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."

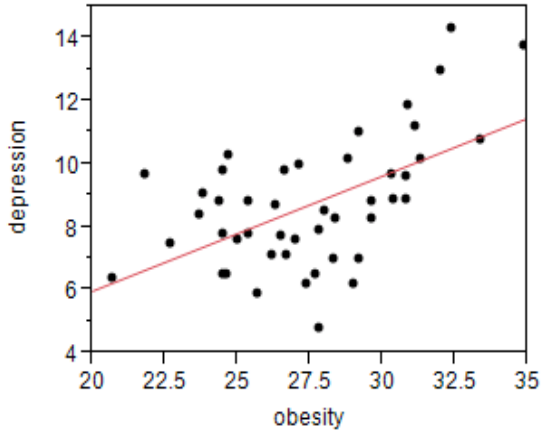
CDC – MMWR 1981 August 28;30:409-10

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.