MSc Applied Data Science

Cuno S.P.M. Uiterwaal, MD PhD

Introduction to epidemiological research

November 13, 2020

The flow of epidemiologic research

1. Starting from research question
2. Design of the occurrence relation
3. Collection of data (empirical)
4. Data analysis and scientific interpretation

Essential descriptors of data collection

Time: t=0, t>0

open population = e.g. district in Utrecht where people can go in and out of the district/sample

Population closed or open

Analysis on all participants (census) or sample

Exposure (determinant) experimental or non-experimental

Main types of data collection

Cohorts

Cross-sections

(randomized) Trials

Case control studies

What is a cohort?

10 cohorts formed a Roman legion

Group of individuals:

- Followed up for specified period of time

Example:

- Users and non-users of antidepressants

Essential descriptors of cohort (non-experimental/observational)

Time: t>0

Population closed or open

Analysis on all participants (**census**)

Exposure (determinant): non-experimental

Cohort studies

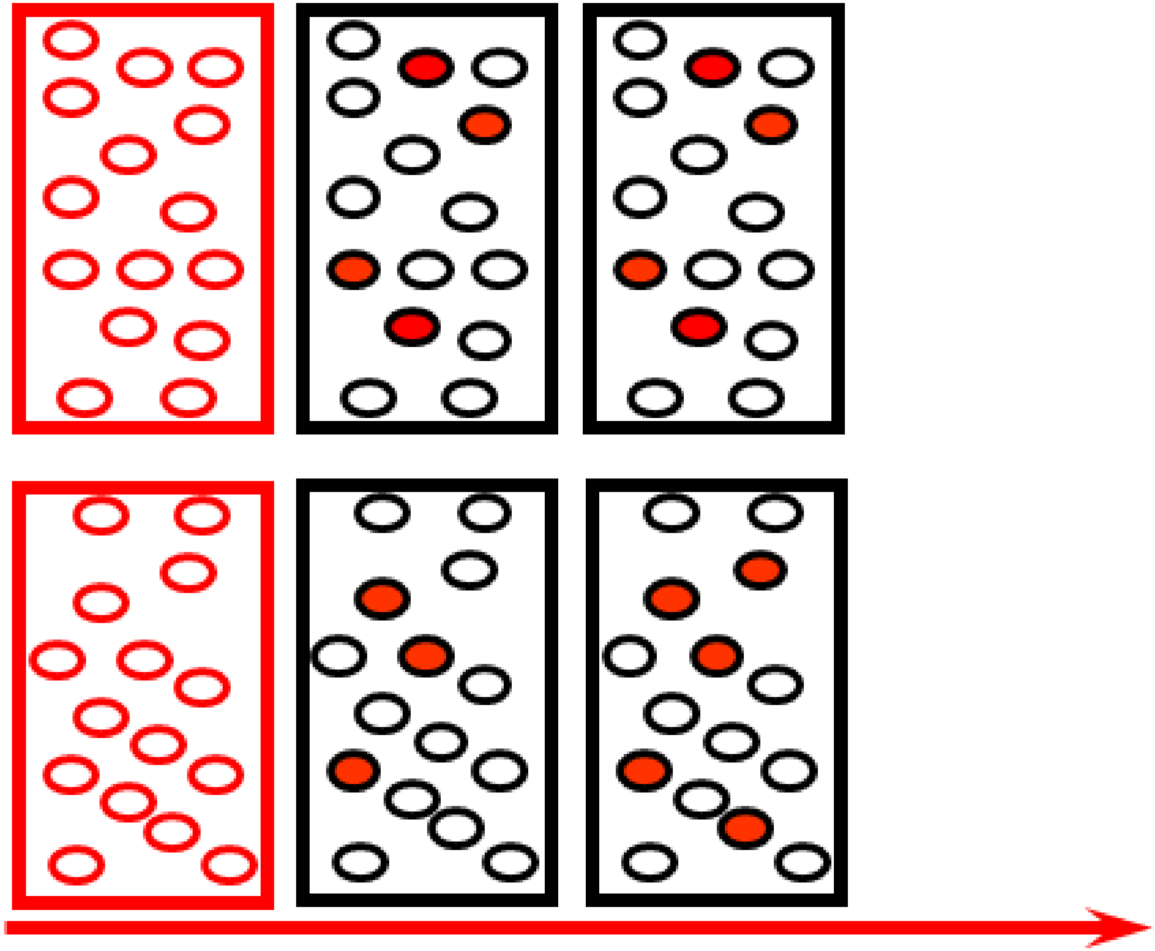
Purpose:

- Study if an exposure is associated with outcome(s)?
- Compare exposure to comparable non-exposure
- Estimate risk of outcome in exposed and unexposed (parts of) cohort (s)

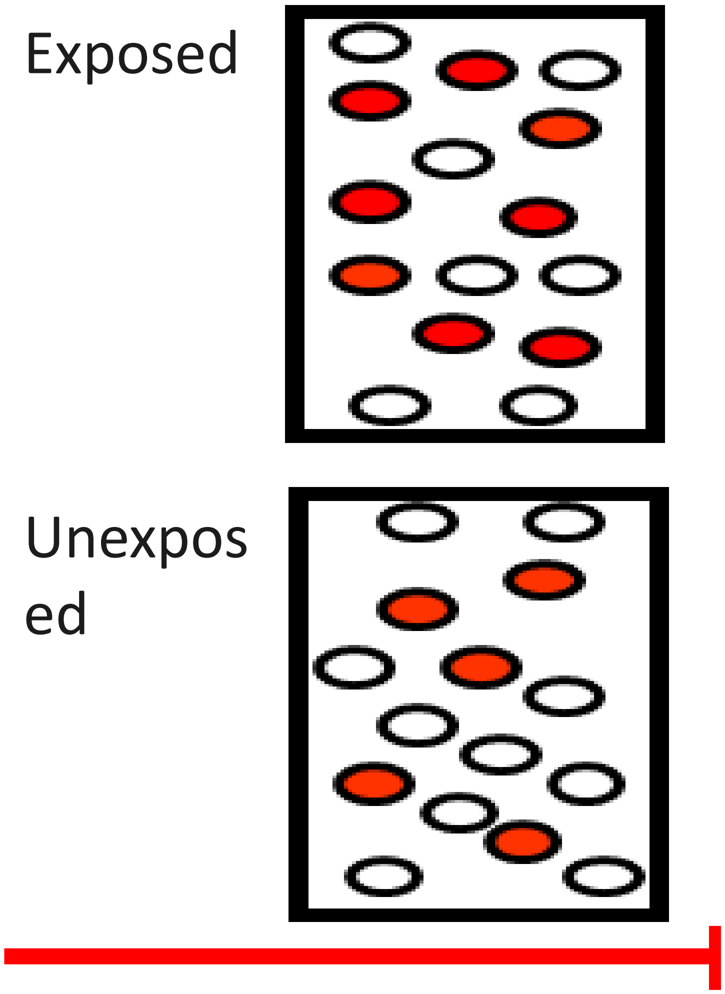
Cohort studies

Exposed →

Unexposed →



Cohort studies



Exposed

Unexposed

measure

Incidence among exposed

measure

Incidence among unexposed

Presentation of cohort data: our example

Anti depressants	Fracture	No fracture	Person - years	Total
Yes	70	1,930	15,930	2000
No	30	1,970	14,800	2000

comparing the measurements

Sources of bias in cohorts

- Confounding

not always and difficult to actively avoid

- Selection bias
- Observation bias

can happen in any type of study design
actively avoid in study

Selection bias in cohorts

1. Happens if subjects (exposed or unexposed) are specifically included based on a known relation between high exposure and increased risk of outcome.

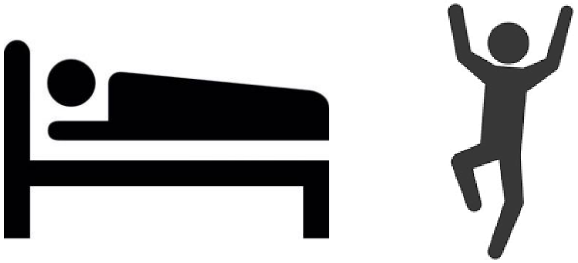
2. Happens if loss to follow-up has to do with the underlying occurrence relation

Remember: cannot be rectified in analysis!!!

to put right by adjustment or calculation, as an instrument or a course at sea.

How could selection bias have occurred in our example?

?



People may get lost to follow-up because they feel too sick or very energetic. This may be due to the (lack of) effects of the antidepressants. The related behaviour (sick in bed or active) can impact fracture risk.

Information bias in cohorts

Happens if observations/measurements of subjects in one group (e.g. exposed) are made different from observations/measurements in the other group (e.g. unexposed).

Remember: cannot be rectified in analysis!!!

How could information bias have occurred in our example?

?

Can happen with interpretation of images. With prior knowledge about use of antidepressants, one may search harder to find a fracture on the x-ray image.

e.g. measure fractures differently or more likely to find something -> information bias

Recipe: Cohort study

Identify group of:

- unexposed subjects
- exposed subjects

Measure incidence of disease

Compare incidence between exposed and unexposed group

Assure comparability as in trials

cohort studies

Limitations

- Latency period
- Loss to follow
- Large sample size
- Exposure can change
- Ethical considerations
- **Cost**
- Time consuming

Strengths

- Incidence in exposed and unexposed groups
- Suited for rare exposure
- **T>0 is clear**
- Less subject to biases: Outcome not known (prospective)

less prone to selection bias
- we don't know what the future brings
- main problem= loss of followup

first exposure then occurrence of outcome - scientific advantage

Essential descriptors of cross-section

Time: t=0

Population closed

Analysis on all participants (**census**)

Exposure (determinant): non-experimental

researches doesn't determine who gets antidepressants or not

Cross-sectional:

Strong points

- Quick, cheap

Limitations

cross-section e.g. when studying a certain gene with older group of people where a lot of the people with that gene may have already died

- Explain selection
- Explain consequences of t=0

Essential descriptors of trials

experimental cohort

Time: t>0 time positive because cohort - waiting for something to occure

Population closed

Analysis on all participants (**census**)

Exposure (determinant): experimental

Trials

trial = (usually) closed cohort

T>0

Experimental

Subjects intervened on with the purpose of learning about intervention effects

How deal with confounding?

In statistics, a confounder (also confounding variable, confounding factor, or lurking variable) is a variable that influences both the dependent variable and independent variable, causing a spurious association.

Does intervention work?

For rational intervention (therapy, revention, practice innovation) unequivocal and quantitative documentation of (relative) efficacy is necessary.

➔ *Question: How effective is intervention?*

Relative to:

- No intervention
- Alternative intervention(s)

Trials

Here limited discussion focused on comparability of ‘natural history’

Note that occurrence relation is

same as for noncoherrent

Outcome = f (intervention | **confounding**)

Randomization

Principle is random allocation to exposures (e.g. drug vs no drug).

Yields groups with comparable (~equal) mean levels of known **and unknown** determinants of outcome (comparable prognosis with respect to outcome)

randomization best way to get good evidence and results

… thus, differences in incidence are due to intervention, not incomparability of prognosis

differences due to intervention and not something else

Randomization

Validity threats:

selection bias (non-differential loss to follow-up)
information bias (blinding etc)

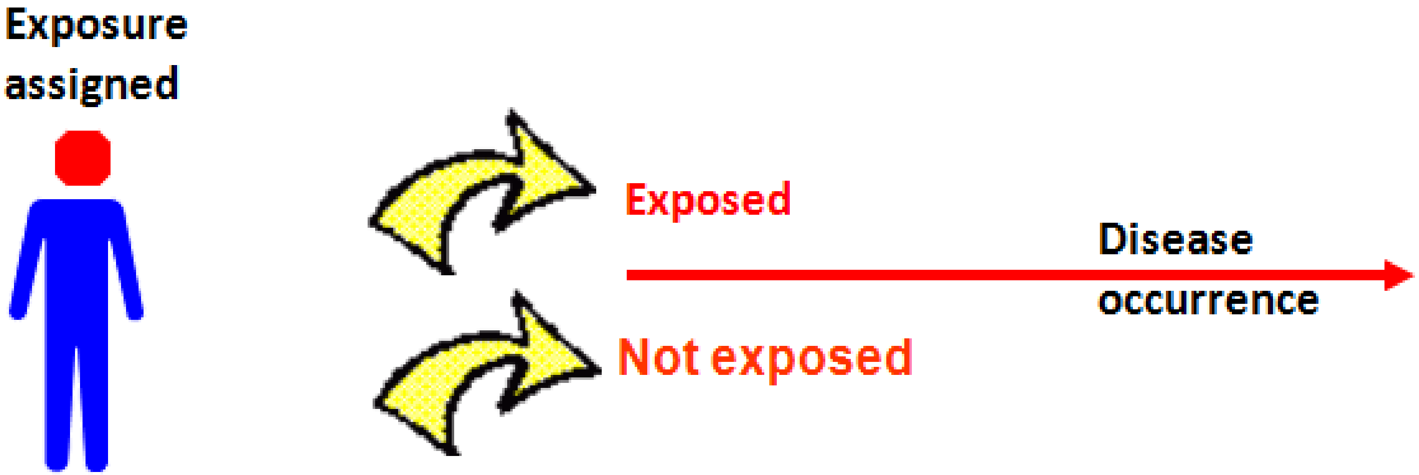
anything that can be done wrong in cohort can be done wrong in trial

like in non-experimental cohorts.

Randomization

Analysis of trial data is usually simple comparison (no need for adjustment)

Experiment



Unethical, unfeasible, unaffordable to perform experiments on people? “observational study”

Essential descriptors of case control

Time: t>0

Population open (sometimes closed/nested)

census or sampling (subsection of participants)

Analysis on **sample** of participants (as opposed to census)

Exposure (determinant): non-experimental

Case Control studies

Short definition: efficient cohort

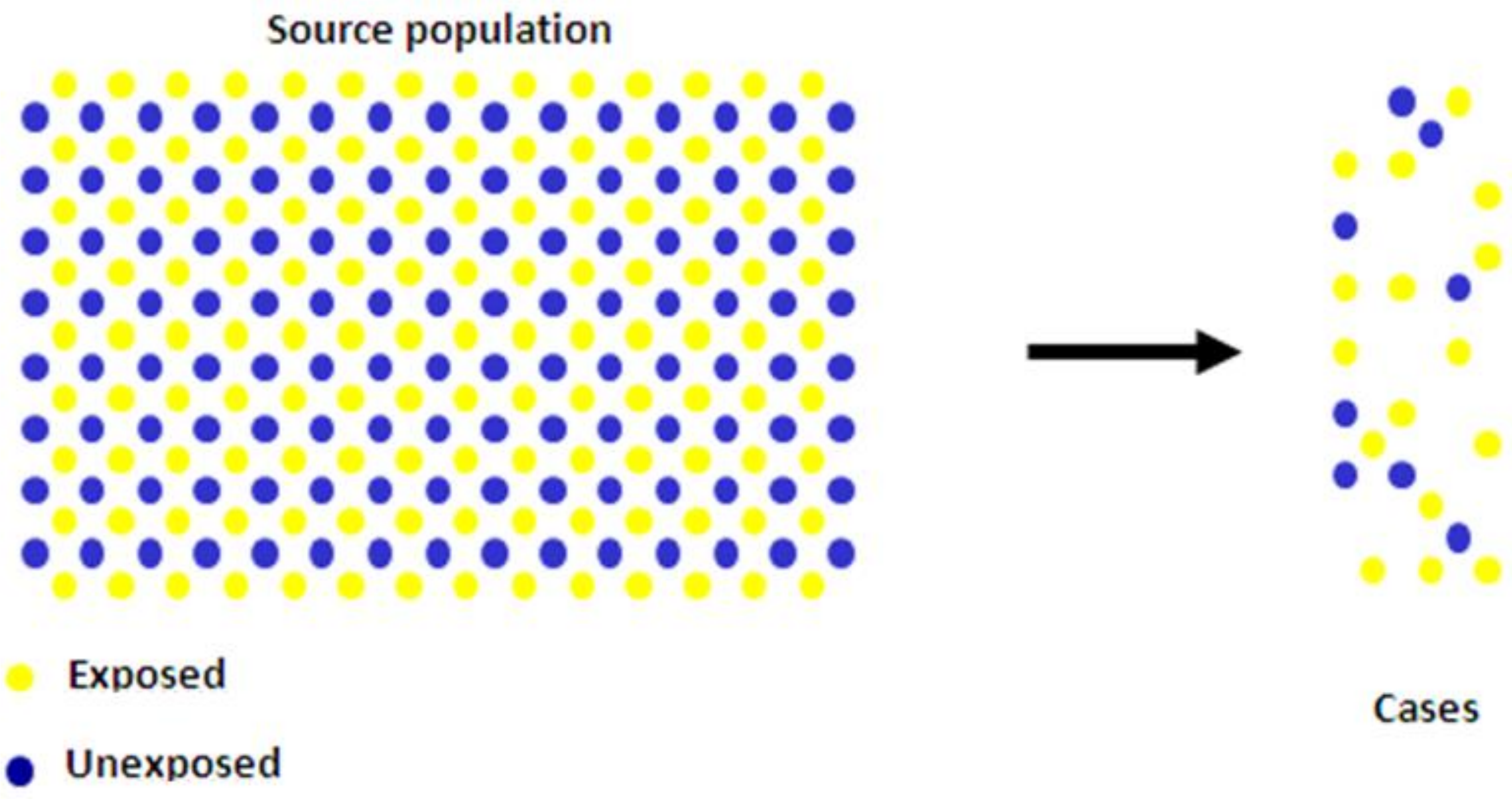
Synonyms:

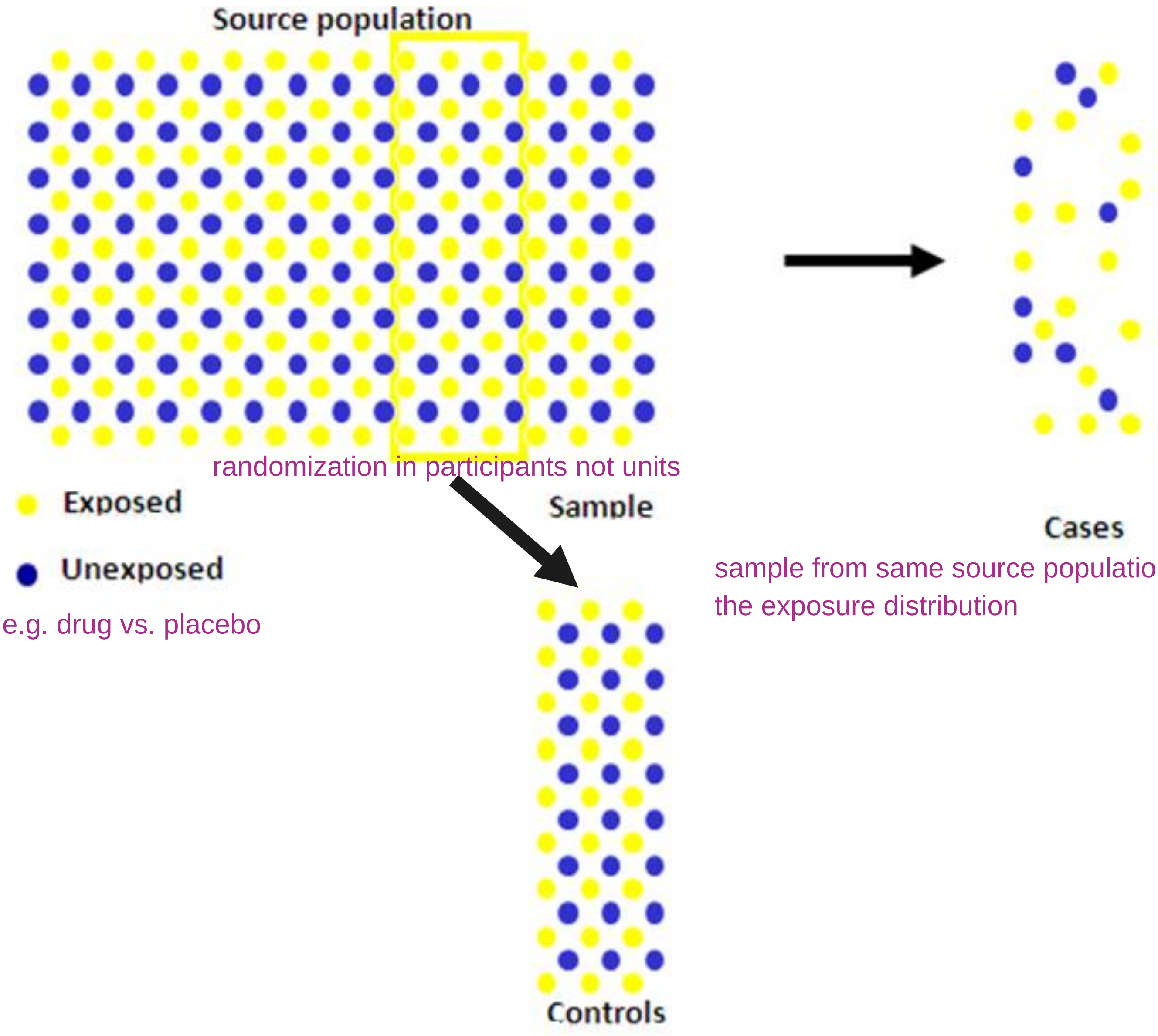
- case-referent, patient- control

Variants of case control:

- nested case-control, case-cohort, case cross-over

Principle of case control studies





randomization in participants not units

e.g. drug vs. placebo

sample from same source population to control the exposure distribution

Purpose of controls sampling

Provide an estimate of the exposure distribution in the source population, so e.g. our study

➔ How many people used antidepressants and how many did not?

Intuitively

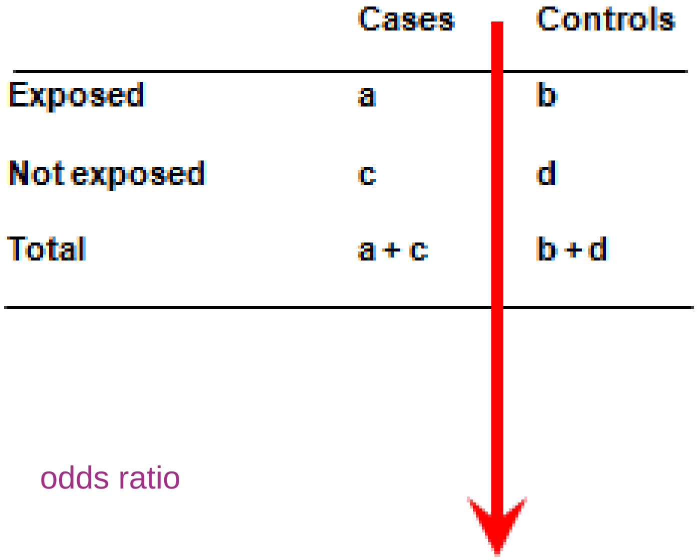
If the frequency of exposure is higher among cases than controls then the incidence rate will probably be higher among exposed than non exposed.

Distribution of cases and controls according to exposure in a case control study

	Cases	Controls
Exposed	a	b
Not exposed	c	d
Total	a + c	b + d
% exposed	a/(a+c)	b/(b+d)

Case control study

odds- prob. of something happening/ prob. of it not happening



	Cases	Controls
Exposed	a	b
Not exposed	c	d
Total	a + c	b + d

odds ratio

$$\text{OR} = \frac{a/c}{b/d} = ad / bc$$

Odds of exposure among cases =

- Probability to be exposed among cases / Probability to be unexposed among cases

$$\textbf{Odds Ecases} = \frac{a / (a+c)}{c / (a+c)} = a / c$$

Odds of exposure among controls =

- Probability to be exposed among controls / Probability to be unexposed among controls

$$\textbf{Odds Econtrols} = \frac{b/ (b+d)}{d/ (b+d)} = b / d$$

Case control study: our example

Our RQ: is the occurrence of (non)osteoporotic fractures related to prior use of antidepressants (in humans)?

Case control study: our example

We have collected cases (people with fractures) over a certain period in time

We have sampled controls (from the source population that cases came from)

We measure prior use of antidepressants in both cases and controls

Case control study: our example

Antidepressants	fractures		
	yes	no	
Yes			
No			
	100	100	200

Case control study: our example

Antidepressants	fractures		
	Cases (yes)	Controls (no)	
Yes	67 **(a)**	47 **(b)**	114
No	33 **(c)**	53 **(d)**	86
	100	100	200

Exposure distribution in controls: 47%

Exposure distribution in cases: 67%

Case control study: our example

Antidepressants	fractures		
	yes	no	
Yes	67 **(a)**	47 **(b)**	114
No	33 **(c)**	53 **(d)**	86
	100	100	200

Exposure odds ratio (OR) =

a*d/c*b = 67*53 / 33*47 =

3551/1551 = 2.3

Remember Incidence Rate Ratio (IRR): our cohort study

Anti depressants	Fracture	Person-years
Yes	70	15,930
No	30	14,800

$$= \frac{70/15{,}930\text{py}}{30/14{,}800\text{py}} = 2.2$$

Case control study: our example

The odds ratio = 2.3 odds ratio close to IRR

In our cohort study on the same RQ: IRR = 2.2

The odds ratio = IRR, if and only if the case control study is executed correctly (time is essential): for details see dedicated study design course

Sources of bias in case control studies

- Confounding
- Selection bias
- Observation bias

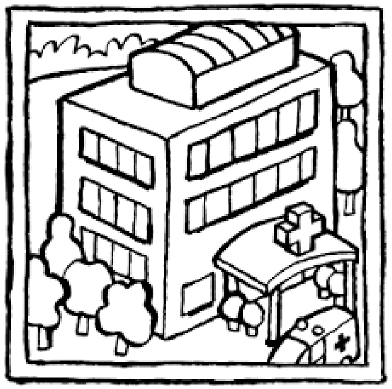
Selection bias in case control studies

Happens if somehow specifically those cases are selected that are exposed

Remember: cannot be rectified in analysis!!!

How could selection bias have occurred in our example?

?



The GP may refer cases to a specialized hospital based on suspicion of a certain fracture based on knowledge of the exposure (use of antidepressants). Controls are selected from general population.

Information bias in case control studies

Happens if exposure measurement in cases is performed different from controls

Remember: cannot be rectified in analysis!!!

How could information bias have occurred in our example?

?

People with fractures may search for reasons, and therefore remember their exposure, in this case, antidepressants use, better .

What defines the choice for case control versus cohort study?

Costs = number of observations (personnel) and time

Nature

Efficiency

case control studies

Advantages

- **Rare diseases**
- Several exposures
- Long latency
- Rapidity
- **Low cost**
- Small sample size
- No ethical problem

Limitations

- Usually no absolute rates/risks
- **Less suitable for rare exposures**
- More prone to bias
- Often performed “quick and dirty”

UMC Utrecht

Diagnostic research

Research into the establishment of presence or absence of diseases

Content

- Diagnostics in practice
  - Example case

- Scientific diagnostic research
  - Study design proper
  - Data-analysis
  - Reporting

Diagnostics in practice

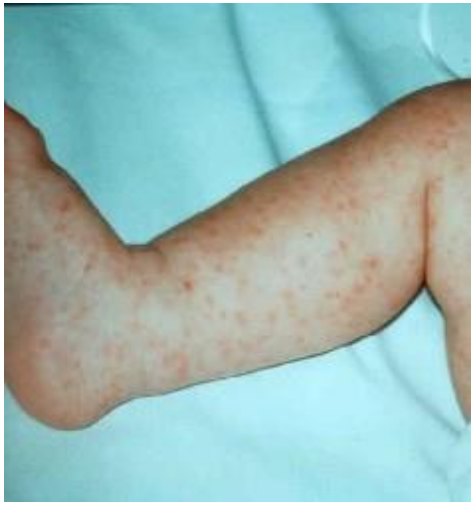
- Diagnostics always start with a patient that has compaint(s)/symptom(s)

- Casus: neck stiffness
  - Child, 2 yrs old, comes to ER with parents
  - Child has very stiff neck

- What is challenge to the physician?
  - Quick and efficient estimation of the correct diagnosis
    - Basis for medical policy making
    - Determines choice of therapy
    - Informs about prognosis

Diagnostics in practice

- Differential diagnoses (DD)
  - Bacterial meningitis
  - Viral meningitis
  - Pneumonia
  - ENT infection
  - Other (e.g. myalgia)



differential diagnosis- what is most likely/most dangerous on top and then alternative explanations (maybe no disease at at all)

- What is the most important diagnosis? Which one does the physician absolutely not want to miss?
  - Bacterial meningitis (BM) very dangerous
  - If missed: often fatal

Diagnostics in practice

- Suppose: 20% of all children at the ER which show neck stiffness have BM
  - 20% with BM in that population = prevalence
  - = prior risk

with no other knowledge the baseline of having BM is 20%

- What would you decide for this child (case)?

Diagnostics in practice

- Decision for child (case)
  - Prior-risk too low to start treatment
  - Prior-risk too high to send home
- Decision: reduce uncertainty → diagnostics
- What is the best test? standard/reference test
- Best test: lumbar puncture (culture of liquor)

Diagnostics in practice

- Gold standard
  - Real disease status; ‘truth’
    - Never 24 carat
  - Reference- / standard test puncture
  - Decisive test in case of doubt

- Do reference test in all children with neck stiffness at ER?

Diagnostics in practice

- Reference test in all? can be dangerous
  - Unethical → too burdening/risky
  - Inefficiency → too expensive
  - Not perform unnecessarily

- Then how establish the probability of presence of disease and what would be ideal?

Diagnostics in practice

- How?

- Simpler diagnostics (than lumbar puncture):
  - Usually medical history, physical examination, simple lab tests, imaging, etc.
  - Ideal: diagnosis without reference test

    hierarchy from less to more invasive method

- Diagnostic proces in practice:
  - Stepwize: less → more invasive
  - No diagnosis based on 1 single test
  - Each item: separate test

Diagnostiek in de praktijk

- Suppose: after medical history & physical examination, 10% chance for BM

- Chance for disease given test results = posterior-risk

  after having done diagnostic test

- The bigger the difference prior - posterior risk, the more the diagnostic value of tests

- We decide for this child that the risk of BM is too high to send it home → next step?

Diagnostics in practice

- Next step:
- Further examinations/tests, e.g.:
  - Blood tests (indicators for infection, etc.)

Diagnostics in practice

- Suppose: 1% posterior-risk after medical history, physical examination and simple blood tests
  - ➢ posterior risk low enough to send home

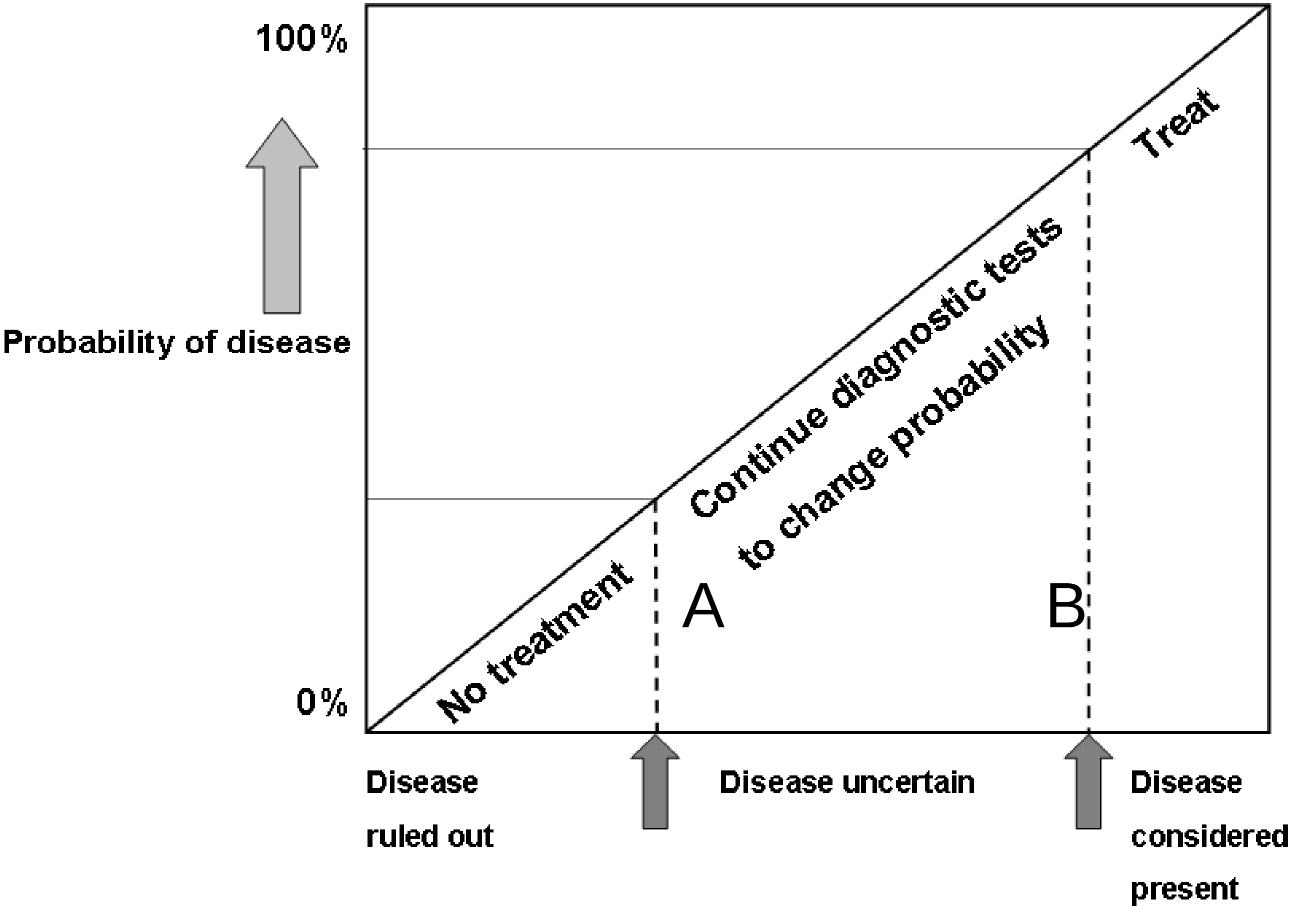
    keeping close eye on child but able to go home

- Ideal diagnostic process: simple tests take posterior risk to 0 or 100% (without doing reference test)

- Usually physician procedes testing until sufficient certainty (sufficient approach towards 0 or 100%)

- Choice when still uncertain: depends on prognosis disease if not treated and costs/risks of treatment

Action thresholds



Diagnostics in practice

- So what is diagnostics in practice?
  - Estimation of risk/chance of presence of disease based on patient's test results
- Why not all possible tests?
  - Too burdening for the child
  - Unnecessary: different tests often provide same info
  - Note: in practice very often more testing than strictly needed!
- What diagnostics is really necessary? → Scientific diagnostic research

Scientific diagnostic research

- Which tests really contribute to estimation of risk (disease presence)?
- This needs to serve practice → so follow practice

Study design

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

Research question/occurrence relation

- Which simple, safe, and cheap tests allow us to estimate the presence/absence of disease?

- What are determinants of presence/absence disease?

  everything that can predict the presence/absence

- Determinant-outcome relation:
  - Chance for disease as a function of test results
  - outcome = chance for disease = % = prevalence
  - Determinants = test results

Occurrence relation

example for diagnostic study - prediction not explanation - combining info that predicts presence of disease

Casus

- %BM = f(age, sex, fevor, indicators blood infection, etc)

Domain

- For whom → domain, generalization
  - = type of patients with particular symptom/ complaint + setting important as prevalence differs "typical patient" changes in setting
  - Study population = 1 sample from domain

- Case:
  - suspicion crucial in study population
  - All children (e.g. in Western societies) suspected of having disease (BM) based on neck stiffnss in general hospitals (setting)

Study population

- Case:
  - 200 children with neck stiffness in 2012 at ER Utrecht MC

neck stiffness and suspicion of BM

Determinants

- = tests to be examined

- Diagnostic determinants:
  - All possibly relevant tests (in domain)

- Case:
  - Items medical history, physical examination and laboratory tests (blood, urine)

Measure of outcome

- ‘True’ presence/absence disease
  - = Diagnostic outcome
  - = Result reference test (gold standard)

- Note: reference is not perfect but at that time best possible test in that setting

- Case:
  - Positive culture of liquor (= liquid surrounding brain tissue, spinal cord tissue) tab the brain - bacteria in brain is real proof but dangerous - outcome measure

so technically we would have to do this "gold standard" test with all children

Size: ‘1 in 10’-rule

- ‘1 in 10’-rule determines size of study population
  - Minimum of 10 participants observed in smallest group to be predicted

ina any study cohort (rule of thumb) you would need 10 cases for 1 determinant as predictor

- Smallest group: almost always cases (having disease)

- 1 determinant → minimum of 10 cases
- 2 determinanten → minimum of 20 cases

Measuring determinants/outcome

- Determinants
  - No knowledge (blinding) of outcome
  - Same measurements in research as in medical practice
    - Do not measure more accurate/precise than possible in medical practice setting (overestimates diagnostic value)

- Outcome measured in all children
  - In everyone
  - Estimate blinded to determinants
  - Use best possible test (in our case lumbar puncture)

to be as close to the truth as possible

Study design proper: descriptors

- Observational = non-experimental

- Descriptive
  - Descriptive= **not causal, confounding not an issue**
  - Is only determinant actually (helps) predict
  - no hypothesis mechanisms Determinant-Outcome

    end product = risk of diseases being present (descriptive!)

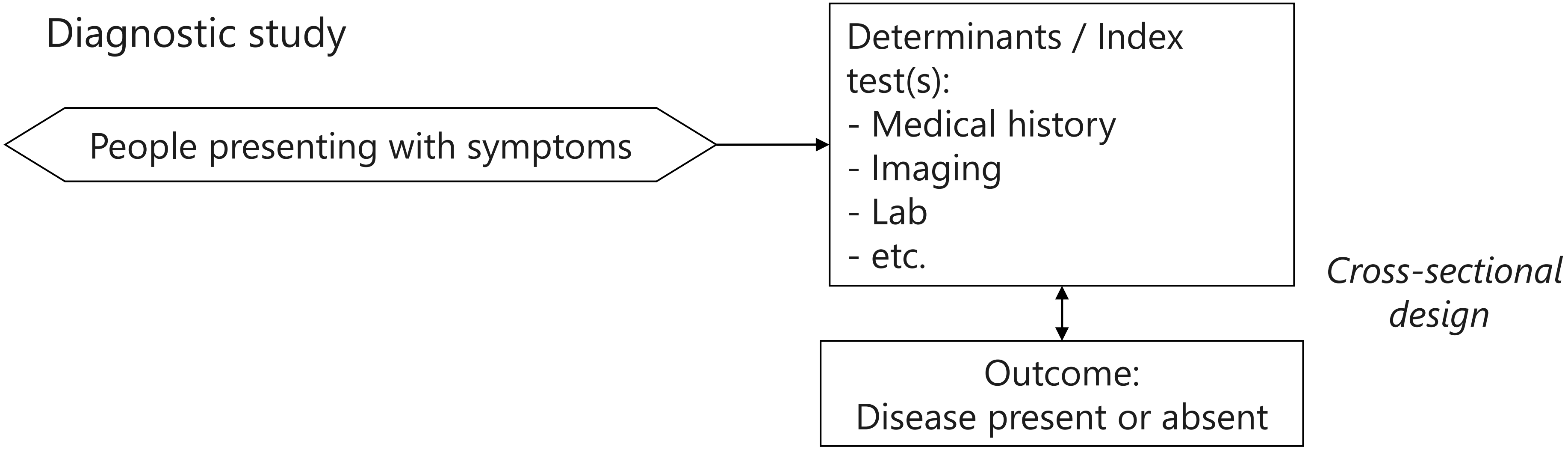
- >1 determinant (diagn research multivariable by nature)

  almost always more that one test - multi variable by nature

Study design proper

- Cross-section
  - Determinants and outcome measurements “same” point in time

Diagnostic study



People presenting with symptoms

Determinants / Index test(s):
- Medical history
- Imaging
- Lab
- etc.

Outcome:
Disease present or absent

Cross-sectional design

T=0

0 because cross-section reflects medical practice that we are in

Data-analysis

- After data collection, per patient
  - Values of determinants (test results)
  - Diagnostic outcome (reference test)

all test results &

all outcomes in dataset

Data-analysis

- Data analysis: 3 steps
  1) Estimate prior risk (without test results)
  2) Compare each index test result with reference test result
     = univariable 2 x2 table and e.g. calculating odds ratio
  3) Compare combinations of index test results with reference test result
     = multivariable (via model, usually logistic regression)
     - Adhere model building to diagnostic order in medical practice
     - Determine added value of index test results as compared to already collected, previously collected index test results

Data-analysis

Casus:

- Data scientific research available:
- 200 patients with neck stiffness at ER
  - Liquor culture positive (BM+) n=40
  - Liquor culture negative (BM-) n=160

Step 1: A priori risk (prevalence) for BM?

Data-analysis

Step 2: Analysis per determinant (univariable)

- Sex (m/f); neck stiffness (y/n); fevor > 38^{0}C (y/n)
- 2 by 2 table --> e.g. fevor > 38^{0}C

- Perfect diagnostic test doesn't happen - never perfect
  - False positive = 0
  - False negative = 0

		BM+	BM-	tot.
Fevor > 38^{0}C	Yes (+)	20	90	110
	No (-)	20	70	90
		40	160	200

Data-analysis: reading 2 by 2 table

in the end predicitve values depending on the tests are what matters not about the certainty of the disease being present or absent

Horizontal

- chance BM+ if fevor+ = 20/110 = 18%

  PV+ = A / A + B
- chance BM – if fevor- = 70/90 = 78%

  PW- = D / C + D

Vertical

- Chance fever+ if BM+ = 20/40 = 50%

  SE = A / A + C
- Chance fever koorts- if BM- = 70/160 = 44%

  SP = D / B + D

	Gouden standaard		
	BM+	BM–	
Koorts +	20 TP A	90 B FP	110
Koorts –	FN C 20	D TN 70	90
	40	160	

- Which of these figures are most useful in medical practice (PW+ and PW- or SE and SP)?

Data-analysis: combine determinants

- In practice no diagnosis based on 1 test
  - Do combined tests distinguish the sick/healthy?
  - Method: statistical modeling (usually logistic regression)

- Moreover: diagnostic process is hierarchical
- (simple --> burdening/costly) --> almost always starts with medical history taking --> see case

Data-analysis

Case:

- model including all medical history items, physical examination tests, + sex + age + fevor + pain etc
  - %BM = *ƒ* (sex, age, tests, ……..., ……..., etc)

- Statistical model to be interpreted as 1 test

- Quantify diagnostic value of model(s) using area under ROC curve (Receiver Operating Characteristic =Area Under Curve (AUC))

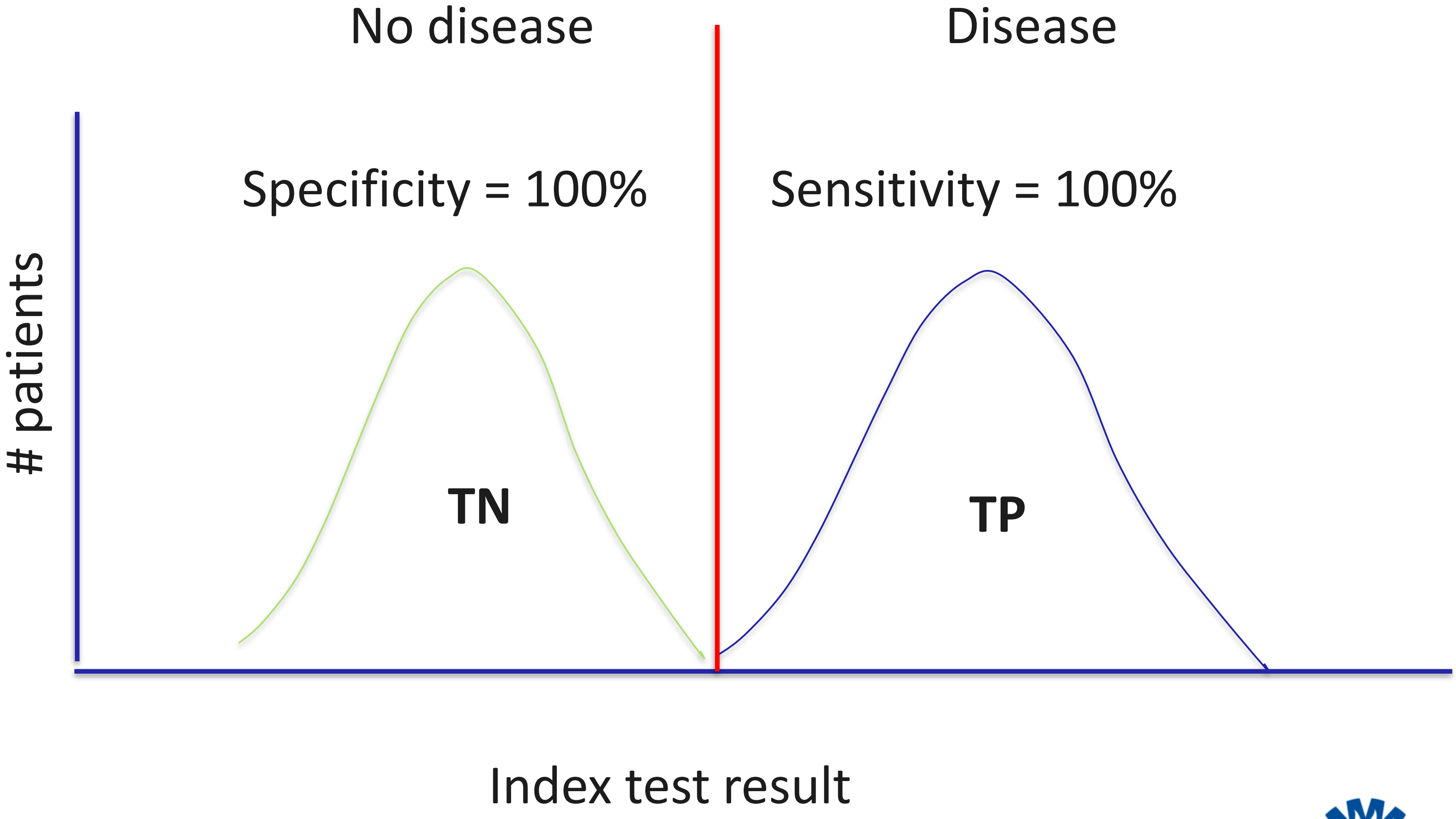
ROC-curve

- axes:
  - Y axis: sensitivity (true positivesTP)
  - X axis: 1 – specificity (false positives FP)

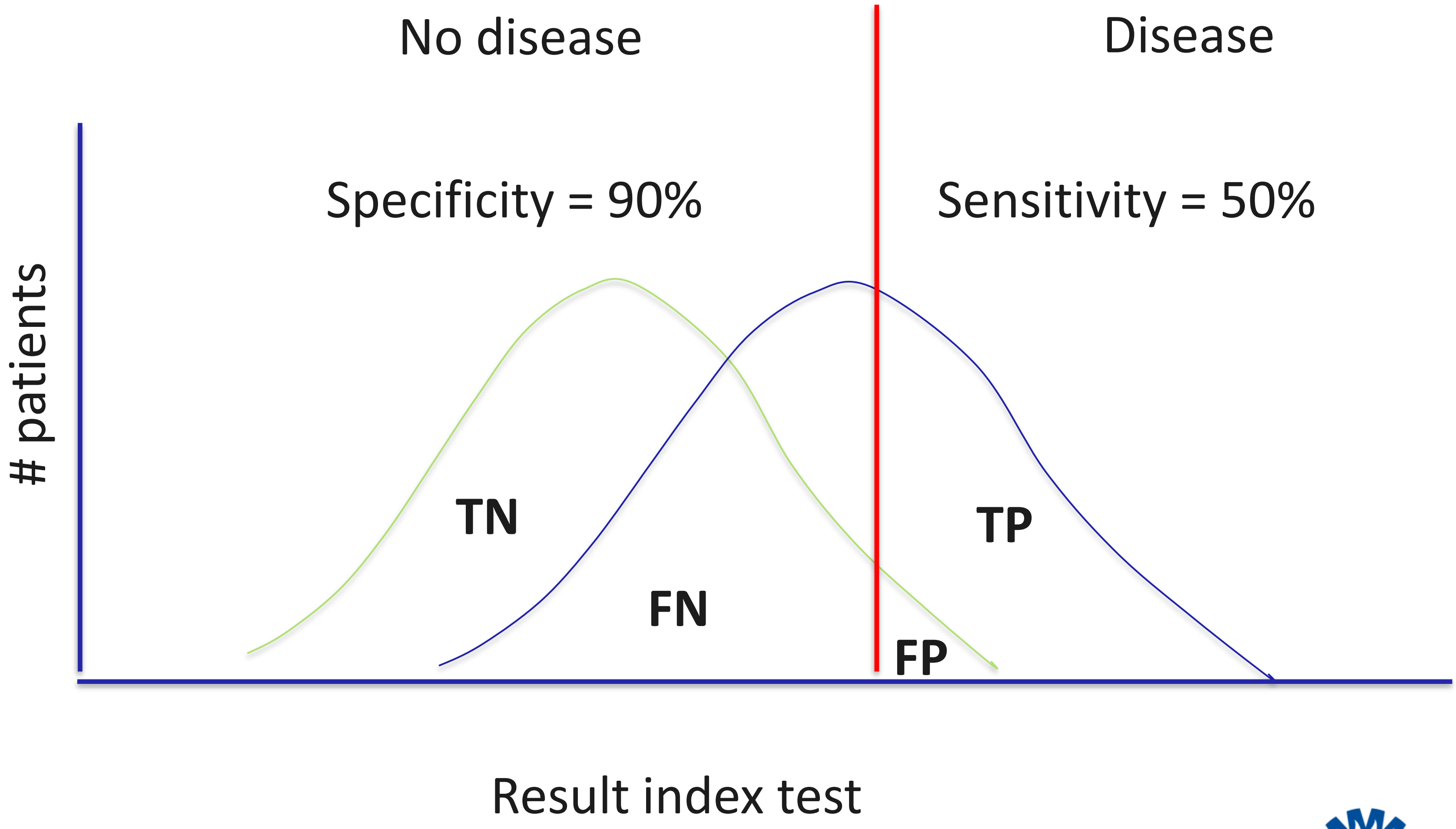
- Various points on curve:
  - Cut off values of test
  - Other cut off → different TP and FP

- Same holds for statistical model

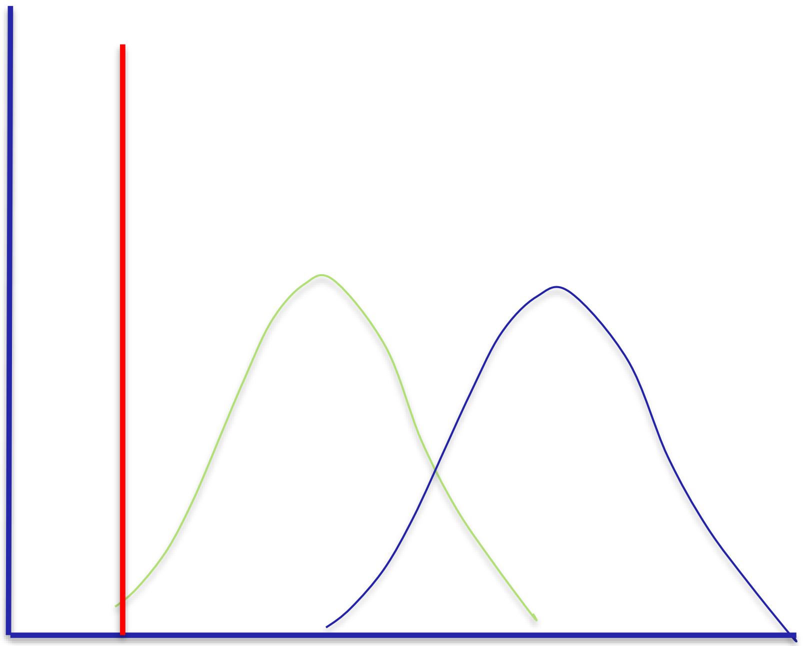
Cut off value index test

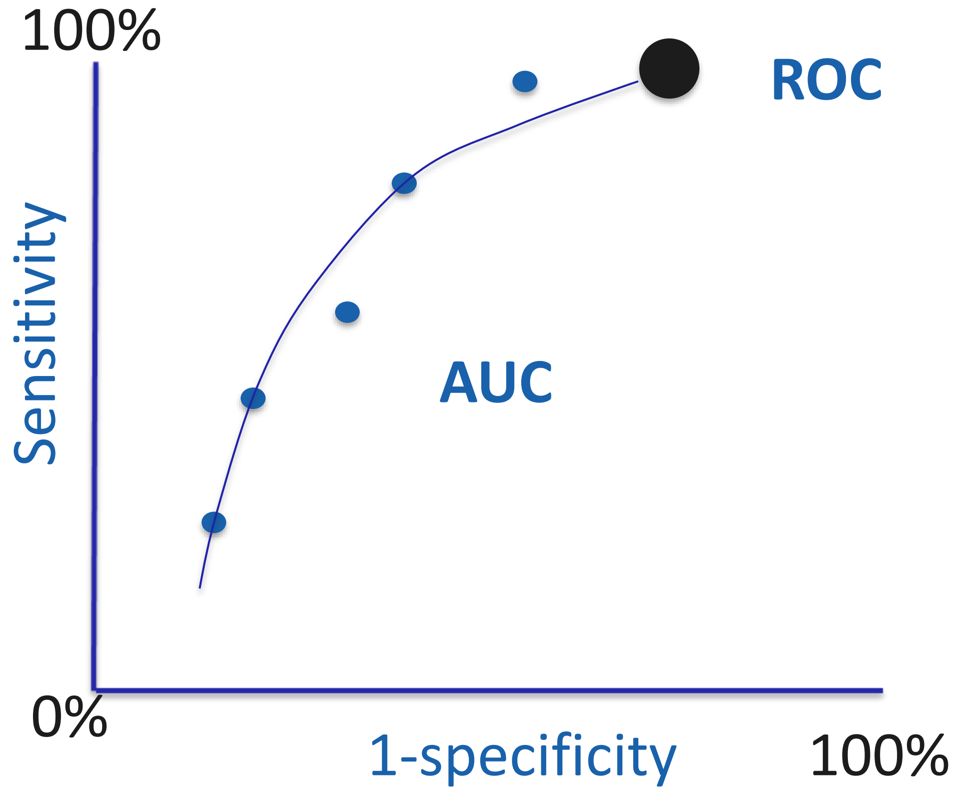


Cut off value index test

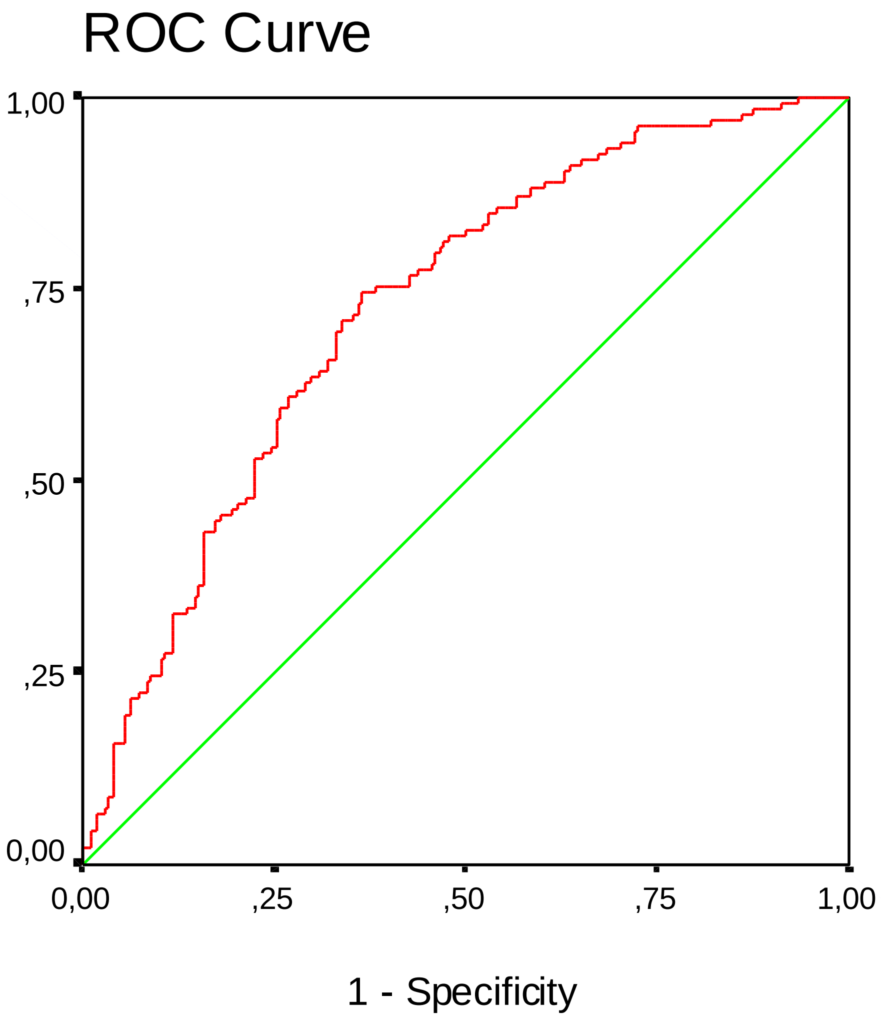


ROC-curve





Data-analysis

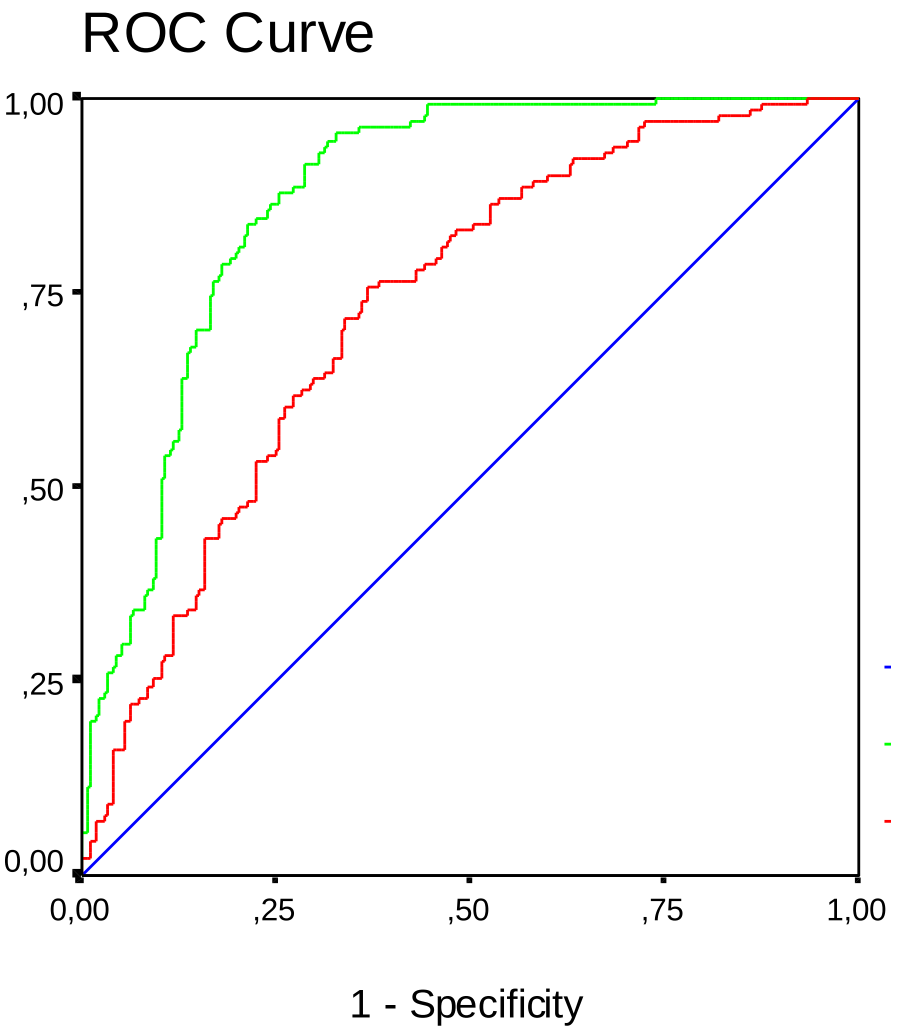


Data-analysis

- Toss a coin
  → *AUC = 0.50*

- *Model 1:*
  med history +
  phys exam
  (5 extra tests)
  → *AUC = 0.72*

- *Model 2:*
  med history +
  3 blood tests
  → *AUC = 0.90*



An easily applicable diagnostic score

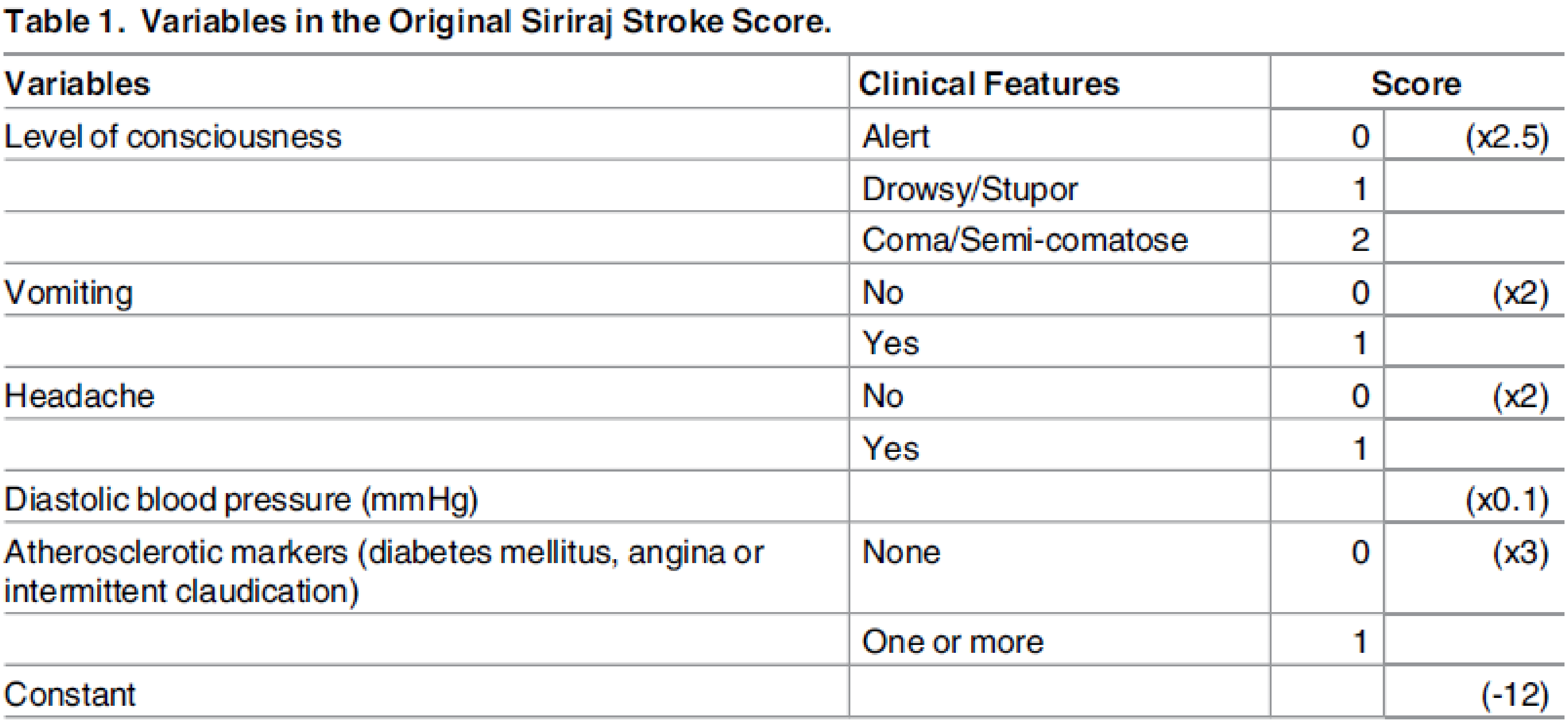


Table 1. Variables in the Original Siriraj Stroke Score.

Variables	Clinical Features	Score	
Level of consciousness	Alert	0	(x2.5)
	Drowsy/Stupor	1	
	Coma/Semi-comatose	2	
Vomiting	No	0	(x2)
	Yes	1	
Headache	No	0	(x2)
	Yes	1	
Diastolic blood pressure (mmHg)			(x0.1)
Atherosclerotic markers (diabetes mellitus, angina or intermittent claudication)	None	0	(x3)
	One or more	1	
Constant			(-12)

Kelder et al. Circulation 2011; 124:2865-

For each patient we estimate the absolute risk

- Risk Score: Absolute risk of bleeding for each patient
  - $1 / 1+ e^{-(b0 + (b1*LVEF) + (b2*coumarin) + (b3*type) + (b4*elective))}$

- Next is the question whether this approach does lead to a good differentiation between those who develop and do not develop the disease.

- ROC curve (total discriminative power)
  - Here the model is looked upon as a test.
  - For each value of the risk score a sensitivity (true positive rate) and 1 - specificity (false positive rate) is estimated.

Calibration of the model

- Create groups of patients based on predicted risk (probability) with the risk score

- Compare observed outcome with expected risks

Probability group	Number of patients	Number with outcome	Observed risk %
< 20 %	1092	89	8.1
20-40 %	365	111	30.4
40-50%	375	167	45.1
50-70%	131	76	57.2
> 70%	129	102	78.1

4. Results easily applicable in clinical practice

- diagnostic algorithm / rule

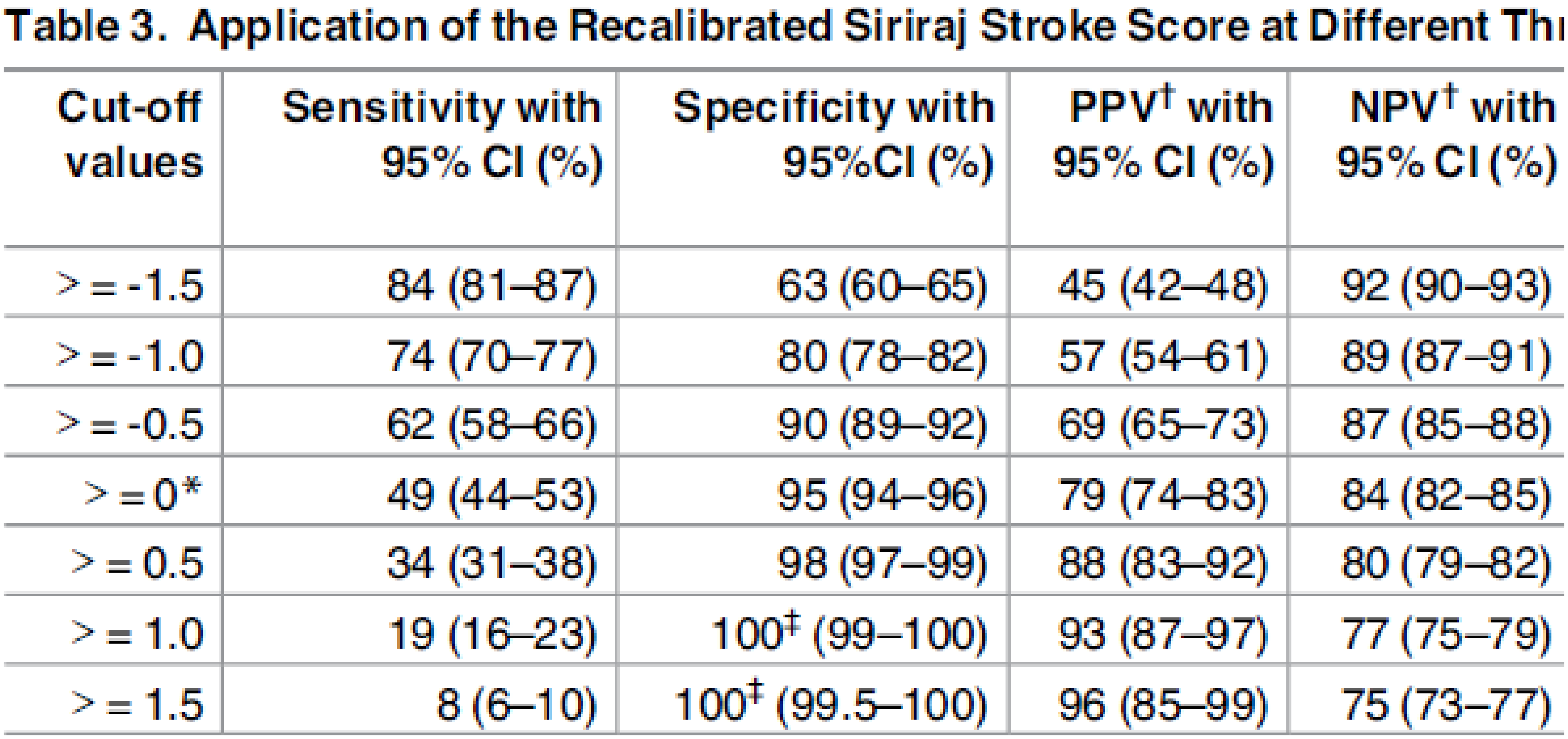
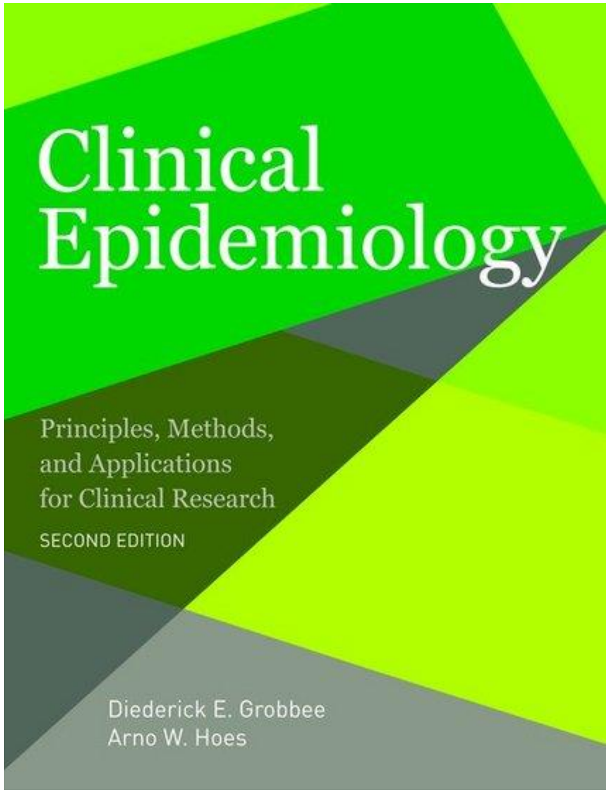


Table 3. Application of the Recalibrated Siriraj Stroke Score at Different Th

Cut-off values	Sensitivity with 95% CI (%)	Specificity with 95%CI (%)	PPV† with 95% CI (%)	NPV† with 95% CI (%)
> = -1.5	84 (81–87)	63 (60–65)	45 (42–48)	92 (90–93)
> = -1.0	74 (70–77)	80 (78–82)	57 (54–61)	89 (87–91)
> = -0.5	62 (58–66)	90 (89–92)	69 (65–73)	87 (85–88)
> = 0*	49 (44–53)	95 (94–96)	79 (74–83)	84 (82–85)
> = 0.5	34 (31–38)	98 (97–99)	88 (83–92)	80 (79–82)
> = 1.0	19 (16–23)	100‡ (99–100)	93 (87–97)	77 (75–79)
> = 1.5	8 (6–10)	100‡ (99.5–100)	96 (85–99)	75 (73–77)

Next,

- Internal validation of the model
  - Correction for ‘overfitting’
- External validation of the model
  - Apply the prediction rule in other datasets
  - Does the model work for all China, India, Brazil, The Netherlands



Reporting

- Note: diagnostic research serves one purpose only, improve medical practice (doctors, patients)

- Report practically useful findings, always provide predictive values

- 'STAndards for Reporting of Diagnostic accuracy' (STARD-2015)

Prognostics

A man of 65 yrs old just had a heart attack. Physician takes medical history, physical examination and takes various other tests. Treatment is streptokinase. Patient's 'high sensitive C-Reactive-Protein' is elevated.

What are from patient's perspective important prognostic research questions?

- Is CRP an independent predictor of mortality after 30 days?
- What is risk of death within 30 days?

Prognostics

- Patient has disease diagnosed
  - What happens in future?
- What determinants predict natural history (future)
- think of patient characteristics, disease characteristics, test results etc
- Outcome: death, disease recurrence, complications, quality of life etc. Outcomes that are relevant to patients
- Starting point / moment of prediction: after establishing diagnosis

Motive

- Reliable info about future
  - Patients and doctors

- Choice of treatment
  - If patient has high risk (predicted) of bad outcome

Prognostic research

- Cohort of diagnosed patients T > 0 (domain patients like these in this particular setting)

- I death 30 days = f(patient chars, med hist, phys exam, lab +other tests)

- Choice of det’s: much like diagnosis; simple – complicated; cheap - expensive; easy – invasive, etc

- Follow order of clinical practice

Prognostic research

- Analytical steps much like in diagnostic research
- But
- Different models because cohort (Cox regression etc)

Reporting of prognostic research

- Same considerations as with diagnostic research

- Dont provide as end product: relative risks, hazard ratios, odds ratios, average 5 year survivals

but

- Do provide as end product: individual estimates of absolute risks for patient relevant endpoints