Definition of Epidemiology

A commonly cited definition for Epidemiology is "*the study of the distribution and determinants of disease frequency in man*" (MacMahon and Pugh. Epidemiology: Principles & Methods. Second Edition. Little, Brown and Company; Boston. 1970).This definition implies two quantitative aspect of epidemiology:

1. Measuring **disease distribution** in regards to person, place, and time (**descriptive epidemiology**), and

2. Measuring the association between a **disease and it determinants** (**analytic epidemiology**).

The first aspect of this definition refers to measuring disease (outcomes). Common examples include measuring the existence of disease in a population (**prevalence**) and the occurrence of disease in a population when it is followed over time (**incidence**). The second aspect of this definition refers to measuring associations between determinants (risk factors) and disease. The primary concern of analytic epidemiology is etiology by identifying the causes (risk factors) of a disease and quantifying the magnitude of their effects on the disease. Such a process begins with a **measure of association** between a risk factor and a disease, followed by an argument that this measure reflects the causal effect of that risk factor on the disease. In practice, there are often many explanations for an observed association between a risk factor and a disease. Arguing that an observed measure of association reflects a causal effect involves eliminating other possible explanations for the observed association. Hennekens and Buring (Epidemiology in Medicine. Little, Brown and Co; Boston: 1987) state that these alternative explanations fall under the broad headings of

1. **Confounding**
2. **Bias**
3. **Chance**

Clinical Epidemiology often focuses on the (causal) outcomes of an intervention (e.g. treatment). It also involves the prediction of disease (diagnosis and prognosis). The Final Report of the HSPH Department of Epidemiology Committee on the Status of Clinical Epidemiology (Singer at al, Fall, 2009 – personal communication) defined **Clinical Epidemiology** as follows:

> *Clinical epidemiology applies the concepts and techniques of epidemiology, statistics, and decision analysis to clinical problems. Clinical epidemiology emphasizes the study of patients, physicians, and systems of health care.* ***The focus is on diagnosis, prognosis, and treatment of disease but it also studies the etiology of disease.*** *Exposures and outcomes are informed by clinical and biologic knowledge as well as broader environmental and societal determinants.* ***Distinctive areas of clinical epidemiology are the development of diagnostic and prognostic models.*** *The gold standard for the study of medical treatments is the*

randomized clinical trial. Observational analyses of treatments are necessary to gain estimates of their impact in real world clinical practice and to screen for rare effects. These observational analyses of medical interventions face the distinctive hurdle of confounding by indication and/or contraindication

Historical Development of Epidemiology

The methods used in epidemiology research have evolved over time and newer methods are being developed. We owe much to key people in our past, who developed and refined these methods. Much of the following was taken from the textbook by Aschengreau and Seage *(Essentials of Epidemiology in Public Health, Second Edition. Jones and Bartlett Publishers. Sudbury 2008)* but can be found in other texts on epidemiology.

The oldest record of an epidemiology study that I could find is from the Bible and the **Book of Daniel** (1:8 -16). Daniel was one of those young Jewish men who were taken off to Babylon during the Babylonian Captivity, about 600 BC. The plan was to train Daniel and some of his companions for the king's service. As part of this training program, Daniel and his colleagues were offered the same food and wine that was served at the King's table. For religious reasons, Daniel and some of his colleagues resisted eating that particular diet, and requested a more traditional diet. The steward for the chief chamberlain resisted this request for fear of punishment by the king if Daniel and his colleagues looked "wretched by comparison with other young men of your age". In response Daniel requested a trial:

"*Please test your servants for 10 days. Give us vegetables to eat and water to drink. Then see we how we look in comparison with the other young men who eat from the royal table*."

By no means does this example imply that the Bible should be read as a text in epidemiology, but this passage contains many of the elements of a classical epidemiology study.

First, a clear research question is specified: does a diet on vegetables and water lead to better outcomes than a traditional diet. The exposure of interest is the diet of the young men put into the king's service. The exposure category of interest is the diet of vegetables and water and the non-exposure category is a standard diet of food and wine from the king's table. The outcome of interest is the appearance of the study participants after 10 days on either of these two diets. The study is limited to young men in the king's service. This limitation increases the baseline comparability of the two groups under study (but also limits the generalizability of the results). No mention is given on how the two groups might differ in other baseline characteristics that might influence the outcome (potential confounders). In addition, the outcome appears is a subjective evaluation of the overall appearance of all study subjects by a single person (steward of the chief chamberlain). No mention is given about the criteria used for this assessment or if the

steward is “blinded” to the diet of a study participant when making an outcome assessment. It is possible that the knowledge of the diet of a participant might influence the outcome classification, introducing a potential for a measurement bias in this study. The topics of confounding and bias will be discussed in future lectures of this course.

Despite the short duration of follow-up, the results from this study were striking:

“*after 10 days, they looked healthier and better-fed than any of the young men who ate from the royal table*”

About 200 years later, the teachings of the Greek physician **Hippocrates**, the father of modern medicine, (http://en.wikipedia.org/wiki/Hippocrates) were recorded and included the following quotation:

> *"Whoever wishes to investigate medicine properly should process thus-- in the first place, consider the seasons of the year, and the effects of each of them. Then the winds, the hot and the cold. One should consider most attentively the waters in which the inhabitants use, and the mode in which the inhabitants live, and what are their pursuits, whether they are fond of drinking, and eating to excess, and given to indolence, and are fond of exercise and labor." (Hippocrates. On Air, Water, and Places. Translated and republished in Medical Classics 3:19-42, 1938)*

In this passage Hippocrates points out that there are characteristics of people and their environment that influence their risks of developing disease. Disease development is not a totally random event and the goal of epidemiology is to identify those risk factors that influence the risk of disease and quantify their effects.

Roughly 2000 years, the British investigator **John Graunt** (http://en.wikipedia.org/wiki/John_Graunt), published the "Natural and Political Observations Mentioned in a Following Index, and Made Upon the Bills of Mortality". The bills of mortality were routinely collected and reported mortality data, describing who was dying, and from what cause. John Graunt used those weekly counts of death to look at associations between characteristics of people and their causes of death. He reported yearly and seasonal mortality trends. He reported the common and the uncommon causes of disease. He reported other findings, including that men were at higher risk of dying than women. An important feature of Graunt’s work was the use of routinely collected data to identify associations between characteristics of individuals and mortality.

About 100 years later, a Scottish physician, **James Lind**, performed one of the first clinical experiments (clinical trial). He performed this experiment to identify the appropriate treatment for the disease scurvy. We know today that scurvy is caused by a deficiency of vitamin C in a person's diet, but in the 1700s, when Lind was treating such people on British ships, he decided to do a study to test possible treatments for scurvy. He enrolled 12 people who had been diagnosed and were sickened with this disease. He

divided them into six treatment groups, each containing two people. All 12 participants received a common diet, but each group received something in addition. One group received a quart of cider. One group received the elixir vitriol (sulfuric acid). One group received vinegar. One group received seawater. One group received a spicy paste made of mustard and garlic, along with some barley water. The last group received two oranges and one lemon. Lind followed all groups and reported that recovery was best for the group that was given the two oranges and the lemon.

A key point of Lind's study is that it is an experimental study. Unlike Daniel and his colleagues who self-selected to take the diet of vegetables and water, the participants in Lind's study were assigned by Lind to the treatment groups. In an experiment, the investigator decides which treatment a person receives, not the individual themselves. This often raises ethical issues. In addition, unlike Lind's trial, assignment in experimental studies is often determined in a random fashion. Randomization, ethics and other features of experimental studies will be discussed in future lectures.

Two major figures in the development of the epidemiology methods, **James Farr and John Snow**, lived in the next century and identified the cause of cholera, a major health problem at that time. Information about these investigators can be found at

http://www.ph.ucla.edu/epi/snow.html

http://en.wikipedia.org/wiki/William_Farr

William Farr was essentially the chief statistician for Great Britain. He was the statistical superintendent of the general register office of England and Wales in the 1800s. He was responsible for collecting and routinely distributing information about the health of the population, for example reporting causes of mortality for different regions of the country. During a cholera outbreak in London he published weekly reports, describing the number of deaths from cholera by such factors as age, sex, air temperature, wind, rainfall, day of the week, elevation, crowding, and property value.

The reported relationship between elevation and mortality is particularly noteworthy. At the time, when Farr was making these reports, the common belief was that the primary cause of cholera was essentially polluted air, miasma. The mechanism for contracting cholera was by breathing harmful particles into your body. Farr believed in the miasma theory. He supported this belief by showing people who lived at low elevations where the air might be denser, had higher mortality rates compared to people who lived at higher elevations, where the air was less dense.

At the same time Farr was reporting these results, John Snow, an anesthesiologist, was developing evidence to support an alternative hypothesis about the cause of cholera. As an anesthesiologist, Snow had knowledge about gases. He didn't believe that it was the bad air that was causing people to contract and die cholera. As a physician, he cared for people who contracted cholera. He noticed their signs and symptoms, in particular their complaints of belly pains. He believed that the cause of disease was not breathing

polluted air, but ingesting polluted water. Snow was able to use prior knowledge and observations based on data to argue his hypothesis. Using data from Farr's reports, he was able to perform two important landmark studies, to convince people, including Farr, to disprove the miasma theory and successfully argue for polluted water as the cause of cholera.

One of the studies was "The Great Experiment," was initially based on data published by Farr showing the mortality rates of cholera for areas of London served by different water companies. Some of these water companies took their water from the polluted portions of the Thames River. Others took it from a less polluted area upstream from London. It was Snow who noticed that people who were receiving water from the less polluted sources had a lower mortality rate from cholera than those who were receiving their water from the polluted areas of the Thames River

In another study Snow investigated an outbreak of cholera in a particular neighborhood of London, the Soho District, during 1854. Reading the weekly results that were reported by Farr, Snow noticed that this region of London suddenly had a huge outbreak of cholera, whereas the mortality from cholera during prior weeks was very low. Snow gathered additional data on the inhabitants of this region (primary data collection), and noticed a clustering of deaths near a public water pump in that area. He was able to build a convincing argument that the likely source of cholera in that region was polluted water from the well served by that pump.

Snow and Farr show the combination that's needed for epidemiology and clinical epidemiology research. On one hand, we need hypotheses suggesting potential risk factors for causing disease. Often these hypotheses are based on routine reports, other investigations, or personal experience. Snow’s background as a physician and an anesthesiologist provided the background for his hypothesis. Snow believed that polluted water was the potential culprit for cholera, but he needed data, partially furnished by Farr, to support hypothesis.

A more recent contributor to epidemiology methods is **Austin Bradford Hill**. In the 1940s, Hill was the statistician on a clinical trial investigating a new treatment for tuberculosis, streptomycin. Hill’s study enrolled homogeneous cases of tuberculosis. One group of patients received the standard treatment of bed rest. Another group received bed rest plus a new potential treatment, streptomycin. Hill used a random device to assign patients to these two groups so that roughly half the people got streptomycin, and half did not. He had only a limited supply of streptomycin available, enough for approximately half the participants in the study, so he could argue that random allocation to patients is an ethical way to decide who should get a new treatment. Hill also introduced the notion of blinding when assessing outcomes. The outcomes were determined by looking at x-ray films, and readers of these ray films, the investigators interpreting these outcomes, were blinded on whether or not a study participant received streptomycin or didn't receive streptomycin. Therefore, the knowledge of the assigned treatment could not influence the outcome decision. We'll discuss the role of randomization, blinding, and ethics in experimental studies in a future lecture.

A colleague of Bradford Hill, **Richard Doll**, did a series of landmark epidemiology studies in the 1950's. Doll examined the association between smoking and lung cancer in two important studies. The first was a very large **case control study**, where Doll enrolled compared past smoking habits for people who developed lung cancer (cases) and for people who didn't develop lung cancer (controls). The second study was a **prospective cohort study**, where he enrolled people who were free of lung cancer, smokers and non-smokers, follow them forwards in time, to see which group developed lung cancer more rapidly, more commonly. These study designs will be discussed in detail in later lectures in this course.

A final example of an important contribution to the development of epidemiology methods is not a person but a study, the **Framingham Heart Study**. This study will be briefly discussed in these series of lectures and discussed in more detail later in this course.

The Framingham Heart Study

The Framingham Heart Study started in 1948. Framingham, Massachusetts is located about 10 miles west of Boston. In 1948 it had both urban and rural aspects. It had been involved in previous tuberculosis study, so the people might be willing to be interviewed, and be involved in another research study to identify the risk factors for developing cardiovascular disease (CVD). The phase "**risk factor**" was coined by the investigator in the Framingham Heart Study. The study enrolled 5,209 men and women, aged 30 to 62. The selected age range of the participants reflects the "dark side" of epidemiology. Epidemiologists typically observe disease occurring among people and identify the risk factors that cause these outcomes. This requires enrolling enough participants into a study to observe a sufficient number of disease cases for finding associations between risk factors and disease. The plan for the Framingham heart study was to enroll people in 1948, follow them for 20 years, and observe who developed cardiovascular disease (CVD). To observe a sufficient number of cases of CVD during follow-up may require a large number of low-risk study subjects, or a smaller number of higher risk subjects. Subjects under 20 years of age are expected to be at low risk for developing CVD and therefore may provide few cases of disease if followed for 20 years. Somewhat ironically, in 1948 many older people were considered to already have early signs of CVD. Following only elderly subjects might show few of them <u>not</u> developing CVD, making it difficult to find associations between risk factors and CVD among such subjects.

The plan was to follow them for 20 years and record outcomes. Outcome and risk factor information were recorded during the follow-up by a series of biennial examination. Every two years, participants are asked to return to a testing center, where they are examined. Their risk factors are updated and their outcomes recorded. This method of follow-up provides the opportunity to collect detailed information on study subjects, but comes at an expensive price tag.

The Framingham Heart Study was initially planned to last for 20 years, but it is still going on today. Survivors from that original cohort of 5,209 people still retrun every two years to be examined. Unfortunately, many members of the original cohort have passed on, and the number of survivors is dwindling each year. There have been several "spin-off" studies of Framingham Heart Study. A second cohort (offspring cohort) of 5124 participants was developed in 1972 using children of the original cohort and their spouses. In 2002 a third cohort of 4095 participants was developed, using the grandchildren of the original cohort and their spouses. (Oppenheimer GM. Becoming the Framingham Heart Study: 1947-1950. *Am J Public Health* 2005;95:602-610.).

Much more information about the Framingham Heart Study can be found at:

Oppenheimer GM. Becoming the Framingham Heart Study: 1947-1950. *Am J Public Health* 2005;95:602-610..

http://www.framinghamheartstudy.org/

There is also an excellent video that was created by CBS news:

http://www.cbsnews.com/8301-3445_162-3358673.html

The investigators from the Framingham Heart Study identified the roles of smoking, blood pressure and total cholesterol, as risk factors for increasing your risk of CVD. They showed that physical activity lowers your risk. They showed the relationships between the components of total cholesterol, HDL and LDL, and the risk of heart disease. They also develop prediction rules that are available at their website to predict your risk for developing various CVD outcomes. We'll be talking about the Framingham Risk Model in a later lecture.

The Framingham Heart Study is a classic example (if not the classic example of a **cohort study**. This type of study design will be discussed in more detail in future lectures.

This course will also use a teaching data set that is available from the National Heart, Lung, and Blood Institute. The data set is derived from the actual Framingham Heart Study data, but it has been perturbed to protect the identity of the original cohort. The teaching data set contains 4,434 participants from the original 5,209 participants. The data pertains three examination cycles from the original Framingham cohort. The first exam cycle refers to the 1956 exam, and two follow-up exams, roughly six years apart are also provided. In addition, all subjects were followed for 24 years and multiple outcomes were recorded.

The teaching data set contains data on the age of each person at that exam, their sex, whether they were a smoker or not, and how much they smoked (in terms of cigarettes per day), their blood pressure (both on the systolic and the diastolic scale),

whether they were taking medications to treat hypertension at that time, their total cholesterol levels (and their HDL and LDL levels), whether they had diabetes , their glucose levels, their heart rate, and whether they had previously been diagnosed with coronary heart disease, stroke, or hypertension. Outcomes include death during those 24 years, evidence of developing coronary heart disease, stroke, or hypertension, and also the number of years from the 1956 exam until they developed these outcomes.

Role of Measurement in Epidemiology

As previously stated, this definition implies two quantitative aspect of epidemiology:

1. Measuring disease distribution in regards to person, place, and time (**descriptive epidemiology**), and

2. Measuring the association between a disease and it determinants (**analytic epidemiology**).

The first quantitative aspect of this definition reflects the domain of **descriptive epidemiology** and requires the specification of appropriate **outcome measures**. Proportions, rates, percentiles and means are examples of outcome measures. The second quantitative aspect of this definition reflects the domain of **analytic epidemiology** and requires the specification of a **measure of association** to compare the values for the outcome measures in different subgroups that are defined by the exposure. Usually these measures involve taking ratios or differences of the values for the outcome measures in the different subgroups. Odds Ratios, Risk Ratios and Mean Differences are examples of commonly used measures of association. For a valid epidemiologic study, these measures of association are used as estimates of the **causal effect of a risk factor on the outcome**.

The choice of an outcome measure depends on the properties of the numerical values that are assigned to the outcome. One common categorization of measurements (exposures and outcomes) is:

Possible Types of Measurements.

1. **Nominal**
2. **Ordinal**
3. **Interval**
4. **Ratio**

Nominal measurements are categorical with no intrinsic ordering. Examples of nominal measurements include integer values assigned to categories of race, religion, and marital status. The numerical values that are assigned to categories of a nominal variable are arbitrary labels and do not reflect the relative locations of the corresponding categories on any scale.

The simplest example of a nominal variable is the **binary indicator variable** (**dummy variable**) that uses integer values to indicate membership in one of two categories of a factor of interest. For example, the following is an example of an indicator variable (Male) representing the sex of a subject.

Male = 1 if a subject is a male
= 0 if a subject is a female

Ordinal variables have categorical responses with a well-defined order among categories. The numerical values assigned to the categories of an ordinal variable reflect the relative positions for these categories but may not necessarily reflect the distance between these categories on an underlying continuous scale. For example, the New York Heart Association Functional Classification Scale assigns patients to one of four categories of cardiac disability as described by the following table (Goldman L, et al. Comparative reproducibility and validity of systems for assessing cardiovascular function class: Advantages of a new specific activity scale. *Circulation* 1981;64:1227-1234:

New York Heart Association Functional Classification Scale.

Class	Description
I	Patients with cardiac disease but without resulting limitations of physical activity. Ordinary physical activity does not cause undue fatigue, palpitation, dyspnea, or anginal pain.
II	Patients with cardiac disease resulting in slight limitation of physical activity. They are comfortable at rest. Ordinary physical activity results in fatigue, palpitation, dyspnea, or anginal pain.
III	Patients with cardiac disease resulting in marked limitations of physical activity. They are comfortable at rest. Less than ordinary physical activity causes fatigue, palpitation, dyspnea, or anginal pain.
IV	Patients with cardiac disease resulting in inability to carry on any physical activity without discomfort. Symptoms of cardiac insufficiency or of the anginal syndrome may be present at rest. If any physical activity is undertaken, discomfort is increased.

A subject who is classified as Class II is more disabled than a subject classified as Class I. However, this classification scheme does not assume that subjects in Class II are half-way on a disability scale between Class I and Class III subjects, or are twice as disabled as subjects in Class I, or are they half as disabled as subjects in Class IV. For this reason, averaging scores from an ordinal outcome across subjects may not be appropriate.

Values for **interval variables** reflect not only the order among categories/levels of the variable, but also reflect the distances between these categories/levels on a specified (but sometimes arbitrary) scale. Therefore calculations based on addition and subtraction are appropriate. However, unless the variable possesses a natural anchoring point (zero value), mean scores must be interpreted relative to the assigned range of values.

For example, different temperature scales possess different anchoring points. On a Celsius scale, a value of 0 represents the temperature at which water freezes on this scale. The same temperature translates to 32 degrees on a Fahrenheit scale and the value of 0 on the Fahrenheit scale represents a sub-freezing temperature. Therefore a ratio of the temperature on two days will be different on these scales. Doubling the temperature on a Celsius scale does not correspond to a doubling of the corresponding temperatures on a Fahrenheit scale.

Since many health status instruments used in clinical research are transformed to a 0 - 100 point scale, an average score of 50 represents a point that is half way between the two extreme scores. If the same scale is put on a 100 - 200 point scale, then the score of 50 on the original scale becomes a value of 150 on the new scale. This value is still half way between the two extremes but no longer is equal to 50% of the upper extreme.

Values for **ratio variables** not only reflect order and distances between categories/levels, but these variables also contain a natural zero value, representing the absence of the quantity that is being measured. This provides a natural anchoring point for the interpretation of values on this scale. Examples of ratio scales include temperature measured on the Kelvin Scale (where 0 = lack of molecular movement and the absence of temperature) and the age of individuals. Not only is a 50 year-old half way between the age of a 40 year-old and a 60 year-old, but also he/she is also twice as old as a 25-year old.

A **continuous variable** may be interval or ratio in scale, depending on the existence of a natural zero value.

The values for the categories of an ordinal variable are often represented by consecutive integers. These values reflect the order of the categories but may not reflect perceived distances between these response options. For example, a commonly used simple measure of health status is the following question:

How would you describe your health status?

Excellent	5
Very Good	4
Good	3
Fair	2
Poor	1

The consecutive integer values may not necessarily reflect distances between these health state categories, since the distance between fair and good health states may be perceived by many to be greater than the distance between very good and excellent health states. For example following values reflect the median locations of these 5 health states as reported by a group of students at HSPH taking EPI241 when asked to place the categories on a line with poor health assigned the value 1 (left end of the line) and excellent health assigned a value of 5 (right end of the line):

Median location of health states by HSPH students on 5-point scale

5.	Excellent	→	5
4.	Very Good	→	4.4
3.	Good	→	3.3
2.	Fair	→	2
1.	Poor	→	1

The students reported that very good health should be recorded with a value of 4.4 rather then 4.0, reflecting their view that very good health was closer to excellent health than to good health. Good health was recorded with a value of 3.3, reflecting their views that this state of health was closer to very good health than to fair health. Although the numbers on the right or the left both work fine for reflecting the order among the health states, only the numbers on the right (1, 2, 3.3, 4.4, 5) reflect both order and distance and could be used to calculate the average health status for a group of people.

Averages are often used to summarize continuous outcome in epidemiology (e.g. average blood pressure in a group of people). However, one should keep in mind the following quotation of a former baseball manger (Booby Bragen: http://sportsillustrated.cnn.com/vault/article/magazine/MAG1074702/index.htm)

> *"Say you were standing with one foot in the oven and one foot in an ice bucket. According to the percentage people, you should be perfectly comfortable*

If I have a group of people and half of them report having excellent health (5) and the other half report having poor health (1), then is it informative to report that the average health status of that group is 3? This value suggests slightly less than good health, when it underestimates the health status for half of the group and over estimates it for the other half?

Outcome Measures for Binary Variables

Since the values assigned to the categories of nominal variables are arbitrary labels, an outcome measure for a nominal outcome should be independent of the values that are assigned to the categories of the variable. Mean and percentile values are not appropriate outcome measures for nominal variables. However, percentages (proportions)

of subjects in categories provide a valid description of the distribution of a nominal variable.

Proportions relate the size of the population that have (or develop) the disease to the total size of the population of interest (**part/total measure**). **Odds** are an alternative measure of the likelihood of belonging to a particular category of a nominal variable. For example, the odds of disease measures the size of the population who have (or develop) the disease to the size of the population that does not have (or does not develop) the disease (a **part/non-part measure**). The value for an odds is easily calculated from the value for a proportion by the following transformation

odds = (proportion)/(1 - proportion)

Alternatively, the value for a proportion can be calculated from the value for an odds by the following formula:

proportion = (odds)/(1 + odds)

Values for proportions range from 0.0 to 1.0. Values for odds range from 0.0 to positive infinity. If the value for the proportion is small, then the corresponding odds will also be small. However, as the magnitude of the proportion increases, so does the difference between values for the odds and the proportion as shown in the following table

Relationship between Proportions and Odds.

Prevalence	Prevalence Odds
0.01	0.01
0.02	0.02
0.03	0.03
0.05	0.05
0.10	0.11
0.20	0.25
0.30	0.43
0.40	0.67
0.50	1.00
0.60	1.50
0.70	2.33
0.80	4.00
0.90	9.00
0.95	19.00
0.97	32.33
0.98	49.00
0.99	99.00

The previous table demonstrates that values for two proportions that are very close to one another (e.g. (0.01 and 0.02), or (0.98 and 0.99)) may have values for the corresponding odds that are also close to one another (0.01 and 0.02 for the first pair) or far apart (49.00 and 99.00 for the second pair). This implies that a measure of association that suggests large differences between values for two odds may not necessarily imply large differences between values for the corresponding proportions. Although a proportion may be a more intuitive outcome measure for a nominal variable, a particular analysis may require that the odds be the outcome measure of choice (e.g. logistic regression).

Prevalence

Prevalence measures how much outcome (disease) exists in the population **at a point in time**. Time can be measured in different dimensions. For example, one could think of a **chronologic date** and report the prevalence of AIDS in the United States today. Alternatively, time could refer to a person's **age** and you could consider the prevalence of low back pain among men at their 65^{th} birthday. Finally, time could be recorded in terms of **life time events** and consider the prevalence of cataracts at the time of retirement among men and women. In all cases, prevalence refers to a **snapshot** of a population at a point in time and quantifies the amount of a disease (or some other characteristic) that exists in the population at that time. It's involves examining people, perhaps taking a survey about their health, and reporting how much disease exist in that population at that one point in time.

Prevalence usually is defined as the proportion of people who have the disease at that point in time.

Prevalence = (# with disease)/(# examined)

If I had a camera and could take a picture of everyone taking this course, then I could see how many of you were wearing eyeglasses. I could report the prevalence of eye problems among people taking this course. If there were 100 people watching me right now, and 30 were wearing eyeglasses, then the prevalence of wearing eyeglasses (prevalence of having eye conditions) would be

30/100 = 30%

Alternatively, I could turn that prevalence into an odds and report the prevalence odds

30/70 = 43%

The following table is from the Framingham Heart Study teaching data set. It shows the prevalence of Coronary Heart Disease (CHD) at the 1956 exam, for groups defined by age and sex.

Sex	Age Group	Number at Exam	Number with CHD	Prevalence
Female	30-40	415	0	0/415= .00
	41-50	908	6	6/908= .01
	51-60	795	38	38/795=.05
	> 60	372	26	26/372=.07
Male	30-40	339	6	6/339=.02
	41-50	731	24	24/731=.03
	51-60	584	37	37/584=.06
	> 60	290	57	57/290=.20

The last column of this table reports the prevalence of CHD for the various groups. Not surprisingly, the prevalence of CHD increases with age for both men and women. In addition, for any age, the prevalence of CHD is higher for men than for women.

These conclusions reflect the curse of being an epidemiologist. We look at data and we report associations. For example, we observe that prevalence of CHD increases with age and is great for men than women. In addition to reporting associations, epidemiologists try to identify the reason for an observed association. For example, what are potential reasons for the higher prevalence of CHD existing for men, compared to women?

One possible explanation for the higher prevalence of CHD among men is that men might be at higher risk for developing coronary heart disease. Therefore a group of men would show a higher **incidence** for developing CHD than a comparable group of women. Perhaps, the reason we see a higher prevalence of CHD among men at the 1956 examination is that more new cases of CHD occurred among men in the previous weeks, months and years prior to the 1956 examination than among women. The topic of incidence will be discussed in the next sequence of lectures.

Another reason for the higher prevalence of CHD among men might be that men live longer with their heart disease. Suppose that the incidence of heart attacks is the same on both men and women, but men survive their heart attacks. When I measure the prevalence of CHD in a population I'll find men who developed heart disease a month ago, a year ago, two years ago, and are still alive. On the other hand I will not find the women who developed CHD in the past because they may not have survived. So another reason you might see higher prevalence of disease in one population than another has nothing to do with the incidence of developing disease but with the duration of disease.

Unfortunately, there are even more possible explanations, beyond just having higher incidence or having higher duration, which might explain the higher prevalence of CHD among men. Recall that we epidemiologists measure associations. We hope those associations represent the true, causal effect of a risk factor on an outcome. However,

before we can conclude that an association reflects a causal effect of a risk factor we have to exclude three possible alternative explanations: **bias, confounding, or chance.**

For example, it may be that men and women really do have the same prevalence of disease, but maybe men see their physicians more often and are tested more often for CHD. If so, the reason for the higher prevalence of CHD among men might be due to a measurement bias: the underreporting of CHD among women.

A fourth possible explanation for the higher prevalence of CHD is that men and women have the same risk of developing disease based on their sex, but men have more additional risk factors (e.g. smoking hypertension, …) that increases their risk of CHD and leads to a higher incidence of CHD (and their higher prevalence). The higher prevalence of CHD among men is not a reflection of the effect of sex on the incidence of CBD, but the effect of confounding factors (e.g. smoking, hypertension, …) that exist more commonly among men. The topic of confounding that we'll be talking about in future sessions.

Finally, suppose that in general, men and women have the same prevalence of CHD. However, your data contain only a sample of men and a sample of women. Because of sampling variability (chance), the observed prevalence of CHD among men in our data may be higher than what exists in general among all men. Suppose that the opposite is true for women: the observed prevalence of CHD among women in our data may be lower than what exists in general among all women. Therefore, the association that we see ion our data is an artifact of sampling variability.

In summary, there are many plausible explanations that could explain why the prevalence of CHD is higher among men than among women. The challenge for epidemiologists is to try to rule out the unlikely explanations, to conclude with what is the most likely explanation. In this manner, *epidemiologists* are often thought as detectives as suggested by the following quotation:

"How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth? We know that he did not come through the door, the window, or the chimney. We also know that he could not have been concealed in the room, as there is no concealment possible. When, then, did he come?"

A. C. Doyle: *The Sign of the Four (1890)*

Unfortunately, there is still another possible explanation (reverse causation) for observing and association that occurs if some studies (but not this example) when using prevalence data. This will be discussed in a future lecture. However, for the present, the main take-home message for interpreting prevalence data is that prevalence is influenced by both **incidence**, the development of new cases of disease, and the **duration** of disease. This issue is demonstrated by the following example.

The following quotation is from the headline of a CNN report (CNN.Com Monday, June 13, 2005 Posted: 1:42 PM EDT (1742 GMT*):*

1 million living with HIV in U.S."

"Statistics provide good and bad news"

Why is it good news that one million people were living with HIV in the United States in 2005? The answer is found later in the article.

"for the first time since the height of the AIDS epidemic in the 1980s, more than a million Americans are believed to be living with the virus that causes AIDS, the government said on Monday."

"The latest estimate is both good news and bad news-- reflecting the success of the drugs that keep more people alive." The duration was increasing.

The good news implied by this quotation is that treatment for patients with HIV appeared to be working, in that is was extending the lifetime of patients with this condition. The bad news was the incidence of HIV was still increasing. There were two reasons why the prevalence was high: **higher incidence** of developing HIV (bad news), and **longer duration** (survival) with disease (good news).

Incidence

Incidence refers to the occurrence of new cases of a disease (outcome) in a population over a period of time. It involves following a group of people who are **at risk for developing the developing the disease** and recording the outcomes that occur. Unlike prevalence, which is measured at a point in time, incidence requires a period of follow-up.

For example, one goal of the Framingham Heart Study (FHS)) was to observe the incidence of Coronary Heart Disease (CHD) among all participants and among participants with certain risk factor characteristics (e.g. smokers, non-smokers, …). The teaching data set for this class contains 4,434 participants who attended the 1956 biennial examination. However, 194 of them were previously diagnosed with CHD, leaving 4,240 participants who are at risk for developing a first CHD event during the 24 years of follow-up. The following results were observed for these 4,240 subjects:

- 1406 died from any cause
- 88,389.45 person-years were observed before subjects died or the study ended

- 824 died from non-CHD causes (no longer at risk of developing CHD),
- 32 lost-to-follow-up and had not developed CHD at their last contact,
- 2338 completed 24 years of follow-up and did not develop CHD,
- 1046 developed CHD during the 24 years of follow-up (including 582 deaths)
- 80,925.16 person-years were observed before subjects were lost-follow-up, died, developed CHD, or the study ended

If you only published that 1406 participants died during this study and gave no other information, then a reader could draw no conclusion about whether this is a large number of deaths (reflecting a high risk population) or a small number (reflecting a low risk population). A reference point is needed and two are available to report:

1. The size of the population at risk (1406 deaths occurred in 4420 subjects during 24 years of follow-up)
2. The amount of follow-up time observed (1406 deaths occurred during 88,389.45 person-years of follow-up)

These findings reflect the two types of calculations used by epidemiologists to measure the incidence of an outcome in a population:

1. **Cumulative Incidence**
2. **Incidence Rate**

C**umulative Incidence** is defined as the proportion of people who develop a disease (outcome) during a fixed period of follow-up. As a proportion, the value for the Cumulative Incidence can range for as low as 0.0 (when there are no deaths) and 1.0 (when everyone dies). The period of follow-up must also be specified. 1406 death occurred during 24 years of follow-up. Far fewer deaths would be expected during only 1 year of follow-up, and 4420 deaths would be expected during 100 years of follow-up. From the data given above, the value for the **24-year Cumulative Incidence of Death is:**

(# Deaths)/(# At Risk) =1406/4420 = 0.33

The direct calculation of the Cumulative Incidence as shown above requires knowledge of the mortality outcomes for all of the 4420 subjects who were followed. For example, if a subject were lost-to-follow-up, then his/her mortality outcome would be unknown and you would not know whether or not to include that person in the numerator of the calculation for the Cumulative Incidence. Fortunately, we have complete follow-up for mortality in these data. The investigators determined mortality status for all subjects, even those who failed to attend scheduled exams, by monitoring reported deaths through a search of a national death index. They contacted previously specified relatives and friends of any subject who was lost-to-follow-up. However, the above calculation could not be preformed to measures the 24-year Cumulative Incidence of CHD as there were 32 subjects who were lost-to-follow-up and had unknown CHD status after becoming lost. In addition, there were 824 other subjects who died from non-CHD causes and we would have no way to know if these subjects would have developed CHD during the remaining part of the 24-year follow-up time had they not died. In such instances where the Cumulative Incidence cannot be calculated directly and alternative measure of incidence, the **Incidence Rate**, is typically calculated.

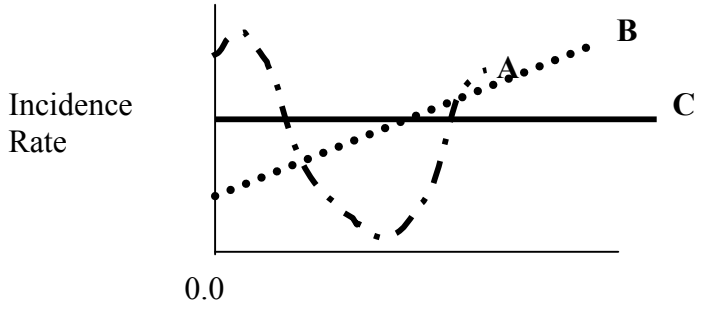
One limitation of the Cumulative Incidence is that is does not account for the timing of the outcomes during the follow-up period. The value for the Cumulative Incidence would be 0.33 if all 1406 deaths occurred at the very beginning of the follow-up period or at the very end of the follow-up period. On the other hand, the values for the Incidence Rate would differ under these two scenarios.

Finally, Cumulative Incidence is often labeled as the **Estimated Risk**, or the **Average Risk.** Risk refers to an individual's probability of developing an outcome. A risk for an individual is estimated by the Cumulative Incidence that is calculated for a group of people who are assumed to share the same common risk as the individual in question. As with the Cumulative Incidence, a risk requires a specification of a time period. My risk of dying in the next few moments is (hopefully) very small, but my risk of dying in the next 100 years is 100% (1.0).

An alternative measure of incidence is the **Incidence Rate** (also labeled as the **Instantaneous Risk**, **Hazard Rate**, or **Incidence Density**). It refers to the instantaneous risk of developing the outcome at a point in time during the follow-up period, among

those at risk at that time. The value for the Incidence Rate may be constant over the period of follow-up or may vary over time, showing periods of high and low values. This is depicted by the three Incidence Rate functions, IR(t), that are shown in the following figure:

Examples of Incidence Rate functions (IR(t)).



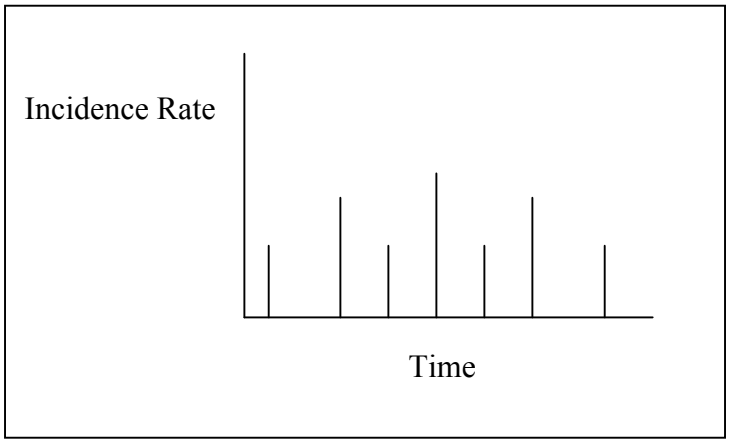
The Incidence Rate function depicted by the "roller-coaster" graph (A) may pertain to a study that records deaths among patients undergoing a major surgical procedure (e.g. coronary bypass surgery or transplant surgery). Patients may be at high risk of dying during and initially after the surgery, as depicted by the higher values for the Incidence Rate function. However, if patients survive this critical period, then they may go through a period of lower risk, as depicted by the lower values for the Incidence Rate function. Finally, as time elapses, these subjects will age and develop risk factors that may increase their risks of dying during subsequent periods, as depicted by the increasing values for the Incidence Rate function.

The Incidence Rate function depicted by the increasing line (B) might be expected from following subjects for a long period of time. As the follow-up time increase, the subjects at risk for developing an outcome at a point in time become older and are at higher risk for developing the outcome.

The Incidence Rate function depicted by the flat line (C) might be expected from a study with a short period of follow-up, where the risk of developing the outcome might not change over the period of follow-up.

As a measure of Instantaneous Risk, the value for the Incidence Rate could be estimated at each point in time during the follow-up by calculating the instantaneous Cumulative Incidence. However, because of the limited number of observed outcomes, values for the Cumulative Incidence would equal zero except for those times at which an outcome occurs. This would lead to a series of spikes to describe the Incidence Rate function as shown in the following figure. Spikes occur only at times when outcomes

occur and the height of each spike corresponds to the portion of outcomes that develop at that time among the subjects at risk. Although this accurately describes the data, the underlying Incidence Rate function is probably continuous in nature and better described by examining the outcome incidence during intervals of time.



The value for Incidence Rate function during an interval of time is estimated by the following formula

IR (# cases of disease during the interval)/(amount of person-time during the interval)

This expression is different from the formula for the Cumulative Interval in the denominator refers to the amount of person-time observed during the interval rather than the number of subjects at risk at the start of the interval. This accounts for the length of the interval. Suppose two investigators (A and B) wished to calculate the value for the Incidence Rate function at a time, t. Suppose that the Incidence Rate function was constant around that time. If investigator B's interval around t was half as wide as Investigator A, then she would expect two see roughly half as many outcome cases and half as much person-time during her interval as were seen by Investigator A. Hence their Incidence Rate calculations should provide similar values. However, this would not be true for calculations of Cumulative Incidence, which are dependent on the length of follow-up (length of the interval).

If the Incidence Rate function is constant during the follow-up period (Graph C in the above figure) then its value for any time, t, can be calculated by the following formula:

IR = (total number of cases of disease)/(total amount of person-time)

This assumption justifies treating 10 person-years of follow-up from following 10 subjects for 1 year as equivalent from following 1 person for 10 years. If the Incidence

Rate function varies over time (Graphs A and B in the above figure), then the value for the Incidence Rate from this formula is an average of the different values for the Incidence Rate function at different point in time. This is described by the data in the following table.

Year	Population A			Population B			Population C		
	# At Risk	# Deaths	Person-years	# At Risk	# Deaths	Person-years	# At Risk	# Deaths	Person-years
1	100	10	95	100	2	99	100	10	95.0
2	90	10	85	98	2	97	91	9	85.5
3	80	10	75	96	2	95	83	8	77.0
4	70	10	65	94	34	77	74	7	69.5
Total		40	320		40	368		34	327

Each population has 100 subjects at risk at the beginning of the study. 40 deaths occur in Population A and in Population B, leading to identical values for the 4-year Cumulative Incidence of Death

$$40/100 = 0.40$$

However, the Incidence Rate functions for Populations A and B differ markedly. If we assume that the deaths occurred at the midpoint of each year then the values for the Incidence Rate function (cases/1 person-year) for each year are shown in the following table

Year	Population A			Population B			Population C		
	# Deaths	Person-years	Incidence Rate	# Deaths	Person-years	Incidence Rate	# Deaths	Person-years	Incidence Rate
1	10	95	0.11	2	99	0.02	10	95.0	0.11
2	10	85	0.12	2	97	0.02	9	85.5	0.11
3	10	75	0.13	2	95	0.02	8	77.0	0.10
4	10	65	0.15	34	77	0.02	7	69.5	0.10
Total	40	320	0.13	40	368	0.44	34	327	0.10

The Incidence Rate function for Population A increases over time despite the constant number of deaths each year. The reason for the increase is the decreasing number of subjects who are at risk at the beginning of each year.

Although Populations A and B show a common value for the 4-year Cumulative Incidence of death, they show markedly different Incidence Rate functions. The Incidence Rate function for Population A increases slightly over time, where it is very low during the first three years for Population B and then increases dramatically during the fourth year of follow-up.

The Incidence Rate function for Population C is roughly constant of the four years of follow-up. The value for the constant Incidence Rate is

$$IR = 34/(327 \text{ person-years}) = 0.10(\text{cases/1 person year})$$

Similar calculations can be performed for Populations A and B but the resulting values ,(0.13cases/person-year) for Population A and (0.44cases/person-year) for Population B, are weighted averages for the different values for the Incidence Rate functions for these populations. The weights for these calculations are the amounts of person-time for each age-specific Incidence Rate. Mathematically, this is shown for Population B by the following equation:

40/368py = [99**(2/99py)** + 97**(2/97py)** + 95**(2/95py)** + 77**(34/7py)**]/ [(99+97+95+77)py]

It should be also noted that the dimension for an Incidence Rate is cases/person-time. This can be simplified to 1/time (by cancelling cases in the numerator with persons in the denominator). This implies that the value for the Incidence Rate depends on the chosen unit of time. For example:

1 case / 1 person year = 1 (cases/person year)

= 1 (cases/52 person-weeks) = 0.0192 (cases/person-week)

= 1 (cases/365 person-days) = 0.0027 (cases/person-day)

= 10 (cases/person-decade)

= 100 (cases/person-century)

This example underscores the meaningless for a reported Incidence Rate without the specification of the time period.

As expected, there is a mathematical relationship between the Cumulative Incidence for a period of time and the Incidence Rate function that exists during that period. For example, if the Incidence Rate is constant during a period of time and is equal to 1case/(100person-years), then 100 persons at risk at the beginning of that interval would approximately provide 100 person-years of observation and 1 case of the outcome if followed for 1 year. Hence the one-year Cumulative Incidence approximately would be 1/100 = 0.01. This is an approximation because the one subject who becomes a case would not provide a full year of follow-up (unless the case developed at the very end of the year) and the 100 subjects would provide slightly than 100 person-years of follow-up. In general, if the value for the Incidence Rate is small or the time period of follow-up is short, then the value for Cumulative Interval during this time period can be calculated from the following formula

$$CI \approx IR \times (\text{length of time period})$$

If the time period is not short or the Incidence Rate is not small this approximation does not hold. For example, the Incidence Rate for Population C is .10(deaths/person year). The previous formula would yield the following results when applied to different time periods

One-year CI = .10 x 1 = 0.1
Four-year CI = .10 x 4 = 0.4
Ten-year CI = .10 x 10 = 1.0
Twenty-year CI = .10 x 20 = 2.0

Although the One-year Cumulative Incidence listed above is similar to what would be calculated from the data in the above Table (10/100 = 0.10), the four-year cumulative incidence overstates what would be calculated from that data (34/100 =0.34). More importantly, the value for the 20-year Cumulative Incidence is mathematically not possible, since the value for a Cumulative Incidence cannot exceed 1.0. The overstated values listed in this calculation reflect a problem of not accounting for the diminishing size of the population at risk as subjects are followed over time. The following formula shows the general association between the Cumulative Incidence and the Incidence Rate when the latter is constant over time

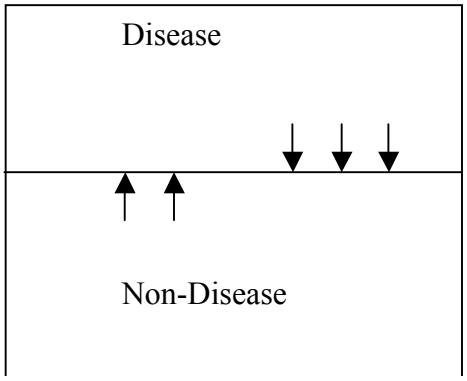
$$CI = 1 - e^{-IR \times (\text{time period})}$$

Using this general formula for Population C yields the following values for the Cumulative Incidence for the different time periods list above

One-year CI = $1 - e^{-.10 \times 1}$ = 0.10
Four-year CI = $1 - e^{-.10 \times 4}$ = 0.33
Ten-year CI = $1 - e^{-.10 \times 10}$ = 0.63
Twenty-year CI = $1 - e^{-.10 \times 20}$ = 0.86

The values for the one-year and four-year Cumulative Incidence (0.10 and 0.33, respectively) are very similar to those obtained from the data in the above Table (0.10 and 0.34, respectively).

There is also a mathematical relationship between the Prevalence of a disease and the Incidence Rate of that disease when we have a **steady state**. The steady state assumption implies that the Prevalence and the Incidence Rate of the disease are constant over a time period for a population of fixed size. This is depicted by the following figure:



If there are N people in the Population and P = the Prevalence of disease then at any time

diseased subjects = N(P)
non-disease subjects= N(1-P)

Suppose the disease in question is non-fatal and IR is the Incidence Rate that determines the number of new cases of disease that develop from the non-diseased group. To retain a steady state, this flow would need to be balanced by the number of cures of the disease that develop from the diseased group. If the D = average duration of disease, then the Cure Rate would equal 1/D. Thus for the steady state to remain, the following equation must hold

$$N(P)(1/D) = N(1-P)(IR)$$

This implies

$$P/(1-P) = IR \times D$$

It follows that is the Prevalence of disease (P) is small that

$$P \approx P/(1-P) = IR \times D$$

The important message from this calculation is not the mathematical relationship between Prevalence and Incidence in a steady state, but the general principal that Prevalence is function of Incidence and disease duration.

Measures of Association

Causal Inference

The **Causal Effect** of an exposure for an individual is the difference in the outcomes (Y) for that individual if given the exposure (Y^1) versus not given the exposure Y^0). These two outcomes are referred to as the **counterfactual outcomes** for that individual. For example, the causal effect of lifetime smoking starting at age 20 on the development of CHD for a person is the difference in the CHD counterfactual outcome if that person started smoking at age 20 (smoker), compared to the CHD counterfactual outcome if that person did not start smoking at age 20 (non-smoker). Since only one of these outcomes can be observed in reality (depending on whether the individual actually smoked or did not smoke), the causal effect on the individual level cannot be measured.

The **average causal effect of a risk factor for a population** is the difference in average counterfactual outcomes, $E(Y^1)$, when all members of the population receive the exposure and the average counterfactual outcomes ,$E(Y^0)$, when none of the members of the population receive it. **Average Causal Measures of Effect** can be defined as

Casual Risk Difference = $E(Y^1) - (E(Y^0)$.

Causal Risk Ratio = $E(Y^1) / E(Y^0)$.

Causal Odds Ratio = $[E(Y^1) /(1-E(Y^1))] /[E(Y^0) /(1-E(Y^0))]$

(N.B. E(Y) refers to the statistical term of "expected value" of Y and represents the average (mean) of y)

If the population is comprised of some members who receive the exposure and others who do not, then it may be possible to estimate each average counterfactual outcomes and the average causal effect of the exposure in the absence of **confounding** and **bias**. For example, in the absence of confounding and bias, the incidence of coronary heart disease from a group of non-smokers (**a factual outcome**) is a valid estimate for the counterfactual outcome for a group of smokers had they not smoked. This assumes an ability to "exchange" factual outcomes of the non-smokers with the counterfactual outcomes for the smokers and vice versa (Greenland S, Robins JM. Identifiability, Exchangeability, and Epidemiological Confounding. *Intl J Epidemiol* 1986;15:412-419.). This implies that the observed difference (ratio) in incidence of coronary heart disease in the two comparison groups (**a measure of association**) is an estimate of the average causal effect of smoking.

Hernan and Robin (Estimating causal effects from epidemiological data. J *Epidemiol Community Health* 2006;60:578-586.) demonstrate how these causal effect measures can be estimated by measures of association from epidemiologic studies that are free of bias and confounding (e.g. randomized experiments) or from observation

studies by conditioning the values of the confounding. They proposed using the method of inverse probability weighting (IPW) to estimate causal effect to control for known confounders. This method is similar to standardization and will be discussed in future lectures.

Measures of Association

In epidemiology a **measures of association** compares the outcome measurement (Prevalence, Incidence Rate, …) in groups of subjects that are defined by categories of a risk factor of interest (exposure**).** Mathematically, measures of association are either ratios or differences of an outcome measure in these groups. In the absence of bias and confounding, these measures of association provide estimates for the causal effect of the risk factor on the outcome (measure of effect). The following table presents some commonly used measures of association in epidemiology. As mentioned previously, the outcome measurement of choice is typically determined by the scale of the outcome.

Table Some Common Measures of Association.

Outcome	Outcome Measure	Measure of Association
Nominal	Proportions (Cumulative Incidence, Estimated Risk) Incidence Rates Odds	Risk Ratio/Risk Difference Rate Ratio/Rate Difference Odds Ratio
Ordinal	Above Measures Medians (Percentiles)	Above Measures Ratio/Difference of Medians (Percentiles)
Continuous	Above Measures Averages	Above Measures Ratio/Difference of Averages

Binary Outcome

A binary outcome may reflect a natural dichotomy (e.g. dead versus alive) or can be created from nominal, ordinal, or continuous outcome by collapsing categories (e.g. New York Heart Association Classes III and IV versus Classes I and II (Goldman et al.,1981)) or specifying a threshold value for a continuous variable to separate high from low outcomes (e.g. hypertension (SBP $\geq$ 140) versus no hypertension). If the exposure of interest is also binary (e.g. current smoker vs. non-smoker), then data relating the exposure to the outcome can be displayed in a 2x2 table as follows"

2x2 Table Displaying the Relationship between a Binary Exposure (E) and Disease (D).

	D		
	+	-	Total
E +	a	b	N_1
E -	c	d	N_0
Total	M_1	M_0	T

Data displayed in 2x2 tables arise from a variety of study designs: cohort studies, case control studies, cross-sectional studies, and experimental studies. Often the choice of the appropriate measure of effect may depend on the type of study design. For example, the odds ratio or some function of the odds ratio is typically the only measure of effect that is calculated from most case control studies.

If there are no losses-to-follow-up or losses due to competing risks, then the Cumulative Incidence of disease for the exposure groups provides estimates for the risk of disease for each group.

Estimated Risk of Disease

Exposed Subjects (E+) $R_1 = a/N_1$

Non-exposed Subjects (E-) $R_0 = c/N_0$

Common measures of association based on ratio or differences of estimated risks are shown in following table.

Measure of Effect	**Formula**
Risk Ratio (Relative Risk)	$RR = R_1 / R_0$
Risk Difference (Attributive Risk)	$RD = R_1 - R_0$
Disease Odds Ratio	$OR = [R_1/(1-R_1)]/[R_0/(1-R_0)]$

Similar measures can be calculated with data from cross-sectional studies using prevalence rather than incidence as an outcome measures.

Incidence Rate data for a binary exposure can be displayed in a slightly different table

	D+	Person-time	Incidence Rate
E +	a	K_1	$R_1 = a/K_1$
E -	c	K_0	$R_0 = c/K_0$
Total	M_1	T	

Measures of Association can be calculated by taking the ratio or difference of the Incidence Rates

$$\text{Rate Ratio} = R_1 / R_0$$

$$\text{Rate Difference} = R_1 - R_0$$

The value for a Rate Ratio is a pure number. However, Rate Difference, like Incidence Rates, is measured in case/person-time.

Example: The following tables display the relationship between current smoking and the incidence death during 24 years of follow-up from the FHS teaching data set.

	Death			
	+	-	Total	Estimated Risk
Smoker	788	1393	2181	0.3613
Non-Smoker	762	1491	2253	0.3382

Risk Ratio = RR = 0.3613 / 0.3382 = 1.0683

Risk Difference = RD = 0.3613 – 0.3382 = 0.0231

	Deaths	Person-years	Incidence Rate (deaths/10,000py)
Smoker	788	44,440.38	177.3162
Non-Smoker	762	46,675.20	163.2559

Rate Ratio = RR = 177.3162 / 163.2559 = 1.0861

Rate Difference = RD = (177.3162 – 163.2559)cases/(10,000py)
= 14.0603 (cases/10,000py)

It should be noted that the label RR may refer to either a Risk Ratio calculation or a Rate Ratio calculation. Often the term **Relative Risk** (also labeled RR) is used to describe either calculation. However, as shown in the previous example, the value for the Risk Ratio and Rate Ratio will tend to differ, and the use of a single tem (Relative Risk) to describe two different results may be confusing.

Attributable Proportions

If R_1 is the risk of developing an outcome among an exposed subject, then a question of interest might be how much of the magnitude of R_1 is actually caused by the exposure. If R_0 is the risk that an exposed subject would have had in the absence of the exposure then $(R_1 - R_0)$ is the extra risk that is caused by the exposure and the proportion of R_1 that is attributed to the exposure is

$$(R_1 - R_0)/R_1 = (R_1/R_0 - R_0/R_0) / R_1/R_0 = (RR - 1)/RR$$

This quantity can be estimated using the estimated risks (cumulative incidence) from population data and is referred to as the **Attributable Proportion among the Exposed** (Ashcengrau and Seage), **Attributable Fraction** (in Stata and in Rothman KJ. Epidemiology: An Introduction 2nd Edition. Oxford University Press 2012), and **Attributable Risk Percent (**Hennekens and Buring**).**

A similar question for consideration is what proportion of the **average risk** in a population is attributed to some members of that population having the exposure of interest. This is a more of a public health consideration as the answer is linked to a specified population with a specific prevalence of exposure (p). The average risk is a population is

$$R_T = pR_1 + (1-p)R_0$$

The portion of the average risk that is attributable to the exposure is

$$\begin{aligned}
(R_T - R_0)/R_T &= [pR_1 + (1-p)R_0 - R_0]/ [pR_1 + (1-p)R_0] \\
&= [pR_1 - pR_0]/ [pR_1 - pR_0 + R_0] \\
&= [p(R_1 - R_0)] / [p(R_1 - R_0) + R_0] \\
&= [p(R_1/R_0 - R_0/R_0] / [p(R_1/R_0 - R_0/R_0) + R_0/R_0] \\
&= [p(RR - 1)] / [p(RR - 1) + 1]
\end{aligned}$$

This quantity also can be estimated using the estimated risks (cumulative incidence) from population data and is referred to as the **Attributable Proportion in the Total Population** (Ashcengrau and Seage), **Attributable Fraction for the Population** (Stata, Rothman), **Population Attributable Risk Percent (**Hennekens and Buring**).**

Example: The following table displays the relationship between current smoking and the incidence death during 24 years of follow-up from the FHS teaching data set.

	Death			
	+	-	Total	Estimated Risk
Smoker	788	1393	2181	0.3613
Non-Smoker	762	1491	2253	0.3382
Total	1550	2884	4434	0.3496

Risk Ratio = RR = 0.3613 / 0.3382 = 1.0683

p = Prevalence of Smoking = 2181/4434 = 0.4919

Attributable Fraction – Exposed = (1.0683 – 1)/1.0683 = 0.0639

Attributable Fraction – Population = [0.4919(1.0683-1)]/ [0.4919(1.0683-1) +1]
= 0.0325

Number Needed to Harm and Number Needed to Treat

The **Number Needed to Harm** is the number of subjects, if given a harmful exposure ($R_1 > R_0$), would cause the one case of disease. For example, the following table displays the relationship between current smoking and death during 24 years of follow-up from the FHS teaching data set.

	Death			
	+	-	Total	Estimated Risk
Smoker	788	1393	2181	0.3613
Non-Smoker	762	1491	2253	0.3382

Risk Difference = RD = 0.3613 – 0.3382 = 0.0231
= 3613/10,000 –3382/10,000 = 231/10,000

This result implies that if 10,000 subjects smoked rather than not smoking then 231 extra deaths would occur. Therefore, if 43.29 subjects smoked (rather than not smoking) then 1 extra death would occur. This number of based on the following formula:

231/10000 = (231/231 / 10,000/231) = 1/43.29 = 1/RD

The general formula for the **Number Needed to Harm (NNH)** is

$$NNH = 1/RD$$

Similarly is an exposure (treatment) lowers the risk of developing an outcome ($R_1 < R_0$) then the **Number Needed to Treat (NNT)** to prevent one case of disease is

$$NNT = 1/(R_0 - R_1) = 1/|RD|$$

Regression Coefficients

Estimates for most of the effect measures that were presented in these notes can also be obtained from **regression models** that describe an outcome measure as a function of the exposure of interest.

The general formula for a straight line to describe linear relationship between two variables X and Y is

$$Y = mX + b$$

where

b = Y-intercept = value for Y when X=0

m = slope = change in Y when X changes by one unit

In epidemiology research, Y (dependent variable) represents a function of outcome measure and X (independent variable) represents an exposure (E) and the model is usually written as

$$f(\text{outcome measure}) = B_0 + B_1E$$

where

B_1 = slope = change in f(outcome measure)/ unit change in E

When the outcome is binary scale and the outcome measure is a proportion (P: prevalence, cumulative incidence) then the **logistic regression model** to describe the relationship between and exposure (E) and the outcome. The logistic regression model uses a function of P (logit: natural logarithm of the P/(1-P)) for Y, yielding the following formula

$$\log(P/(1-P)) = B_0 + B_1E$$

where

$$B_1 = \Delta\log(P/(1-P)) \text{ / unit } \Delta \text{ in E}$$

If the exposure is binary, labeled 1 for exposed subjects and 0 for non-exposed subjects, then

$$\begin{aligned} B_1 &= [\log(P_1/(1-P_1)) - \log(P_0/(1-P_0))] / (1-0) \\ &= \log[P_1/(1-P_1)) / (P_0/(1-P_0))] \\ &= \log(OR) \end{aligned}$$

where P_1 = outcome measure for exposed subjects and P_0 = outcome measure for non-exposed subjects.

Example: The following table, taken from the FHS teaching data set, shows the relationship between smoking at the 1956 exam and the incidence of death during 24 years of follow-up

	Death			
	+	-	Total	Odds
Smoker	788	1393	2181	788/1393 = 0.5657
Non-Smoker	762	1491	2253	762/1491 = 0.5111

Odds Ratio = OR = 0.5657 / 0.5111 = 1.1068

If a logistic regression model were fit to these data the resulting fitted model is

$$\log(P/(1-P)) = -0.6713 + 0.1015(\text{Smoker})$$

$$B_1 = 0.1015 = \log(OR)$$

$$OR = e^{0.1015} = 1.1068$$

Study Designs

A study design is a plan (proposal) to enroll subjects, collect data, and analyze the data to draw inference. In epidemiology, studies can be characterized by their design. A commonly sited list of study designs that will be discussed in this course is

1. Case Reports
2. Ecological Studies
3. Cross Sectional Studies
4. Experimental Studies
5. Cohort Studies
6. Case Control Studies

Case Reports

Case Reports are detailed descriptions of unexpected and unusual symptoms, disease, treatments, and outcomes of individual patient(s). Because of the unexpected nature of their findings, case reports often serve as a basis for **hypothesis generation** and for springboards for future studies.

For example, the following report was published during one of the major cholera outbreak in Great Britain during the 1800's (Craigie D. *An account of the epidemic cholera of Newburn in January and February 1832.* Edinburgh Medical Surgical Journal 1832;37:337-384). It describes the death of Rev. John Edmonston, who had visited the sick of his congregation and unfortunately contracted the disease despite taking all known precautions at that time. The report concentrates on his dinner of pickled salmon on the night before he developed symptoms and perished.

> *"I venture to assert … that the intercourse with the sick went in this case for nothing; and, had Mr. Edmonston secluded himself with his garden wall, the **pickled salmon** would have produced precisely the same effect."*

The potential causal hypothesis that can be developed from this report is that a diet of pickled salmon might be a cause of cholera. This hypothesis could be addressed with one of the study designed options that will be discussed in future lectures.

Perhaps one of the more famous examples of Case Reports are the series of publications in 1981 from the **Centers for Disease Control (CDC)** Mortality **and Morbidity Weekly Report (MMWR).** These are weekly reports about health information and recommendations from state departments of public health in the United States. Information about these reports can be found at http://www.cdc.gov/mmwr. In the summer of 1981 the following MMWR reports were published:

"In the period October 1980 – May 1981, 5 young men, all active homosexuals, were treated for biopsy-confirmed *Pneumocystis carinii* pneumonia at 3 different hospitals in Los Angeles, California"

Editorial Note: "*Pneumocystis* pneumonia in the United States is almost exclusively limited to severely immunosuppressed patients. The occurrence of Pneumocystis in these 5 previously healthy individuals without a clinically apparent underlying immunodeficiency is unusual"

CDC – MMWR June 5 1981 /30(21); 1-3
www.cdc.gov/hiv/resources/reports/mmwr.1981.htm

"During the past 30 months, Kaposi's Sarcoma (KS), an uncommonly reported malignancy in the United States, has been diagnosed in 26 homosexual men (20 in New York; 6 in California)."

Editorial Note: … "The occurrence of this number of KS cases during a 30 month period among young homosexual men is considered highly unusual."

CDC – MMWR July 4;30:306-8

"Twenty-sex cases of Karposi's sarcoma (KS) and 15 cases of *Pneumocystis carinii* pneumonia (PCP) among previously healthy homosexual men were recently reported. … Since July 3, 1981, CDC has received reports of an additional 70 cases of these 2 conditions in persons without known underlying disease."

Editorial Note: "KS is a rare, malignant neoplasm seen predominantly in elderly men in this country."

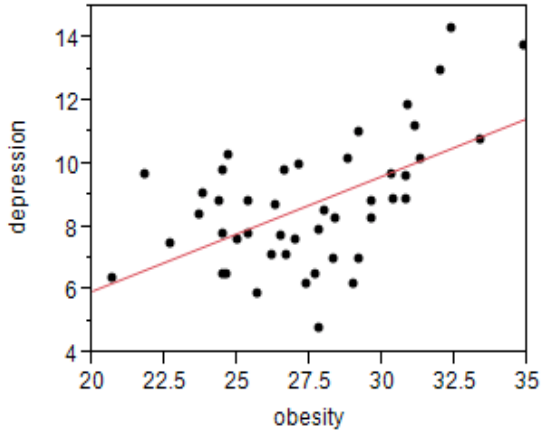
CDC – MMWR 1981 August 28;30:409-10

These reports deal with an unexpected occurrence of Karposi'e Sarcoma (KS) and Pneumocystis pneumonia among young, previously healthy homosexual men. The editorial notes comment that Pneumocystis pneumonia is a disease that usually occurs among severely immunosuppressed patients and that Karposi's sarcoma usually occurs among the elderly. These finding suggest that the cause of these unexpected disease might be related to some characteristic of the individuals (homosexual men) and impact their immune system. These reports lead to a series of studies that identified HIV as the cause AIDS.

Ecologic Studies

Ecologic Studies (correlation studies) examine the relationship between two factors on a population level, rather than on an individual level. The unit of the analysis is a population, rather than an individual subject.

For example, the following figure shows the relationship between the prevalence of depression and the prevalence of obesity in the United States. Each point in the figure corresponds to the prevalence of these two characteristics for a particular state. Data was obtained from CDC publications, using information from the **Behavioral Risk Factor Surveillance System (BRFSS).** The prevalence of depression was reported for the period 2006 – 2008 (http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5938a2.htm), and the prevalence of obesity for 2011 (www.cdc.gov/obesity/data/adult.html).



In general, states with higher prevalence of obesity tend to have higher prevalence of depression. However, does this imply that obesity causes depression or that depression causes obesity **in individuals**? The problem is that these data do not tell us if the inhabitants of a particular state who suffer from depression and the same individuals who are obese. The **Ecologic Fallacy** refers to the potential for incorrectly assuming that an association that exists on a population level reflects an association on an individual. This potential is demonstrated by the following example:

Suppose that the following data show the relationship between obesity and depression in 3 states:

	Prevalence of Obesity	Prevalence of Depression	Odds Ratio

State A:

Obesity	Depression Yes	No	Total
Yes	1	3	4
No	3	3	6
Total	4	6	10

	0.4	.04	0.33

State B

Obesity	Depression Yes	No	Total
Yes	2	3	5
No	3	2	5
Total	5	5	10

	0.5	0.5	.44

State C

Obesity	Depression Yes	No	Total
Yes	3	3	6
No	3	1	4
Total	6	4	10

	0.6	0.6	0.33

On the state level, we see a positive relationship between obesity and depression (as the prevalence of obesity increases over states, so does the prevalence of depression. However the opposite relationship is seen among individuals within states (obese individuals are less likely to be depressed).

Cross Sectional Studies

Cross-Sectional Studies (Survey Studies) report the prevalence of an exposure and a disease in a population at a point in time. For example, the following table from the FHS teaching data set reports the cross-sectional relationship between smoking status and the existence of coronary heart disease at the 1956 examination.

	CHD		Total	Prevalence of CHD
	Yes	No		
Smokers	86	2095	2181	86/2181 = 0.0394
Non-Smokers	108	2145	2253	108/2253 = 0.0479

Cross Sectional studies report prevalence outcomes. These data show that the prevalence of CHD is lower among smokers compared to non-smokers. As noted in earlier lectures on prevalence, the challenge is with the interpretation of any association found in such studies. For example, since prevalence is a function of incidence and duration of disease, two possible explanations for this association are:

1. Smokers have lowers risk (incidence) of developing CHD (unlikely)
2. Smokers who develop CHD have shorter duration of survival

For, any association, there are three additional generic explanations to consider:

3. **Bias**
4. **Confounding**
5. **Chance**

Bias refers to a flaw in a study design that leads to an invalid result. Biases can be characterized into two major types: **selection bias** and **measurement bias**.

A selection bias may occur in a Cross-Sectional study when the disease of interest might differentially influence the selection exposed and non-exposed subjects (or the exposure might differentially influence the selection of diseased and non-disease individuals). For example, the Framingham Heart Study initially enrolled 5209 subjects but only 4434 of them are included in the above table. Some of the 775 participants who are not included in this table may have died before the 1956 exam, but it is possible that others chose not to attend this exam. Perhaps the non-attending smokers had a higher prevalence of CHD and not attend the exam because of limitations due to the CHD.

A measurement bias pertains to the errors and measurement or classification of the exposure or the disease (or any other factor in a study). For example, perhaps smokers see their physicians less often and are tested less often for CHD. This might lead to an under-reporting of the true prevalence of CHD among smokers in a study. There are two general types of measurement bias:

1. **Random Misclassification** (non-differential misclassification) occurs when the errors in classification of disease are the same in the exposed and non-exposed groups (or the errors in misclassification of exposure are the same in diseased and non-diseased groups)

2. **Non-Random Misclassification** (differential misclassification) occurs when the errors in classification of disease are different in the exposed and non-exposed groups (or the errors in misclassification of exposure are different in diseased and non-diseased groups)

In general, random misclassification tends to bias results towards the null, meaning that the observed association in the data underestimates the magnitude of the association that would exist without this bias. On the other hand, non-random misclassification can lead to underestimates or overestimates of the true association between an exposure and a disease.

The challenge to the epidemiologist is to identify the potential sources of bias in a study, to indicate the potential direction of the bias (would it likely lead to an underestimate or an overestimate of a measure of association), and report some indication of its magnitude on its effect (the magnitude of the underestimation or overestimation in the reported measure of association).

Confounding refers to the existence of a third factor that has different distributions in the exposed and non-exposed groups and is also a risk factor (or a determinant) of the disease. For example suppose that the 2181 smokers in the above table are younger than the 2253 non-smokers. Younger people have lower risk (incidence) for developing CHD, leading to lower prevalence of CHD. Hence, the lower prevalence among the smokers in the table is not due to smoking but to the younger age of the smokers. The topic of confounding will be discussed in a future lecture.

Chance refers to sampling variability in the selection of subjects for a study (a sample) from a larger population of potential subjects. For example, the 2181 smokers in this study can be considered as a sample of a larger population of smokers who could have enrolled in the Framingham Heart Study. Although we expect the prevalence of CHD within a sample of subjects will estimate the prevalence of CHD in the larger population, the estimate from one sample may overestimate or underestimate the prevalence of CHD in the population.

In addition to these potential reasons for the observed association, there is another possible explanation for an association like this in a Cross Sectional study:

6. The disease outcome may influence the incidence of the exposure (**reverse causation**).

Perhaps the most likely explanation for the lower prevalence of CHD among smokers, is the reason for the lower prevalence of smoking among cases of CHD, compared to non-cases. It is very likely that smokers who developed CHD stopped smoking soon after the disease was diagnosed.

Examples of Survey Data Sets

Cross Sectional studies are often based on routinely collected survey data. For example the **National Center for Health Statistics (NCHS)** is part of CDC and performs both annual and periodic survey in this country through personal interviews or examinations and by data collected from vital and medical records. Four major survey programs of the NCHS are

1. **National Health and Nutrition Examination Survey (NHANES)**
2. **National Health Interview Survey (NHIS)**
3. **National Health Care Surveys** (survey of health care providers and organizations
4. **National Vital Statistics System (NVSS)** (records information on births and deaths

Information and the National Health and Nutrition Examination Survey (NHANES) can be found at http://www.cdc.gov/nchs/nhanes/about_nhanes.htm . This site contains both a video history of the study and a video tour of the Mobile Examination Centers that used as part of this survey. NHANES assesses health and nutritional status of adults and children in the US. It involves a representative sample of 15 counties of the Untied States with 5000 people each year. It collects data by both interview and examination, using a Mobile Examination Centers (MEC) involving 4 connecting trailers. Limited data from this survey is publicly available for analysis.

Information on the National Health Interview Survey (NHIS) can be found at http://www.cdc.gov/nchs/nhis.htm. It interview members from a representative survey of households and involves a multi-stage sampling scheme to identify these households. First, Primary Sample Units (PSU) are chosen, comprising counties and metropolitan areas of the nation. Next, approximately 35,000 households are chosen within PSU for an interview.

Finally, the data used above to describe an ecologic study was obtained from reports for the **Behavioral Risk Factor Surveillance System (BRFSS).** Information about this survey can be found at http://www.cdc.gov/brfss/. The BRFSS is a state-based system of telephone health surveys, and more than 350,000 adults are interviewed each year as part of this survey, making it the largest telephone health survey in the world.

Experimental Studies

Two major categories of epidemiologic studies are:

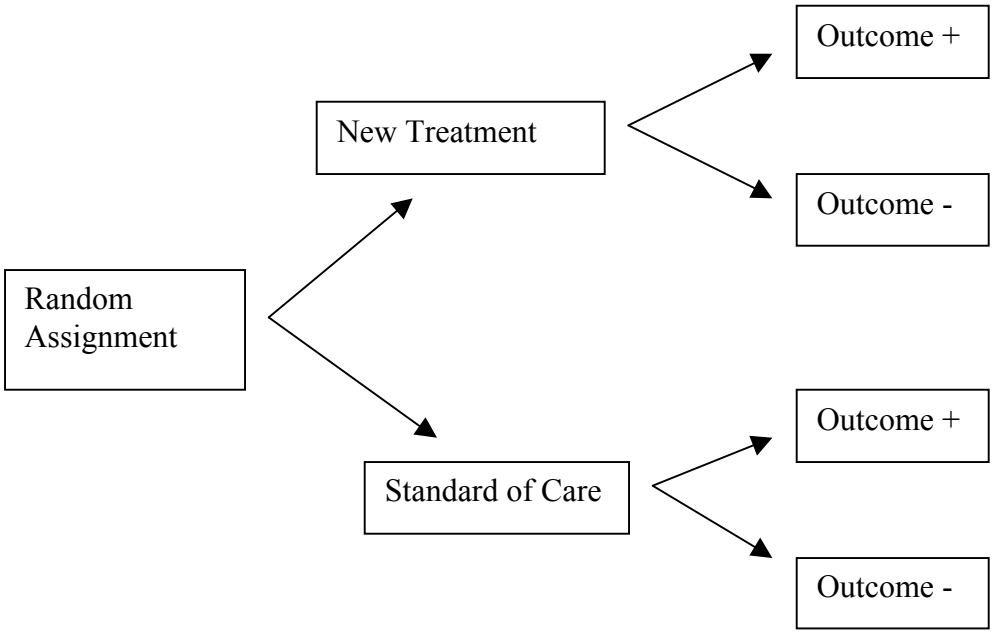
1. Experimental Studies
2. Observational (Non-Experiment Studies

The key distinction between these two types of study designs is the role of the investigator. In an experimental study, the investigator has the active role of assigning subjects to treatment (exposure) groups for the primary purpose of evaluating the effect of that treatment on an outcome. On the other hand, in observational studies exposures (including treatments) are self-selected or determined by someone else (e.g. one's physician) for a primary reason other than evaluating the effect of the treatment. The role of the investigator is passive, to observe outcomes and measure the association between an exposure and an outcome.

For example, suppose and investigator wished to evaluate the effect of daily consumption of a low dose aspirin on the risk of developing a myocardial infarction (heart attack). In an experimental study the investigator might assign some subjects to a regimen of daily low aspirin and assign others to a placebo drug. Usually the assignment is dictated by some type of randomization, whereby each subject has an equal chance of assignment to either group. The use of a **placebo** and **randomization** will be discussed later in these lecture notes.

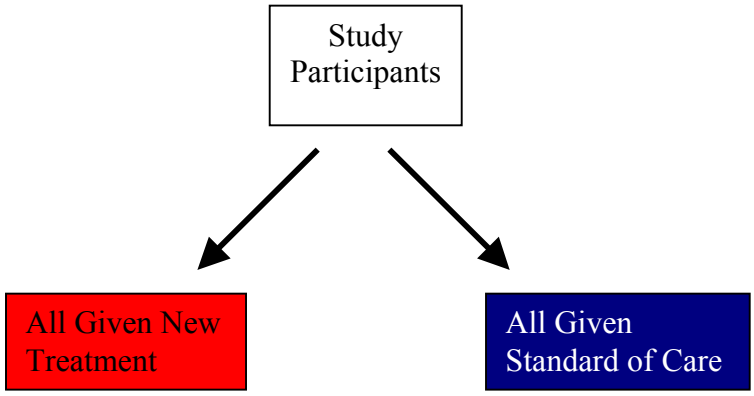
In an observational study the investigator might enroll a series of subjects who report taking daily aspirin and a comparison group of non-aspirin users. The reasons for aspirin use in the first group might because of a personal decision (subjects anticipating some benefit from daily aspirin use) or at the recommendation by one's physician (for disease prevention). One concern with such a study is the group taking aspirin might differ with respect to many indications (confounders) that are risk factors for the outcome. The topic of **confounding** will be discussed in future lectures in this course.

The type of experimental study described above is often referred to a randomized controlled trial or a **randomized clinical trial (RCT)** if it performed in a clinical setting. The following figure describes the basic principles of a Randomized Clinical Trial comparing outcomes for patients who are randomized to receive a new treatment versus receiving standard of care.

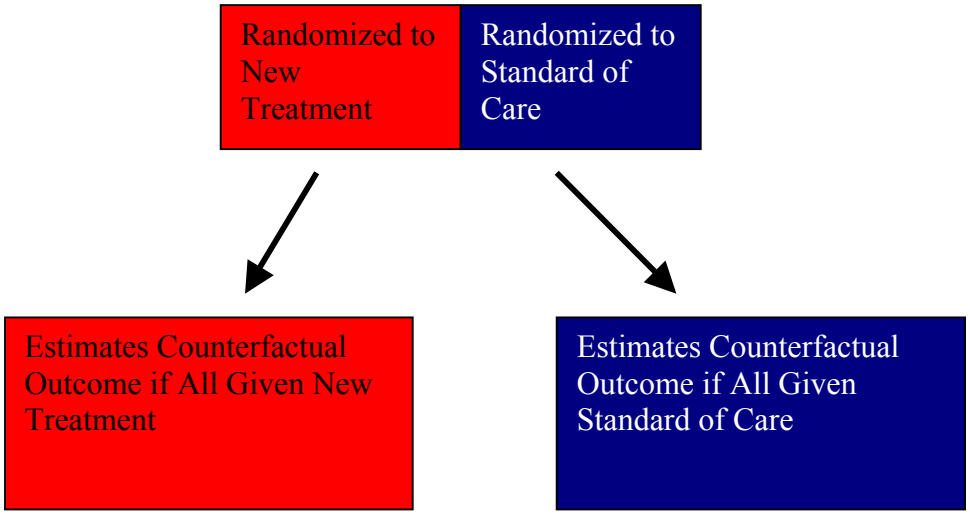


RCT and Causal Inference

Recall in a previous lecture we described the causal effect of a treatment in a group of study participants to be the difference between counterfactual outcome when everyone is given the treatment (Y^1) and when no one is given the treatment (Y^0). This is depicted in the following figure, where the alternative to receiving a treatment is receiving a standard of care rather than no treatment at all.



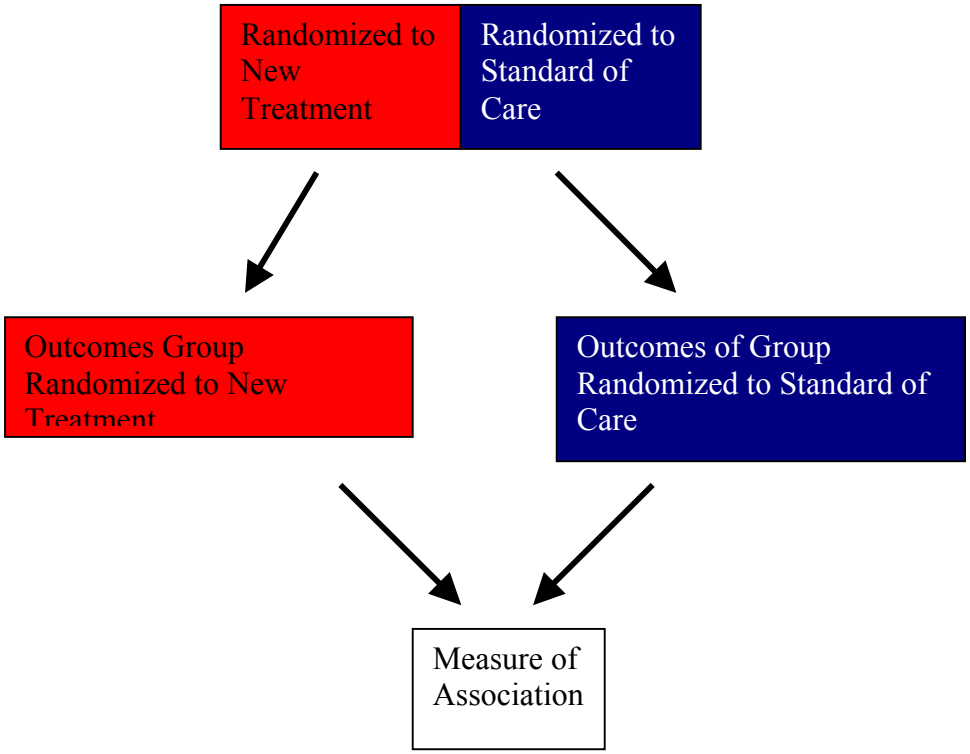
As mentioned previously, the practical problem in measuring the causal effect of a treatment is that only one counterfactual outcome can be observed for any study participant. However, subjects who are randomized to receive the new treatment can be considered a **random sample** of all study subjects. Therefore, their factual outcomes can estimate the counterfactual outcomes for all study subjects, if everyone received the new treatment. Similarly, the subgroup of subjects who are randomized to receive the standard of care can be considered a random sample of all study subjects, and their factual outcomes estimate the counterfactual outcomes of all study participants if none were received the new treatment (i.e. all received the standard of care). This is depicted in the following figure



It follows that a Measure of Association comparing the actual (factual) outcomes in the two comparison groups will estimate the causal effect of the treatment, provided that

1. there is no confounding
2. there are no bias

There should be no confounding for large studies because of randomization, as discussed later in these notes. However, failure to comply with the assigned treatment (**non-compliance**) and biased recording of outcomes are still potential problems that could result in a measure of association not reflecting the causal effect of the treatment.


Randomized to New Treatment
Randomized to Standard of Care
Outcomes Group Randomized to New Treatment
Outcomes of Group Randomized to Standard of Care
Measure of Association

Ethics

Since the investigator assigns treatment, usually at random, in an experimental study, ethical issues are a major concern. If the assignment was based on a random coin flip then approximately half of subjects will receive the new treatment and the other half will receive the standard of care. Therefore, the investigator must have enough confidence in the new treatment to justify giving it to half of the study participants; while at the same time have enough confidence in the new treatment to withhold it from the other half of the patients. This balance opinion about the risk and benefits of a treatment is referred t**o as equipoise.**

Imagine that you have just suffered a myocardial infarction and have been rushed to the emergency room of a nearby hospital. How would you react if you observed a physician reaching into his/her pocket and flipping a coin to make a treatment decision. One would hope that treatment decisions are grounded on solid clinical judgment; however a coin flip is a reasonable action to take when a treating physician is in a state of equipoise.

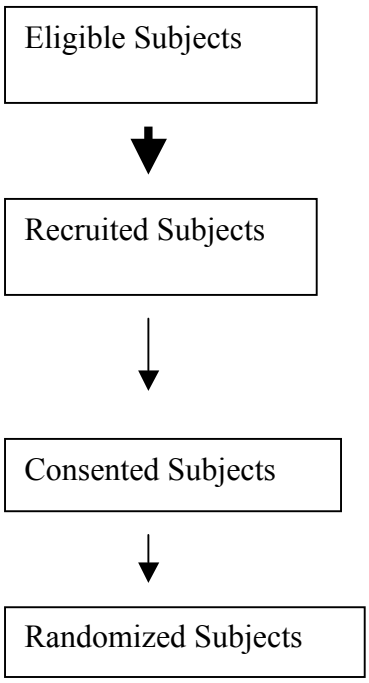
The investigator of an experimental study needs to be in a state of equipoise to justify using a random device to determine treatment assignments. However, others may also need to be in equipoise. For example, if the investigator needs permission from a treating clinician to approach his/her patients to potential enrollment in the study, then the treating clinician must also be in equipoise to grant permission. Often the window of opportunity for performing a RCT is narrow. Once a treatment becomes well accepted as a standard of care, a treating physician may not agree to let his/her patients receive this care.

In addition, Internal Review Boards of a hospital require that an investigator obtain informed consent of each study participant in the study. This involves informing the patient of the benefits and harms of the various treatment options in the study. Hence, the study participants also need to be in a state of equipoise to agree to have equal chances of receiving either treatment option.

Finally, most RCT's are monitored by a Data Safety and Monitoring Board (DSMB). The role of this board is to project the safety of the study participants, Once evidence from the data or external knowledge convince the members of the superiority of one of the treatment options, the members are no longer in a state of equipoise to justify continuing the study.

Efficacy and Effectiveness

Participants in a RCT are often very different for patients who are not eligible for the study. Inclusion and exclusion criteria for recruitment into the study may limit the participants of the study to a small and fraction of eligible subjects. This is described in the following figure



Eligible patients may be all patients diagnosed with a particular disease but only those from a specific set of hospitals might be recruited for the study. Only a portion of the recruited subjects may consent to enroll in the study. Finally, a rule-in phase (discussed later in these notes) may further restrict who might enroll in the study. Each layer of selection may limit the investigator's ability to generalize the results of the study to a wider audience of patients.

In addition the careful monitoring of participants may cause the treatment that patients receive in a clinical trial to be different to what is received outside of the trial, even among patients receiving the standard of care. This means the effect of the treatment in a closely monitored clinical trail (**Efficacy**) might be different from the treatment's effect in actual practice (**Effectiveness**).

Randomization

The main motivation for randomization is to remove the effect of participants and their physician in treatment decision. Random assignment is not influence by patients' characteristics. Therefore, especially in a large randomized trail, the investigator should expect the comparison groups to be comparable with respect to other factors that might influence the risk of developing the outcome. However, for small trials it is possible that some imbalance in the distribution of risk factors may occur and require adjustment in the analysis.

Simple randomization implies that the probability of treatment assignment for one study participant is not influenced by that of another participant. Although randomization is usually determined by a table of random number or a computer generated series of random numbers, simple randomization could be thought of as determined by a coin flip for each participant. The outcome of a coin flip for one participant does not influence the outcome of another participant. An interesting web site for obtaining a series of random assignments under the appearance of coin flips can be found at http://www.random.org/coins/. One limitation of simple randomization, although the statistical expectation if for equal numbers of participants in each comparison group, this might not happen because of the randomness of the coin flip.

An alternative to simple randomization is **block randomization**. Here subjects are assigned to treatment groups in blocks with the condition that equal numbers of participants within each block may be assigned to each treatment group. For example, if the treatments under consideration are labeled A and B, then possible sequences of treatment assignments with a block size of 4 are

AABB ABAB ABBA BBAA BABB BAAB

Treatment assignment begins with a random selection of one of these six possible sequences of treatment. This determines the treatment assignment for the first four participants in the study. Additional treatment assignments are determined by repeating this process. Each time one of the six possible sequence assignments is selected at random to determine the assignments of the next four participants.

One potential drawback of block randomization is the potential to determine the treatment assignment of the next participant who enrolls in a study. For example, knowledge that a block size is four along with knowledge of the treatment assignments of the first three participants in the study completely specifies the treatment assignment of the four subjects in the study. If such knowledge were know to a potential participant, his physician, or the investigator might impact the enrollment of that participant if the treatment assignment was not the one that was desired. This potential problem can be avoided by blinding everyone (except for the person performing the randomization) from the block size or by changing the block size during the enrollment period.

Blocking almost guarantees equal number of participants in each treatment arm in an experimental study. For example, if a block size of four is used, then at the completion of the enrollment process, one of the treatment groups can have at most two more participants than the other treatment group.

Blocking is often used along with stratification to essentially guarantee the similarity of the distribution of a risk factor in the treatment groups. Suppose a study was performed at multiple sites and an investigator wished to have similar distribution of study site within each treatment group to balance the baseline care offered at each group. This can bed accomplished by using a separate block randomization within each site

(stratum). For example the following table demonstrates the use of stratification and blocking in a study performed at three sites.

Study Site	Number of Participants	Selected Assignment Sequence
I	12	ABBA
		ABAB
		BABA
II	8	AABB
		ABBA
III	4	BBAA

The following table displays the distribution of Study Site within each treatment group after the completion of the enrollment period

Study Site	Treatment A	Treatment B
I	6	6
II	4	4
III	2	2
Total	12	12

Blinding

As mentioned above, a measure of association from a RCT estimates the causal effect of a treatment in the absence of confounding and bias. Randomization reduces the potential for confounding, especially in large trials. It also implies that each treatment group is a random sample of the study population, reducing the potential for selection bias at enrollment. However, post randomization selection-bias can still occur due to losses-to-follow-up. In addition, measurement bias can occur in a RCT because of non-compliance with treatment assignment or with outcome detection. **Blinding** is one means for limiting the potential for these biases.

Blinding refers to masking knowledge of the treatment assignment from individuals who might influence the compliance with treatment or the detection with the outcome. Blinding the investigator reduces the potential from having treatment assignment influence the detection and recording of outcomes. The potential for a detection bias is less of a problem for a “hard” outcome like death, but greater for a somewhat subjective outcome like quality of life or performance status. Treatment assignment should not be related to such errors made by the investigator. If any misclassification of outcomes does occur, it is likely to be of the non-differential type.

Blinding the study subjects also eliminates the potential for errors in self-reported outcomes to be related to treatment assignments. In addition, a study subject with a preference for one of the treatment options might be less likely to comply with the assigned treatment is he/she knew that it was not the preferred treatment. This, blinding study subjects may also reduce the potential for non-compliance.

Blinding other physicians who are providing ancillary care to study subjects may also limit the potential prescription of other therapies that might influence compliance of assigned treatment by the study subjects.

In non-clinical trails examining the effect of a preventive agent, the comparison to treatment is often no treatment. In such instances, a **placebo** agent is often used to enhance the blinding of study subjects. A placebo may also be used in a clinical trial, but a standard of care is typically chosen to evaluate the effect of a new treatment because of ethical considerations. A placebo pill resembles the active treatment in appearance but contains no active agents that should influence the outcome. The use of a placebo reduces the potential for knowledge of treatment assignment influencing self-reported outcomes.

Awareness of participating in a RCT may affect the behavior of the study subjects and even influence their outcomes. A series of studies conducted in the 1920s and 1930s at the Hawthorne Plant of the Western Electric Company demonstrated that changes in workplace behavior because of subjects' awareness of participating in a study (Am J Soc 1992;98:451-468). These finding gave rise to the notion of a "Hawthorne Effect", implying that study participants may change behavior during a study and that such changes may influence outcomes that are observed. This is related to the notion of a "placebo effect", which implies that compliance, even with an assigned placebo, may also influence outcomes. Perhaps the strongest evidence to support a placebo effect is found in the results of the Coronary Drug Project (N Engl J Med 1980; 303 -1038-41). This was an RCT comparing the effect of clofibrate (versus a placebo) in the long term treatment of coronary heart disease. Some of the results from this study are shown in the following table.

	5- Year Mortality Risk	
Clofibrate	20.0%	
Compliers		15.0%
Non-Compliers		24.6%
Placebo	20.9%	
Compliers		15.1%
Non-Compliers		28.3%

The first column of data show little overall benefit form Clofibrate (20.0% versus 20.9% mortality risk). On the other hand, the top part of the second column suggests a clear reduction in mortality among compliers of Clofibrate versus non-compliers (15.0% versus 24.6%). However, examining results in the comparison group shows that this reduction is not be attributable to a benefit of Clofibrate but rather reflect the change in the behavior of the participants due to awareness of being in the study and complying

with study medication. The placebo effect is demonstrated by observing a similar reduction in outcome among placebo compliers versus not compliers (15.1% versus 28.3%).

Run-In Phase

Another means to increase assigned treatment compliance is to enroll only participants who are likely to adhere to their assigned treatment. The identification of such subjects might be accomplished during a pre-randomization trial period (**Run-In Phase**), where potential subjects are given a trial period of active or placebo treatment. Non-compliers during this period would be likely to remain non-compliers after randomization, and therefore are not enrolled in the actual trial. The use of a Pun-In Phase will be discussed in more detail later in these notes when referring to the Physicians' Health Study.

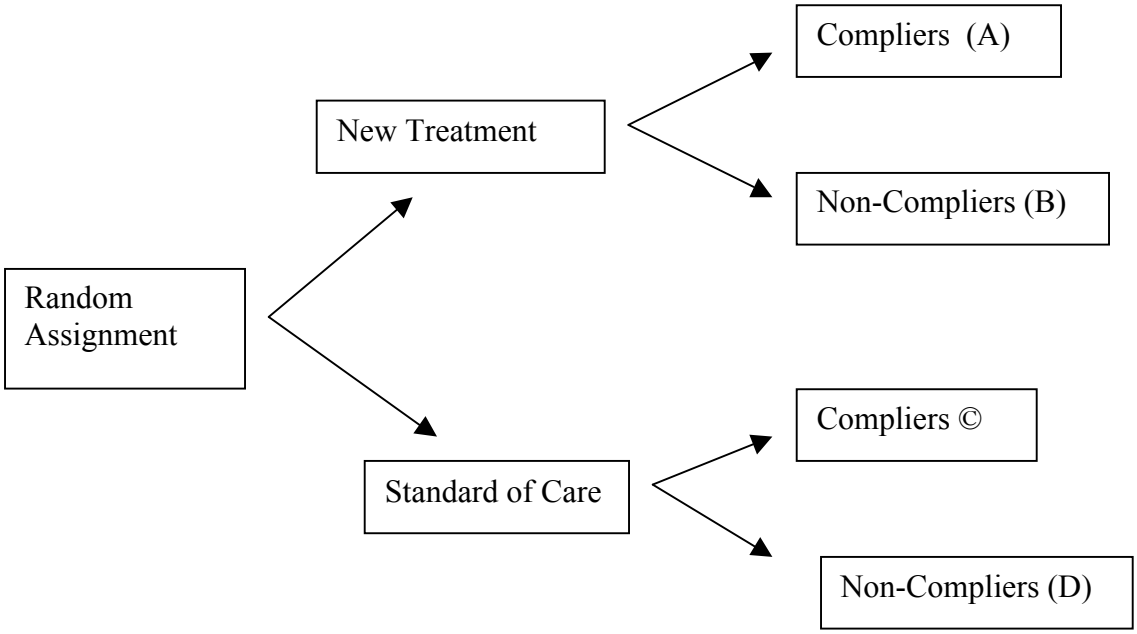
Data Safety Monitoring Board

The role of a Data Safety Monitory Board (DSMB) in a RCT is to monitor and protect the safety of the study participants. This board is usually comprised of outside scientists who are not involved in the study at hand. The DSMB can recommend termination of the study when it is no longer in a state of equipoise and feels that the study participants would be better served with the superior treatment. Possible reasons for terminating the study early include;

1. External evidence of the superiority of one of the treatments
2. Convincing internal information form the data of the superiority of one treatment
3. Severe side effect from one treatment
4. Problems with the operation of the study (including poor recruitment, poor retention of subjects, poor compliance of study subjects, and poor quality of the data
5. Convincing evidence of no likely treatment effect if the study were to continue

Analytic Issues

The following figure displays the potential comparison groups for evaluating the effect of a new treatment in a RCT.



An Intention to Treat (ITT) Analysis compares outcomes for subjects who are randomized to receive a new treatment (A+B) to those who are randomized to receive a standard of care (C+D). This analysis capitalizes on the anticipated effect of randomization to provide two groups that are comparable with respect to other risk factors for the outcome and therefore should be free of confounding (especially in large studies). However, the presence of non-compliers in the treatment group (B), weakens the ability to detect an effect of the new treatment as not all subjects in the group who assigned to the new treatment (A+B) will actually receive the benefit of the new treatment.

An **On-Treatment Analysis** compares outcome only among those who comply with treatment assignment (i.e. compares outcomes in groups A versus C). This analysis might have the advantage over the ITT analysis to detect an effect of the treatment since all subjects in this analysis are compliers. However, this analysis no longer has the advantage of randomization to provide comparison groups that are free of confounding.

Table 1 in a RCT typically displays the distribution of demographic features and potential risk factors for the outcome in the two treatment groups. This table describes the study population and provides a point of reference for the generalizability of the results from the study. More importantly, this table shows the expected comparability of the distribution of risk factors in the two groups from randomization.

Table 2 of a RCT typically displays the overall effect of the treatment on the primary outcome(s). It is usually based on an Intention-to-Treat Analysis. Additional tables may be added to display the effect of treatment on secondary outcomes and the association between treatment and side effects. Finally, results from subgroup analyses,

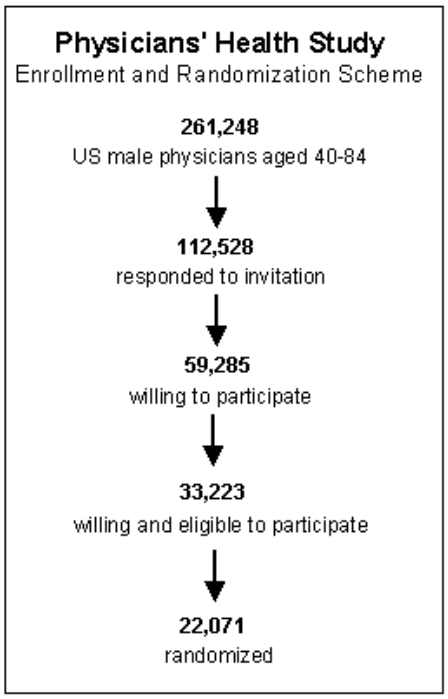
examining the effect of the treatment in subgroups of study subjects are typically performed. One criticism for subgroup analysis is that it may be prone to data dredging (searching exhaustively and reporting subgroups that demonstrate a benefit of the treatment) and problems of interpretation due to multiple testing. Therefore these analyses should be limited to hypotheses that are pre-specified prior to the initiation of the study.

Physicians' Heath Study

The **Physicians' Health Study (PHS)** was a randomized prevention trial with the primary aims to examine if

1. 325mg of aspirin taken every other day reduces cardiovascular disease mortality, and
2. 50 mg of beta carotene taken every other day reduces incidence of cancer

Information about the PHS can be found at http://phs.bwh.harvard.edu . In 1981 the investigators sent invitation letters, consent forms, and enrollment questionnaires to 261,248 male physicians, 40 - 84 years of age, living in the US and registered with the American Medical Association. The following table (provided by Julie Buring) displays the final enrollment numbers in this trial

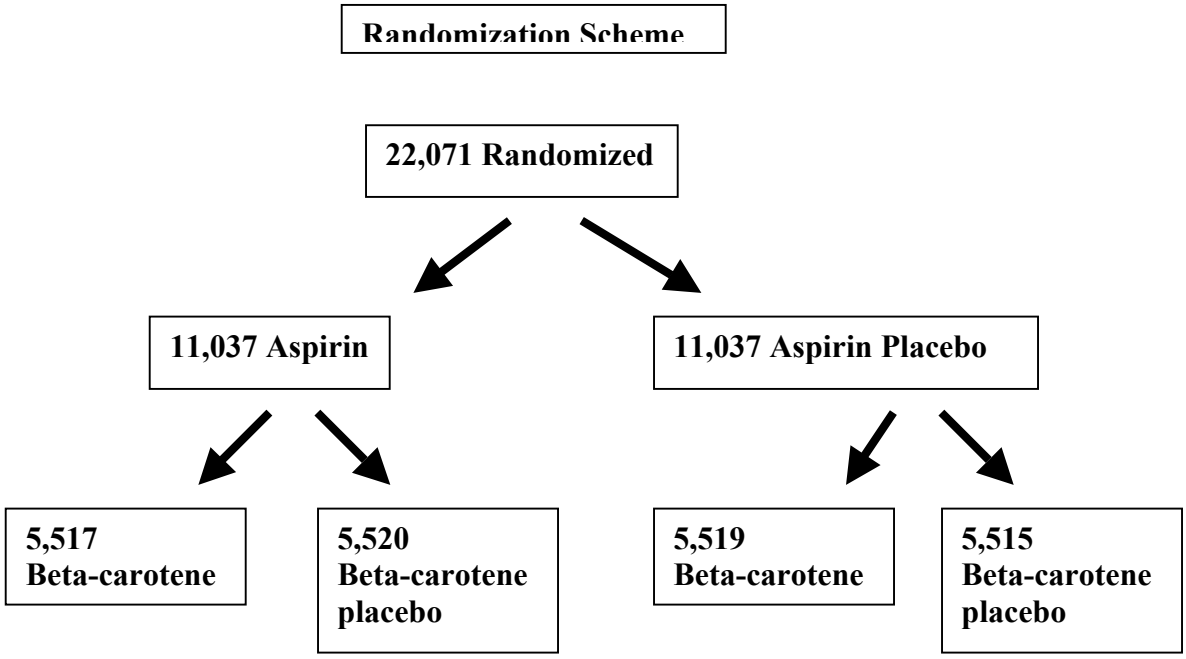


Only 112,528 physicians responded to the invitation, and 59,285 expressed a willingness to participate in the study. However, 26,062 were ineligible because if a variety of

reasons including a history of CVD or cancer; current renal or liver disease; peptic ulcer; gout; or contraindication to or current use of either aspirin or beta-carotene.

The remaining 33,223 were enrolled in a Run-In Phase during which all received active aspirin and placebo beta-carotene. After 18 weeks, participants were sent a questionnaire asking about their health status, side effects, compliance, and willingness to continue in the trial. 11,152 changed their minds, reported a reason for exclusion, or did not reliably take the study pills. This resulted in 22,871 remaining participants for randomization

The PHS uses a 2x2 factorial design whereby each subject underwent two levels of randomization as depicted by the following slide (provided by Julie Buring).

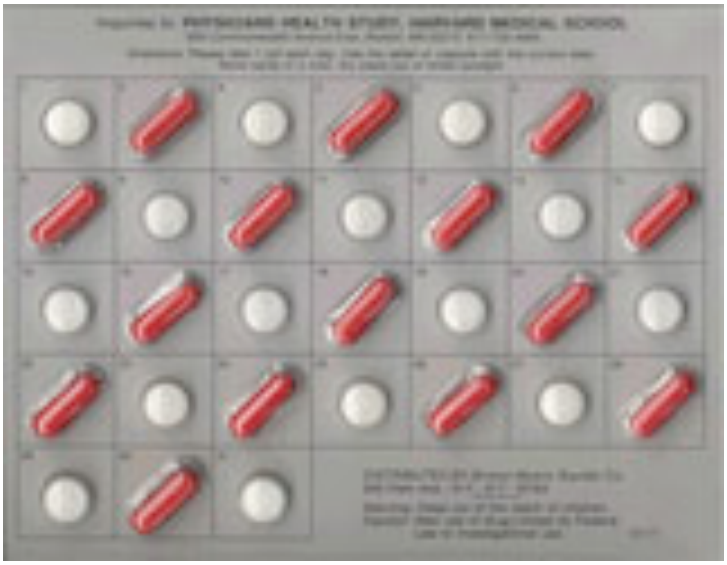


A factorial design allows for the estimation of multiple treatments on the same subjects in a single treatment. The following table displays the comparison groups to examining the two aims of the PHS.

		Treatment Option # 1 :Aspirin		
		Treatment 1	Placebo 1	Total
Treatment Option # 2: Beta-Carotine	Treatement2	A	B	A+B
	Placebo 2	C	D	C+D
	Total	A+C	B+D	Total

The effect of aspirin on CVD mortality can be examined by comparing CVD outcomes among the aspirin users (A+C) versus the aspirin-placebo users (B+D). Similarly the effect of beta-carotine can be examined comparing cancer outcomes among the beta-carotine users (A+B) versus the beta-carotine-placebo users (C+D). A potential problem of factorial design is if the effect of one treatment is modified by the presence or absence of the other treatment (effect modification). For example, if the effect of a treatment is manifested only among subjects given placebo for other treatment then the study may have reduced power for detecting this effect. The topic of Effect Modification will be discussed in a future lecture.

Participants were sent pill packets each month and instructed to take one pill each day. On odd days of the month participants would take either aspirin or aspirin-placebo. On even days of the month, participants would take either beta-carotine or beta-carotine-placebo. The following figure (provided by Julie Buring) displays a picture of the pill packets used in the PHS.



As expected in large RCT, the distribution of demographic features and risk factors were very similar in the comparison groups. This is demonstrated by the series of tables (provided by Julie Buring)

Assignment	**ASA + BC (n = 5,517)**	**ASA + PL (n = 5,520)**	**PL + BC (n = 5,519)**	**PL + PL (n = 5,515)**
Demographics				
Mean age (years)	53.2±9.5	53.3±9.6	53.2±9.5	53.2±9.5
Geographic region				
Northeast	27.8	27.9	27.3	28.8
Midwest	22.4	21.9	22.1	20.3
South or Southwest	26.8	26.1	27.0	27.9
Mountain or Far West	22.0	23.3	22.8	22.1
Other	1.0	0.9	0.7	0.9
Medical history				
Reported hypertension	11.8	11.8	12.4	11.5
Mean systolic BP	126.0±11.9	126.3±12.1	126.1±11.7	126.1±11.7
Mean diastolic BP	78.8±7.6	78.9±7.5	78.9±7.5	78.7±7.5
High cholesterol	7.9	7.5	8.0	8.0
Mean cholesterol (mg/dl)	211.7±43.9	212.3±45.0	212.7±46.5	211.4±44.8
Diabetes mellitus	2.1	2.5	2.2	2.2
Mean BMI	24.9±3.1	24.9±3.0	24.9±3.0	24.9±3.0
Overweight	13.5	14.0	13.6	13.6
History of angina pectoris	1.2	1.4	1.1	1.2
Parental history of Ml	12.7	13.3	12.7	13.6

Health habits				
Cigarette smoking				
Never	49.8	48.8	49.6	50.1
Past only	39.3	40.1	39.1	39.0
Current	10.9	11.1	11.4	10.9
Mean cigarettes/day	21.7±13.0	22.1±12.9	22.4±13.7	22.8±13.5
<15 (of current smokers)	28.8	27.4	27.3	24.6
15-24 (of current smokers)	32.5	33.7	34.7	35.4
25+ (of current smokers)	38.7	38.8	37.9	40.0
Current cigar smoking	4.7	4.2	4.3	4.3
Current pipe smoking	8.8	7.7	8.2	8.0

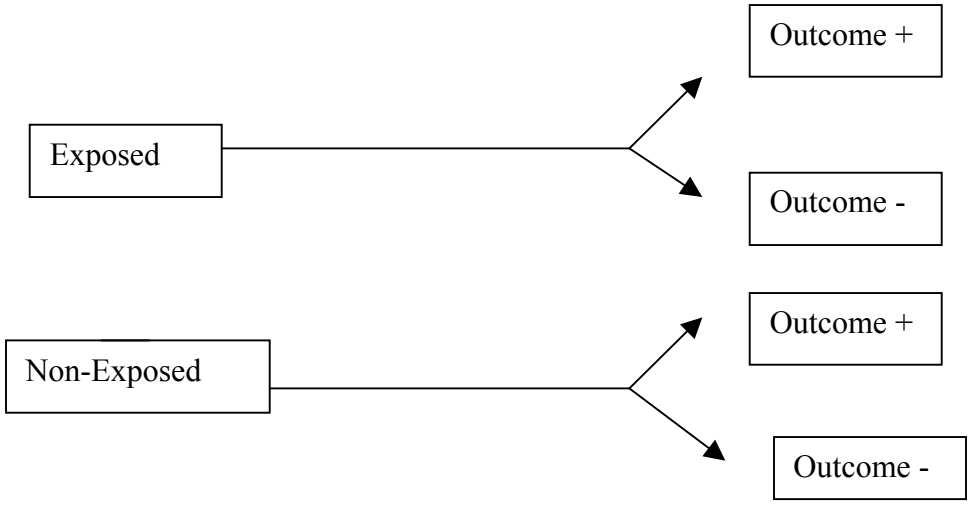
The DSMB stopped aspirin arm of the PHS in 1989 (ahead of schedule) because of clear effect of aspirin decreasing the risk of Myocardial Infarction (heart attacks). However, too few strokes or deaths occurred upon which to base sound clinical judgment regarding aspirin and stroke or mortality. In 1996 beta-carotene arm stopped showing neither benefit nor harm

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.



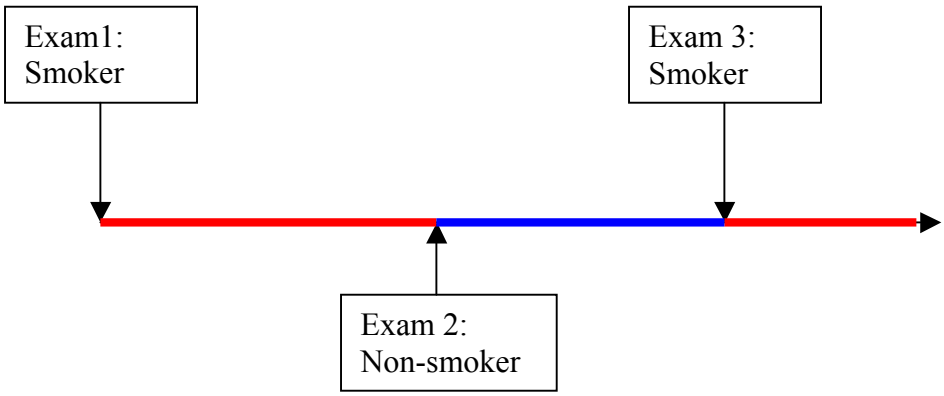
Suppose an investigator wished to study the association between smoking (exposure) and the incidence of a first CHD event. This would require enrolling subjects who are free of CHD and at risk for a first CHD event. There are numerous options for measuring the exposure of interest, smoking. Some are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

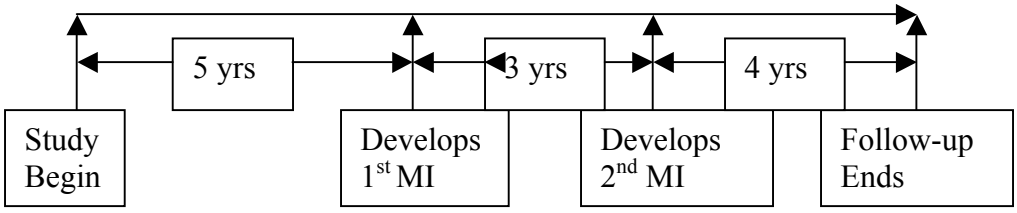
Having enrolled at-risk subjects and determined the appropriate method for measuring exposure, the investigator then needs to decide how to follow subjects over time and record outcomes. Follow-up is recorded in the Framingham Heart Study by having each subject return to an examination center every two 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:

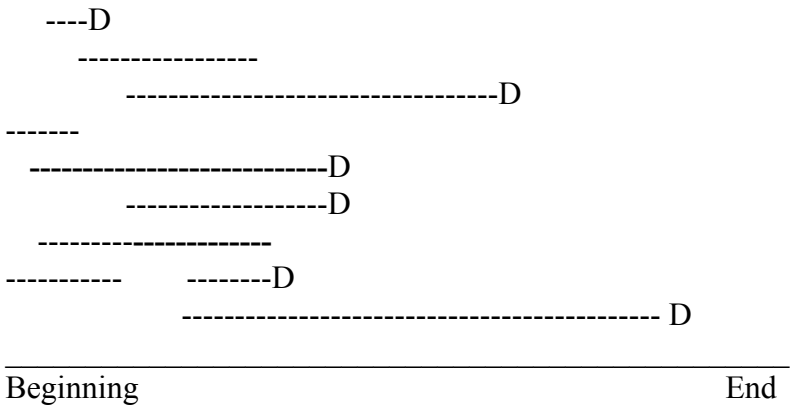
1. Cumulative Incidence
2. Incidence Rate

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.

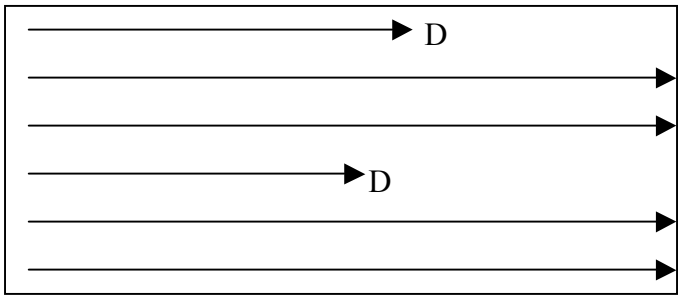
Open vs. Closed Cohort

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.

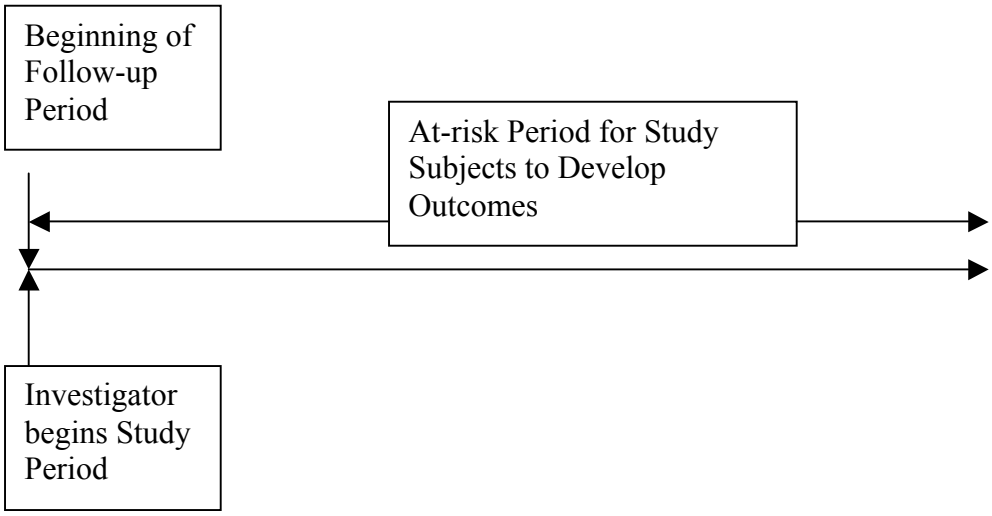


A Fixed Cohort is similar to a closed cohort with the exception that there are some subjects who are lost-to-follow-up. The Framingham Heart Study with the outcome of all-cause mortality and sex as an exposure is an example of a Closed Cohort as there is complete ascertainment of this outcome on all 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.

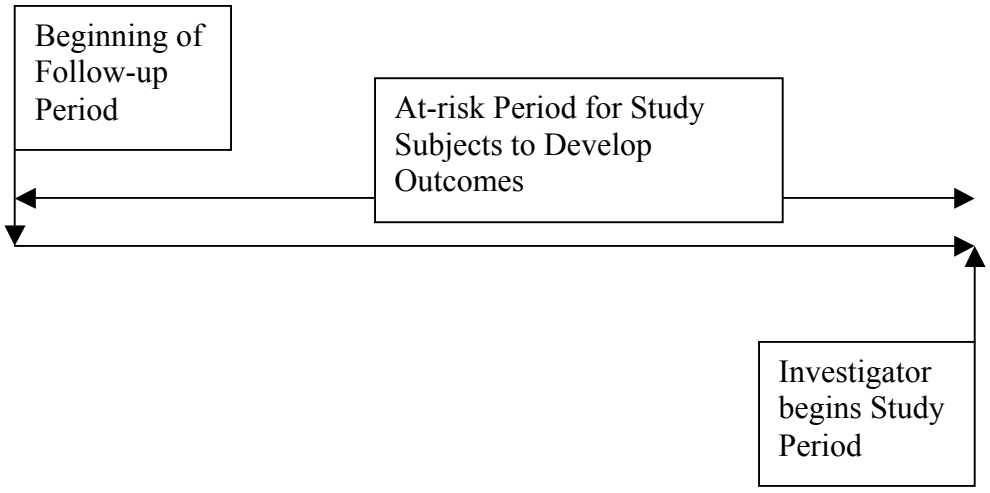
Prospective versus Retrospective Cohort Studies

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.

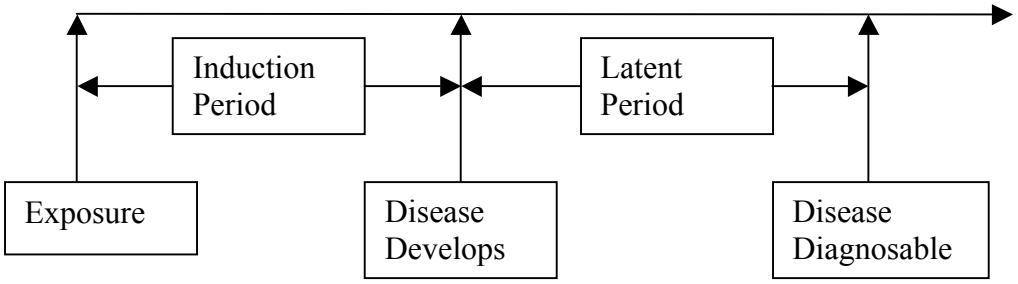
Given the growth of electronic medical records, the potential for creating historical cohorts and performing Retrospective Cohort Studies has increased. However, a limitation of a Retrospective Cohort Study is the quantity and quality of the existing data. On the other hand, the quality and quantity 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.

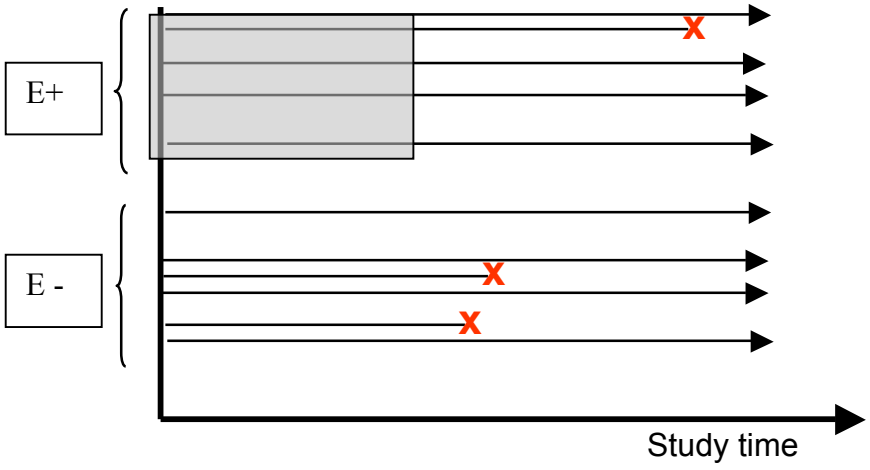
Induction Period

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.

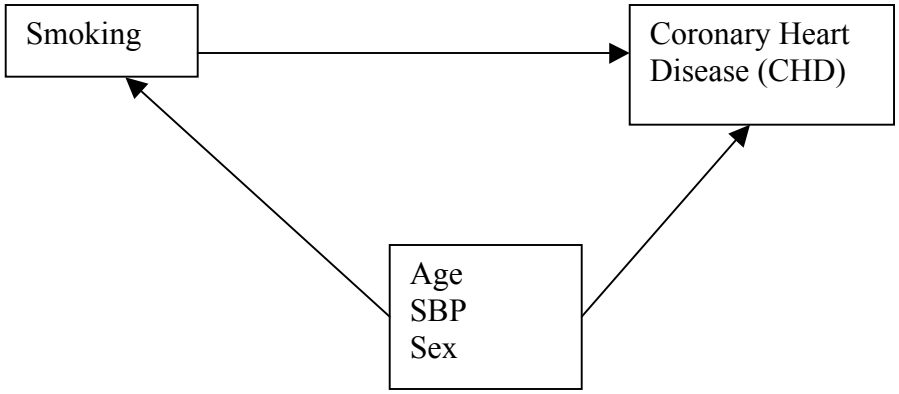
Confounding

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.

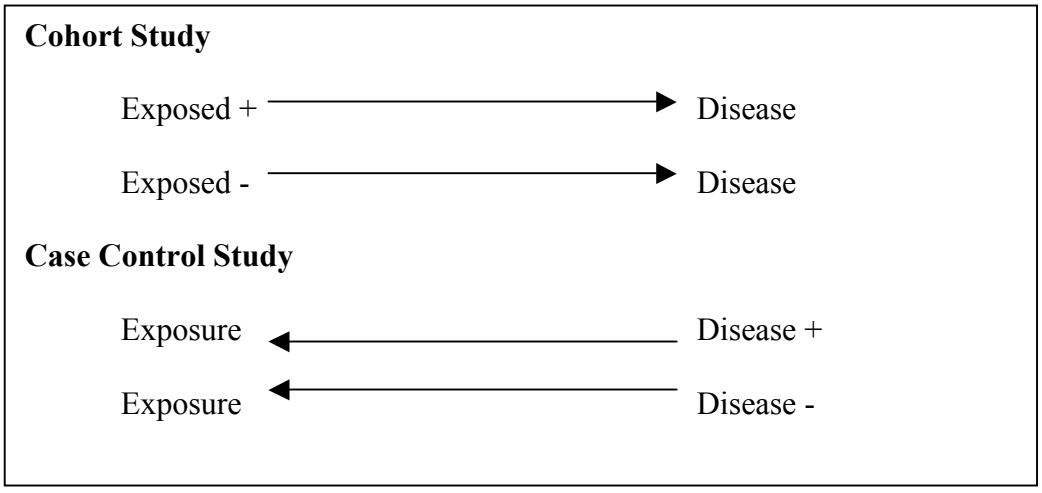
Other the other hand, Cohort Studies may not be efficient for study rare diseases because of the need to enroll large number of subjects and follow them for long periods to time to record enough cases of the disease. Prospective Cohort Studies can be expensive in terms of time and money. Retrospective 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.

Case Control Studies

A Prospective Cohort Study is not efficient for investigating a rare disease outcome because of the large number of study subjects and/or the long period of follow-up that are needed to obtain a sufficient number of cases of disease. In this situation the Case Control Study is a more efficient alternative design to consider.

Alvin Feinstein, a clinical epidemiologist, proposed the phrase "trohoc" to describe a Case Control Study. The word "trohoc is the reversed spelling of the word "cohort", reflecting the timing relationship between a Case Control Study and a Cohort Study.

The classical description of a Case Control Study is a study that compares previous exposure histories among a group of study subjects who have the disease in question (cases) and a group of subjects who do not have the disease (controls). This description and its relationship to a Cohort Study are depicted in the following figure:



Two important questions related to this description of the Case Control Study design are:

1. What criteria should be used to select the controls?

2. Does comparing exposure history among cases and control result in a measure of association that estimates the causal effect of the exposure on the disease, as in a Cohort Study?

To address the first question, suppose an investigator wished to examine the effect of oral contraceptive use on the risk of breast cancer with a Case Control Study. According to the above description the investigation would enroll cases (women with breast cancer) and controls (women without breast cancer). Suppose the cases were women diagnosed with breast cancer at local hospitals and, for convenience, the

investigator wanted to enroll controls from the same hospital. Would any group of women without breast cancer be appropriate controls? For example, would newborn baby girls in the nursery be an appropriate control group? They are free of breast cancer but almost everyone would agree that they are not appropriate for examining the effect of oral contraceptive use on the risk of developing breast cancer. Therefore, some other characteristic is needed to define an appropriate control group.

The answer to what is an appropriate control group lies with the link between Case Control Studies and Cohort Studies. This link is formed by considering the cases in a Case Control Study to be the outcomes from a corresponding Cohort Study. Sometimes this is true by definition, if the cases were taken from a registry of outcomes in a previously documented Cohort Study. However, in many situations the cases are selected from a hospital or health care plan and were not part of a previous performed Cohort Study. Nevertheless we can still entertain the notion that a cohort of subjects existed in the past and if it were followed over time, then its outcomes would be the cases in our Case Control study. Under this assumption, the role of the controls in a Case Control Study is to provide in estimate of the prevalence of exposure in that Cohort Study. If this holds, then comparing previous exposure history among cases and controls yields estimates of measures of association that were previously discussed for Cohort Study.

The demonstrate the link between the controls in a Case Control Study and a corresponding Cohort Study, consider the following table, displaying data from a Case Control Study:

	Case	Control
Exposure +	a	b
Exposure -	c	d
Total	M_1	M_0

The usual measure of association from a Case Control Study is the Exposure Odds Ratio, comparing the odds of exposure among the cases (a/c) to that of the controls (b/d).

$$\text{Exposure Odds Ratio} = \text{EOR} = (a/c) / (c/b)$$

Mathematically, many of the measures of association from Cohort Studies can also be expressed as ratio of exposure odds. For example, the following table displays the results from a closed Cohort Study and the formulas for calculating two common measures of association: the Risk Ratio and the Disease Odds Ratio.

	Disease		
	+	-	Total
Exposure +	A	B	N_1
Exposure -	C	D	N_0

Risk Ratio = RR = $(A/N1)/(C/N_0)$
= $\mathbf{(A/C)/(N_1/N_0)}$

Disease Odds Ratio= DOR = $(A/B) / (C/D)$
= **(A/C) / (B/D)**

The second equation for the Risk Ratio (RR) demonstrates that is can also be calculated by dividing the odds of exposure among the cases of disease (A/C) by the odds of exposure in the source population (closed cohort) (N_1/N_0).

Also, by the symmetry of the odds ratio, the Disease Odds Ratio (DOR) is equal to the ratio of the odds of exposure among the cases (A/C) divided by the odds of exposure among the subjects who did not develop the disease (B/D).

Similarly the following table displays the results from an open cohort and the calculation for the Rate Ratio

	Disease	Person-Time
Exposure +	A	K_1
Exposure -	C	K_0

RR = $(A/K_1)/(C/K_0)$
= $\mathbf{(A/C)/(K_1/K_0)}$

The second equation for the Rate Ratio demonstrates that is can also be calculated by dividing the odds of exposure among the cases of disease (A/C) by the ratio of exposed to non-exposed person-time in source population.

Control Selection

The role of the controls in a Case Control Study is to estimate the prevalence of exposure in the Cohort Study whose outcomes would be the cases at hand. From the above formulas, this allows the Exposure Odds Ratio from a Case Control study to estimate common measures of association from the corresponding Cohort Study. This cohort could be closed or open.

Corresponding Cohort Study: Closed Cohort

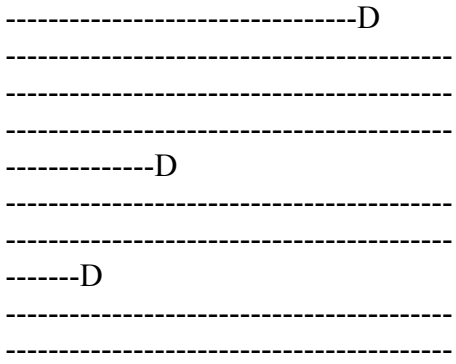
If the corresponding cohort is closed, then the selection of controls is often referred to as cumulative incidence sample and there are two options for selecting controls:

1. Selecting controls from subjects in the cohort who did not develop the outcome during the period of follow-up (this is method used in the Classic Nested Case Control Study)

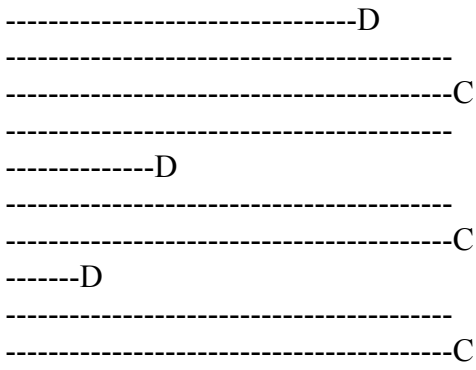
2. Select controls from everyone in the cohort at the beginning of the follow-up (this is an example of a Case Cohort Study).

These options are describes in the following figure:

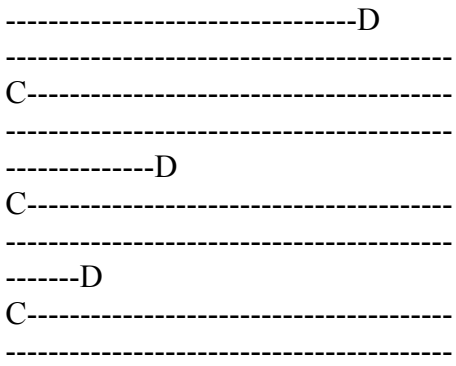
Closed Cohort (D indicates the development of disease)



Nested Case Control Study (C indicates Controls selected from non-disease group)



Case Cohort Study (C indicates Controls selected from full cohort)



Classic Nested Case Control Study

The following example describes the classic Nested Case Control Study (Willett. Lancet 1983 Jul 16;2(8342):130-4). The study involves data from the Hypertension Detection and Follow-up Program (HDFP). This was a previously performed randomized clinical trail that investigated different treatments for hypertension. However, 4480 participants in this RCT provided blood sample, which were frozen for future use. The RCT also created a registry that recorded the names of study subjects who developed cancer from 4480 participants.

111 of these subjects developed cancer and were chosen as the cases in a Case Control Study to examine the relationship between selenium levels and cancer. Controls should be chosen to reflect the prevalence of exposure in the in the corresponding cohort of 4480 subjects. However, in this classic Nested Case Control Study controls were selected from the members of the cohort who did not develop the disease (4480 – 111 = 4369 subjects). 210 controls were selected form these 4369 study subjects. The investigator measured selenium levels from the frozen blood samples of the 111 cases and the 210 controls. 57 of the cases and 84 of the controls had low levels of selenium. The results of the case control Study are displayed in the following table:

	Case	Control
Low Selenium	57	84
High Selenium	54	126
Total	111	210

EOR = [57/54] / [84/126] = 1.6

The following tables displays the results of the Cohort Study had the investigator measured the blood specimens for all 4480 study subjects.

	Cancer		
	Yes	No	Total
Low selenium	57	B	N_1
High selenium	54	D	N_0
Total	111	4369	4480

$RR = [(57/N_1)]/[(54/N_0)] = [57/54]/[N_1/N_0] = ?$

DOR= [57/B]/[54/D] = [57/54]/[B/D] = ?

The values for the Risk Ratio (RR) and Disease Odds Ratio (DOR) would require analyzing the blood specimen on the remaining 4369 subjects. If the selenium distribution of the 210 controls reflects the distribution of all 4369 potential controls then

the Exposure Odds Ratio (EOR) from the Nested Case Control Study will estimate the Disease Odds Ratio from the Cohort Study as demonstrated in the following calculation:

$$\begin{aligned} \text{DOR} &= [(57/B)]/[(54/D)] \\ &= [57/54] / [B/D)] \\ &\approx [57/54] / [84/126] = 1.6 = \text{EOR} \end{aligned}$$

Furthermore, since the disease is rare, the number of potential controls (4369) is almost the same as the number of subjects in the cohort (4480). Therefore, the odds of exposure among the 4369 potential controls (B/D) should be similar to the odds of exposure in the full cohort (N_1 / N_0). Under this **rare disease assumption,** it follows that the Exposure Odds Ratio approximates the Risk Ratio from this Cohort Study.

$$\begin{aligned} \text{EOR} &= [57/54] / [84/126] = 1.6 \\ &\approx [57/54] / [B/D] = \text{DOR} \\ &\approx [57/54] / [N_1 / N_0\} = \text{RR} \end{aligned}$$

Case Cohort Study

In the classic Nested Case Control Study controls are chosen from subjects who did not develop the disease in the corresponding closed cohort (4,369 subjects in the previous example). The Exposure Odds Ratio from the Case Control Study estimates the Disease Odds Ratio from the corresponding Cohort Study, and under the rare disease assumption also estimates the Risk Ratio from the Cohort Study.

An alternative option is to select controls from the 4,480 members of the original cohort. The resulting Case Control Study is usually referred to as a Case Cohort Study. The exposure odds among the selected controls (b/d) should estimate the exposure odds in the full cohort (N_1/N_0). Furthermore, the Exposure Odds Ratio from the Case Cohort Study estimates the Risk Ratio from the Cohort Study without any assumption about the rarity of the disease.

Since the outcomes in a Cohort Study at part of the at-risk subjects at the start of a study, it is possible disease case might also be selected as a control in a Case Cohort Study. This presents some problems in performing tests of significance and confidence interval estimation, but does not invalidate the Exposure Odds Ratio from the Case Cohort Study estimating the Risk Ratio from the Cohort Study.

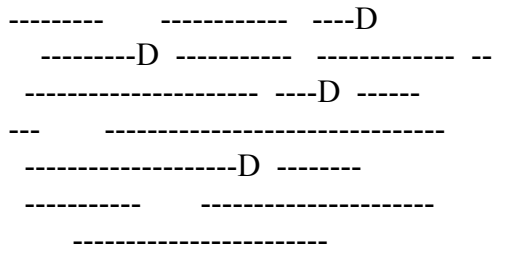
Corresponding Cohort Study: Open Cohort

Returning to a previous example, suppose that an investigator plans a Case Control Study examining the relationship between oral contraceptive use and the risk of developing breast cancer. Furthermore, suppose that the cases are women diagnosed with

breast cancer at a local hospital in the past two years. Since the purpose of the controls is to reflect the prevalence of exposure (oral contraceptive use) in the cohort study that gave rise to the cases, the challenge to the epidemiology if to formulate this cohort study and determine an appropriate control to describe the prevalence of exposure in that cohort.

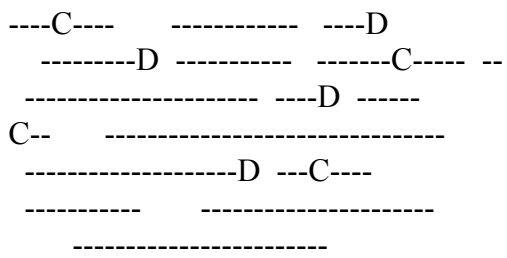
Since the cases are chosen from a single hospital, the corresponding cohort would be the population living in the catchment area of that hospital. Membership in this population may be defined by residential area and also by other factors such as a women's primary care physician and health plan that might influence her being referred to that hospital for testing and ultimately for the diagnosis of breast cancer. Unfortunately a list of women living in the population would not exist. However, if it did exist then it would probably be a dynamic population (open cohort) with women moving in and out of this population. A description of this open cohort is given in the following figure:

Open Cohort (D represent the development of disease)



Since the cohort is open, the appropriate measure of disease incidence is the Incidence Rate. Another term for an Incidence Rate is the Incidence Density. (proposed by Olli Miettinen), and the corresponding Case Control Study is called a Density Type Case Control Study. The controls are chosen so that their odds of exposure will reflect the ratio of the amount of person time in the open cohort that was contributed by oral contraceptive users to the amount of person time in the open cohort that was contributed by non-oral contraceptive users. Therefore controls should be selected from the person-years of the cohort study. The following figure displays this type of density type sampling for controls:

Density Type Sampling of Controls (C represents a selected control)



The measure of association in an open cohort study is the Rate Ratio (RR) as described in the following table

	Cases of Disease	Person-Time
Exposed	A	K_1
Non-Exposed	C	K_0

$$RR = (A/K_1) / (C/K_0)$$
$$= (A/C) / (K_1/K_0)$$

The display of data from the corresponding Density Type Case Control Study is

	Case	Control
Exposure +	A	B
Exposure -	C	D
Total	M_1	M_0

$$EOR = (A/C) / (B/D)$$

If the exposure odds among the controls (B/D) estimates the amount of person time in the open cohort that was contributed by exposed subjects divided by the amount of person time in the open cohort that was contributed by non-exposed subjects (K_1/K_0), then is follows that the Exposure Odds Ratio from the Density Type Case Control Study estimates the Rate Ratio from the corresponding Open Cohort Study

$$EOR = (A/C) / (B/D)$$
$$\approx (A/C) / (K_1 / K_0) = RR$$

Sources of Controls

If the cases in a Density Type Case Control Study are a list of all cases that develop in a geographical population (e.g. state of Massachusetts) then the corresponding open cohort is a census of individuals living in that population in the past. Such cases are referred to as **Population-Based Cases** and the selected are referred to as **Population-Based Controls.**

If the cases are chosen from one (or more) hospitals with a specified diagnosis, then they are referred to as **Hospital Based Cases**. Controls are typically patients selected from the same hospital but with a different diagnosis. The reason for this choice is the assumption that the cases and controls from the same hospital come from the same catchment area. Therefore the controls can be considered a sample from the catchment area. However, for the prevalence of exposure among the controls to reflect the prevalence of exposure in the catchment area, the diagnosis for the controls should not be

one that is caused or prevented by the exposure. For example, if the cases are women who are diagnosed with breast cancer and the exposure of interest is the use of oral contraceptive, then an inappropriate control group would be women diagnosed with venous thrombosis since this might be caused by oral contraceptive use. Controls selected from the same hospital as the cases are referred to as **Hospital Based Controls.**

Example: Density Type Case Control Study

The following results are from a hospital-based Density Case Control Study measuring the association between a series of potential risk factors and the development of Aortic Stenosis (Hoagland. Am J Med 1986;80(6):1041-50). Aortic Stenosis "is a disease of the heart valves in which the opening of the aortic valve is narrowed. The aortic value is the valve between the left ventricle of the heart and the aorta, which is the largest artery in the body" (http://en.wikipedia.org/wiki/Aortic_valve_stenosis).

The cases for this study were 105 subjects with Aortic Stenosis documented by cardiac catheterization (gold standard test). The suspected risk factors of interest (exposures) included smoking, diabetes, hypertension, cholesterol, and a family history of CHD. Three Control Groups were considered for this study:

1. **Group 1**: Patients who underwent cardiac catheterization, which showed no Aortic Stenosis but did show another type of valvular heart disease (n=110)

2. **Group 2**: Patients who underwent cardiac catheterization, which showed no Aortic Stenosis and no other type of valvular heart disease (n=170)

3. **Group 3**: Surgical patients whose reason for surgery was not known to be associated with risk factors of interest (n=269)

All data was obtained from medical record reviews. If no mention of a risk factor was indicated in the medical record, then it was assumed to be absent (i.e. non-exposed). This may results in a large potential for a misclassification bias.

The following table shows the relationship between hypertension and Aortic Stenosis using control Group 3.

	Case	Control
Hypertension	43	91
No Hypertension	62	178
Total	105	269

$$\text{EOR} = (43/62)/(91/178)$$
$$= 1.4$$

One limitation of a Case Control Study is that it does not allow for the estimation of exposure-specific risks or rates for developing the outcome. Since the investigator usually determines the relative sizes of the case and control groups, it follows that the overall prevalence of disease in the data does not reflect the incidence of the disease in the corresponding cohort study. For example, the prevalence of disease, P(D), among the exposed and non-exposed groups from the previous table is

P(Aortic Stenosis| History of Hypertension) = 43/134 = .32
P(Aortic Stenosis| No History of hypertension) = 62/240 = .26

These proportions are somewhat arbitrary and do not reflect the risk of developing hypertension. To demonstrate this, suppose that the investigator selected twice as many controls for the study. The expected results from this study are shown in the following table:

	Case	Control
Hypertension	43	182
No Hypertension	62	356
Total	105	538

EOR = (43/62)/ (91/178)
= 1.4

The value for the Exposure Odds Ratio does not change but the prevalence of Aortic Stenosis in each group changes to

P(Aortic Stenosis| History of Hypertension) = 43/225 = .19
P(Aortic Stenosis| No History of Hypertension) = 62/418= .15

Case Control Studies are sometimes referred to as “quick and dirty” studies. They are labeled as “quick” compared to prospective Cohort Studies in that the follow-up period for the study subjects has happened in the past. On the other hand, they are labeled as “dirty” in part because their potential for selection bias, due to the use of an incorrect control group. This may hold true for the first two control groups considered for this study.

Control Group 1 included patients who underwent cardiac catheterization, which showed no Aortic Stenosis, but did show another type of valvular heart disease. It is very possible that the same risk factors, which cause Aortic Stenosis, may also cause these other types of valvular heart disease. Therefore the exposure history in Control Group 1 may over estimate that for the source population

Control Group 2 included patients who underwent cardiac catheterization, which showed no Aortic Stenosis and no other type of valvular heart disease. It is very possible that the risk factors being considered as exposures in this study may have influenced the

decision for cardiac catheterization for Control Group 2. If so, then the exposure history in Control Group 2 may over estimate that for the source population. This is demonstrated by the suggestion of a protective effect of hypertension in the following table that uses Control Group 2.

	Case	Control
Hypertension	43	89
No Hypertension	62	81
Total	105	170

$$\text{EOR} = (43/62) / (89/81) = .63$$

Measurement bias is a potential in any study but may be a particular problem in the study at hand. All exposure information was recorded from medical records. Control Group 3 included surgical patients whose reason for surgery was not known to be associated with the risk factors of interest. Exposure information on the Cases, members of Control Groups 1, and members of Control Group 2 where obtained from interview by cardiology fellows at the time of cardiac catheterization, which would include detailed questions on Coronary Heart Disease (CHD) risk factors. On the other hand, subjects in Control Group 3 were interviewed by different hospital staff prior to surgery and may have had less detailed questions on CHD risk factors. For, example, it may be that subjects in Control Group 3 were not asked detailed questions about family history of heart disease or such information was not completely recorded in their medical records. This might explain the possible protective effect of this factor that is shown in the following table.

	Case	Control
Family History	42	53
No Family History	63	216
Total	105	269

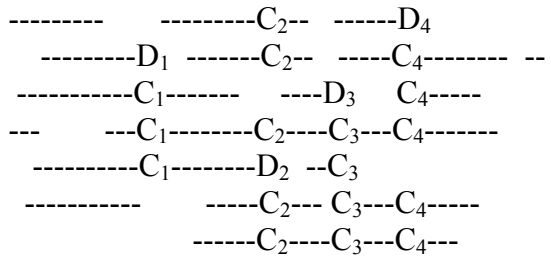
$$\text{EOR} = (42/63) / (53/216) = 2.72$$

Risk Set Sampling

A Case-Cohort Study is also an option when the corresponding when the cases are considered to be the outcome of an open Cohort Study. This would mean that cases has the potential for being selected as controls when the latter are selected to reflect the amounts of person-time from the exposed and non-exposed groups in the open cohort. For example, if Control Group 3 in the previous example were appropriate to reflect this information, then it is possible that this group may contain some cases of Aortic Stenosis since its members did not undergo cardiac catheterization.

Risk Set Sampling is another option for selection controls, in which the selected controls are matched the follow-up times of cases. The risk-set for a case is the members of the cohort study who were also at risk for developing the disease at the time a case developed the disease. Risk-set sampling involves selecting one of more members of that set as controls. The resulting matched analysis is similar to a survival analysis that could be performed on the full cohort. Risk-set sampling is depicted in the following figure.

Risk-Set Sampling of Controls (C_i represents a potential control for D_i)



Concept of Confounding

Suppose an investigator performed a Cohort Study to investigate the association between smoking and the risk of developing Coronary Heart Disease and found that the incidence of CHD among smokers was twice that of non-smokers (Risk Ratio = RR = 2.0). As noted previously before we can conclude that this measure of association reflects the causal effect of smoking, we need to rule out the alternative explanations of

1. Bias
2. Confounding
3. Chance

Suppose that the investigator is confident that there is little potential for bias in this study and that the large sample size limits the role of chance as possible reason for this association. However, a closer look at the data reveals that the smokers are much older than the non-smokers. Does the measure of association (RR = 2.0) reflect the effect of

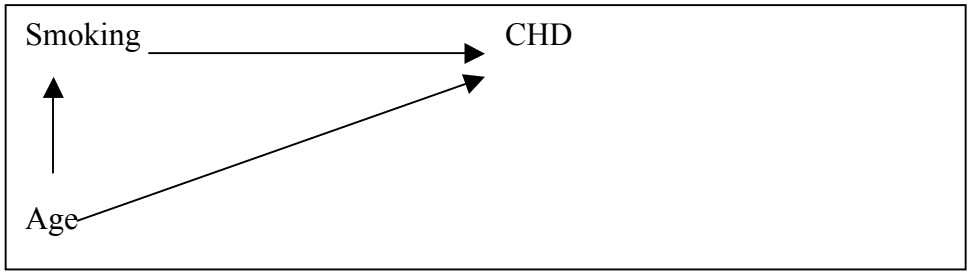
1. Smoking ?
2. Older Age ?
3. Both ?

Recall that the **causal effect of a risk factor for a person** reflects the change in the outcome that is observed when that person is exposed to that risk factor and when that person is not exposed to that risk factor, **under identical situations**. These two outcomes are referred to as **counterfactual outcomes** (Hernan and Robins. J *Epidemiol Community Health* 2006;60:578-586), but only one of them can be observed in reality. For example, the causal effect of lifetime smoking starting at age 20 on the development of CHD for a person is the difference in the CHD counterfactual outcome when that person spends a lifetime as a smoker, compared to the CHD counterfactual outcomes when that person spends a lifetime as a non-smoker. If a person is a lifetime smoker then the observed CHD outcome (factual outcome) matches the counterfactual outcome following a lifetime of smoking. However, we are unable to observe the other CHD counterfactual outcome for that person had that person never smoked. Hence, it is not possible to measure the causal effect of a risk factor for a specific person.

The **average causal effect of a risk factor for a population** is the difference in average counterfactual outcomes when all members of the population receive the risk factor and when none of the members of the population receive it. If the population is comprised of some members who receive the risk factor and others who do not, then it may be possible to estimate each counterfactual outcome and the average causal effect of the risk factor. For example, in the absence of confounding and bias, the incidence of coronary heart disease from a group of non-smokers (**a factual outcome**) may be a valid estimate for the counterfactual outcome for a group of smokers had they not smoked. This assumes an ability to "exchange" factual outcomes of the non-smokers with the counterfactual outcomes for the smokers (Greenland and Robins, *Intl J Epidemiol*

1986;15:412-419). This implies that the observed difference (ratio) in incidence of coronary heart disease in the two comparison groups (**a measure of association)**, is an estimate of the causal effect of smoking.

Causal diagrams in epidemiology are called **Directed Acyclic Graphs (DAGS).** They are directed in that they contains arrows that reflect causal assumptions between potential risk factors and outcomes and they acyclic in that the direction of the arrows do not contains loops from outcomes back to their causes. For example, the following DAG describes the Cohort Study mentioned above and shows two pathways (explanations) that could account for the crude association between smoking and CHD in a data set.



The top arrow reflects of potential direct effect that smoking may have on the risk of developing CHD. However, the other two arrows suggest that there is also a second, **backdoor pathway** to explain this association. Smokers might be older than non-smokers and older age also influences the risk of developing CHD. The existence of a backdoor pathway involving a third factor (e.g. age) provides a **DAG-based definition of confounding.** A risk factor, whose control blocks a backdoor pathway, is a **confounder (confounding factor)**. The challenge of an epidemiologist is to avoid confounding in a study though aspects of the study design (e.g. randomization, restriction, or matching) or to adjust (control) for confounding in the analysis (through stratification, standardization, regression modeling, and the various methods involving propensity scores).

Adjusted measures of association control for factors (confounders) that account for a difference in the factual outcomes in the non-exposed group and the counterfactual outcomes in the exposed group. The lack of **exchangeability** is the **counterfactual-based definition of confounding**. For example, if the smokers are older than the non-smokers in our study, then the non-observed counterfactual outcomes of the smokers (e.g. their incidence of disease under the condition that they did not smoke) would be greater than the factual outcome of the non-smokers (e.g. their observed incidence of disease). The reason for this difference in incidence is the influence of the older age distribution of the smokers. The crude measure of association would mix (confound) the two causal influences (the effect of smoking and the effect of older age).

The existence of a backdoor pathway in a causal diagram, leading to a lack of exchangeability leads to a commonly sited definition of confounding.

1. The confounder must be associated with the exposure (in this example smokers have an older age distribution than non-smokers).

2. The confounder must be associated with the disease, independent of the exposure (in this example, older age increases the risk of disease among the non-smokers).

3. The confounder must not be part of the causal pathway connecting the exposure to the disease.

Suppose that the following data reflect the associations found in the Cohort Study mentioned above.

Crude Analysis

	CHD		
	+	-	Total
Smokers	240	760	1000
Non-Smokers	120	880	1000

Stratified analysis (by Age)

	Young			Old		
	CHD			CHD		
	+	-	Total	+	-	Total
Smokers	60	340	400	180	420	600
Non-Smokers	80	720	800	40	160	200

Do these data reflect confounding by age? The first criterion for confounding states that the confounder must be associated with the exposure. This holds in these data as the prevalence of old age, P(old age), among smokers and non-smokers differ.

P(Old|Smoker) = 600/1000 = 60%
P(Old|Non-Smoker) = 200/1000 = 20%

The second criterion for confounding refers to the relationship between age and CHD, independent of smoking. This can be examined by examining the relationship between age and CHD among the non-smokers. The data support such a relationship as

P(CHD|Young Non-Smoker) = 80/800 = 10%
P(CHD|Old Non-Smoker) = 40/200 = 20%

The third criterion for a confounder states that age should not be in the causal pathway that link smoking with CHD. This criterion can not be examined by the data but can be logically ruled out as smoking does not cause old age.

When confounding exists, the causal graph implies that the value for the crude measure of the association reflects both the effect of the exposure and of the confounder (direct pathway and backdoor pathway). On the other hand, adjusting for a confounder blocks the backdoor pathway and the value for the adjusted measure of association reflects only the direct effect of the exposure. Therefore, when confounding exists, one would expect to observe different values for the crude and adjusted measures of association. This results in commonly used working definition of confounding:

> A confounder is a factor that when adjusted in the analysis results in a value for the adjusted measure of association that is meaningfully different from the value for the crude measure of association.

Although easy to implement and often used in practice, this **"change-in-estimate" definition** of confounding is a necessary but not a sufficient property of confounding and examples have shown it may lead to incorrect conclusions. Nevertheless it is often taken as a method for detecting confounding. This is demonstrated by the following analyses performed on the above data:

Crude Analysis:

	CHD		
	+	-	Total
Smokers	240	760	1000
Non-Smokers	120	880	1000

$$RR_{Crude} = (240/1000)/(120/1000)$$
$$= .24/.12$$
$$= 2.0$$

Stratified analysis (by Age)

	Young			Old		
	CHD			CHD		
	+	-	Total	+	-	Total
Smk	60	340	400	180	420	600
Non-Smk	80	720	800	40	160	200

$$RR_{Young} = (60/400)/(80/800)$$
$$= .15/.10$$
$$= 1.5$$

$$RR_{Old} = (180/600)/(40/200)$$
$$= .30/.20$$
$$= 1.5$$

$$RR_{adjusted} = RR_{Young} = RR_{Old} = 1.5$$

The change-in-estimate method is practical for detecting confounding and displays the "result" of confounding, while the conceptual definition of confounding describes the "mechanism" for the change in estimates.

Stratification

The implication of confounding is that the crude measure of association reflects a mixture of the effect of the exposure and the effect of the confounding factor(s). When confounding exists in a data set, analytical **methods of adjustment** must be used to separate the effect of the exposure from the effect(s) of the confounding factor(s). There are two general approaches for adjusting for confounding factors in the analysis:

1. **Stratification**,
2. **Regression Modeling**.

Regression modeling is the more common method for controlling confounding and stratification can be considered as a special case of modeling. However, because of its intuitive appeal, controlling confounding by stratification will be discussed initially.

Stratification involves dividing the data set into disjoint subgroups (strata) based on categories/values of the confounder(s). There are two methods for adjustment based on stratification:

1. **Pooling (weighted averaging)**
2. **Standardization**.

Stratification with pooling involves the following steps:

1. Create subgroups (strata) defined by categories or sub-ranges of the confounding factor, which are free of residual confounding by that factor,

2. Estimate the value for the measure of association within each stratum, and

3. **When appropriate**, average (pool) these estimates over strata to determine the adjusted measure of association.

The justification for this method is reflected in its first step. If all subjects within a stratum possess (essentially) a common value for a risk factor, then that factor cannot satisfy either of the first two criteria for confounding defined above within that stratum. For a non-continuous confounder, strata defined by distinct categories of the confounder automatically satisfy this situation. For example, when stratifying by sex, the exposed subjects and the non-exposed subjects will have the same sex distribution within the male stratum (all will be males). However, stratification by a continuous confounder requires the specifications of sub-ranges of the confounder to define the strata.

Depending on how sub-ranges of a continuous are defined, there may still be residual confounding by the stratifying factor within a stratum. Suppose that age is a confounder in a study. On one hand, narrowly defined sub-ranges (for example, one-year age intervals) are more homogenous and are less prone to contain residual confounding, but this approach may result in a large number of strata, with little information (individuals) contained within each strata. On the other hand, broadly defined sub-ranges (for example, decades of age) result in fewer strata containing more information, but also have the potential for within-stratum residual confounding by the stratifying factor. In the extreme, the broadest sub-range will result in a single stratum containing the entire data (crude analysis), with no control of confounding.

Given the creation of strata that are free of residual confounding by the stratifying factor, the second step in the stratification calls for estimating the chosen measure of association within each stratum. Averaging (when appropriate) the stratum-specific measures of association into a single number (adjusted measure of association) is usually not based on a simple arithmetic mean, but is based on a method of **weighted averaging** or **pooling** that takes into account the amount of information associated with each stratum-specific estimate.

The most commonly used method for averaging stratum-specific estimates of effect is the method proposed by Mantel and Haenszel (JNCI 22:719-748, 1959). Suppose that the following tables displays the data for the i^{th} stratum

	Disease		
	+	-	Total
Exposure+	a_1	b_i	N_{1i}
Exposure-	c_i	d_i	N_{01}
Total	M_{1i}	M_{01}	T_i

The formula for the **Mantel-Haenszel Weighted Average** is:

Risk Ratio estimate:

$$RR_{MH} = [\Sigma\{a_i N_{0i}/T_i\}]/[\Sigma\{c_i N_{1i}/T_i\}]$$

$$= [\Sigma\{w_i RR_i\}]/[\Sigma\{w_i\}] \text{ if } w_{i.} \neq 0$$

where $w_i = c_i N_{1i}/T_i$

Odds Ratio estimate:

$$OR_{MH} = [\Sigma\{a_i d_i/T_i\}]/[\Sigma\{b_i c_i/T_i\}]$$

$$= [\Sigma\{w_i OR_i\}]/[\Sigma\{w_i\}] \text{ if } w_i \neq 0$$

where $w_i = b_i c_i / T_i$

The weight (w_i) for RR_{MH} can be re-expressed as

$$w_i = [c_i/N_{0i}][(N_{1i}/T_i)(N_{0i}/T_i)][T_i]$$

From this representation, it follows that the value for the weight reflects the amount of information contained within the stratum by it being a function of the frequency of the outcome among non-exposed subjects (the first bracketed term), the balance of the relative sizes of the comparison groups (second bracketed term), and the overall size of the strata (the third bracketed term). These are three components that reflect the amount of information in the table.

Example # 1

The following tables show the crude and age-adjusted measures of association between sex and mortality among patients diagnosed with trigeminal neuralgia (Rothman. Modern Epidemiology Little Brown and Company 1986 and Rothman KJ, Monson RR. J Chron Dis 1973;26;303-309).

Crude Analysis:

	Deaths	Person-yrs	Mort. Rate (per 100 py)
Males	90	2465	3.65
Females	131	3946	3.32

$$RR_{Crude} = 3.65/3.32 = 1.10$$

Stratified (by aged) Analysis

	age < 65		age 65+	
	Deaths	py	Deaths	py
Males	14	1516	76	949
Females	10	1701	121	2245
Total	24	3217	197	3194

$RR_{age<65} = (14/1516)/(10/1701) = 1.57$

$RR_{age65+} = (76/949)/(121/2245) = 1.49$

$$RR_{MH} = [(14)(1701)/3217 + (76)(2245)/3194] / [(10)(1516)/3217 + (121)(949)/3194]$$

$$= 1.50$$

These data suggest that age is a confounder. Age satisfies the first criterion for confounding (1516/2465 = 62% of the male person-years are in the younger group, compared to 1701/3946 = 43% of the female person-years). Age also appear to satisfy the second criterion for confounding (the mortality rate among old females, 5.39 deaths/100py, is much greater than that for young females, .59 deaths/100py). This confounding by age is reflected be the difference between the crude (RR_{Crude} = 1.10) and adjusted (RR_{MH} = 1.50) measures of association.

Example # 2

The following tables show the crude and age-adjusted association between smoking and the 24-year risk of death in the FHS teaching data set.

Crude Analysis

	Died	Survived	Total
Smokers	788	1393	2181
Non-Smokers	762	1491	2253
Total	1550	2884	4434

$$RR_{Crude} = (788/2181) / (762/2253) = 1.07$$

Stratified Analysis

Age ≤ 40	Died	Survived	Total
Smokers	67	385	452
Non-Smokers	25	277	302
Total	92	662	754

Smokers	266	689	955
Non-Smokers	110	574	684
Total	376	1263	1639

50 < Age ≤ 60	Died	Survived	Total
Smokers	286	281	567
Non-Smokers	312	500	812
Total	598	781	1379

Age > 60	Died	Survived	Total
Smokers	169	38	207
Non-Smokers	315	140	455
Total	484	178	662

$$RR_{MH} = [\ \sum(a_i)(N_{0i})/Ti\ ] / [\ \sum\ (c_i)(N_{1i})/T_i]$$

$$= [(67)(302)/754 + (266)(684)/1639 + (286)(812)/1379 + (169)(455)/662] / [(25)(452)/754 + (110)(955)/1639 + (312)(567)/1379 + (315)(207)/662$$

$$= 1.38$$

Age appears to be a confounder in these data. The following table shows that age satisfies the first criterion for confounding (associated with the exposure).

	Non-Smokers (N= 2253)		Smokers (N=2181)	
Age	N	%	N	%
Age ≤ 40	302	13.40	452	20.72
40 < Age ≤ 50	684	30.36	955	43.79
50 < Age ≤ 60	812	36.04	567	26.00
Age > 60	455	20.20	207	9.49

The following table suggests that age satisfies the second criterion for confounding (associated with the outcome independent of the exposure).

Age	Estimated Risk Among Females (Non-Exposed)
Age $\leq$ 40	25/302 = .0828
40 < Age $\leq$ 50	110/684 = .1608
50 < Age $\leq$ 60	312/812 = .3842
Age > 60	315/455 = .6923

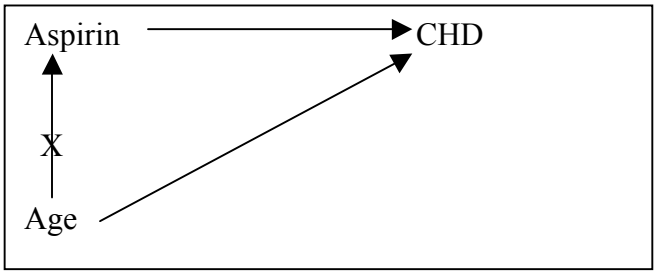
Confounding by age is reflected by the difference between the crude (RR_{Crude} = 1.07) and adjusted (RR_{MH} = 1.38) measures of association

Standardization

Standardization is a second method for adjusting for confounding through stratification. It is also used as a method for summarizing the effect of an exposure when there is **Effect Modification**. Standardization and Effect Modification will be discussed in the next set of lecture notes.

Design Methods of Avoiding Confounding

Randomization in experimental studies reduces for the potential for confounding. For example, in a large RCT examining the effect of aspirin on the risk of Coronary Heart Disease, age should not be a confounder as it would be expected to have very similar distributions in the aspirin and non-aspirin groups as depicted by the following DAG



Restriction is one way to avoid confounding in observational studies. For example, enrolling only subjects of a certain narrow age range in a Cohort Study would avoid confounding by age in a study comparing the incidence of CHD among aspirin and non-aspirin users. However, it may be difficult to generalize the result of this study to other age groups.

Matching is a less rigid for of restriction and may avoid confounding in a Cohort Study. For example, suppose for every aspirin user enrolled in the study, the investigator enrolled a non-aspirin user of the same age. As a result of this matching, the age distribution of the aspirin users would be identical to that of the non-aspirin users and age would not satisfy the first criterion for confounding (as depicted in the DAG above). The topic of matching will be covered in the next sequence of lecture notes.

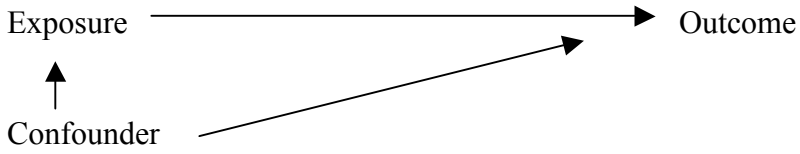
Matching

Matching in a cohort study usually involves selecting non-exposed subjects to have the same the distribution of the matching factor that exists in the exposed group. For example, matching by age may involve selecting a non-exposed subject with the same age as each exposed subject in the study. If the same number of matched non-exposed subjects is enrolled for each enrolled exposed subject (**fixed matching ratio**), then the distribution of the matching factor among the non-exposed will be identical to that of the exposed subjects.

Matching in a case control study involves selecting a control with the same value for the matching factor as each case. If the same number of matched controls is enrolled for each case, then the distribution of the matching factor among the controls will be identical to that of the cases.

The main motivation for matching is typically to avoid confounding in the resulting data that are collected. The simple DAG in the following figure shows the anticipated associations in a study in the presence of confounding and displays of the required relationship of the confounder with the exposure and the outcome. These are the targets of matching in order to avoid confounding in the resulting data.

Directed Acyclic Graph (DAG) reflecting the relationship between a confounder and the exposure and the outcome in the absence of matching in a study.



Matching in Cohort Studies with a fixed matching ratio guarantees that the matching factor will have identical distribution among the exposed and among the non-exposed subjects in the data. This entails eliminating the arrow from the confounder (matching factor) and the exposure in this figure, thereby eliminating the backdoor pathway for confounding. It follows that matching in a Cohort Study, with a fixed matching ratio, avoids confounding by the matching factor in the resulting data.

Matching with a fixed matching ratio in case control studies forces cases and controls to have the same distribution of the matching factor. For example, matching on age would mean that the percentage of elderly people would be the same for cases and controls. However, the causal diagram shows that there are two factors that influence the age distribution of the cases: the direct influence of age (depicted by the arrow connect age with the outcome) and the direct influence of the exposure (which is related to age). Therefore, matching in a case control study does not directly relate to only the pathway from the confounding factor to the outcome. As a result, matching in a Case Control

Study does not block the backdoor pathway and avoid confounding by the matching factor.

Matching by a confounder in a Case Control Study builds similar distributions of the matching factor among the cases and among the matched controls. It will also build similar distributions of other factors that are correlated with the matching factor, including the exposure of interest. Therefore, matching in a Case Control Study tends bias the value for the crude Odds Ratio by towards its null value (1.0). This is demonstrated in the following example.

Example

Large Population:

A. Sex distribution among exposed and among non-exposed subjects in the source population

	Exposed	Non-exposed
Males	8,000 (80%)	2,000 (20%)
Females	2,000 (20%)	8,000 (80%)
Total	10,000	10,000

B. Exposure and sex-specific risks of outcome

	Exposed	Non-exposed
Males	.06	.02
Females	.03	.01

C. Expected number of outcomes cases (# subject x risk)

	Exposed	Non-exposed
Males	480	40
Females	60	80
Total	540	120

D. Expected sex-specific data

Males	Outcome +	-	Total
Exposed	480	7520	8000
Non-exposed	40	1960	2000

RR = 3.0

Females	Outcome +	-	Total
Exposed	60	1960	2000
Non-exposed	80	7920	8000

RR = 3.0

E. Expected crude data

	Outcome		
	+	-	Total
Exposed	540	9460	10000
Non-exposed	120	9880	10000

RR = 4.5

In these data, sex is a confounder since it is associated with the exposure in the source population (Panel A) and is an independent determinant of the outcome (Panel B). Moreover, the common stratum-specific value for the Risk Ratio (RR = 3.0, Panel D) differs from the crude value of the Risk Ratio (RR = 4.5, Panel E).

Matched Cohort Study

A Cohort Study that matches on sex with a fixed matching ratio will enroll a sample of exposed subjects and a sample of non-exposed subjects whose sex distributions are identical. The following table presents the expected results from a matched cohort study that enrolled 1000 exposed subjects selected at random from all exposed subjects in the original large population described in Panel A of the original data. These data also contain 1000 non-exposed subjects who are matched by sex to the exposed subjects

A. Sex distribution among exposed and among matched non-exposed subjects

	Exposed	Non-exposed
Males	800 (80%)	800 (80%)
Females	200 (20%)	200 (20%)
Total	1000	1000

B. Exposure and sex-specific risks of outcome

	Exposed	Non-exposed
Males	.06	.02
Females	.03	.01

C. Expected number of outcomes cases

	Exposed	Non-exposed
Males	48	16
Females	6	2
Total	54	18

D. Expected sex-specific data

	Males				Females		
	Outcome				Outcome		
	+	-	Total		+	-	Total
Exposed	48	752	800	Exposed	6	184	200
Non-exposed	16	784	800	Non-exposed	2	198	200
	RR = 3.0				RR = 3.0		

E. Expected crude data

	Outcome		
	+	-	Total
Exposed	54	946	1000
Non-exposed	18	982	1000
	RR = 3.0		

Panel A of this table shows that the impact of matching in a cohort study is to create a study population to one that cannot support confounding due to the lack of association between the matching factor and the exposure. As a result, the crude Risk Ratio (3.0 from Panel E) is equal to the adjusted values (from Panel D).

Matched Case Control Study

Matching in a Case Control Study impacts the selection of controls. Although a fixed matching ratio will guarantee the lack of a crude association between the matching factor and the outcome, unless the exposure has no effect on the outcome (Odds Ratio =1.0), matching on a confounder will not result in the lack of a conditional association between the matching factor and the outcome . This is demonstrated by the example in the following table. The data are from matched (by sex) case control study based on all 660 outcome cases that developed from the original large population and 660 sex-matched controls.

A. Sex distribution among cases and matched controls

	Cases	Controls
Males	520 (79%)	520 (79%)
Females	140 (21%)	140 (21%)
Total	660	660

B. Prevalence of exposure among cases and matched controls

Cases (Panel C of table for original large population):

	Exposed	Non-exposed	Total
Males	480 (92%)	40 (8%)	520
Females	60 (43%)	80 (57%)	140
Total	540 (82%)	120 (18%)	660

Control (Based on Panel A for original large population):

	Exposed	Non-exposed	Total
Males	416 (80%)	104 (20%)	520
Females	28 (20%)	12 (80%)	140
Total	444 (67%)	216 (33%)	660

C. Expected sex-specific data

Males

	Exposure +	-	Total
Case	480	40	520
Control	416	104	520

OR = 3.0

Females

	Exposure +	-	Total
Case	60	80	140
Control	28	112	140

OR = 3.0

D. Expected crude data

	Exposure +	-	Total
Case	540	120	660
Control	444	216	660

OR = 2.2

E. Expected exposure-specific data:

Exposed

	Male Sex +	-	Total
Case	480	60	540
Control	416	28	444

OR = .54

Non-exposed

	Male Sex +	-	Total
Case	40	80	120
Control	104	112	216

OR = .54

This example demonstrates that matching by sex in this case control study did not avoid confounding. The association between sex and exposure among the controls in

Panel B suggests that the matching factor (sex) satisfies the first criterion for confounding. More importantly, Panel E demonstrates that despite the equal sex distribution among cases and controls caused by matching, conditional on exposure status sex remains an independent determinant of the outcome, although its measure of effect, OR = .54, is different from that suggested in the original large population, RR = 2.0). Therefore by the results of Panels B and E, sex satisfies the two criteria for confounding. In addition, the stratum-specific measure of the effect of the exposure in Panel C (OR=3.0) is different from the crude measure in Panel D (OR = 2.2), suggesting that sex is a confounder by the change-in-estimate criterion for confounding.

Because of its independent relationship to the outcome, stratifying by a confounder in an unmatched Case Control Study would show varying ratios of cases to controls over the strata. If the factor is a strong determinant of the outcome, then some of these strata may contain many cases but few controls (or vice versa), suggesting an inefficient basis to try to measure the effect of the exposure. This is especially true for small studies or when the confounder is nominal is scale with many distinct categories. Since matching tends to make this ratio constant over strata, one would expect that matching and stratification on a confounder might lead to a more efficient analysis than not matching but stratifying by a confounder.

Matched Analysis

The basic principle that underlies a matched analysis is that the association between the exposure and the outcome is first performed within each matched group and then pooled over groups to obtain a summary average. As an example of a matched analysis, the following table presents the basic layout of the results from a Case Control study with a one-one matching design.

		Exposure Status of Control	
		+	-
Exposure Status of Case	+	A	B
	-	C	D

Matched pairs (A and D) with identical exposure status for both the case and the matched control are referred to as concordant pairs. These pairs provide no information on the relationship between exposure and outcome. Matched pairs (B and C) show different exposure status for the case and the matched control, and thus provide information about the relationship between the exposure and the outcome. These matched pairs are referred to as discordant pairs and are the basis for estimates of the effect of the exposure on the outcome and for tests of significance concerning this effect.

The relative sizes of the two types of discordant pairs provide the basis for the measure of the effect of the exposure. An estimate for the odds ratio (OR) from 1-1 match data is

$$OR = B/C$$

This estimate is identical to the Mantel-Haenszel estimator from a stratified analysis with each stratum corresponding to a separate matched group containing one case and 1 control. This demonstrated by the following example presenting an analysis of Case Control Study (Hosmer and Lemeshow. *Applied Logistic Regression Second Edition*. John Wiley & Sons; New York: 2000) that examined risk factors for low birth weight babies. The data for this analysis pertain to 56 matched pairs. Each matched pair contained one case (low birth weight infant) and one control (normal birth weight infant), with the controls matched by the age of mothers of the cases.

		Maternal Smoking of Control	
		+	-
Maternal Smoking of Case	+	8	22
	-	8	18

$$OR = 22/8 = 2.75$$

The following table displays the stratified analysis from the 56 strata.

Strata		Frequency	AD/T	BC/T
D+	D-			
E+ 1	1	8	0	0
E- 0	0			
E+ 1	0	22	1/2	0
E- 0	1			
E+ 0	1	8	0	1/2
E- 1	0			
E+ 0	0	18	0	0
E- 1	1			

$$OR_{MH} = 22(1/2) / 8(1/2) = 22/8 = 2.75$$

Effect Modification

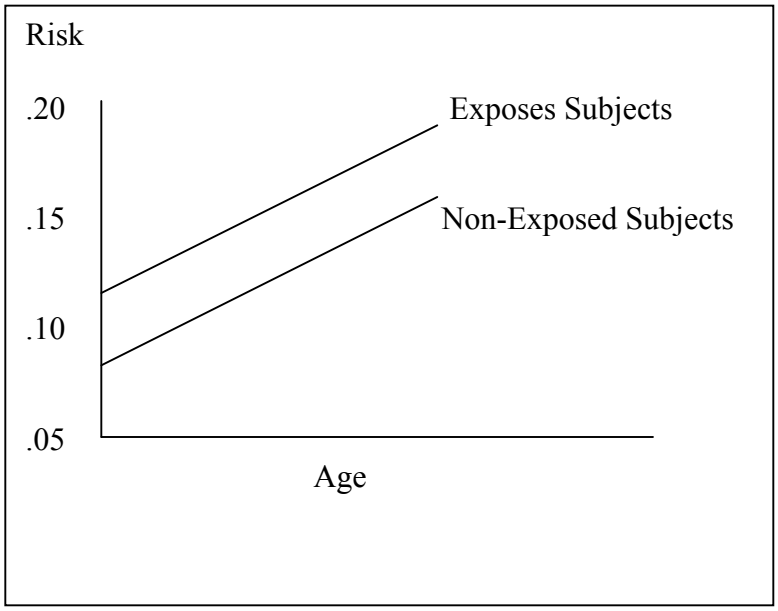
Effect Modification refers to the situation where the effect of the exposure is modified or changed according to the value or level of another factor. Effect Modification is detected by examining sub-group analyses, examining the association between and exposure and an outcome with sub-groups defines by categories of the candidate effect modifier.

For example, the following data are from a Retrospective Cohort Study examining the relationship between perioperative beta-blocker use and in-hospital mortality among 663,535 patients undergoing non-cardiac surgery at 329 Hospitals (Lindenauer et al: N Engl J Med 2005;353:349-61) The data are stratified by the Revised Cardiac Risk Index Score (RCRI), a measure of a patient's risk of developing a cardiac complication during surgery.

RCRI Score	Odds Ratio	Confidence Interval
0	1.36	(1.27,1.45)
1	1.09	(1.01,1.19)
2	0.88	(0.80,0.98)
3	0.71	(0.63,0.80)
≥ 4	0.58	(0.50,0.67)

These data shows the effect of beta-blocker use on mortality depends on the level of the RCRI. Among patients with low risk of a cardiac complication (RCRI scores of 0 or 1), beta-blocker use tends to increase the risk of in-hospital death (RR = 1.36 for patients with RCRI=0 and RR= 1.09 for patients with RCRI=1). On the other hand, for patients with higher values of RCRI, beta-blocker use tends to decrease the risk of in-hospital death (RR= 0.88 for patients with RCRI=2, RR= 0.71 for patients with RCRI=3 and RR= 0.58 for patients with RCRI $\geq$ 3).

The detection of effect modification depends on the choice of the measure of association. For example, the following figure suggests that age does not modify the value for the Risk Difference (the distance between points on the two lines is the same for any value for age), but does modify the value for the Risk Ratio (the ratio of the points on the two lines becomes less with increasing age.)



The detection of Effect Modification is challenging. Before concluding that results like those presented in the previous numerical example reflect the presence of effect modification, an investigator must rule out the roles of

1. Bias
2. Confounding
3. Chance

In addition, as with subgroup analyses in experimental studies, the investigator should have a clinical argument to justify examining for effect modification by a risk factor

Tests for detecting effect modification are based on comparing stratum-specific estimates of effect, which are each based on only a portion of the original data. Therefore, they may have limited power. A comparison of confidence intervals around measures of association for various sub-group analyses also provides evidence about chance being the explanation for the data suggesting the existence of effect modification. For example, each of the RCRI-specific Odds Ratio in the beta- blocker example is not contained the confidence intervals for any of the other sub-group analysis.

Presenting the Effect of an Exposure in the Presence of Effect Modification

When effect modification exists, the single average measures of association cannot be expected to estimate the differing values for the measure of association that exist in the various strata. In this situation, the presentation of stratum-specific results or a method of standardization is superior to the method of weighted averaging (Mantel-Haenszel Estimate) as described in the previous lecture notes.

When effect modification exists, probably the best manner to present the effect of the exposure is by displaying the different measures of association for different subgroups or strata of the effect modifier. For example, if age is an effect modifier, then one might display the separate effects of the exposure for young, middle-aged, and old subjects.

An alternative to presenting stratum-specific estimates of effect is to present a summary (average) measure of association that is linked to a specified population with a known distribution of the effect modifier. This involves the method of **direct standardization** and compares two standardized measures of disease frequency: one under the assumption that everyone in the standard population has the stratum-specific risks of the exposed subjects, and the other under the assumption that everyone in the standard population has the stratum-specific risks of the non-exposed subjects. Therefore, standardization estimates the counterfactual outcomes that were described in the previous series of lecture and estimates the average causal effect of the exposure in the standard population.

The formula for the standardized risk ratio is

$$SRR = [\Sigma\{n_i R_{1i}\}] / [\Sigma\{n_i R_{0i}\}]$$

$$= [\Sigma\{w_i RR_i\}] / [\Sigma\{w_i\}]$$

where

n_i = number of subjects in the standard population in the i^{th} stratum

R_{1i} = risk (rate) of the outcome among the exposed subjects in the i^{th} stratum

R_{0i} = risk (rate) of the outcome among the non-exposed subjects in the i^{th} stratum

$w_i = n_i R_{0i}$

If the exposed subjects are chosen as the standard population (i.e. $n_i = N_{1i}$), then this formula simplifies to

$$SRR = a / [\Sigma\{N_{1i} R_{0i}\}]$$

where a = total number of exposed subjects who develop the outcome.

In this special case where the exposed subjects are taken as the standard population, the standardized risk ratio becomes the ratio of the observed number of exposed cases to the expected number of exposed cases. This ratio is usually referred to as the SMR

(standardized mortality ratio or standardized morbidity ratio) and is a component of the method of indirect standardization.

A standardized risk ratio reflects the overall, unconfounded effect of the exposure in a specific population (the chosen standard) with a specific distribution for the effect modifier. Choosing a different standard with a different distribution for the effect modifier should result in a different value for the standardized risk ratio, reflecting the overall effect of the exposure in the new standard population.

Example

The following data can be used estimate the age-standardized risks of death among smokers and non-smokers in the FHS teaching data set. The standard population is the total number of study subjects (4434) in the data and the age strata are the same that were used in the previous lecture notes to describe stratification.

Stratified Analysis

Age ≤ 40	Died	Survived	Total
Smokers	67	385	452
Non-Smokers	25	277	302
Total	92	662	754

40 < Age ≤ 50	Died	Survived	Total
Smokers	266	689	955
Non-Smokers	110	574	684
Total	376	1263	1639

50 < Age ≤ 60	Died	Survived	Total
Smokers	286	281	567
Non-Smokers	312	500	812
Total	598	781	1379

Age > 60	Died	Survived	Total
Smokers	169	38	207
Non-Smokers	315	140	455
Total	484	178	662

The following table displays the calculation of the expected number of deaths under the two scenarios that all members of the population smoked and that none of the members of the population smoked. The expected number of deaths for each age group (columns 4 and 6) are estimated by multiplying the number of subjects in the standard population who are in that age group by the correspond risk of dying for that age group.

Standard Population(Age Group)	Number	Risk if all were Exposed (Smoker)	Expected. # Cases	Risk if all Non-Exposed (Non-Smoker)	Expected # of Cases
≤ 40	754	67/452 = .1482	754(.1482) = 111.74	25/302 = .0828	754(.0828) = 62.43
(40, 50]	1639	266/955 = .2785	1639(.2785) = 456.46	110/684 = .1608	1639(.1608) =263.55
(50, 60]	1379	286/567 = .5044	1379(.5044) = 695.57	312/812 = .3842	1379(.3842) = 529.82
> 60	662	169/207 = .8164	662(.8164) = 540.52	315/455= .6923	662(.6923) = 458.30
Total	4434		1804.29		1314.10

The Standardized Risks of Death (estimated counterfactual outcomes) and Standardized Risk Ratio from this table are

Smokers: 1804.29/4434 = 0.4069

Non-Smokers: 1314.10/4434 = 0.2964

Standardized Risk Ratio: .4069/.2964 = 1.37

Inverse Probability Weighting

Standardization involves inverse probably weighting. For example, the previous table shows that the expected number of deaths in the youngest age group if all 754 subjects smoked is111.74. The following calculation shows two formulas for the calculation of this value.

111.74 = (smokers risk) x (size of standard population)
= (67/452) x (754)
= (67/452) x (452) x (754/452)
= (smokers risk) x [(# smokers) x (weight)]

The bottom equation implies that the expected number of death can be obtained by multiplying the number of young smokers by a weight and then multiplying this value by the risk of death among the young smokers. The weight (754/452) equals

754/452 = 1/P(Smoking| Age $\leq$ 40)
= 1/P(Exposure| Age $\leq$ 40)
= 1/P(Exposure|Confounder)

The P(Exposure | Confounder(s)) is called the **Propensity Score.** This topic will be discussed in the next series of lecture notes. The previous calculation demonstrated that the estimation of the number of deaths if everyone in the standard population has the exposure involves weighting the exposed subjects by the inverse of their propensity scores. Similarly, the estimation of the number of deaths if no one in the standard population has the exposure involves weighting the non-exposed subjects by the inverse of (1- propensity score).

Regression Models

Role of Regression Models in Clinical Research:

The practical goal of epidemiology is to measure and interpret associations between suspected risk factors and outcomes. For causal research, the usual measurement goal of the investigation is to quantifying the effect of a single risk factor of interest (exposure) on the outcome while controlling for confounding by other factors (**explanation**). However, a second measurement goal of epidemiology (especially of clinical epidemiology) may be to quantify the joint effect of many risk factors to estimate an individual's risk of developing or possessing an outcome (**prediction**). Regression Models are commonly used to achieve either of these goals. However, the steps used to develop these models and the focus on their results depends on whether explanation or prediction is the goal of the analysis.

Regression Coefficients

Regression coefficients can be interpreted as slope coefficients, reflecting the change in an outcome, per unit chance in a risk factor. For example, the following data show the relationship between smoking (measured in packs smoked per day) and the log(odds of dying), the **logit**, during the 24 years of follow-up in the Framingham Heart Study (FHS) Teaching Data Set. (N.B. 32 of the 4434 participants have missing values for cigpday1 and therefore have missing values for packs smoked per day.)

Packs Smoked Per Day	Number at Risk	Number of Deaths	Estimated Risk	Logit
0	2253	762	762/2253 = .34	log(.34/.66) = -.66
1	1671	573	573/1671 = .34	log(.34/.66) = -.66
2	398	169	169/398 = .42	log(.42/.58) = -.32
3+	80	36	36/80 = .45	log(.45/.55) = -.20

The data in this table shows evidence of an increasing log(odds of death) with increasing number of packs of cigarettes smoked, which might be approximated by the following linear equation

$$\log(\text{Odds of Death}) = B_0 + B_1(\text{Packs})$$

A regression equation that describes the relationship between the log(odds of an outcome) as function of one or more risk factors is called a **Logistic Regression Model**. The slope of this linear equation (B_1) measures the change in the log(Odds of Death) per smoking one additional pack of cigarettes per day. For example, if P_x is the risk of death when smoking "x" packs of cigarettes per day then

$$B_1 = \log(P_{x+1}/(1-P_{x+1})) - \log(P_x/(1-P_x))$$

$$= \log[(P_{x+1}/(1-P_{x+1})) / \log(P_x/(1-P_x))]$$

$$= \log (\text{Odds Ratio})$$

The last equation shows that regression coefficients for a logistic regression model can be interpreted as the logarithm of a common measure of association, the Odds Ratio. In addition if the logistic regression model contains multiple risk factors, the coefficient for any risk factor has the interpretation as the log(Odds Ratio), measuring the association between that risk factor and the outcome, conditional on all other risk factors in the model.

For example, the following formula describes the risk (P) of death during 24 years of follow-up in the FHS Teaching Data Set as a function of five risk factors: current smoking status (CURSMOK1: 1=yes, 0=no), age in years (AGE), male sex (MALE: 1= yes, 0=no), hypertension (HIGHBP1: 1 if sysbp1 $\geq$ 140 or diabp1 $\geq$ 90, 0 otherwise) and diabetes (DIABETES1: 1=yes, 0=no).

$$\log(P/(1-P)) = B_0 + B_1\text{CURSMOK1} + B_2\text{AGE} + B_3\text{MALE} + B_4\text{HIGHBP1} + B_5\text{DIABETES1}$$

When fit to the FHS teaching data set, the fitted model becomes

$$\log(P/(1-P)) = -7.5869 + 0.5522(\text{CURSMOK1}) + 0.1181(\text{AGE}) + 0.7759(\text{MALE}) + 0.6386(\text{HIGHBP1}) + 1.5834(\text{DIABETES1})$$

If this model correctly describes the relationship between the 5 risk factors and the risk of death, then it follows that the 24-year risk of death for a 50-year-old female, without hypertension and without diabetes is

$$\log(P/(1-P)) = -7.5869 + 0.5522(\text{CURSMOK1}) + 0.1181(50) + 0.7759(0) + 0.6386(0) + 1.5834(0)$$

$$= -1.6819 + 0.5522(\text{CURSMOK1})$$

This equation resembles the linear equation presented above. Therefore, the coefficient of CURSMOK1 (0.5522) is the log(Odds Ratio) describing the relationship between smoking and death, for a 50-year-old female, without hypertension and without diabetes and

$$OR = e^{0.5522} = 1.74$$

On the other hand, the 24-year risk of death for a male, 50-year-old male with hypertension and with diabetes is

$$\log(P/(1-P)) = -7.5869 + 0.5522(CURSMOK1) + 0.1181(50) + 0.7759(1) + 0.6386(1) + 1.5834(1)$$

$$= 1.3160 + 0.5522(CURSMOK1)$$

The coefficient of CURSMOK1 (0.5522) again is the log(Odds Ratio) describing the relationship between smoking and death, but now for a 50-year-old male, with hypertension and without diabetes and

$$OR = e^{0.5522} = 1.74$$

In summary, for any combination of (AGE, MALE, HIGHBP1, and DIABETES1), the model's estimate of the effect of CURSMOKE1 is OR = 1.74.

In general, a regression coefficient estimates the effect of a predictor on the outcome, controlling for all other factors in the model. This property is the basis for controlling for confounding by a regression model requires. This method requires including sufficient terms in the model to represent the confounders. It also requires valid **modeling assumptions**, and wrong modeling assumptions can lead to wrong conclusions.

Multiple Regression Models

A multiple regression model is a mathematical expression that postulates a relationship between an outcome and a set of predictors. Typically the predictors in the model represent the exposure of interest, potential confounders, and possibly effect modifiers. Perhaps the most commonly used model in epidemiologic research is the **logistic regression model**. This model is appropriate for a **binary outcome** (e.g. dead versus alive) and describes the risk (P) of developing the outcome as a function of the predictors ($X_1, X_2, \ldots, X_n$) by the following formula:

$$\log(P/(1-P)) = B_0 + B_1X_1 + B_2X_2 + \ldots. + B_nX_n$$

The unknown coefficients in this model are the intercept term (B_0) and the coefficients

(B_i) of the predictors (X_i). The intercept term (B_0) specifies the value for the outcome (log(P/(1-P))) when all predictors are set equal to zero (i.e. $X_i = 0$, $i = 1,2, \ldots, n$). More importantly, each coefficient (B_i) is a slope coefficient, describing the change in log(P/(1-P)) per unit change in X_i (log(OR)) when all other predictors are fixed. This coefficient is often interpreted as an estimate of the "effect" of the corresponding predictor, controlling for all of the other factors in the model. In this way multivariate models are the most common method for measuring the effect of an intervention, controlling for confounding by other factors.

Another multivariable model often used in clinical research is the **linear regression model**, which describes the relationship between a **continuous outcome** (Y) and a set of predictors. This model assumes that the average outcome (expected value, E(Y)) is related to the predictors according to the following formula

$$E(Y) = B_0 + B_1X_1 + B_2X_2 + \ldots + B_nX_n$$

Other models that are sometimes used in clinical research are the **Cox Regression Model** for survival time outcomes with censoring and the **Poisson Regression Model** for count outcomes. The main differences in these models are the type of outcome (binary outcome, continuous outcome, survival time outcome, and count outcome) and the method for fitting a model to a data set. However, all model share many principles in common, including the interpretation of any regression coefficient as representing the change in the outcome per unit change in a predictor, controlling for all other predictors in the model.

Methods for estimating regression coefficients will not be discussed in detail in these lecture notes. Algorithms for fitting a model to a data set differ, depending on the type of the regression model. For example, maximum likelihood estimation is used to estimate the coefficients of a logistic regression model. Least estimation is used to estimate the coefficients in a linear regression model. Partial likelihood methods are used to estimate the coefficients in a Cox regression model. All methods have the goal that outcome estimates for individuals from the model should agree as much as possible with the actual outcomes. For example, risk estimates from a logistic regression model should be as high as possible (close to 1.0) for those who develop the outcome and as low as possible (close to 0.0) for those who do not develop the outcome.

Model Assumptions

The following table shows the equation for the Mortality Prediction Model (MPM_0) (*Lemeshow et al. Predicting the Outcome of Intensive Care Unit Patients. J Am Stat Assoc 1988;83(402):348-356.*). The MPM_0 was developed from a logistic regression model to predict the risk of in-hospital death, P, from a set of clinical measures for patients admitted to an ICU.

Table. Model predicting the risk of hospital mortality (P) for patients admitted to an intensive care unit based on a logistic regression model containing seven predictors

Predictors

CONS	: level of consciousness (1 if coma or deep stupor, 0 otherwise)
TYPE	: type of admission (1 if emergent, 0 if elective)
CANCER	: cancer as part of present problem (1 if yes, 0 if no)
CPR	: prior CPR (1 if yes, 0 if no)
INFECT	: infection (1 if probable, 0 otherwise)
AGE	: age in ten year increments
SBP	: systolic blood pressure
SBP2	: SBP squared.

Model :

$$\log(P/(1-P)) = -1.370 + 2.44(CONS) + 1.81(TYPE) + 1.49(CANCER) + .974(CPR) + .965(INFECT) + .0368(AGE) - .0606(SBP) + .000175(SBP2)$$

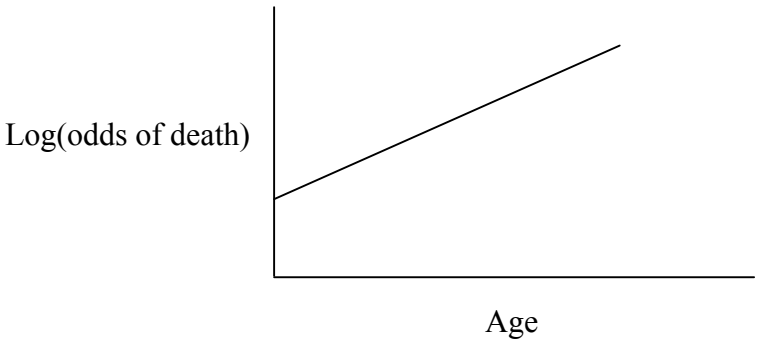
Most of the factors in the MPM model are binary predictors, each representing the presence or absence of a risk factor. However, the model also contains terms for two continuous risk factors: age and systolic blood pressure. The model assumes that each predictor has a single effect on the outcome as measured by its coefficient and that this effect holds over all subgroups of subjects that are defined by the other predictors in the model. For example, this model assumes that the effect prior CPR (as estimated by its regression coefficient (0.974) is not modified by any of the other predictors in the model. This condition is known as the **assumption of additivity**.

The model in this table also contains a single term for the age of each subject. If we fix the values for the other predictors, then this model assumes that the conditional linear relationship between age and the log odds of dying, which is described by the following equation and represented by a straight line as depicted in the following figure.

$$\log(P/(1-P)) = B_0^* + .0368(AGE)$$

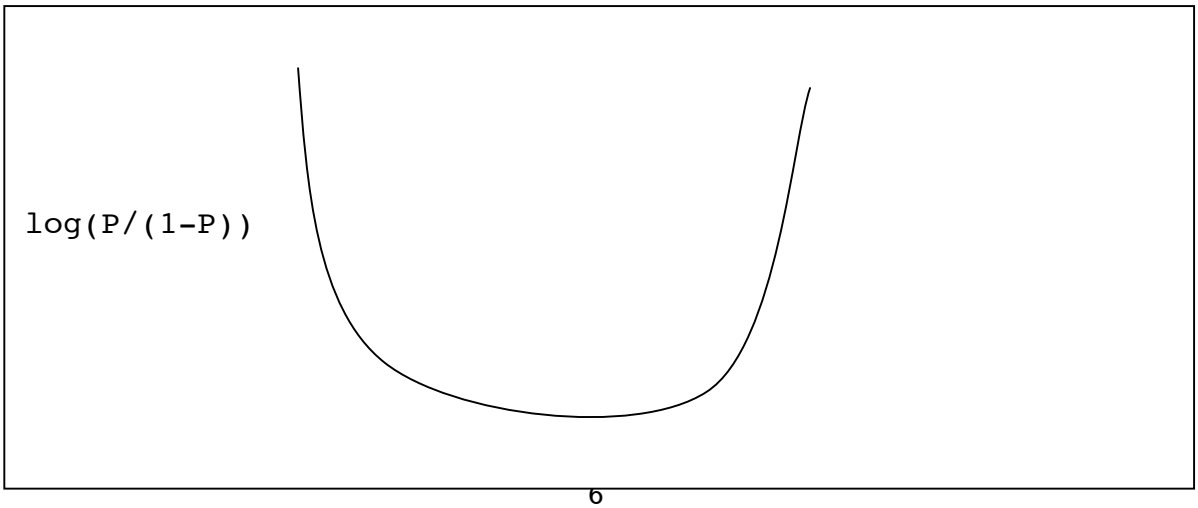
where the value for B_0^* depends on the fixed values for the other predictors in the model.

Figure Assumed conditional linear relation between age and Log(odds of deaths)



The relationship displayed in this figure refers to an **assumption of linearity**. The slope of the line reflects the increase in the outcome (log odds of dying) per unit increase in the age. For example, the model presented in previous table assumes that the log odds of death increases by 0.0368 for every increase in the year of age. This corresponds to an odds ratio of $e^{.0368} = 1.04$ for each one-year increase in of age.

It is important that the model's assumption properly reflect the true relationship between a continuous predictor and the outcome. If the relationship between a continuous predictor and an outcome is not linear, then the model may need to contain additional terms to reflect its non-linear relationship to the outcome. For example, one might expect that the risk of death might be high for patients with very high blood pressure (hypertension), but also be high for patients with very low blood pressure (hypotension). Thus one might expect a U-shape relationship between blood pressure and the log(odds of death as shown in the following figure:



Fixing the values of the other predictors, the MPM_0 simplifies to the following quadratic equation to describe the conditional association between systolic blood pressure (SBP) and the log(odds of death):

$$\log(P/(1-P)) = B_0^* - .0606(SBP) + .000175(SBP^2)$$

where the value for B_0^* depends on the fixed values for the other predictors in the model.

Relationship Between Stratification and Regression for Controlling Confounding

The simplest method for controlling a confounder is through stratification according to categories or sub-ranges of the confounder. Strata are sub-groups of patients with common values for the confounder. Since the value for the confounding factor is constant (or nearly constant) within a stratum, all subjects with a stratum should have homogeneous risk of developing the outcome as influenced by the confounder. For example, stratifying by sex will create two strata, one containing only males and the other only females. Even if males may have higher risk for developing the outcome than females, comparing males who received the exposure to males who did not receive the exposure will provide an estimate for the effect of the intervention that is free of confounding by sex. The same is true when the analysis is performed within the female stratum. If the effect of the intervention within males is similar to that within females (no effect modification by gender), then the sex-specific effects can be pooled over strata to provide a single adjusted measure of the effect of the intervention, as demonstrated in a previous series of lecture notes.

This argument can be generalized to confounders with more than two categories to define strata. It can also be applied to continuous confounders by defining strata based on sub-ranges of the confounders. For example, stratification by age often involves strata that are based on decades of age. This approach assumes that the risk of developing the outcome does not vary much by the level of the confounder within each stratum. However, using too wide of a sub-range of the confounder to define a stratum could result in residual confounding within that stratum.

Control of Multiple Confounders

Potentially stratification can adjust for the joint confounding by a set of factors through multiple levels of stratification. For example, if age is divided into three age groups (young, middle-aged, and old), then the joint stratification by age and sex will result in six strata (young men, young women, middle-aged men, middle-aged women, old men, and old women). Although simple and straightforward, this method is not practical for adjusting for many confounders. For example, simultaneous stratification by only six binary confounders results in 64 strata, which may be too many for most data sets. Therefore alternative adjustment strategies must be considered. A regression model is the most commonly method for controlling multiple

confounders. This is accomplished by fitting a model containing a term for the intervention/exposure and additional terms for the confounding factors.

If age is treated as a categorical variable (as in the previous paragraph) then controlling for age and sex in a model is analogous to stratification if the model contains separate terms to represent the 6 strata defined by age and sex. However, if age possesses a linear relationship with the log(odds of death) and the effect of age is not modified by sex, then the following simple model may provide a better control of confounding by age and sex than stratification

$$\text{Log}(P/(1-P)) = B_0 + B_1(\text{Exposure}) + B_2(\text{Age}) + B_3(\text{Sex})$$

A regression model that controls for many confounders would be very large and complex. Fixing large models to small or moderate sized data sets is a challenge. One way to resolve this problem is to summarize the confounders into a single score. For example, the following section provides another methods for controlling confounding, not by including terms for the individual confounders in the model, but by summarizing the confounders into summary score, a **propensity score (***Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometika 1983;70:41-55 and Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984;79:516-524.***)** If conducted properly, controlling for the propensity score should also control for the individual confounders that define the propensity score.

Propensity Scores

Typically observational studies (i.e. non-randomized trials) that investigate the effect of clinical interventions (treatments or triage decisions) are characterized by a large number of factors that influence both the choice of the intervention and the outcome. This problem is labeled as **confounding by indication** by epidemiologists, and is demonstrated by the following examples.

Example # 1: Measuring the Effect of Right Heart Catheterization (SUPPORT)

Connors et al examined effect of right heart catheterization in the care of critically ill patients (*Connors et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. JAMA. 1996;276:889-897*.). This study examined the survival and health care utilization outcomes for 5735 ICU patients. 2184 of these patients received a right heart catheter (RHC) during the first 24 hours of care in an ICU and another 3551 ICU patients did not receive a RHC. The following table displays the distributions for a sample of patient characteristics.

Patient Characteristics	Received (RHC) (n=2184)	Did Not Receive RHC (n= 3551)
Percent over 80 Years of Age	8%	14%
Mean SBP	68	85
Mean Heart Rate	119	112
Mean Creatinine	221	168
Mean of Apache Score (Measure of Disease Severity)	61	51
Mean Albumin	29	32
Mean of Estimate for 2-Month Survival from Prediction Rule	56	61

This table demonstrates that RHC patients differed from the non-RHC patients on a number of factors that can influence outcomes. Therefore, any difference in the outcomes for the two groups of patients might be attributed to the effect of the right heart catheter and/or to the effects of these other factors. This potential for confounding is depicted by the following causal graph (Directed Acyclic Graph, DAG).

Figure: Directed Acyclic Graph (DAG) displaying potential for confounding in the study by Connors et al.

The following table shows the outcomes (6 month survival, mean total cost, and mean ICU Length of Stay (LOS)) for these patients. Although RHC patients showed worse outcomes, the potential for confounding by the patient characteristics in the previous table raises the question about whether the worse outcomes for the RHC patients reflects the effect of RHC or the type of patient who is given a RHC.

Outcome	Received (RHC) (n=2184)	Did Not Receive RHC (n= 3551)	P-Value
6-Month Survival Probability	53.7%	46.3%	< 0.001
Mean Total Cost	$131,900	$74,300	< 0.001
Mean ICU LOS (days)	15.5	10.3	< 0.001

The following table displays the same patient characteristics for a sample of 1008 RHC and 1008 non-RHC patients who were matched by a propensity score (to be defined later in these notes). Contrary to the previous table for all 2187 RHC patients and 3551 non-RHC patients, the following table shows near identical distributions of these characteristics, suggesting less potential for confounding.

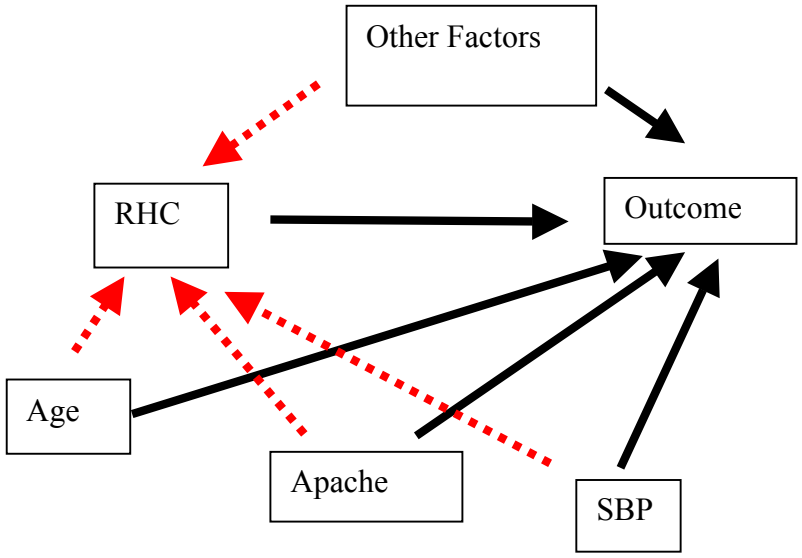
Patient Characteristics	Received (RHC) (n=2184)	Did Not Receive RHC (n= 3551)
Mean Age	60	60
Percent Male	59%	54%
Mean SBP	71	73
Mean Heart Rate	111	111
Mean Creatinine	203	203
Mean of Apache Score (Measure of Disease Severity)	57	57
Mean Albumin	30	30
Mean of Estimate for 2-Month Survival from Prediction Rule	.59	.58

The following table displays the outcomes for the 1008 RHC patients and the 1008 non-RHC patients who were matched by the propensity score. Although the difference in outcomes are attenuated compared to those for all 5735 patients, the outcome for the RHC group remain worse than those for the non-RHC group. These may reflect the true effect of Right Catheterization or possibly confounding by other factors.

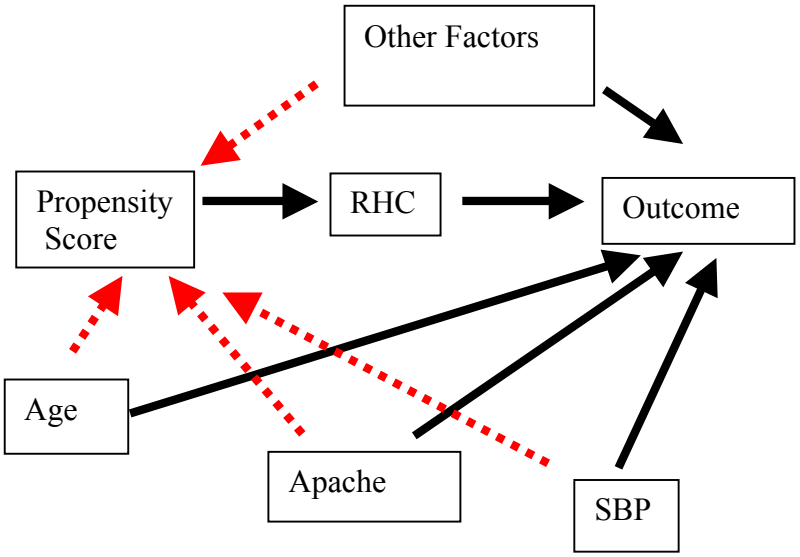
Outcome	Received (RHC) (n=1008)	Did Not Receive RHC (n= 1008)	P-Value
6-Month Survival Probability	51.2%	46.0%	< 0.001
Mean Total Cost	$49,300	$35,700	< 0.001
Mean ICU LOS (days)	14.8	13.0	< 0.001

The motivation for a Propensity Score Analysis is to control for a large number of confounders by combining them into a single summary score. Although the theory of propensity scores was developed over 30 years, their use to control confounding was seldom used until recently.

The Propensity Score is the probability of receiving the exposure as a function of the confounders. The following causal diagram (DAG) displays the anticipated relationship for the factors mentioned in the previous example. The solid arrows connect the suspected confounders and the exposure to the outcome. These arrows reflect the regression coefficients for these predictors in a regression model that predicts the outcome, like the models mentioned earlier in these notes.



On the other hand, the dashed arrows connecting the suspected confounders to the exposure are the basis for the propensity score. The role of the propensity score is to summarize the role of the individual confounders in influencing the treatment decision as shown the following DAG. If the propensity score captures the influence of all of the individual confounders on the treatment decision, then controlling for the propensity score will block the backdoor pathways through the individual factors to the outcome, thereby controlling for any confounding attributed to them.



Example # 2: Effect of Hypertension Treatment

The following analyses examine the association between hypertension treatment (BPMEDS1) at the 1956 exam and the 24-year risk of death in the FHS teaching data set. The analysis is restricted to 1372 participants with a diagnosis of hypertension at the 1956 exam. The following table shows the crude association between hypertension medication use and death

	Died	Survived	Total
BPMEDS1=1	91	48	139
BPMEDS1=0	627	606	1233
Total	718	654	1372

$$OR_{Crude} = (91/48) / (627/606) = \mathbf{1.83}$$

These results suggest a somewhat surprising harmful effect from hypertension medication use. However the following table displays the imbalance of the distribution for 10 potential confounders.

TABLE	On Hypertension Medication (N=139)	Not on Hypertension Medication (N=1233)
Male (%)	30%	47%
Age (Mean)	56.29	53.55
Cholesterol (Mean)	257.43	246.34
SBP (Mean)	165.08	154.26
DBP (Mean)	96.45	93.43
Obese/Overweight (%)	73%	72%
Smoker (%)	35%	42%
Diabetes (%)	6%	4%
Prevalence of CHD (%)	14%	7%
Prevalence of Stroke(%)	5%	1%

. The table shows that participants taking hypertension medication have greater average Age, SBP, DBP, Total Cholesterol, Diabetes, History of CHD and Stroke. Participants not on hypertension medication have a higher percentage of Males and Smokers. As mentioned above, the usual method for controlling multiple confounders is by an outcome regression model that contains terms for the individual confounders. For example, the following logistic regression model depicts the risk of death, P, as a function of the exposure (BPMEDS1) and the 10 potential confounding factors

$$\begin{aligned}\log(P/(1-P)) = -9.4873 + \mathbf{0.3405(BPMED1)} &+ 0.1117(AGE1) + 0.000239(TOTCHOL1) \\ &+ 0.0232(SYSBP1) - 0.00894(DIABP1) + 1.2178(MALE) \\ &+ 0.4763(CURSMOKE1) + 1.6317(DIABETES1) + 1.0511(PRECCHD1) \\ &+ 0.9972(PREVSTRK1) + 1.3087(UNDERWT) - 0.3166(OVERWT) \\ &+ 0.0301(OBESE)\end{aligned}$$

As described earlier, the coefficient of BPMED1 (0.3405) is the log(Odds Ratio) measuring the association between medication use and death, controlling for the other factors in the model. It follows that the adjusted Odds Ratio is

$$OR_{Logistic} = e^{0.3405} = 1.41$$

An alternative method to control for the potential confounders in this model is by an analysis based on a propensity score, reflecting their relationship of the confounders with the

exposure (BPMEDS1). The following logistic regression model describes the risk of hypertension medication use, PS, as a function of the potential confounders

$$\begin{aligned}\log(PS/(1-PS)) = {} & -7.0445 + 0.0207(AGE1) + 0.00347(TOTCHOL1) \\ & +0\ .0145(SYSBP1) + 0.00592(DIABP1) - 0.4586(MALE) \\ & - 0.0289(CURSMOKE1) + 0.0697(DIABETES1) \\ & + 0.5283(PRECCHD1) + 1.3750(PREVSTRK1) \\ & + 0.8119(UNDERWT) + 0.0849(OVERWT) + 0.1080(OBESE)\end{aligned}$$

The propensity score is a balancing score. If it captures the relationship between the potential confounders and the exposure, then conditioning on it should eliminate any association between the individual confounders and the exposure. The following table shows the adjusted relationship between the individual confounders and hypertension medication use after adjusting for the propensity score. The first two columns repeat the crude imbalance of the confounders shown in a previously presented table. The last two columns show the balance of the same potential confounders in an analysis that adjusts for the propensity score by we-weighting the data by a function of the propensity score (as demonstrated in the last series of lecture notes). This table demonstrates very similar distributions of the confounders (better balance) after adjusting by the propensity score.

TABLE	Crude		Propensity Score Adjusted	
	Meds	No Meds	Meds	No Meds
Male (%)	30%	47%	43%	45%
Age (Mean)	56.29	53.55	53.91	53.83
Cholesterol (Mean)	257.43	246.34	245.77	247.43
SBP (Mean)	165.08	154.26	153.86	155.37
DBP (Mean)	96.45	93.43	92.66	93.73
Obese/Overweight (%)	73%	72%	71%	66%
Smoker (%)	35%	42%	41%	41%
Diabetes (%)	6%	4%	4%	5%
Prevalence of CHD (%)	14%	7%	9%	8%

Prevalence of Stroke(%)	5%	1%	2%	2%

Confounding can be controlled with a propensity score by the following methods

1. Matching by the propensity score (as in the RHC analysis above)
2. Stratifying by ranges of the propensity score
3. Including the propensity score in an outcome regression model in place of individual confounders
4. Re-weight the data by a function of the propensity score (similar to standardization as described in the previous lecture notes)

The following analysis uses stratification by ranges of the propensity score (second option) to control for confounding in the problem at hand. The following table shows the distribution of the propensity score in seven strata defined by ranges of the propensity score. Not surprising, 33.81% of the participants taking hypertension medication belong to the last four strata, compared to only 12.90% of the participants not on medication.

Propensity Score	BPMEDS=1	BPMEDS=0
$0 \leq PS \leq 0.05$	12 (8.63%)	234 (18.98%)
$0.05 < PS \leq 0.10$	43 (30.94%)	574 (46.55%)
$0.10 < PS \leq 0.15$	37 (26.62%)	266 (21.57%)
$0.15 < PS \leq 0.20$	20 (14.39%)	82 (6.65%)
$0.20 < PS \leq 0.25$	9 (6.47%)	32 (2.60%)
$0.25 < PS \leq 0.30$	8 (5.76%)	28 (2.27%)
$PS > 0.30$	10 (7.19%)	17 (1.38%)

Creating seven strata (defined by ranges of the propensity score), measuring the association between hypertension medication use and death within each stratum, and average

these results over the strata using the Mantel-Haenszel formula (presented in the previous series of lecture notes), yields an adjusted Odds Ratio

$$OR_{MH} = 1.37$$

This value for the Odds Ratio from stratifying by the propensity score ($OR_{MH} = 1.37$) is very similar to the adjusted Odds Ratio that was obtained from the outcome logistic regression model ($OR_{Logistic} = 1.41$). The similarity of results is not surprising as both methods are valid options for controlling confounding.

In general, a correctly specified outcome model and an appropriately performed propensity score analysis should result in similar adjusted measures of association to estimate the effect of the exposure. Both methods use regression models, either to prediction the outcome (outcome regression model) or to predict the exposure (propensity score). Using computer simulations, Drake (*Effects of misspecification of the propensity score on estimators of treatment effects. Biometrics, 1993;49:1231-1236*) and Cepeda et al. (*Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol 2003;158:280-287.*) demonstrated that incorrect modeling assumption may have less influence in a propensity score analysis. Cepeda et al. also suggested that a propensity score analysis is superior to an outcome model analysis when the number of outcome events per confounder is small (< 7 events/confounder).

Propensity score analyses are used more frequently that in the past and provide an alternative method for controlling confounding, compared to the commonly used outcome regression model. However, the propensity score analysis (nor the traditional analysis) controls for unmeasured confounders, unless they tend to be highly correlated with those measured confounders present in the model. Even after correctly controlling for multiple confounders by a propensity score analysis (or by an outcome model), the reported adjusted measure of association can still be confounded by other unknown or unmeasured confounders (as might be the case with the adjusted measures of association for the RHC example).

Screening

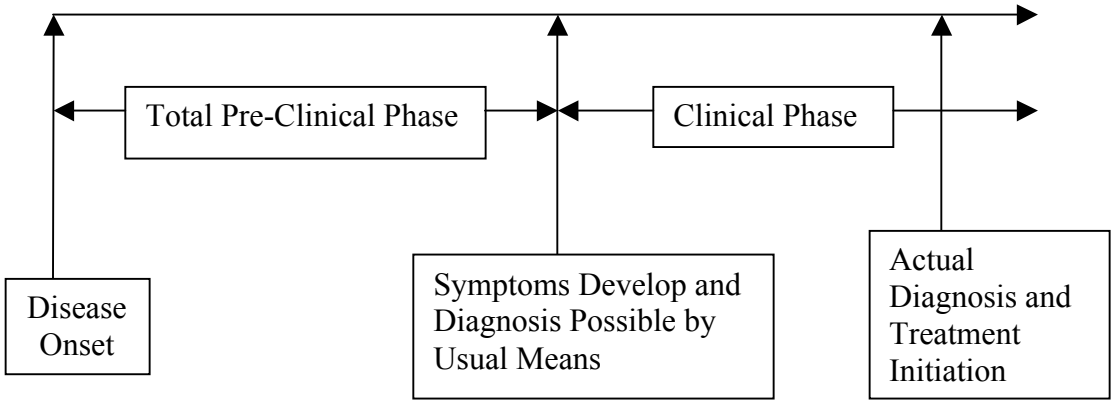
The goal of a screening program is to apply a **simple and inexpensive test** to a large number of persons in order to classify them as likely or unlikely to have a disease of interest. The ultimate goal is to reduce the morbidity and mortality from that disease in the persons that are screened by detecting disease at an earlier stage, where a treatment might have a more beneficial effect. The components of a beneficial screening program include a suitable

1. Disease
2. Test
3. Population

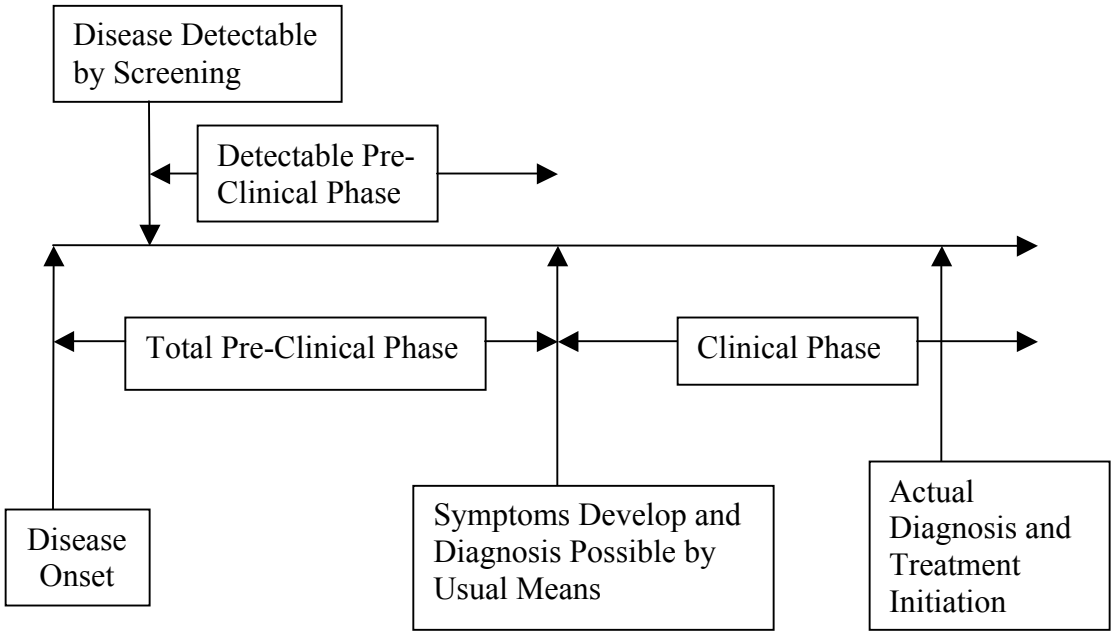
Time Periods

` A disease that is suitable for a screening program should have **serious consequences** (e.g. fatal or severe/prolonged morbidity) to merit the time and cost as the target of a screening program. It must **have a treatment** that, when applied to a case of disease that is detected by screening, is more effective than treatment applied after symptoms when the case is detected by usual means. Finally, the disease should have a **high prevalence in the Detectable Preclinical Phase** (**DPCP).**

Clinical disease begins with the development of signs or symptoms that would lead to a diagnosis of that disease through usual means. Pre-Clinical Disease refers to the existence of disease that has not advanced to a stage where it could be detected by usual means. The **Total Preclinical Phase** begins with the first development of disease (biologic onset) and ends with the development of signs/symptoms and the diagnosis of disease by usual means. These time periods are depicted in the following figure:



The **Detectable Preclinical Phase (DPCP)** refers to a portion of the TCPP and begins when the disease can be detected by the screening test. The length of the DPCP depends on the screening test's ability to detect the disease before signs and symptoms develop. A long DPCP enhances a screening program's chances to detect disease well before it would normally be diagnosed. The following figure shows the timing of the DCPC.



Screening Test

A suitable screening test is one that accurately detects the presence of absence of pre-clinical disease. This is usually measured by two performance measure

1. Sensitivity = P(Test + | Pre-Clinical Disease Exists)

2. Specificity = P(Test - | Pre-Clinical Disease Does Not Exist)

These measures can be estimated from the following 2x2 table

	Pre-Clinical Disease		
	Present	Absent	Total
Test +	A (True Positives)	B (False Positives)	A+B
Test -	C (False Negatives)	D (True Negatives)	C+D
Total	A+C	B +D	

Sensitivity = A/(A+C)

Specificity = D/(B+D)

In addition to having a high Sensitivity and Specificity, a suitable screening test should be low in cost, painless, and not cause morbidity or mortality.

As an example, shows the performance of a screening test of physical examination and mammography for the detection of breast cancer from the Health Insurance Plan (HIP) of Greater New York (Shapiro et al. *Lead Time in Beast Cancer Detection and Implications for Periodicity of Screening*. Am J Epidemiol 100:357-66. 1974 and Hennekens and Buring. *Epidemiology in Medicine* Little Brown 1987 **),**

	Pre-Clinical Disease		
	Present	Absent	Total
Test +	132	983	1115
Test -	45	63650	63695
Total	177	64633	64810

Sensitivity = 132/177 = 0.75
Specificity = 63650/64633 = 0.98

The Sensitivity and Specificity of a Test are characteristics of the screening test. Therefore, one might expect that these values would not change is a screening test were applied to different populations. However, it may be possible that the screening test's ability to detect pre-clinical disease may depend on the severity of the pre-clinical disease. For example, pre-clinical disease states that are about to become clinical disease may be more easily detected by the screening test than less advanced pre-clinical disease. Therefore, the value for the sensitivity of a test may vary of populations that differ in severity of the cases of pre-clinical disease.

Screening Program

A Screening Program involves using a particular screening test in a particular population of asymptomatic individuals. Process measures that reflect the suitability of a screening program include process measures of the screening test (Sensitivity and Specificity) as well as the following

1. Number of people examined
2. Detected prevalence of disease in DPCP
3. Cost
4. Follow-up treatment and outcome
5. Positive Predictive Value of the Test
6. Negative Predictive Values of the Test

The Positive (PV+) and Negative (PV-) Predictive Values of a test refer to the tests ability to predict the presence and absence of the disease. These measures are defined as

Positive Predictive Value = P(Pre-Clinical Disease Present| Test +)

Negative Predictive Value = P(Pre-Clinical Disease Absent| Test -)

The following table shows these values for the HIP data shown above

	Pre-Clinical Disease		
	Present	Absent	Total
Test +	132	983	1115
Test -	45	63650	63695
Total	177	64633	64810

PV+ = 132/1115 = 0.12
PV- = 63650/63695 = 0.999

The Positive and Negative Predictive Values are **posterior probabilities**. They are predictions of an outcome (pre-clinical disease) that take data (test results) into account. They also depend on the base prevalence of disease in the screened population (**prior probability**). Therefore they depend on characteristics of the screening tests (Sensitivity and Specificity) and also characteristics of the population being screened (prevalence of pre-clinical disease). These dependencies are demonstrated by the following expression of Bays Theorem.

(PV+)/(1-(PV+))= [P(D)/(1-P(D))][sensitivity/(1-specificity)]

(PV-)/(1-(PV-)) = [(1-P(D))/P(D)][specificity/(1-sensitivity)]

Posterior Odds = (Prior Odds)(Likelihood Ratio)

These formulas demonstrate that Positive Predictive of Test depends on prevalence of preclinical disease in a population. Screening in a high risk population will results in higher PV+ and enhanced the suitability of a screening program. One means to increase the prevalence of pre-clinical disease in DPCP of a population is to restrict enrollment of participants in the screening program to those with one of more risk factors for the disease. Another option is to screen a population at an optimum frequency. An initial screen of a population will remove prevalent cases of detectable pre-clinical

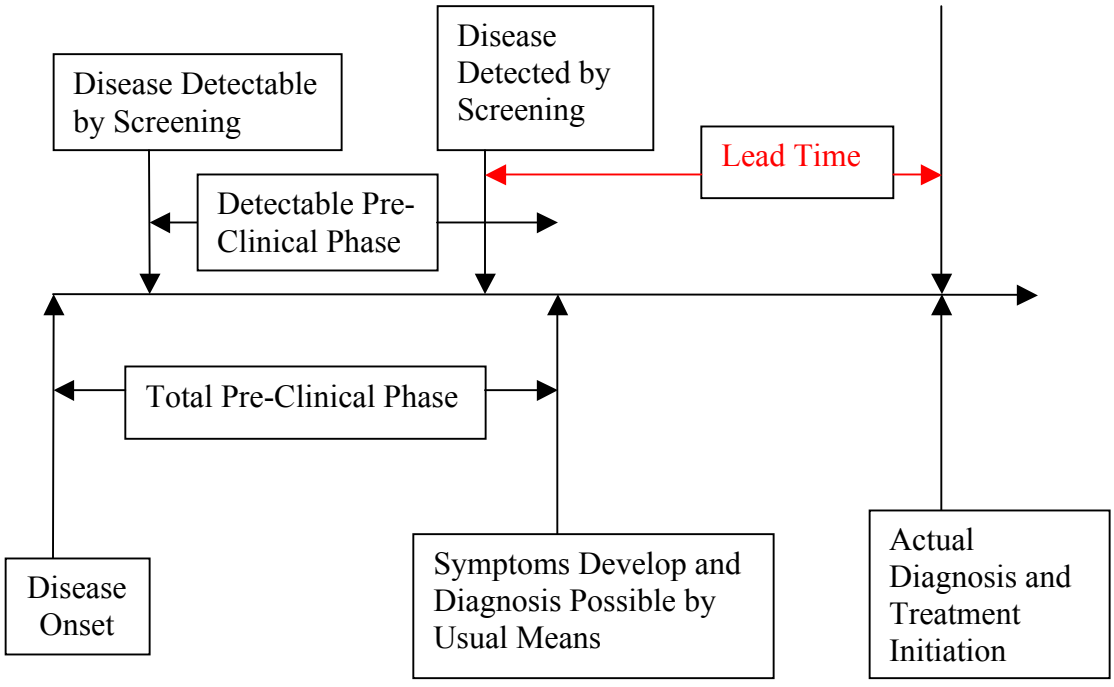
disease. Sufficient time should elapse for incident cases of pre-clinical disease to develop in that population before it re-screened.

Evaluation

The ultimate goal of a screening program is to reduce the morbidity and mortality from that disease in the persons that are screened by detecting disease at an earlier stage, where a treatment might have a more beneficial effect. Therefore, the effect of a screening program can be detected by comparing outcomes from participants in a screening program to similar other individuals who were not part of a screening program. This can be done using the study design options that were discussed in previous lecture notes. For example, with an experimental design, study participants can be randomized to participate in a screening program or to usual care, then mortality rates in each group can be compared from the point of randomization.

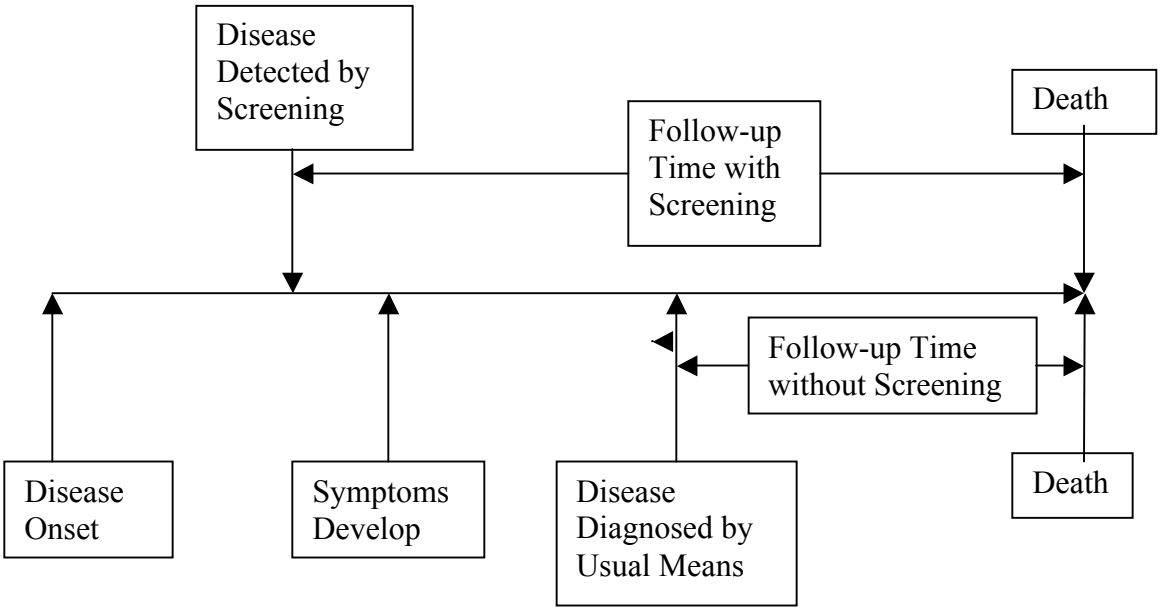
On the other hand, evaluation a screening program with a non-experimental Cohort Study has some potential problems. Individuals who agree to participate in a screening program may not be comparable to those who refuse to participate, providing a potential for confounding. In addition the potential for **Lead Time Bias** and **Length Bias** can be considered.

Lead Time is the additional time an individual lives with knowledge of disease because of earlier diagnosis from screening. A long Lead Time is desirable property for a screening program if early treatment results in better outcomes. The following figure depicts the lead time from a screening program.



Lead Time Bias occurs when the follow-up time for the screened participants does not account for the Lead Time, since individuals in the comparison group will not benefit from a lead time. For example, suppose follow-up for all individuals began at the time of diagnosis. For the screening group follow-up would begin at the time of screen-detected disease diagnosis. This group would show longer follow-up times (and lower incidence rates of death) even if treatment (at any time) had no effect, because of the benefit of Lead Time.

This is depicted in the following figure. The bottom part of the figure describes the life course of an individual without screening. The top part of the figure describes the life course with early detection of disease through screening. However, the figure assumes that early treatment has no benefit and that the individual dies at the same time under both scenarios.



Length Bias occurs when evaluating a screening program because of the expectation that screening may detect prevalent cases of disease in the Detectable Pre-Clinical Phase that have a favorable prognosis. Since prevalence is a function of incidence and duration, an initial screening in a population may detect cases of pre-clinical disease with a long duration. Some of these cases may never develop into clinical cases of disease. Cases of disease detected from a screening program may have better prognosis that those detected by usual means. Therefore, better outcomes in the screened groups may not reflect the effect of the screening program but rather the types of cases that are detected by screening. This potential bias lessens when evaluating the second application of a screening program, since many of the slow growing cases of disease may have been detected by the initial screen.

Ethical issues may also be important to consider when implementing a screening program. For example, screen-detected cases are often subjected to more invasive diagnostic tests or might be treated with therapies with potential serious side-effects. False positive cases receive no benefit from such additional testing or treatment. In addition, the impact on quality of life for true positive cases might not be worth the extra testing and treatment if the expected benefit from early treatment is minimal. For this reason, screening for disease in the very elderly might not be desirable.

Clinical Prediction Rules

Prediction is an integral part of clinical medicine. Prognostic models quantify the likelihood of developing (prognosis) or possessing (diagnosis) an outcome of interest. A **prediction rule** is an algorithm for estimating this likelihood based on values of selected predictors for this outcome. Thus, the main motivations for developing clinical prediction rules include:

1. Predicting a future state of health based on information that is currently available (**Prognosis**), and

2. Predicting a current state of health that is not easily observable based on other information that is available and more easily observable (**Diagnosis**).

Examples of Clinical Prediction Rules

Although a clinical prediction rule could be based on expert opinion, usually these rules are based on empirical relationships (associations in a data set) between other identified predictors and the outcome. Some Common types of clinical prediction rules include:

1. Stratification Patterns
2. Regression Models
3. Point Scoring Systems
4. Neural Networks
5. Other Data Mining Methods

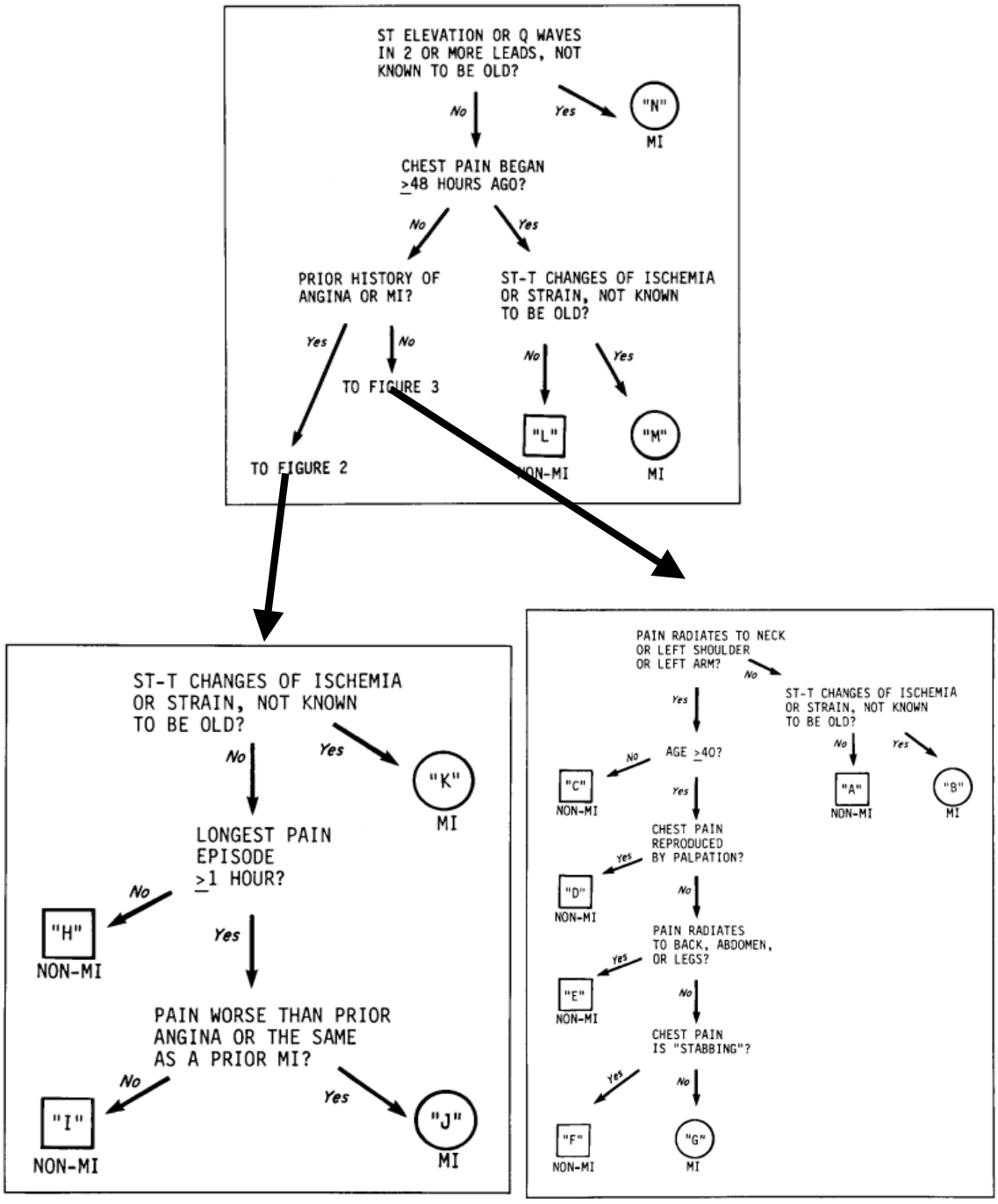
The choice of which type of prediction rule to adopt for a given situation depends on a number of factors, including the number and nature of candidate predictors, the size of the data set, the interrelationships among the predictors, and the potential acceptability of the rule by clinicians.

Stratification Patterns involves grouping the subjects into subgroups (strata) defined by combinations of categories of categorical predictors. The stratum-specific cumulative incidence or the prevalence of the outcome is used as the estimated risk for all patients within that stratum. When the number of categorical predictors is large, then cross-stratification by all predictors usually results in too many subgroups with insufficient number of subjects within most subgroups to provide accurate risk estimates.

Recursive Partitioning (**Classification Trees**) creates an asymmetric stratification pattern by applying a step-wise stratification algorithm to sequential divide subgroups of subjects into smaller subgroups with different estimated outcome risks. This method initially divides the data into two subsets by the single "best" predictor. Each subset in then divided into two additional subsets by the best predictor for that subgroup. The process continues in a recursive manner until there is no evidence that further division of each existing subset would improve prediction. The resulting stratification is

asymmetric, not only in that some portions of the data set undergo more levels of stratification than others, but also that different predictors can be used in different parts of the stratification patterns to determine the subsets. An example of a prediction rule based on a clinical prediction rule in presented in the following figure.

Figure. Recursive Partitioning algorithm for prediction the incidence of myocardial infarction for patients who present to an emergency department with a chief complaint of chest pain. (Goldman L, et al. *A compute protocol to predict myocardial infarction in emergency department patients with chest pain*. *N Engl J Med* 1988;318:797-803)



The classification tree presented in this figure involves 11 different binary predictors. Cross-stratification by all 11 binary factors would have resulted in 2,048 strata, rather than the 14 that are displayed in that figure. Therefore, Recursive Partitioning retains much of the simplicity of stratification while avoiding the problems of having too many strata from cross-stratification by all predictors.

The performance of the partitioning tree displayed in the following table from reference (Goldman L, et al. *A compute protocol to predict myocardial infarction in emergency department patients with chest pain*. ***N Engl J Med*** 1988;318:797-803).

Table 1. Myocardial Infarctions as Predicted Retrospectively (1379 Patients) and Prospectively (4770 Patients) by the Computer Protocol.

Patient Subgroup*	Retrospective Group	Prospective Groups						
		University Hospitals		Community Hospitals				Total
		1	2	1	2	3	4	
		number of infarctions/total number of patients (percent)						
A	1/259 (0.4)	5/290	4/288	11/380	2/145	0/66	2/49	24/1218 (2)
B†	2/22 (9)	9/27	7/43	10/45	9/18	3/10	1/7	39/150 (26)
C	0/40	0/31	1/35	2/29	0/16	0/9	0/4	3/124 (2)
D	0/25	0/32	1/25	0/13	0/3	0/2	0/4	1/79 (1)
E	0/17	0/11	2/19	1/14	1/10	1/5	0/6	5/65 (8)
F	1/20 (5)	0/14	0/15	1/21	0/2	0/4	0/1	1/57 (2)
G†	30/91 (33)	12/69	4/55	24/104	3/19	4/24	4/23	51/294 (17)
H	0/96	5/100	3/99	5/132	1/30	1/26	1/17	16/404 (4)
I	3/57 (5)	0/39	0/49	0/68	2/20	0/11	0/7	2/194 (1)
J†	17/118 (14)	14/86	5/63	7/80	3/27	2/35	3/13	34/304 (11)
K†	38/152 (25)	27/119	7/61	27/96	10/20	11/25	7/22	89/343 (26)
L	1/234 (0.4)	2/338	0/327	8/201	4/101	3/45	0/27	17/1039 (2)
M†	8/50 (16)	12/50	3/22	8/40	4/12	4/21	1/4	32/149 (21)
N†	158/198 (80)	60/82	20/36	102/132	34/38	32/46	14/16	262/350 (75)
	259/1379 (19)	146/1288 (11)	57/1137 (5)	206/1355 (15)	73/461 (16)	61/329 (19)	33/200 (17)	576/4770 (12)

Each row of this table refers to a subgroup in the partitioning tree in the previous figure. The first column of table reports the cumulative incidence of MI (estimated risk) in each of the subgroups in the data set that created the partitioning tree (**Training Set** or **Derivation Set**). The last column reports the cumulative incidence for each subgroup using a different data set (**Testing Set** or **Validation Set**) that were not used in the construction of the partition tree.

In general, there is high agreement between risk estimates from the two data sets. However, the reported risk from the training set for a low- group (e.g. subgroup A: 0.4%) underestimates what is reported for that subgroup in the testing set (2%). Similarly, the reported risk in the training set for a high-risk subgroup (e.g. subgroup N: 80%) overestimates what is reported for that subgroup in the testing sets (75%). This reflects the potential problem (**overtraining**) of using a single data set to create a prediction rule and then using that same data to describe the performance of the rule in future patients.

The estimates in the first column of the previous table 1 are **optimistic** in that they underestimate the risk for low-risk subgroups and overestimate the risk for high-risk subgroups. When the training set is used to find the "**best"** fitting prediction rule for that training set data set, then chances are that this rule will not perform as well in a testing set. Therefore is it important to have a more valid way to estimate the performance of a rule on future patients, and using an independent testing set if often the desired approach.

When a testing set is not available, then re-sampling methods such as **bootstrapping** and **cross-validation** (described later in these notes) provide alternative methods for assessing the optimism of a prediction rule.

The most common method for developing clinical prediction is a **Regression Model**, describing the risk (P) of being in the outcome category of interest as a function of the predictors. For example, the Framingham Risk Model is a prediction rule to estimate the 10-Year Risk for developing Coronary Heart Disease based on six risk factors: age, sex, diabetes, smoking, cholesterol and blood pressure. The model was based on a Cox Regression Model that was fit to 2489 men and 2856 women from the original Framingham Cohort Study and from the Framingham Offspring Study. The model incorporated categories of total cholesterol, LDL cholesterol and HDL cholesterol as defined by the National Cholesterol Education Program (NCEP). Blood pressure was also represented by categories by the Joint National Committee (JNC-V). (Wilson PWF, et al. *Prediction of coronary heart disease using risk factor categories.* Circulation 1998:97:1837-47). The estimated risk of a subject is determined by substituting that subject's values for the risk factors into the regression model.

An example of a prediction rule based on a logistic regression model is the Mortality Prediction Model (MPM). This rule predicts the risk of in-hospital death for patients admitted to an intensive care unit (Lemeshow S, et al. *Predicting the Outcome of Intensive Care Unit Patients.* J Am Stat Assoc 1988;83(402):348-356.). Panel A of the following table displays the seven predictors that are used in this rule. As is typically the case, these predictors were chosen from a larger pool of potential predictors. Panel B of this table displays the formula for the Logistic Regression Model that defines this prediction rule. Panel C demonstrates the calculation of the estimated risk of death from this model for a subject with specified values for the predictors in the model.

Table Mortality Prediction Model (MPM) for Predicting In-Hospital Mortality among Patients admitted to an Intensive Care Unit.

A. Predictors

CONS	Level of Consciousness (1 if coma or deep stupor, 0 otherwise)
TYPE	Type of Admission (1 if emergent, 0 if elective)
CANCER	Cancer as Part of Present Problem (1 if yes, 0 if no)
CPR	Prior CPR (1 if yes, 0 if no)
INFECT	Infection (1 if probable, 0 otherwise)
AGE	Age in Years
SBP	Systolic Blood Pressure
SBP2	SBP squared.

B. Prediction Rule

$$\log(P/(1-P)) = -1.370 + 2.44(CONS) + 1.81(TYPE) + 1.49(CANCER) + .974(CPR) + .965(INFECT) + .0368(AGE) -.0606(SBP) + .000175(SBP2)$$

C. Sample Estimated Risk Calculation

CONS	= 1 (patient has coma or deep stupor)
TYPE	= 1 (emergent admission)
CANCER	= 0 (cancer part of present problem
CPR	= 0 (no prior CPR)
INFECT	= 0 (no probable infection)
AGE	= 50 (50 years of age)
SBP	= 150 (systolic blood pressure = 150)
SBP2	= 22500 (SBP squared)

$$\text{Log}(P/(1-P)) = -.1370 + 2.44 + 1.81 + .0368(50) - .0606(150) + .000175(22500) = .8005$$

$$P/(1-P) = \exp(.8005) = 2.227$$

$$P = 2.227/3.227 = 0.69$$

Panel C demonstrates the amount of calculations needed to obtain risk estimates from a logistic regression model. Although easy to program on a computer, the amount of calculations may limit the use of such a model in actual practice. However, when the predictors are binary in scale and the regression coefficients are proportional in scale, the predictive information in a regression model can be approximated by a simpler **Point Scoring System.** A scoring system assigns a weight (number of points) to each predictor. The weights are proportional to the regression coefficient of the predictors. Ranges of the sum of weights define risk categories.

As an example of a scoring system, Panel A of the following table displays the predictors that were chosen for a logistic regression model to predict the probability of bacteremia in blood samples of 1007 hospitalized patients (Bates DW, et al. *Predicting bacteremia in hospitalized patients*. Ann Intern Med 1990;113:495-500). Panel B displays the logistic regression model that defines the prediction rule. The regression coefficients in the logistic regression model show a pattern of similarity. For example, the coefficients for TEMP, CHILLS, POSEXAM, and COMORB are similar in value. Dividing the coefficients in the model by .32 and rounding these values to the nearest integer results in the scoring system displayed in Panel C of the table.

Table Scoring System to Predict the Risk of Bacteremia among Blood Tests.

A. Predictors

TEMP	Indicator variable for having a maximum temperature $\geq$ 38.3 C (1 = yes, 0 = no)
DCLASS2	Indicator variable specifying that a subject's diagnosis falls in a predefined category of rapidly fatal diseases (1 = yes, 0 = no)
DCLASS3	Indicator variable specifying that a subject's diagnosis falls in a predefined category of ultimately fatal diseases (1 = yes, 0 = no)
CHILLS	Indicator variable for the presence of chills (1 = yes, 0 = no)
DRUG	Indicator variable for intravenous drug abuse (1 = yes, 0 = no)
POSEXAM	Indicator variable for a positive focal abdomen examination (1 = yes, 0 = no)
COMORB	Indicator variable for a having one of a specified set of comorbid conditions (1 = yes, 0 = no)

B. Prediction Rule

$$\log(P/(1-P)) = -4.14 + .91*TEMP + 1.40*DCLASS1 + .65*DCLASS2 + .96*CHILLS + .287*DRUG + 1.03*POSEXAM + .96*COMORB$$

C. Scoring System

$$\text{Number of Points} = 3*TEMP + 4*DCLASS1 + 2*DCLASS2 + 3*CHILLS + 4*DRUG + 3*POSEXAM + 3*COMORB$$

The intercept term in the original regression model (Panel B of the previous table) is not included in the scoring system, since this would only add a constant to the total number of points calculated for any subject. The scoring system was used to develop a classification rule (shown the following table) based on ranges of points.

Table: Classification rule based on an Integer-Based Scoring System for predicting the presence of bacteremia among hospitalized patients.

	Risk Score			
	0-2 Points	3 Points	4-5 Points	$\geq$ 6 Points
Training Set (n=1007)	4/303= 0.01	11/236 = 0.05	18/204 = 0.09	41/264=0.16
Testing Set (n=509)	3/155=0.02	8/121=0.07	9/88=0.10	21/145=0.14

An examination of the results for the Training Set in this table shows that the risk of bacteremia ranges from 1% (4/303) in patients with 0-2 points to 16% (41/264) in patients with ≥ 6 points. However, as mentioned previously the performance of the prediction rule in these patients may be optimistic and over-estimate the performance in future patients. This is seen measuring the performance of the rule in a Testing Set of 509 different patients. The estimated risk of patients with 0-2 points is low (3/155 = 2%), but not as low as in the Training Set. Similarly, the estimated risk of patients with ≥ points is high (21/145 = 14%), but not as high as in the Training Set.

Although the integer-based scoring system in Panel C of the previous table may be easier to use than the original logistic regression model in Panel B, the gain in simplicity has its price. First, the regression coefficients from Panel B undergo some degree of rounding to create the weights of the scoring system, thereby losing some of the predictive information incorporated in the original model. In addition, because the weights in Panel C only reflect the relative importance of the predictors, they can no longer be used to generate estimated risks for the outcome. Categories defined by low number of points have lower risk of developing the outcome. However, the estimated risk for any category must be based on the cumulative incidence of the outcome among the subjects in that category .

Evaluation of a Clinical Prediction Rule

Performance Measures

The validity of a prediction rule is often quantified by various performance measured by how well its estimates agree with the actual outcomes of subjects. Measures of **discrimination** refer to the ability of the prediction rule to separate subjects with different outcomes into categories according to their values for the prediction rule. Measures of **calibration** refer to the ability of the model's estimated risk to agree with actual outcomes within groups of subjects.

Measures of Discrimination

One common method for assessing a prediction rule's ability to discriminate between the categories of a binary outcome is to determine the distribution of the outcome categories when subjects are ranked by their estimated risks. Ideally, cases of the outcome will tend to have high ranks of estimated risk, while non-cases will ten to have low ranks. The most common way to display the separation of outcome positive subjects from outcome negative subjects in this ranking is to calculate the ROC curve (Receiver Operating Characteristic). This curve is created by defining a classification rule based on a threshold of estimated risk. Subjects above that threshold are classified as

"high risk" subjects and subjects below that threshold are classified as "low risk" subjects. A **classification table** can then be displayed comparing the categories of the classification rule to those of the actual outcome. The following table displays the format of a confusion matrix along with the formulas for the **sensitivity** and **specificity** of the classification.

Table: 2x2 Table Displaying the Validity of a Binary Classification Rule.

	Outcome +	Outcome -
High-Risk	A	B
Low-Risk	C	D
Total	A+C	B+D

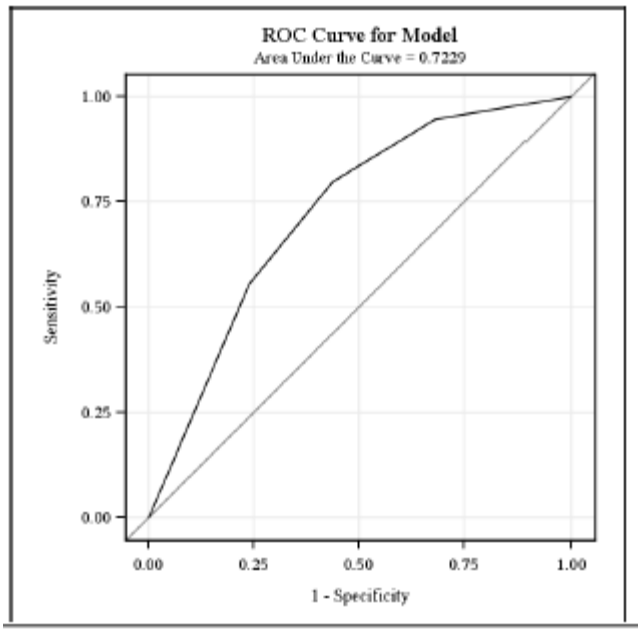
Sensitivity = A/(A+C)
Specificity = D/(B+D)

Varying the threshold for defining high risk generates a series of tables like that displayed in the previous table, with corresponding values for sensitivity and specificity. The ROC curve depicts the overall relationship between the prediction rule and the outcome by graphing the value for the sensitivity from each classification table on the vertical axis and for (1-specificity) on the horizontal axis. An ideal curve is one with points in the top left-hand region of the graph, reflecting a classification rule with high sensitivity and high specificity. Since an axis of the ROC curve ranges from 0.0 to 1.0, the total area in a box bounded by these axes is 1.0. Therefore, the ideal curve is one whose area under the curve (AUC) is close to 1.0. The following table shows the classification tables for the scoring system presented in a previous table for the bacteremia data according to three different thresholds of high risk.

Table: Classification tables showing the risk of bacteremia according to different definitions of high risk for the data

	Bacteremia		Sensitivity	Specificity
	+	-		
High Risk (3 + points)	70	634		
Low Risk (0-2 points)	4	299		
Total	74	933	70/74=0.95	299/933=0.32
High Risk (4 + points)	59	409		
Low Risk (0-3 points)	15	524		
Total	74	933	59/74=0.80	524/933=0.56
High Risk (6 + points)	41	264		
Low Risk (0-5 points)	33	669		
Total	74	933	41/74=0.55	669/933=0.72

The following figure shows the ROC Curve and it corresponding area for the 1009 subjects in the training set for the bacteremia data set.



The area under the ROC curve (AUC) is 0.72. However, the ROC curve for the testing set in the bacteremia data has an AUC 0.69. The difference in area (0.72 – 0.69 = 0.03) estimates the amount of optimism that is obtained in the value from the AUC based on the model's performance in the training data set.

Measures of Calibration

Calibration refers to the degree of agreement between a subject's estimated outcome from a prediction rule and the subject's actual outcome. Measures of the degree of calibration commonly take on the form of "observed versus expected" comparisons of the outcome. A common approach is to assign subjects to risk categories according to sub-ranges of predicted risk. Within each risk category, the average predicted risk is compared to the observed cumulative incidence of the outcome. Alternatively, the sum of the predicted risks in a category provides an estimate of the expected number of outcomes for that category (E), which can be compared to the actual number of outcomes for that category (O).

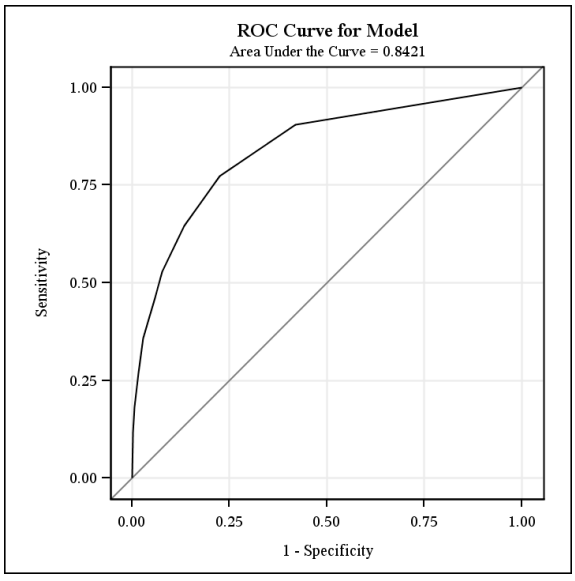
A common summary measure of calibration for binary outcomes is the Hosmer and Lemeshow goodness-of-fit statistic (Hosmer DW, Lemeshow S. *Applied Logistic Regression Second Edition.* John Wiley & Sons; New York: 2000). An example of a calibration table is given in the following table, which shows the performance of the MPM prediction algorithm described in a previous table. This algorithm was developed on an initial data set of 775 admissions to an intensive care unit. The results listed in the following tables describe the performance of this algorithm in another set of 1997 admissions to the intensive care unit.

Table :. Calibration of the Mortality Prediction Model (MPM) in 1997 patients admitted to an intensive care unit.

Range of MPM P(Dying)	Number of Subjects (N)	Deaths Observed	Deaths Expected
.00 - .09	967	38	41.9
.10 - .19	365	52	52.2
.20 - .29	194	50	48.2
.30 - .39	139	47	48.8
.40 - .49	88	41	39.0
.50 - .59	56	26	30.0
.60 - .69	56	35	35.9
.70 - .79	48	35	36.0
.80 - .89	35	26	29.5
.90 - 1.0	49	46	46.7
Total	1997	396	408.2

$X^2_{HL} = 4.94$, $p = .90$

In general, the previous table shows good agreement between expected and observed outcomes for each risk category, suggesting good calibration of risk estimates. Furthermore, the following ROC curve and its corresponding AUC show good discriminating ability of the risk estimates.



Determining the Predictors to Include in a Prediction Rule

Often the pool of potential predictors is large and the choice is to include all or a representative subject of the predictors in a model. The optimal number of predictors to include in a rule should be guided by the amount of information that is contained in the data set. For a binary outcome, a very rough rule of thumb for model stability is to require a minimum number of subjects in each outcome category for every predictor considered for the analysis. The suggested minimum number of subjects has ranged from 5 to 10 (Wasson JH, et al. *Clinical prediction rule: application and methodologic standards.* N Engl J Med 1985;313:793-799.). For continuous outcomes, the suggested rule of thumb is to require 10 subjects for every candidate predictor (Harrell FE, et al. *Regression modeling strategies for improved prognostic prediction*. ***Stat in Med*** 1984).

When the number of potential predictors exceeds the limit suggested by these guidelines then a model based on all predictors is not only unstable and complex but is also likely to be optimistic and not generalize to other data sets. One solution to this problem is to select only a subset of the predictors for the model based on associations found in the data.

Parsimony

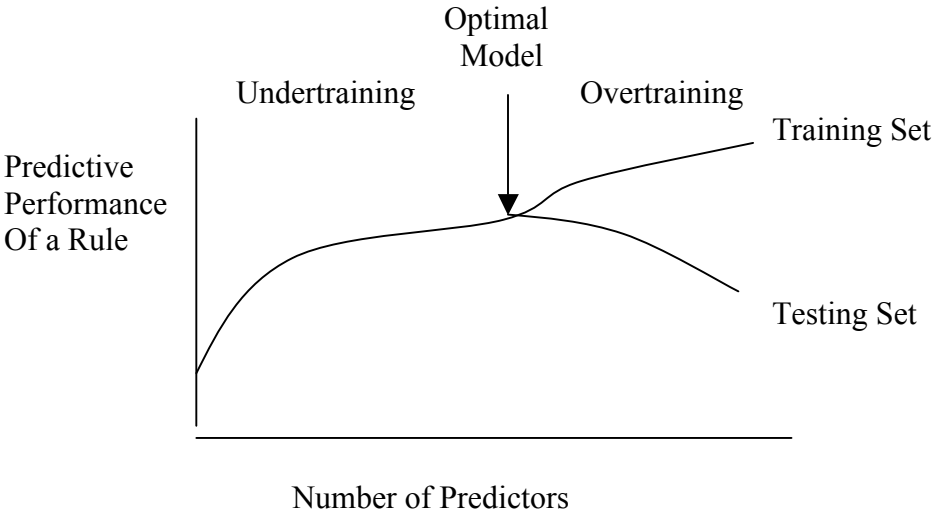
Parsimony pertains to the number of predictors to include in a prediction rule. **Occam's Razor (**William of Occam) states that "*Pluralitas non est ponenda sine neccesitate*" (plurality should not be posited without necessity). **Albert Einstein** stated a similar theme when saying “Everything should be made as simple as possible, but no simpler”.

For model building, parsimony implies including only those factors that are true predictors, not only in the data on which the model is created (training set) but also on other data sets on which it is to be applied (testing sets). This has direct relevance to models that are created by variable selection algorithms. A **forward selection algorithm** builds a model in steps, each time adding a factor that adds the more statistically significant, incremental predictive information to the model. **A backwards elimination algorithm** builds a model by starting with a large model that contains all of the factors and then reduces the model in steps, each time eliminating the least statistically significant factor.

Although commonly used for developing prediction rules, variable selection algorithms possess serious potential problems. Since they involve a large number of tests of significance, these methods increase the likelihood for overtraining through multiple testing. Although a single non-predictor has a 5% chance of demonstrating a statistically significant relationship to the outcome in a dataset, 10 independent non-predictors have a 40% chance of at least one of them reaching statistical significance. In the extreme, 100 independent non-predictors have a 99% chance of one reaching statistical significance.

Parsimony suggests that the model building process should cease at the point when selected factors would not be true predictors in other data sets. This is demonstrated by following figure showing the expected pattern of performance of series models created by a forward selection algorithm in training data set and evaluated in a testing data set. Models created in the early stage of the selection process tend to be based on predictors whose strong associations tend to generalize to other data sets. Therefore, the performance of these models in the training set also reflects the expected performance in other data sets. However, because of the search to identify factors related to the outcome in the training set, a point is often reached where factors are selected based on associations that are particular to the training set and are not repeated in a testing set. Including such factors in the model results in worse performance in a testing set. This process is often labeled as **overtraining** or **overfitting.** The challenge when using variable selection algorithms is to determine the optimal stopping point to avoid both problems.

Figure Expected change in predictive accuracy as the number of predictors in a prediction algorithm increases.



Validation

Validation of a clinical prediction rule involves obtaining "honest" estimates of the rule's performance in actual practice. Perhaps the simplest means for assessing the validity of a prediction rule is by examining its performance in an independent testing set. This is often implemented by randomly splitting a data set into a training data set and a testing data set. The prediction rule is developed in the training set and validated in the testing data set.

Alternatively, **cross-validation** divides the data set into a series of (usually disjoint but exhaustive) testing sets. For example, **10-fold cross-validation** divides the

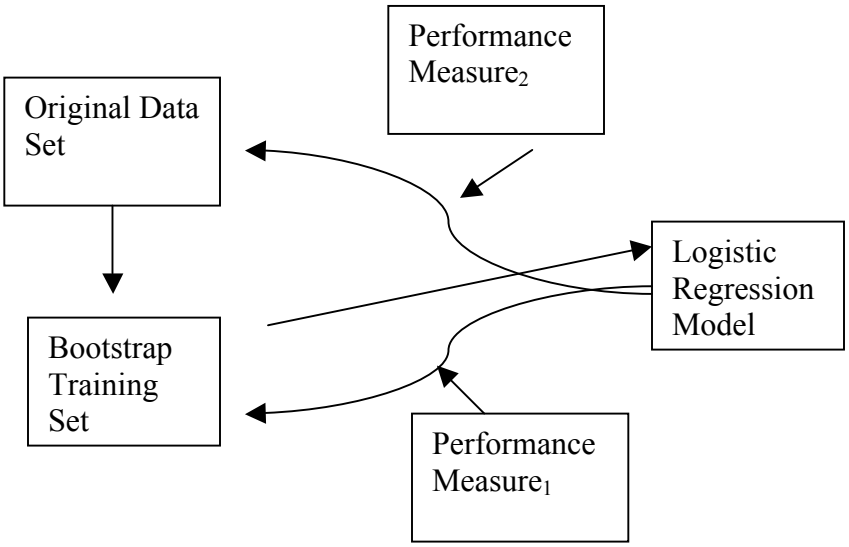
data set into 10 mutually exclusive testing sets with each subject in only one of these sets. A prediction rule is fit in the complement of each of these testing sets (90% of the data) and then evaluated in the corresponding testing set. Finally, the performances of the 10 prediction rules in the 10 testing sets are averaged and used as an estimate of the performance of the single prediction rule built on the entire data set. The following figure presents a graphical display of this method.

Figure: Ten-Fold Cross Validation.

Subset #1	**Subset #2**	**Subset #3**	**Subset #4**	**Subset #5**	**Subset #6**	**Subset #7**	**Subset #8**	**Subset #9**	**Subset #10**
Test Set # 1	Train Set # 1	Train Set # 1	Train Set # 1	Train Set # 1	Train Set # 1	Train Set # 1	Train Set # 1	Train Set # 1	Train Set # 1
Train Set # 2	**Test Set # 2**	Train Set # 2	Train Set # 2	Train Set # 2	Train Set # 2	Train Set # 2	Train Set # 2	Train Set # 2	Train Set # 2
Train Set # 3	Train Set # 3	**Test Set # 3**	Train Set # 3	Train Set # 3	Train Set # 3	Train Set # 3	Train Set # 3	Train Set # 3	Train Set # 3
Train Set # 4	Train Set # 4	Train Set # 4	**Test Set # 4**	Train Set # 4	Train Set # 4	Train Set # 4	Train Set # 4	Train Set # 4	Train Set # 4
Train Set # 5	Train Set # 5	Train Set # 5	Train Set # 5	**Test Set # 5**	Train Set # 5	Train Set # 5	Train Set # 5	Train Set # 5	Train Set # 5
Train Set # 6	Train Set # 6	Train Set # 6	Train Set # 6	Train Set # 6	**Test Set # 6**	Train Set # 6	Train Set # 6	Train Set # 6	Train Set # 6
Train Set # 7	Train Set # 7	Train Set # 7	Train Set # 7	Train Set # 7	Train Set # 7	**Test Set # 7**	Train Set # 7	Train Set # 7	Train Set # 7
Train Set # 8	Train Set # 8	Train Set # 8	Train Set # 8	Train Set # 8	Train Set # 8	Train Set # 8	**Test Set # 8**	Train Set # 8	Train Set # 8
Train Set # 9	Train Set # 9	Train Set # 9	Train Set # 9	Train Set # 9	Train Set # 9	Train Set # 9	Train Set # 9	**Test Set # 9**	Train Set # 9
Train Set # 10	Train Set # 10	Train Set # 10	Train Set # 10	Train Set # 10	Train Set # 10	Train Set # 10	Train Set # 10	Train Set # 10	**Test Set # 10**

Bootstrapping is another method for estimating the validity of a prediction rule in the absence of an independent testing set. It involves re-sampling the original data set with replacement in order to obtain a new "bootstrap" training set of the same size as the original data set. A prediction rule is then developed on the bootstrap training set and a measure of its performance is calculated both on that data set and on the original data set, using the original data set as a testing set for the prediction rule developing on the bootstrap training set. The difference a performance measure on the two data set provides an estimate of the **optimism** of the performance of the prediction rule on the bootstrap training set. This process is then repeated multiple times and the average of the optimism values is used as an estimate of the optimism of a single prediction rule that is developed on the full data set. The following figure presents a graphical display of this method.

Figure Bootstrap estimate of the optimism of the Area of the ROC Curve (AUC) from a logistic regression model.



Optimism = $(\text{Performance Measure})_1 - (\text{Performance Measure})_2$