**Epidemiology week 1**

**Lecture 1: Introduction to epidemiological research**

Epidemiology = occurrence research

= concerned with the **frequency** and **pattern** of health events in a population

Present: shift in focus from **single** determinants & disease occurrence to **combined** causes (determinants)

Future: measurements capacity is growing -> more complex topics

* More data (data being collected in hospitals, devices, exposure in people, genetics)

1. **Research question**

* Outcome outcome = f(Determinants)
* Determinants e.g bleeding =f(Aspirin)
* Domain = situation for which the relation and its occurrence is studied
  + Goal: generalize the empirical relation to a larger group
  + E.g. findings in study about study including elderly generalized to the findings being true for older (healthy) adults

Example:

* + Outcome: myocardial infarction
  + Determinant: variation in intake of SFAs
  + Domain: humans

MI = f(SFA variation | confounders)

1. **Design of the occurrence relation**

Fractures = f(antidepressants | confounding)

**DEPTH model**

* Diagnostic knowledge – what’s wrong
* Etiologic knowledge – why ill
* Prognostic knowledge – what if I don’t intervene
* Therapeutic (prognostic) knowledge – what if I intervene

1. **Collection of data (Lecture 2)**
2. **Data Analysis and scientific interpretation**

**Measures:**

* Frequency (measures: Prevalence & Incidence)
* Association
* Impact

|  |  |  |
| --- | --- | --- |
| **Frequency measures** | **Association measures** | **Impact measures** |
| Prevalence |  | Attributable risk |
| Cumulative incidence | Relative risk | Population attributable risk |
| Incidence density | Relative rate |  |
| Odds | Odds ratio |  |

**Prevalence** = estimate of probability/risk that one will be ill at some **point in time** (%) -> severity

= incidence (# new cases) \* duration

= proportion of a particular population found to be affected by a medical condition at a specific time

Example: Fractures =f(antidepressants)

P= number of people using antidepressants / population at risk

**Incidence** = How often does XXX occur?

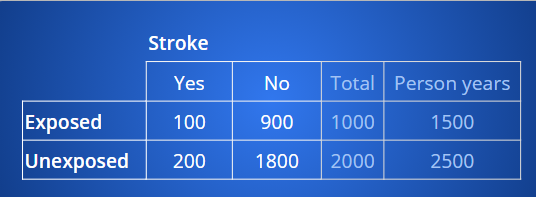
Example: **Crude (naïve) estimation**

10 year risk of fracture when 20 people at risk and 2 get a fracture within 10 years

* 2/20 = 10% = cumulative risk
* Underestimation of real risk
* Need to take complete follow-up in consideration!
  + Instead of 20\*10= 200 years
  + Due to missing values only 176 years observed -> **2/176 > 2/200**

e.g. incidence: how many **new** corona patients in a day

prevalence: proportion of how many sick patients are in hospital and for how long (duration) – total people also like vaccination

**Cumulative Incidence**

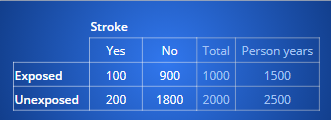
CI of outcome/disease for exposed:

100/1000 = 0.1

**Exposed and stroke/total**

* **Developing a disease in a specified period (total time)**
* Proportion %
* Probability of XXX (0-1)
* Definition of
  + Time frame
  + Population ‘at risk’
  + **Underestimates** the risk of being sick because it accounts for the whole time period and doesn’t take person years into account

**Incidence density (ID) = Incidence Rate:**

IR of outcome among the exposed:

100/1500=0.067

**Exposed and stroke/person years**

* How quickly new cases develop over a specified period (Corona numbers in a certain day)

**# cases / person years of population at risk**

ID= 2/176

* Prevents underestimating the probability of developing a disease over a specific time period
  + Uses person years instead of years

**Cumulative incidence Incidence Density**

|  |  |
| --- | --- |
| * Proportion * Easy to use * Not ideal with long time periods (not equal follow-up for all) * Use for small time units * **years** | * Density * Harder to use * Better for large time units * **Person years** |

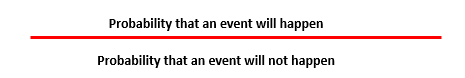
|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Prevalence** | **Cumulative incidence** | **Incidence density** |
| Numerator | All cases counted on a single occasion | New cases occurring during a specified follow-up period | New cases occurring during a specified follow-up period |
| Denominator | All individuals examined - cases and non-cases | All susceptible individuals present at the start of the study | Sum of time periods during which all individuals could have developed disease |
| Time | Single point or period | Defined period | Measured for each individual from beginning of study until disease event |
| Interpretation | Probability of having disease at a point in time | Probability of developing disease over a specified period | How quickly new cases develop over a specified period |

**Measure of association & Measures of impact**

**Odds**

**Odds of developing a stroke are**

**100/200**



10 year outcome odds = (2/20) / (18/20) = **0.11**

* Used for logistic regression if real risk cannot be estimated
* **Real risk** preferred whenever possible

**Odds Ratio**

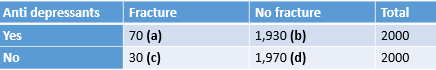
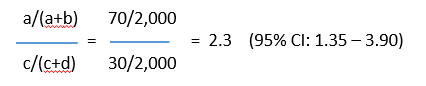


**(100/200)/(900/1800)**

**Risk Ratio (RR):**

**(100/1000)/(200/2000)=**

* Short follow up -> easy interpretation
* Long follow up -> limited because of CI estimation
* Cannot be applied on case control study (case control referring to odds ratio)
* Rate Ratio often preferred
* Over TOTAL not person years



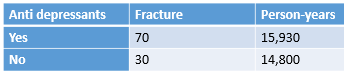
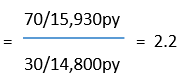
**CI refers to accuracy**

**2.3 prefers to precision**

* **1 =** no association between exposure and disease
* **>1 =** positive association (exposure increases risk)
* **<1 =** negative association = protective effect (exposure decreases risk)

**Rate Ratio**

**(100/1500)/(200/2500)=**



**Interpretation:**

2.2 times higher risk of fractures when use of anti-depressants

**Etiology** = the study of causation or origination

**Etiologic Research**

* Inference is either correct or incorrect = biased

**Etiologic research**= Why is the patient sick?

* Requires knowledge and judgment of confounders/confounding

**Bias** = result of any process that confounds the relation

* Selection bias
* Observation bias
* …
* **Confounding** (mixing together)

= a distortion (inaccuracy) in the estimated measure of association that occurs when the primary exposure of interest is mixed up with some other factor that is associated with the outcome (etiological knowledge)

**Confounder** = a variable that influences both the dependent and independent variable, causing a spurious association

* + Intrinsic – can be handled during:
    - Analysis (when measured correctly) – after getting data
    - Type of data collection (design phase) – beforehand
  + Predicts occurrence of outcome & is associated with determinant
  + Has NO direct effect in the occurrence relation -> extraneous
  + Conditions for confounding factor:
    - Associated with exposure (but not consequence)
    - Associated with outcome (independently of exposure)

Example:

* + Physical activity as confounding factor for research in effect of antidepressants on fractures
  + Risk Ratio needs to be adjusted according to the confounder

Carrying matches not confounder for smoking and then cancer because it cannot be a result of the exposure itself.

Confounding can be removed by randomization

**Validity and Precision**

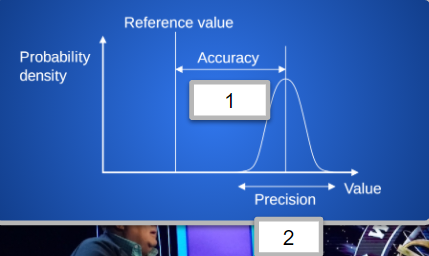


**Precision**

* Reproducibility, Reliability, Consistency
* More precise => the greater the statistical power (at the same sample size) – smaller CI
* Effected by random error

**Accuracy = Valicity**

* The degree it actually represents what it is intended to represent.
* Important influence on the internal & external validity of the study
* Accuracy =><= Systematic error (bias)
* The more accurate – the wider the CI?
* Usual interpretation for CI if we repeated the research exactly the same the estimate would be within the limits (accuracy)



**Question underlying any occurrence relation**:

*What would have happened to the same people if they had not been exposed?*

* Compare exposed to non-exposed
* Non experimental cohorts/case-controls: hard to compare design & analysis
* Experimental cohort: hard to compare design (randomization of the groups)

**Lecture 2: Introduction to epidemiological research**

**Descriptors of data collection**

* + Time
  + Population closed/open (= e.g. district in Utrecht where people can move in/out of)
  + Analysis on all participants (census) or sample
  + Experimental/non experimental

**Types of data collection**

* + **Cohorts**
  + **Cross-sections**
  + **Randomized trials**
  + **Case control studies**

**Cohorts**

* + - Non-experimental (= researcher does not decide who gets treatment) /observational
    - Group of individuals followed up for specific period of time
    - Measuring outcome (waiting for someone to get sick)
    - Census
    - Population closed/open

Purpose:

* + - * Find out if exposure is associated with outcome
      * compare to non-exposure
      * estimate risk of outcome – measure incidence among exposed & unexposed

Source of Bias:

* + - * selection bias
        + when subject specifically included based on known relation to exposure/risk
        + when loss of follow-up has to do with occurrence relation
      * observation/information bias
        + e.g. prior knowledge effects interpretation of X-ray (expecting a fracture) (blinding used to counter selection bias = doctor doesn’t know additional information – all people part of the study are unaware of whether the participant is part of the real study or control group)
      * (confounding)

Cannot be rectified (=adjusted/calculated) in analysis

**Recipe:**

1. Identify groups – unexposed & exposed
2. Measure incidence
3. Compare incidence (exposed & unexposed)
4. Assure comparability

|  |  |
| --- | --- |
| * Latency period * Loss of follow-up * Large sample size * Exposure can change * (un-)ethical? * Cost & time | * Incidence in both exposed & unexposed groups * **Works with rare exposure** * Time frame is clear (t>0) * Less subjected to bias as we don’t know outcome (don’t know what the future holds) – scientific advantage |

**Cross-sections**

* + - Time=0
    - Population closed
    - Census
    - Non-experimental exposure
    - Example asking on the street random people if they got a vaccine or not – one measurement

|  |  |
| --- | --- |
| * Quick * cheap | * Explain selection * Consequence of time   e.g. when studying certain gene with old people where a lot of people with this gene have already died |

**Trials** (experimental cohort)

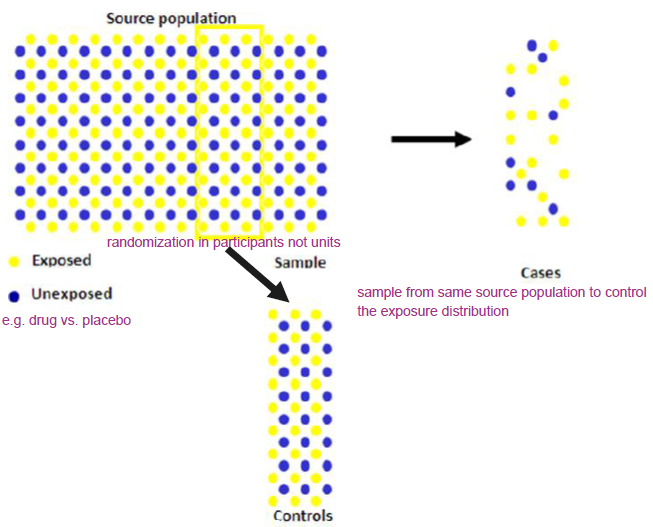
* + - Time>0 (because cohort -> waiting for occurrence)
    - Population closed
    - Census
    - Experimental exposure
      * Subjects intervened to find effects of intervention
      * Q: How effective is intervention?
      * E.g. drug vs. placebo

**Randomized Trials**

= random allocation to exposures

* Groups with comparable means levels of known/unknown determinants of outcome
* That way differences have to be due **to intervention**
  + Best way for good evidence/results

**Randomized Controlled Trial**

* Useful to study effectiveness of interventions
* **Not observational – all others are observational**
* Experimental exposure
* Never for e.g. testing if smoking affects lungs cancer because one would have to decide who gets exposed which is unethical (not causality)
* **Confounding unlikely** (randomization) especially when large study
* Threats: selection bias when follow up

**Case Control**

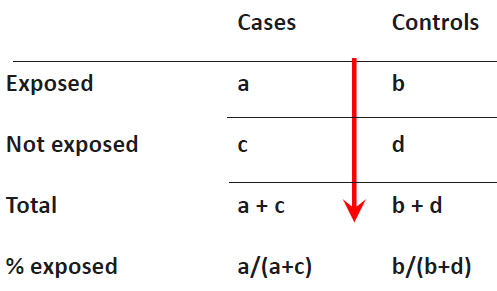
= compares patients who have a disease/outcome (=case) with patients who do not have the disease/outcome

* Retrospectively compares how frequent the exposure to a risk factor is present in each group to determine relationship between risk & disease
  + Time >0
  + **observational**
  + Population usually open (can be closed/nested)
  + Sample of participants (should never have connection to other participants)
  + Non-experimental exposure
  + Always **retrospective** so risk ratio has no meaning in this study type
  + Select people on disease case= people with disease control people without disease

Purpose:

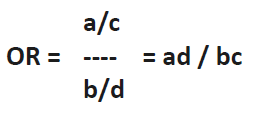
* Provide estimate of exposure in the source population
* Good for **rare diseases** (find exposure and analyze for cause) select people that have the disease

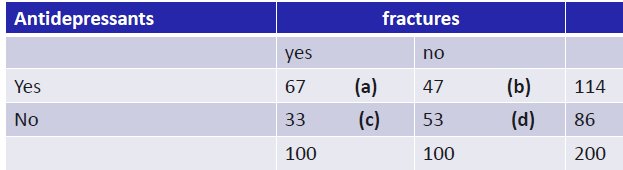
**Distribution of exposure in cases & controls**

**Odds of exposure** =

Prob. to be exposed/ prob. to be unexposed

**Odds ratio=**



**Previous example:**

**Exposure odds ratio (OR)=**

67\*53/33\*47=3351/1551=2.3

Odds ratio (**OR**) **=** incidence rate ratio (**IRR**) **IF** case control study is executed correctly

Source of Bias:

* + - Selection bias
      * When exposed cases are selected specifically
      * E.g. when suspected cases are send to specialized hospital based on knowledge of exposure
    - Information bias
      * Exposure measurement performed different for cases

|  |  |
| --- | --- |
| * No absolute rates/risk * Not good for rare exposures * Prone to bias * Often performed too quick/sloppy | * Research rare diseases * Long latency (=time between exposure & symptoms) * Low cost * Small sample size * ethical |

**Diagnostic Research**

**Diagnostics in practice**

Start: Patient with complaint/symptom

Differential Diagnosis (DD)

* + what is the most important/dangerous possibility?

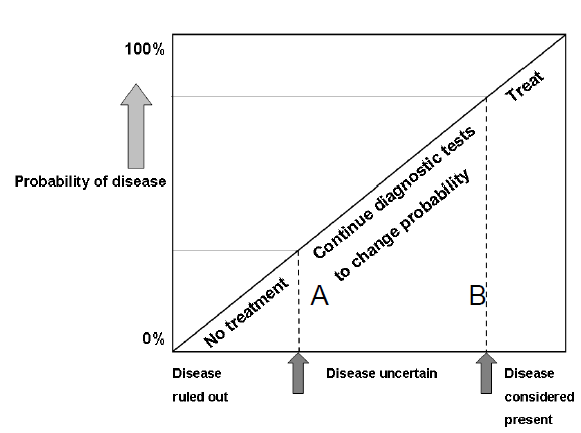
**Prevalence** = prior risk = baseline likelihood of having disease with no other knowledge

**e.g.** child with neck stiffness – 20% prior risk of it being bacterial meningitis

prior risk too low to start treatment for BM but too high to send home -> reduce uncertainty

**Gold standard**

* + Real disease status; ‘ truth
  + Reference / standard test
    - Would bring certainty but oftentimes too invasive/espensive/ineffeicient
* Simpler diagnostics (medical history, simple lab…) *from less to more invasive*
  + Decisive test in case of doubt



**prior risk -> testing -> posterior-risk**

* The bigger difference between prior & posterior risk = the higher value of test
* testing until posterior-risk is acceptable – acceptable uncertainty dependent on disease

**Diagnostics in practice** is an **estimation of risk/change of the presence** of a disease based on patients test results

**Study Design**

* + Research question
  + Domain
  + Determinant(s)
  + Outcome
  + Study design
  + Data analysis, interpretation + reporting

**Research question/ occurrence relation**

* Which test to estimate presence/absence of disease
* What determinants (predictors)
* Determinant- outcome relation
* Occurrence relation = Prediction not explanation
  + %BM = ƒ( age , sex , indicators, etc.)

**Domain**

* Who? = generalization
* Study population = domain sample
* all participants suspected of disease based on symptom in setting
* Study population

**Determinants**

* Diagnostic determinants = all possible/relevant tests
* No prior knowledge of outcome
* Same measurement for all cases

**Size 1 to 10 rule**:

* Size of study population should be a minimum of 10 for 1 determinant (2 det. = 20 cases)

Measure of **Outcome**

= diagnostic outcome/ result of reference test (= gold standard)

* Predict/try to be as close as possible to the truth
* Gold standard test (e.g. very invasive but confirming test would have to be done with all patients for real proof of whether disease present or not)

**Study design**

**Descriptors**

* Non- experimental -> **observational** (because end product will be likelihood/risk ratio)
* Confounding not an issue as we are not trying to explain whether test result is causally related – we just want to know what action to take as a doctor

**Ideal study**

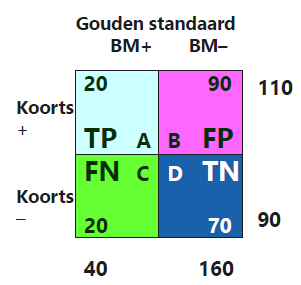
* **Cross-section** = determinants & outcome measurements at the same time (t=0)
* **T=0** because at point of testing the patient already has/ does not have the disease
  + **Testing** outcome & determinants for ideal study to see what determinants predict the outcome accurately

**Data-analysis**

* All outcomes and diagnostic measures

1. **Estimate prior risk**
2. **Compare each test with reference test – univariable (for each case)**
   * Often with odd ratio
3. **Compare combinations index test results with reference test – multivariable**
   * Model building
   * Find added value of certain test or set of tests if it predicts a result better

* Predictive values matter not the certainty of disease being present or absent

****

**Horizontal**:

chance of disease + symptom/test result = predicted value

A/A+B = 20/20+90 = 18%

* 18% chance that disease when symptom

**Predicted value is more important than certainty**

**Vertical**:

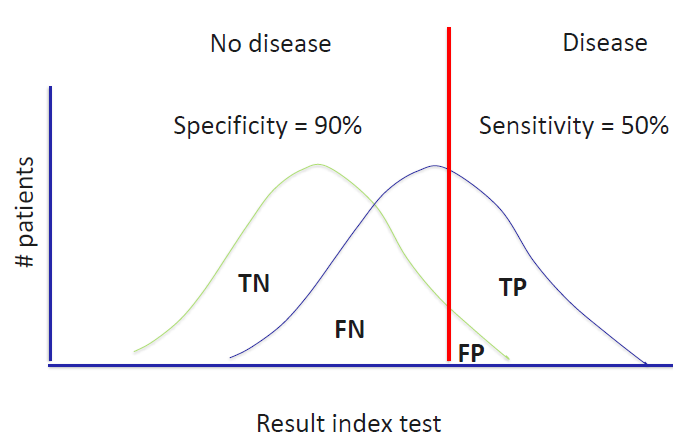
calculating the change of a symptom (fever) when certainty of disease

**Diagnostic Process**

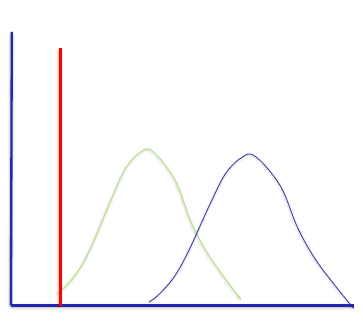
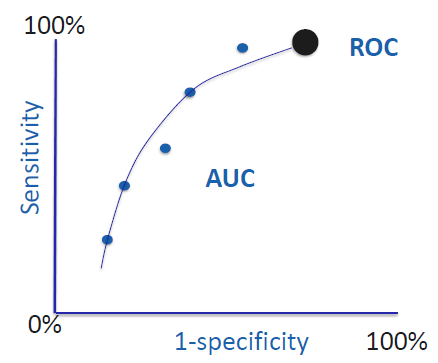
* Hierarchical (often starting with medical history = test 1)

**Data Analysis**

* ROC (Receiver Operating Characteristic) – Area under the curve (AUC)



**Red line = cut off value**

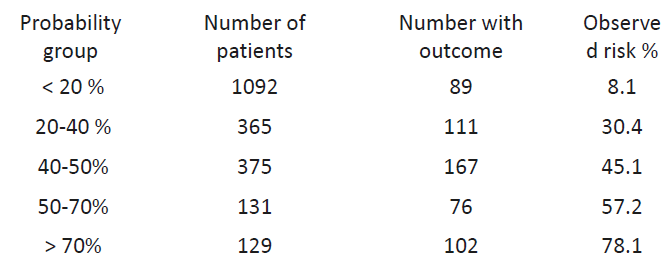


**Tossing coin AUC 0.5**

**The more tests the bigger AUC and the better prediction**

**Diagnostic score**

* Calculating by weighting information through testing
* Absolute risk score for each patient
* Build ROC curve
* Probability sorted by groups and check with observed values for risk of disease



After internal validation of the model -> external validation

**Reporting**

* Purpose of diagnostic research is ONLY to improve medical practice

**Prognostics**

* Purely clinical
* “Will I die?”
* What will happen if I don’t intervene

**Similar in steps to diagnostic research but:**

* What determinants predict future for patient
* not “guess what’s wrong with me” but “what will happen to me in the future?”
* Cohort with characteristics (determinants) to find relevant outcomes

**Risk = always over the TOTAL not over the person years**

**Odds= exposed/unexposed**