SYSTEMATIC AND RANDOM ERRORS

OBJECTIVES

After completing this session you should be able to :

1. Define and describe the concept of systematic and random errors.
2. Describe types of systematic error.
3. Define and describe misclassification.
4. Define and identify the confounding variables.
5. List the methods to control confounding in the data.

The overall goal of epidemiologic study is accuracy in measurement of both the factors and outcome of interest. Our objective is to estimate the value of the parameter with little error. The term "error" is defined as a false or mistaken result obtained in a study or experiment (Last, 1988). The sources of error in measurement may be classified as either **random** or **systematic**.

ERRORS = SYSTEMATIC ERROR + RANDOM ERROR

RANDOM ERROR (PRECISION PROBLEM)

The observations about disease occurrence in the population are usually made on a sample of population than all the populations in question. A single set of observations, even if selected in unbiased way, may misrepresent the truth because of error arising from random variation. In fact, actual observations on single sample are unlikely to corresponds exactly to the true state of affairs in larger groups of all populations. However, if the observations were repeated on may such samples, they would be found to vary about the true value.

For example, suppose we have a bag of 100 marbles, of which half are red and half blue and we wish to infer the proportions of the different colored marbles in the entire bags by drawing a sample. If we were to draw only two marbles, there would be one chance in four [$(½)^2$] that both marbles would be blue. This means that 25 percent of the time we would incorrectly conclude that all marbles in the bag were blue given that only half were, based on a sample size of two. By drawing even five marbles instead of two, the probability of observing all five to be blue drops to about 2 in 100 [$(½)^5$]. Thus as the sample since increases, the likelihood that we will incorrectly infer from the sample the true characteristics of the entire population decreases.

Similarly, in epidemiology, investigators rarely evaluate every member of the entire population. Most often, an inference concerning the true relationship between an exposure and disease in made from observations on a sample. Thus, for example, the magnitude of an association between malnutrition and diarrhea among children under five years of age would not be assessed in a case-control design by obtaining anthropometric measurement on every children under five in a community who has had diarrhea and every children who has not. More commonly, a sample of those children with diarrhea would be taken as well as a sample of those who had not experienced diarrhea during the specified

time period, and the nutritional status for those children would be compared. As a result, as in previous example with the marbles, there is always a possibility that the resultant estimate will differ from the true magnitude of association between malnutrition and diarrhea simply because of sampling variation or chance. Again, the smaller the sample on which the inference is made, the more variability there will be in the estimate and less likely the findings will be to reflect the experience of the total population. Conversely, the larger the sample on which the estimate is based, the less variability and the more reliable or precise the inference.

Random error is the fluctuation of an estimate around the population value and is essentially attributed to sampling variation, the extent of which may depend on aspects of the study design (e.g., sample size considerations) and statistical characteristics of the estimator (e.g., its variance). It is a difference between an estimate computed from the study data and parameter actually being estimated (KKM, 1982). In an epidemiologic study, random error has many components, but a major contributor is the process of selecting the specific study subjects. This process is usually referred to as ***sampling***; the attendant random error is known as ***sampling error.*** The estimate of effects obtained from the subset of all study subjects (sample) would differ in value when the entire population were studied.

Precision is epidemiologic measurements corresponds to the reduction of random error (Rothman, 1986). The primary means to increase precision of the study is to increase the size of the study. The way to assess the adequacy of the size of a study is to calculated study size based on statistical "**sample size**" formulas. The precision in measurement also implies and ability to obtain replicate values that differ little because the random error that affects them is small. Concern about precision is concern about random variability of one's measure. The width of the confidence interval also help us to determine the precision of our estimate which depends on the amount of variability in the data. The wider confidence interval generally means lower precision or reliability of the estimate but it also depends on an arbitrarily selected value of level of confidence level (90% CI or 95% CI). Intervals with greater confidence are wider and always include within them any intervals of lesser confidence. The following example would help you to understand better. Consider the two hypothetical sets of study results of the same hypothesized cause-effect relation shown below.

	Study A	Study B
Study Sample	450	90
90% CI (RR)	2.1 – 3.4	1.1 – 10.8
95% CI (RR)	1.6 – 4.3	0.9 – 12.1

Here we observe that study A consisting of large study subjects and gave precise results, that is, a narrow confidence interval, whereas study B was consisting of smaller study subjects and led to a large confidence interval. Study B was considered imprecise when compared to study A. At the same time we also see that as the level of confidence increases from 90% to 95% the width of the interval also increases for both the studies.

SYSTEMATIC ERROR (VALIDITY PROBLEM)

Systematic error is an error due to factors that inherent in the study design, data collection, analysis and interpretation to yield results or conclusions that depart from the truth. It is attributable to methodical aspects of the study design or analysis other than sampling variation, particularly the selection of subjects, the quality of information obtained, and variables of importance other than the disease and study factor of interest. A term "**bias**" is used synonymously with systematic error and is defined as any systematic error in an epidemiologic study that results in an incorrect estimate of the association between exposure and risk of disease. When misrepresentation of effect can be identified, it will be called bias. When there is no misrepresentation we will say that our estimate of effects is "**valid**" or, equivalently, that there is "**no bias**" (KKM, 1982). The validity of study is usually separated into two components: the term internal validity concerns the validity of inference about the target population, using information from the study population; and the validity of the inferences as they pertain to people outside the study population or concerns inferences to an external population beyond the study's restricted interest (external validity or generalizability).

There general types of bias (systematic error) can be identified as follows:

1. Selection bias
2. Information bias
3. Confounding

1. SELECTION BIAS

It is refers to a distortion in the estimate of effect resulting from the manner in which subjects are selected for the study population. Among many sources of selection bias are flaws in the study design, most notably concerning the choice of groups to be compared (in all types of studies) and choice of sampling frame (in case-control and cross-sectional studies). Selection bias can also result in case-control studies when the procedure used to identify disease status varies with exposure status. In follow-up studies, efforts to prevent or to minimize selection bias prior to analysis are primarily involved with ensuring complete follow-up of initial cohort and obtaining as large response rate as possible. When persons lost to follow-up differ from those who remain with respect to both the exposure and outcome, any observed association will be biased. For example, in a cohort study using mailed questionnaires to evaluate the relation between smoking and myocardial infraction (MI), to the extent that those who smoke and develop MI are less (or more) likely to respond and outcome will be obtained. The potential for bias due to losses to follow-up is present no matter how small the population of loss is, as long as such loss is related to both exposure and disease.

2. INFORMATION BIAS

It is referred to a distortion in the estimate of effect due to measurement error or misclassification of subjects on one or more variables. Major sources of information bias include invalid measurement, incorrect diagnostic criteria, and omissions and inadequacies in previously recorded data. Information bias from misclassification may result the follow-up data studies when there is unequal diagnostic surveillance among exposure groups.

The determination of an individual's exposure status of disease status may be subject to error. This can occur for diagnoses based on physical examination or laboratory findings, or for the estimation of past exposures based on recall in personal interviews or abstracted from written records. For example the diagnosis of many tropical diseases such as malaria or dengue hemorrhagic fever based on clinical symptoms alone involves considerable error because there is high percentage of patients who had inapparent infection. In case-control studies, the most likely source of classification error will occur in determination of exposure. Interviewing, technique and the phrasing of question, the ages and educational backgrounds of the respondents, the time interval between the exposures of interest and the interview, and the degree of detail sought will affect the validity of the respondents. For example, a mother may to certain degree or

inaccurately recall the aspects of recent episode of diarrhea of the child such as type of feeding before the illness, sanitation practices, frequency, amount and characteristics of stool, and feeding practices and use of ORS during the episode. The information extracted from the interview would be used to classify the levels of exposure or severity of disease. Estimates of the relative risk will be biased if there is misclassification of either disease or exposure status.

The misclassification is said to be differential if the magnitude of the error for one variable differs according to the actual value another variable (or there is differing sensitivities and differing specificity's of classification of exposure over disease categories and vice versa).

For example, consider a case-control study of the radiation and the occurrence of congenital malformations. If information on exposure to pelvic radiation during pregnancy is obtained retrospectively from reports of the mothers and if the mothers of infants with congenital malformations recall radiation to which they were exposed to a greater extent than do the control mothers, then the measurement error is said to be differential with respect to disease status. In another example, suppose a follow-up study were undertaken to compare incidents rates of emphysema among smokers and nonsmokers. Emphysema may go undiagnosed without usual medical attention. If smokers, because of concern about health-related effects of smoking or as a consequence of other health effects of smoking (such as bronchitis), seek medical attention to a greater degree than nonsmokers, then emphysema might be diagnosed more frequently among smokers than nonsmokers simply as a consequence of medical attention. This is an example of differential misclassification, since the under diagnosis of emphysema, a misclassification error, occurs more frequently for nonsmokers than for smokers.

The bias resulting from differential misclassification may be either toward or away from the null. That is, the apparent magnitude of the measure may be increased or reduced, or the direction of association may even be reversed. Spurious associations may be found between an exposure and a disease when in fact no association exists.

Measurement error is said to be **nondifferential** if the magnitude of error for one variable does not vary according to the actual value of other variables. For a binary variable, measurement error is nondiffential if both sensitivity and specificity remain constant irrespective of the values of other variables. Such nondifferential misclassification has generally been considered a lesser treat to validity than differential misclassification,

since the bias introduced by nondifferential misclassification is always in a predictable direction: toward the null condition.

The effect of nondifferential misclassification bias in case-control study can be illustrated as follows (derived from Breslow & Day, 1980).

Let P_1 = P (E- | case)

P_0 = P (E+ | control)

π_1 = P (E+ misclassified as E-)

(for both cases & control i.e. random misclass)

π_2 = P (E- misclassified as E+)

(for both cases & control i.e. random misclass)

P_1^* = P (case classified as E+) = $P_1(1-\pi_1) + (1-P_1)\pi_2$

P_0^* = P (control classified as E+) = $P_0(1-\pi_1) + (1-P_0)\pi_2$

$$\text{True OR} = \frac{P_1(1-P_0)}{P_0(1-p_1)}$$

$$\text{Misclassified OR} = \frac{(P_1+\pi_2/d)\{(1-P_0)+\pi_2/d\}}{(P_0+\pi_2/d)\{(1-P_1)+\pi_1/d\}}$$

Where $d = 1 - \pi_1 - \pi_2$

True situation $P_1 = 0.3$, $P_0 = 0.1$

		Case	Control	
Exposure	+	30	10	40
	-	70	90	160
		100	100	200

Odds ratio = 3.86 (1.67 < OR < 9.10)

Misclassified OR A. $\pi_1 - \pi_2 = 0.1$, then $P_1 = 0.34$, $P_0 = 0.18$, d = 0.8

		Case	Control	
Exposure	+	34	18	52
	-	66	82	148
		100	100	200

Odds ratio = 2.35 (1.67 < OR < 4.78)

B. $\pi_1 - \pi_2 = 0.2$, then $P_1 = 0.38$, $P_0 = 0.26$, d = 0.6

		Case	Control	
Exposure	+	38	16	64
	-	62	74	136
		100	100	200

Odds ratio = 1.74 (0.92 < OR < 3.33)

From the above example we see that when there is nondifferential misclassification of exposure status among the cases and controls then the magnitude of odds ratio is deflated or there is bias toward the null value (which is 1).

REMEMBER		
Non differential misclassification	→ →	Bias toward the null value
Differential Misclassification	→ →	Bias toward or away the null value

3. CONFOUNDING

It is a bias that results when a study factor effects is mixed, in the data, with the effects of effects of extraneous variables or the third variables. The effect of an extraneous variable may wholly or partially accounts for the apparent effect of the study exposure, or masks the underlying true association. Thus, an apparent association between an exposure and disease may actually be due to another variable. Confounding

can lead to an overestimate or underestimate of the true association between exposure and disease and can even change the direction of the observed effect.

To be confounding the extraneous variable must have the following characteristics:

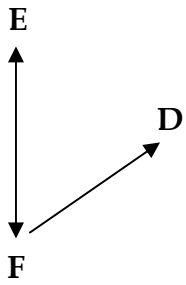
a. A confounding variable must be a risk factor for the disease (or associated with the outcome independently of exposure).
b. A confounding variable must be associated with exposure under study.
c. A confounding variable must not be an intermediate step in the causal path between the exposure and the disease.

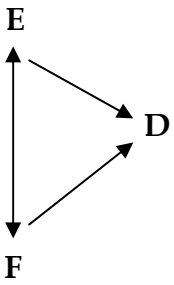
For example, consider a study that showed the relationship between increased level of physical activity and decreased risk of myocardial infarction (MI). One additional variable that might affect the observed magnitude of this association is age. People who exercise heavily tend to be younger, as a group, than those who do not exercise. In this circumstance, age would confound the observe association between exercise and MI.

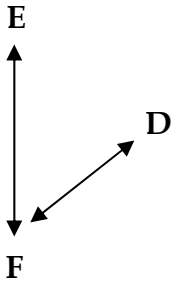
Another example, suppose that we are investigating the effectiveness of an educational program aimed at improving the reading ability of elementary school children. Two classes are being compared, one receiving the new program and one utilizing the standard curriculum (existing program). Suppose that the class receiving the new program contains higher proportion of poor children. Then if poverty is thought to have direct influence on reading ability, it can be considered a confounding factor.

Third example would make you understand better. An investigator would like to study the association between alcohol consumption and low birth weight of babies. A variable related to the outcome and not to the exposure will not confound the exposure-outcome association. It is known, for example, that short mothers tend to have small babies. Unless short mothers differ in their drinking habits from taller mothers, however, the association between alcohol consumption and low birth weight will remain unbiased. Cigarette smoking, on the other hand, might indeed confound and association between maternal drinking and birth weight. Not only is smoking known to be an independent risk factor for low birth weight, but it also associated with drinking, since drinkers are more likely to smoke than nondrinkers. Thus, criteria (a) and (b) are both satisfied. Since the causal path by which maternal alcohol consumption reduces fetal growth does not involve cigarette smoking, criterion (c) is also met.

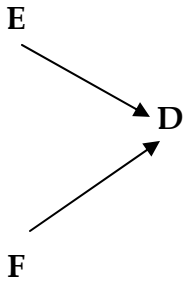
** Situation in which F is a **confounder** for D-E association **

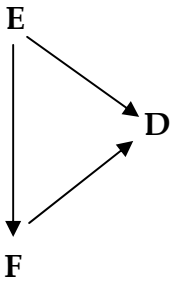


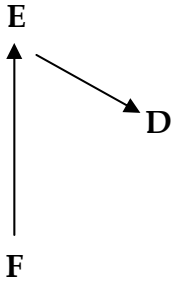




** Situation in which F is a **not confounder** for D-E association **







D = Disease (↔) non causal association

E = Exposure

F = Confounder (→) causal association

Control of confounding

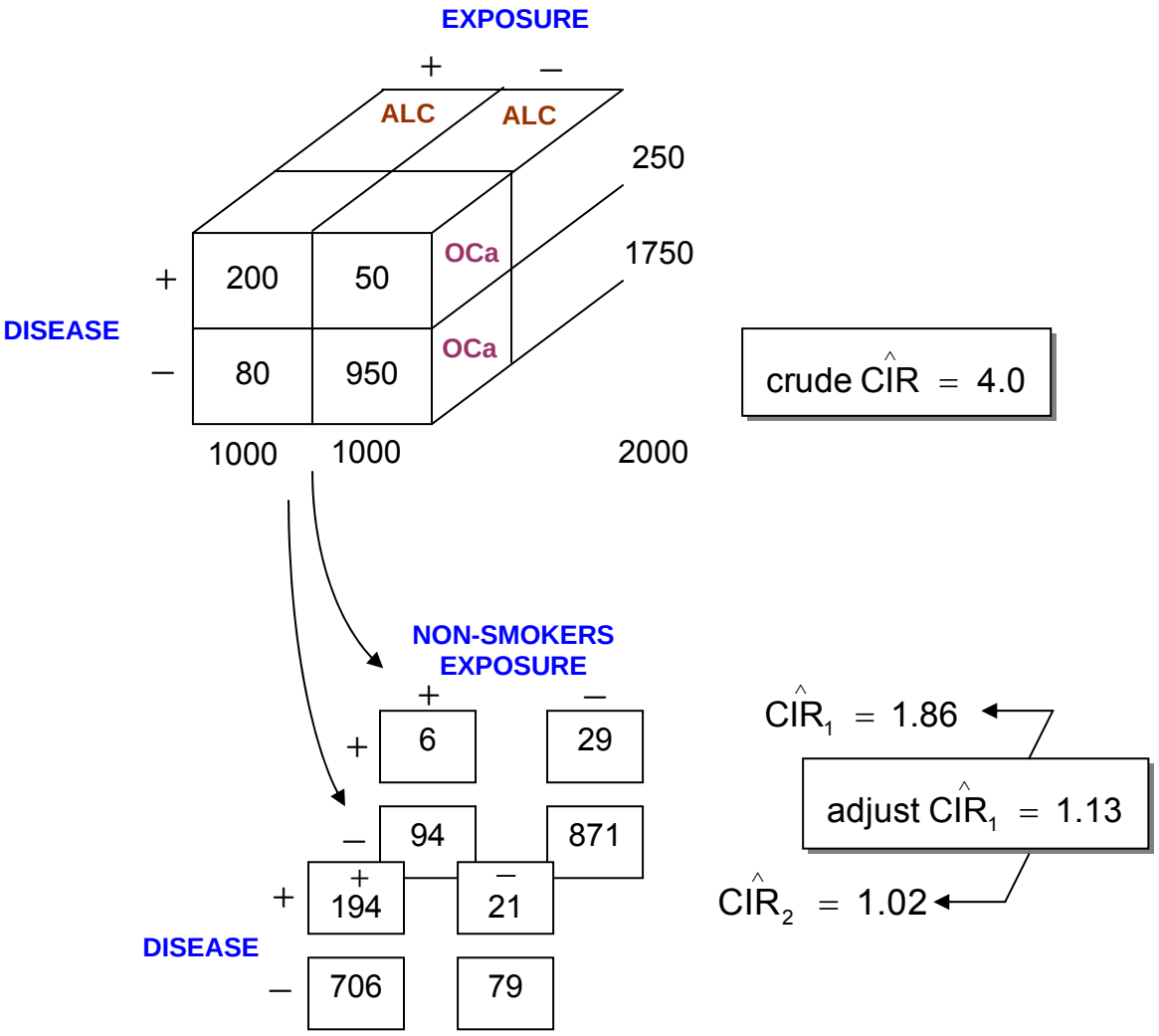
Confounding can be controlled in either research design or data analysis phase. There are three methods that can be used to control confounding during the design phase of the study: **randomization, restriction and matching**. Randomization is applicable only to intervention studies while restriction and matching can be considered for all analytic study designs. In the analysis phase we can use **stratified analysis** and **multivariate analysis** to control for the confounding variables.

- Restriction
- Matching

} **DESIGN**

- Stratified analysis
- Multivariate analysis

} **ANALYSIS**



The data-based criterion for establishing the presence or absence of **confounding involves the comparison of a CRUDE EFFECT MEASURE with an ADJUSTED EFFECT MEASURE**.

CONFOUNDING is acknowledged to be present when the crude and adjusted differ in value

Mantel-Haenszel Adjusted Relative Risk (RR).

$$mRR = \frac{\sum a_g n_{0g} / n_g}{\sum b_g n_{1g} / n_g}$$

$$= \frac{(194 \times 100)/1000 + (6 \times 900)/1000}{(21 \times 900)/1000 + (29 \times 100)/1000}$$

$$= 1.1376$$

Mantel-Haenszel Adjusted Odds Ratio (RR)

$$aOR = \frac{\sum a_g d_g / n_g}{\sum b_g c_g / n_g}$$

$$= \frac{(194 \times 79)/1000 + (6 \times 871)/1000}{(21 \times 706)/1000 + (29 \times 94)/1000}$$

$$= 1.1709$$

References

1. Last J.M. A dictionary of epidemiology. New York : Oxford University Press 1983.

2. Hennenkens C.H. and Buring, J.E. Epidemiology in Medicine. Boston Toronto : Little, Brown and Company 1987.

3. Kelsey J.L.,Thompson W.D. and Evans S.E. Methods in Observational Epidemiology. New York Oxford : Oxford University Press 1986.

4. Rothman K.J. Modern Epidemiology. Bonston Toronto : Little, Brown and Company 1986.

5. Schlesselman J.J. Case-Control Studies; Design, Conduct, Analysis. New York Oxford: Oxford University Press 1982.

6. Kleinbaum D.G., Kupper L.L., Morgenstern H. Epidemiologic. Research; Prinicples and Methods. London, Singapore, Sydney, Toronto, Mexico City: Lifetime Learning Publications 1982.

7. Breslow N.E., Day N.E. Statistical Methods in Cancer Research (Vol.1); Analysis of Case-Control Studies. Lyon: IARC Scientific Publications 1980.