Data management for epidemiological analysis in R

Bendix Carstensen Steno Diabetes Center Copenhagen
Herlev, Denmark
http://BendixCarstensen.com

Paula Bracco Post Graduate Program in Epidemiology
Federal University of Rio Grande do Sul, Brazil
paula.abracco@gmail.com

IDEG workshop, Porto, Portugal, 2 December 2022

From c:\Bendix\teach\Epi\IDEG2022\slides/slides.tex Tuesday 29th November, 2022, 16:00

purpose of data collection

- epidemiology is all about how **time**
- influences health phenomena
- and how to **describe** and
- **quantify** this
- define a **measure**(ment scale) to use:
  - rate (observation scale)
  - probability (integrated scale)
  - time (integrated probability)
- analysis scale and reporting scale need not be the same

register / cohort characteristics

- records and variables
- persons as records
- events as records
- time intervals as records

records and variables

- **record**: a line in the dataset
  contains the **key**
- **key**: a set of variables needed to uniquely identify each record
- **variable**: column with the same piece of information in each record

persons as records

- cancer register
- diabetes register
- ...
- **dates** for each person:
  - date of birth
  - date of diagnosis
  - date of death
  - date of end of FU:
    event or last time seen (censoring)
- basis for creation of analysis data

events as records

- diagnoses of (recurrent) disease
  dates of diagnoses are events
- measurements
  dates of measurements are events
- **key**: (person, date)

time intervals as records

- representation of follow-up:
  - time span
  - event type (possibly “none”)
- **key**: (person, interval)
- basis for calculation of likelihood for **rates**
- each interval is an empirical rate: (event, time)
- ⇒ statistical models for rates
- `Epi` package uses `Lexis` frames for this

merging data frames

- same key?
- different keys
  $\Rightarrow$ decide which key to match on
  $\Rightarrow$ decide which key will be the new one
- the key is the most important characteristic of your analysis data frame
- most statistical models in epidemiology assume that (part of) the key represents independent observations (mostly person-id).