TG3: Initial Data Analysis

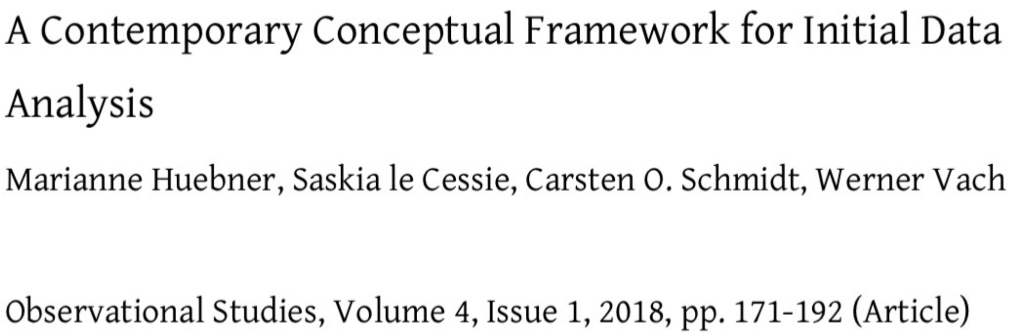
Chairs: Marianne Huebner (Michigan State University, USA), Carsten Oliver Schmidt (University Medicine Greifswald, Germany)

Members: Mark Baillie (Novartis, Switzerland), Saskia le Cessie (Leiden University , Netherlands), Lara Lusa (University of Primorska, Slovenia)

Website: https://www.stratosida.org

STRATOS
INITIATIVE

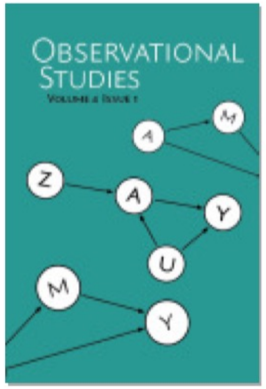
TG3 Papers

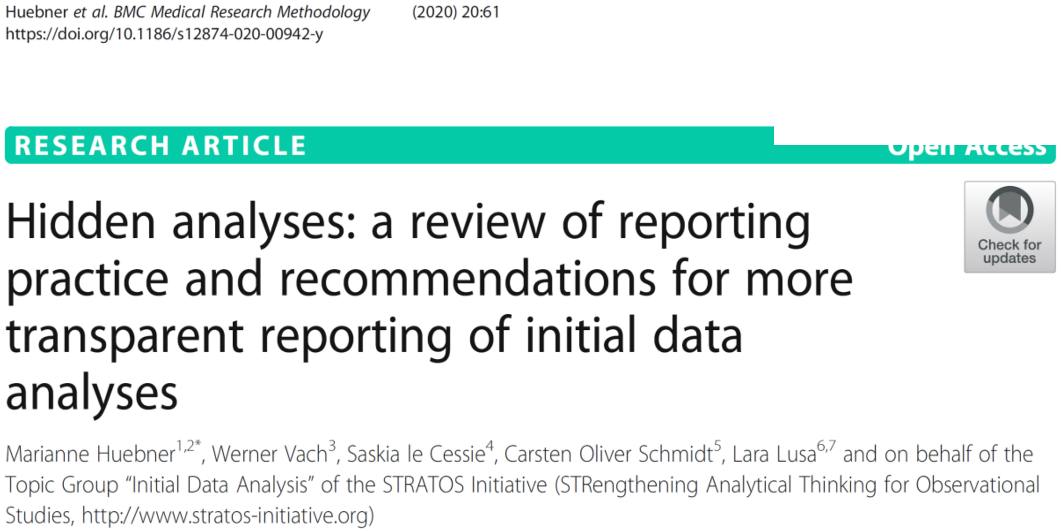


A Contemporary Conceptual Framework for Initial Data Analysis

Marianne Huebner, Saskia le Cessie, Carsten O. Schmidt, Werner Vach

Observational Studies, Volume 4, Issue 1, 2018, pp. 171-192 (Article)





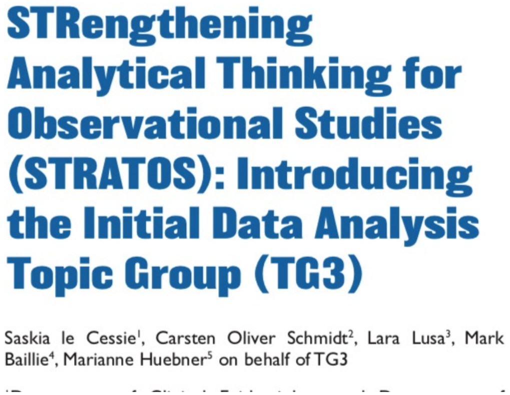
Huebner *et al. BMC Medical Research Methodology* (2020) 20:61
https://doi.org/10.1186/s12874-020-00942-y

RESEARCH ARTICLE Open Access

Hidden analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses

Check for updates

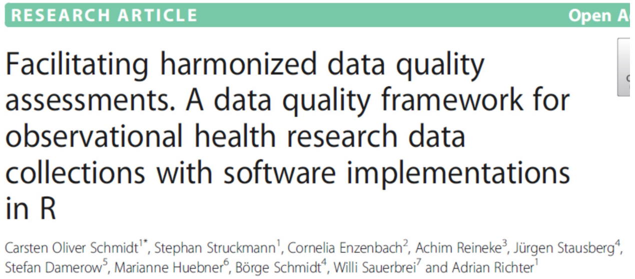
Marianne Huebner[1,2*], Werner Vach[3], Saskia le Cessie[4], Carsten Oliver Schmidt[5], Lara Lusa[6,7] and on behalf of the Topic Group "Initial Data Analysis" of the STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies, http://www.stratos-initiative.org)



STRengthening Analytical Thinking for Observational Studies (STRATOS): Introducing the Initial Data Analysis Topic Group (TG3)

Saskia le Cessie[1], Carsten Oliver Schmidt[2], Lara Lusa[3], Mark Baillie[4], Marianne Huebner[5] on behalf of TG3

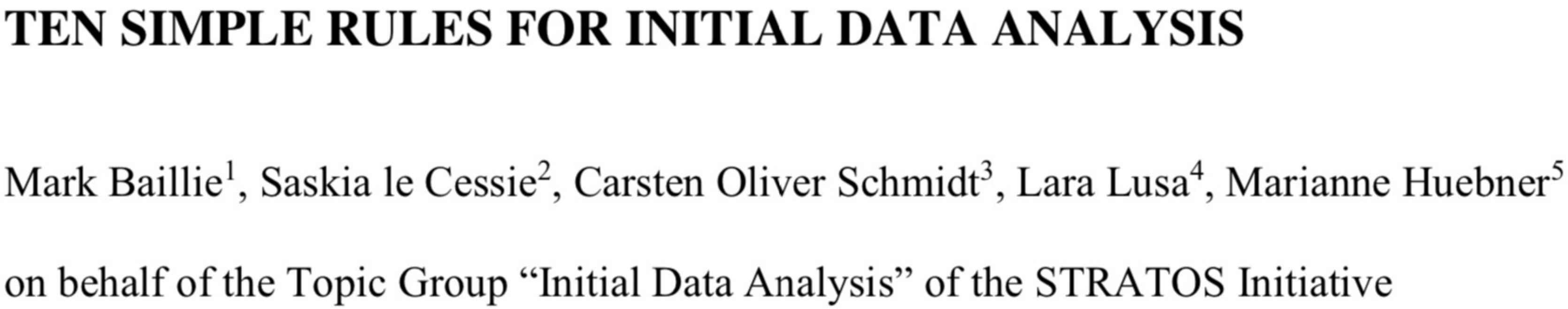
Associated with TG3:



RESEARCH ARTICLE Open A

Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R

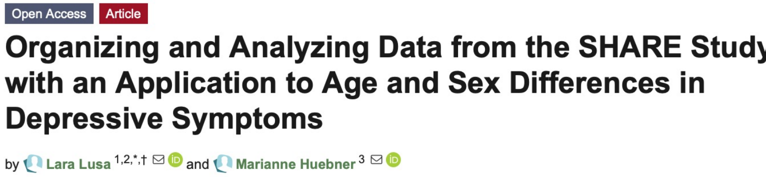
Carsten Oliver Schmidt[1*], Stephan Struckmann[1], Cornelia Enzenbach[2], Achim Reineke[3], Jürgen Stausberg[4], Stefan Damerow[5], Marianne Huebner[6], Börge Schmidt[4], Willi Sauerbrei[7] and Adrian Richter[1]



TEN SIMPLE RULES FOR INITIAL DATA ANALYSIS

Mark Baillie[1], Saskia le Cessie[2], Carsten Oliver Schmidt[3], Lara Lusa[4], Marianne Huebner[5]

on behalf of the Topic Group "Initial Data Analysis" of the STRATOS Initiative

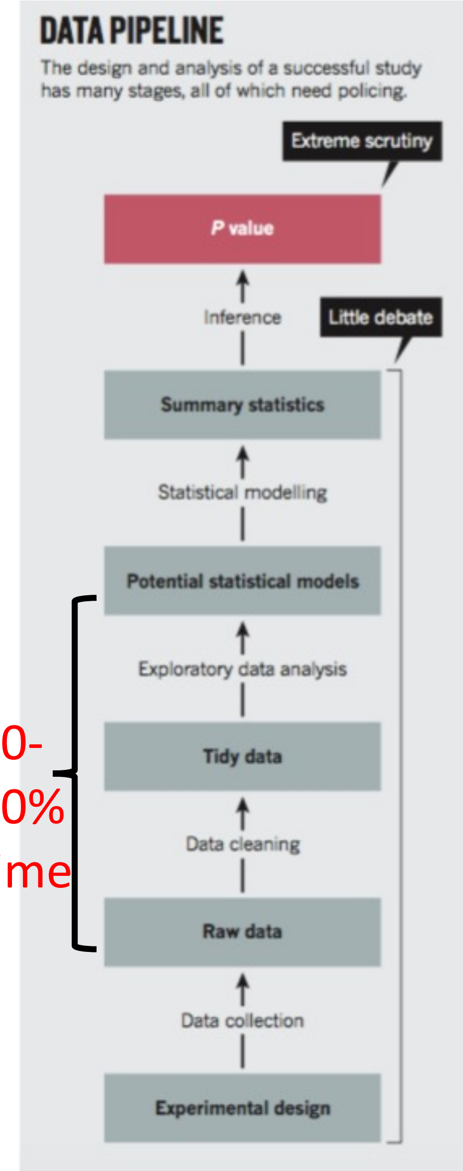


Open Access Article

Organizing and Analyzing Data from the SHARE Study with an Application to Age and Sex Differences in Depressive Symptoms

by Lara Lusa[1,2,*,†] and Marianne Huebner[3]

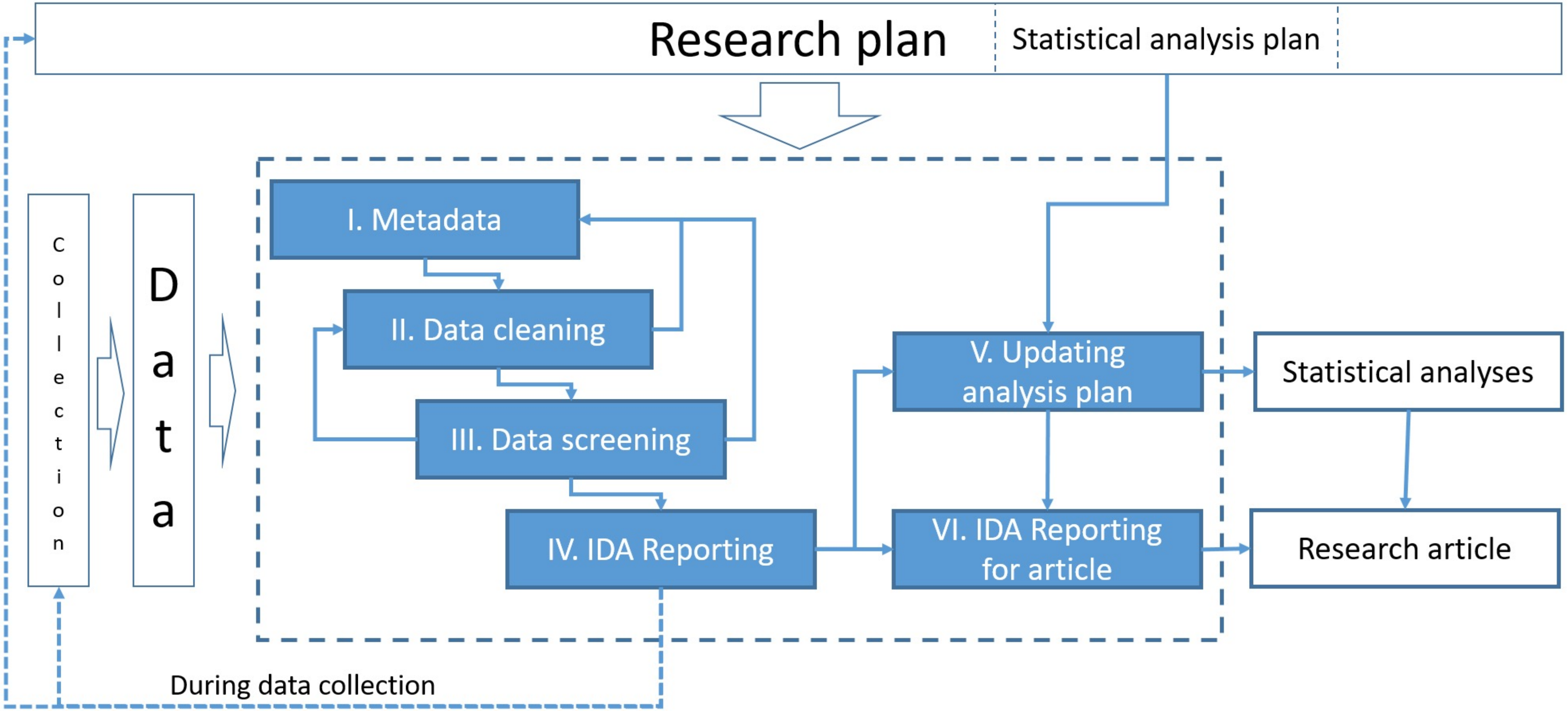
What is Initial Data Analysis?



DATA PIPELINE

The design and analysis of a successful study has many stages, all of which need policing.

50-80% time



Research plan

Statistical analysis plan

Collection

Data

I. Metadata

II. Data cleaning

III. Data screening

IV. IDA Reporting

V. Updating analysis plan

VI. IDA Reporting for article

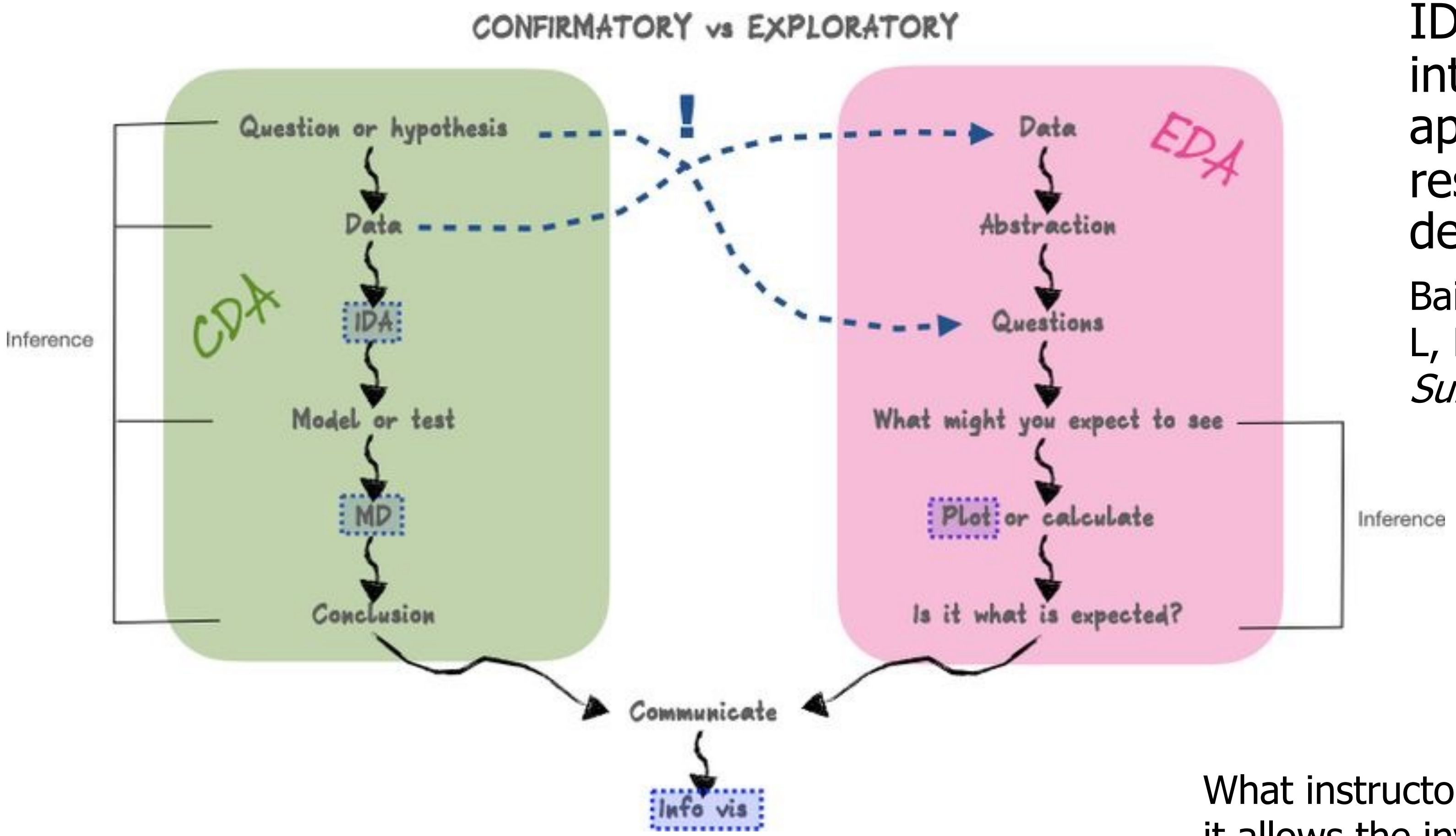
Statistical analyses

Research article

During data collection

Huebner M, le Cessie S, Schmidt CO, Vach W . A contemporary conceptual framework for initial data analysis. Observational Studies 2018; 4: 171-192.

EDA vs CDA vs IDA



Cook, D., Reid, N., & Tanaka, E. (2021). The Foundation is Available for Thinking about Data Visualization Inferentially. *Harvard Data Science Review*

We disagree about how IDA is depicted.

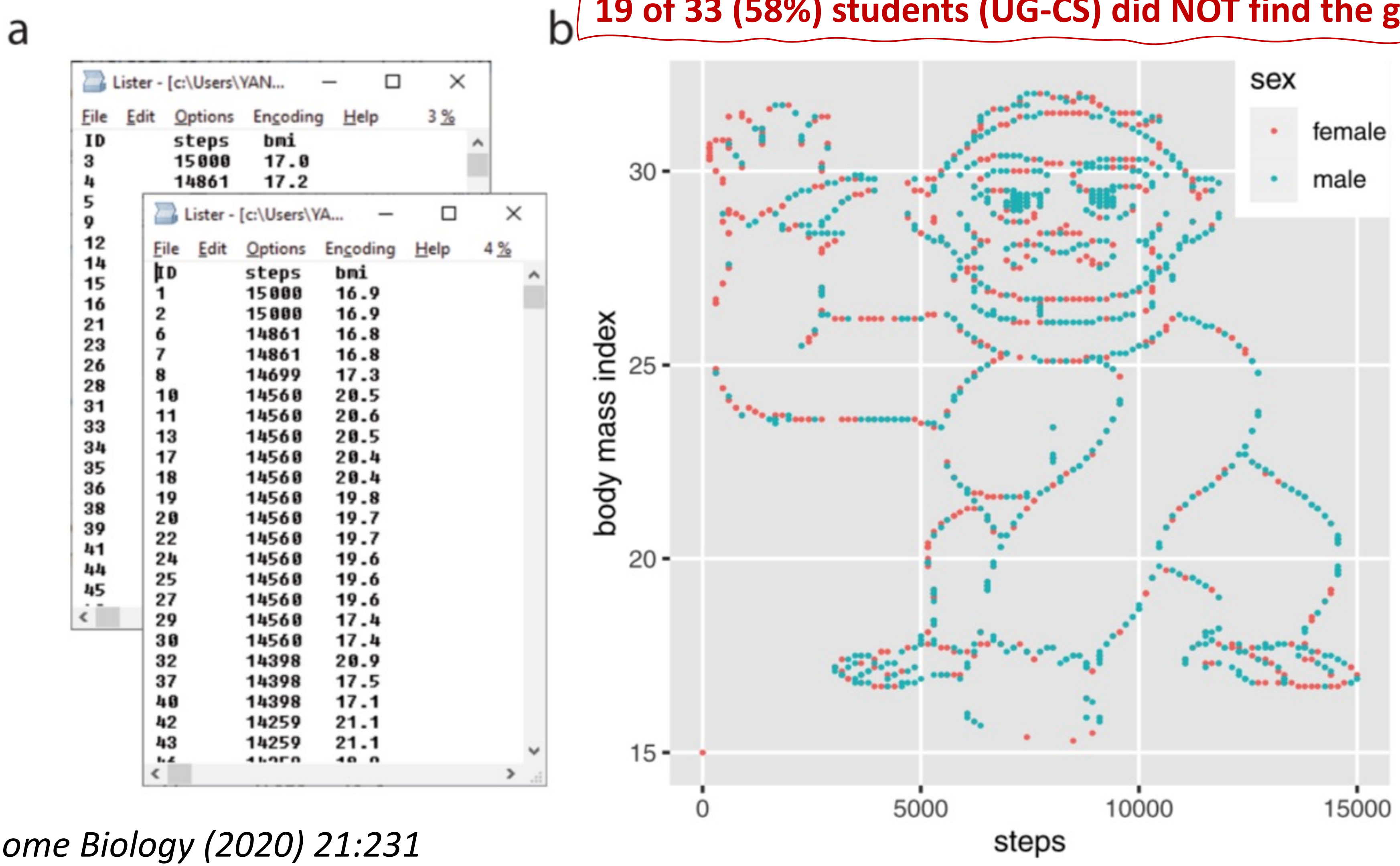
IDA primarily ensures transparency and integrity of preconditions to conduct appropriate statistical analyses in a responsible manner to answer pre-defined research questions.

Baillie M, le Cessie S, Schmidt CO, Lusa L, Huebner M. Ten simple rules for IDA. *Submitted*

What instructors say about EDA: "EDA is important because it allows the investigator to make critical decisions about what is interesting to follow up on and what probably isn't worth pursuing because the data just don't provide the evidence" R. Peng (Coursera course on Data Science)

Can you find the gorilla hidden in your data?

19 of 33 (58%) students (UG-CS) did NOT find the gorilla



```
ID          steps     bmi
3           15000     17.0
4           14861     17.2
5
9
12
14
15
16
21
23
26
28
31
33
34
35
36
38
39
41
44
45
```

```
ID          steps     bmi
1           15000     16.9
2           15000     16.9
6           14861     16.8
7           14861     16.8
8           14699     17.3
10          14560     20.5
11          14560     20.6
13          14560     20.5
17          14560     20.4
18          14560     20.4
19          14560     19.8
20          14560     19.7
22          14560     19.7
24          14560     19.6
25          14560     19.6
27          14560     19.6
29          14560     17.4
30          14560     17.4
32          14398     20.9
37          14398     17.5
40          14398     17.1
42          14259     21.1
43          14259     21.1
```

Yanai et al. Genome Biology (2020) 21:231

Without IDA we have to trace backwards

- Reconstruct the entire sequence of events
- Starting with the output, trace back through the system diagram or sequence of code statements
- Summarize the root causes

Roger D. Peng, Athena Chen, Eric Bridgeford, Jeffrey T. Leek & Stephanie C. Hicks (2021) Diagnosing Data Analytic Problems in the Classroom, Journal of Statistics and Data Science Education

Problems might not be discovered at all

Current projects:
IDA check lists and R code for different settings

1. **Regression without regrets (one time point)**

Leads: G. Heinze, M. Baillie, M. Huebner, a TG2-TG3 collaboration

Scope: Descriptive, explanatory or predictive regression model to relate an outcome variable with a set of independent variables (3-50)

Outcome: Continuous, binary or count

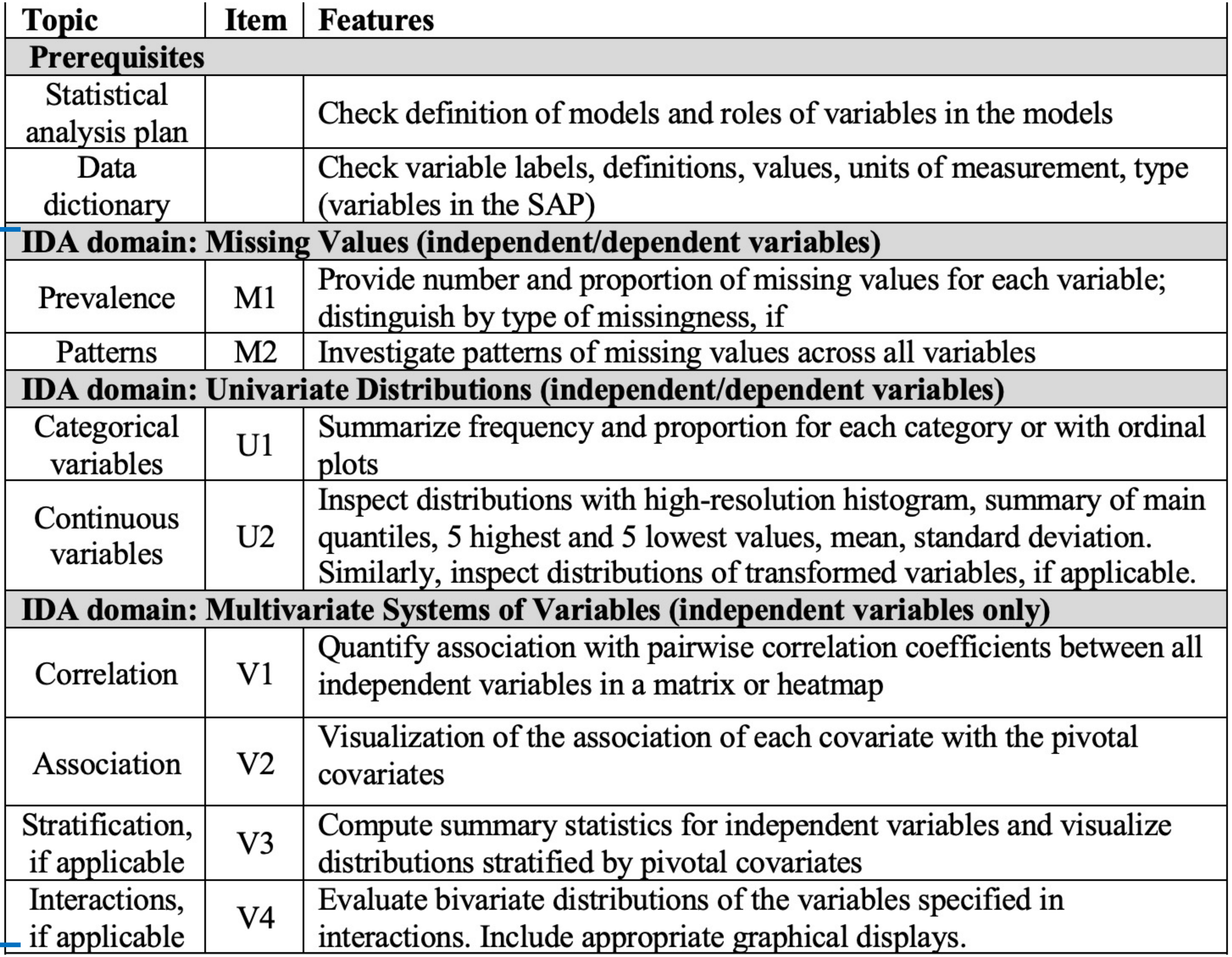
2. **IDA for longitudinal data**

Leads: L. Lusa and M. Huebner (collaboration with Kate Lee, TG1)

Scope: Regression model that uses repeated measurements obtained for individuals

PRE-REQUISITES: SAP + metadata exist; data cleaning has been performed

“Generic“ IDA Plan for a cross-sectional study

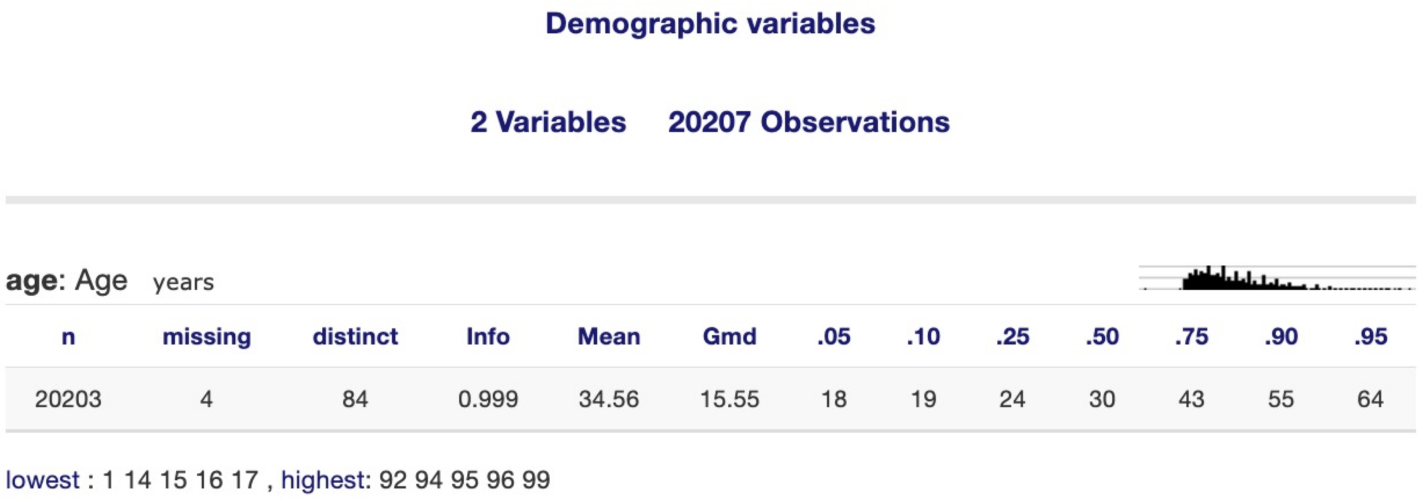


Topic	Item	Features
Prerequisites		
Statistical analysis plan		Check definition of models and roles of variables in the models
Data dictionary		Check variable labels, definitions, values, units of measurement, type (variables in the SAP)
IDA domain: Missing Values (independent/dependent variables)		
Prevalence	M1	Provide number and proportion of missing values for each variable; distinguish by type of missingness, if
Patterns	M2	Investigate patterns of missing values across all variables
IDA domain: Univariate Distributions (independent/dependent variables)		
Categorical variables	U1	Summarize frequency and proportion for each category or with ordinal plots
Continuous variables	U2	Inspect distributions with high-resolution histogram, summary of main quantiles, 5 highest and 5 lowest values, mean, standard deviation. Similarly, inspect distributions of transformed variables, if applicable.
IDA domain: Multivariate Systems of Variables (independent variables only)		
Correlation	V1	Quantify association with pairwise correlation coefficients between all independent variables in a matrix or heatmap
Association	V2	Visualization of the association of each covariate with the pivotal covariates
Stratification, if applicable	V3	Compute summary statistics for independent variables and visualize distributions stratified by pivotal covariates
Interactions, if applicable	V4	Evaluate bivariate distributions of the variables specified in interactions. Include appropriate graphical displays.

Data screening

IDA discoveries

CRASH-2 http://crash2.lshtm.ac.uk/
Data available on Vanderbilt website



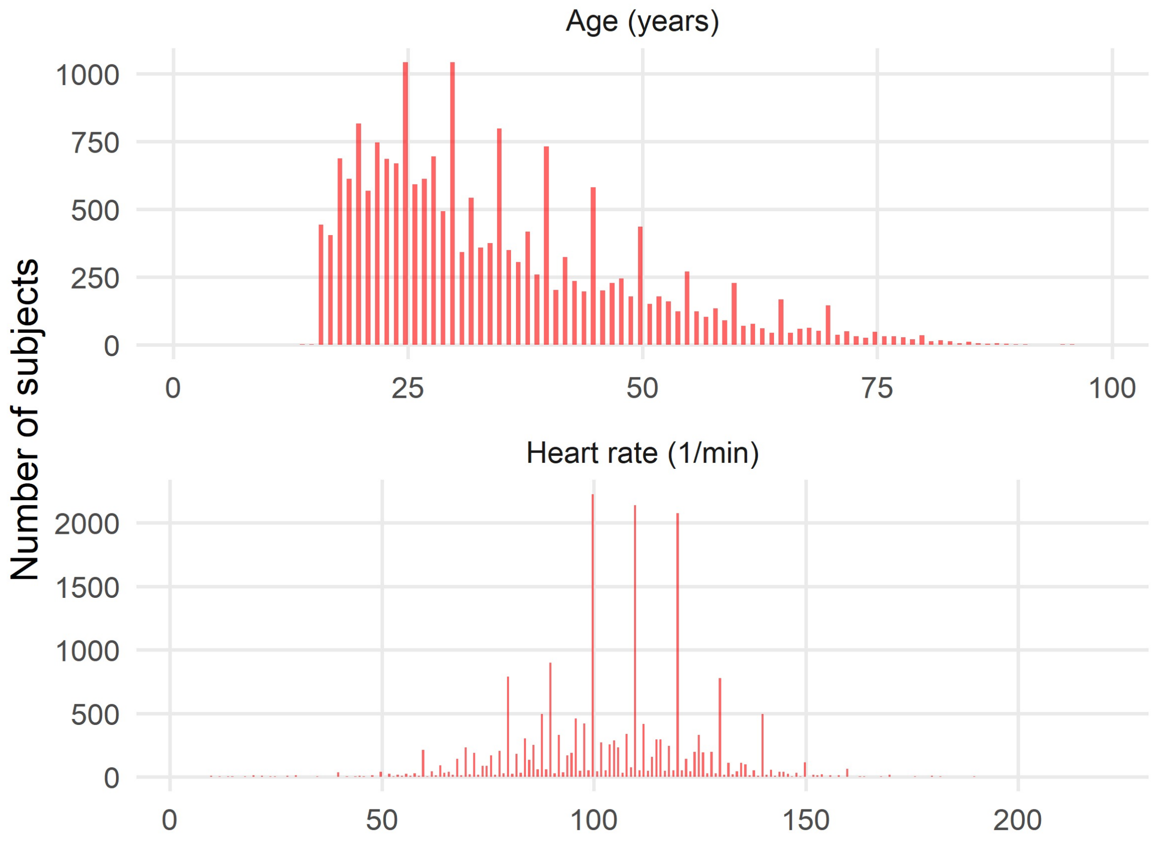
Demographic variables

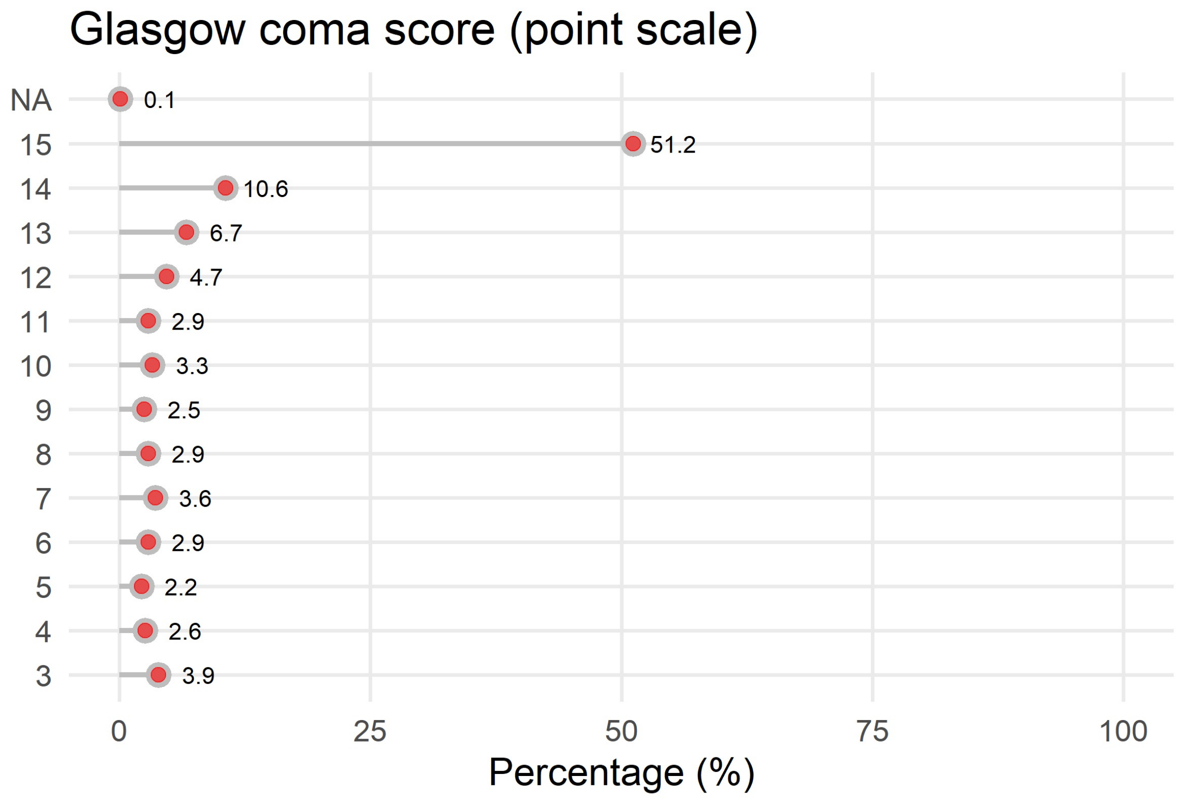
2 Variables   20207 Observations

age: Age years

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
20203	4	84	0.999	34.56	15.55	18	19	24	30	43	55	64

lowest : 1 14 15 16 17 , highest: 92 94 95 96 99

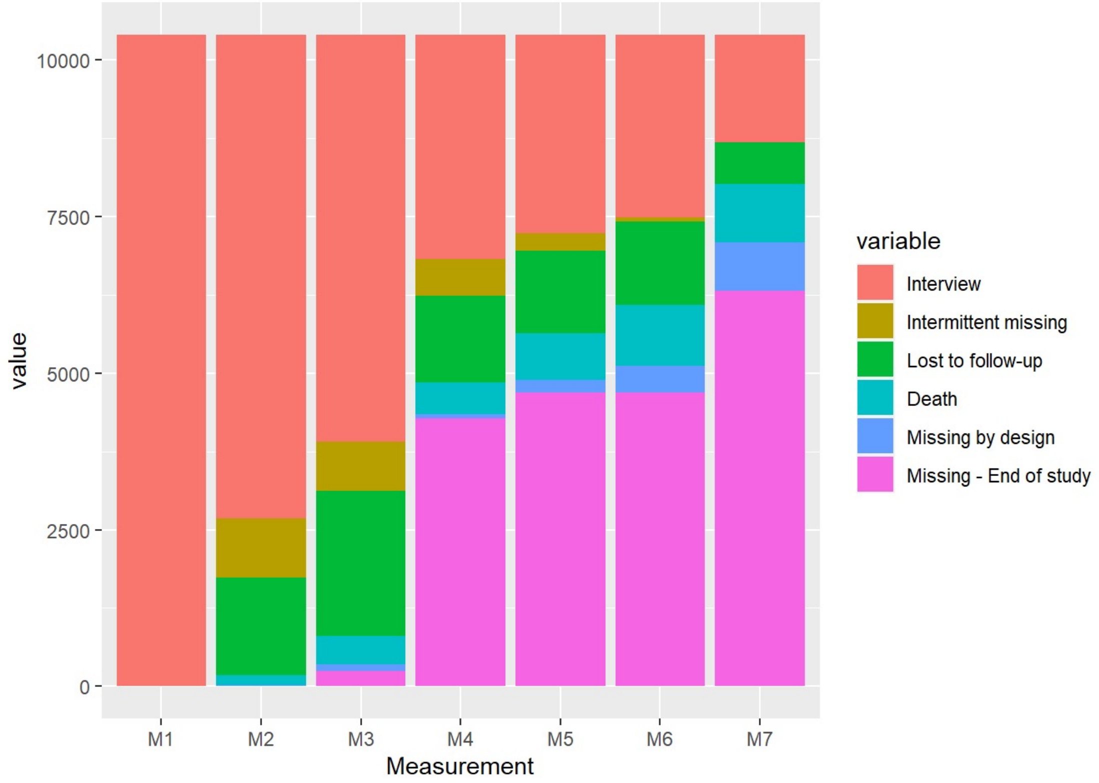




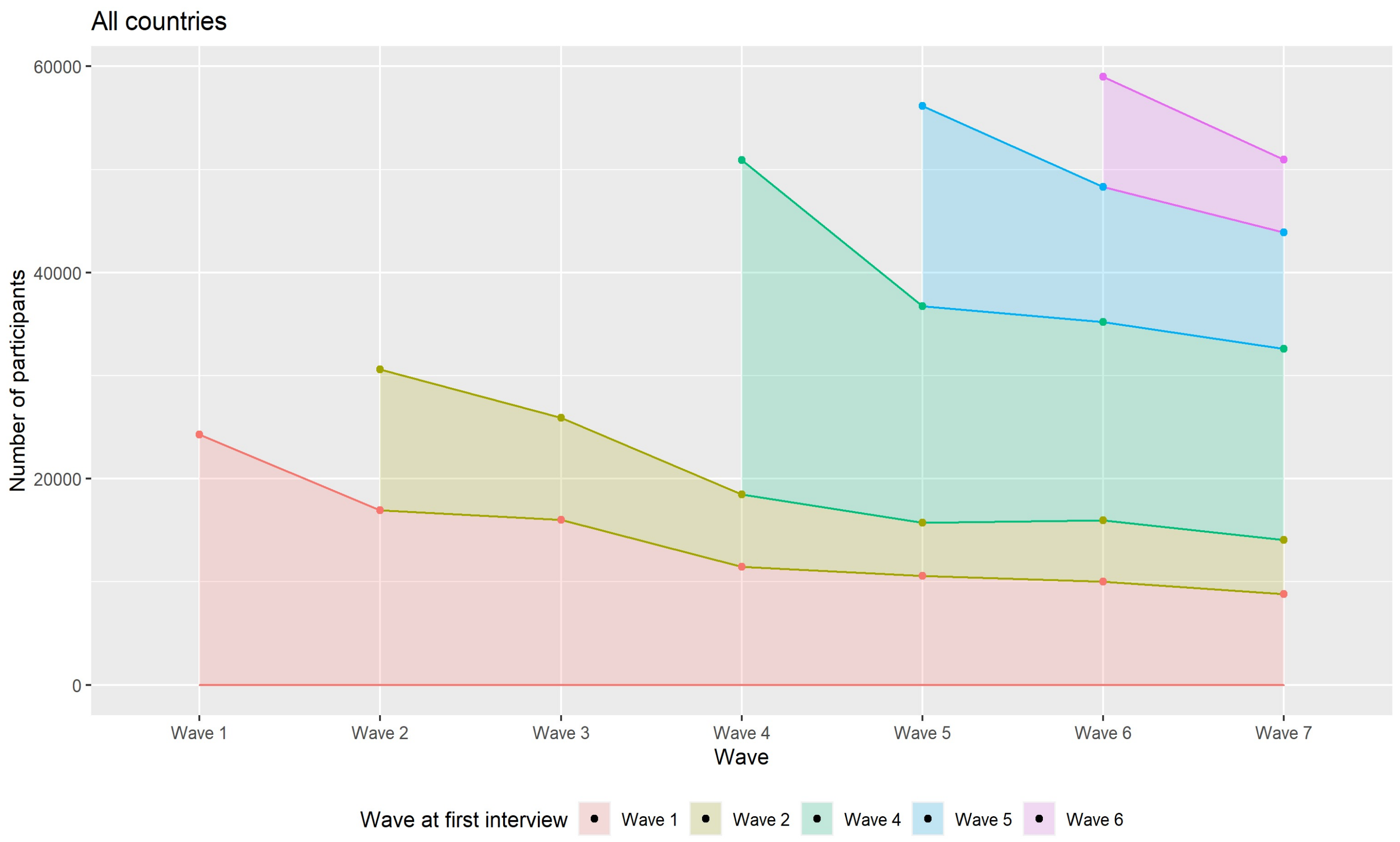
NA = missing

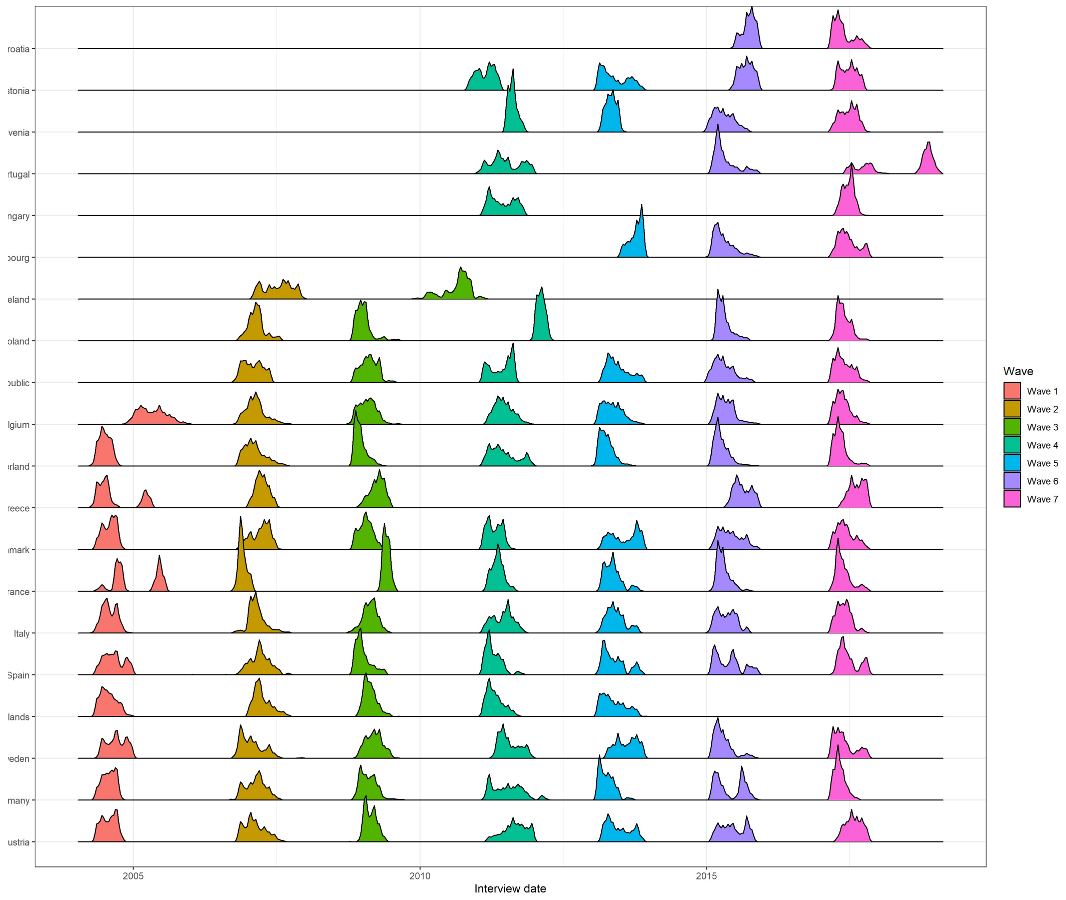
IDA plan for longitudinal data adds challenges

- Participation profile (participation over time)
- Time metrics (calendar time, measurement occasion, age, etc)
- Time-varying covariates
- Missing data (intermittent missingness, death, lost to follow-up, missing by design, missing end of study)



Participation





Possible consequences of IDA

Updating the analysis plan

- Variable omission

Interpretation of results

- Large effect sizes in the context of variable distributions

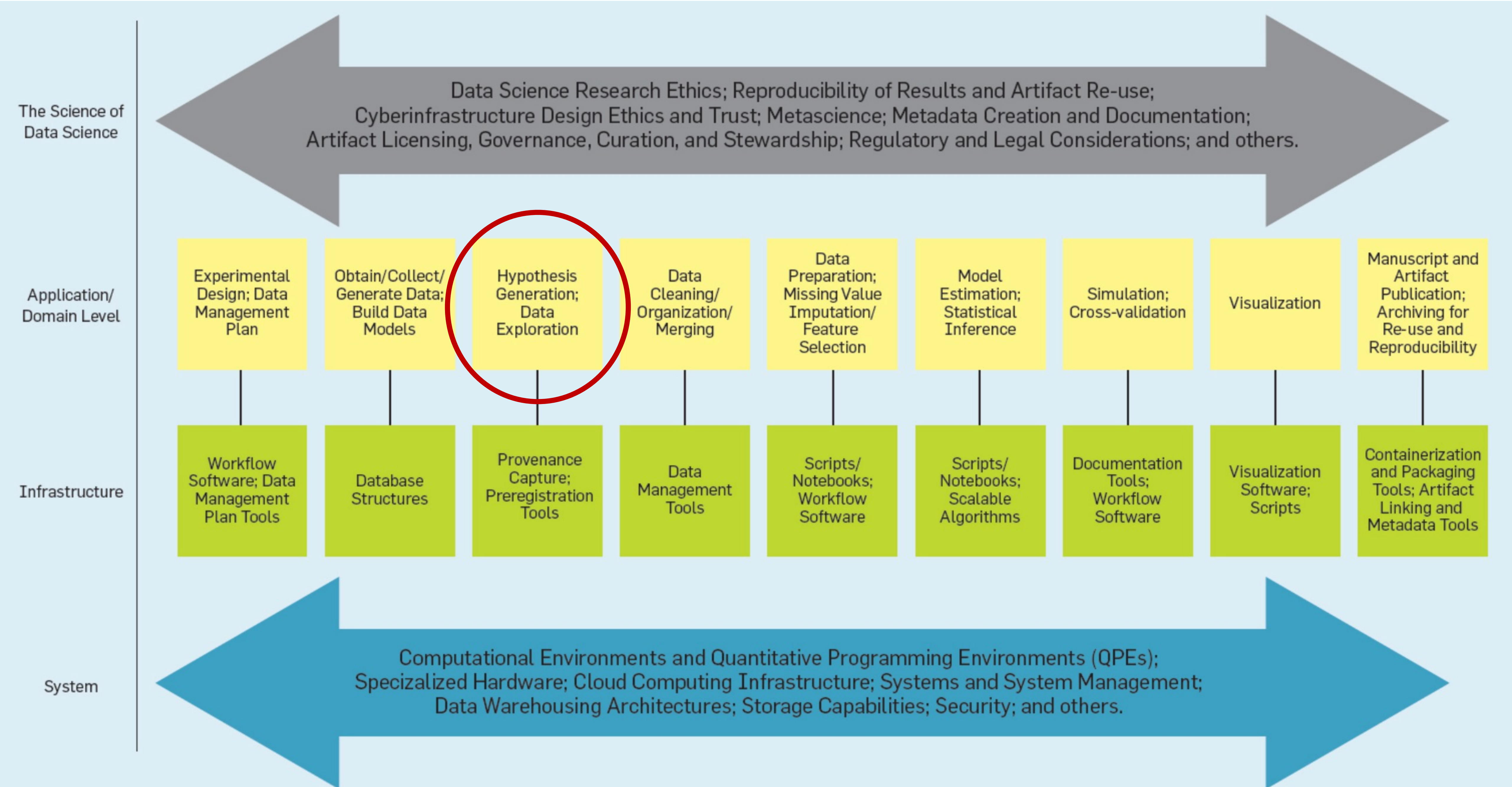
Presentation of data

- Updated data dictionary
- Updated flow diagram with exclusion criteria due to missingness, impossible values, or other reasons
- Table of characteristics of study participants
- Description of IDA findings of data properties that affect the interpretation

Transparency in reporting

- IDA methodology
- Relevant IDA results
- Impact of IDA on interpretation
- IDA driven alterations of the analysis plan
- Full reporting of missingness

Huebner M, Vach W, le Cessie S, Schmidt CO, Lusa L. Hidden Analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses. BMC Med Res Meth 2020; 20:61



The Science of Data Science

Data Science Research Ethics; Reproducibility of Results and Artifact Re-use;
Cyberinfrastructure Design Ethics and Trust; Metascience; Metadata Creation and Documentation;
Artifact Licensing, Governance, Curation, and Stewardship; Regulatory and Legal Considerations; and others.

Application/ Domain Level

- Experimental Design; Data Management Plan
- Obtain/Collect/ Generate Data; Build Data Models
- Hypothesis Generation; Data Exploration
- Data Cleaning/ Organization/ Merging
- Data Preparation; Missing Value Imputation/ Feature Selection
- Model Estimation; Statistical Inference
- Simulation; Cross-validation
- Visualization
- Manuscript and Artifact Publication; Archiving for Re-use and Reproducibility

Infrastructure

- Workflow Software; Data Management Plan Tools
- Database Structures
- Provenance Capture; Preregistration Tools
- Data Management Tools
- Scripts/ Notebooks; Workflow Software
- Scripts/ Notebooks; Scalable Algorithms
- Documentation Tools; Workflow Software
- Visualization Software; Scripts
- Containerization and Packaging Tools; Artifact Linking and Metadata Tools

System

Computational Environments and Quantitative Programming Environments (QPEs);
Specizalized Hardware; Cloud Computing Infrastructure; Systems and System Management;
Data Warehousing Architectures; Storage Capabilities; Security; and others.

Stodden V. Communications of the ACM, July 2020, 63(7)

In Summary

- **IDA is the foundation for statistical modeling**:
  presentation, checking expectations, interpretation, model decisions
- **IDA takes time and planning**
- **IDA needs to be conducted systematically**
- **IDA needs to be reported**

Research studies need both:
Statistical analysis plan + IDA plan

Proposed project TG3-TG8: IDA for survival analysis

Andersen et al. Analysis of time-to-event for observational studies: Guidance to the use of intensity models. Stat Med 2020

Section 2.2 Check list #1, Section 3.3 Check list #2

“the source of the data, what population it represents, what variables are relevant and which among these are available, and data completeness, both with respect to inclusion of subjects and missing data for those that are included.”