Causation and error

Yousef Alimohamadi
PhD of Epidemiology

Associations may be due to

- **Chance (random error)**
  - statistics are used to reduce it by appropriate design of the study
  - statistics are used to estimate the probability that the observed results are due to chance
- **Bias (Systematic error)**
  - must be considered in the design of the study
- **Confounding**
  - can be dealt with during both the design and the analysis of the study
- **Causation**

Associations may be due to

- Chance (random error)
  - statistics are used to reduce it by appropriate design of the study
  - statistics are used to estimate the probability that the observed results are due to chance
- Bias (Systematic error)
  - must be considered in the design of the study
- Confounding
  - can be dealt with during both the design and the analysis of the study
- Causation

Dealing with chance error

- **During design of study**
  - Sample size
  - Power
- **During analysis (Statistical measures of chance)**
  - Test of statistical significance (P value)
  - Confidence intervals

Statistical measures of chance I

(Test of statistical significance)

Association in Reality

Observed association

	Yes	**No**
Yes		Type I error
No	Type II error	

P-value

- the probability the observed results occurred by chance
- the probability that an effect at least as extreme as that observed could have occurred by chance alone, given there is truly no relationship between exposure and disease (Ho)
- statistically non-significant results are not necessarily attributable to chance due to small sample size

Statistical Power

- Power = 1 – type II error
- Power = 1 - β

P value

0.00001

Clinical Importance

VS

Statistical Significance

Statistical measures of chance II
(Confidence intervals)

Question?

- 20 out of 100 participants: 20%
- 200 out of 1000 participants: 20%
- 2000 out of 10000 participants: 20%
- What is the difference?

Answer: Confidence Interval

- **Definition:** A range of values for a variable of interest constructed so that this range has a specified probability of including the true value of the variable for the population
- **Characteristics:**
  - a measure of the precision (stability) of an observed effect
  - the range within which the true magnitude of effect lies with a particular degree of certainty
  - 95% C.I. means that true estimate of effect (mean, risk, rate) lies within 2 standard errors of the population mean 95 times out of 100
  - Confidence intervals get smaller (i.e. more precise or more certain) if the underlying data have less variation/scatter
  - Confidence intervals get smaller if there are more people in your sample

95% Confidence Interval (95% CI)

- 20 out of 100 participants: 20%
  95% CI: 12 to 28
- 80 out of 400 participants: 20%
  95% CI: 16 to 24
- 2000 out of 10000 participants: 20%
  95% CI: 19.2 to 20.8

How to Estimate CI?

- Standard Error (SE)
- 95% CI = statistic ± 1.96 SE
- Example:
  - 95% CI of mean = sample mean ± SEM
  - $$SEM = \frac{SD}{\sqrt{n}}$$

How to Estimate CI? (example)

- A sample of 100 participants
- Mean of their age 25 years
- SD of age: 10
- CIM?

- CIM = $25 \pm 1.96 * 10 / \sqrt{100}$
- CIM ~ from 23 to 27

Confidence Interval

vs

P value

Associations may be due to

- **Chance (random error)**
  - statistics are used to reduce it by appropriate design of the study
  - statistics are used to estimate the probability that the observed results are due to chance
- **Bias (Systematic error)**
  - must be considered in the design of the study
- **Confounding**
  - can be dealt with during both the design and the analysis of the study
- **Causation**

Associations may be due to

- **Chance (random error)**
  - statistics are used to reduce it by appropriate design of the study
  - statistics are used to estimate the probability that the observed results are due to chance
- **Bias (Systematic error)**
  - must be considered in the design of the study
- **Confounding**
  - can be dealt with during both the design and the analysis of the study
- **Causation**

Bias

- Any systematic error that results in an incorrect estimate of the association between risk factors and outcome

BIAS: threats to validity and interpretation

- Bias is the result of systematic error in the design or conduct of a study; a *tendency* toward erroneous results
- Systematic error results from flaws in either the (1) *method of selection of study participants,* or

  (2) in the *procedures for gathering relevant exposure and/or disease information*
- Hence - the observed study results will *tend* to be different from the true results

Bias results from systematic flaws

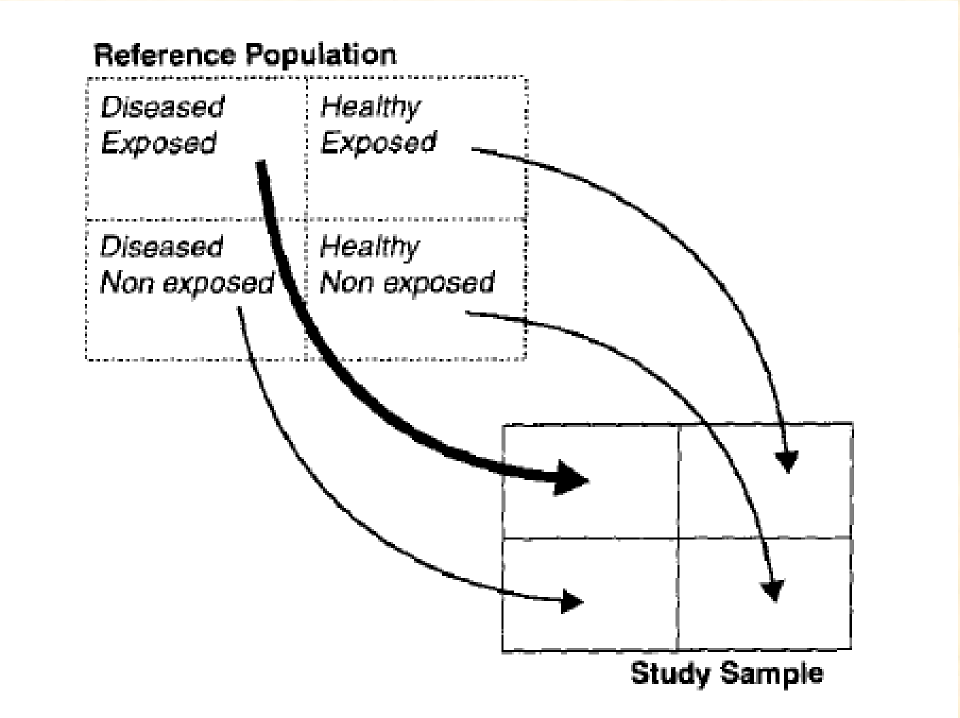
- study design,
- data collection,
- analysis
- interpretation of results

Types of Bias

- Selection bias – identification of individual subjects for inclusion in study on the basis of either exposure or disease status depends in some way on the other axis of interest
- Observation (information) bias – results from systematic differences in the way data on exposure or outcome are obtained from the various study groups

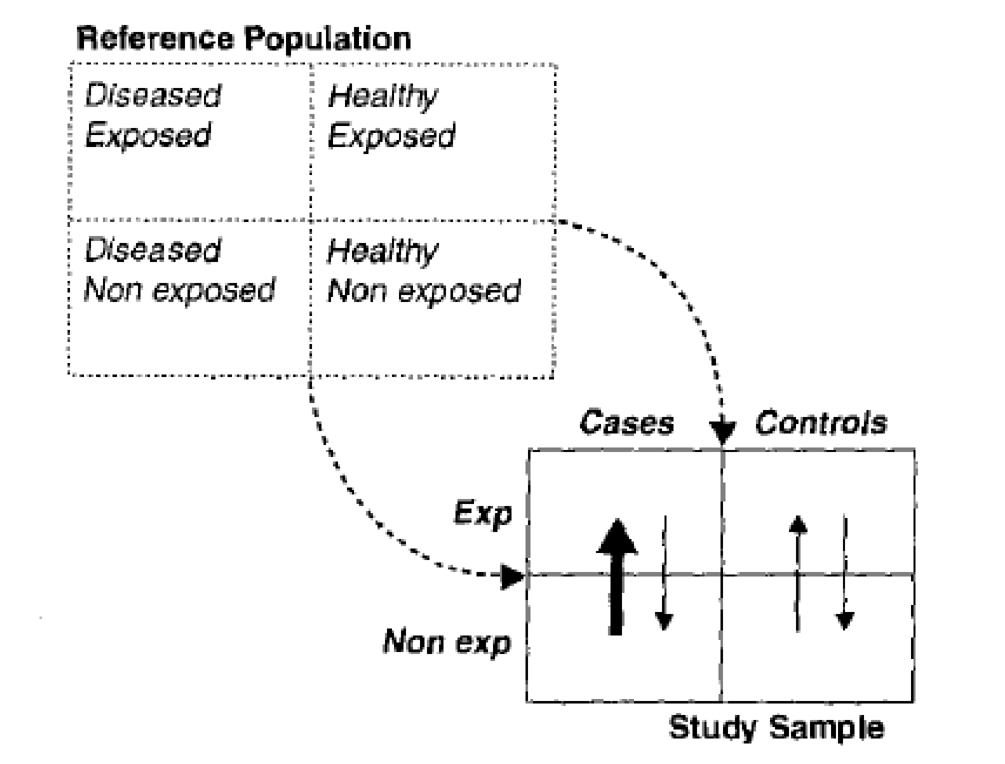
Selection bias

- is present when individuals have different probabilities of being included in the study sample according to relevant study characteristics-namely, the exposure and the outcome of interest.



Observation (information) bias

- results from a systematic tendency for individuals selected for inclusion in the study to be erroneously placed in different exposure/outcome categories, thus leading to misclassification.



Selection Bias

- Selection bias occurs when a systematic error in the ascertainment of study subjects (cases or controls in case-control studies, or exposed or unexposed subjects in cohort studies) results in a tendency toward distorting the measure expressing the association between exposure and outcome.

Selection bias in cohort

- *Healthy worker effect*: In a cohort study, because study participants (exposed or unexposed) are selected *before* the disease actually occurs, differential selection according to disease status is less likely to occur. Nevertheless, selection bias may occur at the outset of a cohort study when, for example, a group of persons exposed to an occupational hazard is compared with a sample of the general population.
- *Differential losses to follow-up*

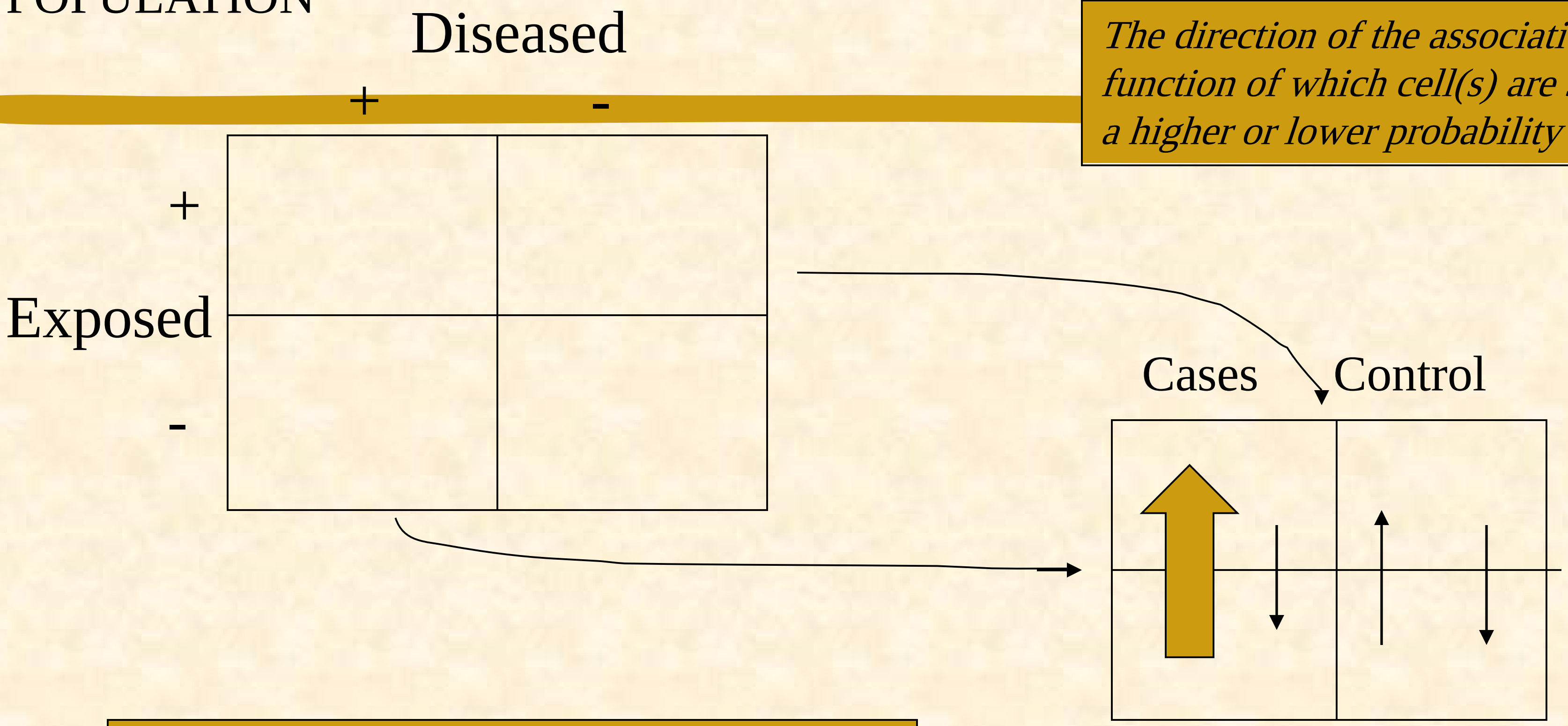
Information Bias

- Information bias in epidemiologic studies results from either imperfect definitions of study variables or flawed data collection procedures. These errors may result in misclassification of exposure and/or outcome status for a significant proportion of study participants.
- A classic example is: recall bias, in which the ability to recall past exposure is dependent on case or control status. Cases may be more likely than controls to overstate past exposure

Misclassification of EXPOSURE

REFERENCE
POPULATION

The direction of the association is a function of which cell(s) are subjected to a higher or lower probability

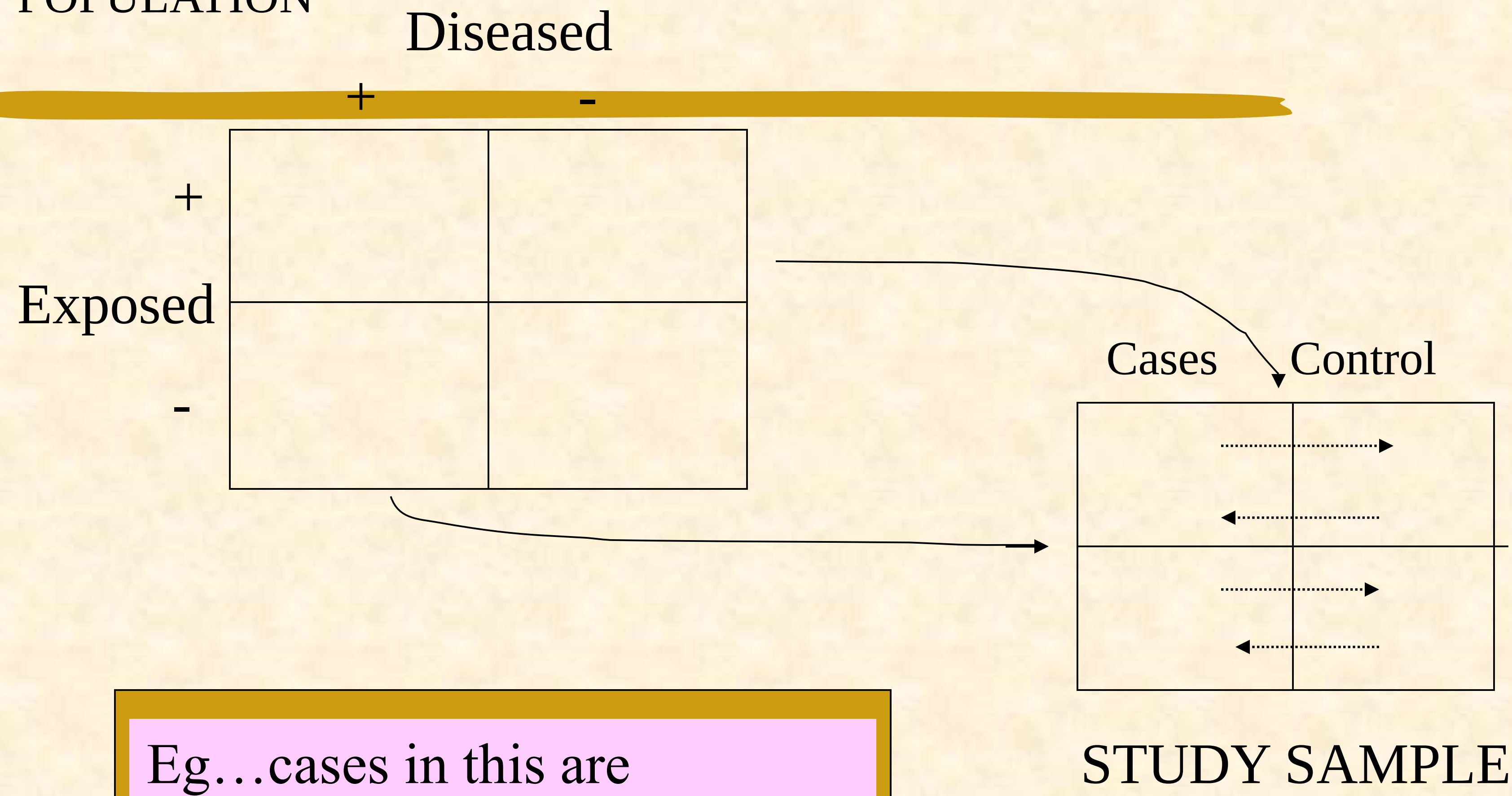


STUDY SAMPLE

Eg...unexposed cases in this example tend to mistakenly report past exposure to a greater extent than do controls

Misclassification of OUTCOME

REFERENCE POPULATION



Eg…cases in this are mistakenly classified as controls due to low sensitivity on a screening test

Types of Information Biases

- Exposure Identification Bias
  - Recall bias
  - Interviewer bias
- Outcome Identification Bias
  - Observer bias
  - Respondent bias

Exposure Identification Bias

- Problems in the collection of exposure data or an imperfect definition of the level of exposure.
- 2 main examples:
  - Recall bias
  - Interviewer bias

Recall Bias

- Most cited: inaccurate recall of past exposure (may be due to temporality, social desirability or diagnosis).

How to Prevent Recall Bias

1. Verification of exposure information from participants by review of pre-existing records
2. Selection of diseased controls and compensating this bias
3. Objective markers of exposure or susceptibility (for example- genetic markers).
4. Nested case-control studies allow evaluation of exposures prior to “case” status

Interviewer Bias

- May occur when interviewers are not blinded to disease status.
  - They may probe more
- Interviewers may be biased toward the study hypothesis (or have other biases).
  - They may ignore protocols

Outcome Identification Bias

- may be due to an imperfect definition of the outcome or to errors at the data collection stage .
- Two main examples:
  - Observer bias
  - Respondent bias

Observer Bias

- In a Cohort study: decision to classify outcome may be affected by knowledge of exposure status. Especially "soft" outcomes such as migraine, or psychiatric symptoms

Preventing Observer Bias

- Mask observers in charge of classifying outcome with respect to exposure status
- Multiple observers

Respondent Bias

- Synonym of recall bias in cohort studies.
- outcome ascertainment bias may occur during follow-up of a cohort when information on the outcome is obtained by participant response: for example, when collecting information on events for which it is difficult to obtain objective confirmation, such as episodes of migraine headaches.

Respondent Bias

- In a Cohort study: respondents may respond with little consistency to un-standardized questions or to "subjective" questions.
- Eg. Questions about depression may be very subjective. A solution is to use a standardized instrument.

The result of information bias: Misclassification

- Nondifferential misclassification
- Differential misclassification

Nondifferential misclassification

- Nondifferential misclassification occurs when the degree of misclassification of exposure is independent of case-control status (or vice versa).
- When there are two categories, nondifferential misclassification tends to bias the association toward the null hypothesis.

Differential Misclassification

- Occurs when the degree of misclassification of exposure (outcome) differs between the groups being outcome (exposure) groups

- Effect is: bias toward or away from the null

Combined selection/information biases

- biases related to
  - medical surveillance
  - cross-sectional studies
    - *Incidence-Prevalence Bias*
    - *Temporal Bias*
  - evaluation of screening
    - *Selection Bias*
    - *Incidence-Prevalence Bias*
    - *Lead Time Bias*

Associations may be due to

- **Chance (random error)**
  - statistics are used to reduce it by appropriate design of the study
  - statistics are used to estimate the probability that the observed results are due to chance
- **Bias (Systematic error)**
  - must be considered in the design of the study
- **Confounding**
  - can be dealt with during both the design and the analysis of the study
- **Causation**

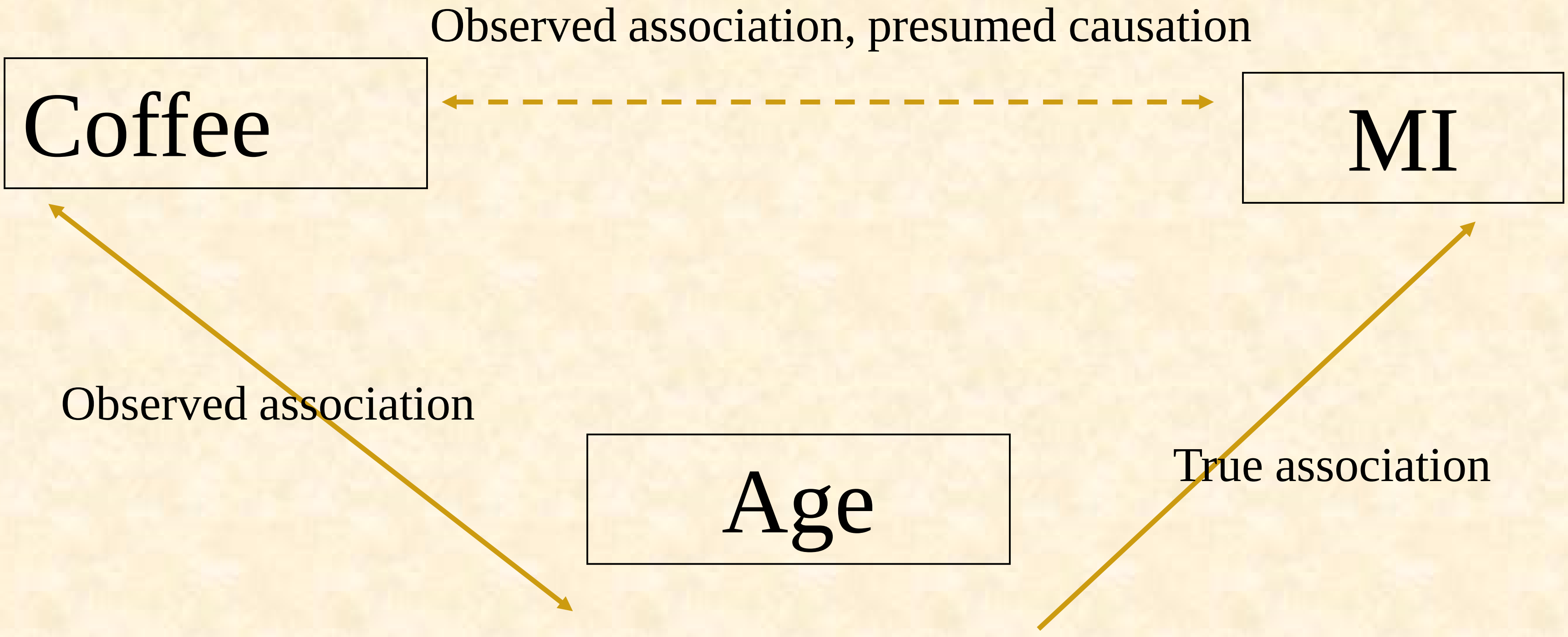
Associations may be due to

- **Chance (random error)**
  - statistics are used to reduce it by appropriate design of the study
  - statistics are used to estimate the probability that the observed results are due to chance
- **Bias (Systematic error)**
  - must be considered in the design of the study
- **Confounding**
  - can be dealt with during both the design and the analysis of the study
- **Causation**

Confounding

- Confounding results when the effect of an exposure on the disease (or outcome) is distorted because of the association of exposure with other factor(s) that influence the outcome under study.

Confounding



The confounding variable is causally associated with the outcome

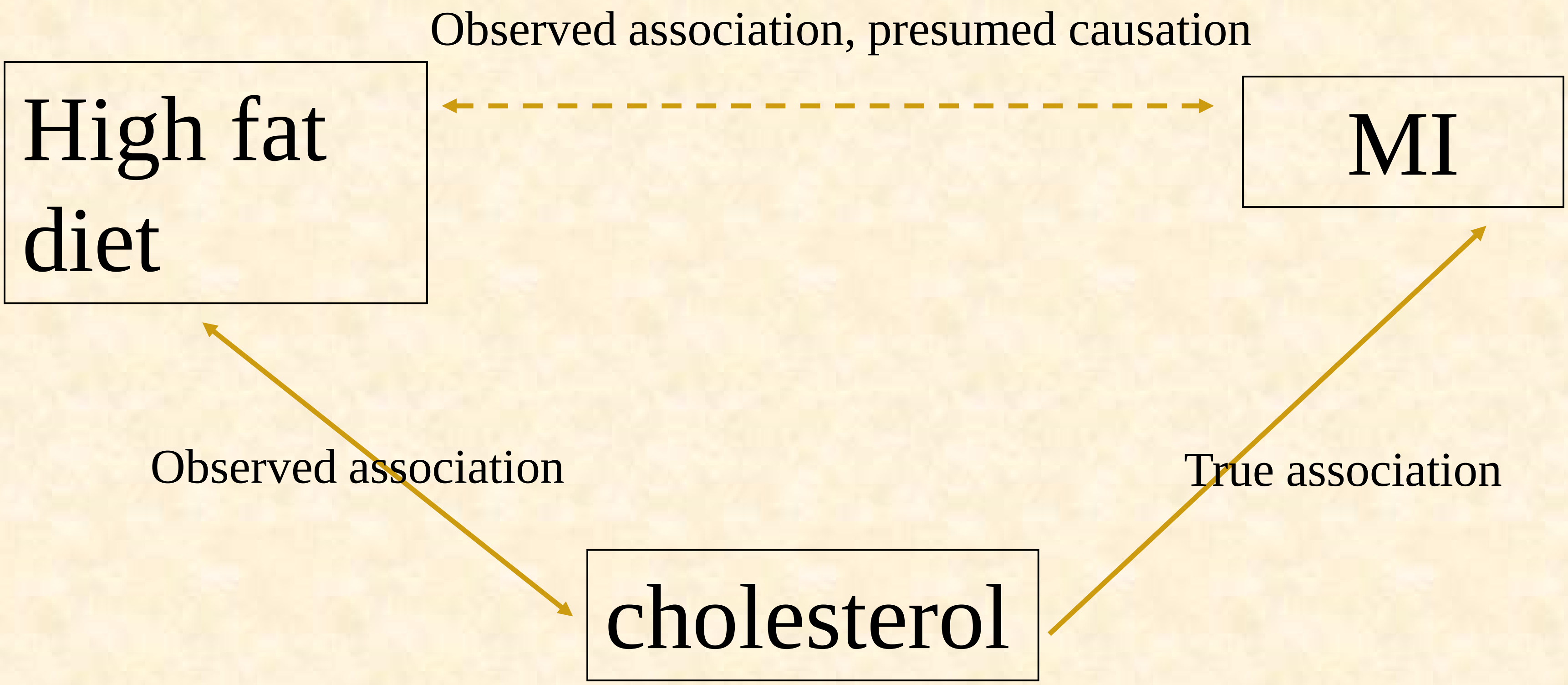
and

noncausally or causally associated with the exposure

but

is not an intermediate variable in the causal pathway between exposure and outcome

Confounding



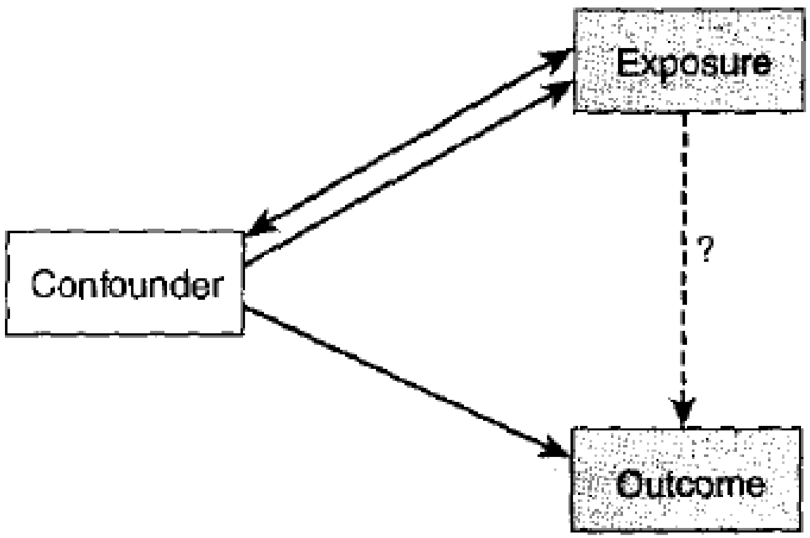
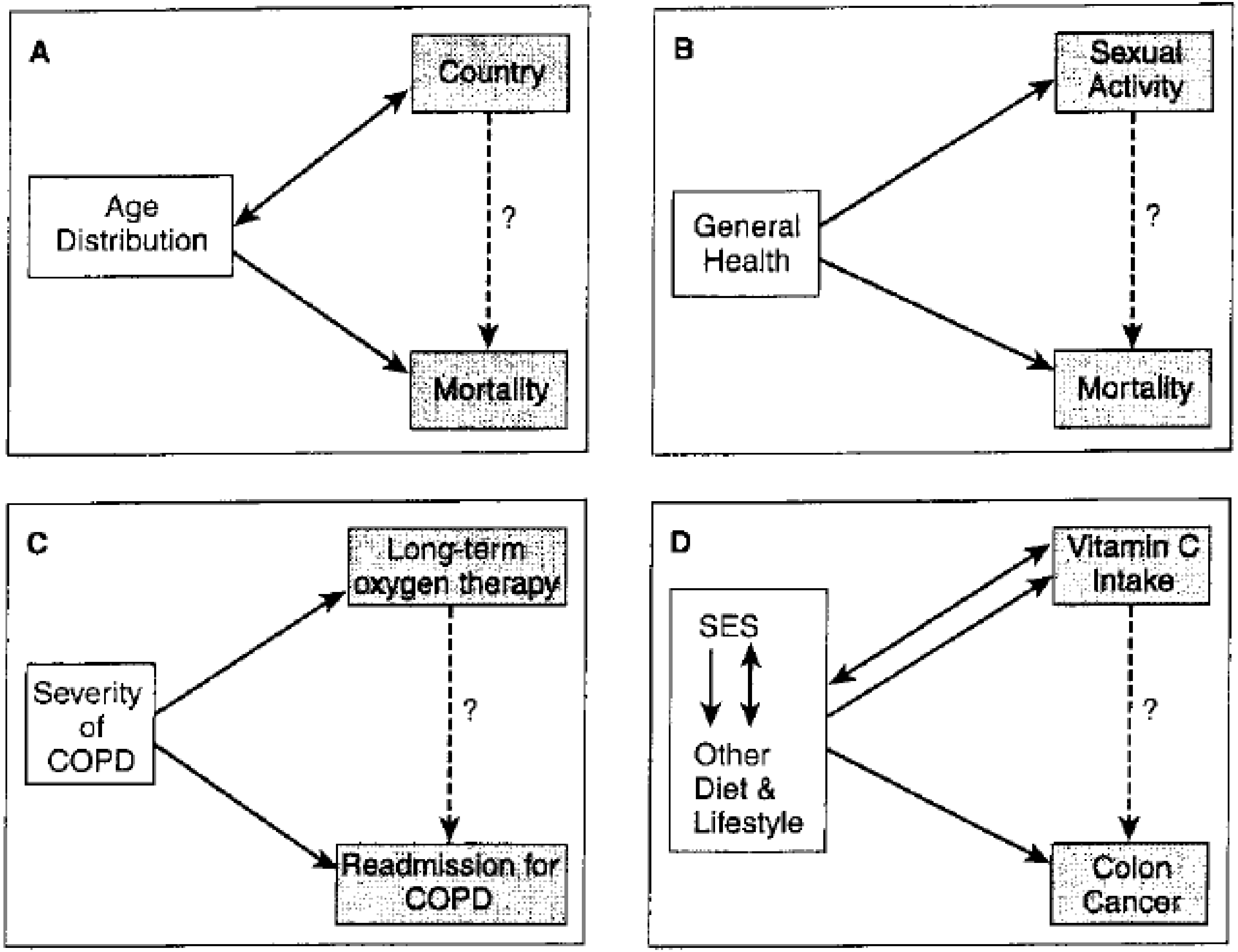
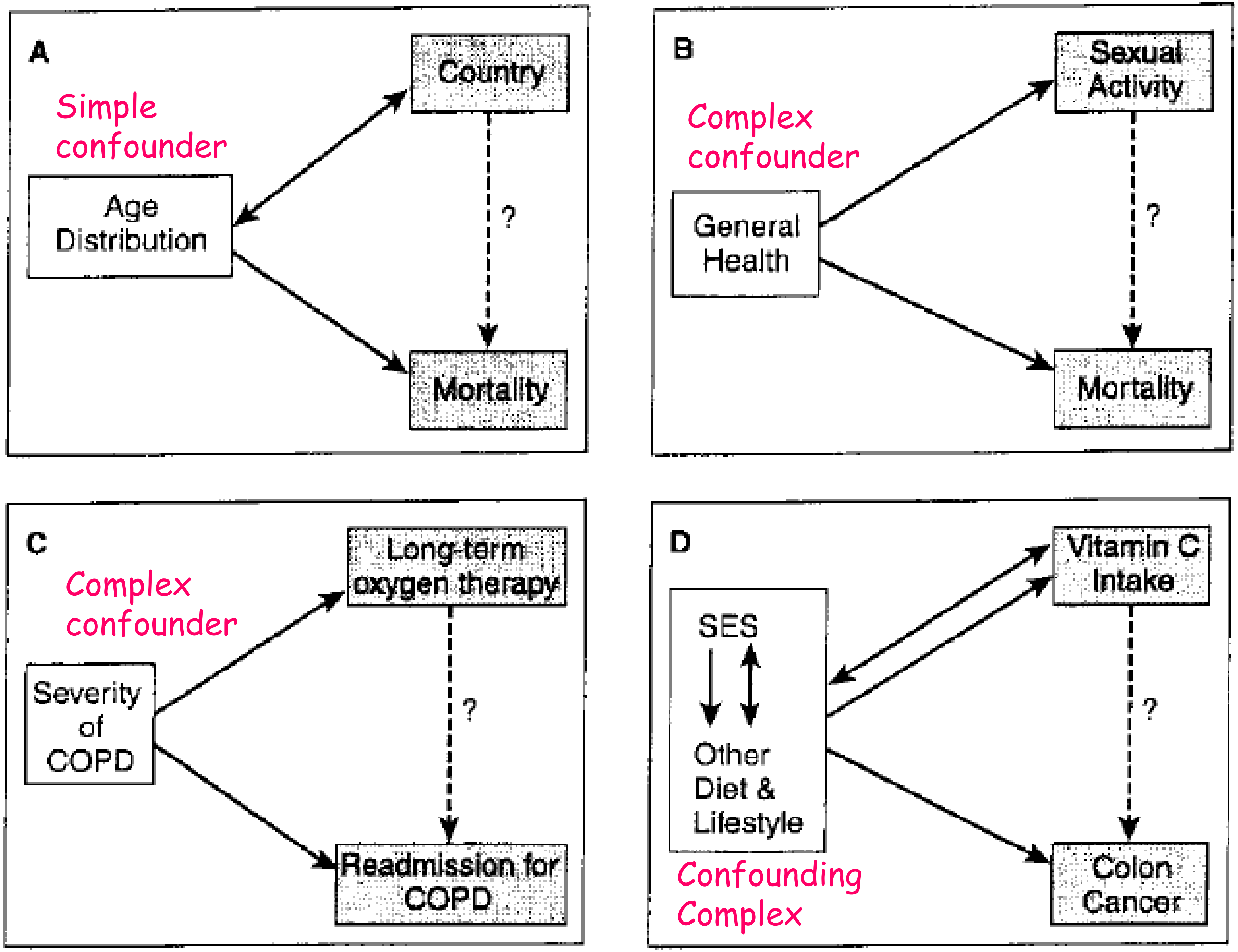


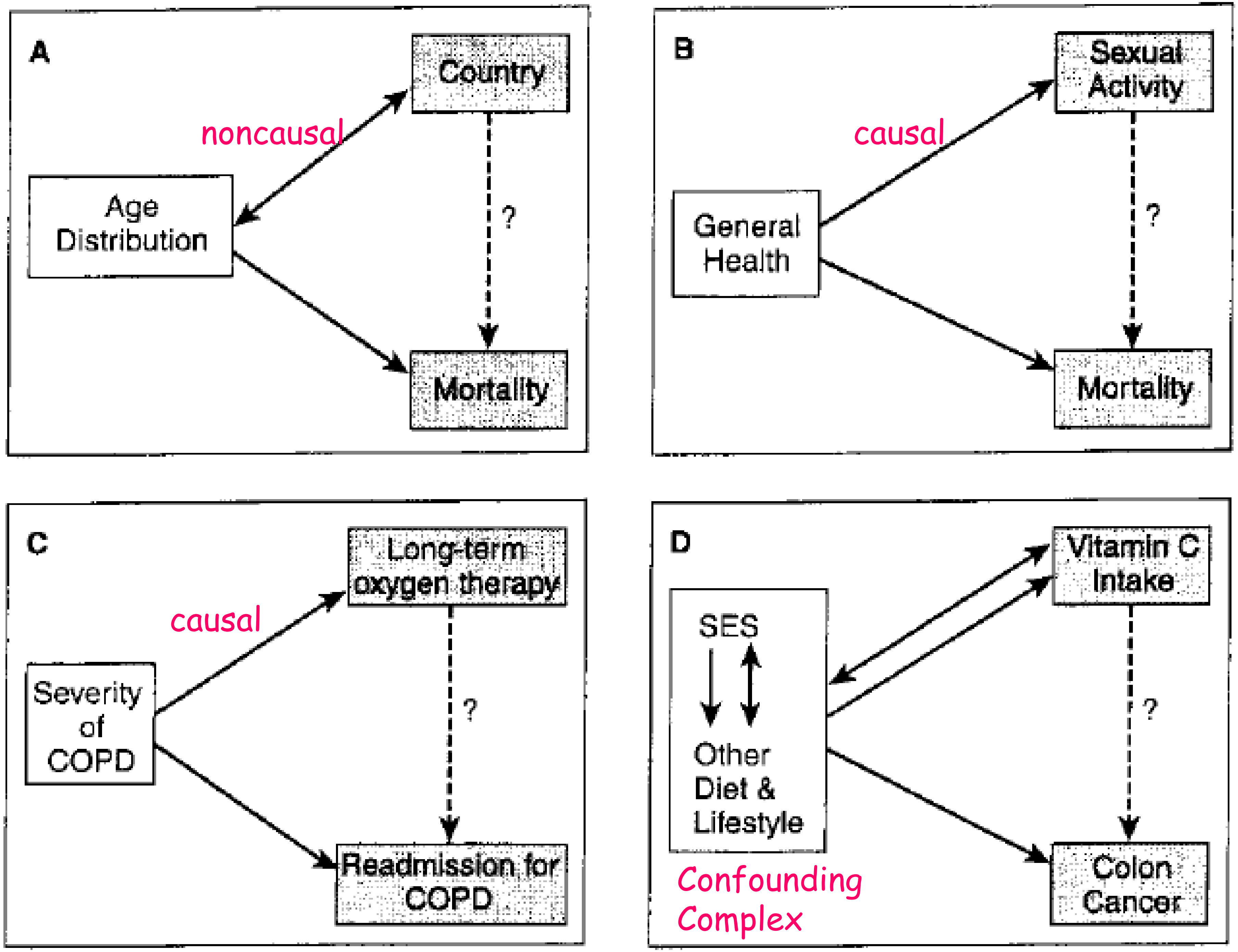
Figure 5–1 General definition of confounding. The confounder is causally associated with the outcome of interest and either causally or noncausally associated with the exposure; these associations may distort the association of interest: whether exposure causes the outcome. A unidirectional arrow indicates that the association is causal; a bidirectional arrow indicates a noncausal association.

General Rule

- The common theme with regard to confounding is that the association between an exposure and a given outcome is induced, strengthened, weakened, or eliminated by a third variable or group of variables (confounders).


A
Country
Age Distribution
?
Mortality
B
Sexual Activity
General Health
?
Mortality
C
Long-term oxygen therapy
Severity of COPD
?
Readmission for COPD
D
Vitamin C Intake
SES
Other Diet & Lifestyle
?
Colon Cancer


A
Simple confounder
Country
Age Distribution
?
Mortality
B
Complex confounder
Sexual Activity
General Health
?
Mortality
C
Complex confounder
Long-term oxygen therapy
Severity of COPD
?
Readmission for COPD
D
Vitamin C Intake
SES
Other Diet & Lifestyle
?
Confounding Complex
Colon Cancer


A
Country
noncausal
?
Age Distribution
Mortality
B
Sexual Activity
causal
?
General Health
Mortality
C
Long-term oxygen therapy
causal
?
Severity of COPD
Readmission for COPD
D
Vitamin C Intake
SES
Other Diet & Lifestyle
?
Colon Cancer
Confounding Complex

Assessing the presence of confounding

- There are several approaches to assess the presence of confounding, which are related to the following questions:
  1. Is the confounding variable related to both the exposure and the outcome in the study?
  2. Does the exposure-outcome association seen in the crude analysis have the same direction and similar magnitude as the associations seen within strata of the confounding variable?
  3. Does the exposure-outcome association seen in the crude analysis have the same direction and similar magnitude as that seen after controlling (adjusting) for the confounding variable?

Overadjustment (overmatching)

- A related issue is *overadjustment* (or *overmatching),* which occurs when adjustment is carried out for a variable so closely related to the variable of interest that no variability in the latter is allowed.

Control of Confounding

- **During design of study**
  - Restriction
  - Matching
  - Randomization
- **During analysis**
  - Stratified analysis
  - Multivariate analysis

Associations may be due to

- **Chance (random error)**
  - statistics are used to reduce it by appropriate design of the study
  - statistics are used to estimate the probability that the observed results are due to chance
- **Bias (Systematic error)**
  - must be considered in the design of the study
- **Confounding**
  - can be dealt with during both the design and the analysis of the study
- **Causation**

Associations may be due to

- **Chance (random error)**
  - statistics are used to reduce it by appropriate design of the study
  - statistics are used to estimate the probability that the observed results are due to chance
- **Bias (Systematic error)**
  - must be considered in the design of the study
- **Confounding**
  - can be dealt with during both the design and the analysis of the study
- **Causation**

DETERMINATION OF CAUSATION

- The general QUESTION:Is there a cause and effect relationship between the presence of factor X and the development of disease Y?
- One way of determining causation is personal experience by directly observing a sequence of events.

Scientific Evidences

The answer is made by inference and relies on a summary of all valid evidence.

Nature of Evidence:

1. Replication of Findings –
   - consistent in populations
2. Strength of Association –
   - significant high risk
3. Temporal Sequence –
   - exposure precede disease

Nature of Evidence:

4. Dose-Response –
   - higher dose exposure, higher risk
5. Biologic Credibility –
   - exposure linked to pathogenesis
6. Consideration of alternative explanations –
   - the extent to which other explanations have been considered.

Nature of Evidence

7. Cessation of exposure (Dynamics) –
   - removal of exposure – reduces risk
8. Specificity
   - specific exposure is associated with only one disease
9. Experimental evidence

Bradford Hill Criteria (1965)

criteria for assessing causality:

- Consistency
- Strength
- Specificity
- Temporality
- Plausibility
- Coherence
- Dose Response
- Analogy
- Experimental evidence

Bradford Hill Criteria

- Hill stated
  - None of my criteria can bring indisputable evidence for or against the cause-and-effect hypothesis
  - None can be required as sufficient alone

H. pylori

- Consistency
  - association has been replicated in other studies
- Strength
  - H. pylori is found in at least 90% of patients with duodenal ulcer
- Temporal relationship
  - 11% of chronic gastritis patients go on the develop duodenal ulcers over a 10-year period.
- Dose response
  - density of H.pylori is higher in patients with duodenal ulcer than in patients without

H. pylori

- Biologic plausibility
  - originally – no biologic plausibility
  - then H. pylori binding sites were found
  - know H. pylori induces inflammation
- Cessation
  - Eradication of H Pylori heals duodenal ulcers

SMOKING AND LUNG CANCER

1. Strength of Association:
   - The relative risks for the association of smoking and lung cancer are very high
2. Biologic Credibility:
   - The burning of tobacco produces carcinogenic compounds which are inhaled and come into contact with pulmonary tissue.

SMOKING AND LUNG CANCER

3. Replication of findings:

- The association of cigarette smoke and lung cancer is found in both sexes in all races, in all socioeconomic classes, etc.

4. Temporal Sequence:

- Cohort studies clearly demonstrate that smoking precedes lung cancer and that lung cancer does not cause an individual to become a cigarette smoker.

SMOKING AND LUNG CANCER

5. Dose-Response:
   - The more cigarette smoke an individual inhales, over a life-time, the greater the risk of developing lung cancer.
6. Dynamics (cessation of exposure):
   - Reduction in cigarette smoking reduces the risk of developing lung cancer.

- Smoking is cited as a cause of lung cancer, however. . .
  - . . . smoking is not necessary (is not a prerequisite) to get lung cancer. Some people get lung cancer who have never smoked.
  - . . . smoking alone does not cause lung cancer. Some smokers never get lung cancer.
- Smoking is a member of a set of factors (i.e., web of causation) which cause lung cancer.
  - The identity of all the other factors in the set are unknown. (One factor in the web of causation is probably genetic susceptibility.)

necessary / sufficient

- necessary and sufficient
  - the factor always causes disease and disease is never present without the factor
    - most infectious diseases
- necessary but not sufficient
  - multiple factors are required
    - cancer
- sufficient but not necessary
  - many factors may cause same disease
    - leukemia
- neither sufficient nor necessary
  - multiple cause

necessary / sufficient

- Few causes are necessary and sufficient
  - High cholesterol is neither necessary nor sufficient for CVD because many individuals who develop CVD do not have high serum cholesterol levels