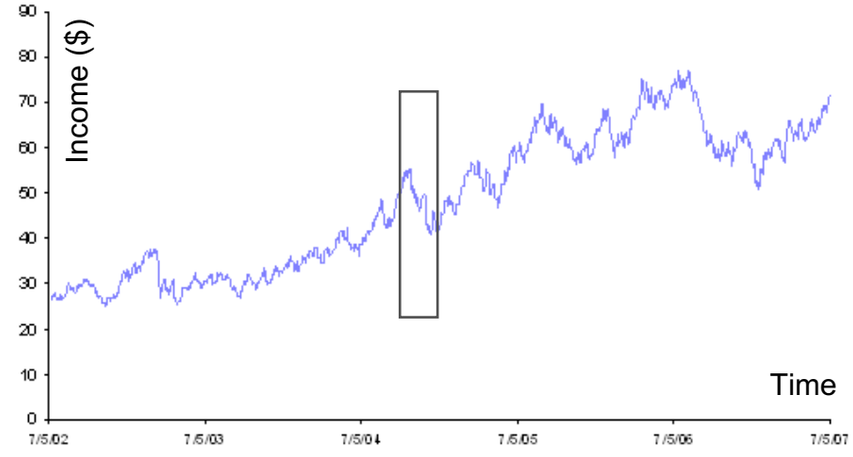
Medical statistics and informatics

Quantitative medicine

Miklós Kellermayer

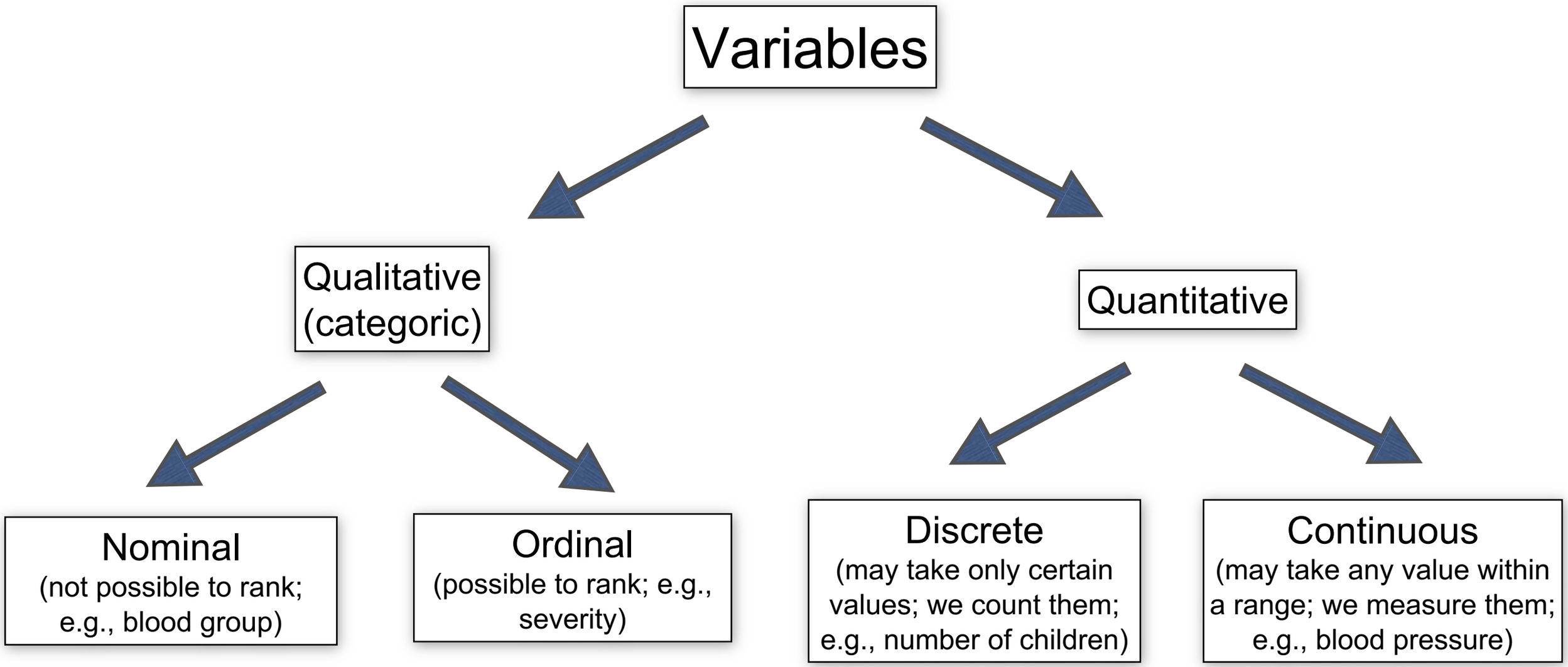
Fundamentals of statistical thinging

- Data are not “only numbers” (3850 vs. 3850-gram infant)
- Data beat anecdotes (5 year, $5 million National Cancer Institute study vs. TV interview with the mother of a leukemic child)
- Beware of the lurking (hidden) variable (Students playing music excel in other studies...?)
- Source of data is important
  (we draw conclusions from samples; representativity)
- Variation is everywhere (role of random variation)



- Be careful with conclusions (correlation is not cause-and-effect)
- Statistical data: information (can be encoded, stored, transmitted, analyzed)
- Medical data, medical knowledge (giant information pool)

Data: values of stochastic variables



There is random variation in the values of the variable.

Objectives of clinical studies

- Estimation

  Estimation of certain **features** of a population.
  E.g., frequency of diarrheal episodes in children under 5, incidence of Covid infection in pregnant women, etc.

- Associations

  Investigation of the **association** (correlation) between a factor of interest (environmental parameter) and a particular outcome (disease, death).
  E.g., does smoking increase the incidence of respiratory infections, does Covid infection increase mortality, etc.?

- Evaluation of intervention

  Evaluation of the **efficiency** of a drug therapy or other medical (e.g., surgical, vaccination) intervention.
  E.g., does the use of sleeping nets reduce the risk of malaria, does Covid immunization reduce mortality/morbidity, etc.? The efficiency of a diagnostic test is evaluated the same way.

Methods of studies - Vital statistics

John Graunt, 1662
Natural and Political Observations upon the Bills of Mortality
First analysis of vital statistics

Edmund Halley, 1693
Natural and Political Observations upon the Bills of Mortality
First survival table

William Farr, 1807-1883
Registrar General, England and Wales
Systematic use of vital statistics

NB: *Vital statistics - data of births and deaths*

Methods of clinical studies I.

- ## A. Vital statistics analysis

  Often provides the first clues to to the anassociation between a disease and its cause. E.g., increase in mortality from lung cancer and its possible association with increased frequency of cigarette smoking was initially noted from vital statistics data.

- ## B. Observational studies

  The disease is observed without actually intervening. Sampling methods are important: sample size, probability of selection into observed group.

  ### 1. Cross-sectional studies.

  Relatively inexpensive, easily executed. Measures the prevalence but not the incidence of the disease. Therefore, associations are difficult to interpret.

  N.B.:

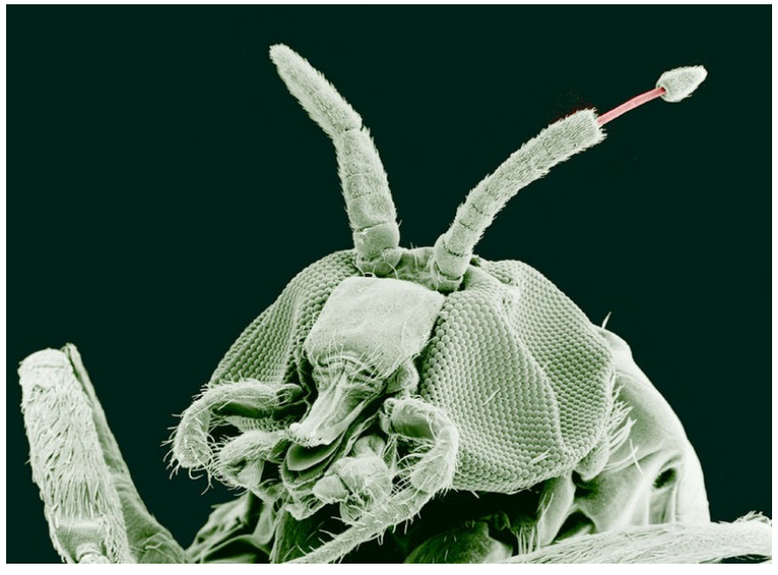
  Prevalence - frequency of diseased in total population at a given point in time.

  Incidence - number of new patients in the diseased population within a time interval.

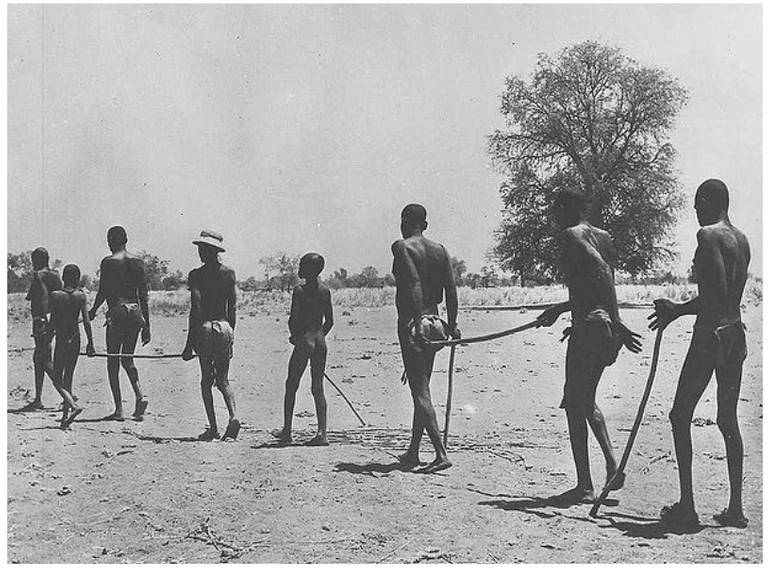
Problem of cross-sectional studies

Onchocerciasis study: blind persons are of lower nutritional status

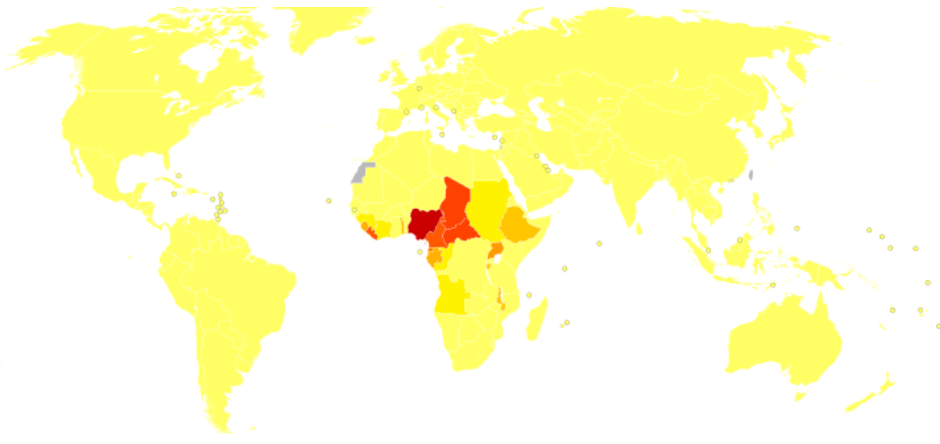
Onchocerciasis: river blindness, Robles'-disease
Pathogen: *Onchocerca volvulus* (nematode, roundworm), survives for up to 15 years as a parasite in the human body.
Transmitted to humans by the bite of a female blackfly (*Simulium yahense*).
Upon worm necrosis, the endosymbiont (*Wolbachia*, bacterium) is released, evoking tissue necrosis (eg., in the eye).
Second leading cause of infectious blindness.



Onchocerca volvulus worm, as it is released from the antenna if the blackfly.



Children leading blind adults in sub-Saharan Africa.



Global distribution of onchocerciasis.

Low nutrition - defective immune response.
But: Blindness - interferes with nutrition.
Cause or effect? Only longitudinal (time-dependent) studies can resolve the issue.

Methods of clinical studies II.

- ## B. Observational studies (cont'd.)

The disease is observed without actually intervening.

2. Longitudinal studies

Individuals are followed as a function of time.
Continuus: from birth to death.
Retrospective / prospective: past records / present or future records.
Simplest type: periodiclaly repeated cross-sectional studies.
Period (interval): depends on the type of disease (e.g., diarrhea repeats in short episodes).
Patient group may be ***dynamic*** or ***fixed.***
Dynamic group: individuals may leave or enter the group (e.g., diarrhea in under-5-year-old children).
Fixed group (cohort): group remains unchanged throughout the study.

3. Case-control study

One group: ***diseased (case group)***. The other group: ***control (control group)***
E.g.,: does breast feeding reduce infant mortality? (Case group: infants died in first year; control group: children of identical gender living in the same area)
Highly effective in the investigation of rare diseases and large effects. Study design can be difficult.

Methods of clinical studies III.

- C. Experimental studies

  Individuals are allocated *a priori* (by the investigator) into groups: control, treated.

  Considerations: randomization, pairing, single- and double-blind studies, use of placebo, ethical issues (withholding therapy).

  1. Clinical trials

  Investigation of the efficiency of pharmaceuticals.

  2. Vaccination trials

  Investigation of the efficiency of vaccines.

  3. Intervention trials

  a.) Evaluation of profilactic (prevention) protocols (e.g., antimalarials).

  b.) Evaluation of non-pharmaceutical preventive measures (e.g., sleeping nets - malaria).

Clinical trials - History

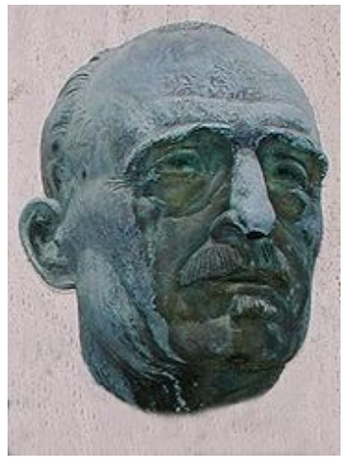
- Egypt - Imhotep (~3000 BC, surgery, herbal medicine)
- China (~2700 BC, herbal medicine)
- Ancient Greeks and Rome (Hippocrates, 460-370 BC, Galenus, 130-200 AD)
- Middle ages - Renaissaince (“Consilia”, Leonardo Da Vinci - anatomy)
- Edward Jenner (1749-1823, smallpox)
- Oliver Wendel Holmes (1809-1894, anaesthesia, puerperal fever)
- **Ignatius Semmelweis (1818-1865, savior of mothers)**
- Louis Pasteur (1822-1895, fermentation, anthrax, rabies)
- Robert Koch (1843-1910, tuberculosis)
- Emil von Behring (1854-1917, diphtheria)
- Elie Mechnikov (1845-1916, phagocytosis)
- Paul Ehrlich (1854-1915, complement system)
- Florence Nightingale (1820-1910, modern nursing)
- Alexander Fleming (1928 penicillin)
- Banting and Best (1921 insulin)
- World War II - nazi human experiments, Nürenberg Code 1947
- 1953 National Institutes of Health, USA: Principles of the practice of medical experiments on humans

Semmelweis Ignác Fülöp (1818-1865)

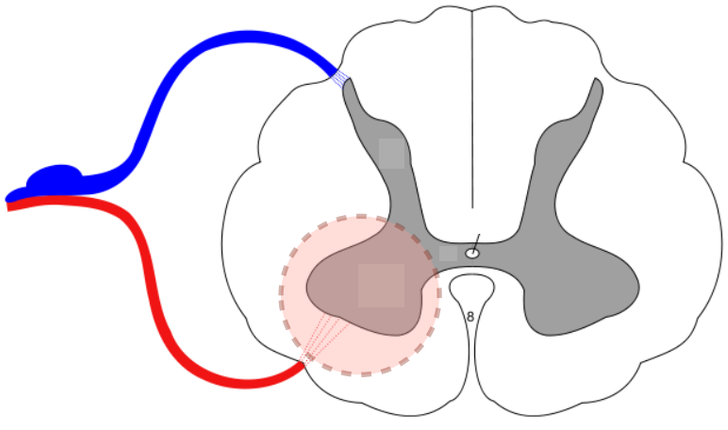
Poliomyelitis

Poliomyelitis anterior acuta, Heine-Medin disease, infantile paralysis

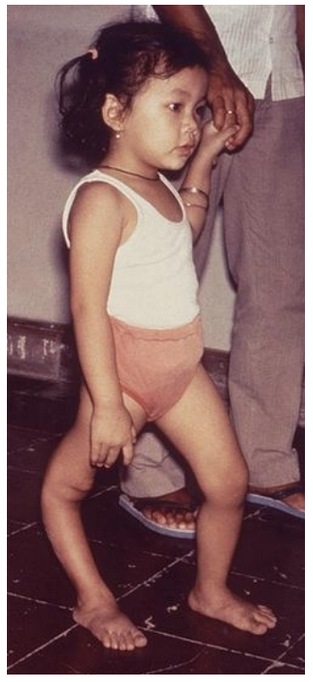
Jakob Heine, 1840



Oskar Medin, 1890



Poliovirus preferably attacks the motor neurons of the anterior horn in the spinal chord.



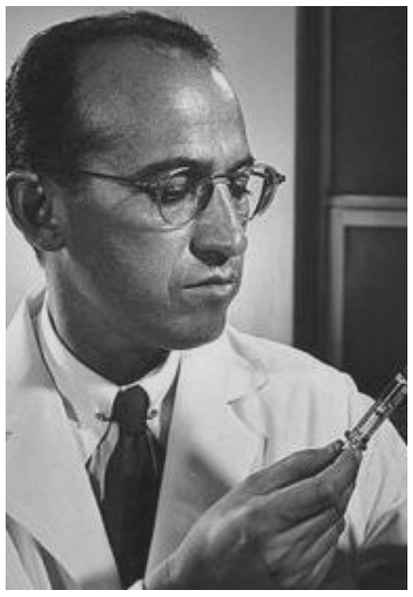
Flaccid paralysis of limb muscles, muscle atrophy, deformation of limbs.

“Iron lung” (negative-pressure ventilator)

In severe cases (in bulbispinal polio), respiratory failure occurs, requiring life-long respiratory assistance.

Randomized, controlled double-blind experiments

Polio vaccine trials



Jonas Salk, 1955
IPV: Intravenous Polio Vaccine

Albert Sabin, 1962
OPV: Oral Polio Vaccine (“Sabin drops”)

Is the polio vaccine effective?

Consideration	***Problems***
The vaccine is simply provided.	Intensity of epidemic varies by itself (solution: comparative study).
Establishment of a **Control group**	Ethical concerns (reassurance: treatment also carries risks)
Comparison	Different size of control and treated groups (solution: calculate ratios)
Group selection	Lurking variable (e.g., financial background, hygiene) (solution: similar groups - **randomization**)
Selection of administration method	Effect of subconscious factors (solution: placebo)
Diagnostics	Driven diagnosis (solution: **double blind** experiment)

	Group size	Incidence
Treated group	200 000	28
Control group	200 000	71

Sources of error

Random error:

Stochastic effects. Lead to measurement uncertainty.
Reduces precision, but does not lead to invalidity.

Systemic error (bias):

1. Selection bias

Systemic, relevant differences exist between the selected and non-selected individuals. E.g., the most severe diarrhoic patients do not get selected into clinical groups in certain countries.

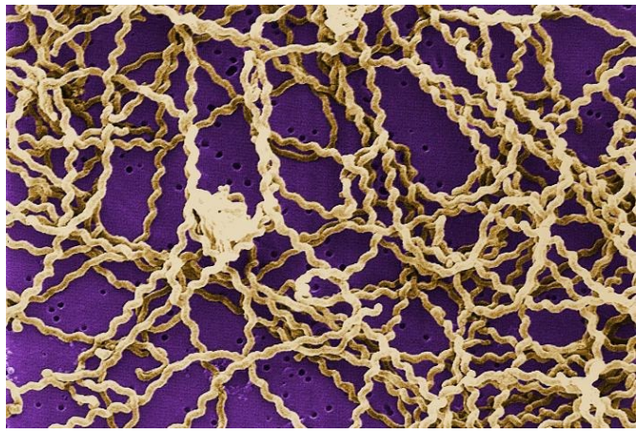
2. Confounding bias

Participating groups differ ***a priori*** in terms of the investigated parameter.
E.g., prevalence of leptospirosis differs between city and village residents.
However, the ***gender is a confounding*** parameter: leptospirosis prevalence differs according to gender, but gender composition also differs between city and village.

3. Information bias

Errors caused by questionnaire problems, the examiner, the responder, or instruments.

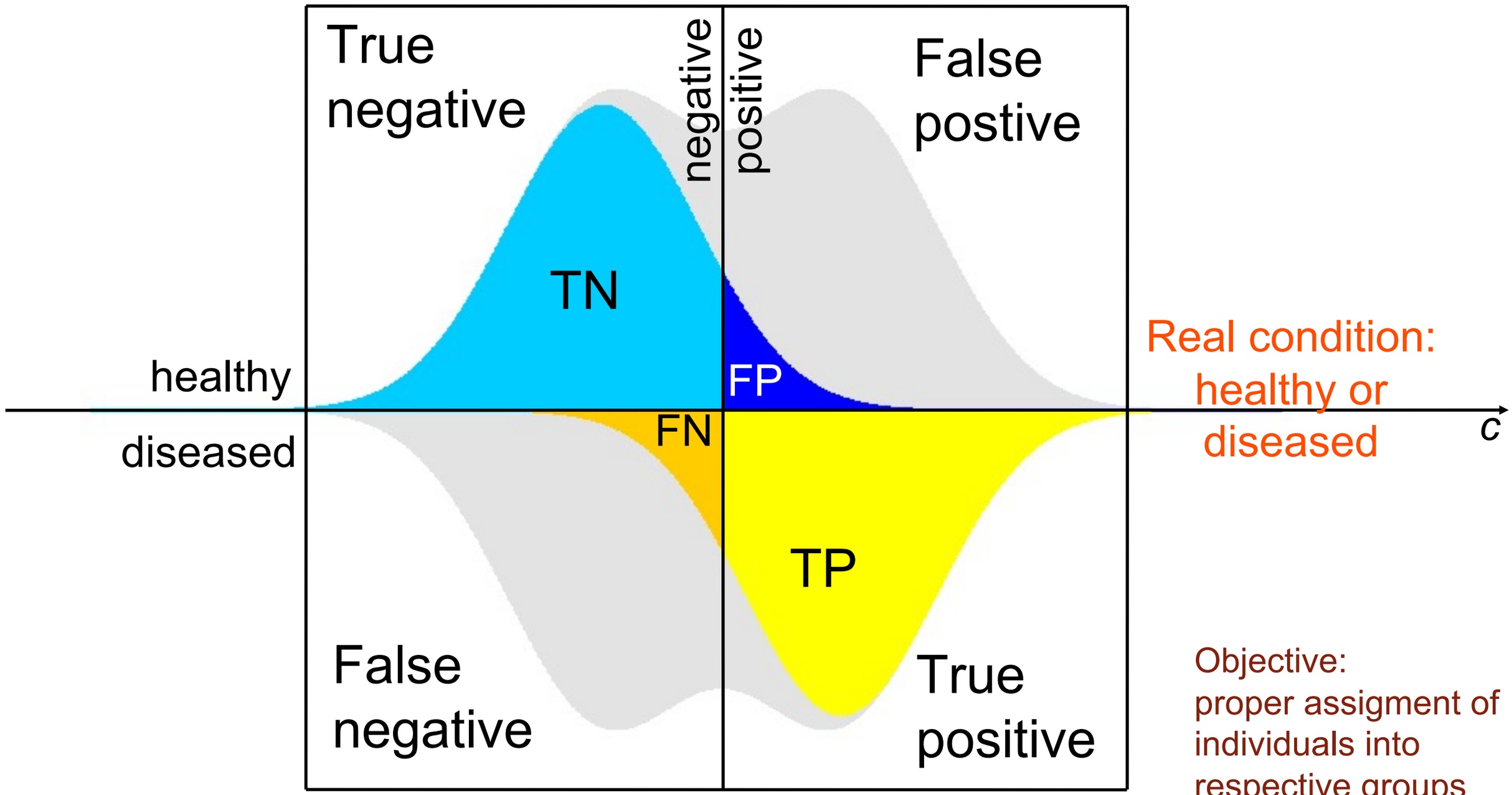
Leptospirosis: most frequent zoonosis (disease spreading from anmals to humans). Infectious disease caused by Leptospirae of the Spirochaeta family



Leptospira bacteria, SEM image.

Parameters of correct group selection

Contingency table: Confusion matrix



Objective:
proper assigment of individuals into respective groups (e.g., patients, risk groups, etc.)

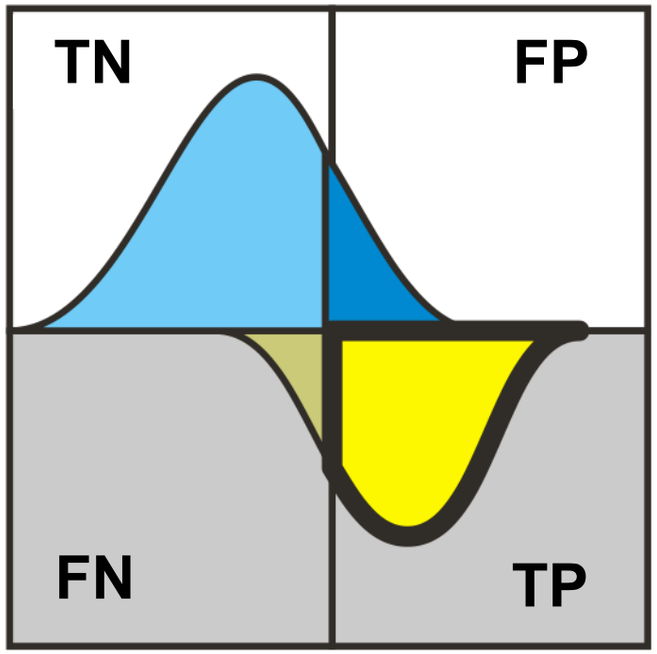
Diagnostic (selection) sensitivity

Also called:

-True Positive Rate (TPR)

-Hit Rate

-Recall



Probability that the test finds the diseased positive.

Positive within diseased.

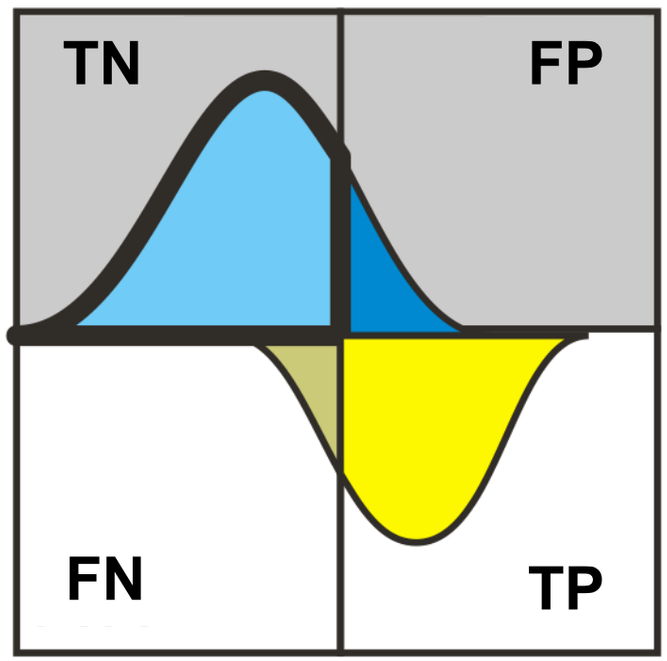
$$\frac{\text{▼}}{\text{▼}} = \frac{\text{true positives}}{\text{total diseased}}$$

Large-sensitivity tests (100%) are required in early diagnosis (screening) so that few patients remain unrecognized. E.g., in the beginning of a vaccination trial it is important to identify (end exclude) all the diseased.

Diagnostic (selection) specificity (SPC)

Also called:

-True Negative Rate (TNR)



Probability that the test finds a healthy negative.

Negative among healthy

$$\frac{\text{}}{\text{}} = \frac{\text{true negatives}}{\text{total healthy}}$$

High-specificity tests (near 100 %) are important when the false positive values have severe consequences (e.g., surgery). E.g., in a post-vaccination follow-up all the healthy should be identified.

Data acquisition

- Census
- Sampling

Simple - Sampling frame, random table

Complex

Layered (age groups, gender)

Multi-step (school > classes > child groups)

Clusters

Size of sample?

-Ethical, financial issues

-Standard error, power

Medical activity

Series of decisions!

The logic of a research scientist and a physician are similar:

Observation	Symptoms
Consideration, hypothesis	Preliminary (target) diagnosis
Experiment	Tests (laboratory, imaging)
Theory	Diagnosis

Therapy

Diagnostics, differential diagnostics

The physician meets the individual, not an abstract group.

Diagnosis: identified disease of the patient.

Diagnostics: intellectual process during which the physician arrives at the diagnosis.

dia = apart, *gnosis* = knowledge.

Differential diagnostics: selection of correct diagnosis from many alternative choices.

Diagnosis is not a fact, but a possibility.

Steps of differential diagnosis:
1) data collection, 2) evaluation, contemplation, 3) differentiation.

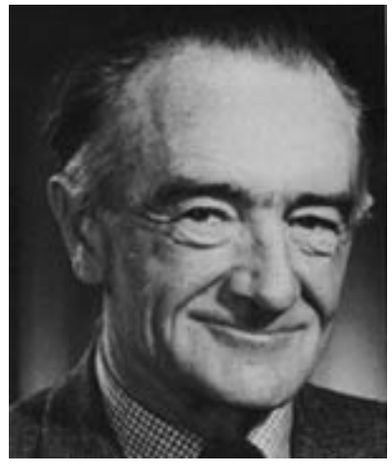
“Evidence-based medicine”

“The sole criterion of scientific truth is the experiment.” *(Richard P. Feynman)*

Application of the best available evidence gained from the scientific method in medical decision making.

History:

- Ancient Greeks (?)
- Ancient Chinese medicine (?)
- Avicenna (*Ibn Sīnā*) (XI. sz.): *Canon medicinae* (1025); 14 volume medical encyclopedia
- Ignatius Semmelweis (1818-1865): “savior of mothers”
- Archie Cochrane: Scottis physican epidemiologist. *‘Effectiveness and Efficiency: random reflections on health services’ (1972)*
- Introduction of the term “Evidence-based medicine”: Gordon Guyatt, 1992.
- Cochrane Centers, Cochrane Collaboration, 1993. International network, Cochrane library, reviews.

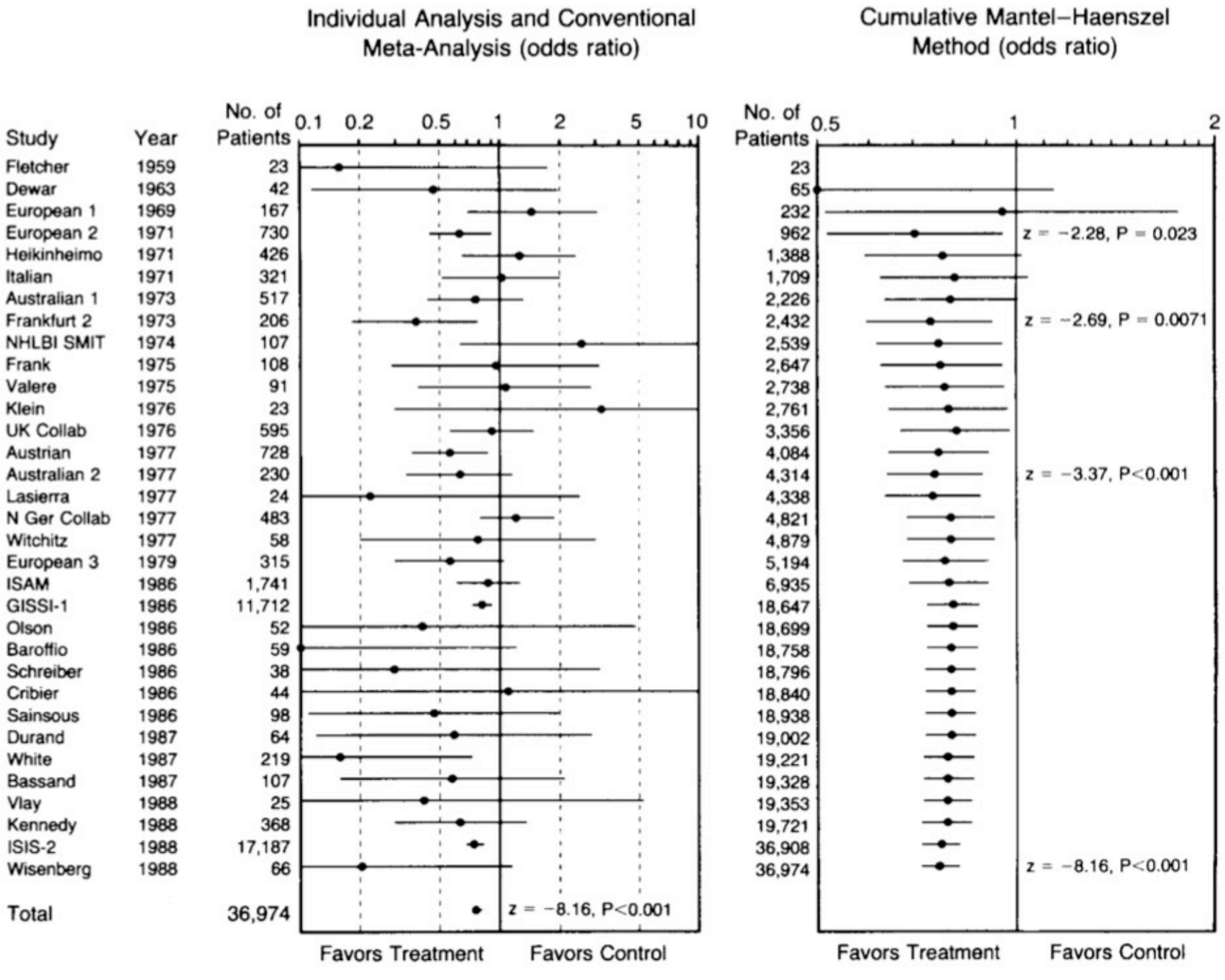


Archie Cochrane (1909-1989)

The Cochrane Collaboration
Working together to provide the best evidence for health care

“Evidence-based medicine”

Effect of streptokinase treatment in acute myocardial infarction



N.B.:

- *meta-analysis:* combined analysis examining several hypotheses.
- *odds ratio:* one parameter of probability. For odds = 1 the probability of the given outcome is identical in both groups.
- With a proper evaluation of clinical trial results the efficiency of streptokinase treatment could have been identified by 1973.

“Evidence-based medicine”

Practice:

1. **Evidence-based guidelines** - practice of evidence-based medicine at the organizational or institutional level.
2. **Evidence-based individual decision making** - evidence-based medicine as practiced by the individual health care provider

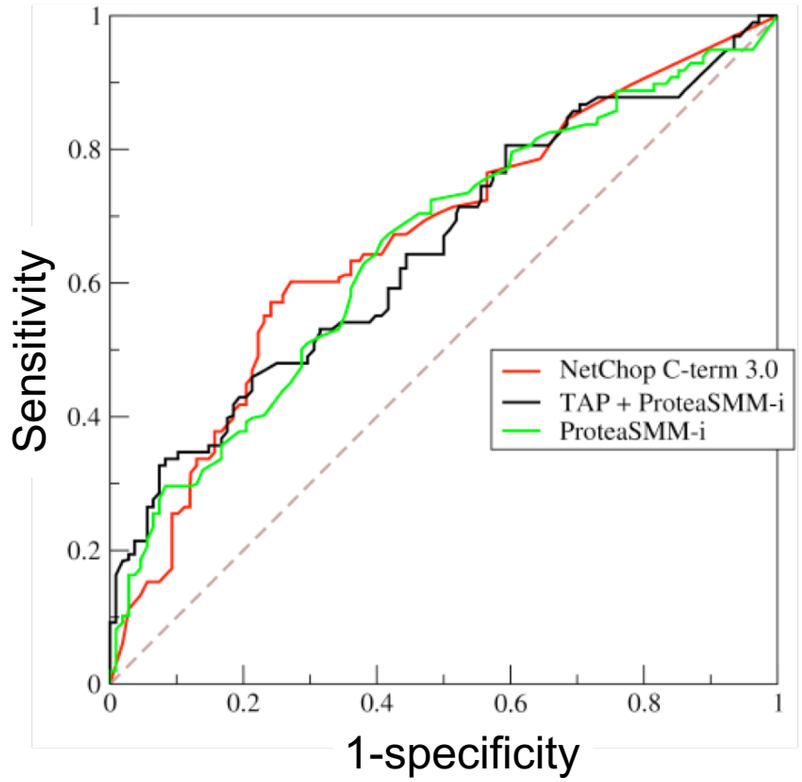
Types:

1. Application of the recommendation of original medical literature.
2. Application of the recommendation of review literature.
3. Application of the recommendations of medical organizations.

How good is the evidence?

1. Based on criteria of professional organizations. E.g.:
   I. Evidence obtained in properly executed double blind randomized trial.
   II. Evidence obtained in properly designed controlled trial (but with no randomization).
   III. Opinions of selected authorities.
2. Statistical criteria.

Mathematical analysis of the efficiency of diagnostic and therapeutic methods. E.g., ***AUC-ROC curve*** (“area under the receiver operating characteristic curve”).



Computer-aided medical decisions I.

Diagnostic steps (or therapeutic decisions) assisted by computer algorithms.

Medical knowledge: datasets of symptoms and formalized diseases.

Symptoms: sum of information characterizing the health status of the patient (anamnesis, physical signs, laboratory tests, diagnostic imaging tests)

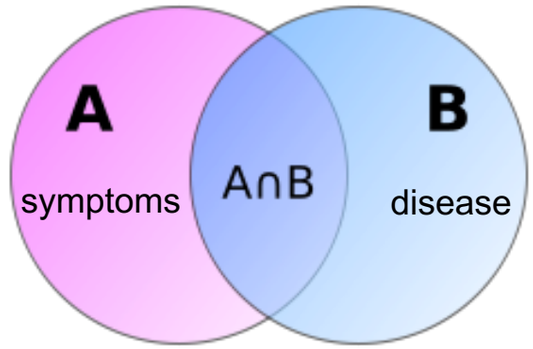
Formalized diseases: diagnostic categories organized into logical order (e.g., upper respiratory diseases, malignant tumors, etc.)

Computer-aided medical decisions II.

- Computer Aided Diagnosis (CAD), diagnostics supported with artificial intelligence.
- **Objectives**:
  Simulation of specialist (medical) arguments.
  Reduce guessing (reduce number of hypotheses or target diagnoses)
  Consideration of pathophysiological argumentation.
- Generally employed **logical iteration**:
  1. Do the observed symptoms occur in the considered diseases?
  2. Assign weighting points to the diseases according to the number of observed symptoms.
  3. Rank the diseases according to the points.
  4. Ivestigate whether the observed symptoms contain ones which are not present in the most highly-ranked disease.
  5. If yes, consider the next disease in the ranking order.
  6. In case of **new symptoms** the iteration is started from the beginning (stage 1); if not, the diagnosis is established according to the ranking.
- **Problems:**
  The **frequency** and **severity** of the symptoms are difficult to evaluate correctly.
  New symptoms make the iteration very difficult.

Computer-aided medical decisions III.

Objectives: set-theory associations are investigated between symptoms and diseases.



Boolean operators:

A OR B: union

A AND B: intersection

A XOR B: union-intersection

Medical **arguments**:
“Runny nose is *almost always* present in common cold.”
“Acute pyelonephritis is *usually* accompanied by cystitis and inflammation.”
“Acute pyelonephritis is *sometimes* accompanied by fever, shiver and malaise (discomfort).”

Common cold, acute pyelonephritis: diseases ($D_{1\text{-}2}$)
Runny nose, cystitis, inflammation, shiver, malaise: symptoms ($S_{1\text{-}6}$)
“Almost always, usually, sometimes”: mathematical conditional operators

$D_1 \xrightarrow{\text{almost always}} S_1$

$D_2 \xrightarrow{\text{usually}} S_3 \text{ AND } S_4$

$D_2 \xrightarrow{\text{sometimes}} (S_3 \text{ OR } S_4) \text{ AND } S_6$

Feedback

https://chart.googleapis.com/chart?chs=450x450&cht=qr&chl=https://feedback.semmelweis.hu/feedback/index.php?feedback-qr=KX3JU6T85AOHUDYJ