurement Bias can occur in a Cohort Study, as in any study. One example is if the detection of the outcome is performed differently for exposed and non-exposed groups, resulting in a potential for a non-differential misclassification (**detection bias**). For example, suppose an investigator wished to examine the relationship between Oral Contraceptive use and the incidence of Breast Cancer. If women or their physicians suspected a causal link between these factors, then women taking oral contraceptive might see their physicians more often and be tested more often for breast cancer, compared to women not taking oral contraceptives. More frequent testing may result in some cases of breast cancer being detected among oral contraceptives users that may not be detected among the women not taking oral contraceptives with less frequent testing. One method for avoiding this bias is through uniform testing procedures for both groups.

Another example of a measurement bias in a cohort study might occur when knowledge of the exposure status for an individual may influence (consciously or unconsciously) the classification of outcome status by an interviewer. One method for avoiding this bias is by blinding the interviewer to the exposure status of the study subjects or, if this is not possible, blinding the interviewer to the study hypothesis.

Selection bias can also occur in a cohort study due to losses-to-follow-up. This may occur if the reason for the loss if related to the risk of the developing the outcome and also to the exposure. For example, if smokers who develop early signs of developing CHD fail to attend a scheduled follow-up visit, then the incidence rate that is based on only those smokers who attend the visit will underestimate the true incidence rate for all smokers. If this problem does not occur among non-smokers then the resulting Rate Ratio will be biased.

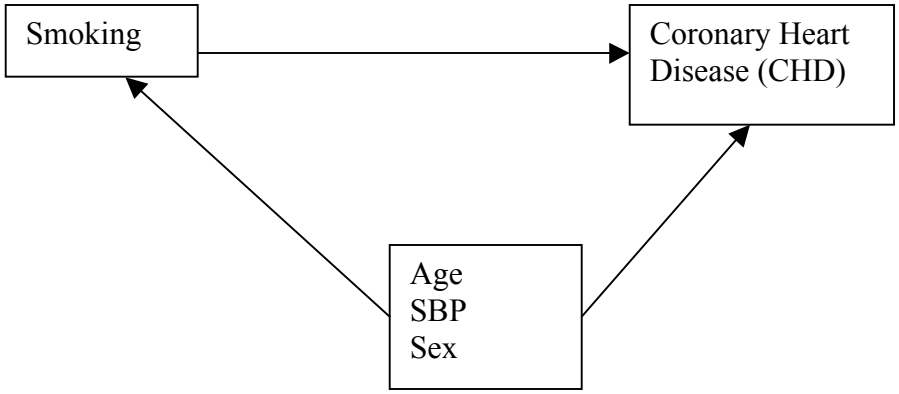
Confounding

Confounding can occur in a Cohort Study if risk factors (determinants) of the disease are related to the exposure of interest. For example, the following table describes the age, blood pressure, and sex distribution among current smokers and non-smokers in the FHS teaching data set.

	Smokers	Non-Smokers
Mean Age	48.1	51.7
Mean SBP	129.8	135.9
% Male	53.9	34.1

Smokers are younger than non-smokers but also have higher average blood pressure and a great percentage of males than non-smokers. These imbalances provide alternative

explanations for any difference in CHD outcomes that might be observed when comparing smokers to non-smokers. The potential for these alternative explanations is shown in the following causal graph. Confounding and causal graphs will be discussed in future lecture notes.



Framingham Heart Study

The Framingham Heart Study was initiated in 1948 by enrolling 5,209 study subjects, age 30-60 and following them for the primary outcome of Cardiovascular Disease (CVD). Detailed information about this study can be found at

- http://www.framinghamheartstudy.org/
- http://www.cbsnews.com/video/watch/?id=3365580n

At that time, the town of Framingham had both rural and urban aspects. The people of Framingham were also involved in a previous study on tuberculosis. Study subjects were asked to return to an examination center every two years (biennial examinations), where risk factors were updated, tests were performed, and outcomes that happened since the last exam were recorded. Initially 84 medical tests were performed and more were added over time. Medical care was not performed but test results were sent to a subject's physician. Mortality was also assessed through search of a National Death Index, providing complete follow-up for that outcome.

An Offspring Cohort was established in 1971, enrolling the children of the original cohort and their spouses. A third generation cohort was established in 2001 using the grandchildren of the original cohort and their spouses.

Strengths and Limitations of Cohort Studies

Cohort Studies can examine the effects of single exposure on multiple outcomes. Unlike Cross Sectional Studies, Cohort Studies can elucidate temporal relationship

between exposure and disease. They allow direct measurement of incidence of disease in exposed and unexposed groups, as well as calculating various measures of association. Carefully planned and implemented Prospective Cohort Studies may reduce the potential for measurement and selection bias.

Other the other hand, Cohort Studies may not be efficient for study rare diseases because of the need to enroll large number of subjects and follow them for long periods to time to record enough cases of the disease. Prospective Cohort Studies can be expensive in terms of time and money. Retrospective Cohort Studies are more efficient but require the existence of previously collected data of adequate quantity and quality. Biases due to losses-to-follow-up are a potential problem to both Prospective and Retrospective Cohort Studies.