International initiative

# Guidance for key issues of design and analysis of observational studies

# TG 3: Descriptive and initial data analysis

Saskia le Cessie (Leiden, The Netherlands),
Maria Blettner (Mainz, Germany),
Werner Vach (Freiburg, Germany)

# Our topic is different from others:

1. It is not specialized (“everyone knows this”)
2. There are not many statistical papers available
   - Most applied statistics book have a chapter on descriptive statistics/ initial data analysis
   - STROBE (on quality of reporting)

# And it is important since:

3. First steps of data analysis are often forgotten
4. Descriptive statistics are often performed unorganized
5. Initial data analysis is sometimes all that is needed.
6. Young researchers are often reinventing the wheel

# Members of TG 3: Initial Data Analysis

- Currently:
  - Saskia le Cessie (Leiden, The Netherlands)
  - Maria Blettner (Mainz, Germany)
  - Werner Vach (Freiburg, Germany)

# Who are the experts on this field?

- Most of the statisticians with experience in applying statistics in medical research

# The approach of our TG group

- We start as small group
  - Identify relevant topics
  - Perform literature search
  - Combine literature with own experience

- Write first draft of guidance document

- Ask other “experts”, i.e. experienced applied statisticians and non-statisticians with ample experience on data analysis for feedback

- Finish guidance document

# Initial data analysis: Topics

- Research question and analysis plan
- Data cleaning.
- Description of the sample ("getting to know the data")
- Basic inference on the study population in Tables and Figures.
- Deciding what to do with peculiarities of the data (outliers, missing data) . Preparing the data for further analyses.

# Study protocol/Analysis plan

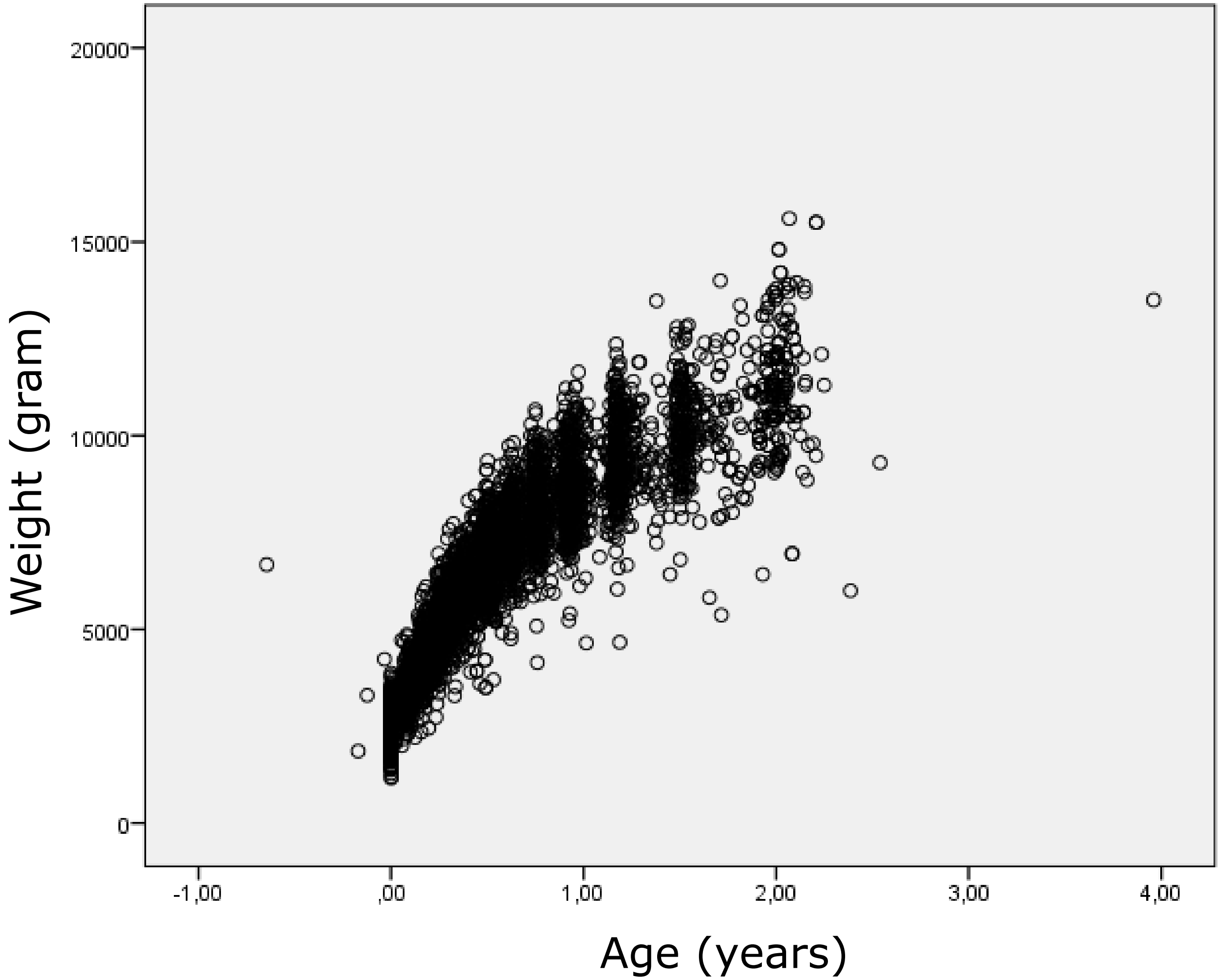
- Background is very crucial

- Research questions and analytic strategy plan should be formulated prior to initial data analysis (IDA) and IDA is part of the analytic strategy
- This to prevent fishing expeditions

- However IDA could sometimes change the analytic strategy (and even change the research question)

- Be transparent if such things happen

# Data cleaning



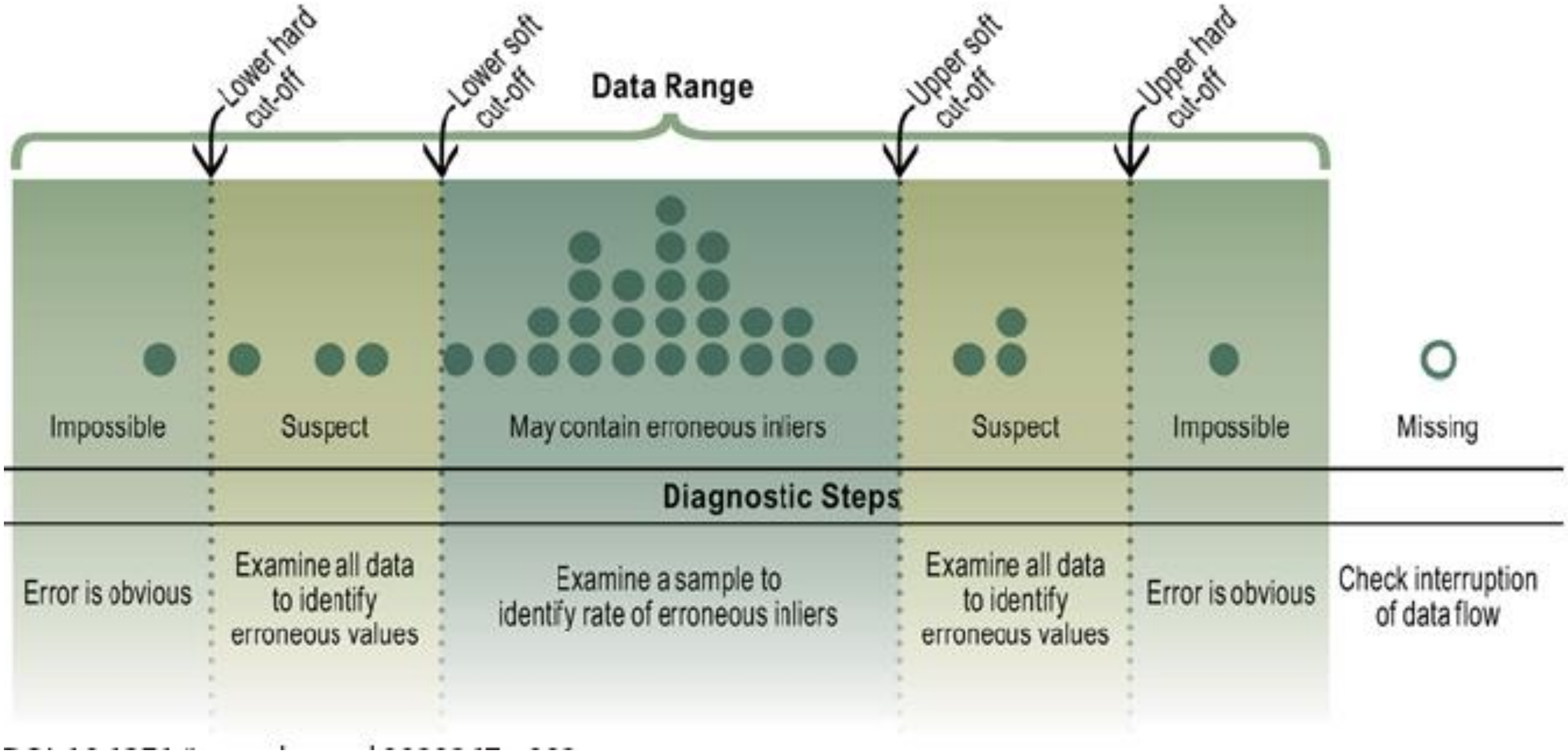
# Data cleaning

- Try to ensure to some degree, that data was collected and typed "error free".

- Identify: impossible values, strange patterns, inconsistencies,  outliers, missing values, impossible order of dates, etc.

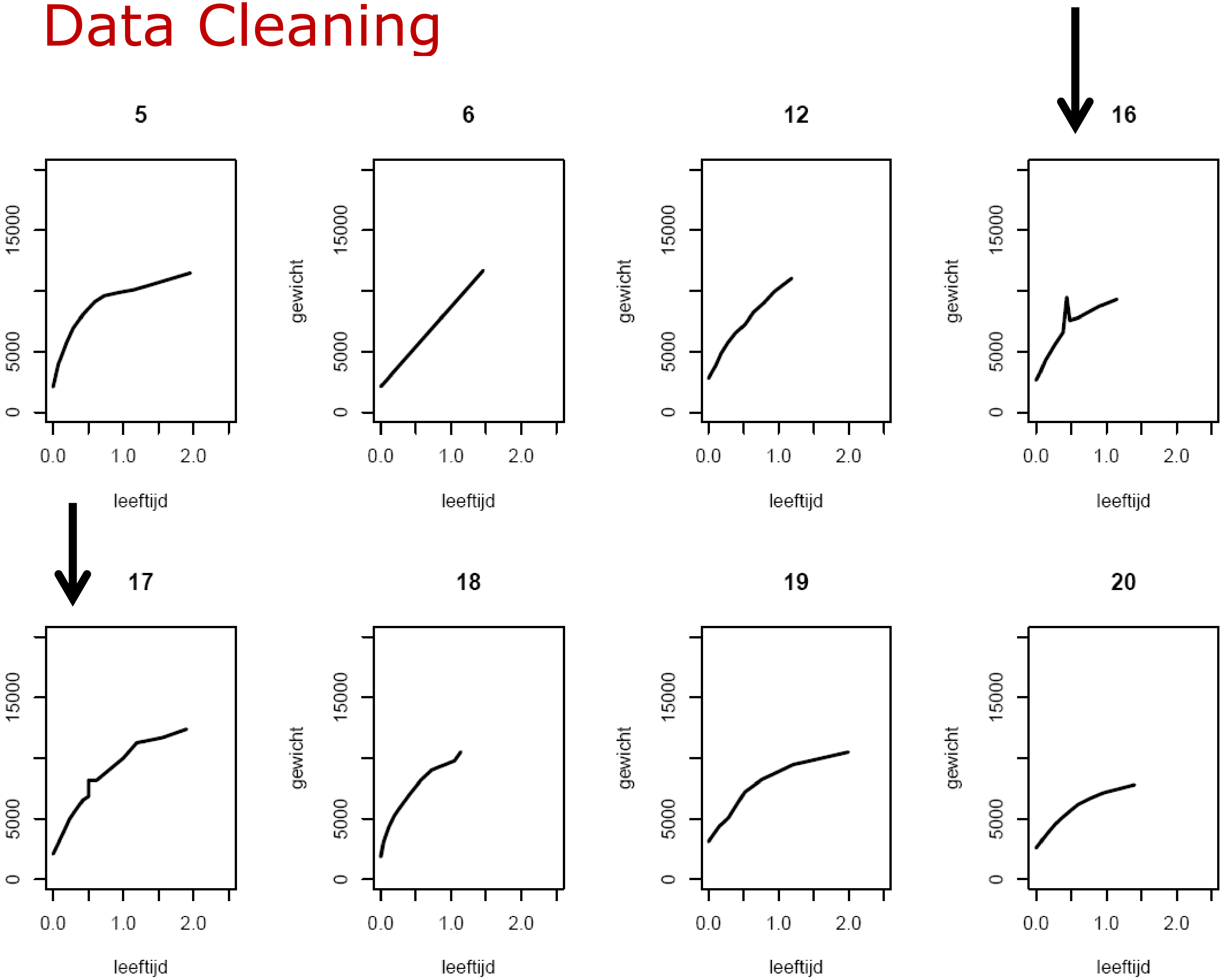
- Data cleaning is quite time consuming, not very exiting
- All changes should be documented (syntax, do-file, r-script)
- Never make changes in original database

# Data Cleaning


Lower hard cut-off
Lower soft cut-off
Data Range
Upper soft cut-off
Upper hard cut-off
Impossible
Suspect
May contain erroneous inliers
Suspect
Impossible
Missing
Diagnostic Steps
Error is obvious
Examine all data to identify erroneous values
Examine a sample to identify rate of erroneous inliers
Examine all data to identify erroneous values
Error is obvious
Check interruption of data flow

- Van den Brouck et al , PLOS medicine 2005

# Data Cleaning



# Exploring the data

- Examine data for particularities:
  - skewness of continuous variables,
  - outliers
  - limited variation,
  - number and patterns of missing values,
  - distributions of categorical variables (empty categories)

- Inclusion and flow of the study
  - Overview of baseline characteristics
  - overview of missing measurements and follow-up data
  - Flow chart

# Preparing data for subsequent analyses: dealing with the peculiarities of the data

- Continuous variables which are very skewed and/or extreme outliers
- Danger of very influential points
  - Transform
  - Categorize (using percentiles or clinically sensible values)

- Categorizing (heaping) continuous variables is not always such a bad idea.
- Context is important.
  - Confounders are handled differently from exposures
  - Prognostic questions versus causal questions

# Preparing data for subsequent analyses

- Very few observations in some categories
  - Pool categories (based on prior knowledge, not to obtain the smallest p-value)

- Variables with limited information
  - Could imply that variable is not usable ( and corresponding research questions cannot be handled)

- Missing data
  - Bias?
  - Can missing data techniques be used?

# Tables and Figures

- Table 1: Description of sample, or description of population?
  - Impute missing data in table 1?
  - Sampling weighting if certain groups are oversampled?

- Correct summary measures
  - use SD, not SE or 95% CI in population descriptions
  - Mean/SD versus median IQR

- In many situations, it may be usual/useful to split table 1 already in two groups
  - Treated/non treated
  - Responders/ non responders (NOT: total versus responders)

# Tables and Figures

- Should be understandable on its own, without reference to the text.

- Layout, rounding of numbers, rounding of p-values

# Relevant literature

- Data cleaning :
  - Van den Brouck et al Plos Medicine 2005
  - Several departments/ study groups have guidelines
- Initial data analysis
  - Chatfield The statistician 2001, JRSS-A 1985
  - Cox, Donnely 2011 Principles of Applied Statistics, Ch 5
- Descriptive statistics/ figures and tables
  - Basic statistical texts books and papers on statistics for non statisticians
- Reporting : STROBE
- Preparing data for subsequent analyses