**Epidemiology: week 2**

**Practical:**

* Missind Data practical I
* Missing indicator method practical

**Reading:**

* Groenwold et al 2011 Dealing with missing outcome data in RCT and observational studies
* Janssen 2010 Missing covariate data in medical research
* Liublinska 2011 Re -dealing with missing outcome data in RCTs and observational studies
* Naaktgeboren e.a. - 2016 - Anticipating missing reference standard data when planning diagnostic studies

**Summary:**

**Lecture 1: Types of missing data**

**NAs occur in any study**

* Retrospective & prospective
* Missings = problems for precision (stat. power) and validity (bias)

**Epidemiological studies**

* Most use complete case analysis -> ignoring NAs by kicking all cases with NA out

**Type of missing values**

**MCAR** = missing completely at random

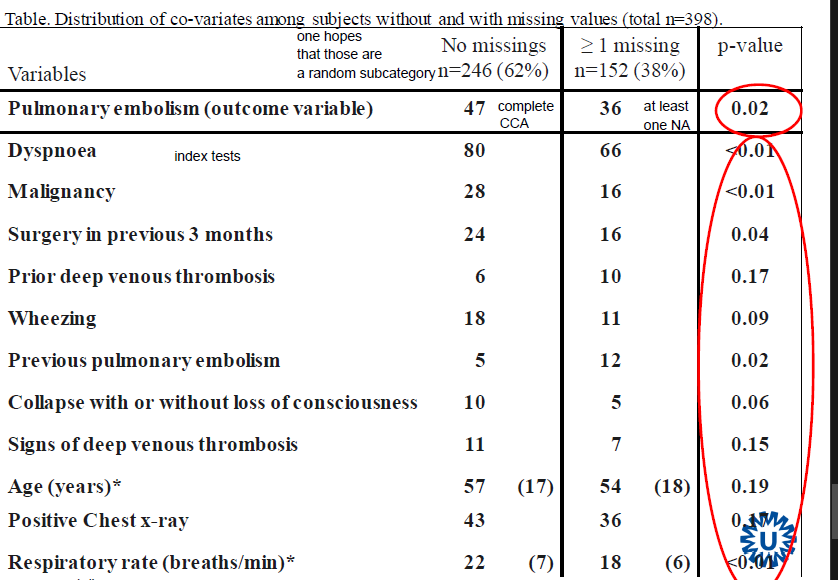
* Probability that observation is missing only dependent on change
* Probability that observation is missing is constant for all subjects and not related to any other characteristics
* Analytical methods to handle NAs unbiased
* Often unrealistic

**MAR** = missing at random

* Probability that observation is missing fully dependent on other observed values
* Most analytical methods work well (unbiased)

**MNAR** = missing not at random

* Probability that observation is missing dependent on unobserved values themselves
* Missingness is dependent on the unknown variable itself
* Always problematic
* E.g. people with higher income leave out the info about income



* if 246 would have exactly same values as the 152 it would be MAR and we could use CCA
* Values differ between group with missing values and group without missing values
* we have to assume that the difference is dependent on the missingness
* variables are associated with missingness
* Missing data NOT MCAR because subset is unequal in observed characteristics
* If missingness related to observed characteristics – the observed characteristics can be used to estimate/predict NA
* Could still be MNAR or MAR
  + No test for MNAR but we can reduce the MNAR-part by including observed characteristics

**Lecture 2: Methods to deal with missing data**

Methods to handle missing data

* Complete case analysis (CC)

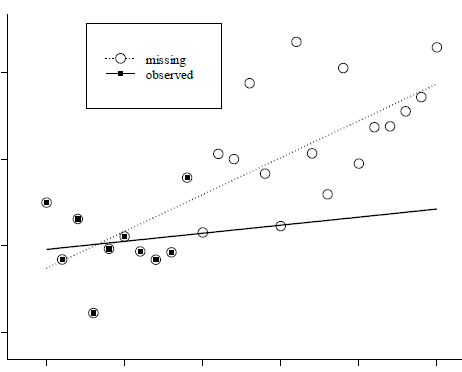
= Available case analysis (AC)

* Missing indicator method (See exercise)
* Overall mean/median imputation
* Subgroup mean/median imputation
* Single (multivariable) regression based imputation
* Multiple regression based imputation

**Complete Case Analysis**

* Completely observed X & Y
* If MCAR less precise but unbiased -> valid results
* If MAR less precise and biased -> invalid results
* Redefines study sample

-> changes domain

e.g. Study population of all **completely observed** patients/subjects

<- Not MCAR because missing values are not equally distributed as observed values – all NAs are in upper area and change the regression line

**Available Case**

Main difference to CCA: complete case for each sub analysis – instead of deleting entire row when NA only delete the missing values for the analyzed column

* Unbiased if MCAR
* Biased if MAR
* Less precise but more efficient because sub-analysis uses more records/data
* Problem for prediction modeling
  + Number of cases differs for each model
  + Comparison on different subjects (included/excluded depending on NA)

**Imputation (replacement) methods**

* possible under MAR (dependent on observed values)
* use observed values to gather information on missingness and predict NAs

**Overall mean/median imputation**

* overall mean/median if variable from observed values is imputed for NAs
* all imputations have same value
  + distributions of variable have bigger overlap -> bias
  + **standard deviation** appears lower than reality
  + **standard error** underestimated -> appears more significant (small p-value)

**Subgroups mean/median imputation**

* relevant subgroups for variable with missing values are defined
* estimate mean/median for each group
* imputation of mean/median for NA in each subgroup
  + less (but still) biased **standard error**
  + requires **categorization of continuous variables** -> information loss

Single regression based imputation & multiple regression based imputation in week 3