A contemporary conceptual framework for initial data analysis

Werner Vach, Marianne Huebner, Saskia le Cessie
on behalf of Topic Group 3 of the STRATOS initiative

Institute of Medical Biometry and Statistics,
University of Freiburg, Germany
Department of Statistics and Probability,
Michigan State University, USA
Department of Clinical Epidemiology,
Leiden University Medical Center, The Netherlands

IDA ?

- often done
- informal and unorganised
- content unclear
  - data cleaning
  - basic data summaries
  - exploratory analysis
  - preparation of main analysis
- informal character $\rightarrow$ nontransparent impact on final results
- Chatfield C (1985): *The Initial Examination of Data*. J R Stat Soc Ser Gen 148: 214-253
- condition and scope of IDA has changed ...

Change in working situation

- Data sets have grown in size and complexity
- Data sets may include data from different sources
- Applied researchers are performing standard statistical analyses
- They may rush to perform sophisticated analyses, **without**
  - systematically checking for errors in the data,
  - a clear understanding about the underlying features of the data
  - knowledge on the suitability of the data for the intended analyses,
  - knowledge whether the data actually could provide answers to the research questions of interest.

Change in opportunities

- We can create hundred of histograms, boxplots and scatterplots in a second
- We have sophisticated tools for data visualisation
- In some areas we have established data preprocessing tool

New concerns

- IDA may have undesired or negative consequences on the final result
- unjustified removal of “disturbing” observations
- data driven hypotheses
- “optimising” of analysis strategies similar to trial and error approaches
- “Data dredging”, “Data snooping”
- nontransparent changes in the statistical analysis plan
- lack of reproducibility

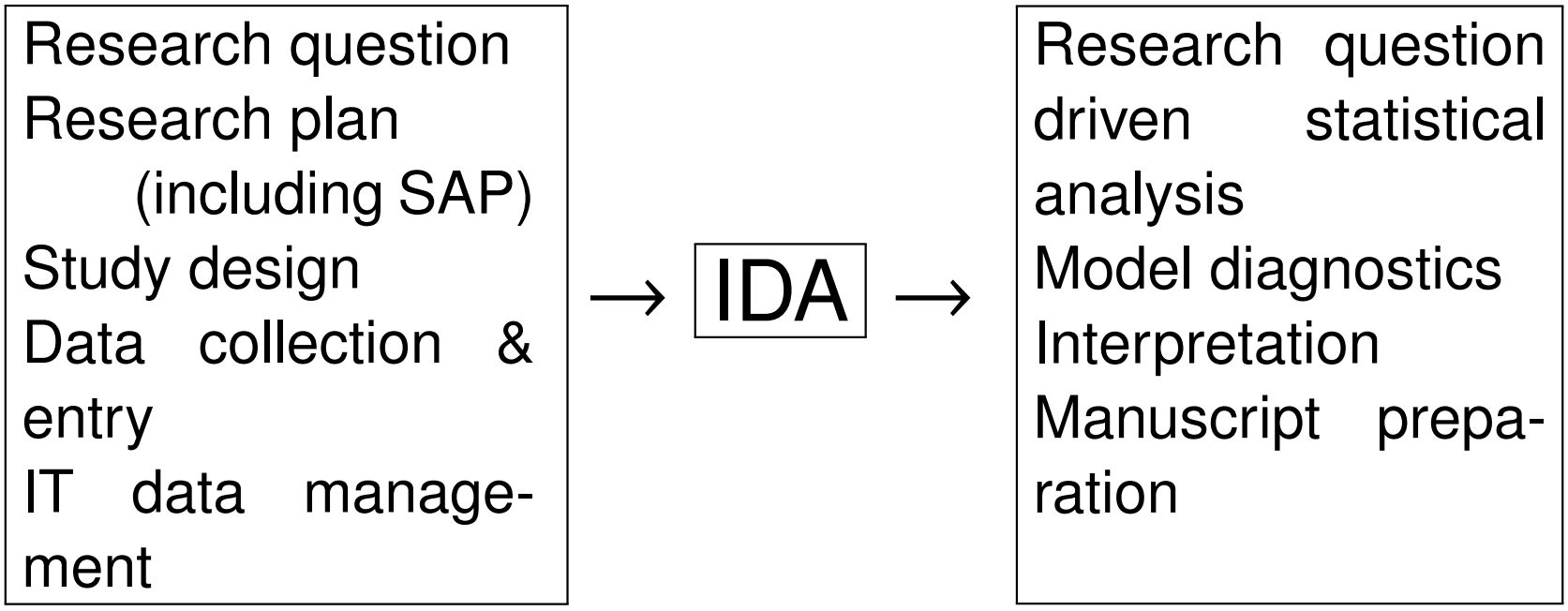
Our mission

IDA has to be established
as a necessary and legitimate step
in the mind of all researchers,
and how to perform IDA
in a structured and strategic way
needs to be discussed

Definition

IDA comprises all steps performed on the data of a study
between the end of the data collection/entry
and
start of those statistical analyses
in which research questions are addressed.

Definition


Research question
Research plan
(including SAP)
Study design
Data collection & entry
IT data management
IDA
Research question driven statistical analysis
Model diagnostics
Interpretation
Manuscript preparation

The aim of IDA

- to provide a data set and reliable background information on this data set
  - allowing researchers to work with this data set in a responsible manner, minimising the risk of producing numerical results and/or interpretations which are misleading or incorrect due to overlooking properties of the data

The 5 step model

1. Data cleaning
2. Data screening
3. Initial data reporting
4. Refining and updating the analysis plan
5. Reporting IDA in research papers

The 5 step model

1. **Data cleaning** is aimed at identifying and correcting errors in the data.
2. **Data screening** consists of understanding the properties of the data that may affect future analysis and interpretation.
3. **Initial data reporting** aims at informing all potential collaborators working with the data in future about all relevant insights obtained from the data screening.
4. **Refining and updating the analysis plan** translates the relevant findings from the data screening into corresponding adaptations of the analysis plan.
5. **Reporting IDA in research papers** ensures that all findings and actions from the previous steps that impact the interpretation of results are documented for the reader.

Data Screening – Aims

- Checking explicit or implicit expectations about certain properties of the data that need to be met (or are highly desirable) in order that the intended analysis of the study is applicable and can yield convincing results.
- Identifying properties of the data which may be regarded as a potential threat to the correct application of statistical methods or the adequate interpretation of results

Data Screening - Topics

- Distribution of single variables
- Missing data
- Association between variables
  - higher correlations than expected
  - lower correlations than expected
- Individual trajectories in longitudinal data
- Unintended factors
  - centres
  - observers
  - treatment providers
  - place of residence
  - time of day
  - day of the week
- Measurement error
  - deviations from expectations in variance or correlation
  - variables involved in logical inconsistencies
  - pre- and postvalues in usual care/placebo group

Organisational frame of IDA

- IDA team
- IDA plan
- The limits of manual inspection
- IDA subtasks as independent research
- Multi purpose studies / registers / EHR
- Legal frame of IDA
- IDA costs

Discussion

- Crucial decisions in defining the 5 steps
  - Data Cleaning vs. Data screening ?
  - Initial data reporting as a step of its own?
  - Refining and updating the statistical analysis plan?
- Statistical methodology
  - checklists and suggestions for adequate tools/techniques
- contemporary view $\rightarrow$ modern view

Outlook

Next steps

- Review about reporting IDA in research paper
- Guidance / Checklist for Data Cleaning
- Guidance for Data Screening
- Paper on graphical tools
- Paper on handling of skewed distribution of covariates
- Paper on topics prior to fitting a regression model

Who we are

- Maria Blettner (Mainz, Germany)
- Dianne Cook (Melbourne, Australia)
- Heike Hofmann (Iowa, USA)
- Hermann-Josef Huss (Bayer Health Care, Germany)
- Marianne Huebner (co-chair) (Michigan, USA)
- Saskia le Cessie (co-chair) (Leiden, Netherlands)
- Lara Lusa (Ljubljana, Slovenia)
- Werner Vach (co chair) (Freiburg, Germany)