Recent and future work of
Initial data analysis Topic Group (TG3)

3	Initial data analysis	Chairs: Members:	**Marianne Huebner**, **Saskia le Cessie**, **Werner Vach** Dianne Cook, Heike Hofmann, Lara Lusa, Carsten Oliver Schmidt

Observational Studies 4 (2018) 171-192 Submitted 7/17; Published 4/18

A Contemporary Conceptual Framework for Initial Data Analysis

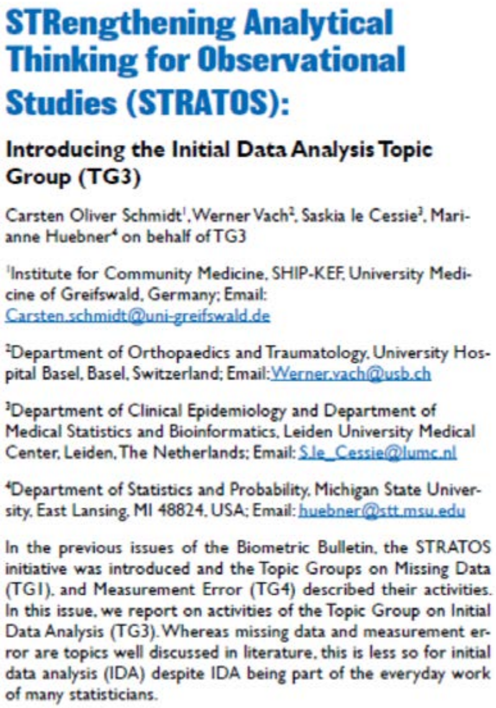
Marianne Huebner **huebner@stt.msu.edu**
Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824, USA

Saskia le Cessie **S.le Cessie@lumc.nl**
Department of Clinical Epidemiology and Department of Medical Statistics and Bioinformatics
Leiden University Medical Center
Leiden, The Netherlands

Carsten O. Schmidt **Carsten.schmidt@uni-greifswald.de**
Institute for Community Medicine, SHIP-KEF
University Medicine of Greifswald
Greifswald, Germany

Werner Vach **Werner.vach@usb.ch**
Department of Orthopaedics and Traumatology
University Hospital Basel
Basel, Switzerland

on behalf of the Topic Group "Initial Data Analysis" of the STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies, http://www.stratos-initiative.org). Membership of the Topic Group is provided in the Acknowledgments.



STRengthening Analytical Thinking for Observational Studies (STRATOS):

Introducing the Initial Data Analysis Topic Group (TG3)

Carsten Oliver Schmidt[1], Werner Vach[2], Saskia le Cessie[3], Marianne Huebner[4] on behalf of TG3

[1]Institute for Community Medicine, SHIP-KEF, University Medicine of Greifswald, Germany; Email: Carsten.schmidt@uni-greifswald.de

[2]Department of Orthopaedics and Traumatology, University Hospital Basel, Basel, Switzerland; Email: Werner.vach@usb.ch

[3]Department of Clinical Epidemiology and Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands; Email: S.le_Cessie@lumc.nl

[4]Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA; Email: huebner@stt.msu.edu

In the previous issues of the Biometric Bulletin, the STRATOS initiative was introduced and the Topic Groups on Missing Data (TG1), and Measurement Error (TG4) described their activities. In this issue, we report on activities of the Topic Group on Initial Data Analysis (TG3). Whereas missing data and measurement error are topics well discussed in literature, this is less so for initial data analysis (IDA) despite IDA being part of the everyday work of many statisticians.

Content

1. **Problem statement, mission, scope**
2. Initial data analysis (IDA) conceptual framework
3. Projects and papers in progress

Aims and scope of Initial Data Analysis (IDA)

IDA aims to

improve

Research quality

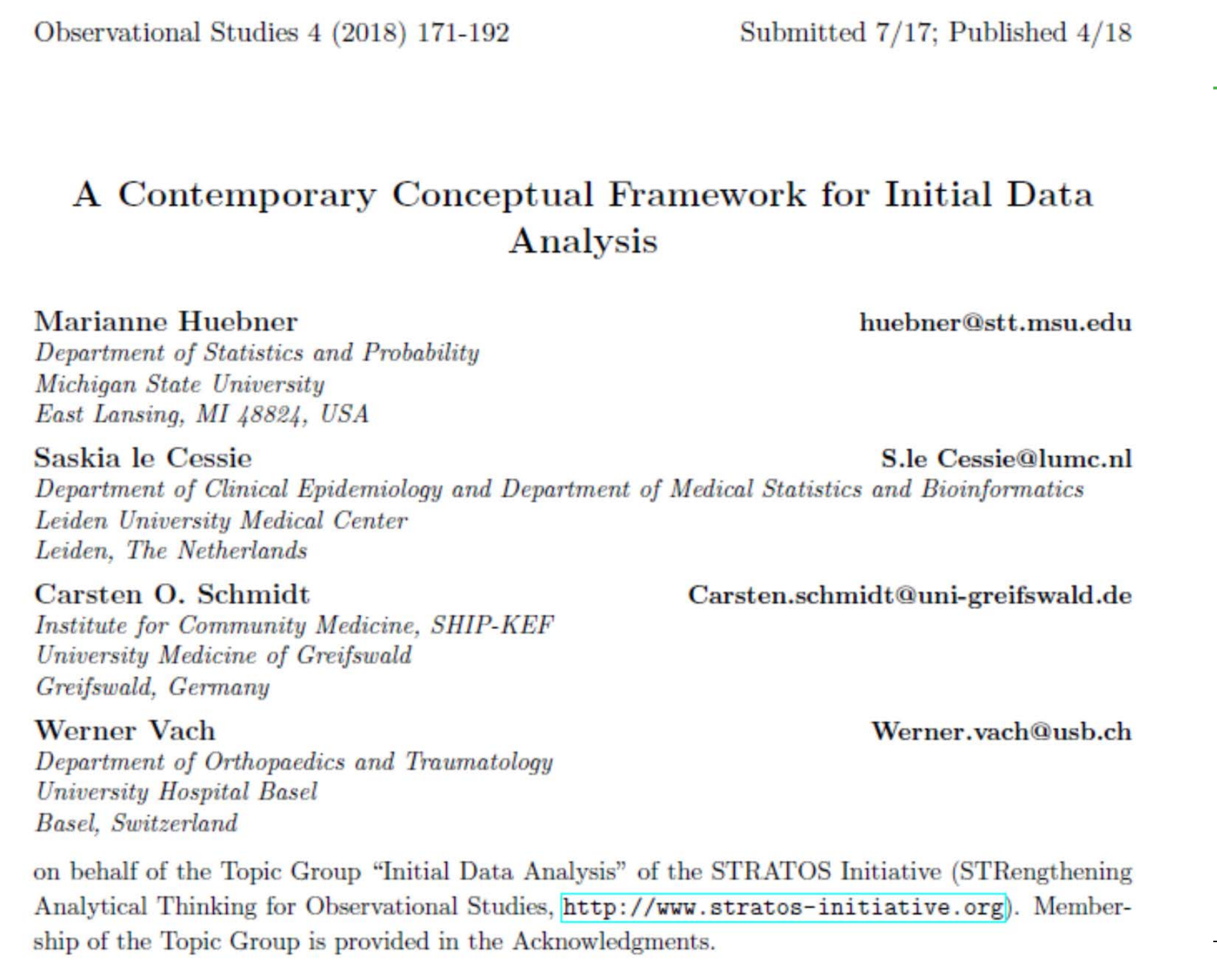
and

Research efficiency

Content

1. Problem statement, mission, scope
2. **Initial data analysis (IDA) conceptual framework**
3. Projects and papers in progress

STRATOS INITIATIVE



Observational Studies 4 (2018) 171-192 Submitted 7/17; Published 4/18

A Contemporary Conceptual Framework for Initial Data Analysis

Marianne Huebner **huebner@stt.msu.edu**
Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824, USA

Saskia le Cessie **S.le Cessie@lumc.nl**
Department of Clinical Epidemiology and Department of Medical Statistics and Bioinformatics
Leiden University Medical Center
Leiden, The Netherlands

Carsten O. Schmidt **Carsten.schmidt@uni-greifswald.de**
Institute for Community Medicine, SHIP-KEF
University Medicine of Greifswald
Greifswald, Germany

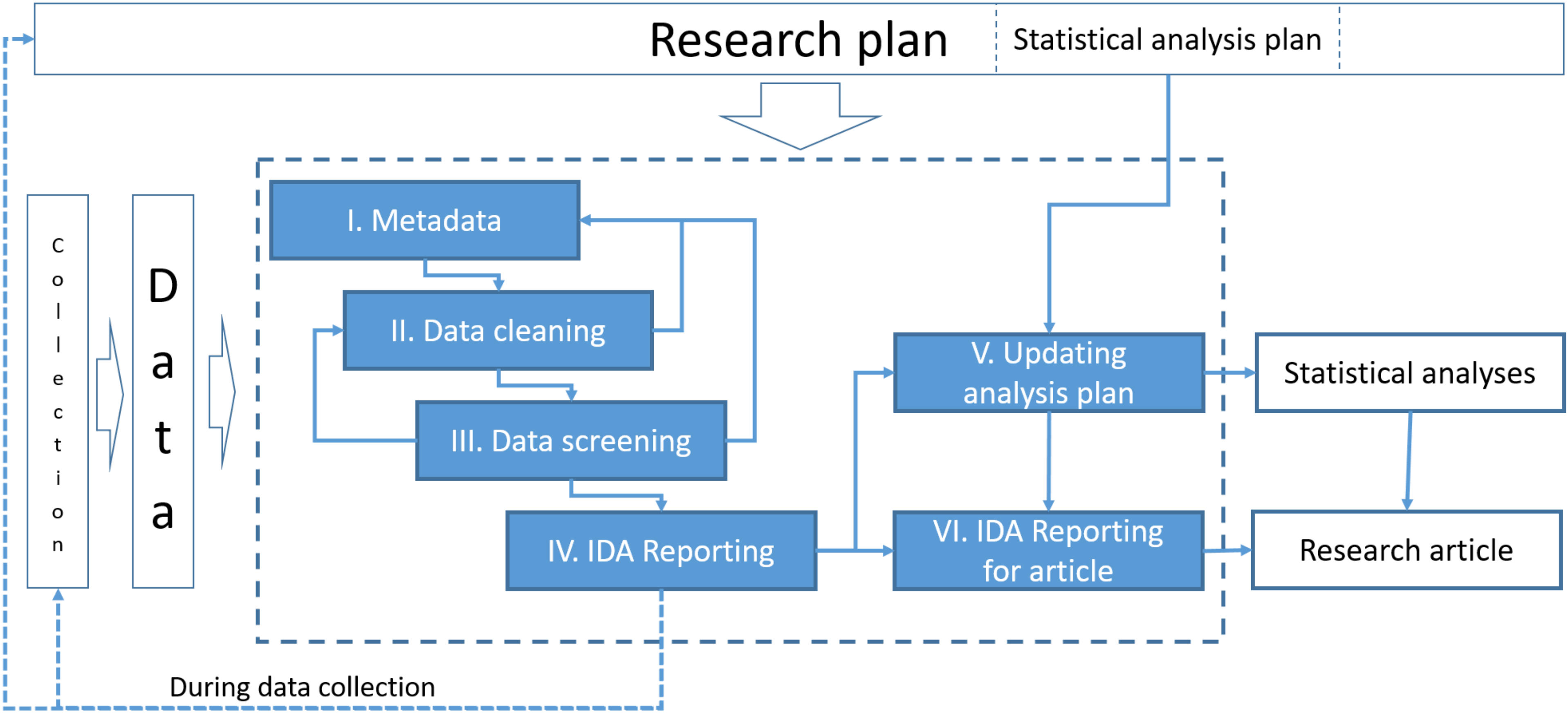
Werner Vach **Werner.vach@usb.ch**
Department of Orthopaedics and Traumatology
University Hospital Basel
Basel, Switzerland

on behalf of the Topic Group "Initial Data Analysis" of the STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies, http://www.stratos-initiative.org). Membership of the Topic Group is provided in the Acknowledgments.

IDA steps in the research workflow

1. **Metadata setup:** background information to properly conduct all following IDA steps.
   - Labels, data dictionary, study protocol, information sources
2. **Data cleaning:** identify and correct data errors.
3. **Data screening:** exploring data properties that may affect future analysis and interpretation.
4. **Initial data reporting:** document insights from previous steps for researchers who will work with the data.
5. **Refining and updating the analysis plan**
6. **Reporting IDA in research papers**

IDA steps in the research workflow



STRATOS INITIATIVE

IDA steps in the research workflow

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 be performed during ongoing data collections to detect data issues as early as possible.

IDA should refrain from touching research questions.

IDA- Topics

- Data integrity
- Data inconsistencies
- Compliance of data with study design
- Compliance of data with population properties
- Missing data (unit, item; missing variables; missing mechanisms)
- Association between variables
  - higher / lower correlations than expected
- (Joint-) Distributions / unexpected heterogeneity / measurement error
  - Outliers, suspicious values
  - Centres, observers, devices, treatment providers
  - Nondifferential error, differential error
  - Individual trajectories / carry over effects…
- Compliance between data properties and proposed statistical methods

IDA- Topics

Data Cleaning

- Data integrity
- Data inconsistencies

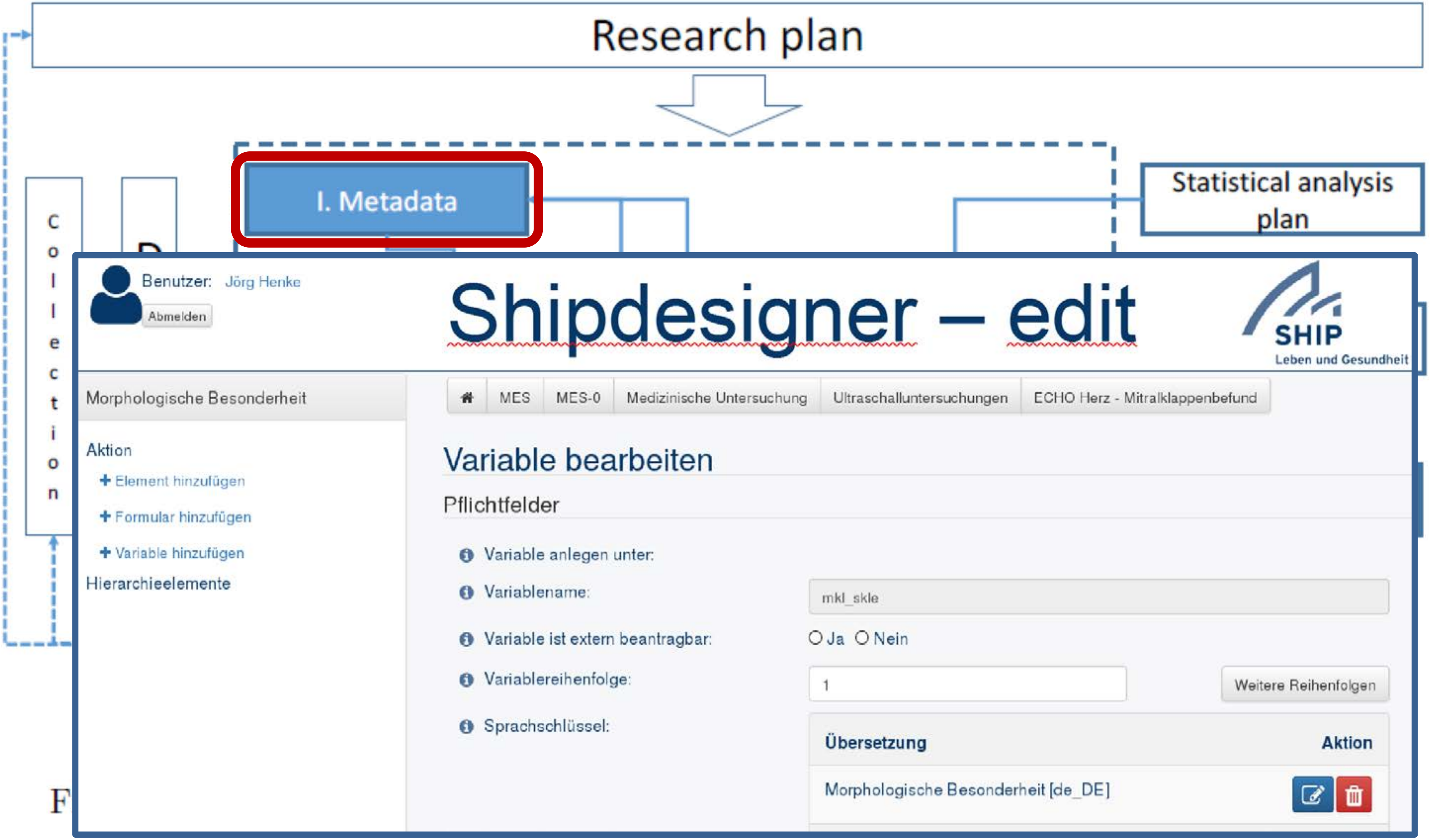
Data Screening

- Compliance of data with study design
- Sampling issues
- Missing data (missing values, missing units, patterns of missing)
- Association between variables
  - higher / lower correlations than expected
- (Joint-) Distributions / unexpected heterogeneity / measurement error
  - Outliers, suspicious values
  - Centres, observers, devices, treatment providers
  - Nondifferential error, differential error
  - Individual trajectories …
- Compliance between data properties and proposed statistical methods

Areas of discussion…

- Defining the IDA steps
  - Metadata concept
  - Demarcation data cleaning vs. data screening
  - Refining and updating statistical analysis plan
- Methodology and reporting
  - Appropriate / efficient setup of IDA analyses
  - checklists and suggestions for adequate tools/techniques
- Organizational aspects
  - IDA Team vs. analysis team
  - Manual vs. automated processses
  - IDA as part of data monitoring

Implementation example: SHIP


Research plan
I. Metadata
Statistical analysis plan
Collection
D
F

Benutzer: Jörg Henke

Abmelden

Shipdesigner – edit

SHIP
Leben und Gesundheit

Morphologische Besonderheit

Aktion

- ✚ Element hinzufügen
- ✚ Formular hinzufügen
- ✚ Variable hinzufügen

Hierarchieelemente

MES
MES-0
Medizinische Untersuchung
Ultraschalluntersuchungen
ECHO Herz - Mitralklappenbefund

Variable bearbeiten

Pflichtfelder

Variable anlegen unter:

Variablename: mkl_skle

Variable ist extern beantragbar: ○ Ja ○ Nein

Variablereihenfolge: 1

Weitere Reihenfolgen

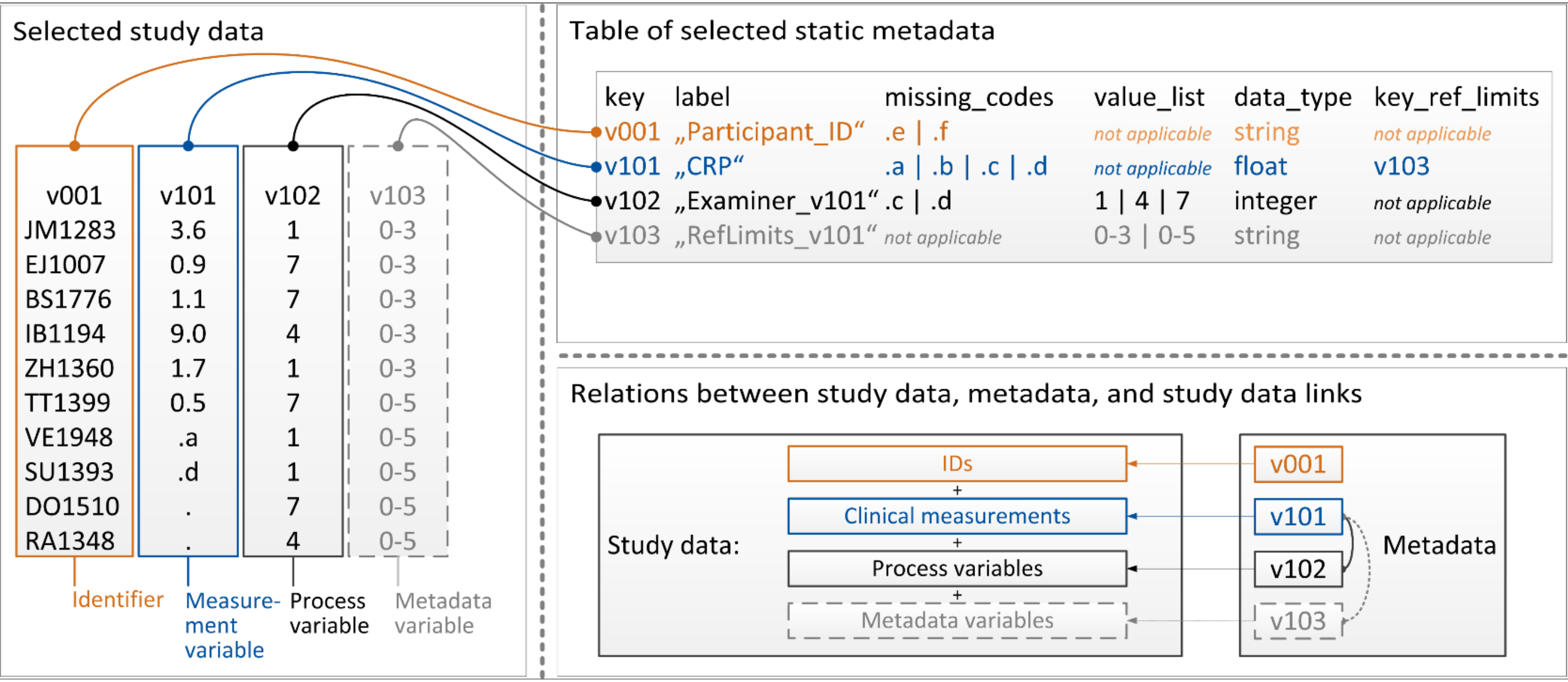
Sprachschlüssel:

Übersetzung	Aktion
Morphologische Besonderheit [de_DE]	

STRATOS INITIATIVE

Huebner, Cessie, Schmidt, Vach et al. 2018 Obs. Studies

How should metadata be organized?



Selected study data

v001	v101	v102	v103
JM1283	3.6	1	0-3
EJ1007	0.9	7	0-3
BS1776	1.1	7	0-3
IB1194	9.0	4	0-3
ZH1360	1.7	1	0-3
TT1399	0.5	7	0-5
VE1948	.a	1	0-5
SU1393	.d	1	0-5
DO1510	.	7	0-5
RA1348	.	4	0-5

Table of selected static metadata

key	label	missing_codes	value_list	data_type	key_ref_limits
v001	„Participant_ID“	.e | .f	*not applicable*	string	*not applicable*
v101	„CRP“	.a | .b | .c | .d	*not applicable*	float	v103
v102	„Examiner_v101“	.c | .d	1 | 4 | 7	integer	*not applicable*
v103	„RefLimits_v101“	*not applicable*	0-3 | 0-5	string	*not applicable*

Relations between study data, metadata, and study data links

Richter et al. Metadata in SHIP data quality monitoring 2018 submitted

Implementation example: SHIP

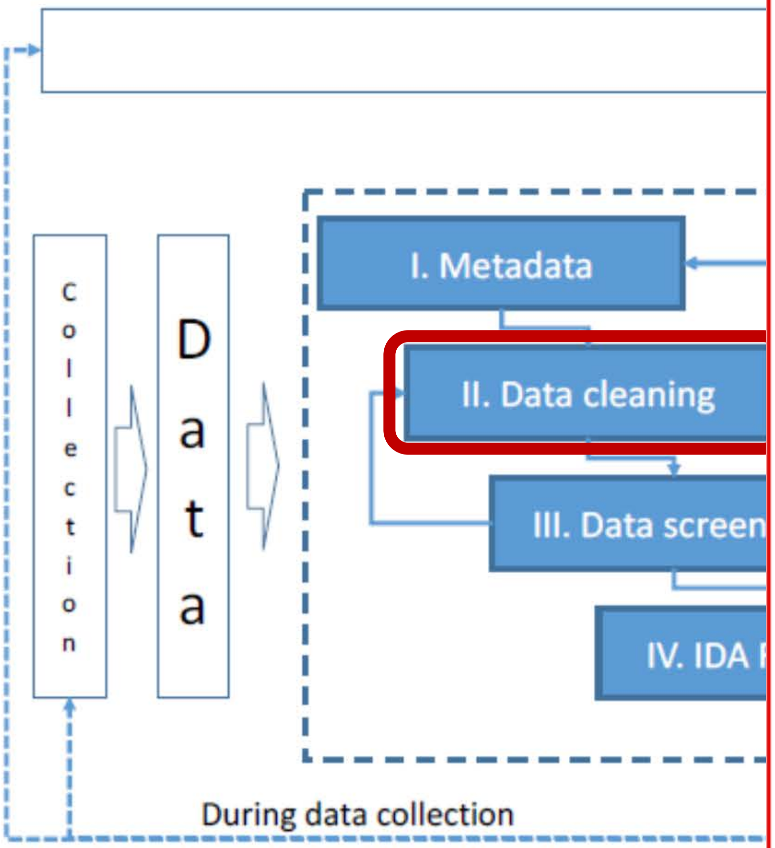
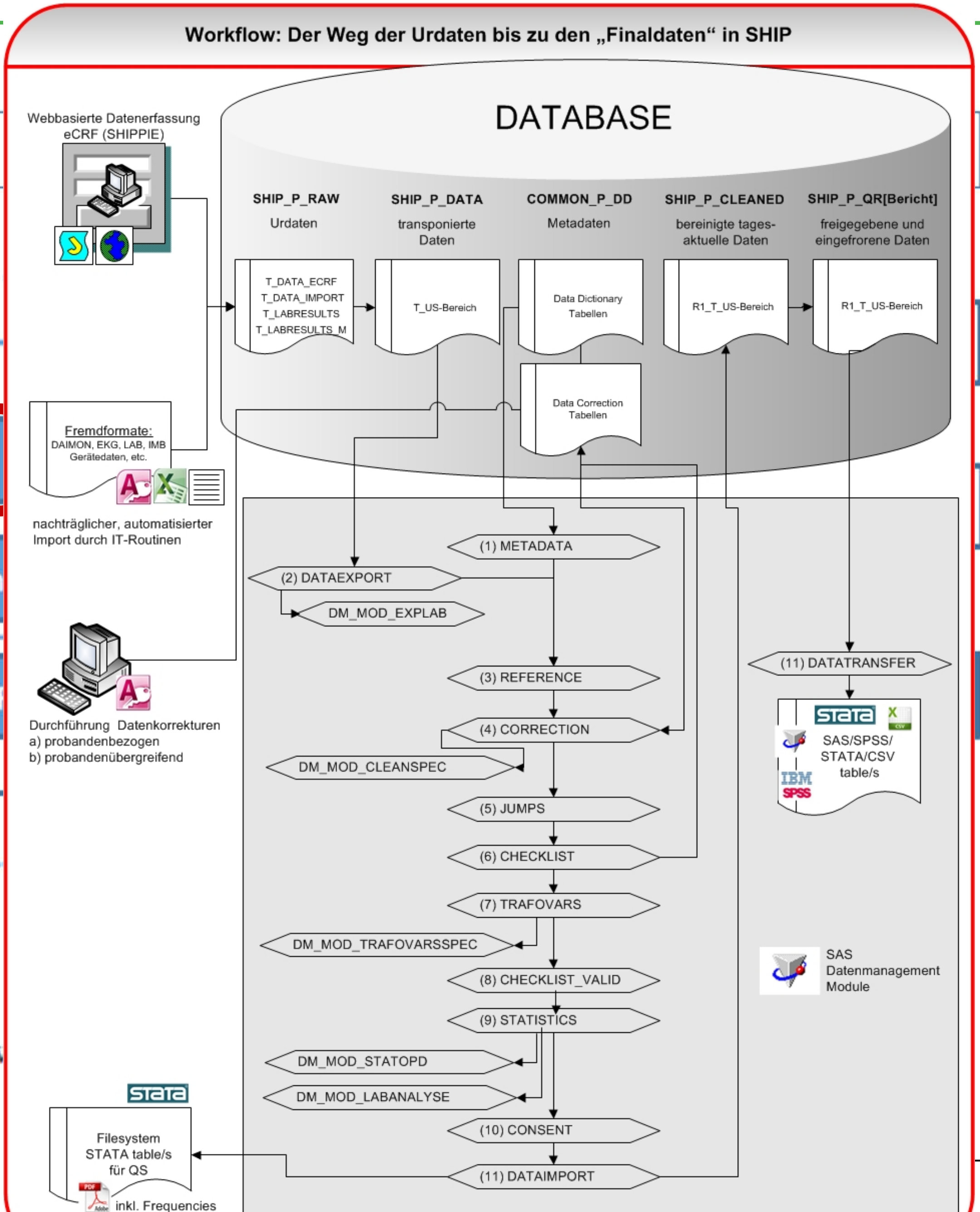
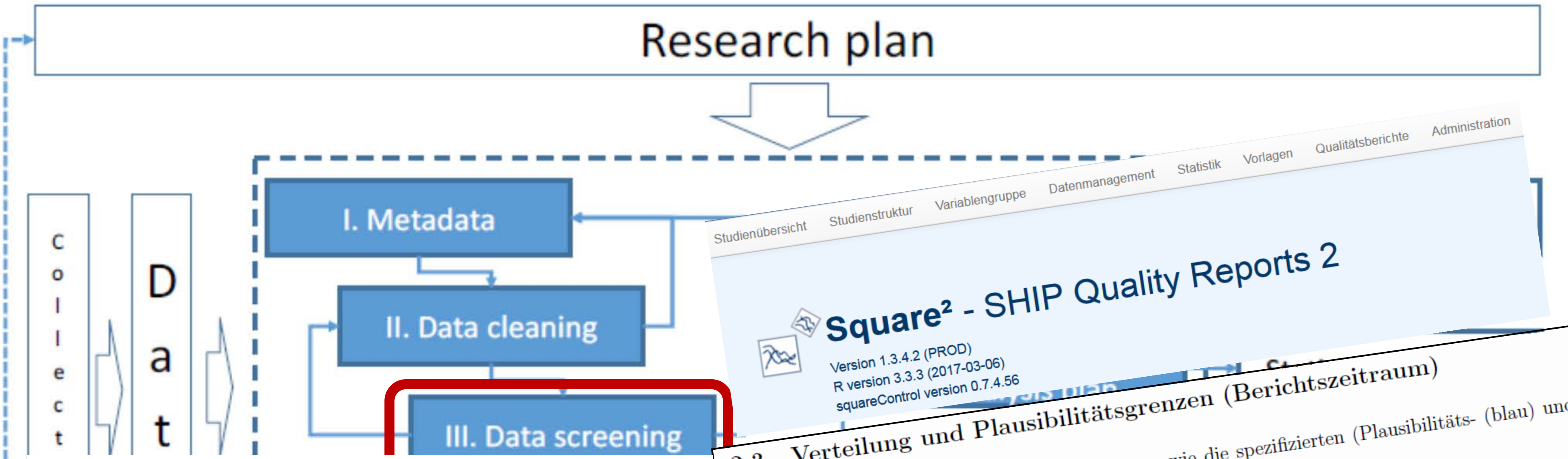


Figure 1: The main connections betw



Werner et al. 2016 KSFE

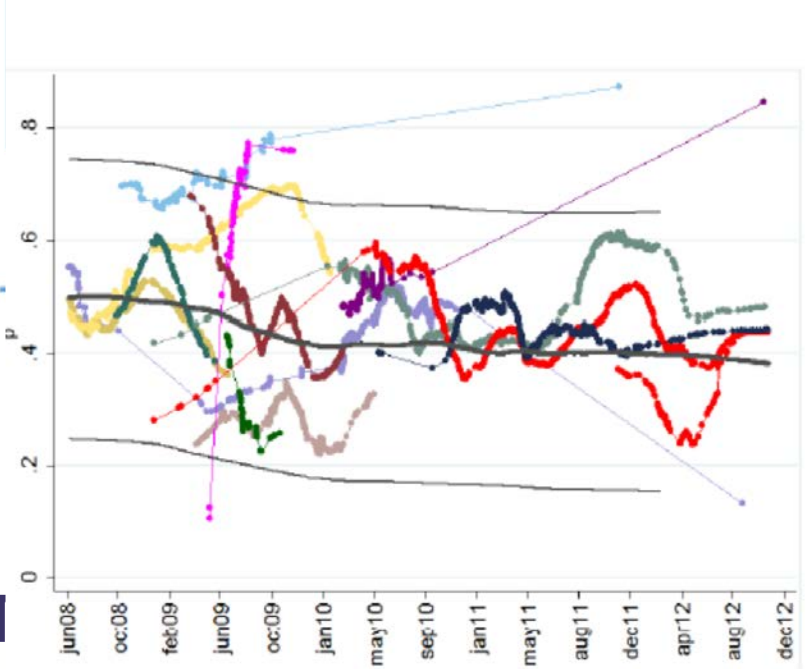
Implementation example: SHIP



Studienübersicht Studienstruktur Variablengruppe Datenmanagement Statistik Vorlagen Qualitätsberichte Administration

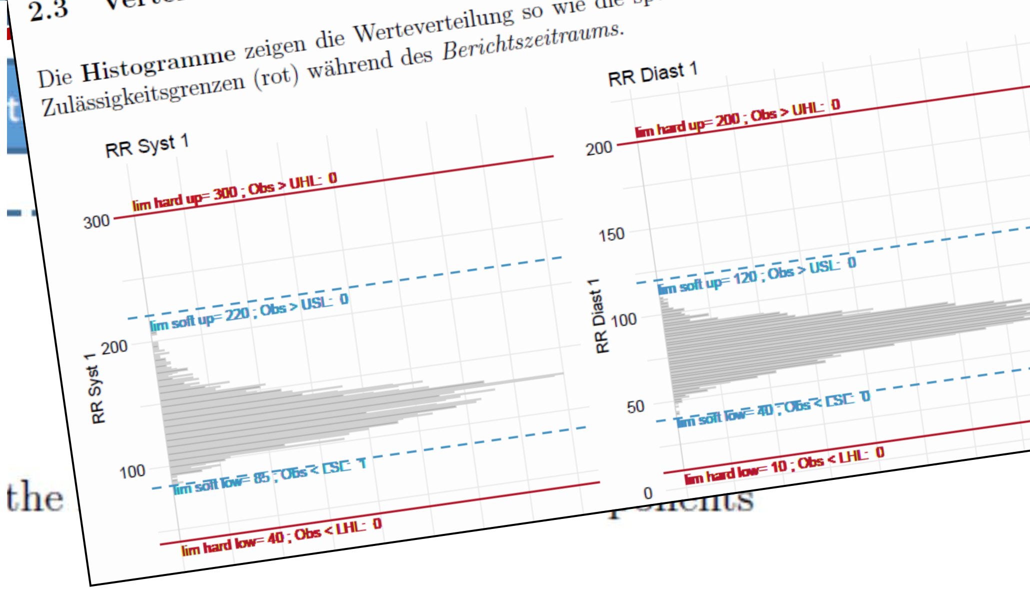
Square² - SHIP Quality Reports 2

Version 1.3.4.2 (PROD)
R version 3.3.3 (2017-03-06)
squareControl version 0.7.4.56



2.3 Verteilung und Plausibilitätsgrenzen (Berichtszeitraum)

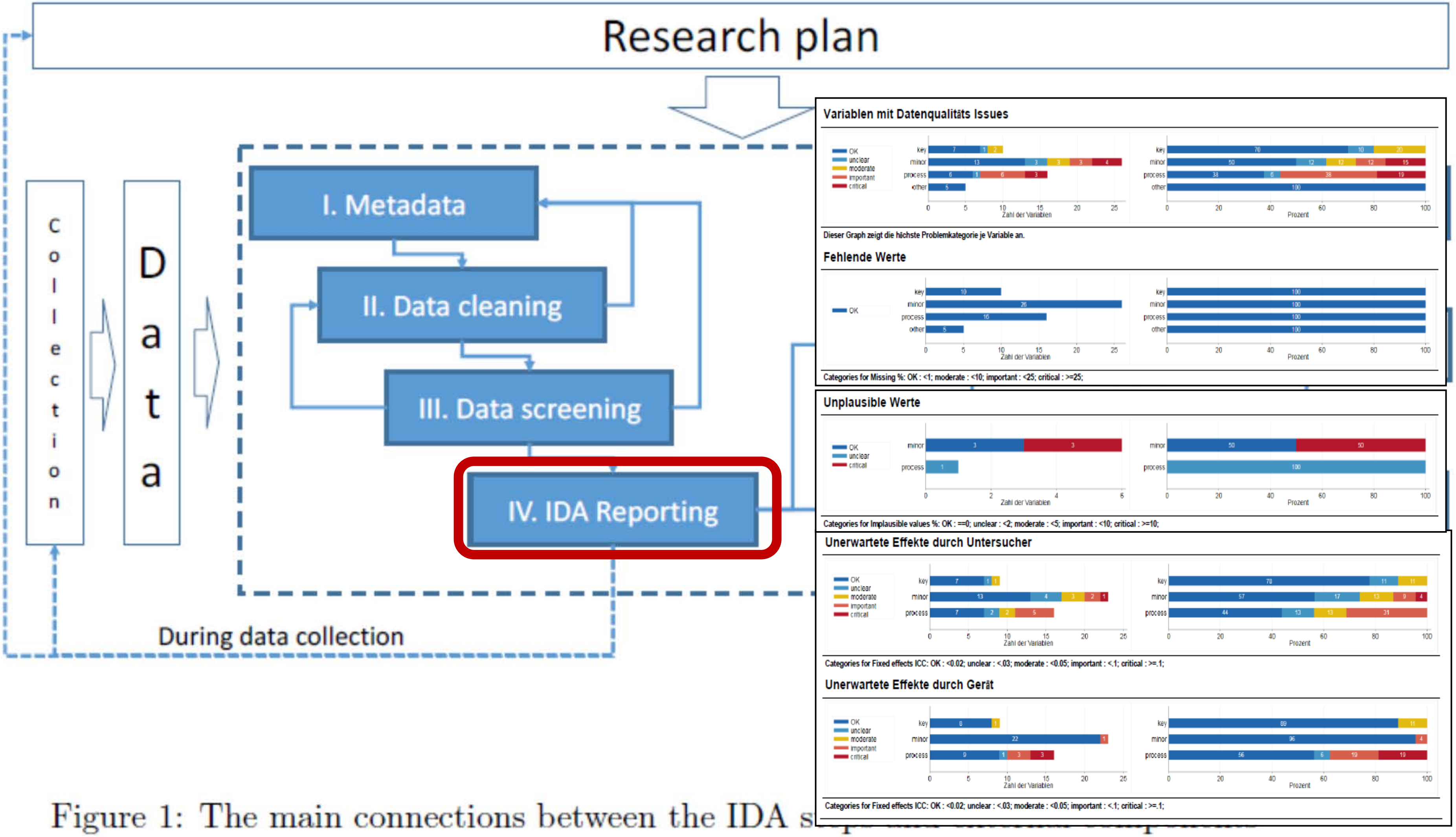
Die **Histogramme** zeigen die Werteverteilung so wie die spezifizierten (Plausibilitäts- (blau) und Zulässigkeitsgrenzen (rot) während des *Berichtszeitraums*.



the

hmidt, Vach et al. 2018 Obs. Studies

Implementation example: SHIP



Variablen mit Datenqualitäts Issues

Dieser Graph zeigt die höchste Problemkategorie je Variable an.

Fehlende Werte

Categories for Missing %: OK : <1; moderate : <10; important : <25; critical : >=25;

Unplausible Werte

Categories for Implausible values %: OK : ==0; unclear : <2; moderate : <5; important : <10; critical : >=10;

Unerwartete Effekte durch Untersucher

Categories for Fixed effects ICC: OK : <0.02; unclear : <.03; moderate : <0.05; important : <.1; critical : >=.1;

Unerwartete Effekte durch Gerät

Categories for Fixed effects ICC: OK : <0.02; unclear : <.03; moderate : <0.05; important : <.1; critical : >=.1;

Figure 1: The main connections between the IDA s

Huebner, Cessie, Schmidt, Vach et al. 2018 Obs. Studies

Some issues to think about in IDA

1. Efficient IDA begins at the start, not at the end of a study
2. Efficient background IT infrastructure mandatory

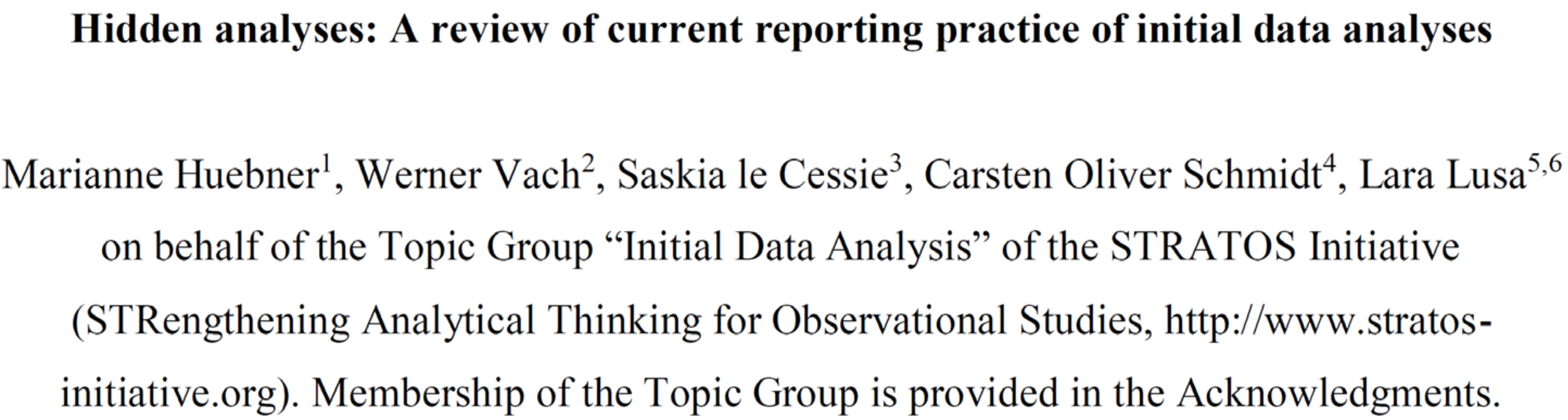
Content

1. Problem statement, mission, scope
2. IDA framework paper
3. **Projects and papers in progress**

Projects and papers

- **Review about reporting IDA in research papers**
- **Paper on IDA prior to fitting a regression model**
- Guidance for Data Screening
- Paper on graphical tools
- “Real life IDA examples”
- Standards and tools for data quality assessments

Review IDA reporting



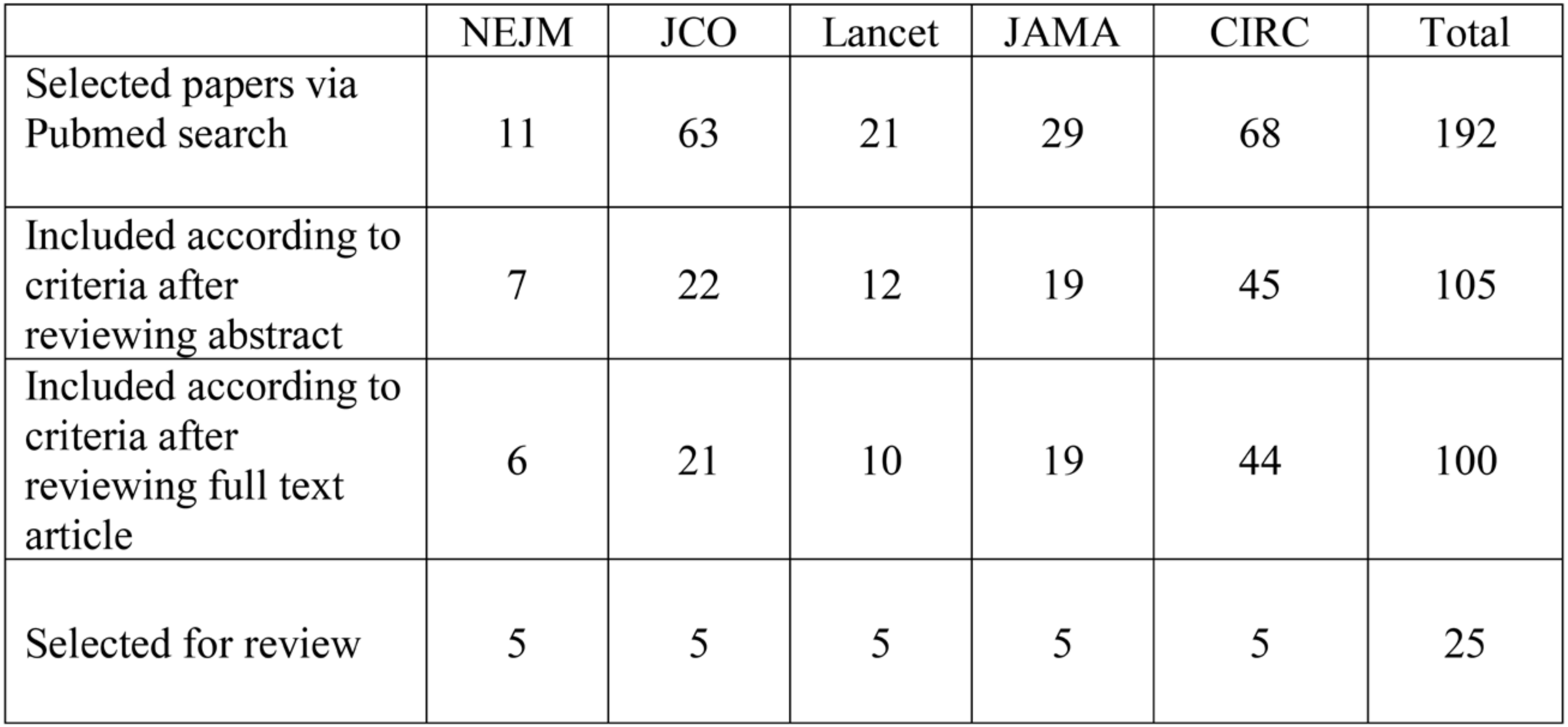
Hidden analyses: A review of current reporting practice of initial data analyses

Marianne Huebner[1], Werner Vach[2], Saskia le Cessie[3], Carsten Oliver Schmidt[4], Lara Lusa[5,6]
on behalf of the Topic Group "Initial Data Analysis" of the STRATOS Initiative
(STRengthening Analytical Thinking for Observational Studies, http://www.stratos-initiative.org). Membership of the Topic Group is provided in the Acknowledgments.

Aim: to describe current practice of IDA reporting in observational studies

Review IDA reporting – Article selection

Table 1: Search and selection of articles



	NEJM	JCO	Lancet	JAMA	CIRC	Total
Selected papers via Pubmed search	11	63	21	29	68	192
Included according to criteria after reviewing abstract	7	22	12	19	45	105
Included according to criteria after reviewing full text article	6	21	10	19	44	100
Selected for review	5	5	5	5	5	25

Each paper reviewed by two people

Review IDA reporting – data cleaning

- Ten papers (40%) included a statement about data cleaning

- Most statements were rather generic, e.g “Clinically improbable laboratory values were removed.”

- Some papers provide details in supplements

- One paper included the computer code used for data cleaning in the Supplement

Review IDA reporting - Data screening

- All papers give summary of non outcome and outcome variables
  - Most often in results
  - In Tables and text
- Missing values often reported (> 75%)
- Study flow/missing units often reported
- Transformations of non outcome variables mentioned in 10 papers (most often categorisation)

Review IDA reporting - Change in analysis plan

- Categories were grouped or numerical variables were categorized
  - Because few women were underweight (1.2%), we combined underweight with normal BMI (normal/underweight) and performed a sensitivity analysis excluding the underweight group.”
  - “patients had Hospital Frailty Risk Scores ranging from 0 to 99, but this was heavily skewed to the right”→ 3 categories
- Change of inclusion criteria
  - Exclude centers with bad data quality
  - Exclude centers with few events
- Dealing with missing data
  - complete case analysis/imputation

Review IDA reporting- Discussion

1. Information on IDA is sparse.
   - Relevant findings and subsequent decisions should be reported
2. Information on IDA can be found in all sections of a paper.
   - IDA methodology to be described in Methods;
   - IDA results to be described in Methods or Results;
   - Impact of IDA on interpretation discussed in discussion.
3. Distinction between pre-planned decisions and IDA-driven decisions are unclear.
4. Characteristics of participants are listed without comments.
5. Reporting on missingness is often incomplete.

IDA- Finishing remarks

More in:

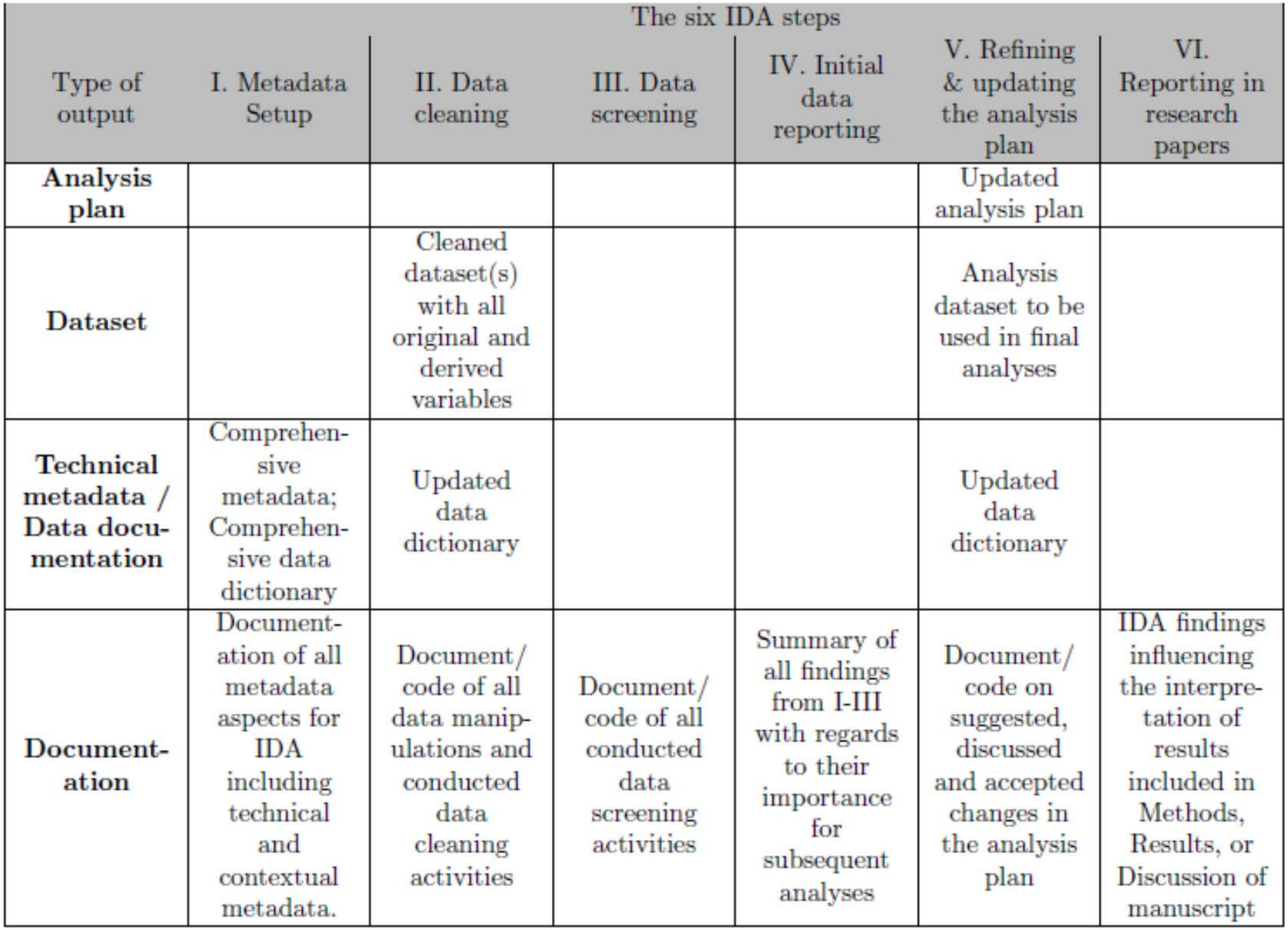
- Huebner, le Cessie, Schmidt, Vach et al. 2018 Obs. Studies
- Biometric Bulletin

IDA in teaching

website

- https://www.stratosida.org/home

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 metadata; Comprehensive data dictionary	Updated data dictionary			Updated data dictionary	
Documentation	Documentation of all metadata aspects for IDA including technical and contextual metadata.	Document/ code of all data manipulations and conducted data cleaning activities	Document/ code of all conducted data screening activities	Summary of all findings from I-III with regards to their importance for subsequent analyses	Document/ code on suggested, discussed and accepted changes in the analysis plan	IDA findings influencing the interpretation of results included in Methods, Results, or Discussion of manuscript

Table 1: Key output elements of the six IDA steps