

Introductory Biostatistics:

***Epidemiological Studies***

# What is Epidemiology?

- Greek origin:  
***epi*** (upon) + ***demos*** (the people)
- Definition: “the study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to control of health problems” (*Last JM: A Dictionary of Epidemiology 4<sup>th</sup> Ed 2000*)
- Research Aim: to find causes of disease
- Public Health Practice: disease prevention

# Why is Epidemiology Needed?

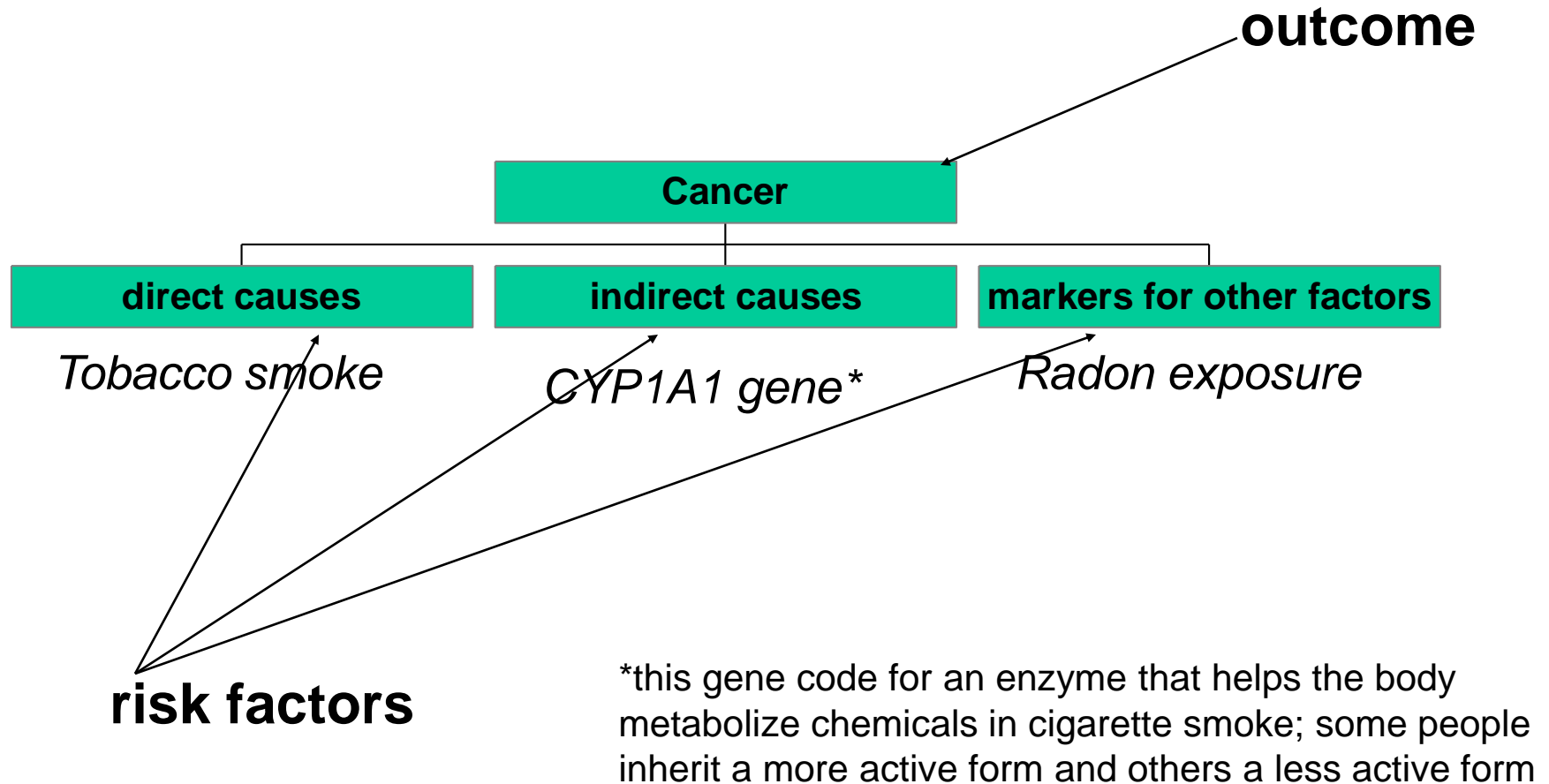
- Epidemiology seeks to discover causes of disease in different groups of people.
- This knowledge is needed for effective prevention strategies.
- Human behavior is complex, affecting how diseases are caused, distributed, detected, and prevented.
- Due to this complexity, the true causes of diseases can be masked by other factors.
- Epidemiology uses many types of study designs to deal with these issues.

# How/Can Do We Do Experiments to Find Causes of Diseases?

- Animal models are useful but extrapolation problems limit their predictive value for humans.
- Many types of human experimentation are clearly unethical.
- In other instances (e.g. vaccine trials) the experimental method may be ethical and valid.
- Where experimental trials cannot be done, then observational studies are the alternative.
- This lecture focuses on observational studies (later we'll discuss RCTs)

# Vocabulary: risk factors and outcome

Example: lung cancer

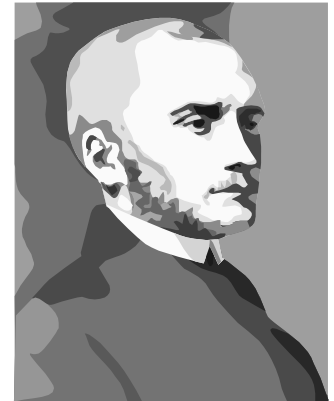


# Causality

- Since the pathway from risk factor to disease can be complex, epidemiology uses standard criteria for accepting the conclusion that a disease is caused by a specific factor.
- In some instances this seems obvious (e.g. HBV causes liver cancer) but in many situations (e.g. multifactorial disease such as diabetes) the evidence for causality may be murky.
- A set of rules for determining the causality of a risk factor has become accepted (the Bradford Hill criteria).

# Hill's Criteria for Evaluating Evidence for Causal Associations

- Strength of the association
- Specificity of the association
- Temporal sequence of events
- Consistency across studies
- Dose-response relationship
- Biological plausibility
- Experimental evidence



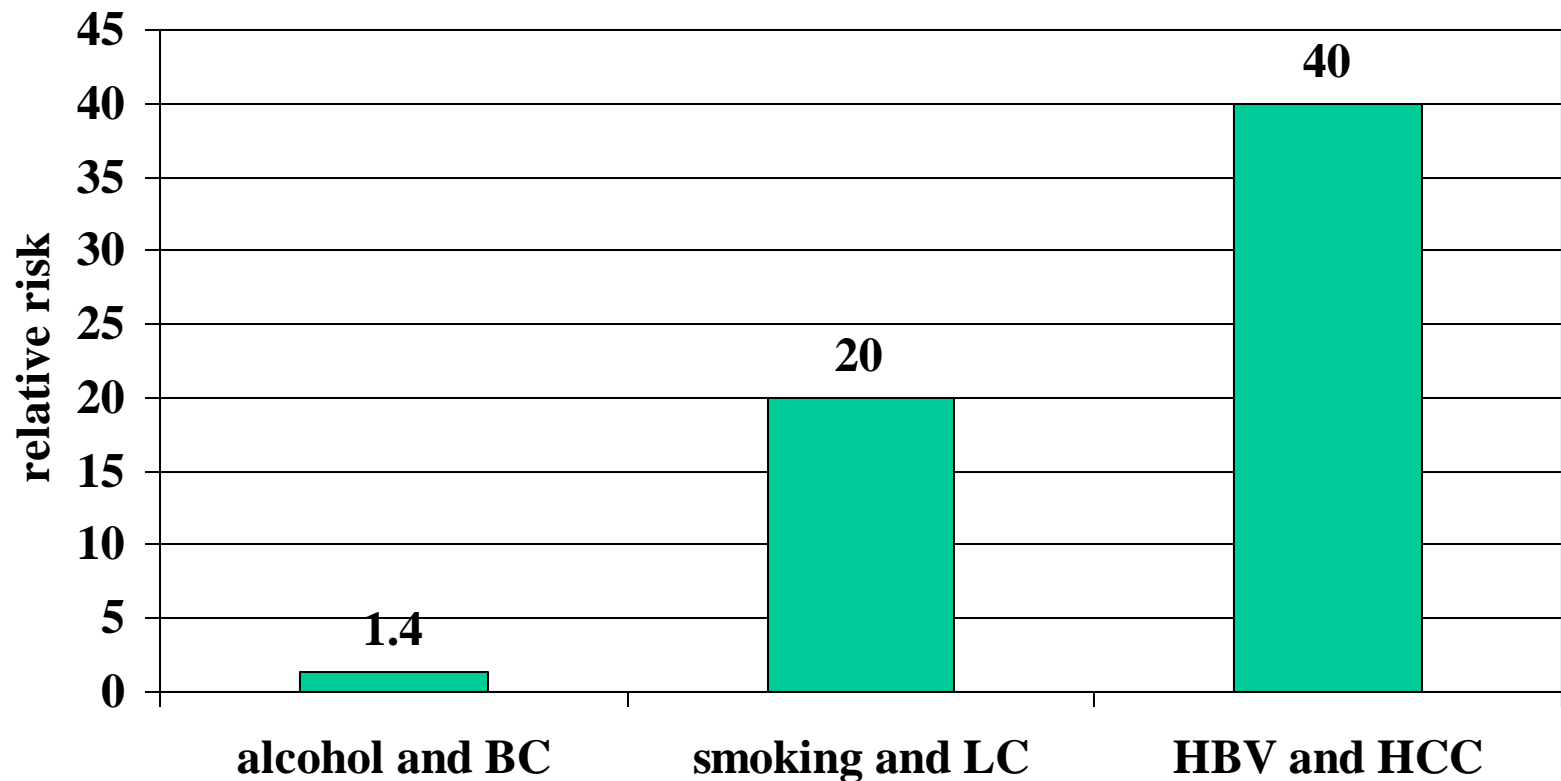
Sir Austin  
Bradford Hill

# Strength of Association

- If the risk estimate (odds ratio, relative risk) for the disease due to the putative agent is high, there is an increased expectation that the association is real, and not due to confounding by other factors.
- Weak associations (odds ratio or relative risk less than 1.5 for example) require further study as they may be due to confounding.



# Examples of Strong and Weak Associations



BC=bladder cancer, LC=lung cancer, HCC=hepatocellular carcinoma

# Specificity of the Association

- The belief that a factor truly causes a disease is increased if exposure seems to result in a narrow spectrum of disease, or indeed only one disease.
- A good example is asbestos, which causes cancer of the membrane around the lungs but not other cancers.
- Tobacco smoke seems *NOT* to fit the criterion very well, as it is associated with dozens of diseases. (But is it just one agent or a mixture?)



# Temporal Sequence of Events

- The exposure to the agent should precede the development of the disease.
- This is not always obvious, due to the types of study designs that are sometimes used in epidemiology.
- For example, if a group of newly HIV-infected people is enrolled into a study and then followed over time, and a certain Dz of interest is diagnosed ten years later in some of them, it is clear that the agent preceded the disease.



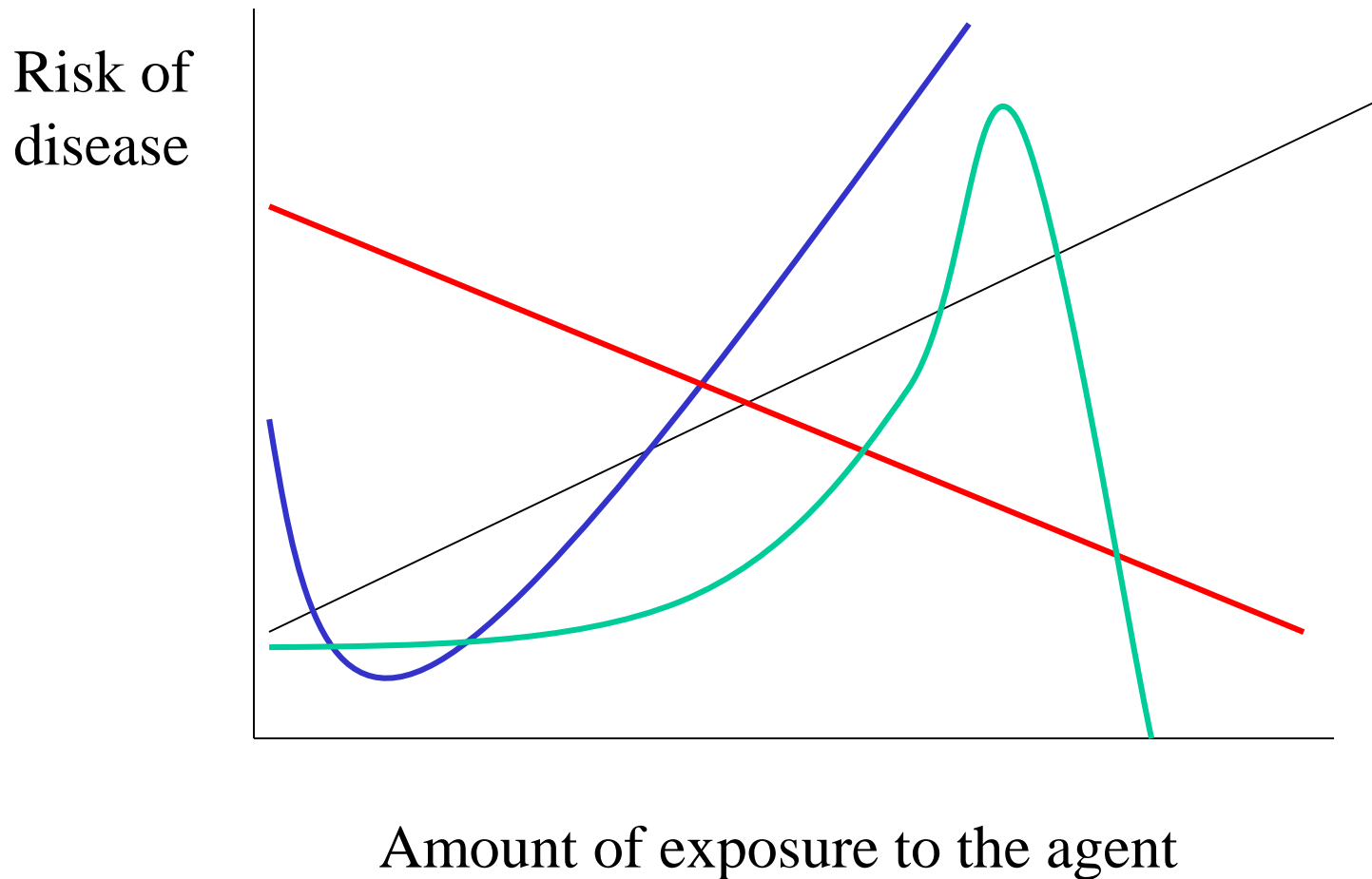
# Temporality (continued)

- On the other hand, if a group of cancer patients is recruited today and compared to a group of non-cancer patients, and blood samples are taken to measure vitamin C levels for instance, it would be difficult to establish the temporal relationship between the agent and the disease.
- Suppose that the cancer patients had not been eating well due to symptoms and complications of the cancer. If they had **lower levels of vitamin C** compared to the control subjects, did the apparent low vitamin C “cause” the cancer (as the researcher might have claimed), or did the cancer cause the levels to be lower?

# Dose-Response Relationship

- The risk of developing the disease should increase as the amount of exposure to the putative risk factor increases.
- There are several aspects of “dose” that can be considered: duration, peak level, usual daily level, intensity of exposure, age at onset of exposure.
- Be cautious when interpreting unusual patterns (inverse associations, inconsistent, U- or J- shaped curves).

# Some Dose-Response Curves



# Biological Plausibility

- This criterion asks whether the association fits within an accepted biological theory.
- However, biological plausibility is sometimes later proven when more biochemical or medical evidence has accumulated, and it may not be known at the time the association is first observed.
- For example, the association between **oral contraceptives and cardiovascular disease in women** did not fit any known biological mechanism when it was first discovered. Today there is much more known about how hormones affect the circulatory system.

# Experimental Evidence


- The belief that an agent causes a disease is increased considerably if true experiments (animal models or randomized controlled trials in humans) demonstrate the association.
- For example, clinical trials showed that using a drug to reduce serum lipid levels led to a decrease in heart attacks, strengthening the hypothesis that high lipid levels cause heart disease (and conversely, that controlling the exposure can lower the risk).





# Types of Epidemiological Studies

- Clinical case series
- Cross-sectional study
- Case-control study
- Cohort study
- Randomized controlled trial



Increasing  
value of the  
evidence for  
disease causes

*Descriptive type of studies: Case Reports (or case series); Cross Sectional*  
*Analytical: Cross Sectional, Case Control, Cohort, RCT (experimental)*

# Statistical Inference in Epidemiology

- In general, even in the cross sectional study, we're interested in some characteristic of a certain population (e.g. people who exercise), but it is impossible to study every such person.
- We draw a random sample and study the persons in the sample in order to draw some conclusions about the population.
- This process is called statistical inference.
- There will be errors when making inferences based on samples, so statistical procedures are designed to minimize the chances of such errors.

# Clinical Case Series

- This the least important type of study for establishing causality
- A clinician describes a group of patients that seem to share some pattern of characteristics, suggestive of a cause of the disease.
- Hypothesis generation is the goal.
- Example: a clinician in Germany in 1961 wrote a paper about a sudden increase in his clinic of a very rare birth defect that seemed to be related to maternal use of a new drug (thalidomide) during pregnancy. In 8 of 9 cases in his practice, the mom had used this drug. This association was later confirmed in case-control studies.

# Descriptive Cross-Sectional Studies

- A snapshot at a point in time to describe the occurrence of disease or risk factors in a specified population
- The measure of disease occurrence is prevalence proportion
- Also called prevalence surveys

# Analytical Cross-sectional Studies

- The goal is to infer an association between the risk factor and disease by comparing two or more groups of people at one point in time.
- Compares prevalence of disease or of risk factors among different groups (e.g. exposed and unexposed or diseased and non diseased)
- Measures of association: Prevalence Ratio, Odds Ratio

# How to Minimize Errors of Inference in a Cross Sectional Survey

---

- Draw a random sample of the population, rather than a highly selected sample.
- Make sure that the sample is large enough to minimize the effects of sampling errors.
- Avoid all forms of bias when sampling.
- Use the correct statistical test and be sure that its assumptions are met.

# Problems with Inference in the Cross Sectional Study Design

- Antecedent-consequent bias:  
Since cross-sectional surveys collect data only as they exist now, there can be no unbiased estimate of the timing of the association (did the risk factor precede the disease, or did the disease cause the appearance of the risk factor?). We really cannot tell, although we may have a good guess.

*Exposure and Disease are ascertained at the same point in time*

# Example of a Cross-Sectional Study

## HIV and male sex workers (Shinde et al., 2009)

**Objective of the study:** assess the prevalence of HIV and risk behaviour in male sex workers. **Evaluated** the association between HIV and sociodemographic factors.

**Data collection:** through interviewer-administered questionnaires for sociodemographic and behaviour data, clinical evaluation for sexually transmitted infections and serological evaluation for STIs (including HIV).

### What they found:

- The prevalence of HIV in male sex workers was 33%
- Male-to-female transcended people were significantly more likely to be HIV-infected compared with male
  - Prevalence ratio: 3.5 with 95% CI 1.0,11.7
- Prevalence of HIV is higher among those who choose sex work to be their main occupation (40% vs. 7%,  $p=0.02$ )



# Prevalence

- Numerator: all cases in the population (i.e. those that occurred in the past and are still in the population and those cases that are new)
- Denominator: all persons in the population of interest, whether or not they are pre-existing cases of the disease
- A proportion: number of persons with the disease as a proportion of the total population

# Prevalence

- The proportion of the population having a certain disease
  - at a point in time (point prevalence)
  - during a period of time (period prevalence)
- The prevalence of asthma depends on how many children who originally developed asthma still have asthma at the time prevalence is measured.
- Prevalence is a static measure of the amount of disease in a population at a point in time (point prevalence) or during a specific time period during which the number of cases are counted (period prevalence).

# Analysis of Cross-Sectional Studies

- Different statistical approaches can be used to analyze the data.
- Measures of association in cross-sectional studies:
  - **Prevalence Ratio (PeR)**
  - **Odds Ratio (OR)**

# Example

- Suppose a survey is conducted of the willingness of physicians to give out medical advice about weight control to their patients. Doctors are asked to report their own height and weight. Here are the results:
- 30 obese doctors: 10 gave out advice (rate=10/30)
- 60 non-obese doctors: 40 gave out advice (rate=40/60).
- Null hypothesis:  $PeR=1$
- Prevalence ratio =  $(10/30) / (40/60) = 0.33/0.66 = 0.50$
- Interpretation: the prevalence of giving out medical advice about weight among those obese doctors is 0.5 times to their non-obese doctors.
- We need more information to decide whether we should reject the null hypothesis or not.

# Confidence Interval (CI)

- There are several different methods, but the one we will use is based on the table method.
- Diagram the study results in a two-by-two table.

Example:

	Gave advice	Didn't give advice	
Obese doctors	10	20	30
Non-obese doctors	40	20	60

# 95% CI for Prevalence Ratio

Disease      No Disease

A	B	A+B Exposed
C	D	C+D Unexposed

**Pe Ratio =**

$$\{A/(A+B)\} / \{C/(C+D)\}$$

It is the ratio of Pe of disease among exposed to Pe of dz among unexposed.

We can also calculate the Pe of exposure among those with or without the disease.

$$95\% \text{ CI} = e^{(\ln \text{PeR} \pm 1.96V)}$$

where  $V = \text{square root } [B/A/(A+B) + D/C/(C+D)]$

Exercise: calculate the PeR of Dz and then PeR of Exp, the OR of Dz and the OR of Exp, and their corresponding 95% CI for the example on the previous slide.

# Examples of Cross-sectional Studies

- **U.S. National Health Survey (NHS)** includes  
National Health Interview Survey (NHIS)  
National Health and Nutrition Examination Survey  
(NHANES)  
National Health Record Survey (NHRS)
- These are repeated every 5 to 10 years
- They are based on sampling the whole US population

# Textbook

Online at  
Dahlgren Library

eBooks

