BASIC CONCEPTS OF EPIDEMIOLOGY

Tellervo KORHONEN, PhD
Adjunct Professor in Public Health Science
University Lecturer
Hjelt Institute

TRANSMED METABOLIC SYNDROME 2013:
SECTION 2: EPIDEMIOLOGY
TEACHING SESSIONS

1. BASIC CONCEPTS OF EPIDEMIOLOGY I
   ■ 18.2.2013

2. BASIC CONCEPTS OF EPIDEMIOLOGY II
   ■ 20.2.2013

3. BRIEFING FOR THE SEMINAR
   ■ 20.2. 2013

4. SEMINAR:
   Student presentations on epidemiological papers
   ■ 11.3.2013
WHAT IS EPIDEMIOLOGY?

Epidemiology is the study of:

- health/disease-related events
- health/disease-related characteristics
- disease risk factors in a certain population
WHY DO WE NEED EPIDEMIOLOGY?

1. Identifying risk factors or preventive factors
2. Helping to make good diagnosis criteria
3. Helping to make policy decisions
4. Guiding how to screen for diseases
5. Conducting population level evaluation of treatments
AREAS OF EPIDEMIOLOGY

- Pharmacological
- Genetic (Lecture by Docent Silventoinen)
- Infection
- Cardiovascular disease (CVD) (Lecture by Prof. Salomaa)
- Diabetes
- Psychiatric
- Nutritional
- Cancer
- Environmental
- Occupational
- Methodological
- Etc...
HISTORY OF EPIDEMIOLOGY (1)

John Snow
(1813-1858)

- “father of epidemiology”
- water pollution → cholera epidemic

http://www.ph.ucla.edu/epi/snow/fatherofepidemiology.html
Sir Austin Bradford-Hill
(1897 – 1991)

He defined nine criteria of causation between an environmental factor and a disease.

Bradford-Hill Criteria / (1)
Strength of Association

- The stronger the association is, the more likely it is that the relation of "A" to "B" is causal.

- Percivall Pott (1775): scrotal cancer (kivespussisyöpä) 200 times more common among chimney sweepers than among other occupations

- Association of cigarette smoking and lung cancer: 9-10 times

- How strong is strong enough?
The association is consistent, if

1. similar results are replicated in studies in different settings using different methods

OR

2. several populations and different study designs show similar results.
Bradford-Hill criteria / (3)
Specificity (??)

- This criterion is fulfilled, when a single putative cause produces a specific effect.

- However, absence of specificity does not negate a causal relationship, if other criteria are fulfilled.
Bradford-Hill criteria / (4)
Temporality

If risk factor "A" is believed to cause a disease, then it is clear that this factor "A" must necessarily always precede the occurrence of the disease.

- Obesity → Diabetes
  NOT
- Diabetes → Obesity
Dose-Response Relationship:
An increasing amount of exposure increases the risk of disease

However, as with specificity, the absence of a dose-response relationship does not rule out a causal relationship.
Dose-Response Relationship: Retrospective case-control vs. prospective cohort study

Relation between amount smoked by continuing smokers and mortality from lung cancer obtained in case-control and cohort studies: mortality in men aged 45-74 years, standardised for age, expressed as percentage of unweighted average in different smoking categories and amount smoked per day in grams of tobacco (all methods of smoking combined)
Bradford-Hill Criteria / (6)
Biological Plausibility

- The association agrees with currently accepted understanding of pathological processes.

- However, research that disagrees with established theory is not necessarily false; it may, in fact, force a reconsideration of accepted beliefs and principles.
The association should be compatible with existing theory and knowledge.

It is necessary to evaluate claims of causality within the context of the current state of knowledge within a given field and in related fields.
Bradford-Hill Criteria / (8)  
Experiment

- The disease can be altered (prevented or cured) by an appropriate experimental regimen.

Example:  
If a non-experimental study identifies that diabetes can be prevented by avoiding excess overweight, causal association can be proved in an experiment, where a weight control intervention is compared to a control condition (no intervention).
In judging whether a reported association is causal, it is necessary to determine the extent to which researchers have taken other possible explanations into account and have effectively ruled out such alternate explanations.

In other words, it is always necessary to consider multiple hypotheses before making conclusions about the causal relationship between any two items under investigation.
Measuring diseases

In epidemiology we are usually interested in:

- **incidence** of disease in a population
- **prevalence** of disease in a population
"Incidence" is the probability of a new event (disease) in a certain period of time:

\[
\text{Incidence} = \frac{\text{number of new cases in a period of time}}{\text{length of the follow-up time}}
\]
INCIDENCE: Example

- New cases in certain period of time divided by length of follow-up.
- Let us follow-up 8 persons
- What is incidence in period \([1.5 - 5.5]\), length 4a (a=year)
- \# of new cases = 3
- Calculate follow-up time for each person:
  \[(5*4) + (1*0.5) + (1*2.5) + (1*3.5) = 26.5 \text{ a (years)}\]
- Incidence = \[3 / 26.5 = 0.11/\text{a}\]
PREVALENCE = Proportion of disease in population in a certain time point

- Let us assume a chronic disease, and no deaths during the follow-up

- Prevalence in different time points (e.g. years):
  - 3.0: $1/8=0.125=12.5\%$
  - 5.5: $3/8=0.375=37.5\%$
  - 7.5: $4/8=0.500=50.0\%$
Prevalences of Substance (Ab)use

ESPAD International Survey for 15-16-aged adolescents in 2007
Association between risk and disease
**RISK FACTOR**

**Risk factor** is a factor which increases the likelihood to have the disease under investigation when compared to those who are not exposed to that factor.

For example, smoking is a risk factor of lung cancer, because smokers have a higher likelihood to get cancer compared to non-smokers.

**Protective factor** is the opposite of risk factor: it decreases the likelihood to have the disease under investigation.
OUTCOME

- OUTCOME is used in longitudinal investigation
- DISEASE under investigation
- EVENT

- RISK FACTOR → OUTCOME / DISEASE / EVENT

- Dependent variable in the analysis
- Variable to be explained in the analysis
- Used in cross-sectional analysis where we cannot be sure which was first and which after that
TYPES OF ASSOCIATION

- **Cross-sectional**
  All measurements - both 'risk factor' and 'event' have been conducted at the same time point. Hence, only associations between them can be assessed, but no conclusions about causal associations can be done.

- **Longitudinal**
  - Here, the measurements have been done at 2 or more time points.
  - There are 2 types of longitudinal associations:
    - **Prospective**: Risk factor → Follow-up → Event
    - **Retrospective**: Event + restrospectively investigated exposure to risk factor
ESTIMATES OF ASSOCIATION (1)

- Correlation Coefficient (r)
  - Range: 0 → 1
  - Can be positive or negative
  - Reflects the association between two continuous variables
    - Let’s assume that correlation between Weight (kg) and Height (cm) $r=0.75 \rightarrow$ are highly positively correlated
    - Let’s assume that correlation between number of cigarettes smoked per day and Weight (kg) $r=-0.43 \rightarrow$ are moderately negatively correlated

- NOTE: Correlation alone does not prove causal relation between 2 variables!
Positive and Negative Correlation Between Variables $X$ and $Y$
ESTIMATES OF ASSOCIATION (2)

- Odds Ratio (OR) = Estimate of association between two dichotomous (0 vs. 1) variables

- Outcome: 0 vs. 1
  - Disease 0=no disease vs. 1=disease

- OR reflects the effect size of the risk for '1' to 'happen'

- OR = 1,00 (reference category of the exposure)

- OR=1,6 = for smokers the disease risk is elevated by 60%, compared to never smokers (OR=1,00)

- OR=0,8 (the risk for disease is decreased by 20%)
ESTIMATES OF ASSOCIATION (3)

- \( \beta \) - coefficient
  - Outcome has to be a continuous variable
  - Regression analysis
<table>
<thead>
<tr>
<th>Model 1&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Model 2&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Model 3&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Final Model&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Model 1&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Model 2&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Model 3&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Final Model&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta )</td>
<td>( \beta )</td>
<td>( \beta )</td>
<td>( \beta )</td>
<td>( \beta )</td>
<td>( \beta )</td>
<td>( \beta )</td>
<td>( \beta )</td>
</tr>
<tr>
<td>Occasional smokers</td>
<td>-3.53***</td>
<td>-3.43***</td>
<td>-3.25***</td>
<td>-2.36***</td>
<td>-4.61***</td>
<td>-4.34***</td>
<td>-4.21***</td>
</tr>
<tr>
<td></td>
<td>(-4.62, -2.44)</td>
<td>(-4.46, -2.41)</td>
<td>(-4.34, -2.21)</td>
<td>(-3.47, -1.26)</td>
<td>(-5.99, -3.23)</td>
<td>(-5.65, -3.04)</td>
<td>(-5.57, -2.87)</td>
</tr>
<tr>
<td>Recent quitters&lt;sup&gt;1&lt;/sup&gt;</td>
<td>-3.35***</td>
<td>-3.76***</td>
<td>-3.87***</td>
<td>-2.88***</td>
<td>-5.27***</td>
<td>-5.93***</td>
<td>-5.53***</td>
</tr>
<tr>
<td></td>
<td>(-5.02, -1.68)</td>
<td>(-5.33, -2.19)</td>
<td>(-5.54, -2.21)</td>
<td>(-4.55, -1.20)</td>
<td>(-7.25, -3.31)</td>
<td>(-7.79, -4.07)</td>
<td>(-7.46, -3.59)</td>
</tr>
<tr>
<td>Former smokers&lt;sup&gt;2&lt;/sup&gt;</td>
<td>-1.11**</td>
<td>-1.47***</td>
<td>-1.51***</td>
<td>-1.17**</td>
<td>-2.41***</td>
<td>-2.51***</td>
<td>-2.55***</td>
</tr>
<tr>
<td></td>
<td>(-1.90, -0.31)</td>
<td>(-2.22, -0.72)</td>
<td>(-2.33, -0.69)</td>
<td>(-1.98, -0.36)</td>
<td>(-3.43, -1.39)</td>
<td>(-3.47, -1.55)</td>
<td>(-3.56, -1.53)</td>
</tr>
<tr>
<td>Observations</td>
<td>899</td>
<td>899</td>
<td>847</td>
<td>847</td>
<td>715</td>
<td>715</td>
<td>683</td>
</tr>
</tbody>
</table>

*** \( p<0.001 \), ** \( p<0.01 \), * \( p<0.05 \)

<sup>1</sup>quit ≤ 6 months ago, <sup>2</sup>quit > 6 months ago
<sup>a</sup>Model 1 adjusted for age
<sup>b</sup>Model 2 adjusted for age and BMI
<sup>c</sup>Model 3 adjusted for age, BMI, education, marital status, physical activity, alcohol use fat and fiber intake
<sup>d</sup>Final model adjusted for age, BMI, education, marital status, physical exercise, alcohol consumption, fat and fiber intake, and nicotine addiction (FTND >=4)

Note: \( \beta \) coefficients with 95% CIs are calculated by using daily smokers as the reference group
COMPARING RISKS (1)

- **Risk to have disease**
  - With exposure E: \( P(D | 'E') \)
  - Without exposure E: \( P(D | 'no E') \)

- **“Relative Risk” = “Risk Ratio” = “RR”**
  - \( RR = \frac{P(D | 'E')} {P(D | 'no E')} \)

- **“Risk difference” = ”Excess Risk” = ER**
  - \( ER = P(D | 'E') - P(D | 'no E') \)

**NOTE:**

E = exposure to a risk factor, 
e.g. tobacco smoke, overweight, etc..
COMPARING RISKS (2)

Risk is usually measured with **cumulative incidence**

Cumulative incidence =

\[
\frac{\text{Number of new cases during the follow-up}}{\text{Size of population in the beginning of the follow-up}}
\]
**RISK RATIO (RR) and EXCESS RISK (ER)**

Let's divide our population \(n=8\) into 2 sub-populations: A (5 men) and B (3 women).

- **Risk (cumulative incidence) in A:** \(\frac{3}{5}\)
- **Risk (cumulative incidence) in B:** \(\frac{1}{3}\)

- **RR A vs. B:** \(\frac{3/5}{1/3} = 1.80\)
- **ER A vs. B:** \(\frac{3}{5} - \frac{1}{3} = 0.27\)

● = event = ‘disease’
INCI DENCE RATE RATIO (IRR)

IRR also called as 'rate ratio' or 'incidence ratio'

Ratio of 2 incidences

Let's assume that Incidence (I)

in A is $I_A$

and

in B is $I_B$

Then:

$$IRR = \frac{I_A}{I_B}$$
**IRR - EXAMPLE**

- **Populations A & B**
  - $I_A = \frac{3}{8+8+4+5+7} = 0.09375$
  - $I_B = \frac{1}{8+8+2} = 0.0556$

- **IRR A vs. B =**
  - $0.09375 / 0.0556 = 1.68$
Exercise / Solution

- Prevalence in the beginning of year 6
  \[ \frac{3}{10} = 0.30 = 30\% \]

- Incidence in period \([2,6]\)
  \[ \frac{2}{(7*4+1+3)} = \frac{2}{32} = 0.0625 \]
STATISTICAL SIGNIFICANCE
**P - VALUE - WHAT DOES IT TELL US?**

- P-value is a probability that the 'correct result' is as unlikely
- \( p < 0.001 \) ( *** ) \( \rightarrow \) The association is **STATISTICALLY VERY SIGNIFICANT**
- \( p < 0.01 \) ( ** ) \( \rightarrow \) The association is **STATISTICALLY MODERATELY SIGNIFICANT**
- \( p < 0.05 \) ( * ) \( \rightarrow \) The association is **STATISTICALLY SIGNIFICANT**

**NOTE:** P-value is related to testing the Ho-hypothesis (sample 1 = sample 2) and it does not tell us anything about the size of the association.

**P-VALUE REFLECTS ONLY STATISTICAL SIGNIFICANCE OF THE ASSOCIATION - NOT STRENGTH OF CLINICAL SIGNIFICANCE!!**
**H₀ - HYPOTHESIS:** Risk (A) = Risk (B)

**Accept? / Reject?**

**H₁ - HYPOTHESIS:**

- Risk (A) ≠ Risk (B)
- Risk (A) < Risk (B)
- Risk (A) > Risk (B)
CONFIDENCE INTERVAL

The 95% Confidence Interval (CI) is an interval where the unknown ‘real’ parameter is located with 95% certainty.

If we could repeat the same study several times, with same sample size and same conditions, there would be 5% risk that the result would be outside of that interval.
Confounding factor (‘Confounder’) is another risk factor of the disease under investigation. In addition, it is associated also with the risk factor under investigation.
Why is confounding a problem?

Confounding factors may be harmful, because many risk factors tend to cluster within the same individuals or population subgroups.

We know, that smoking, heavy alcohol use, and sedentary lifestyle are highly correlated.

Thus, all these factors can confound the effects of each other when investigating, for example, risk factors of cardiovascular disease incidence.
Sekoittavien tekijöiden vaikutus voidaan demonstroida näyttämällä useita malleja riskitekijän yhteydestä tutkittuun sairauteen. Riskiä kuvaava lukuarvo voidaan näyttää esimerkiksi kolmesta mallista: 1) kun vain ikä on vakioitu, 2) kun voimakkaimmat sekoittavat tekijät on vakioitu, 3) kun kaikki mitatut sekoittavat tekijät on vakioitu. Jos tutkitun riskitekijän riskisuhteet ovat samat kaikissa kolmessa mallissa, voidaan päätellä, että tutkitun riskitekijän vaikutus on riippumaton sekoittavista tekijöistä.

How can we test confounding?
<table>
<thead>
<tr>
<th>Smoking 1975-1981</th>
<th>Model 1 (n=4975)</th>
<th>Model 2 (n=4939)</th>
<th>Model 3 (n=4164)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never smokers</td>
<td><strong>1.0</strong> 95% CI:</td>
<td><strong>1.0</strong> 95% CI:</td>
<td><strong>1.0</strong> 95% CI:</td>
</tr>
<tr>
<td>Long term former smokers</td>
<td><strong>1.0</strong> 95% CI:</td>
<td><strong>1.0</strong> 95% CI:</td>
<td><strong>0.9</strong> 95% CI:</td>
</tr>
<tr>
<td>Short term former smokers</td>
<td><strong>1.9</strong> 95% CI:</td>
<td><strong>1.7</strong> 95% CI:</td>
<td><strong>1.7</strong> 95% CI:</td>
</tr>
<tr>
<td>Occasional smokers</td>
<td><strong>1.5</strong> 95% CI:</td>
<td><strong>1.4</strong> 95% CI:</td>
<td><strong>1.2</strong> 95% CI:</td>
</tr>
<tr>
<td>Recurrent smokers</td>
<td><strong>1.9</strong> 95% CI:</td>
<td><strong>1.5</strong> 95% CI:</td>
<td><strong>1.2</strong> 95% CI:</td>
</tr>
<tr>
<td>Persistent smokers</td>
<td><strong>1.9</strong> 95% CI:</td>
<td><strong>1.5</strong> 95% CI:</td>
<td><strong>1.4</strong> 95% CI:</td>
</tr>
</tbody>
</table>

Model 1 = Adjusted for age, marital status, social class (Korhonen et al. 2007)
Model 2 = + alcohol use, physical activity
Model 3 = + somatic health, social network, emotional support, life events, neuroticism and life satisfaction