Unit

3

# Associations and Cause

### 

### Introduction

In this unit of the Quantitative Research Methods module, steps 4 to 6 in the **Seven Steps Approach** to conducting an epidemiologic study will be covered. In an observational analytic study, it is essential to understand some of the concepts covered, which is the approach we will be using to help you learn about quantitative research.

### The Seven Steps are as follows:

1. Define the population of interest
2. Conceptualize and create measures of exposures and health indicators
3. Take a sample of the population

### Estimate measures of association between exposures and health indicators of interest

1. **Rigorously evaluate whether the association observed suggests a causal association**
2. **Assess the evidence for causes working together**
3. Assess the extent to which the result matters, is externally valid, to other populations

**Contents of Unit 3**

In this unit of the module we will cover the following topics:

Session 1: Are exposures associated with disease?

[Session 2: What is a cause?](#_bookmark16)

[Session 3: How do non-causal associations arise?](#_bookmark28)

[Session 4: How can we mitigate against non-causal associations in design and](#_bookmark40) [analysis?](#_bookmark40)

[Session 5: When do causes work together?](#_bookmark52)

**Reading**

Keyes, K.M. & Galea, S. (2014) *Epidemiology Matters: a new Introduction to Methodological Foundations.* Oxford: Oxford University Press.

Session 1 -

Unit

3

Are exposures associated with disease?

**Session Content**

In this first session of Unit 3, we will cover the following topics:

1.1 Associations

1.2 Ratios

1.3 Risk ratios

1.4 Rate ratios

1.5 Odds ratios

1.6 Difference measures

1.7 Population Attributable Risk Proportion (PARP)

**Intended learning outcomes**

|  |
| --- |
| **By the end of this session, you should be able to:** |
| * Understand, explain and work with Ratios, Difference Measures and Population Attributable Risk Proportion.   In addition, you will practice the following **academic skills** in this session:  - Understanding and interpreting a graph;   * Calculating and interpreting ratios and differences |

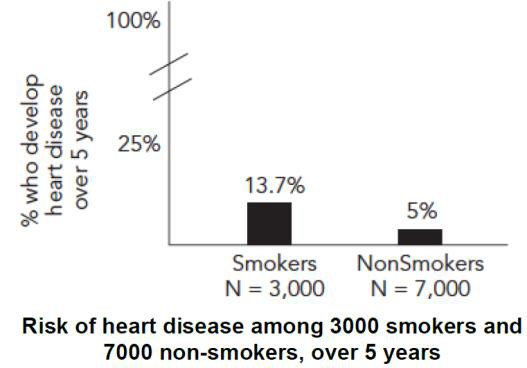
* 1. **Associations**

An observational descriptive study starts with measures of disease occurrence and frequency. In an analytic study measuring the association involves the comparison of two measures in exposed and unexposed populations.

### Example: Pholeleni associations

In our hypothetical community / population, Pholeleni, 10 000 people are enrolled in a study without heart disease. They are followed up for 5 years. In this population, 3000 people smoke and 410 of smokers develop heart disease. We will assume this is a stationary population and there is no loss to follow-up or change in smoking status over time.

The results are presented in the graph on the next page.



***Task 1: Understanding and interpreting a graph***

*Study the graph above and answer these questions:*

1. *What is the risk of disease among the exposed (smokers)?*
2. *What is the risk of disease among the unexposed (non-smokers)?*
3. *How much larger is the risk between exposed and unexposed?*
4. *Is the difference between risk in exposed and unexposed* ***meaningful****?*

***Note****: Check the answers to the tasks in the Unit 3 Feedback document in your Module Resources*

### 

### 1.2 Ratios

Relative ratios are a way to quantify the **magnitude of difference** between two measures of disease occurrence or frequency. We will now revise the different ratios used to measure associations in epidemiology.

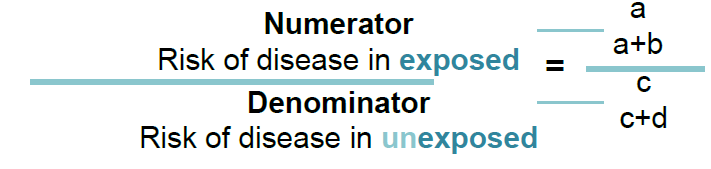
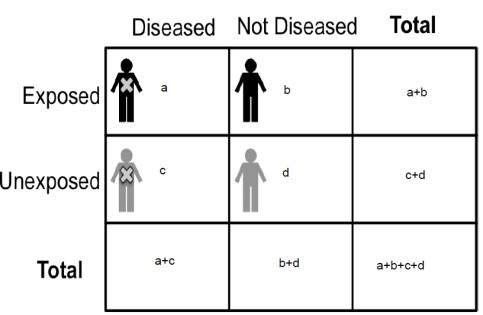
### 

### 1.3 Risk ratios

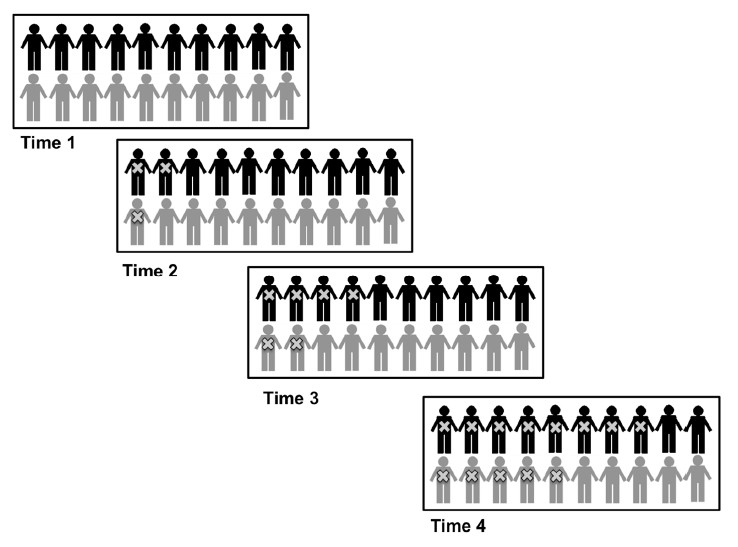
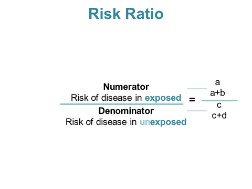
In calculating a Risk Ratio the:

**Numerator =** Conditional risk of disease among **exposed**

**Denominator =** Conditional risk of disease among **unexposed**



An example:



**Figure 1: Disease incidence over 4 time periods in Pholeleni**

* Ratios > 1.0 indicate rate is **higher** among exposed than unexposed
* Ratios = 1.0 indicate **no** association

Ratios < 1.0 indicate rate is **lower** among exposed than unexposed

|  |
| --- |
| ***Task 2 – Calculate and interpret Risk Ratios in a study***  *Draw up a 2x2 table at time 4 in this study assessing disease incidence in an exposed and unexposed population.*  *What is the incidence of disease in the exposed and unexposed? What is the Incidence Risk Ratio?*  *How does one interpret a Risk Ratio?* |

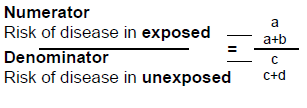
### 1.3.1 Risk Ratio - 95% Confidence Interval (CI)

Because a sample of the study population is used in most studies, chance is introduced. A sample, by chance, will often not represent **exact** disease and exposure experience of population. Confidence intervals help to understand variability in study estimates **due to chance** in sampling process between the results obtained from the sample and the confidence that this reflects the measure in the Study Population.

### Steps in calculating the Confidence Interval:

1. Take natural log of Risk Ratio = ln (Risk Ratio)
2. Estimate standard error (SE)

Incidence Risk ratio =

 = (8/10)/ (5/10) = 1.6

1. Estimate upper and lower bounds on log scale - 95% confidence interval **upper** bound

ln (Risk ratio) + 1.96 (SE [ln (Risk ratio)]) - 95% confidence interval **lower** bound

ln (Risk ratio) - 1.96 (SE [ln (Risk ratio)])

1. Exponentiate upper and lower bounds
2. Report and interpret estimate and confidence interval

### A sample interpretation would be stated as:

In these data, the exposed individuals had **[risk ratio estimate]** times the risk of the outcome compared with the unexposed, with a 95% confidence interval for the observed risk ratio ranging from **[lower bound]** to **[upper bound].**

### *Example*: Risk Ratio - 95% Confidence Interval

A study was conducted to measure association between family history of Alzheimer’s Disease (AD) and incidence of AD among those aged >70 years.

A random sample of 1000 individuals aged >70 years with no symptoms of AD, were followed for 20 years. Symptoms of AD were measured every year. There were no losses to follow-up.

***Task 3 -***

*What is the incidence Risk Ratio of a history of family history and developing AD in those >70 years?*

*How confident can we be about this finding?*

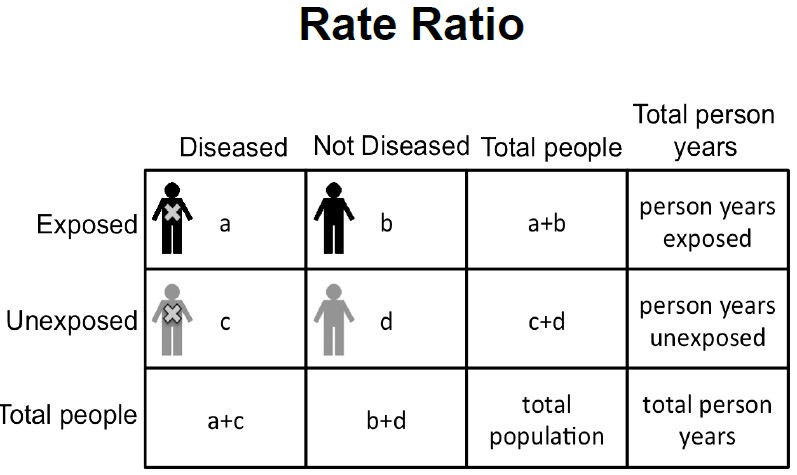
### 1.3.2 Central Limit Theory

The validity of confidence interval relies on Central Limit Theory (CLT) Remember the assumptions of CLT

* + - * Large sample size
      * Each cell in 2 x 2 ≥ 5

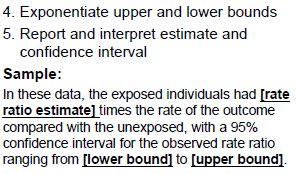
### 1.4 Rate Ratios

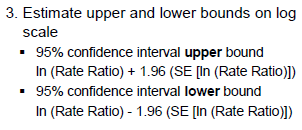
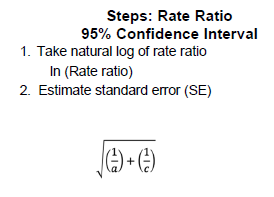
Risk Ratios are ideal if there is little or no **loss to follow-up.** Most studies have substantial loss to follow-up. When this occurs the **Rate Ratio** is a more accurate representation of incidence when loss to follow-up an issue.





Steps to calculate the confidence interval for a Rate Ratio:





The interpretation of a Rate Ratio issimilar to risk ratio

* Ratios > 1.0 indicate rate is **higher** among exposed than unexposed
* Ratios = 1.0 indicate **no** association
* Ratios < 1.0 indicate rate is **lower** among exposed than unexposed

**1.5 Odds Ratio**

An appropriate measure of association for a **prospective** study is a Risk or Rate Ratio.

If however the sample of individuals **with** and **without disease** and the exposure status is assessed **retrospectively** the appropriate measure of association is Odds Ratio.

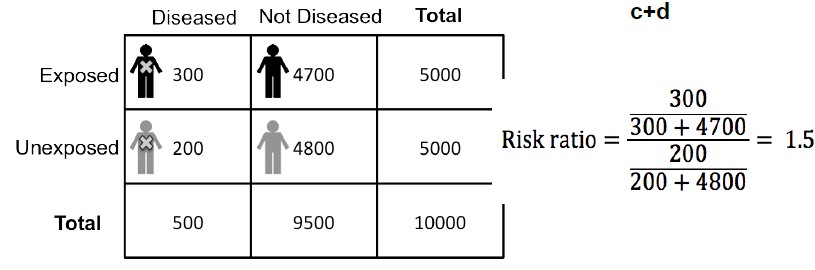
### *Example A*

**Research question**: Is smoking cigarettes during pregnancy a potential cause of offspring attention-deficit hyperactivity disorder (ADHD)?

### Sample:

Recruit 5000 women during pregnancy who are smokers, and 5000 women during pregnancy who are not smokers in Pholeleni. Follow them prospectively assuming no loss to follow-up.

**Measures**: Follow the offspring and at age 10 years, determine which children developed ADHD & which did not.



### Risk Ratio interpretation

From the prospective study, offspring of women who smoked in pregnancy have 1.5 times the risk of developing ADHD over 10 years compared to offspring of women who did not smoke in pregnancy.

***Task 4 – Calculate Risk Ratio and Odds Ratio***

*Draw the 2 X 2 table for this data.*

*What is the risk of ADHD among the exposed? Among the unexposed?*

*What is the Risk Ratio?*

*What is the odds of ADHD among the exposed? Among the unexposed?*

*What is the Odds Ratio?*

### *Feedback*

**Numerator =** Odds of disease in **exposed Denominator =** Odds of disease in u**nexposed**

### Odds Ratio interpretation

The odds of developing ADHD in the first 10 years of life among those exposed are 1.53 times the odds of disease in the unexposed.

### *Example A:* odds and risk ratio

* The estimated Odds Ratio = 1.53
* The estimated Risk Ratio = 1.5

These ratios are similar when outcome is **relatively rare** in the population (refer to Unit 1 Session 3)

### *Example B*: Odds Ratio

**Research question**: Is smoking cigarettes during pregnancy a potential cause of offspring attention-deficit hyperactivity disorder (ADHD)?

### Sample:

* 500 10-year-old children in Pholeleni who are seeking care for hyperactivity
* For each child we find with ADHD, we select two children of the same age from the same clinics who present for routine well visits (do not have ADHD) – a **purposive sample.**

### Case control study

**Measures**: Mothers respond to questions, including whether they smoked cigarettes while they were pregnant.

***Task 5 – Calculate Odds Ratio***

*Draw the 2 X 2 table for this data.*

*What are the odds of ADHD among the exposed and the unexposed?*

*What is the Odds Ratio?*

### *Example A and B*: Odds Ratios

* The estimated Odds ratio using a prospective cohort study design = **1.53**
* The estimated Odds ratio using a case control study design = **1.48**

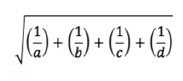
### 1.5.1 Odds Ratio: why we use it

* Exposure odds ratio and disease odds ratio are mathematically equal!
* In the **prospective** study, we estimated the odds of disease among exposed and odds of disease among unexposed
* In the **case control study**, we estimated the odds of exposure among the diseased and the odds of exposure among the non-diseased.
* When we **select our cases and controls correctly,** we get an unbiased estimate of the exposure odds even though we estimate the disease odds.
* This Odds Ratio is approximately equivalent to the Risk Ratio when the disease is rare.

The key thing is selecting the cases and controls **correctly**.

* + In our example, cases and controls were selected from the **same underlying population base** as the sample from the prospective study.
  + When cases and controls are selected from the same population base, the same estimate of the association between exposure and disease from the case control study is same as from a prospective study from the same population base.

### Steps: Odds Ratio - 95% Confidence Interval

1. Take natural log of odds ratio ln (Odds ratio)
2. Estimate standard error (SE)
3. Estimate upper and lower bounds on log scale
   * 95% confidence interval **upper** bound ln (Odds Ratio) + 1.96 (SE [ln (Odds Ratio)])
     + 95% confidence interval **lower** bound ln (Odds ratio) - 1.96 (SE [ln (Odds ratio)])
4. Exponentiate upper and lower bounds
5. Report and interpret estimate and confidence interval

### Sample: In these data, the exposed individuals had [odds ratio estimate] times the odds of the outcome compared with the exposed, with a 95% confidence interval for the observed odds ratio ranging from [lower bound] to [upper bound].

### 1.5.2 Summary: Odds Ratio

When using an Odds Ratio one cannot estimate the risk of disease directly as sampling is based on whether they have the disease or not (case control study).

With Odds Ration one can estimate proportion exposed among diseased and non-diseased

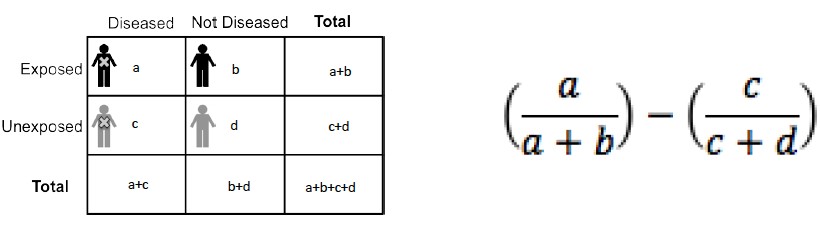
* + - * Estimate Odds Ratio for exposure
      * Odds Ratio for exposure = Odds Ratio for disease

If the disease is rarein population, the Odds Ratio approximates the Risk Ratio from a prospective study.

### 1.6 Difference measures

Difference measures are used in epidemiology as well.

### 1.6.1 Risk Difference between two risks



**Interpretation:** Excess risk due to the exposure

**Example:** If the risk of disease is 10 per 100,000 in the unexposed and 15 per 100,000 in the exposed, then 5 per 100,000 cases is associated with the exposure of interest.

### Example: Nutrition and obesity

**Research question**: Are nutrition classes in middle school associated with the development of obesity in adolescence?

### Sample:

* + - Middle school A, 400 students, receives health education (intervention)
    - Middle school B, 300 students, in neighboring district, does not receive health nutrition class
    - Purposive

**Measures**: Schools collect students’ height and weight yearly for 5 years

***Task 6 – Calculate Risk Ratio and Risk Difference***

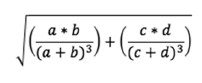
*Draw up the 2 X 2 table.*

*What is the incidence risk of obesity in those who had nutrition class? What is the incidence risk of obesity in those who had nutrition class? What is the Incidence Risk Ratio?*

*What is the Incidence Risk Difference? How would you interpret this?*

### Steps in calculating the Risk Difference- 95% Confidence Interval

1. Estimate standard error (SE)



1. Estimate upper and lower bounds
   * 95% confidence interval **upper** bound

Risk difference + 1.96(SE [Risk difference])

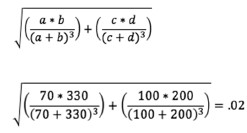
* + 95% confidence interval **lower** bound

Risk difference - 1.96(SE [Risk difference])

1. Report and interpret estimate and confidence interval

* Sample **Positive**: In these data, exposure was associated with an excess of **[risk difference estimate]** cases compared with the unexposed, with a 95% confidence interval for the observed excess cases ranging from **[lower bound]** to **[upper bound]**.
* Sample **Negative**: In these data, exposure was associated with an **[risk difference estimate]** fewer cases compared with the unexposed, with a 95% confidence interval for the observed decrease in cases ranging from **[lower bound]** to **[upper bound]**.

### Steps: Risk Difference: 95% Confidence Interval

1. Estimate standard error (SE)
2. Estimate upper and lower bounds
   * 95% confidence interval **upper** bound Risk difference + 1.96(SE [Risk difference]) - 0.15 + 1.96 (0.02) = - 0.12
   * 95% confidence interval **lower** bound Risk difference - 1.96(SE [Risk difference])

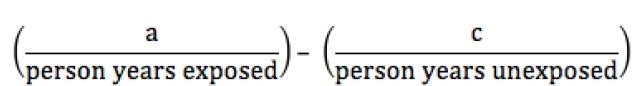
- 0.15 - 1.96 (0.02) = - 0.19

### Steps: Risk Difference 95% Confidence Interval

1. Report and interpret estimate and confidence interval

* **Negative**: Middle high nutrition education is associated with 15 fewer cases of obesity per 100 adolescents over five years, with a 95% confidence interval for the observed decrease in cases from 11.6 to 19.4 fewer cases.
* ***Interpretation:*** *There are approximately 16 fewer cases of obesity during* adolescence for every 100 adolescents associated with nutrition class in middle school.

### 1.6.2 Rate Difference

Rate Difference is the difference between two rates

***Example*:** If the rate of disease is 8 per 100,000 person years in the exposed and 4 per 100,000 person years in the unexposed, then 4 per 100,000 person-years of exposure is associated with the exposure of interest.

***Task 7 – Compare uses of Ratio Measure and Difference Measure***

*When is a ratio measure appropriate versus a difference measure?*

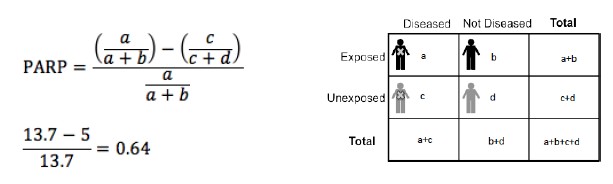
*Why would we use one over the other?*

### 1.7 Population Attributable Risk Proportion (PARP)

A population attributable fraction is a measure of the proportion of the total disease burden associated with exposure.

*Example:* PARP

* Proportion of people who develop heart disease among smokers and nonsmokers
* Risk of heart disease in smokers = 13.7%
* Risk of heart disease in nonsmokers = 5%
* PARP would be calculated as:



Interpretation A: 64% of the heart disease in the population of Pholeleni is potentially attributable to smoking.

Interpretation B: If we were to convince all of the smokers to quit, we would reduce the incidence of heart disease by 64%.

PARP is particularly useful measure in **public health practice.**

Session 2 –

Unit

3

What is a cause?

This session addresses Step 5 of the Seven Step Approach, which is to **rigorously evaluate whether the association observed suggests a causal association**.

**Session Content**

In this session of Unit 3, we will cover the following topics:

2.1 A motivating example

2.2 Counterfactual

2.3 Necessary and sufficient causes for populations

2.4 Disease causation – time and space

2.5 Public health implications

2.6 Disease causation in a non-deterministic world

2.7 Hypothesis testing

2.8 Evaluating causal associations

2.9 Making a decision – in summary

**Intended learning outcomes**

|  |
| --- |
| **By the end of this session, you should be able to:** |
| * Understand, describe and work with concepts related to disease causation in populations, and the public health implications of these. * Evaluate causal associations in analytical epidemiological studies using the Bradford-Hill criteria.   In addition, you will practice the following **academic skills** in this session:   * Interpreting visual data * Applying theory to understand data * Applying theory to familiar situations * Interpreting data to make a public health decision * Identifying types of causes * Thinking critically |

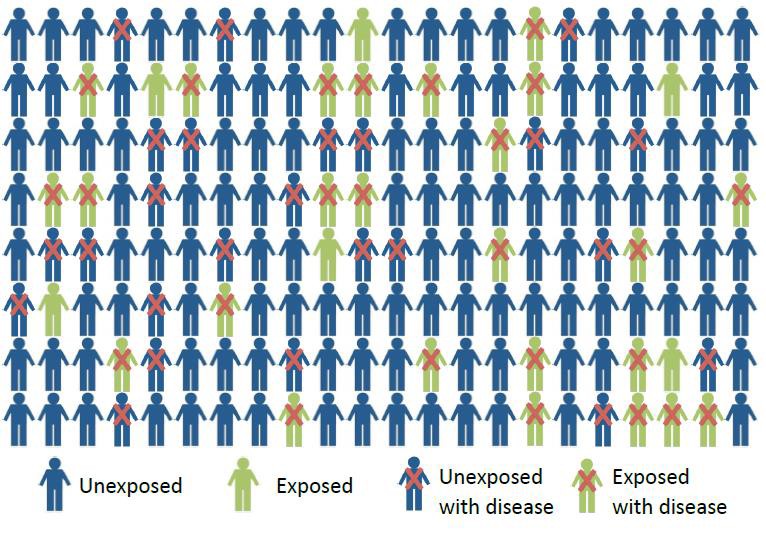
### 

### 2.1 A motivating example

Uncle Joe says:

“*My mother smoked in pregnancy - no one knew the health dangers back then - and I’m just fine. All of these warnings about smoking during pregnancy are overdone*.”

If one person smoked in pregnancy with no adverse consequences for their offspring, can we conclude that smoking in pregnancy is not harmful?



***Task 8 – Interpret visual data***

*How would you describe the population depicted in Figure 1?*



**Figure 1: An example of a population to be studied in epidemiology**

**Key concept to start with**: Epidemiologists understand disease causation to be a multifactorial process.

### 2.2 Counterfactual

One of the current concepts which are used to explain causality in epidemiology is the counterfactual. The counterfactual is the condition that is **counter to the fact**. Put differently:

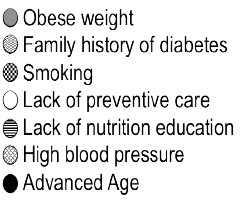
* + A factor is a cause if the **outcome would not have occurred** in the **absence of that factor**, holding all **other things constant**, including space and time.

Here is another example of disease causation:

### *Example*: Individual diabetes

Take Type II diabetes, a multifactorial disease that is increasingly prevalent worldwide. There are many different combinations of types of causes that can bring about diabetes in a person. So, for example there are different risk factors (depicted as marbles in the

schematic) for smoking (a cause of this disease) and for obesity (also a cause of this disease).



In this example there is a combination of factors.

Two individuals have Type 2 diabetes. The third person despite having risk factors does not have the disease:

**Person 1.** Obese weight, lack of preventive care, diabetes family history, 20 pack-years smoking

**Person 2.** Poor nutritional education, diabetes family history, high blood pressure, advanced age

**Person 3.** Obese weight, diabetes family history, high blood pressure

***Task 9 – Apply theory to understanding of data***

*For persons 1-3 above, consider whether each factor is* ***necessary and******sufficient?***

In explaining these concepts the notation used is that of ‘marbles’, where each marble is a

**component cause** or risk factor.

* But one marble is rarely sufficient to cause disease
* A particular **marble set** can be a **sufficient cause** for disease
* There can be more than one marble set that become a sufficient cause

Causes are rooted in a **counterfactual** definition, each marble is a necessary cause of disease **for that particular sufficient cause set.** If every other component cause were kept constant in an individual, the disease would occur if the factor would need to be present.

**2.3 Necessary and sufficient causes for populations**

A risk factor is n**ecessary** if all cases of disease require the cause in order for disease to onset. A risk factor is s**ufficient** if all individuals exposed to the cause will acquire the disease.

* Causes can be
  + necessary and sufficient
  + unnecessary but sufficient
  + necessary but insufficient
  + unnecessary and insufficient

Here are some examples:

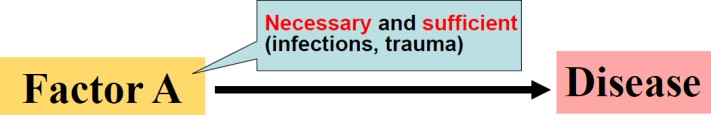
### Necessary and sufficient causes: Trisomy 21 and Down Syndrome

All individuals with three copies of the 21st chromosome will evidence Down Syndrome.

Trisomy 21 is thus sufficient for DS.

All individuals with Down Syndrome have three copies of the 21st chromosome.

Trisomy 21 is thus necessary for DS.



***Apply theory to a familiar situation:***

*Think of another 2 causes that would be necessary and sufficient to cause a particular outcome in your field of interest or expertise*.

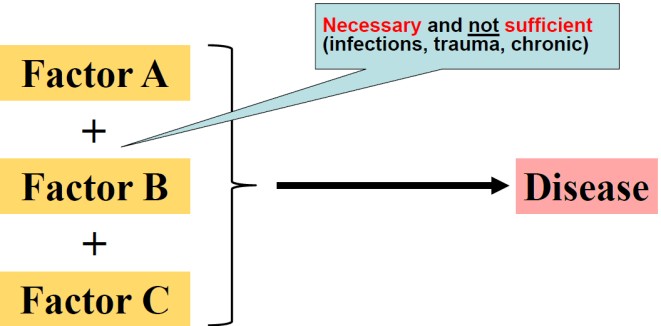
### Necessary but insufficient causes - Alcohol consumption and alcoholism

Not all individuals who consume alcohol will develop alcoholism.

Alcohol consumption is thus insufficient for alcoholism.

However, all individuals with alcoholism will have consumed alcohol.

Alcohol consumption is thus necessary for alcoholism.



***Apply theory to a familiar situation:***

*Think of another 2 causes that would be necessary and but not sufficient to cause a*

*particular outcome in your field of interest or expertise*.

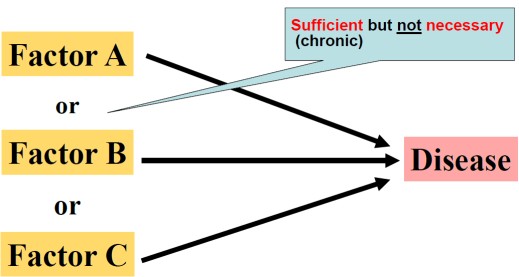
**Unnecessary but sufficient cause - Hysterectomy and pregnancy prevention**:

All women who have a hysterectomy are unable to become pregnant.

Hysterectomy is thus sufficient for pregnancy prevention.

Not all pregnancies are prevented through hysterectomy.

Hysterectomy is thus unnecessary for pregnancy prevention.

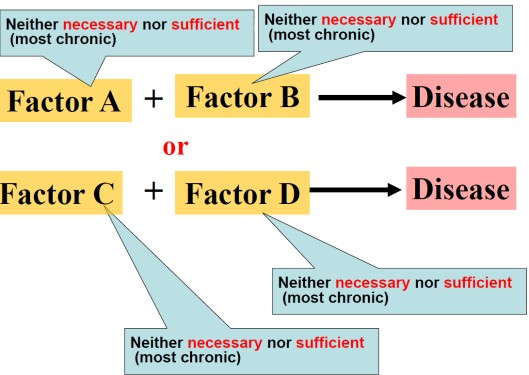


***Apply theory to a familiar situation:***

*Think of another 2 causes that would sufficient but not be necessary and to cause a particular outcome in your field of interest or expertise*.

### Unnecessary and insufficient cause - Smoking and lung cancer:

Not all individuals who smoke will develop lung cancer. Smoking is thus insufficient to cause lung cancer. Not all lung cancer cases are smokers. Smoking is thus unnecessary to cause lung cancer.



***Apply theory to a familiar situation:***

*Think of another 2 causes that would neither be necessary nor sufficient to cause a particular outcome in your field of interest or expertise*.

### 2.4 Disease causation – time and space



This example (of lung disease) shows one individual progressing through his life course, from birth, to adolescence, to young adulthood, to older adulthood. Some risks (marbles) are present at birth, some are acquired at one point, and some slowly accumulate.

For example, this person had one marble prior to birth (the black marble at the top of Figure 2). This marble may represent exposure to tobacco smoke or other toxins in utero.

He also acquired a series of grey marbles in childhood. These may represent ongoing exposures to, for example, chronic poverty or unhealthy home environment.

In adolescence, he began being exposed to the white marbles, which accumulated yearly, with increasing amounts of exposure during adulthood. This might represent cigarette smoking, which begins for this individual in adolescence in a low dose, and then increases in adulthood and remains a consistent exposure throughout his life.

Additional marbles began accumulating during adulthood, until disease onset.

**Figure 3: Collecting risk across the life course**

### Risks and thus causes are not independent

* They may be shared across individuals
* One person’s marble collection may influence another person’s marble collection
* Example, person-to-person infectious disease transmission

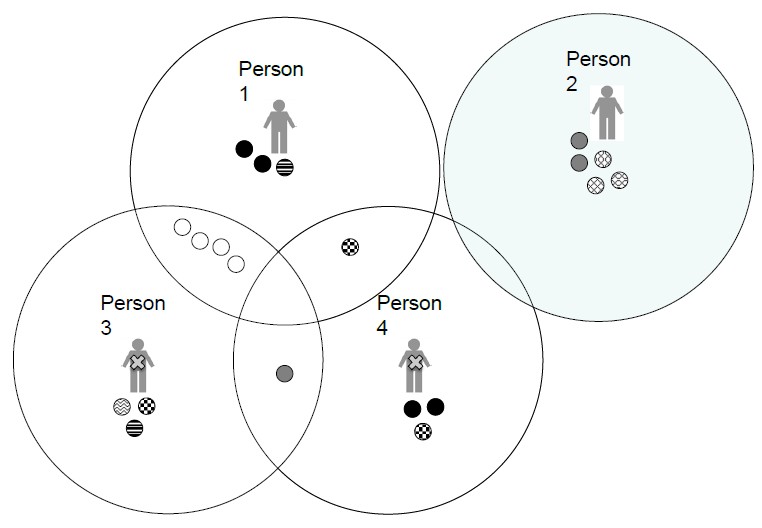
### Examples of common risk factors in a population (shared marbles)

* Unhealthy food environment e.g. People living in a Low Income township may share the same unhealthy food environment
* Community violence
* Social norms around substance use and cigarette smoking

e.g. adolescents are more likely to begin smoking if an influential peer begins

* Policies and laws managing access to quality health care

### Sharing marble exposures



Each marble represents an exposure. Marbles with the same colour and pattern are varying amounts of the same exposure. Some marbles are shared between individuals - for example, Person 1 and Person 3 share four clear marbles, and Person 3 shares one grey marble with Person 4. Individuals also have marbles that are unique and unshared - for example Person 1 has two black marbles that are unshared, as does Person 4.

Examples of shared marbles could be neighbourhoods, social norms, or sex partners.

Examples of unique marbles could be genetic alleles, dietary patterns, and experiences of personal trauma.

### Summary: marble analogy

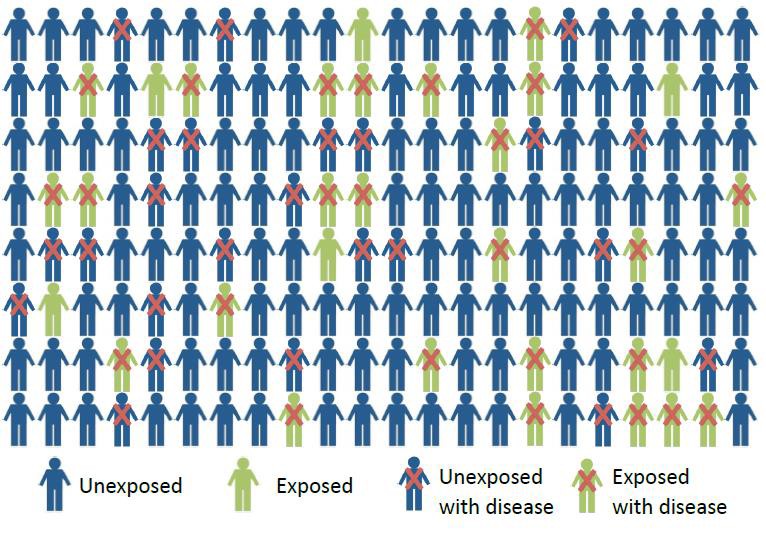
* Individuals share marbles and transmit marbles from one space to another
* Within each person’s space, there remains a complete set of marbles that is necessary to cause disease.
* One person’s complete set of marbles may differ from another person’s set

### Summary: causes in time and space

At a population level we need to understand shared exposure to risk such as unhealthy environments and transmission of disease. If these risks can be understood at a population level, preventive **interventions** can be planned (primordial, primary, secondary prevention) to prevent adverse health conditions. By identifying marbles that are **common** across many marble spaces, we can identify the exposures and environments for intervention and prevention efforts.

### 2.5 Public health implications

Each **individual’s** set of risks (marbles) that caused disease may be unique, with or without overlap across individuals. Each marble was n**ecessary** for that person to develop the disease when and how he or she did.



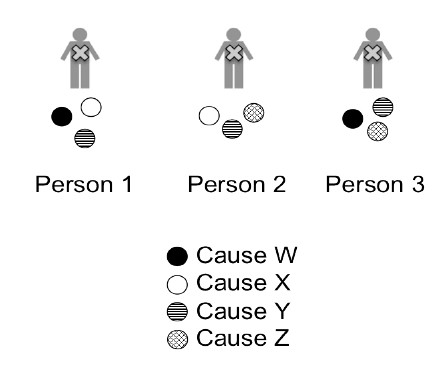
Epidemiologists look for the risks (marbles) that are most common across individuals with disease compared to those without disease **in populations.**

* + Preventing any one of the marbles can prevent disease in that individual.
  + Preventing common (risk / cause) marbles can prevent more disease in more individuals.

### Public health targets

***Task 10 – Interpret data to make a public health decision***

*In the example below, which cause should we try to prevent?*



### Summary of public health implications

* + At the population level causes may be necessary and/or sufficient, but need not be either.
  + Multifactorial and complex diseases are often caused by many factors necessary in at least one person.
  + Identifying factors, i.e. component causes, common to **most individuals** has the greatest impact on reducing disease for largest amount of people.

### 

### 2.6 Disease causation in a non-deterministic world

The process of disease development may begin in utero and continue until the moment that the disease occurs. Often many causes must align for a disease to occur in an individual.

### *Example*: Smoking and lung cancer

Smoking is not sufficient to cause lung cancer; smoking must act with other causes, i.e.:

* One individual smokes, works in an occupation with a high degree of exposure to asbestos, has a genetic predisposition to develop lung cancer;
* Another person who develops lung cancer may have a different constellation of causes;
* Causes can be shared across people or be unique to a certain person

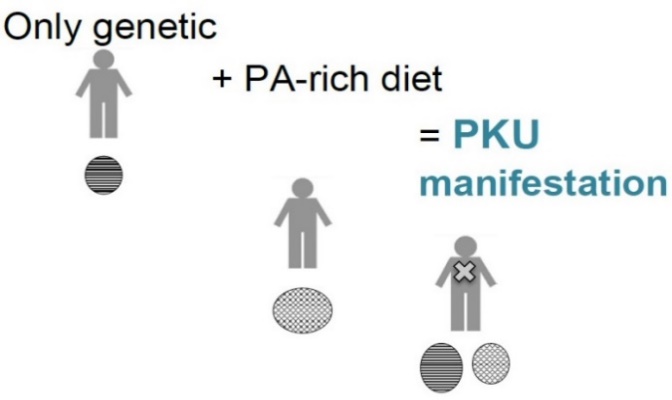
The idea that many causes must accumulate through the life course before the disease manifests is, in epidemiology, expressed as the concept of **interaction.**

* That is, if seven marbles are all necessary to cause disease in an individual, then all of these marbles interact with each other.
* By preventing exposure to even one marble, the disease will not occur.

### *Example:* Diet and phenylketonuria (PKU*)*

Phenylketonuria (PKU) is a rare disorder present at birth characterized by the inability to process a certain amino acid, phenylalanine. If untreated in children it results in altered appearance, hyperactivity, mental retardation, and seizures, among other symptoms. All PKU patients have specific maternal and paternal genetic sequences. However, alone this genetic abnormality will not cause PKU.

The necessary cause for PKU is inheriting a particular genetic sequence from both mother and father. However, this inherited genetic sequence will not result in the expressed phenotype unless an individual eats a diet rich in phenylalanine. Therefore a diet rich in phenylalanine is also necessary for the symptoms of PKU to manifest.



**Figure 4: Diet and phenylketonuria**

***Task 11 – Identifying types of causes***

*What type of cause is genetic sequence?*

### *Example*: Obesity

Obesity is an example of a multifactorial and complex condition. There are genetic variants involved in the process of increasing and maintaining high and weight. The in utero environment influences obesity in childhood and adulthood. Childhood factors including food insecurity, socio-economic position, availability of healthy food and food cost. Health behaviours including high consumption of sugar-sweetened beverages can contribute to obesity.

The potential causal mechanism for obesity could include:

– High consumption of sugar sweetened beverages

– + low physical activity

– + genetic predisposition

### There is no single cause of obesity. We need to conceptualize causes as interacting.

**Summary:**

Interaction: many causes must accumulate through the life course before the disease manifests.

* Necessary but insufficient causes interact for disease to manifest in an individual
* Multifaceted disease causation requires many component causes interacting

### 

***Hypothesis testing:*** *predict and then observe if correct*.

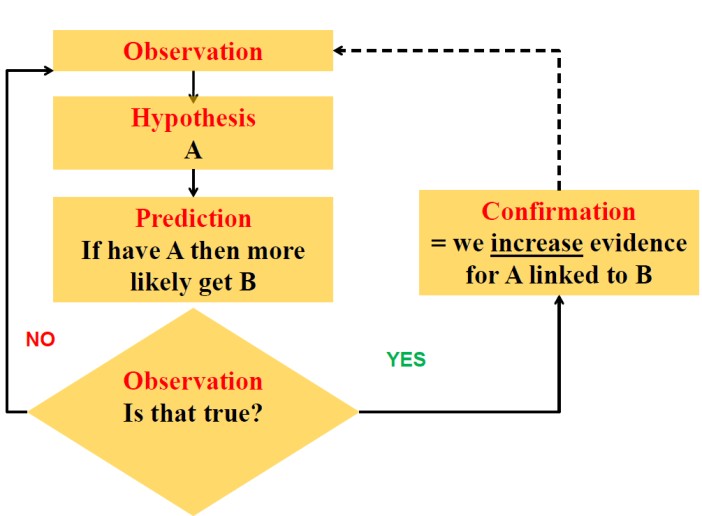
### 2.7 Hypothesis testing

Would the disease have occurred when and how it did without the exposure, or without the amount of exposure that occurred, the timing of exposure, or within the context of multiple exposures? This raises the conceptual (counterfactual) question: *When does exposure cause disease?*

A counterfactual test to see if an exposure is a cause would require us to:

1. Take the same person observed over the same time period, once with the exposure and once without the exposure;
2. Hold all other characteristics of the person, place and time constant;
3. Change only the exposure and observe if the health indicator changes.

This is, of course, impossible!



**The normal hypothesis testing schematic is depicted as:**

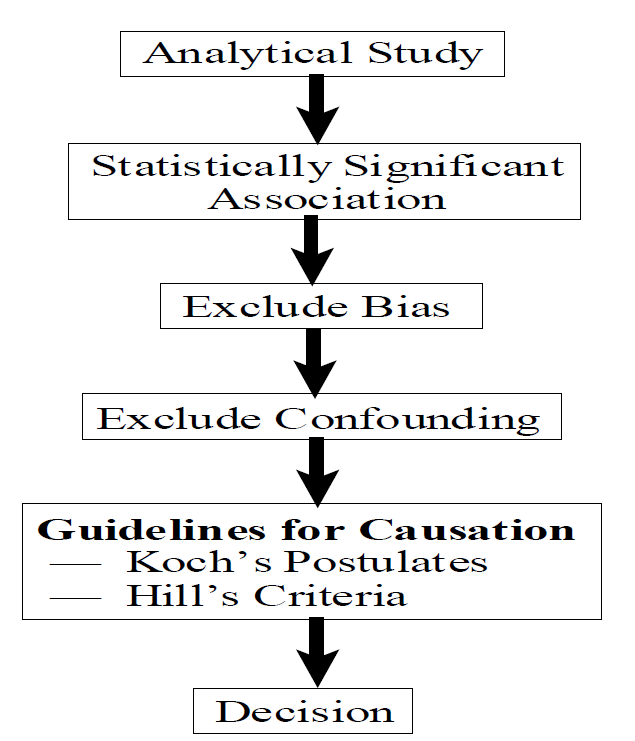
In an analytic epidemiological study the exposure and the outcome are observed. A measure of association is then calculated.

The practical question that needs to be answered is: Does the association that we measure in **our data** reflect the amount of excess disease that occurred due to the effects of the exposure, or could there be alternative explanations for the study findings other than a causal explanation?

2.8 Evaluating causal associations

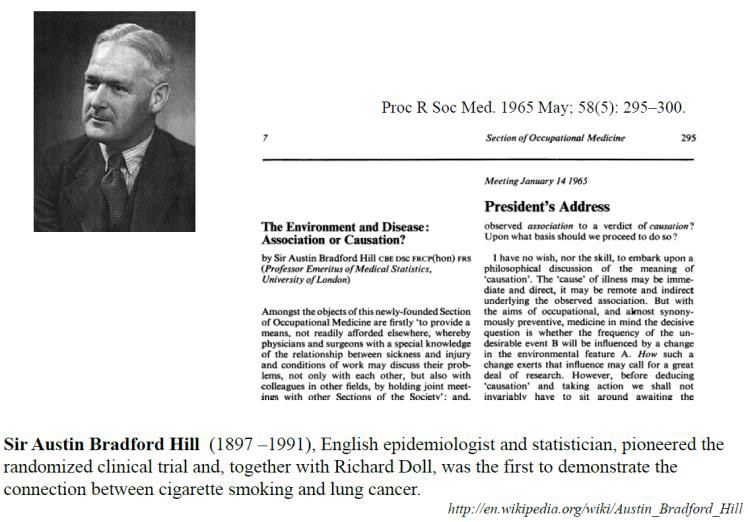
There is a process to determine whether an association is causal. In microbiology deciding whether a microbe caused a disease, uses Koch’s Postulates (1890). Koch’s Postulates require these defined criteria to be met for confirming causality:

1. The microorganism must be found in abundance in all animals suffering from the disease, but should not be found in healthy animals.
2. The microorganism must be isolated from a diseased animal and grown in pure culture.
3. The cultured microorganism should cause disease when introduced into a healthy animal.
4. The microorganism must be isolated from the inoculated, diseased experimental host and identified as being identical to the original specific causative agent.



In observational analytic epidemiological studies the Bradford-Hill criteria are used.

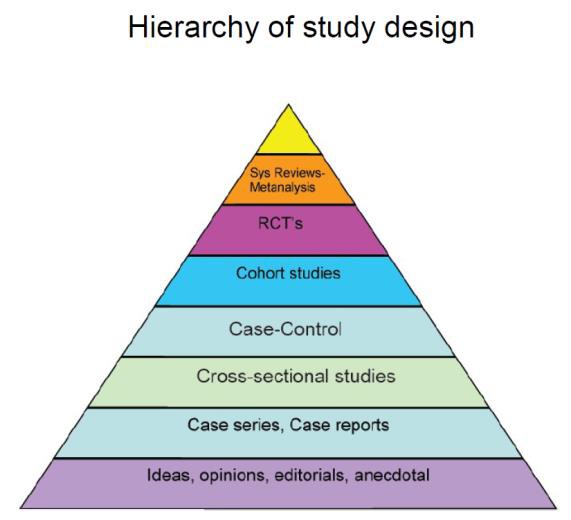
Bradford Hill’s criteria are guidelines for judging whether an association is causal.



The criteria that Hill defined include the ten criteria which are described below:

### Study design

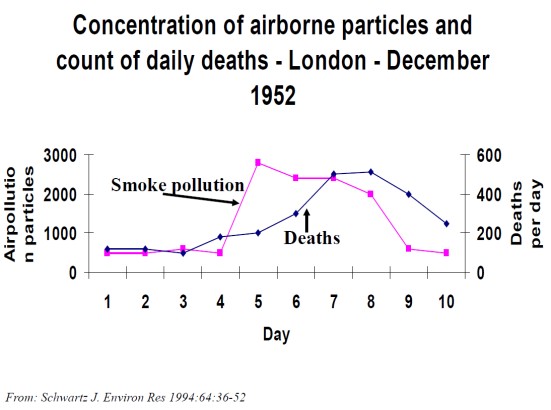
There is a hierarchy of study designs. Strong evidence is based on higher level study design:Experimental > cohort > case-control > cross-sectional.



* 1. **Temporal relationship**

The temporal relationship between the exposure and outcome (risk and disease) is probably the single most important of the Bradford-Hill criteria in determining whether an association is causal. It should be clear that exposure precedes onset of disease.

The length of interval between exposure and outcome needs to be sufficient.



### Strength of association

The stronger (size) and significance of relative measures of association, - the less likely it is to reflect bias, and thus be a non-causal association. Strength of association is linked to the hierarchy of study design. A strong association suggests that there is a pronounced excess of disease associated with the exposure.

***Task 12 – Think critically about measures of association***

*What are the different measures of association? How strong is “strong”?*

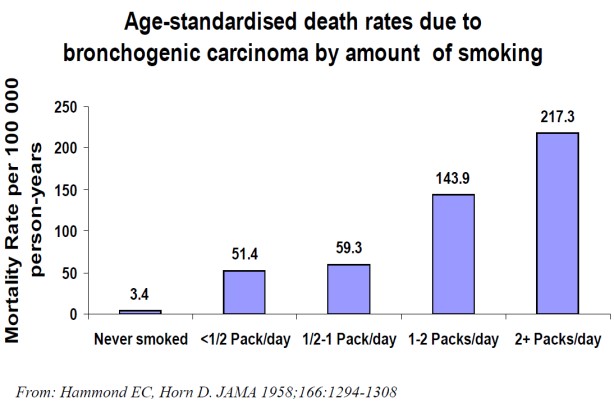
*Are relative or absolute measures of association more important in determining causality?*

* 1. “**Biological gradient**” or Dose-response relationship

Higher levels of exposure are associated with an increased risk of disease. This is a useful criterion but often there is a threshold level for disease.

1. **Plausibility**

There needs to be a plausible mechanism for disease causation, which should be consistent with known biological activity of suspected agent.

****

### 6. Consistency (Replication of findings)

Results need to be replicated in a number of different studies in different situations using different methods.

***Think about this question:***

*How was the principle of consistency applied in verifying that Medical Male Circumcision is efficacious in preventing HIV transmission?*

### 7. Specificity

Although the Bradford-Hill criteria included specificity i.e. that disease would only found in people exposed to suspected agent, we know this is not always applicable. It is not necessary as diseases can have many causes.

### 8. Experimental Evidence

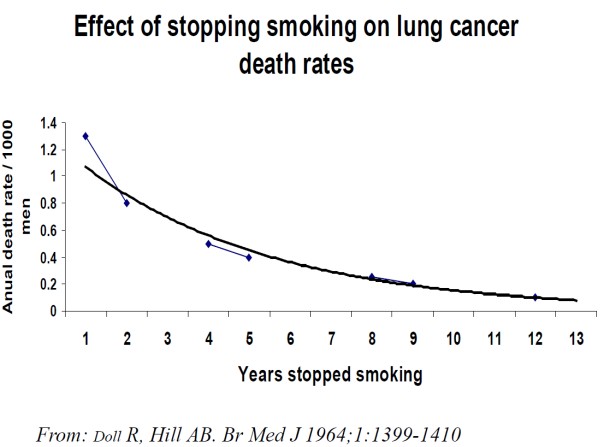
Demonstration that under controlled conditions changing E causes change to O is of great value if this info is available (but usually not available at time association first found).

### 9. Analogy

We are readier to accept arguments that resemble others we accept e.g. similar drugs cause similar things; not helpful just says “we go with what we know” but of course we might not know.

### 10. Removing exposure (cessation effects)

If removing of the exposure is followed by a reduction of disease risk, the likelihood of the association being causal is higher.



### 

### 2.9 Making a decision – in summary

In summary in order to decide whether an association is causal, a number of factors need to be considered:

1. No single criterion is completely reliable alone
2. Evidence may be conflicting
3. Weigh the available evidence
4. Criteria not equally important
5. Need adequate strength of association (how much?)
6. Correct temporal relationship is essential
7. Increased likelihood if different types of evidence point to same conclusion
8. Increased certainty as more evidence becomes available
9. Make a reasonable decision, but keep mind open to new evidence

*Note:*

* Correct temporal relationship is Absolute and Essential  Everything else is Relative and Non-essential*

# Session 3 -

Unit

3

# How do non-causal associations arise?

**Session content**

In this session we will cover the following topics:

3.1 Random chance

3.2 Associated causes

3.3 Confounders

3.4 Selection and follow-up of population

3.5 Bias

3.6 Sampling on disease status

3.7 Sampling based on disease

3.8 Misclassification

**Intended learning outcomes**

|  |
| --- |
| **By the end of this session, you should be able to:** |
| * Explain reasons for associations not being causal * Calculate measures of association and interpret results in terms of causality or non-causality   In addition, you will practice the following **academic skills** in this session:   * Calculating and interpreting data * Using your experience to identify causes * Identifying characteristics of study designs |

### 3.1 Random chance

One of the reasons why an association would not be causal is Random Chance. Variability occurs when taking a sample of a population by chance alone. It is possible that the sample collected may have an unequal distribution of other causes of disease.

Confidence intervals around the measures of association put bounds around the range of associations and reflect possible unequal distributions due to this chance. Investigators do not cause random chance, but can control the effect of random chance on study estimates by increasing sample size and precision of our measures.

### Random chance example

In a study to assess the association between alcohol consumption and the risk of breast cancer, the investigators collected a sample of women (without breast cancer) and recorded their average weekly alcohol consumption. The sample was followed for 10 years. The number of cases of breast cancer that occurred in in both groups was measured. The Rate Ratio was 2.0 (95% CI 1.2 to 3.5).

But the confidence interval does not tell us how likely the estimate is to be causal.

3.2 Associated causes

***Task 13 – Interpreting data***

*What was the denominator of the measure of occurrence? What does the Confidence Interval of the Rate Ratio tell us? What does this tell us about causality?*

Another reason for an association not being causal is that it could be due to Associated Causes. Non-comparability can arise between two or more groups simply because these groups are different from each other in ways that affect health. We must recognize group differences and account for them in epidemiological study design and analysis.

***Task 14 – Use own experience to identify associated causes***

*What might people who eat a lot of leafy green vegetables every day also do (compared to those who do not)?*

*How might the behaviour of women who use oral contraceptives differ from those who do not*?

These associated causes may be:

### 3.3 Confounders

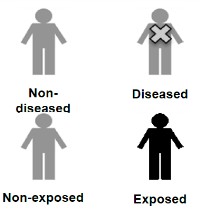
Differences across exposed and unexposed groups are often termed “confounders”. A confounder is a factor that **contributes to non-comparability** between the exposed and unexposed.

### Example:

People who engage in one particular potentially problematic health behavior are more likely to engage in other such behaviors, compared to people who do not engage in that particular health behavior.

We must consider what causes of outcome may ‘travel’ with hypothesized exposure and

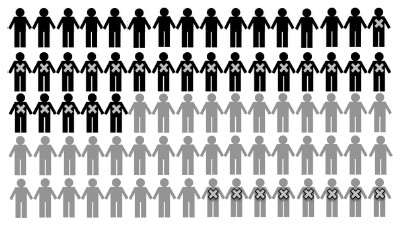
account for ‘traveling partners’ in epidemiological design and analysis.



### Example, alcohol consumption and oesophageal cancer

In this example the Study Population (of interest) is men ≥ 50 years living in Pholeleni. A sample of 800 men without oesophageal cancer was identified at the start of the study. They were followed for 20 years with no loss to follow up.

Out of 800:

* + 370 were heavy drinkers and of these 220 developed oesophageal cancer.
  + 430 are not heavy drinker and 8 developed oesophageal cancer

(In the schematic for this example – alcohol consumption and oesophageal cancer, each person in the figure represents 10 people in the sample.)

***Task 15 – Calculating and interpreting incidence risk***

*Draw a 2 X 2 table for the example above.*

*What is the incidence risk (probability) of disease in the exposed? What is the incidence risk (probability) of disease in the unexposed?*

*How would you interpret these findings? What is the Risk Ratio?*

BUT - How do we know this is due to alcohol consumption?

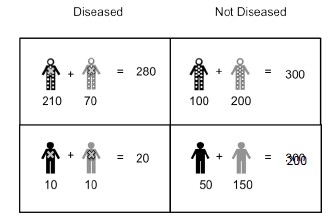
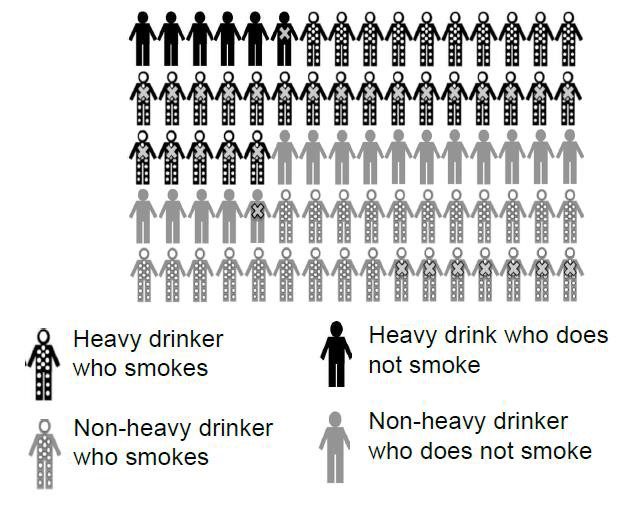
Individuals who consume heavy alcohol and individuals who do not are likely to be different in other ways as well. For example:

* + Heavy alcohol consumers are:
    - More likely to be men
  + Moderate alcohol consumers are:
    - More careful in watching their diet - consuming more fresh fruits and vegetables, less saturated fat
    - Less likely to be smokers
    - More active
    - More likely to go to regular check-ups with their doctor
    - Generally more health conscious

We do not know whether it was the heavy alcohol consumption that **caused** the adverse health outcome.

### Example, alcohol consumption and oesophageal cancer

In this example, men who consumed heavy amounts of alcohol were more likely to be smokers than people who do not. Could cigarette smoking be a potential risk factor for oesophageal cancer?



Could smoking be driving the observed association between alcohol & oesophageal cancer**?** Total smokers (with dots) = 580 (310 are heavy drinkers and 270 non-heavy drinkers)

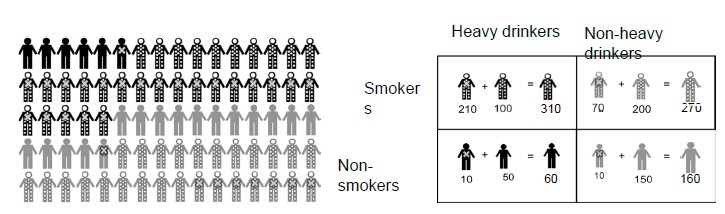
Of these smokers 280 have oesophageal cancer.

So the conditional probability of cancer among smokers = 48%.

Total non-smokers (without dots) = 220 (60 heavy drinkers, 160 non-heavy drinkers) Of the non-smokers - 20 have esophageal cancer

So the conditional probability of cancer among non-smokers = 9%

We can confirm in our study that smoking is associated with oesophageal cancer.

**But is there an association between smoking and heavy drinking?**

**In the schematic showing the association between smoking and drinking:**

Total smokers (with dots) = 580 (310 heavy drinkers, 270 non-heavy drinkers) So the conditional probability of heavy drinking among smokers = 53%.

Total non-smokers (with dots) = 220 (60 heavy drinkers, 160 non-heavy drinkers).

So the conditional probability of heavy drinking among non-smokers = 27%

### We can confirm in our study that smoking is associated with heavy drinking.

### So in this example:

* + Now we know that cigarette smoking is a risk factor for oesophageal cancer
  + We also know that heavy drinkers are more likely to be smokers
  + Therefore, smoking may be explaining the alcohol-oesophageal cancer association

### 3.4 Selection and follow-up of population

This is another reason that an association may not be causal. How we select our sample and how we follow-up our populations can also introduce (or minimize) non-comparability between groups in our studies.

Sampling can be:

* on exposure status; OR
* on disease status.

Follow-up of study subjects may lead to bias:

* Differential loss to follow-up by exposure
* Differential loss to follow-up by outcome
* Differential loss to follow-up by exposure and outcome

### 3.4.1 Exposure status based sample

For these samples we typically select disease-free individuals and follow them forward in time. Non-comparability arises if the unexposed distribution of other causes of disease is not the same as the distribution in the exposed group. This typically happens if we select exposure groups in a way that aggregates other causes of disease in one exposure group.

### 3.5 Bias

A study (or results) are “biased” when we suspect errors that result in deviation from the truth. Biases arise because of non-comparability between the groups being compared. Bias can be caused by selecting exposed and unexposed groups differently.

We never know the “truth” but we endeavor to design studies such that non-comparability is not introduced due to selection factors.

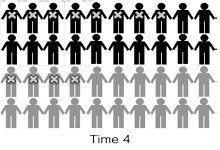
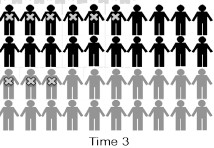
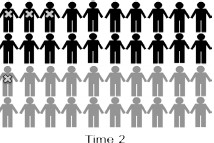
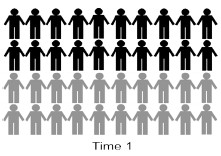
### 3.5.1 Selection bias – due to follow-up of study subjects

1. Differential loss to follow-up by exposure
2. Differential loss to follow-up by outcome
3. Differential loss to follow-up by exposure and outcome

### An example of bias due to loss-to-follow up - Drug education and drug use follow-up

In this study 400 individuals were followed up for 4 years. 200 had education about drug use and 200 did not have the education. The outcome was whether the youth had used illicit drugs.

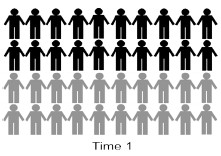
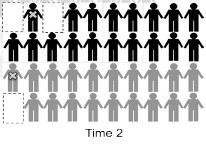
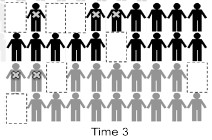
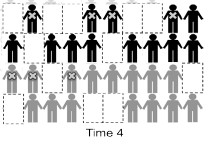
### The “truth”



***Task 16 – Calculating and interpreting measures of association***

*Draw a 2 X 2 table at time 4 for the above example.*

*Calculate the measures of association for the “truth”. What is your interpretation if this is the “truth”?*

Now assume that in this study done in Pholeleni there is loss to follow up.

***Task 17 – Calculating and interpreting based on data***

*Draw a 2 X 2 table at time 4 with loss-to follow up. Calculate the measures of association.*

*What is your interpretation if loss to follow up is the reality?*

### 3.6 Sampling on disease status

* + Follow-up of study subjects
* Differential loss to follow-up by exposure
* Differential loss to follow-up by outcome

### Differential loss to follow-up by exposure and outcome

* + Drug education and drug use loss to follow-up - exposure and outcome
  + Drug education and drug use loss to follow-up - exposure and outcome

In this example there was loss to follow-up associated with exposure and outcome.

When this happens all the measures of association – risk ratio, risk difference and odds ratio are likely to be underestimated.

### 3.7 Sampling based on disease

***Task 18 – Identify characteristics of study design***

*What study design samples are based on disease?*

*What are some of the key characteristics of this type of study design?*

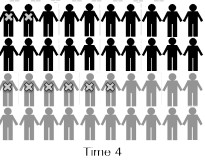
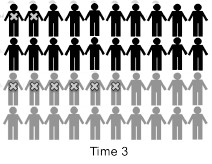
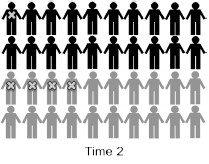
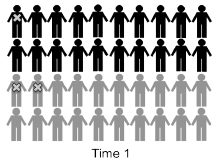
Efficient way to estimate measure of association between exposure and outcome

* + Central consideration of comparability is that exposed and unexposed are comparable on all other causes of disease
  + Conceptually similar to a prospective study with losses to follow-up dependent on disease status; if we do this right, the resulting odds ratios will be unbiased

### Example: dental floss and gum disease

In this study the investigators are interested in whether individuals who use dental floss have a lower incidence of gum disease than individuals who do not.

First, let’s see an example where we follow the whole population forward in time – the “truth”.



### Dental floss and gum disease “truth”

**Example: dental floss and gum disease**

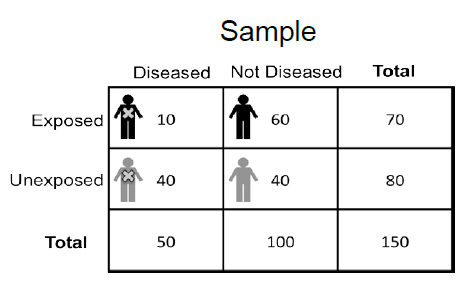
***Task 19 – Using data to calculate and interpret***

* *Draw a 2 X 2 table at time 4 with the “truth”.*
* *Calculate the measures of association.*
* *What is your interpretation of the “truth”?*

In this example there are different ways that this study can be conducted e.g. a case control study design could be used to conduct the study.

A sample of all individuals with gum disease could be collected from several dentists, and sample a group of individuals without gum disease at the same clinics, seeing dentist for another reason.

Doing this and using the Study Population in Pholeleni one could elect 50% of cases (50) and 30% (100) of non-cases from the population at Time 4.



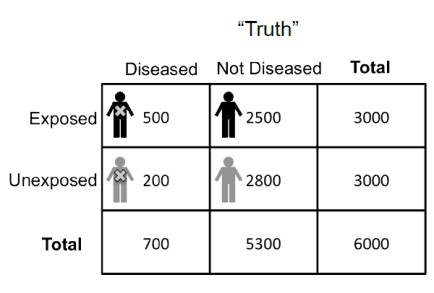
This is the 2 X 2 table using the same study population but the sample obtained using a Case Control Study design.

***Task 20 – Calculating and interpreting***

*What is the best measure of association for a disease status based sample?*

*Calculate the appropriateness of association based on this sample.*

*How do you interpret this?*

This study looks at the association between genetic risk factor and prostate cancer over 20 years among men 60 – 80 years old.

|  |
| --- |
| ***Task 21 – Calculating and interpreting***  *Calculate the measures of association for the “truth” What is your interpretation with the “truth”?* |

### 

### Why did this happen?

* + In underlying population of interest, among those without prostate cancer - 47% had the genetic factor
  + In the study sample, among those without prostate cancer - 64% had the genetic factor
  + There was a greater proportion of exposed individuals among non-cases (in the study) compared with non-diseased in the underlying population of interest
  + Bias occurred because of **non-comparability between exposed and unexposed** – i.e. those unexposed no longer represent the disease experience of the exposed if

they had not been exposed.

### 

### 3.8 Misclassification

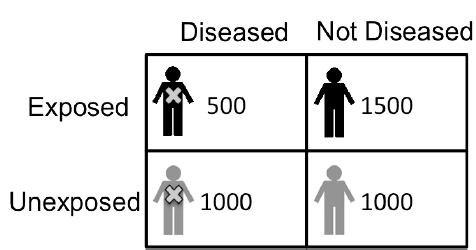
**Example**

A hypothetical study was conducted into the association between coffee consumption and risk of stroke among women in Pholeleni.

A study sample of 4000 women ≥ 60 years, with no history of stroke were recruited and followed for one year. There was no loss to follow-up.

Exposure: compared heavy coffee drinkers (>4 cups a day, on average) to all others.

Outcome: Stroke



### 

### 

### 

### “Truth”

***Task 22 – Calculating and interpreting***

*Which measures of association could be used? Calculate these.*

*How do you interpret them?*

### 

How could **misclassification occur in this sample?**

Take a random sample of 400 women with no history of stroke in Pholeleni. Each woman was asked how coffee they consume on average every day. The women were followed up for one year.

* + - The number of cases of stroke arising in heavy and non-heavy coffee drinkers was counted.
    - At the end of one year the incidence of stroke between groups was compared.

### How will women’s misreport of coffee consumption affect our results?

**Measurement error could occur:**

* + - Mistakes in how we record an individual’s value on a variable of interest in our study
    - Mistakes can arise from numerous sources
      * Respondents may not accurately report
      * Diagnostic test used may have errors
      * Problems coding responses in a database
    - Errors can be
      * Completely random
      * Associated with exposure status but not outcome status (or outcome status but not exposure status)
      * Associated with both outcome and exposure status

### 

### 3.8.1 Fundamental principle of misclassification, by exposure or disease

Misclassification leads to **non-comparability** between exposed and unexposed. Misclassification will usually render results that are **underestimates** of the true association.

**Example, error associated with exposure and disease status**

* + - Causal effect of using over-the-counter cold medication in pregnancy with congenital malformations in neonates
    - Population of interest is neonates over three years

### “Truth”

* + - **Pholeleni sample**
    - All cases of neonates with congenital malformations (hospitals in Pholeleni)
    - Across three years
    - Ascertain all 30 cases in population
    - Sample 300 healthy newborns from the same hospitals in the same time frame
    - Mothers report cold medication use pregnancy

### Sample “truth”

**Sample misclassification of exposure dependent on disease 1**

Errors in classification associated with both exposure and disease status can create non- comparability in observed data, leading to incorrect inference. The direction of the bias depends on which groups are more likely to be misclassified.

Session 4 -

Unit

3

Mitigating against non-causal associations in design and analysis?

### Session Contents

### In this session we will be looking at the following topics:

### 4.1 Comparability

### 4.2 Randomisation

### 4.3 Matching

### 4.4 Stratification

### 4.5 Factors in the causal pathway

**Intended learning outcomes**

|  |
| --- |
| **By the end of this session, you should be able to:** |
| * Identify, calculate and interpret measures of association in relation to randomization, stratification and other ways of avoiding non-causal associations.   In addition, you will practice the following **academic skills** in this session:   * Identifying measures * Calculating and interpreting * Using your knowledge and experience to give examples * Summarising information |

### 

### Comparability

Exposed and unexposed groups should be comparable on all factors associated with the disease other than the exposure. What is wrong with non-comparability?

### An example: In a study, 5000 smokers and 5000 non-smokers are followed for 10 years.

After 10 years, the smokers have 3.0 times the risk of motor vehicle crash fatality compared with non-smokers.

Are you comfortable reporting that smoking causes motor vehicle crash fatality?

Individuals who choose to smoke are **more likely** to engage in **other behaviors** with adverse consequences for health.

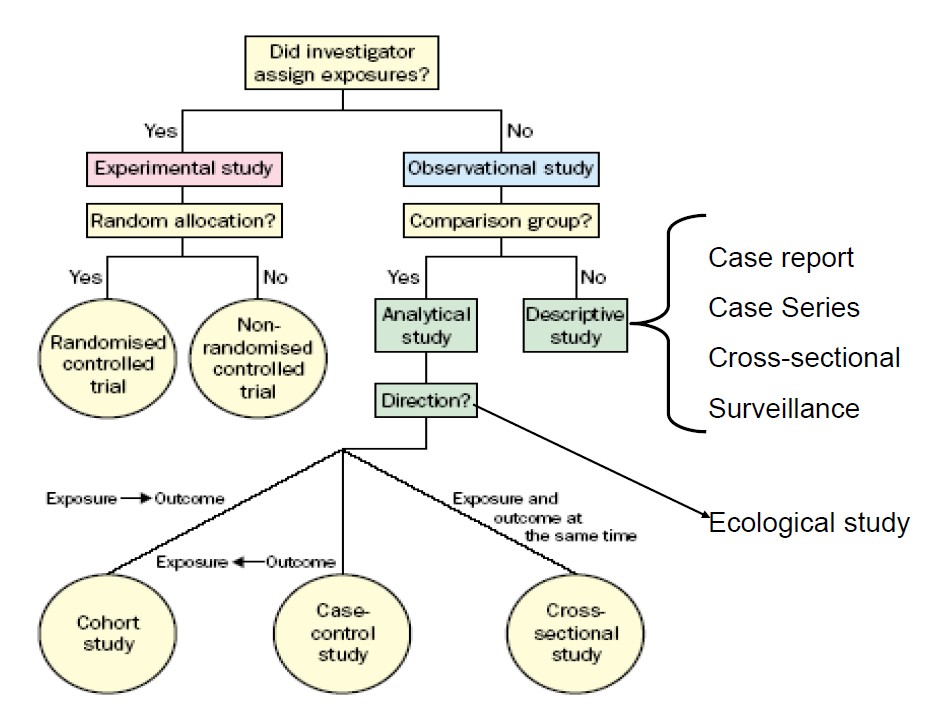
### Randomization

This creates **comparability** between groups, by removing the individual’s ability to **choose** their exposure status.

### 4.2.1 Randomized Control Trial (RCT)

The process of performing and RCT involves the following steps:

* + - Take a purposive sample from the study population
    - Assign individuals to be exposed or unexposed
    - Follow population forward to determine who develops outcome





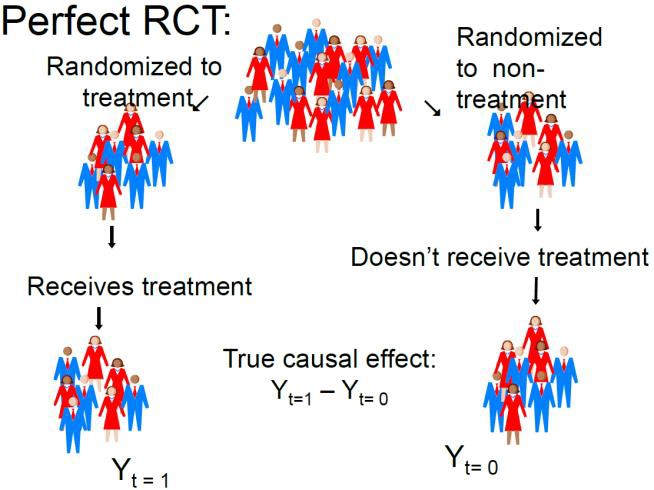
**Figure 5: Algorithm for classification of types of epidemiological study designs**

**The goal of RCT** is to be able to compare groups that are:

* “**different**” on just our main exposure that we are studying in relation to some outcome

### and

* the “**same**” on all the other important co-variates



### Why does randomization control for non-comparability?

**Example**

Two investigators conduct two separate studies, each exploring the effects of regular cardiovascular exercise on incidence of cardiovascular disease. The Study Population is post-menopausal women and the study hypothesis is*: I think exercise is protective against cardiovascular disease.*

### Example, study 1

In this study a purposive sample of 80 post-menopausal women with no history of cardiovascular disease is selected. These women are asked if they engage in ≥ 30 minutes of regular cardiovascular exercise ≥ 3 times/week (regular exercise compared to non-regular exercise). The control and intervention groups are both followed for five years. The outcome is cardiovascular events and the number occurring in each group is counted.

Assume this is a stationary population and there are no losses to follow-up.

These are the findings of Study 1 presented in a 2 X 2 table:



***Task 23 – Identifying measures, calculating and interpreting***

*What measures of association could be measured? Calculate these.*

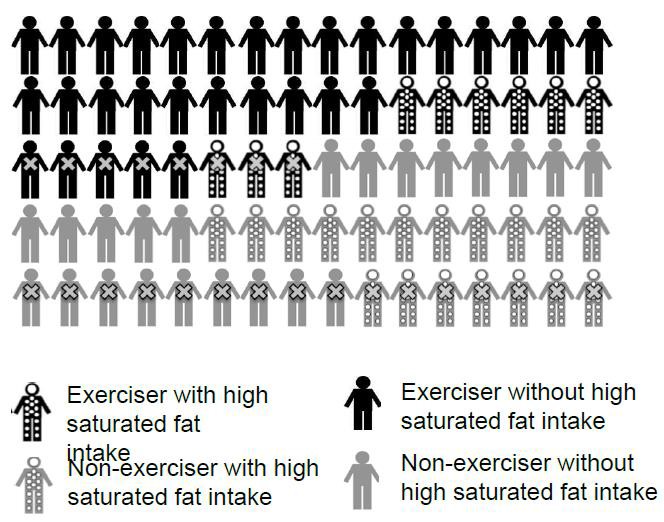
*What is the interpretation of the measures of association?*

### Are the results of Study 1 valid?

Are women who choose to exercise regularly may be more likely to be non-smokers, eat a more healthy diet, take multivitamins, etc.? Actually in this study the women who exercise had much lower average daily saturated fat intake than the non-exercisers.

So, we do not know whether the exercise had any causal effect on their cardiovascular health.

The impact of saturated fat could be represented schematically:



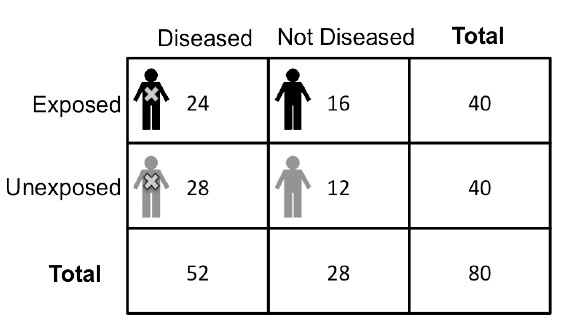
If there are 9 dotted people (high fat consumers) among 40 women who exercise, the total prevalence = **22.5%** of high fat consumption among the exercisers.

If there are 18 dotted people (high fat consumers) among the 40 non-exercisers, the total prevalence = **45%** of high fat consumption among the non-exercisers.

The conclusion is that there is a greater proportion of high fat consumers among the non- exercisers.

### Example, study 2

In this study a purposive sample of 80 post-menopausal women with no history of cardiovascular disease is selected. These women are **randomly assigned** to engage in ≥ 30 minutes of regular cardiovascular exercise ≥ 3 times/week (regular exercise compared to non-regular exercise). The control and intervention groups are both followed for five years. The outcome is cardiovascular events and the number occurring in each group is counted. Assume this is a stationary population and there are no losses to follow-up.

These are the findings of Study 2 presented in a 2 X 2 table:

***Task 24 – Calculating and interpreting***

*Calculate the appropriate measures of association?*

*What is the interpretation of these?*

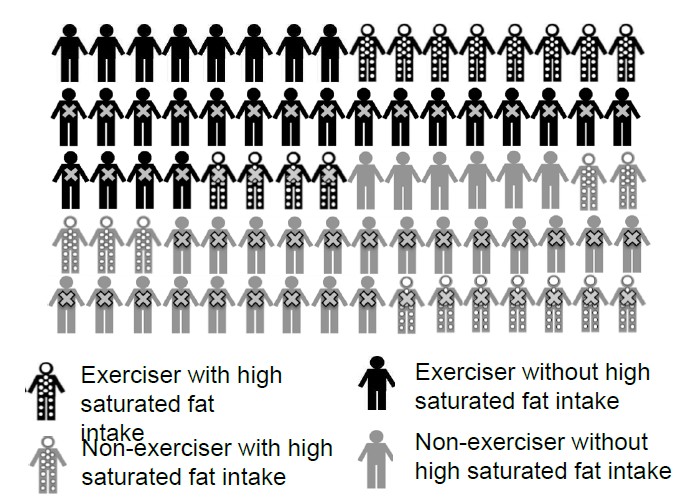
So - in Study 1 risk ratio = 0.5 and risk difference = -0.2 Study 2 risk ratio = 0.86 and risk difference = -0.1. Therefore, the effect is weaker in Study 2 than the effect in Study 1.

**Why?**

The impact of saturated fat could be represented schematically:

There are 12 dotted people (high fat consumers) among 40 exercisers

* Total prevalence = 30% of high fat consumption among the exercisers There are 12 dotted people (high fat consumers) among the 40 non-exercisers
* Total prevalence = 30% of high fat consumption among the non-exercisers

The conclusion is that when randomisation occurs there is the same proportion of excess high fat consumers among both groups.

**The strength of evidence from different study designs varies:**

* **Observational descriptive study designs** – The lowest level of studies use these designs, e.g. case reports, case series, prevalence surveys and cross-sectional descriptive studies.
* **Observational analytic studies** provide stronger evidence, rising from ecological, to analytic cross-sectional, case-control and finally to cohort study design.
* **Non-randomised experiments** including controlled before and after studies and interrupted time series analysis can provide some evidence.
* A good quality **randomised controlled trial** can provide a higher level of evidence.
* Only a **systematic review** of randomised controlled trials and a meta-analysis of RCTs provide higher level of evidence in health and medicine.

### Limitations to randomization

1. Equipoise and ethics
2. Complication and intention to treat analysis,
3. Placebos and placebo effects, a
4. Importance of blinding

### Randomization - summary

1. When randomization works, all factors that would differ between two groups who got to choose their exposure status are, on average, evenly distributed between the groups.
2. This includes all known risk factors for the outcome and a myriad unknown or difficult to measure.
3. Because they are evenly distributed across the groups, factors cannot affect the study estimates.
4. Randomized trials are a powerful way to **achieve comparability** between exposed and unexposed groups on both known and unknown factors that cause the outcome.

### Matching

Randomization is often **unethical** and **infeasible.**

***Use your own knowledge and experience:***

*Think of an example from your field of interest where randomisation may be unethical or infeasible.*

Matching controls non-comparability where randomization is impossible. Participants are usually matched on potential sources of non-comparability. Where randomisation is not possible matching is a common way to control for non-comparability in the design stage when setting up a study. Matching is done in observational studies.

* + - In a cohort study, exposed individuals are matched to ≥ 1 unexposed individuals on ≥ 1 factor(s) of interest.
    - In a case control study, diseased individuals are matched to a sample of disease free individuals from the same study population.

### Example

The research question: Is low regular consumption of fish oil associated with development of depression?

The study sample comprises 25 individuals with a first diagnosis of depression recruited from local mental health treatment center and 25 individuals with no history of depression from community surrounding mental health treatment centre.

In this study we are concerned about **sex** as a potential source of non-comparability:

* + - Women more likely to develop depression compared with men;
    - Women on average have more nutritious diets and more likely to supplement diets with fish oil.

Other potential sources of non-comparability to worry about (though we are not necessarily matching on) are age, alcohol and cigarette use, socio-economic factors.

If we were matching for sex, each time we select a case from the treatment centre, we would select one or more controls of the **same sex.**

### Stratification

**Non-comparability can be controlled at:**

Design stage by:

* Randomization
* Matching

Analysis stage by:

* Stratification

### How is stratification done?

In the study it would be important to collect data on variables that might contribute to non- comparability. Stratification enables us to answer the question: Is a potential factor related to non- comparability **associated** with the **exposure** and the **outcome**? Stratification removes effects of non-comparable variable on an exposure-outcome relation by limiting the variance on that outcome.

NB: The ability to control for non-comparability in analysis stage is only as good as the quality of measures of variables contributing to non-comparability.

### Stratification, example

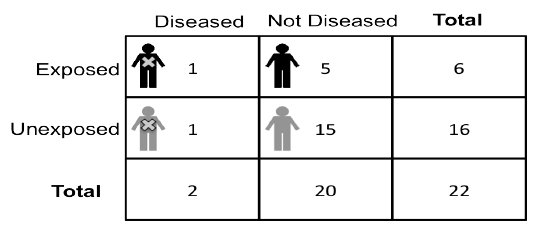
We want to examine the relation between alcohol consumption and oesophageal cancer among two groups:

Non-smokers

* + - Among individuals who have never smoked a cigarette in their lives, what is the relation between heavy alcohol consumption and esophageal cancer?
    - Smoking cannot confound the effect estimate because no individual in this subgroup has engaged in any smoking.

Smokers

* + - Among smokers (presumably around the same duration and average amount), were those who are heavy alcohol consumers more likely to develop esophageal cancer?
    - Smoking cannot confound the estimate because everyone is a smoker.



### Stratification example - non-smokers

Conditional probability of esophageal cancer among heavy alcohol consumers

= 1/6 or 16.7%

Conditional probability of esophageal cancer among not heavy alcohol consumers

= 1/16 or 6.3%

Risk ratio = 16.7/ 6.3 = 2.65

Risk difference = 16.7– 6.3 = 10.4

Interpretation:There is an increased risk of esophageal cancer among heavy alcohol consumers, even in the sub-population of individuals who do not.



### Stratification example – smokers

Conditional probability of esophageal cancer among heavy alcohol consumers = 21/31, or 67.7%.

Conditional probability of esophageal cancer among not heavy alcohol consumers = 7/27 or 25.9%

Risk ratio = 67.7 / 25.9 = 2.61

Risk difference = 67.7 – 25.9 = 41.8

Interpretation: There is an **increased risk** of esophageal cancer among heavy alcohol consumers, even in the subpopulation of individuals who **all smoke**.

### Stratification, example

* + - There is an **increased risk** of esophageal cancer among heavy alcohol consumers, even in the subpopulation of individuals who **do not smoke.**
    - There is an **increased risk** of esophageal cancer among heavy alcohol consumers, even in the subpopulation of individuals who **all smoke.**
    - Therefore, even when we limit variance on the possible source of non-comparability (i.e., smoking) there still remains an increased risk of esophageal cancer among heavy alcohol drinkers.

### What have you learned?

### List or make a mind-map of key points you have noted about stratification.

### Then check the Summary at the end of this session to compare it with your points.

### 4.5 Factors in the causal pathway

Is every variable that is associated with exposure and outcome a potential source of non- comparability? **No.**

Factors in the causal pathway are not non-comparable variables.

1. Factors that are on the causal pathway of interest between the exposure and outcome do not contribute to non-comparability.
2. If we control for them, we will obstruct the ability to observe the true effects of the exposure on the outcome.
3. Factors on the causal pathway of interest should not be controlled.

### 4.5.1 Factors in causal pathway, example

We are interested in prenatal exposure to tobacco smoke and its effect on offspring growth restriction during puberty.

The research hypothesis is that prenatal exposure to tobacco causes low birth weight, and then this low birth weight causes growth restriction in puberty. Because birth weight is in the causal pathway – we should not control for it.

### What if we do control for birth weight through stratification?

1. Among offspring with low birth weight, we **would find** that exposure to tobacco smoke is unrelated to offspring growth restriction:

– We restricted analysis to only those with the intermediary outcome of interest - low birth weight.

1. Among offspring with normal birth weight, we **would not find** an association between the exposure and outcome:

– We restricted analysis to only those without the intermediary outcome – low birth weight.

|  |
| --- |
| Summary - stratification  1. Careful and rigorous measurement of potential non-comparable variables is key to control for non-comparability in data analysis 2. Before stratification, always check that potential non-comparable variables are associated with exposure and outcome under study 3. If a variable is not associated with both exposure and outcome, then stratifying or otherwise controlling for that variable will not change the estimate of the effect of exposure on outcome |

Session 5 -

Unit

3

When do causes work together?

**Session Content**

In this session we will cover the following topics:

5.1 Concept of interaction

5.2 Looking for interaction in data

5.3 Interaction across scales

5.4 Additivity, multiplicativity and interaction

5.5 Random variation

We will also refer to the 6th step in Seven Steps approach to epidemiology

### (epidemiologymatters.org).

### Look at the Seven Steps again to remind yourself of these:

1. Define the population of interest
2. Conceptualize and create measures of exposures and health indicators
3. Take a sample of the population
4. Estimate measures of association between exposures and health indicators of interest.
5. Rigorously evaluate whether the association observed suggests a causal association.

### Assess the evidence for causes working together.

1. Assess the extent to which the result matters, is externally valid, to other populations.

**Intended learning outcomes**

|  |
| --- |
| **By the end of this session, you should be able to:** |
| * Understand and explain the concept and types of interaction of causes. * Identify and interpret interaction in data.   In addition, you will practice the following **academic skills** in this session:   * Checking your understanding of a topic |

### Concept of Interaction

Interaction occurs when multiple component causes work together to produce a particular health indicator. Component causes of disease rarely act in isolation. In epidemiology exposures are typically one of a set of component causes that have to work together in order for a change to occur in the health indicator.

Causes interact when they work together as part of the same sufficient cause (marble set). Causes that interact are causes in which both factors are necessary to cause disease in at least one sufficient cause.

For example, what can ‘cause’ a sprinter to win a 100 meter sprint?

* + - Only trains for years **Does not win**
    - Only has tied running shoes **Does not win**
    - Only reacts promptly to the starter’s pistol **Does not win**
    - Trains for years, tied shoes, prompt reaction **Sprinter wins**

Conceptually this would be an example of interaction.

### Interaction in theory

1. We could determine with certainty who would get disease if we could measure every component cause in a sufficient cause;
2. Those exposed to all component causes would inevitably get disease;
3. Those who do not have all the component causes, would never get disease;
4. However, this is never the case, i.e., we can never know what all the component causes are and we therefore have to assess for causes that work together (i.e. interact) in our data.

We can observe interaction when measure of association for exposure and outcome varies across levels of third variable.

### Interaction example - alcohol consumption

**Question:** Is consuming alcohol before driving associated with risk of dying in a motor vehicle crash?

There may be other factors that can contribute to risk of dying in a motor vehicle crash, including time of day, wearing a seatbelt.

The key questions of interest here are:

* Does alcohol consumption cause a greater risk of dying in a motor vehicle crash?
* Does alcohol consumption interact with either (or both) time of day and seatbelt use in its causing motor vehicle crashes?

How would we answer these questions?

### We would need to know:

* + Amount of alcohol consumed before driving
  + Subsequent death in a motor vehicle crash
  + Time of day that driving occurs
  + Driver wearing a seatbelt

### A. Alcohol consumption and death – seatbelt use

Alcohol use is associated with greater risk of death.

* + - * + Among those who did not wear a seatbelt, the risk of dying in crash was 10% among those who consumed alcohol prior to driving and 6% among those who did not consume alcohol prior to driving

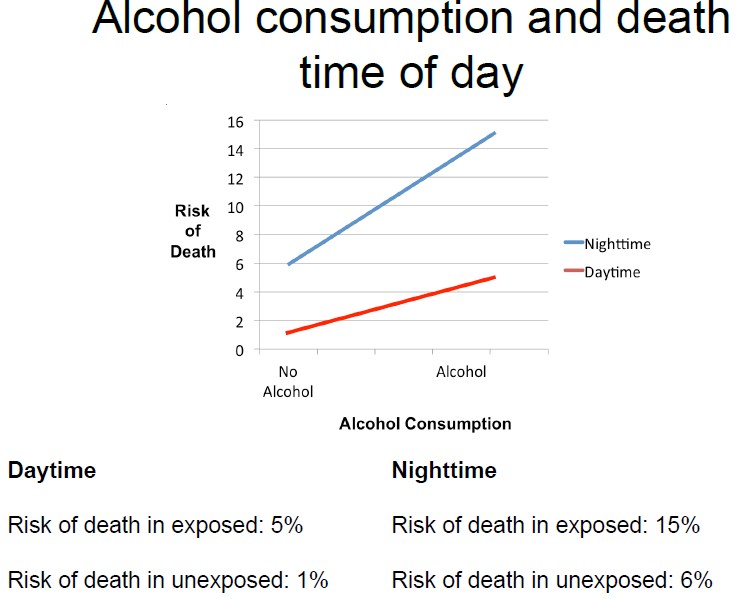
– Risk difference (RD) = 0.10 - 0.06 = 0.04 (95% CI 0.012, 0.06)

* + - * + Among those who did wear a seatbelt, the risk of dying in crash was 5% among those who consumed alcohol prior to driving and 1% among those who did not consume alcohol prior to driving

– Risk difference (RD) = 0.05 – 0.01 = 0.04 (95% CI 0.02, 0.05)

Therefore there is no difference in risk difference between those who do and do not use a seatbelt. Seat belt use and alcohol use are part of different disease causation ‘sets’ and do not operate jointly to cause crash death. This indicates **no interaction.**

### B. Alcohol consumption and death - time of day



Alcohol use is always associated with greater risk of death.

Time of day and alcohol use:

1. Among those who drove at night, the risk of dying in crash was 15% among those who consumed alcohol prior to driving and 6% among those who did not consume alcohol prior to driving

* Risk difference (RD) = 0.15 – 0.06 = 0.09 (95% CI 0.06, 0.12)

1. Among those who drove during the day, the risk of dying in crash was 5% among those who consumed alcohol prior to driving and 1% among those who did not consume alcohol prior to driving

* Risk difference (RD) = 0.05 – 0.01 = 0.04 (95% CI 0.02, 0.05)

Therefore there is a difference in risk differences associated with alcohol consumption for night-time drivers and for daytime drivers; this indicates the **presence of interaction.**

### Looking for interaction in data

1. Examine the evidence for interaction in data by comparing magnitude of association between exposure and disease across a third variable
2. If measure of association **differs** across levels of the third variable, there is **evidence of interaction** for that measure
3. If measure of association does not differ across levels of third variable - is not evidence of interaction

### Interaction across scales

The presence of interaction depends on the measure of association we are examining.

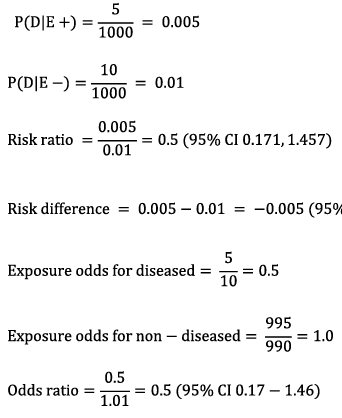
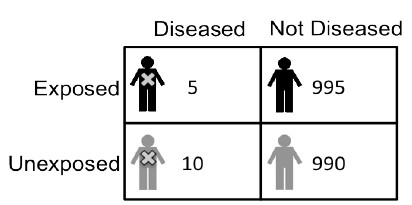
### 

### Interaction across scales, example

**Question:** Is consumption of green tea associated with reduced risk of stomach cancer? Does the relationship vary by whether individuals have diets that are rich in cured food? Purposive sample of 4000 individuals without stomach cancer:

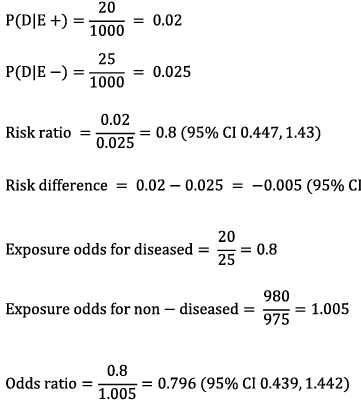
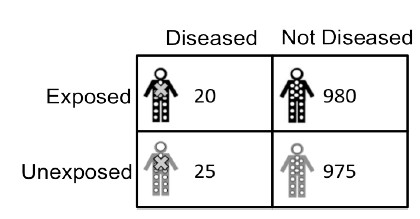
* + - * 1000 drink green tea and do not eat cured foods
      * 1000 drink green tea and eat cured foods
      * 1000 do not drink green tea but eat cured foods
      * 1000 do not drink green tea and eat cured foods

All follow forward for twenty years to determine which individuals develop stomach cancer.



### Green tea and cancer - no smoked/cured food

**Interpretation:** Among those who do not eat smoked/cured foods, green tea consumption is associated with 0.5 times the odds of stomach cancer compared with those who do not consume green tea.

### Green tea and cancer - no smoked/cured food

**Interpretation:** Among those who consume smoked/cured foods, green tea consumption is associated with 0.8 times the odds of stomach cancer compared with those who do not consume green tea.

**Based on the risk ratio and the odds ratio**, green tea consumption has a stronger protective effect among those who do not consume smoked/cured meats than among those who do consume such food. Therefore, there is **evidence of interaction** between green tea and smoked/cured foods.

However, **risk differences** across the two strata indicate that green tea consumption is associated with 5 fewer cases of stomach cancer for every 1,000 individuals who consume green tea, regardless of whether an individual consumes smoked/cured foods or not, i.e. there is **no evidence of interaction** between green tea and smoked/cured foods.

### Interaction is dependent on whether we use relative measures or difference measure.

### Additivity, multiplicativity and interaction

**Additive:** if two exposures do not interact, the risk of disease among exposed to both exposures = **sum** of risk of disease given exposure to one factor + risk of disease given exposure to the other factor.

**Multiplicative:** If two exposures do not interact, the risk of disease among those exposed to both = **product** of risk of disease given exposure to one factor \* risk of disease given exposure to the other factor.

### Interaction is scale dependent, example A

Risk among those exposed to both X and Y: 10% Risk among those exposed to X but not Y: 6% Risk among those exposed to Y but not X: 5% Risk among those exposed to neither X nor Y: 1%

There is **evidence of additive interaction.** The risk of disease among those exposed to both X and Y is = sum of the risk associated with exposure to X alone, plus Y alone, minus the exposure associated with neither exposure (10=6+5-1).

This is **no evidence of multiplicative interaction.** The risk of disease among those exposed to both X and Y to be 30% if there were no multiplicative interaction, because 6 x 5=30 - observed risk is 10% < 30%.

### Interaction is scale dependent, example B

Risk among those exposed to both X and Y: 30% Risk among those exposed to X but not Y: 6%

Risk among those exposed to Y but not X: 5%

Risk among those exposed to neither X nor Y: 1%

There is **evidence of multiplicative interaction.** The risk of disease among those exposed to both X and Y = to product of the risk associated with exposure to X alone, times Y alone (30=6\*5).

There is no **evidence of additive interaction.** 30% is greater than the sum of risks for those exposed to X but not Y (6%) and Y but not X (5%) (Minus the risk among those exposed to neither, 1%).

### Random variation

The appearance of interaction can arise due to chance in sampling process. We may collect a sample in which there were, by chance, a large proportion of individuals with disease in a certain sub-group.

NB: Therefore confidence intervals around interaction measures are important.

### Interaction – quiz

### Check your understanding of this topic by filling in the missing words in the sentences using the words in the box below. You will find the numbered answers underneath.

1. Interaction occurs when two causes are both components of the same \_\_\_\_\_\_\_\_\_\_;
2. When two causes interact this means that at least some individuals become diseased through a certain sufficient cause that includes \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_;
3. We can observe interaction when \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ varies across levels of third variable;
4. Different measures of association will evidence difference variation over a third variable depending on the \_\_\_\_\_\_\_\_ (additive or multiplicative);
5. In epidemiology we are principally concerned with \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

|  |  |
| --- | --- |
| A | both component causes |
| B | additive interaction |
| C | sufficient cause |
| D | scale |
| E | measure of association for exposure and outcome |

|  |
| --- |
| 1 = C; 2 = A; 3 = E; 4 = D; 5 = B |