Overview of a framework for Initial Data Analysis

Saskia le Cessie (Leiden, The Netherlands)
Marianne Huebner (Michigan, USA)
Werner Vach (Freiburg, Germany)

On behalf of
TG 3: Initial data analysis

Topic Group 3: Initial Data Analysis (IDA)

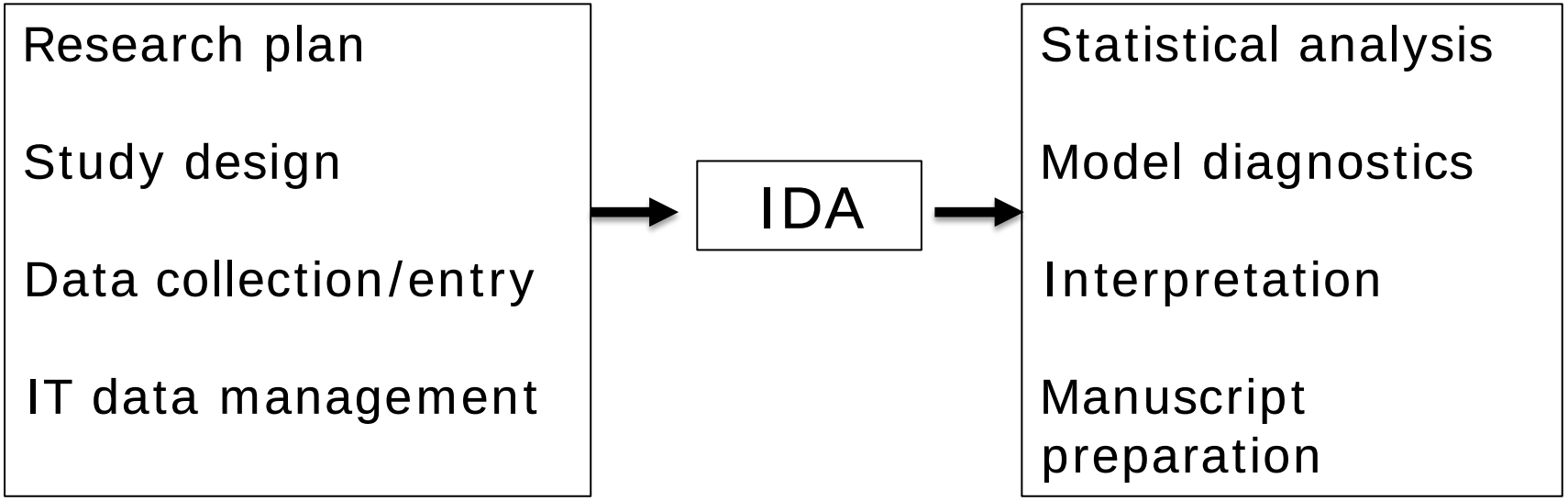
- Co-chairs
  - Marianne Huebner (Michigan State University, USA)
  - Saskia le Cessie (Leiden University Medical Centre, Netherlands)
  - Werner Vach (Medical Center Freiburg University, Germany)
- Members:
  - Maria Blettner (University of Mainz, Germany)
  - Dianne Cook (Monash University, Australia)
  - Heike Hofmann (Iowa State University, USA)
  - Hermann Huss (Bayer Pharma AG, Germany)
  - Lara Lusa (University of Ljubljana, Slovenia)

First TG3 project

- Develop a conceptually oriented, contemporary view on IDA (initial data analysis).

- Formulate the frame and scope of IDA, and different steps involved in IDA

Place of IDA in the overall process of a single study



IDA: The steps performed on the data between the end of data collection and intended start of statistical analyses

Why a systematic approach to IDA?

- Unorganized preparation of data may lead to numerical errors and wrong interpretations
- Need to ensure whether (complex) methods are appropriate for the data available
- Transparency and reproducibility
- Adequate reporting: reviewers and readers of papers have no idea whether (and how) IDA took place.

Initial data analysis

General principle

- IDA should not touch the research question of interest (no exploratory data analysis)

Result of IDA:

- Dataset with background information (meta-data) so that researchers can work with it in a responsible manner

Distinguish 5 steps in IDA

I. **Data cleaning:** aimed at identifying and correcting errors in the data.

II. **Data screening:** understanding the properties of the data that may affect future analysis and interpretation

III. **Initial data reporting:** relevant insights obtained from the data screening, description of complete process

IV. **Refining and updating analysis plan:** translating relevant findings into adaptations of analysis plan

V. **Reporting IDA in research papers**: relevant findings and steps impacting interpretations

1. Data cleaning

- Systematic attempt to find errors and –if possible- correct them
- Often detected indirectly (by observing inconsistencies in data)
- IDA flags (possible) errors and suggests corrections
- Actual corrections should be done in a transparent reproducible manner.

Result: Cleaned dataset, supplemented with record of changes

2. Data screening

- Aim: understanding the properties of data

- Are expectations of data met?
  Are observed distributions in accordance with expectations ?
  (i.e no underrepresentation of certain subgroups, hints of selection bias)

- Do properties of data meet requirements for correct application of statistical methods ?
  (i.e. skewness, missing data)?

Data screening tools

- Frequency tables, histograms
- Association between variables (Note: not touching the research question)
- Individual patterns over time
- Patterns of missing data
- Differences between centers
- Measurement error

3. Initial data reporting

An extensive report with results of data screening including

- Detailed flow chart of the study
- Distribution of all variables for the population and important subpopulation
- All insights which may influence the interpretation of the results
- All insights which may influence the further statistical analysis

4. Refining and updating the analysis plan

- Problematic: bears risk of “optimize findings”
- Clearly motivate changes, resulting from data screening

Topics which may influence statistical analysis plan

Suspicious values (outliers, inconsistent follow up times)

→ Decision on objective rule whether and how to use this information in analysis plan

Unexpected heterogeneity in study population (i.e. deviating centers, large amount of non response in certain subgroups)

→ Restrict the analysis to a well defined subgroup

Very skewed or bimodal variables

→ Transformation, or splitting bimodal variables into a binary variable to improve interpretability of results

Topics which may influence statistical analysis plan (2)

Data properties not in accordance with intended statistical methods (numbers too small for asymptotic methods, assumptions on distributions do not hold)

→ Change of statistical methods

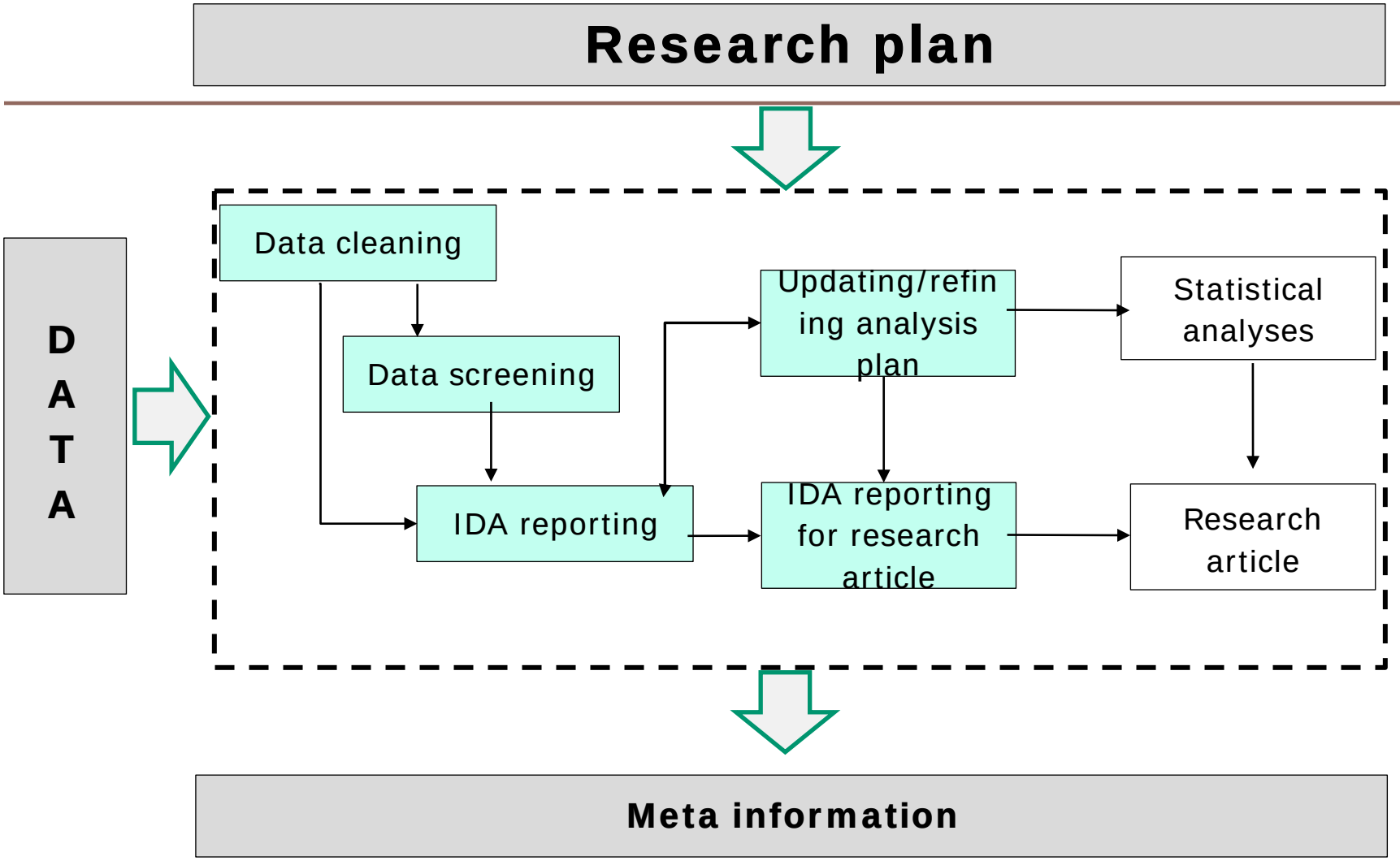
Intended model may be inadequate (zero inflation, strong center effects)

→ Refinements, extensions or reduction of models

5. Reporting IDA in manuscript

- Describe the flow of the data
- Overview about characteristics of the study population.
- IDA findings regarded as potentially influencing the statistical analysis
- Actual changes in the analysis plan
- IDA findings influencing the interpretation of results

- By part covered in reporting guidelines (strobe, consort)
- Details can be added as supplementary material


Research plan
DATA
Data cleaning
Data screening
IDA reporting
Updating/refining analysis plan
IDA reporting for research article
Statistical analyses
Research article
Meta information

Meta-information

"Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource." (*NISO) 2004*

- Information on variables (names, coding, measurement scale)
- Actual mechanisms for collecting the data
- Background and scope of study (study protocol, design, analysis plan)
- Data sources, processes for combining sources
- IDA reporting with suggested, discussed and accepted changes

Overlap between IDA steps?

- **Data cleaning and screening:** same analysis-different interpretation
- **Initial data reporting** included in data cleaning and screening
- In our frame work IDA reporting is separate step to ensure systematic and transparent reporting

Some discussion points

- IDA plan and team needed
- IDA is time consuming
- Limits of manual inspection
- Multi purpose studies (biobanks, registries, big cohorts)
  → danger that findings from IDA generate new research questions

TG3 papers

- A systematic approach to initial data analysis is good research practice. Huebner M, Vach W, le Cessie S. J Thorac Cardiovasc Surg. 2016 Jan;151(1):25-7

- A Contemporary Conceptual Framework for Initial Data Analysis. Huebner M, le Cessie S, Vach W (submitted)

Future projects

- Data visualizations for IDA (leads: Heike Hoffman and Dianne Cook)
- Literature review of reporting practices (lead: Marianne Huebner)
- Covariates with skewed distributions (lead: Werner Vach)
- More ideas:
  - Connection with other TGs for TG specific IDA
  - Electronic health records
  - Consider IDA steps in more depth