Study design

András Keszei

1st Budapest Clinical Epidemiology Course – organized jointly with the 15th Budapest Nephrology School
Elements of design

- Theoretical design
  - Research question
  - Occurrence relation

- Design of data collection
  - Time
  - Study population
  - Experimental or non-experimental
  - Choice of measurement

- Design of analysis
Theoretical design

Does five-day treatment with penicillin in children with acute otitis media reduce the duration of symptoms?

- Outcome
- Determinants
- Domain

Occurrence relation

Outcome  ←  Determinant(s)
Types of epidemiologic research

- Etiologic
  - Does exposure cause disease?
  - One or many extraneous determinants
- Diagnostic
  - Predict probability of disease from (non)clinical profile
- Prognostic
  - Predict course of disease
- Intervention prognostic
  - Predict course of disease given treatment and profile
Etiologic research

• Theoretical design
  – Does exposure cause disease?
  – One or many extraneous determinants
  – outcome as a function of determinant conditional on confounders
Etiologic research

- Excessive body weight
- Determinant
- Diabetes
- Outcome
- Age
- Extraneous Determinant
Etiologic research

• Theoretical design
  – Does exposure cause disease?
  – One or many extraneous determinants
  – outcome as a function of determinant conditional on confounders
  – longitudinal in nature
Etiologic research

• Design of data collection
  – Temporal relationship must be incorporated
  – Cohort study
  – Case-control study
  – Experimental or non-experimental
**Cross-sectional studies**: Exposure and disease are measured at the same moment among study participants.

Examples of national surveys:

- NHANES (National Health and Nutrition Examination Study)  
  USA

- MORGEN (Monitoren Risicofactoren Gezondheid en Ziekte in Nederland) Monitoring Risk factors for Health and Disease  
  The Netherlands

- Hungarostudy 2002  
  Hungary
Ecological studies: correlation between (frequency of) exposure and disease at population-level

Examples:

- Meat consumption and risk of colon cancer
  (Armstrong and Doll, 1975)

- Wine consumption and coronary heart disease mortality
  (St Leger et al., 1979)

- Helicobacter Pylori infection and gastric cancer mortality
  (Forman et al., 1990)
Ecological studies: correlation between (frequency of) exposure and disease at population-level

■ Advantages:
  - Costs very low
  - Quick results, short study duration
    data from population statistics and other registries

■ Disadvantages:
  - Data at population level: are the exposed the ones that are diseased? ecological fallacy
  - Regional differences in diagnostic procedures, exposure measurement
  - Often lack of information on important variables (other exposures)
Cohort design

P = case; Outcome +/-
E = exposure +/-
I = incident cases

t_0: start of observation

end of observation

t_1: time
Cohort design

- Start with population not experiencing the outcome
- Risk of developing outcome should be present
- Exposure measurement at $t_0$
- Follow-up
- Determine incidence of outcome
## Risk of coronary heart disease in men aged 47-55 years

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Case</th>
<th>Cumulative incidence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never smoked</td>
<td>2212</td>
<td>88</td>
<td>4</td>
</tr>
<tr>
<td>Ex-smoker</td>
<td>1530</td>
<td>92</td>
<td>6</td>
</tr>
<tr>
<td>Current smoker</td>
<td>2806</td>
<td>281</td>
<td>10</td>
</tr>
</tbody>
</table>

11.8 year follow-up
# Incidence of myocardial infarction among vasectomized and nonvasectomized men

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Case</th>
<th>Person years</th>
<th>Incidence per 1000 person years</th>
</tr>
</thead>
<tbody>
<tr>
<td>35-44</td>
<td>14</td>
<td>16 806</td>
<td>0,8</td>
</tr>
<tr>
<td>45-54</td>
<td>24</td>
<td>8 133</td>
<td>3,0</td>
</tr>
<tr>
<td>55-64</td>
<td>7</td>
<td>1 700</td>
<td>4,1</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
<td>26 639</td>
<td>1,7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Case</th>
<th>Person years</th>
<th>Incidence per 1000 person years</th>
</tr>
</thead>
<tbody>
<tr>
<td>35-44</td>
<td>56</td>
<td>83 057</td>
<td>0,7</td>
</tr>
<tr>
<td>45-54</td>
<td>110</td>
<td>40 971</td>
<td>2,7</td>
</tr>
<tr>
<td>55-64</td>
<td>49</td>
<td>8 570</td>
<td>5,7</td>
</tr>
<tr>
<td>Total</td>
<td>215</td>
<td>132 598</td>
<td>1,6</td>
</tr>
</tbody>
</table>
Information is available from existing registries or other historical sources
Case-control design

“The sophisticated use and understanding of case-control studies is the most outstanding methodological development of modern epidemiology.”

Rothman, 1986

• Different terms used:
  – Case-referent study, TROHOC, Retrospective study

• Case-control studies have proven their potential value
  – Aspirin use and Reye syndrome  (Hurwitz ES. et al. 1987)
  – Diethylstilboestrol use by pregnant women and occurrence of clear-cell vaginal carcinoma in their daughters  (Herbst AL. et al. 1971)
Case-control design

• Efficiency
  – outcome of interest is rare
  – time between exposure and outcome is long
  – measurement of determinant is expensive, burdening or time-consuming

• Essence

Study *cases* and a *sample* of the *source population*
Case-control design

Source population

E+  
E-

Cases
E+  E-

Sample of the source population
Non-cases

E= exposure (+/-)

t_0  time
Case-control design

• Proper sampling of controls
  – to ensure that measures of association will be similar to what would be obtained from a cohort study
  – controls should be representative sample of the source population from which cases are drawn
Case-control design

<table>
<thead>
<tr>
<th>E+</th>
<th>$d_1$ cases in $T_1$ person time</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-</td>
<td>$d_2$ cases in $T_2$ person time</td>
</tr>
</tbody>
</table>

$RR = \frac{d_1/T_1}{d_2/T_2}$

↓ f sampling fraction

<table>
<thead>
<tr>
<th>e+</th>
<th>$t_1$ person time $\sim fT_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>e-</td>
<td>$t_2$ person time $\sim fT_2$</td>
</tr>
</tbody>
</table>

$RR_s = \frac{d_1/t_1}{d_2/t_2} \sim \frac{d_1/fT_1}{d_2/fT_2} = RR$

Sample of the source population
Swimming pool analogy

Pool - Source population
Lifeguard - Researcher

Net is used to catch those who become cases and to randomly sample other swimmers.

Grobbee DE
Control sampling - dynamic population

Case 1 Case 2 Case 3 Case 4
Control sampling - dynamic population

Sampling of controls
Control sampling - closed population

- Case
Case-cohort design

![Graph showing the number of publications on case-cohort studies from 1983 to 2008.](image-url)
Case-control design

• Retrospective
  – All data on outcome, determinants and other factors are available when study is initiated

• Prospective
  – Cases are identified and recruited until enough cases have been included
  – A control group is sampled during the same time period
Etiologic research

- Data collection
  - outcome
  - determinant
  - confounders

Does dietary nitrate cause bladder cancer? Zeegers MP, 2006
Data collection

• Outcome
  – Registry
  – Hospital or physicians’ charts
  – Morbidity Sentinel Programs

• Determinants
  – Questionnaire
  – Diary
  – Interview
  – Physical measurement
Instruments
1. non-invasive not harmful

2. non-invasive, but potentially harmful

3. invasive
Diagnostic research

• Theoretical design
  – Probability of disease as a function of multiple diagnostic tests
  – Domain
  – Goal
    • Optimal diagnostic strategy
    • Whether new diagnostic test provides added diagnostic value
N-terminal pro B-type natriuretic peptide in the detection of heart failure

<table>
<thead>
<tr>
<th></th>
<th>NT-proBNP positive (T+)</th>
<th>NT-proBNP negative (T-)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart failure (D+)</td>
<td>9</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Heart failure absent (D-)</td>
<td>69</td>
<td>55</td>
<td>124</td>
</tr>
<tr>
<td></td>
<td>78</td>
<td>55</td>
<td>133</td>
</tr>
</tbody>
</table>

Sensitivity $P(T^+|D^+)$=100%
Specificity $P(T^-|D^-)$=44%

Positive predictive value $P(D^+|T^+)$=12%
Negative predictive value $P(D^-|T^-)$=100%

Likelihood ratio of positive test $P(T^+|D^+)/P(T^+|D^-)$=1.8
Likelihood ratio of negative test $P(T^-|D^+)/P(T^-|D^-)$ =0
N-terminal pro B-type natriuretic peptide in the detection of heart failure

\[
P(D+) \cdot \text{sensitivity} \\
\frac{P(D+ | T+)}{P(D- | T+)} = \frac{P(D+)}{P(D-)} \cdot \text{LR}^+ \]
Posterior odds in sequential testing

\[
\frac{P(D^+|T_1^+, T_2^+, T_3^+)}{P(D^-|T_1^+, T_2^+, T_3^+)} = \frac{P(D^+)}{P(D^-)} \times LR(T_1^+|D^+) \times LR(T_2^+|D^+) \times LR(T_3^+|D^+)
\]

Assumes that the results of tests are independent of each other
Diagnostic research

• Design of data collection
  
  **Cross-sectional design**

<table>
<thead>
<tr>
<th>Population suspected of having target disease</th>
<th>D+</th>
<th>D-</th>
</tr>
</thead>
<tbody>
<tr>
<td>D+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Diagonsitic determinants are measured in **all** patients

**Case-control design**

<table>
<thead>
<tr>
<th>Population suspected of having target disease</th>
<th>D+</th>
<th>D-</th>
</tr>
</thead>
<tbody>
<tr>
<td>D+</td>
<td>cases</td>
<td></td>
</tr>
<tr>
<td>D-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Determinants measured on **subset** of controls
Diagnostic research

• Design of data collection
  – Case-control design

Attractive if measurement of determinant is time consuming or expensive
Diagnostic research

• Example

Does results from cardiovascular magnetic resonance imaging (CMR) provide additional diagnostic information for identifying heart failure (HF) in patients with stable COPD. (Rutten FH. 2005)

n=405

Population suspected of having target disease

37 patients with HF

Random sample of 41 patients without HF
Bias in diagnostic study

• Knowledge of results from diagnostic test may influence patient selection
  – Selective disease verification leads to bias

• Determinant influences outcome assessment
  – Incorporation bias, diagnostic review bias
Design of analysis diagnostic study

- Objective will guide analysis
  - Which determinant contributes to estimation of probability of disease
  - To what extent does it change the probability
  - To develop and validate a diagnostic model
Prognostic research

- Theoretical design
  - Outcome in the future as function of determinants measured at one (or more) time points before the outcome occurs.

- Domain
  - Individuals at risk of outcome

- Goal
  - Which determinant contributes to prediction
  - Does a marker provide *additional* predictive value
Prognostic research

• Design of data collection
  – longitudinal
  – cohort design Johansson, 1992 – prostate cancer
  – case-control design does not allow estimation of absolute risk
  – observational
  – prognostic study within a trial
Prognostic research

- Prognostic study within trial
  - Predictors of prolonged course in children with acute otitis media (Rovers MM. et al. 2007)
    - Control arm of randomized control trials
    - Outcome: fever and/or pain at 3-7 days
Prognostic research

• Bias in prognostic study
  – Determinant influences outcome assessment
  – Selective loss to follow-up
    • value of determinant is underestimated if poor prognosis is related to loss to follow-up
Intervention prognostic research

Effect of intervention

Natural history + Observer effect + Extraneous effects
Intervention prognostic research

- Comparability of natural history, extraneous effects and observer effects
  - Quasi-experiment
    - Comparison of two drugs in the treatment of leukemia
      - Select groups to be similar on age, proportion of males, severity, etc.
Intervention prognostic research

Intervention → Outcome

Reasons to initiate the intervention
## Confounding by indication

Comparison of death from cardiovascular causes in untreated and drug treated hypertensive women

<table>
<thead>
<tr>
<th>Rate ratio (95% Confidence interval)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude value</td>
<td>1.0 (0.6 to 1.5)</td>
</tr>
<tr>
<td>Adjusted for:</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.7 (0.4 to 1.1)</td>
</tr>
<tr>
<td>+ Body mass index, pulse rate</td>
<td>0.6 (0.4 to 1.0)</td>
</tr>
<tr>
<td>+ Smoking, lipid levels</td>
<td>0.6 (0.4 to 0.9)</td>
</tr>
<tr>
<td>+ Diabetes</td>
<td>0.5 (0.3 to 0.9)</td>
</tr>
</tbody>
</table>
Intervention prognostic research

- Randomization
  - Treatment is allocated at random
  - Differences in prognosis without treatment is result of random imbalances

Reasons to initiate the intervention
Intervention prognostic research

- Randomized controlled trial
- Cohort study in which allocation of exposure is made by the investigator
Randomized Controlled Trial

Base population
Study population
Measurement outcome
In-/ exclusion criteria
Informed consent
Sampling

Check prognostic factors
Check prognostic factors
Control intervention
Intervention
Measurement outcome
Measurement outcome

Comparable?
Contrast?
Effect?

time
Intervention prognostic research

• Factorial design
  – Two treatment contrasts are studied
  – Efficient
  – Study interactions

A randomized factorial trial of vitamins C and E and beta carotene in the secondary prevention of cardiovascular events in women: results from the Women's Antioxidant Cardiovascular Study. Cook NR et al., 2008
ADVANCE study

Randomization

- Perindopril-indapamide + Intensive glucose control (N=2500)
- Perindopril-indapamide + Standard glucose control (N=2500)
- Placebo + Intensive glucose control (N=2500)
- Placebo + Standard glucose control (N=2500)
Intervention prognostic research

• Crossover design
  – Comparison of effects within individual
  – Efficient
    • Between patient variability is not an issue
  – Limitations
    • Period effect
    • Carry-over effect

– Suitable for quick and reversible treatment effects
Intervention prognostic research

• Cluster randomized design
  – Randomization at group level, ex. GP
    • Minimal intervention strategy in the treatment of low back pain in general practice (Jellema P. et al. 2005)
Intervention prognostic research

• Participants
  – Generalizability of findings
    • Inclusion/exclusion criteria
• Comparability of extraneous and observer effects
  – placebo, blinding
Intervention prognostic research

• Analysis
  – categorical data:
    – RR reduction
    – absolute risk reduction, number needed to treat
  – continuous data
    – e.g t-test
  – survival data
    – relative hazard
  – “intention to treat” or “per protocol”
Key questions regarding the sample size

- What is the main purpose of the trial?
- What is the **principal** measure of patients **outcome**?
- How will the data be analyzed to detect a treatment difference? (The test statistic: **t-test**, **X²**)
- What type of results does one anticipate in the non-exposed group?
- **H₀** and **Hₐ**, How small a **difference** is it important to detect and with what degree of **certainty**? (δ, α and β.)
- How to deal with treatment withdraws and protocol violations.
## Sample size

<table>
<thead>
<tr>
<th>Incidence in non-exposed group</th>
<th>Prevalence of exposure in population %</th>
<th>RR</th>
<th>Sample size per group</th>
<th>Sample size per group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cohort</td>
<td>Case-control</td>
</tr>
<tr>
<td>1/1000</td>
<td>50</td>
<td>1.2</td>
<td>576732</td>
<td>2535</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td></td>
<td>31443</td>
<td>177</td>
</tr>
<tr>
<td></td>
<td>4.0</td>
<td></td>
<td>5815</td>
<td>48</td>
</tr>
<tr>
<td>1/100</td>
<td>50</td>
<td>1.2</td>
<td>57100</td>
<td>2535</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td></td>
<td>3100</td>
<td>177</td>
</tr>
<tr>
<td></td>
<td>4.0</td>
<td></td>
<td>567</td>
<td>48</td>
</tr>
<tr>
<td>1/10</td>
<td>50</td>
<td>1.2</td>
<td>5137</td>
<td>2535</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td></td>
<td>266</td>
<td>177</td>
</tr>
<tr>
<td></td>
<td>4.0</td>
<td></td>
<td>42</td>
<td>48</td>
</tr>
</tbody>
</table>

$\alpha=0.05; \beta=0.1$

J. Tonascia