BASIC CONCEPTS OF EPIDEMIOLOGY

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

TRANSMED METABOLIC SYNDROME 2014
SECTION 2: EPIDEMIOLOGY
AIMS OF THIS SECTION ARE TO...

1. Understand the basic concepts of epidemiology

2. Know the essential epidemiological study designs and their strengths and weaknesses

3. Apply this knowledge so that is able to give a short presentation on one scientific epidemiological article
1. BASIC CONCEPTS OF EPIDEMIOLOGY I
   ▷ 18.3.2014 (Tellervo Korhonen)

2. BASIC CONCEPTS OF EPIDEMIOLOGY II
   BRIEFING FOR THE SEMINAR
   ▷ 21.3. 2014 (Korhonen)

3. EPIDEMIOLOGY OF CVD
   ▷ 1.4. 2014 (Veikko Salomaa)

4. GENETIC EPIDEMIOLOGY
   ▷ 2.4. 2014 (Karri Silventoinen)

5. SEMINAR:
   Student presentations on epidemiological papers
   ▷ 4.4.2014 (Korhonen)
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...
Sir Austin Bradford-Hill (1897 – 1991)

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

1. Bradford-Hill Criterion: **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?
2. Bradford-Hill criterion: **Consistency**

- 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.
3. Bradford-Hill criterion: **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!

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
5. Bradford-Hill criterion: 
**Dose-Response Relationship**

- **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\textsuperscript{12} and cohort\textsuperscript{11} 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)
6. Bradford-Hill Criterion: 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.
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
Measuring diseases

In epidemiology we are usually interested in:

- **Incidence** of disease in a population
- **Prevalence** of disease in a population
INCIDENCE

“Incidence” is the probability of a new event (disease) in a certain period of time:

Incidence =

\[
\text{number of new cases in a period of time} \div \text{length of the follow-up time}
\]
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:
  \[(5\times4) + (1\times0.5) + (1\times2.5) + (1\times3.5) = 26.5 \text{ a (years)}\]
- Incidence = 3 / 26.5 = 0.11/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
- population n=8

Prevalence in different time points:
- time 3.0: \( \frac{1}{8} = 0.125 = 12.5\% \)
- time 5.5: \( \frac{3}{8} = 0.375 = 37.5\% \)
- time 7.5: \( \frac{4}{8} = 0.500 = 50.0\% \)
Association between risk and disease
RISK FACTOR

PROTECTIVE 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 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 $\rightarrow$ 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 + retrospectively investigated exposure to risk factor
ESTIMATES OF ASSOCIATION (1)

- Correlation Coefficient ($r$)
  - Range: $0 \rightarrow 1$
  - 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
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"
  - $RR = P(D|'E') / P(D|'no E')$

- "Risk difference" = "Excess Risk" = ER
  - $ER = P(D|'E') - P(D|'ei E')$

**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: $\frac{1}{3}$

- \textbf{RR A vs. B:} $(\frac{3}{5})/(\frac{1}{3}) = 1.80$
- \textbf{ER A vs. B:} $(\frac{3}{5})-(\frac{1}{3}) = 0.27$
STATISTICAL SIGNIFICANCE
**P-VALUE**

- **P-value** is related to testing the Ho-hypothesis (e.g. Mean of sample 1 = Mean of sample 2)
- \( p < 0.001 \) (***) → The association is **STATISTICALLY VERY SIGNIFICANT**
- \( p < 0.01 \) (**) → The association is **STATISTICALLY MODERATELY SIGNIFICANT**
- \( p < 0.05 \) (*) → The association is **STATISTICALLY SIGNIFICANT**

**NOTE:** **P-VALUE REFLECTS ONLY STATISTICAL SIGNIFICANCE OF THE ASSOCIATION - NOT STRENGTH OF CLINICAL SIGNIFICANCE!!**
<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>INTERVENTION GROUP (N=256)</th>
<th>CONTROL GROUP (N=250)</th>
<th>P VALUE†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean ±SD 95% CI</td>
<td>mean ±SD 95% CI</td>
<td></td>
</tr>
<tr>
<td>Change in weight</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In kilograms</td>
<td>−4.2±5.1 −4.8 to −3.6</td>
<td>−0.8±3.7 −1.3 to −0.3</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Percent change</td>
<td>−4.7±5.4 −5.0 to −4.4</td>
<td>−0.9±4.2 −1.0 to −0.8</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Change in waist circumference (cm)</td>
<td>−4.4±5.2 −5.1 to −3.9</td>
<td>−1.3±4.8 −1.9 to −0.7</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Change in plasma glucose (mg/dl)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fasting</td>
<td>−4±12 −6 to −2</td>
<td>1±12 0 to 2</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>2 Hr after oral glucose challenge</td>
<td>−15±34 −19 to −11</td>
<td>−5±40 −8 to −2</td>
<td>0.003</td>
</tr>
<tr>
<td>Change in serum insulin (µg/ml)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fasting</td>
<td>−2±9 −3 to −1</td>
<td>−1±7 −2 to 0</td>
<td>0.14</td>
</tr>
<tr>
<td>2 Hr after oral glucose challenge</td>
<td>−29±64 −37 to −21</td>
<td>−11±51 −18 to −4</td>
<td>0.001</td>
</tr>
<tr>
<td>Change in serum lipids (mg/dl)‡</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total cholesterol</td>
<td>−5±28 −8 to −2</td>
<td>−4±28 −7 to −1</td>
<td>0.62</td>
</tr>
<tr>
<td>High-density lipoprotein cholesterol</td>
<td>2±7 1 to 3</td>
<td>1±6 0 to 2</td>
<td>0.06</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>−18±51 −24 to −12</td>
<td>−1±60 −8 to 6</td>
<td>0.001</td>
</tr>
<tr>
<td>Change in blood pressure (mm Hg)§</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Systolic</td>
<td>−5±14 −7 to −3</td>
<td>−1±15 −3 to −1</td>
<td>0.007</td>
</tr>
<tr>
<td>Diastolic</td>
<td>−5±9 −6 to −4</td>
<td>−3±9 −4 to −2</td>
<td>0.02</td>
</tr>
</tbody>
</table>
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.
<table>
<thead>
<tr>
<th>VARIABLE</th>
<th><strong>INTERVENTION GROUP</strong></th>
<th></th>
<th><strong>CONTROL GROUP</strong></th>
<th></th>
<th><strong>P VALUE†</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(N=256)</td>
<td>mean ±SD 95% CI</td>
<td>(N=250)</td>
<td>mean ±SD 95% CI</td>
<td></td>
</tr>
<tr>
<td>Change in weight</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In kilograms</td>
<td>−4.2±5.1</td>
<td>−4.8 to −3.6</td>
<td>−0.8±3.7</td>
<td>−1.3 to −0.3</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Percent change</td>
<td>−4.7±5.4</td>
<td>−5.0 to −4.4</td>
<td>−0.9±4.2</td>
<td>−1.0 to −0.8</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Change in waist circumference (cm)</td>
<td>−4.4±5.2</td>
<td>−5.1 to −3.9</td>
<td>−1.3±4.8</td>
<td>−1.9 to −0.7</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Change in plasma glucose (mg/dl)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fasting</td>
<td>−4±12</td>
<td>−6 to −2</td>
<td>1±12</td>
<td>0 to 2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>2 Hr after oral glucose challenge</td>
<td>−15±34</td>
<td>−19 to −11</td>
<td>−5±40</td>
<td>−8 to −2</td>
<td>0.003</td>
</tr>
<tr>
<td>Change in serum insulin (μg/ml)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fasting</td>
<td>−2±9</td>
<td>−3 to −1</td>
<td>−1±7</td>
<td>−2 to 0</td>
<td>0.14</td>
</tr>
<tr>
<td>2 Hr after oral glucose challenge</td>
<td>−29±64</td>
<td>−37 to −21</td>
<td>−11±51</td>
<td>−18 to −4</td>
<td>0.001</td>
</tr>
<tr>
<td>Change in serum lipids (mg/dl)‡</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total cholesterol</td>
<td>−5±28</td>
<td>−8 to −2</td>
<td>−4±28</td>
<td>−7 to −1</td>
<td>0.62</td>
</tr>
<tr>
<td>High-density lipoprotein cholesterol</td>
<td>2±7</td>
<td>1 to 3</td>
<td>1±6</td>
<td>0 to 2</td>
<td>0.06</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>−18±51</td>
<td>−24 to −12</td>
<td>−1±60</td>
<td>−8 to 6</td>
<td>0.001</td>
</tr>
<tr>
<td>Change in blood pressure (mm Hg)§</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Systolic</td>
<td>−5±14</td>
<td>−7 to −3</td>
<td>−1±15</td>
<td>−3 to −1</td>
<td>0.007</td>
</tr>
<tr>
<td>Diastolic</td>
<td>−5±9</td>
<td>−6 to −4</td>
<td>−3±9</td>
<td>−4 to −2</td>
<td>0.02</td>
</tr>
</tbody>
</table>
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.
How can we test confounding?

Compare the risk estimates (e.g. OR) in different models

1) Basic model (sex, age adjusted)
2) Adjusted for most well known confounders
3) Adjusted for all potential confounders

→ If the estimate of the risk factor remains in all models, the association of that factor is independent of the other factors, i.e. robust association

→ If the estimate of the risk factor becomes non-significant, the association is confounded by those factors which were added to the model
SMOKING & DEPRESSION:  
Multiple logistic regression models of 1975-1981 smoking status  
as a predictor of depression in 1990 among Finnish men

<table>
<thead>
<tr>
<th>Smoking 1975-1981</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n=4975)</td>
<td>(n=4939)</td>
<td>(n=4164)</td>
<td></td>
</tr>
<tr>
<td>Never smokers</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>OR 95% CI</td>
<td>0.8 - 1.4</td>
<td>0.7 - 1.3</td>
<td>0.6 - 1.3</td>
</tr>
<tr>
<td>Long term former smokers</td>
<td>1.0</td>
<td>1.0</td>
<td>0.9</td>
</tr>
<tr>
<td>OR 95% CI</td>
<td>0.8 - 1.4</td>
<td>0.7 - 1.3</td>
<td>0.6 - 1.3</td>
</tr>
<tr>
<td>Short term former smokers</td>
<td>1.9</td>
<td>1.7</td>
<td>1.7</td>
</tr>
<tr>
<td>OR 95% CI</td>
<td>1.4 - 2.5</td>
<td>1.2 - 2.3</td>
<td>1.2 - 2.4</td>
</tr>
<tr>
<td>Occasional smokers</td>
<td>1.5</td>
<td>1.4</td>
<td>1.2</td>
</tr>
<tr>
<td>OR 95% CI</td>
<td>1.1 - 2.0</td>
<td>1.0 - 1.9</td>
<td>0.8 - 1.8</td>
</tr>
<tr>
<td>Recurrent smokers</td>
<td>1.9</td>
<td>1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>OR 95% CI</td>
<td>1.2 - 2.9</td>
<td>1.0 - 2.4</td>
<td>0.7 - 2.1</td>
</tr>
<tr>
<td>Persistent smokers</td>
<td>1.9</td>
<td>1.5</td>
<td>1.4</td>
</tr>
<tr>
<td>OR 95% CI</td>
<td>1.5 - 2.3</td>
<td>1.2 - 1.9</td>
<td>1.1 - 1.9</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
<table>
<thead>
<tr>
<th></th>
<th>CROSS-SECTIONAL</th>
<th>LONGITUDINAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NON-EXPERIMENTAL</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Correlational Study</td>
<td>PROSPECTIVE: Cohort Study</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RETROSPECTIVE: Case-Control</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Study</td>
</tr>
<tr>
<td><strong>EXPERIMENTAL</strong></td>
<td></td>
<td>Randomized Controlled Trial</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>
THANK YOU!

tellervo.korhonen@helsinki.fi