Introduction to
• Descriptive studies
• Cross-sectional studies
• Aggregative studies
• Sampling and proportion of participation
• Selection bias

Objectives (1)
• Level 6: Evaluation (critique, appraise, judge)
• Level 5: Synthesize (integrate, collaborate)
• Level 4: Analyze (hypothesis, structure)
• Level 3: Application (utilize, produce)
• Level 2: Comprehension (translate, discuss)
• Level 1: Knowledge (define, enumerate, recall)

Objectives (2)
Follow and apply these six levels for
• Basic principles of epidemiological thinking
• Cross-sectional studies including
  - selection biases
  - proportion of participation
  - analysis of prevalence (PR)
  - confounding and interaction
• Knowledge, application of computer programs, synthesize findings, design and evaluate

Textbook, chapter 2, 4, and 5.
Textbook, chapter 6.
Problem Solving Cycle in Public Health (general and community-related problems)

Descriptive Studies (1)

- Descriptive studies examine the distribution of disease in a defined population.
- Based on existing mortality or morbidity statistics, such as hospital discharge data.
- Examine patterns of health outcome by age, gender or ethnicity, for specified time period or geographical areas.
- Estimates the incidence or prevalence of the disease:
  - Lifetime prevalence
  - Period prevalence
  - Point prevalence
Descriptive Studies (2)

- Do not formally evaluate the association between exposure and health outcome, although they can be helpful in assessing the possibility that an association exists.
- Descriptive data are used to examine patterns of health outcome by:
  - Place
  - Time
  - Person

Epidemiology:  (Stallones 1980, Ann Rev Public Health 1:69-82)

Axiom: Disease does not distribute randomly in human populations.

Corollary 1: Aggregations of human diseases are manifested along axes of:
  - time
  - space
  - individual / group characteristics.

Corollary 2: Variations in the frequency of human disease occur in response to:
  - variation in exposure
  - variation in susceptibility.

Conceptualization of Scientific approaches in Epidemiology

Collect Data

Make Inferences re: Operational Hypotheses

Make Inferences re: Conceptual Hypotheses

Conclusions and Interpretations

Empirical Findings

Analyse Data

Observations Data

Synthesise and Formulate Hunches

Theory/ Knowledge

Conceptual Hypotheses

Design Study

Operational Hypotheses
Hypotheses:

Requirement: **if - condition**
- (risk factor, exposure, predisposition etc.)

**then - condition**
- (disease, marker of a disease, precursor of a disease etc.)

Any information on additional conditions such as susceptible groups or calendar period or continent/nation is informative for the formulation of the hypothesis.

Formulation of hypotheses:

- **exposure** = risk or vector
- **predisposition / genotype** = host factor
- **confounders or competing risks** = environment

Any additional conditions such as ‘only in boys’, ‘only in Afro-Americans’ has to be taken into account in the design of the study.

Hierarchy of exposure indicators

- unspecific (proxy measures)
  - place of residence, occupation, social strata, ...
  - nutrition, working conditions, etc.
  - specific exposures: mercury, noise, ...
- specific
  - number of indicators
disease-specific

unspecific

cause of death

specific disease (well defined disease entities)

unspecific disorders

disease-unspecific changes: headache, increased levels of hormones or enzymes, ...

number of indicators

specific effects, long latency period

specific effects, prognosis?

global effects, latency period

disease-unspecific changes: headache, hormones or enzymes, ...

specific exposures: mercury, noise, ...

nutrition, working conditions, etc.

place of residence, occupation, social strata...

unspecific disorders

outcome of the disease (subsequent change in health status or death)

initiation of the pathologic process (disease starts)

clinical detection of the disease (onset of sign and symptoms)

initiation of the etiologic process (onset of the first cause)

primary prevention

secondary prevention

tertiary prevention
Interference of causal and prognostic associations

Cause → Markers of Exposure

Causal relation

Markers of Effect → Manifestation of the disease

Prognostic relation

Dose-Response Relations:

<table>
<thead>
<tr>
<th>Exposure Risk</th>
<th>Effect Outcome</th>
<th>Relation</th>
<th>(severity of the disease)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure Risk</td>
<td>Prevalence Incidence</td>
<td>Relation</td>
<td>(frequency of the disease)</td>
</tr>
<tr>
<td>Disease</td>
<td>Disease</td>
<td>“Amount” of Exposure</td>
<td>Relation</td>
</tr>
</tbody>
</table>

Criteria, which make causal inference easier:
(old medical paradigm for infectious diseases and accidents)

A factor is assumed to be causal:

- if the effect manifests itself in a limited risk period
- if the manifestation achieves the full extent of a clinical disorder.
Flow of knowledge and its application

Assumptions: knowledge about the appropriate time window of exposure & effect
randomized clinical or intervention trials
analytical observational studies
observational studies

Study design and corresponding analytical models (modified from Pearce, Environ. Epidemiology)

<table>
<thead>
<tr>
<th></th>
<th>Longitudinal study</th>
<th>X-sectional study</th>
<th>Case-control study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Numerator</strong></td>
<td>Cases</td>
<td>Cases</td>
<td>Cases</td>
</tr>
<tr>
<td><strong>Denominator</strong></td>
<td>P*Time</td>
<td>Persons</td>
<td>Persons</td>
</tr>
<tr>
<td><strong>Occurrence</strong></td>
<td>Rate</td>
<td>Risk</td>
<td>Proprions</td>
</tr>
<tr>
<td><strong>measure</strong></td>
<td>(incidence)</td>
<td>(incidence %)</td>
<td>Average values</td>
</tr>
<tr>
<td><strong>Effect</strong></td>
<td>Rate ratio (RR, IDR)</td>
<td>Risk ratio (RR)</td>
<td>Change of means over time</td>
</tr>
<tr>
<td><strong>measure</strong></td>
<td>Prevalence ratio (PR)</td>
<td>Diffferences of means</td>
<td></td>
</tr>
<tr>
<td><strong>Multivariate</strong></td>
<td>Posen regression, Survival analysis</td>
<td>Logistic regression</td>
<td>Logistic regression</td>
</tr>
<tr>
<td><strong>analysis</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Aggregative Study (ecological study)

- Population 1
- Population 2
- Population 3
- Population 4
- Population 5
- Population 6

Mean of a variable
Mean of a variable
Mean of a variable

Association?

Unit of observation: "a" = unit of analysis; "c" = ecologic fecoly

Figure 11.8. Example of an aggregate risk study: relationship between wine consumption and cancer mortality in developed countries. (Data from St. Leger RZ, "Health and Mortality: A Comparison of Alcohol Consumption and Cardiovascular Health in Developed Countries," American Journal of Public Health, vol. 97, no. 1, pp. 1-6, 2007.)
Classification of studies (individual-based and aggregative studies)

<table>
<thead>
<tr>
<th>Direction</th>
<th>longitudinal</th>
<th>historical f-up</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cross-sectional</td>
<td></td>
</tr>
<tr>
<td>backward</td>
<td>case-control</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prospective</td>
<td>retrospective</td>
</tr>
<tr>
<td></td>
<td>Risk or disease occur after the onset of the study.</td>
<td>Risk or disease occur before the onset of the study.</td>
</tr>
</tbody>
</table>

Hybrid studies

- nested follow-up (follow-up after a screening = survey follow-up study)
- nested case-control
- follow-up prevalence study (particular disease is not considered at the beginning of the study)
- backward prevalence study
- repeated surveys
- twin study (monozygotic vs. dizygotic: discordant vs. concordant pairs with respect to the disease)
- aggregative trend study (repeated aggregative surveys)
- ......

WOBURN, MASSACHUSETTS: 37 000 residents
8 municipal wells

EXPOSURE: 2 municipal wells contaminated: Trichloroethylene, tetra-chloroethylene, chloroform

DISEASE: childhood leukemia

TELEPHONE INTERVIEWS WITH THE PARENTS

RELATIVE RISK: 12 cases observed between 1969-79
5.3 cases expected

Relative Risk: 2.3
Lagakos et al. (1986), JASA 81
SELLAFIELD, ENGLAND:
EXPOSURE: Sellafield nuclear site
DISEASE: Childhood leukemia and Non-Hodgkin-Lymphoma (74 cases)
RELATIVE RISK in different zones around the nuclear site:
- 4 km: 1
- 5 - 9 km: 0.21
- 10 - 14 km: 0.17
- ≥ 15 km: < 0.16

Gardner et al. (1990), BMJ 300
See also:
JAMA 1985 Aug 2;254(5):621-6

Proportion of the population who reported that they found the air irritating versus the six-monthly average nitrogen dioxide concentration (Forsberg et al. 1997).

When is a study classified as aggregative (including aggregative cross-sectional or follow-up), when as a cross-sectional investigation?

Unit of analysis (uoa) - Unit of observation (uoo)

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Outcome</th>
<th>classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>uoa = uoo</td>
<td>uoa = uoo</td>
<td>cross-sectional</td>
</tr>
<tr>
<td>uoa ≠ uoo</td>
<td>uoa = uoo</td>
<td>cross-sectional, (exposure misclassification)</td>
</tr>
<tr>
<td>uoa ≠ uoo</td>
<td>uoa ≠ uoo</td>
<td>aggregative</td>
</tr>
</tbody>
</table>
Cross-sectional studies

• **Purpose**
  - This type of study is designed to estimate the prevalence of a disease in a population at a point in time or during a given period.
  - It can also be used to make association between a set of risk factors and the prevalence of a disease.

• **Conceptual design**
  - Data relating to the disease status and risk factors that may be associated with the disease are measured simultaneously.
  - Begins with a defined population

**Sampling strategies for cross-sectional studies**

![Diagram of sampling strategies for cross-sectional studies]

**Cross-sectional study** ⇒ analysis as cohort study

![Diagram of cross-sectional study flowchart]
Cross-sectional study \(\rightarrow\) analysis as case-control study

Begin of the study \(\downarrow\)

Measure/Classify and Compare

- Free of Disease/Outcome
  - Risk/Factor (+)
  - Risk/Factor (-)
- Have Disease/Outcome
  - Risk/Factor (+)
  - Risk/Factor (-)

Study population

Steps in conducting a cross-sectional study:

- Identification of a defined source population
- Choice of a sampling design and sampling frame for selecting the study participants
- Measurement the exposure and health outcome status of the study population

The (underestimated) beauty of cross-sectional studies (1)

- *Instantaneous effects*
  - (marker of) exposure \(\rightarrow\) marker of effects
  - ozone, smoking, hormonal effects, enzymatic effects etc.
- *Accidents (no time for selective forces)*
  - short term exposure \(\rightarrow\) marker of effects
  - trigger of disorders (e.g. antigens), acceleration of disorders, subjective complaints
  - exposure registration and monitoring of signs
The (underestimated) beauty of cross-sectional studies (2)

- no/little selective survival and/or risk factors are also prognostic factors
  » e.g. heavy lifting and low back disorders
- screening and monitoring
  » hip dysplasia, passive smoking, alcohol consumption, etc. (to estimate the prevalence = prevalence study)
- stability of risk factors (gene, ethnic groups, paternal/maternal conditions, and other invariable personal characteristics) and long duration of the disease
  » bronchial asthma and maternal history of asthma etc.

Cross-sectional studies
Advantage:
- general population
- lower costs
- Information for health planning (e.g., # of hospital beds)

Cross-sectional studies
Disadvantage:
- could not detect disease with short duration (exception: accidents)
- higher proportion of diseases with a long duration (length bias sampling)
- sometimes difficulties to separate cause and disease
  (obesity and arthrosis of the knee)
Sample and survey

- Sample = - representative subset of the population covered
  - best use of scientific methods
- Survey = - statistical picture of a population
  - organized measurement
  - some are good

Reasons:
1. save time and money
2. fewer measurements are more accurate than measurements of the whole population
3. logistic and motivation

Sampling unit = basic unit around which sampling is planned (person, household, etc.)
Sampling frame = list of the sampling units in the population
  (list of residents, list of households, list of registered nurses, etc.)
Sample = sample units chosen from the population eligible to be included
  (exclusion criteria)

Sampling (1)
(see: Kelsey et al. Method in observational epidemiology, 1996)

Sampling (2)
Sampling does not necessarily mean random sampling.

Random sampling = each unit in the target population has an equal chance of being included.

Simple random sampling: - with replacement (unit can be samples twice)
  - without replacement
  - unit in schoolchildren to measure family size
Systematic sampling: - sampling units are spaced regularly throughout the sampling frame
Stratified random sampling: - sample is drawn from different strata
   ▶ subgroups of interest are represented
   ▶ in the analysis the sample can be divided into (more homogeneous) subgroups

Cluster sampling: - sample is drawn from a sampling frame of clusters (classroom, departments of a plant) and all units in that cluster are included

Multistage sampling: - sample is drawn from a hierarchical sampling frame (state, county, city blocks, etc.)

sampling (3)

Sampling (4)

NON-random sampling
   ▶ Convenience sampling
   ▶ Volunteers
   ▶ Systematic, non-random sampling: comparison of extreme groups

Random sampling strategies are required to enable probabilistic statements of generalizability.
(random sampling ≠ random assignment in a clinical trial)

Calculating the proportion of participation

There is no standardized way to calculate the proportion of participation. Indeed participation can be calculated to be 90%, which others would calculate to be 60%.

It is not a truth-telling session.
Proportion of cooperation: (Slattery et al. 1995)

\[
\text{Propcoop} = \frac{I + P}{I + P + R}
\]

I = interviewed individuals
P = partially interviewed individuals
R = refusals

Proportion of response: (Slattery et al. 1995)

\[
\text{Propresp} = \frac{I + P}{I + P + R + NCP}
\]

I = interviewed individuals
P = partially interviewed individuals
R = refusals
NCP = no contact because of inability to locate or unable to reach participant (including: moved/died after selection)

Proportion of contact: (Slattery et al. 1995)

\[
\text{Propcont} = \frac{I + P + R + NE}{I + P + R + NCP + NE}
\]

I = interviewed individuals
P = partially interviewed individuals
R = refusals
NCP = no contact because of inability to locate or unable to reach participant (including: moved/died after selection)
NE = not eligible by study criteria (outside the age range, previous history of cancer, pregnant, etc.)
Example:

I = 400 (interviewed individuals)
P = 200 (partially interviewed individuals)
R = 200 (refusals)
NCP = 100 (no contact because of inability to locate)
NE = 100 (not eligible)

Propcoop = 600/800 = 75%  
Propresp = 600/900 = 66.7%  
Propcont = 900/1000 = 90%

Which one would you like to present?

Proportions of participation are typically for studies involving:

(Kelsey et al. 1996)

- Interviewing: 75% or less
- Clinical examinations: 55–65%
- Postal questionnaires: less than 60%

Quantification of internal validity (wrestling with the truth)

Determination of target and actual population & sampling of the study population → selection bias

Collection information from members of the study population → information bias

Analysis with disentangling of effects → mixing of effects = confounding
A bias does not mean that a study is useless.

Accuracy
  \[\text{Precision} \quad \text{Statistical Power}\]

Validity
  \[\text{Selection of the sample} \quad \text{Selection bias} \quad \text{Reliability}
  \quad \text{Validity of an instrument (Specificity, Sensitivity, PPV...) \quad \text{Confounding bias}}\]

Collection of information
  \[\text{Information bias} \quad \text{Mixing of statistical effects}\]

How to “quantify” a selection bias?
Study population \(\neq\) target population

External population \(\}\) external validity, theoretical aspects

Target population \(\}\) selection & info bias

Actual population \(\}\) statistical inference
Study population
Quantification of biases

If Bias < 1
then OR, RR, PR is biased **towards** the null-value

If Bias > 1
then OR, RR, PR is biased **away** from the null-value

Selection biases =
selective forces prior or during recruitment of the sample that change the composition of the sample.

- Selection-in
- Detection bias
- Response-bias
- Selection-within
- Selection-out: Selective survival, infertile worker effect, healthy smoker effect
Selective survival

Selection bias

<table>
<thead>
<tr>
<th>E</th>
<th>Ė</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

Target population
Actual population

Selection odds

<table>
<thead>
<tr>
<th>α=</th>
<th>β=</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/A</td>
<td>B/B</td>
</tr>
</tbody>
</table>

γ= | δ= |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C/C</td>
<td>D/D</td>
</tr>
</tbody>
</table>

No selection bias if

\[ \frac{α \times δ}{β \times γ} = 1 \]

\[ \frac{α \times δ}{β \times γ} > 1 \] = away from the null value

\[ \frac{α \times δ}{β \times γ} < 1 \] = towards the null value

Response bias

<table>
<thead>
<tr>
<th>E</th>
<th>Ė</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

Example:
Individuals who are not exposed and not diseased are less likely to respond to a survey.

Remedy:
- obtain high response
- calculate response proportion
- compare baseline characteristics of responders and non-responders
Example:
Likelihood of an incident case to survive until the investigation is different for exposed and unexposed.

Remedy:
No prevalent cases in cross-sectional studies if selective survival is important.

Typical bias of cross-sectional studies:
Length-biased sampling
Cases with a long disease duration will be over-represented; cases with a short duration will be under-represented.

There are two relationships involved:
⇐ exposure-disease relation
⇐ exposure-duration relation

How to deal with this bias?
- collect diagnosis dates ⇔ duration of the disease
- collect exposure dates ⇒ induction period (retrospective ascertainment)
Selective Survival (prevalence-incidence bias)

<table>
<thead>
<tr>
<th></th>
<th>E</th>
<th>É</th>
<th></th>
<th>E</th>
<th>É</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>5,000</td>
<td>8,000</td>
<td></td>
<td>D</td>
<td>2,500</td>
</tr>
<tr>
<td>ND</td>
<td>15,000</td>
<td>72,000</td>
<td></td>
<td>ND</td>
<td>14,250</td>
</tr>
</tbody>
</table>

OR = 3; RR = 2.5

\[ \frac{\alpha^* \delta}{\beta^* \gamma} = 0.79 < 1 \]

Bias towards the Null.

Selective migration

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

\[ \frac{\alpha}{\beta} \frac{\gamma}{\delta} \Leftrightarrow \frac{\alpha^* \delta}{\beta^* \gamma} = 1 \]

Example:
Individuals migrate out of a polluted area or risk group independently of their disease status.
Effect on Incidence?

Selective migration: selection within or -in

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

\[ \frac{\alpha}{\beta} \frac{\gamma}{\delta} \Leftrightarrow \frac{\alpha^* \delta}{\beta^* \gamma} < 1 \]

Example:
Selection-within:
- Diseased wish to change their condition from exposed to unexposed.

Selection-in:
- Persons with early signs of disease or predispositions are less likely to become part of the exposed group (jobs, joggers, etc.).
Healthy worker effect or healthy smoker effect, etc.:

- selection in
- selection within
- selection out

Ratio of observed to expected deaths from all causes among active and terminated rubber workers (Delzell & Monson 1981)

<table>
<thead>
<tr>
<th>Duration of employment (years)</th>
<th>Years since starting work</th>
<th>2-9</th>
<th>10-19</th>
<th>20-29</th>
<th>30-39</th>
<th>40-49</th>
<th>≥50</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-9</td>
<td></td>
<td>0.5</td>
<td>0.8</td>
<td>0.8</td>
<td>1.3</td>
<td>1.1</td>
<td>-</td>
</tr>
<tr>
<td>10-19</td>
<td></td>
<td>0.6</td>
<td>1.1</td>
<td>1.0</td>
<td>1.1</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>20-29</td>
<td></td>
<td>0.7</td>
<td>1.2</td>
<td>1.1</td>
<td>1.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥30</td>
<td></td>
<td>0.7</td>
<td>0.9</td>
<td>1.1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Proportion of ex-workers increase.

Example:
Women with a life birth tend to terminate their job (maternal leave, Europe). Women with spontaneous abortions stay on the payroll. In jobs with small increased risk of sp. a., women with sp.a. will accumulate.

Remedy:
- only first pregnancy
- exclude housewives
**Example:**

Exposure produces an increased chance to have the disease detected (or vice versa). Women who receive oral contraceptives have more gynecological examinations. Some gynecological diseases might therefore be more frequent in women who use oral contraceptives.

Under-detection-bias

**Lead time bias (resulting from a selection bias)**

- **Latency period**
- **Duration of the disease**
  - without early detection
  - with early detection

**Classification of selection biases**

<table>
<thead>
<tr>
<th>Selection due to</th>
<th>Selection due to</th>
<th>Selection due to</th>
</tr>
</thead>
<tbody>
<tr>
<td>exposure</td>
<td>disease</td>
<td>a confounder</td>
</tr>
<tr>
<td><strong>Initial</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detection bias</td>
<td>Response-bias</td>
<td></td>
</tr>
<tr>
<td>Selection within</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Final</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selection-out</td>
<td>Selective survival</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(length-biased</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sampling)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>In fertile worker effect</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Healthy smoker effect</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Choice of controls</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Berkson's Bias</td>
<td></td>
</tr>
</tbody>
</table>