A Contemporary Conceptual Framework for Initial Data Analysis

Carsten Oliver Schmidt, Werner Vach, Marianne Huebner, Saskia le Cessie

on behalf of *Topic Group 3*

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

Institute for Community Medicine, University Medicine Greifswald, Germany

Challenges in research

- Data sets have grown in size and complexity
- Data sets may include data from different sources
- Applied researchers perform standard statistical analyses, rushing 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 intended analyses,
  - knowledge whether the data actually could provide answers to the research questions of interest.

Initial Data Analysis (IDA)…

- is often done
- informal and unorganised
- content unclear
  - data cleaning
  - basic data summaries
  - exploratory analysis
  - preparation of main analysis
- informal character
- nontransparent impact
- Chatfield C (1985): The Initial Examination of Data. J R Stat Soc Ser Gen 148: 214-253

Our mission

IDA has to be established as a necessary and legitimate step in the mind of all researchers.

Aims and scope of IDA

IDA aims to

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

Aims and scope of IDA

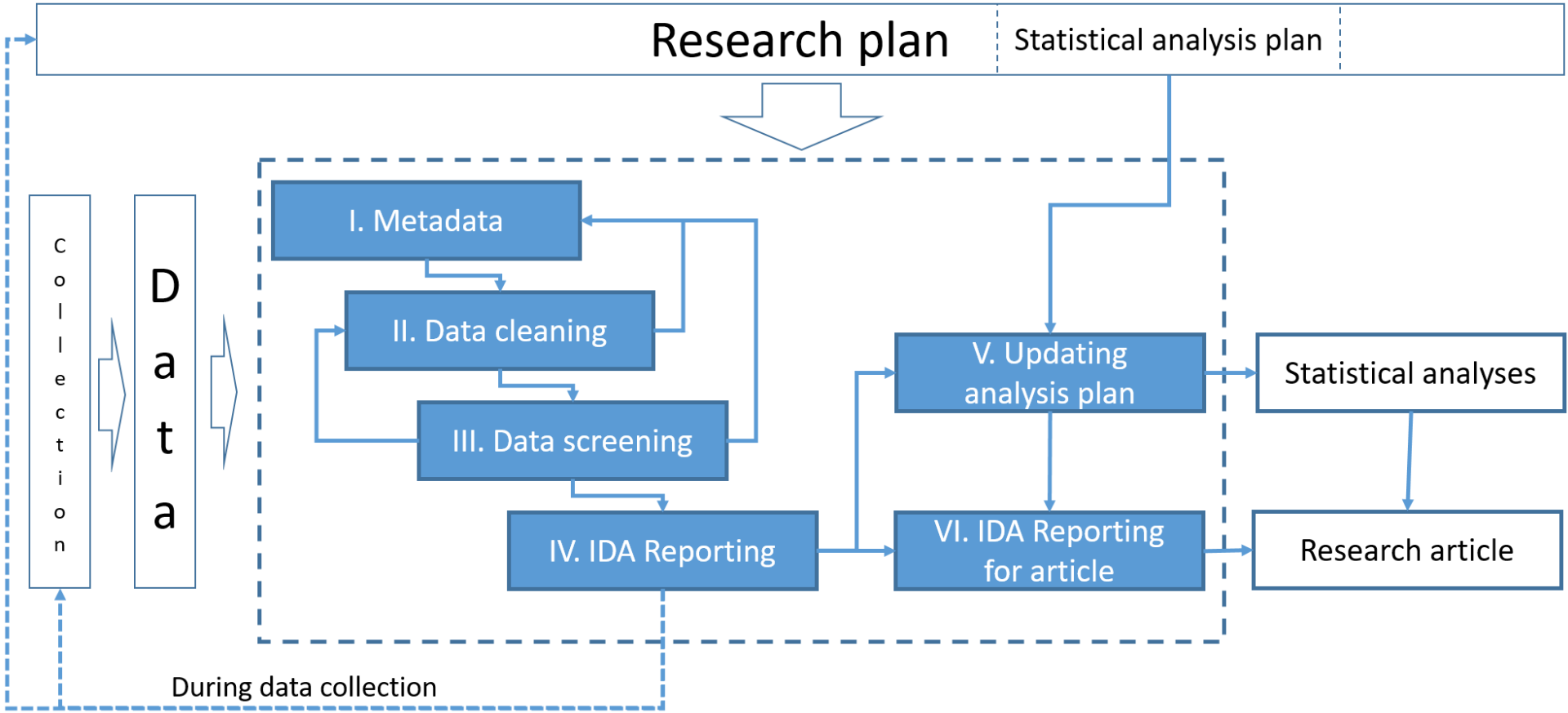
IDA typically takes place between the end of the data collection/entry and start of statistical analyses in which research questions are addressed

Ideally, IDA should already be performed during ongoing data collections to detect and deal with data issues as early as possible.

IDA steps….

1. **Metadata** setup to properly conduct following IDA steps
2. **Data cleaning** to identify and correct data errors
3. **Data screening** to better understand data properties
4. **Initial data reporting** to inform about insights
5. **Refining and updating the analysis** plan that translates the relevant findings into adaptations to the analysis plan
6. **Reporting of IDA** in research papers

IDA in the research workflow



Key output elements of IDA

	The six IDA steps					
Type of output	**I Metadata Setup**	**II Data cleaning**	**III Data screening**	**IV Initial data reporting**	**V Refining & updating the analysis plan**	**VI Reporting in research papers**
Analysis plan					Updated analysis plan	
Dataset		Cleaned dataset(s) with all original and derived variables			Analysis dataset to be used in final analyses	
Technical metadata / Data documentation	Comprehensive meta data; Comprehensive data dictionary	Updated data dictionary			Updated data dictionary	
Documentation	Documentation of all steps undertaken to create metadata	Documentation of all data manipulations and conducted data cleaning activities	Documentation of all conducted data screening activities	Summary of all findings from I-III with regards to their importance for subsequent analyses.	Documentation on suggested, discussed and accepted changes in the analysis plan	- Section in method parts - Section in results part - Discussion of IDA findings influencing the interpretation of results

Data Cleaning- Topics

- Inconsistencies between observed data values and the formal frame given by the data structure or meta-data
  - Date outside the time frame of the study
  - Missing patterns incompatible with the intended data
  - co
- Inco
  - L
  - I
- Log
  - I
  - I
- Indications of typical mistakes in the data collection or entry
  - Duplicate records (indicating double entry)
  - Partial duplicates (indicating unintended copy and paste)
  - A reversal of digits in non-matching key variables.

IDA
- can only generate flags
- needs an independent body to make decisions about changes
- accepted flags and corrections should be documented and reversible

Data Screening - Topics

- Distribution of single variables
- Missing data
- Association between variables
  - higher / 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

Refining the Analysis Plan - Topics

Topic	Examples	Possible influence on statistical analysis plan
Data properties not in accordance with requirements of intended statistical methods	• Unexpected skewness in the distribution of an outcome variable • Subgroups are much smaller than expected and asymptotic tests inappropriate.	Refinement, extensions or reduction of models
Unexpected heterogeneity of the study population	• Confusing information about the education of immigrants due to incompatibility of education systems	Stratified analyses; Restricting the analysis to a well-defined subgroup
Suspicious values	• Suspicious outliers; • Inconsistent follow-up dates	Decision rule on whether and how to use suspicious information in the analysis.

Risks of IDA

IDA may lead to

- “Date dredging” and “Data snooping”
- unjustified removal of “disturbing” observations
- to data driven hypotheses
- to nontransparent changes in the statistical analysis plan, refrain from touching the research questions.

IDA should as much as possible, refrain from touching the research questions.

Extensions

- Multi-purpose studies
- Reusing existing data

- IDA as part of data quality monitoring

IDA should start as early as possible and should continue with the growth of the study.

Areas of discussion…

- Crucial decisions in defining the IDA steps
  - Metadata concept
  - 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
- Organizational aspects
  - IDA Team
  - Manual vs. automated processses

Next steps…

- Review about reporting IDA in research papers
- 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
- Cooperation on project to develop standards for data quality assessments

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)
- Carsten Oliver Schmidt (Greifswald, Germany)