Foundations in Biostatistics and Epidemiology
Session 8: Confounding, Effect Modification, and Sources of Bias in Epidemiologic Studies
The following provides a summary of the content of this module. Some definitions were taken from Last's *A Dictionary of Epidemiology.*

I. Sources of Bias in Epidemiologic Studies
- a. Study Validity:
  - i. Internal validity: study provides unbiased estimate of what it claims to estimate
  - ii. External validity: results from study can be generalized to some other population
- b. Goals of Measurement:
  - i. Accuracy (validity): an expression of the degree to which a measurement measures what it purports to measure
  - ii. Precision (reliability): the quality of being sharply defined or measured
- c. Bias: Any systematic error (not a random error arising by chance or through sampling variability) in design, conduct or analysis of a study that results in a mistaken estimate of an exposure's association with disease. i.e., produces a biased estimate of OR, RR, PRR etc. Two main categories: Selection Bias and Information Bias
  - i. Selection biases – relate to how study groups are chosen; errors due to systematic differences in relevant characteristics between those who are included in study and those who are not
    1. Self-selection or Membership Bias:
       a. Refers to characteristics of individual that may consciously or unconsciously affect membership in a certain group; Individuals may select themselves into certain occupations or choose behaviors because of certain personal characteristics
       b. Example: Healthy worker effect - Morbidity and mortality is lower in workers than in general pop. (which includes non-workers)
       c. May increase or decrease risk estimate
    2. Non-response Bias:
       a. People who agree to participate in study may be different in terms of exposure or other important characteristics from those who do not agree
       b. May increase or decrease risk estimate depending on factors related to non-response
    3. Loss to Follow-up/Withdrawal Bias (cohort or longitudinal follow-up studies):
       a. If incidence of outcome is different among exposed who are followed, compared to those who are lost, or among unexposed who are followed compared to those who are lost, then a biased risk estimate will result
    4. Berkson's Bias
       a. Refers to selective factors that lead hospital cases or controls to be systematically different from all cases or controls in population they represent
  - ii. Information biases – inaccuracy in measurement or classification of exposure, outcome, or covariates; results in measurement error (error in measurement of continuous variable) or misclassification (error in measurement of categorical variables)
    1. Recall Bias
       a. Differences among cases and controls in accuracy or completeness of recall of exposure
       b. Results in OR away from null when cases more accurately remember exposure than controls or when cases over-report exposure history
       c. Avoid by validating exposure history from independent source or objective measures; Use of "sick" (disease unrelated to the outcome or exposure of interest) controls may help equalize this bias
    2. Interviewer Bias
       a. Systematic error due to interviewer's subconscious or conscious gathering of selective information in different groups

b. Results in over-estimate of association when interviewers probe for more information among cases
c. Avoid by blinding interviewers to case/control status; Train interviewers and use standardized forms; Assess by re-interviewing a sample of study subjects

3. Family Information Bias
   a. Occurs when cases are more aware of their family history because occurrence of their disease has stimulated discussion or investigation of past disease history in family
   b. Results in increased (odds) estimate, since cases may be more aware of a positive family history than are controls
   c. Avoid by validating disease status of family members

d. Direction of Bias:
   i. Bias away from the null: observed measure of association is farther from the null value (e.g., a difference of 0 between two means or two proportions, or a ratio of 1 for an odds ratio or relative risk) than the true value (i.e., observed larger effect, either positive or negative, than true effect)
   ii. Bias towards the null: observed measure of association is closer to the null value (e.g., a difference of 0 between two means or two proportions, or a ratio of 1 for an odds ratio or relative risk) than the true value (i.e., observed smaller effect than true effect)

e. Differential vs. Non-differential Misclassification
   i. Non-differential Misclassification:
      1. Classification error NOT depend on disease status or exposure status
      2. If error of exposure classification same for cases and controls (or disease classification is same in exposed and un-exposed), bias usually towards null; Exceptions: if there are mis-measured confounding factors or if > 2 categories of variable
   ii. Differential Misclassification:
      1. Classification error DOES depend on disease status or exposure status
      2. If error of exposure classification NOT same for cases and controls (or disease classification not same in exposed and unexposed), bias can go either toward or away from null depending on type of misclassification

II. Confounding

a. A distortion of the exposure-disease association due to the influence of a third factor
b. A confounder may fully or partially account for the observed effect of the study exposure or mask or hide an underlying true association
c. Results when another factor is unevenly distributed between comparison groups and may account for the observed association
d. Example: In a cross-sectional study comparing lung function (outcome) between smokers and non-smokers (exposure) among pediatric patients aged 3 to 19 years, the apparent increased lung function in smokers versus non-smokers is confounded by age (smokers are older and have higher lung function measures)
e. Criteria for Confounding Variable
   i. It must be a risk factor for the disease, or associated with the disease, but not necessarily causal.
   ii. It must be associated with the exposure under study, but is not a result of it.
   iii. It must not be an intermediate variable on the causal pathway
f. Steps for Evaluating for Confounding:
   i. Calculate crude overall estimation of the exposure-disease association
   ii. Stratify the data by levels of the suspected confounder
   iii. Calculate the stratum-specific estimates
   iv. Compare the stratum-specific estimates to one another and to the crude measure of association

g. Direction of Confounding:
    i. Positive Confounding (Away from the Null): The confounding factor produces an observed estimate of the association between exposure and disease that is an overestimate – either more positive or more negative – than the true association.
    ii. Negative Confounding (Toward the Null): The confounding factor produces an observed estimate of the association between exposure and disease that is an underestimate of the true association.

h. Avoid or Control for Confounding (eliminate its effect)
    i. In designing and conducting study:
        1. Randomization: Randomly assigning exposure to ensure equal distribution of confounders in each exposure category
        2. Restriction: Limiting inclusion criteria to prohibit variation (e.g., limited age eligibility criteria if age is a confounding variable)
        3. Matching: Selecting subjects according to the value of suspected confounders to ensure equal distributions among study groups (e.g., match cases and controls by age and gender in a case-control study if age and sex are confounding variables)
    ii. In analyzing the data:
        1. Stratification: Evaluating the association between exposure and disease within homogeneous categories (strata) of the confounding variable (e.g., analyze data from males and females separately if sex is a confounding variable)
        2. Multivariate analyses: Constructing mathematical models to describe the association between exposure and outcome variables

III. Effect Modification

a. The effect of the exposure on the outcome differs depending on the level of another variable, the effect modifier.
b. Effect Modification = Interaction
c. Example: An association that is stronger in older people than in younger people; age is an effect modifier
d. Types of Interaction:
    i. Synergistic effect (positive interaction) = The effect modifier potentiates or accentuates the effect of the exposure of interest.
    ii. Antagonistic effect (negative interaction) = The presence of the effect modifier diminishes or eliminates the effect of the exposure of interest.
e. Steps for using stratification to identify effect modification
    i. Calculate the crude overall estimation of the exposure-disease association
    ii. Stratify the data by levels of the third factor
    iii. Calculate the stratum-specific estimates
    iv. Compare the stratum-specific estimates to one another and to the crude measure of association
    v. Determine whether the magnitude of the stratum-specific estimates is different across strata
f. Reporting
    i. The aim is to describe and report effect modification, not control it
    ii. Assessed by comparing the magnitude (and direction) of stratum-specific estimates
    iii. Use stratification to evaluate and describe