STRATOS INITIATIVE


https://mossandfog.com/top-five-tallest-skyscrapers-2017-edition/

Building blocks of efficient initial data analysis and data quality assessments Best practice examples

Carsten Oliver Schmidt,
Lara Lusa, Marianne Huebner
on behalf of TG3

45th ISCB 2024, Thessaloniki

Aim of IDA

The aim of IDA is 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.

Huebner et al. 2018

Aspects of data knowledge ……..

Range violations
Contradictions
Inadmissible values

Volume
Unit / item missingness
Missing patterns
Missing mechanisms

Univariate descriptions
Multivariate descriptions
Associations
    Time-trends
    Process Variables

Aspects of data knowledge ……..

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

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

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

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)

Range violations
Contradictions
Inadmissible values

Volume
Unit / item missingness
Missing patterns
Missing mechanisms

Univariate descriptions
Multivariate descriptions
Associations
- Time-trends
- Process Variables

Aspects of data knowledge ....

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

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

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)

Range violations
Contradictions
Inadmissible values

Volume
Unit / item missingness
Missing patterns
Missing mechanisms

Univariate descriptions
Multivariate descriptions
Associations
- Time-trends
- Process Variables

Gaps in the usage and reporting of multiple imputation for incomplete data: Findings from a scoping review of observational studies addressing causal questions

Rheanna M Mainzer
rheanna.mainzer@unimelb.edu.au
The University of Melbourne
Margarita Moreno-Betancur
Murdoch Children's Research Institute
Cattram D Nguyen
Murdoch Children's Research Institute
Julie A Simpson
The University of Melbourne
John B. Carlin
Murdoch Children's Research Institute
Katherine J Lee
Murdoch Children's Research Institute

Aspects of data knowledge in IDA……..

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

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Check for updates

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

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)

Range violations
Contradictions
Inadmissible values

Volume
Unit / item missingness
Missing patterns
Missing mechanisms

Univariate descriptions
Multivariate descriptions
Associations
 Time-trends
 Process Variables

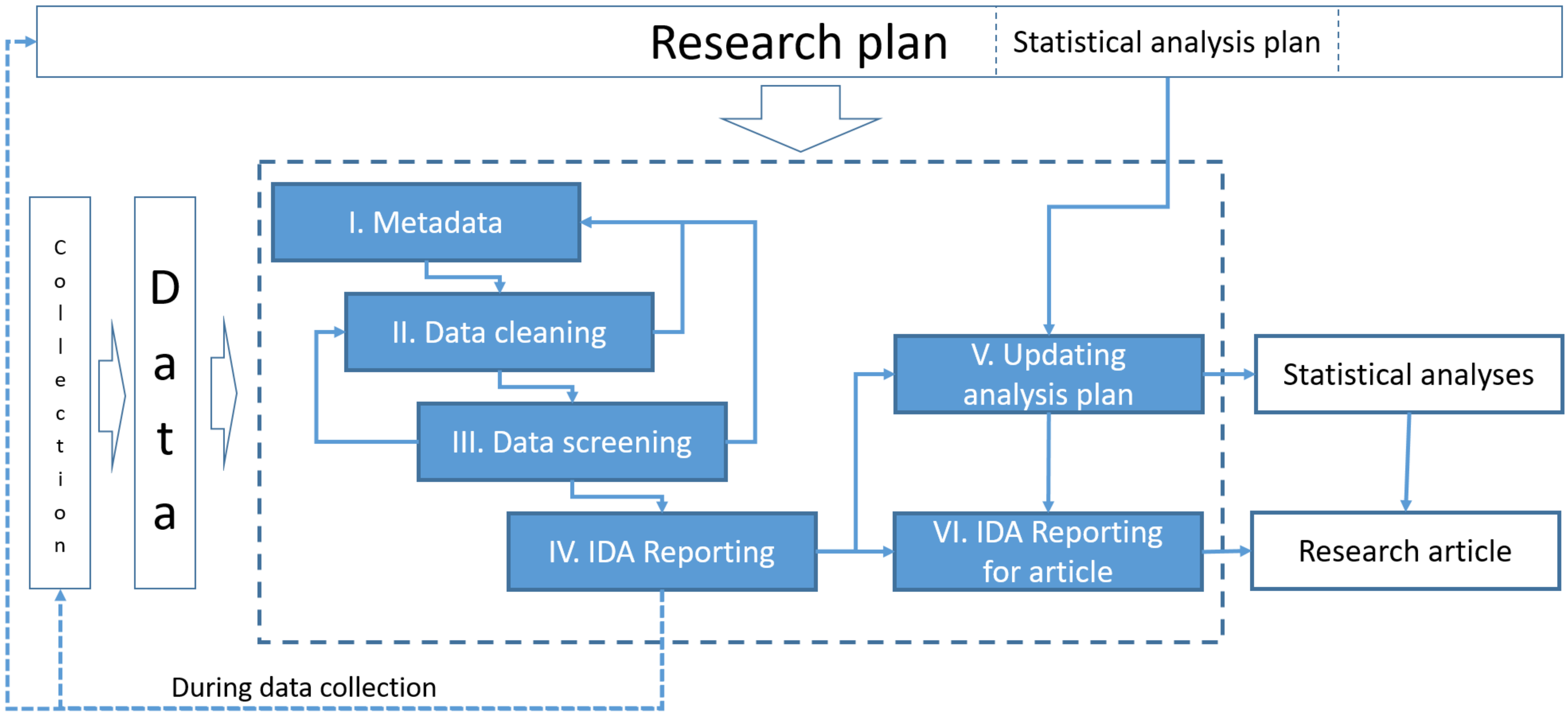
Research Square

Gaps in the usage and reporting of multiple imputation for incomplete data: Findings from a scoping review of observational studies addressing causal questions

Rheanna M Mainzer
rheanna.mainzer@unimelb.edu.au
The University of Melbourne
Margarita Moreno-Betancur
Murdoch Children's Research Institute
Cattram D Nguyen
Murdoch Children's Research Institute
Julie A Simpson
The University of Melbourne
John B. Carlin
Murdoch Children's Research Institute
Katherine J Lee
Murdoch Children's Research Institute

..remain mostly unknown!

IDA framework: Structure the workflow



Huebner et al. 2018

Example 1

PLOS ONE

RESEARCH ARTICLE

Initial data analysis for longitudinal studies to build a solid foundation for reproducible analysis

Lara Lusa[1,2]*, Cécile Proust-Lima[3], Carsten O. Schmidt[4], Katherine J. Lee[5,6], Saskia le Cessie[7,8], Mark Baillie[9], Frank Lawrence[10], Marianne Huebner[10,11], on behalf of TG3 of the STRATOS Initiative[¶]

1 Department of Mathematics, Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Koper, Capodistria, Slovenia, **2** Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia, **3** Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR1219, Bordeaux, France, **4** Institute for community Medicine, SHIP-KEF University Medicine of Greifswald, Greifswald, Germany, **5** Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Melbourne, Australia, **6** University of Melbourne, Melbourne, Australia, **7** Department of Clinical Epidemiology and Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands, **8** Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands, **9** Novartis, Basel, Switzerland, **10** Center for Statistical Training and Consulting, Michigan State University, East Lansing, MI, United States of America, **11** Department of Statistics and Probability, Michigan State University, East Lansing, MI, United States of America

Example 1: IDA in longitudinal studies - Plan

Table 1. Initial data analysis checklist for data screening in longitudinal studies.

Topic	Item	Features
IDA screening domain: Participation profile		
Time frame	P1	Provide number of time points and intervals at which measurements are taken, using the time metric that best reflects the time from inclusion in the study, or calendar time in studies that involve long enrollment times. Highlight the differences between the time of first measurements and follow-up times.
Time metric	P2	Describe the time metric and corresponding time points specified in the analysis strategy, if different from the time metric described in P1.
Participants	P3	Provide the number of participants who attended the assessment by time metric(s).
Optional extensions: Participation Profile		
Other time metrics	PE1	Use different time metric(s) to describe the time frame of the study, if applicable and appropriate, e.g. calendar time or data collection visits.
IDA screening domain: Missing data (outcome variable and explanatory variables)		
Non-enrollment	M1	Describe the non-enrolled, i.e., the participants that were selected but did not enter the study (and the reasons, if available), if applicable.
Drop-out	M2	Describe the participants who dropped out from the study during the follow-up (loss to follow-up and other possible reasons: death, withdrawal, missing by design, if applicable).
Intermittent visit missingness	M3	Describe the participants that have missing data for some of the measurements (intermittent, occasional omission, but do not drop out out of the study).

Example 1: IDA in longitudinal studies - Check

Table 2. Number of interviews per participant.

Interviews per participant	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Frequency	965	966	1508	527	307	685	494
Proportion	0.18	0.18	0.28	0.10	0.06	0.13	0.09

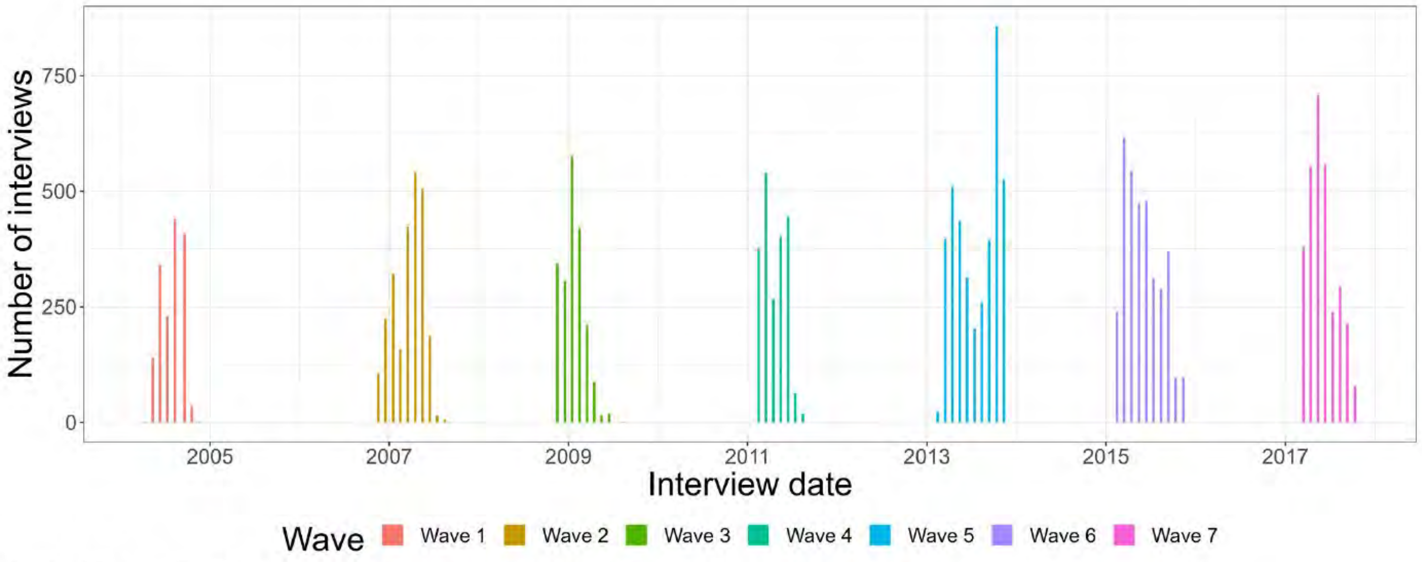
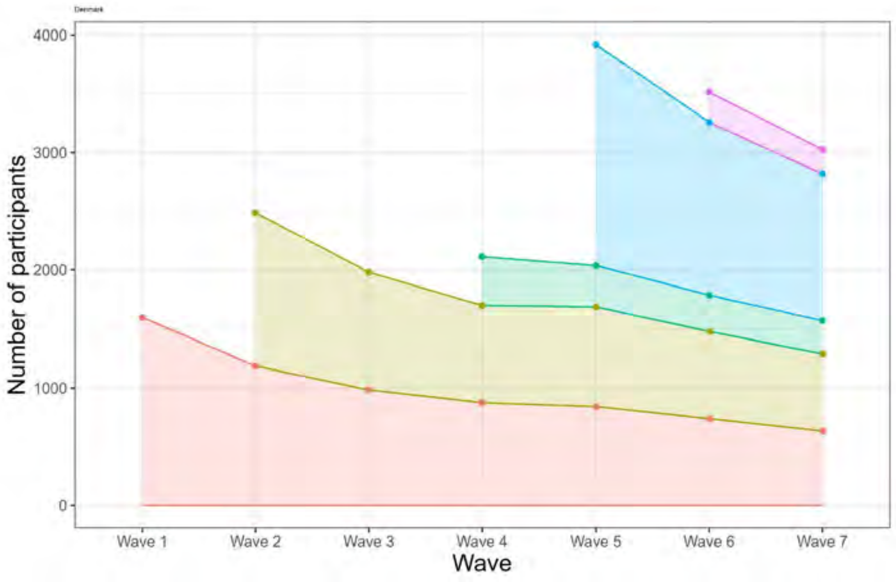
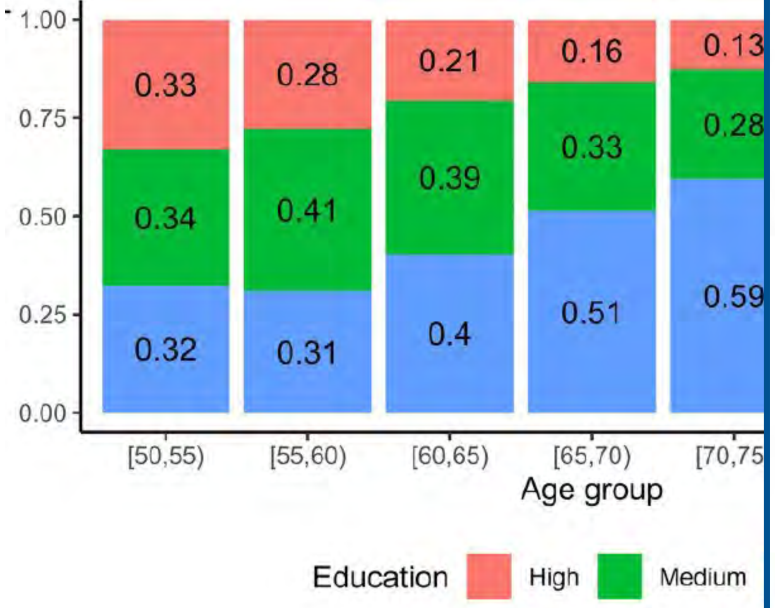


Fig 1. Distribution of the number of interviews carried out in Denmark in the SHARE study in time.



Example 1: IDA in longitudinal studies - Check

General popula



Baseline characteristics by type of missingness.

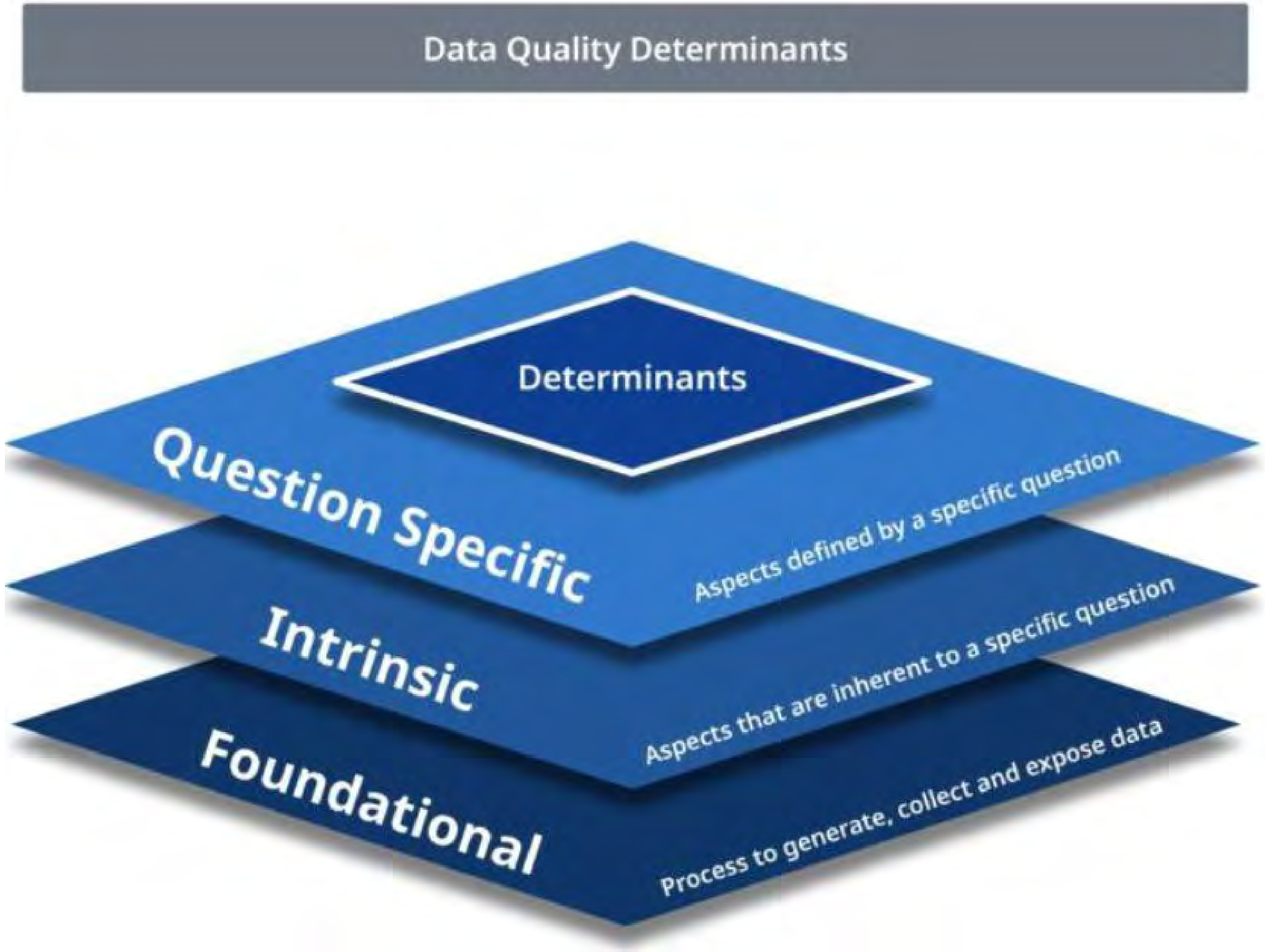
	N	Complete *N=2681*	Death *N=978*	Intermittent missing *N=476*
gender : Female	5452	0.54 1440/2681	0.51 494/978	0.50 240/476
age_int	5452	52.00 58.00 66.00 60.28 ± 8.79	66.00 75.00 81.00 73.29 ± 10.44	52.00 58.00 64.00 59.55 ± 8.18
age_int_cat : 50-59	5452	0.54 1452/2681	0.12 120/978	0.59 282/476
60-69		0.29 780/2681	0.21 202/978	0.27 127/476
70-80		0.14 384/2681	0.41 399/978	0.13 62/476
80+		0.02 65/2681	0.26 257/978	0.01 5/476
weight	5361	66.0 76.0 86.0 77.2 ± 15.2	62.5 71.0 81.0 72.7 ± 15.0	65.0 76.0 85.0 77.1 ± 15.6
height_imp	5418	165.00 172.00 178.00 171.82 ± 9.04	163.00 169.00 175.00 169.34 ± 8.80	165.00 172.00 178.00 171.66 ± 8.98
education_imp : Low	5428	0.17 447/2678	0.38 371/969	0.19 90/472
Medium		0.38 1019/2678	0.39 375/968	0.41 195/472
High		0.45 1212/2678	0.23 223/968	0.40 187/472
pa_vig_freq	5423	0.67 1798/2677	0.35 339/965	0.66 311/473
pa_low_freq	5422	0.94 2512/2677	0.73 707/964	0.95 447/473
cusmoke_imp : Yes	5423	0.22 590/2679	0.34 327/963	0.27 126/472
maxgrip	5272	29.0 36.0 48.0 38.5 ± 12.5	21.5 29.0 38.0 30.3 ± 11.9	28.0 37.5 49.0 38.5 ± 13.1

a b c represent the lower quartile *a*, the median *b*, and the upper quartile *c* for continuous
is the number of non-missing values.

Example 1: IDA in longitudinal studies - Act

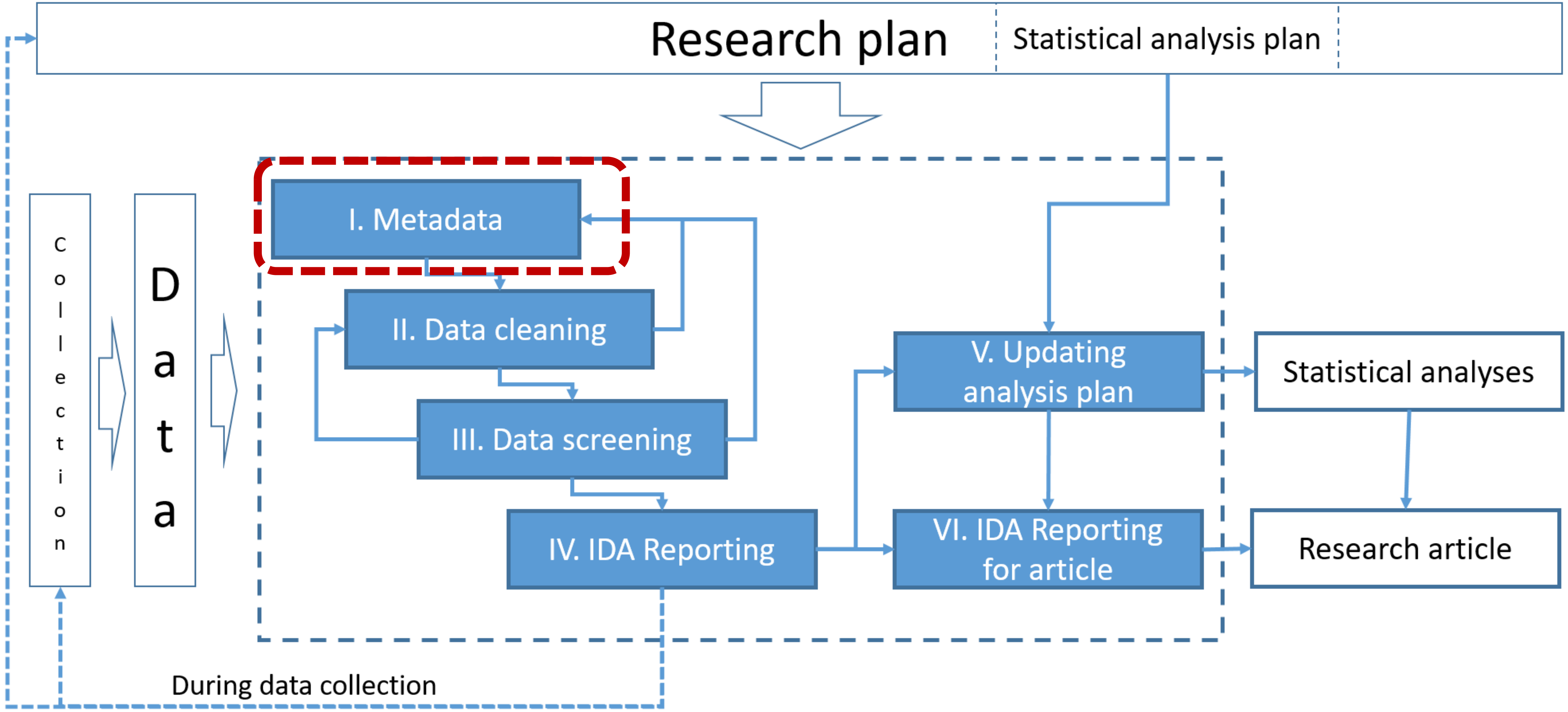
Item	Topic	Consequences	Actions
Participation profile			
P1	Most participants had four or less measurement occasions (74%), 19% were measured only once.	Lack of information for the identification of very flexible shapes of trajectories at the individual level.	The number of random effects that can be included in the mixed model should be limited to three at most. The small number of repeated measurements may prevent the inclusion of an autocorrelation process.
Missing data			
M1 and ME1	Responders had substantially higher education than the target population, even when age and sex were taken into account.	If sampling bias is not taken into account, this could lead to lack of generalization to the entire population.	Statistical models need to account for the selection bias; this could be weighting approaches or adjustment for education.
M2	About 20% of participants were lost to follow-up after first interview, about 35% after 12 years. Participants who dropped out of the study for reasons other than death had lower education and less healthy habits than those that remained in the study.	If the attrition mechanism is not appropriately taken into account in the statistical model, this could lead to biased results.	Methods that are robust to missing data mechanism are needed. With mixed models, the results will be robust to missing data predicted by the observations. Otherwise, joint models may be explored [25].
ME2 and ME3	Deaths were common during the follow-up period in the study that includes an ageing population. For example, about 50% of the participants aged 80 or more at inclusion were dead after 6 years of follow-up. The trajectories of the outcome variable of participants that died differed from those that survived during follow-up. The characteristics of the participants that died were as expected, the quality of reporting of deaths was good.	If the deaths are not appropriately taken into, this could lead to biased results.	Random effect models can be used if deaths are assumed to be predictable by the observed outcome trajectories, while joint models with death as an event may assume a dependency based on unobserved outcomes values. Joint models for competing causes of drop-out might be used if both loss to follow-up and deaths are assumed to depend on the underlying outcome. A model assessing jointly the risk of drop-out (possibly by nature of drop-out—loss of follow-up or death) could be envisaged as a sensitivity analysis.

IDA and data quality



Data Quality Framework for EU medicines regulation 2023

IDA framework: Structure the workflow



Huebner et al. 2018

Information perspective

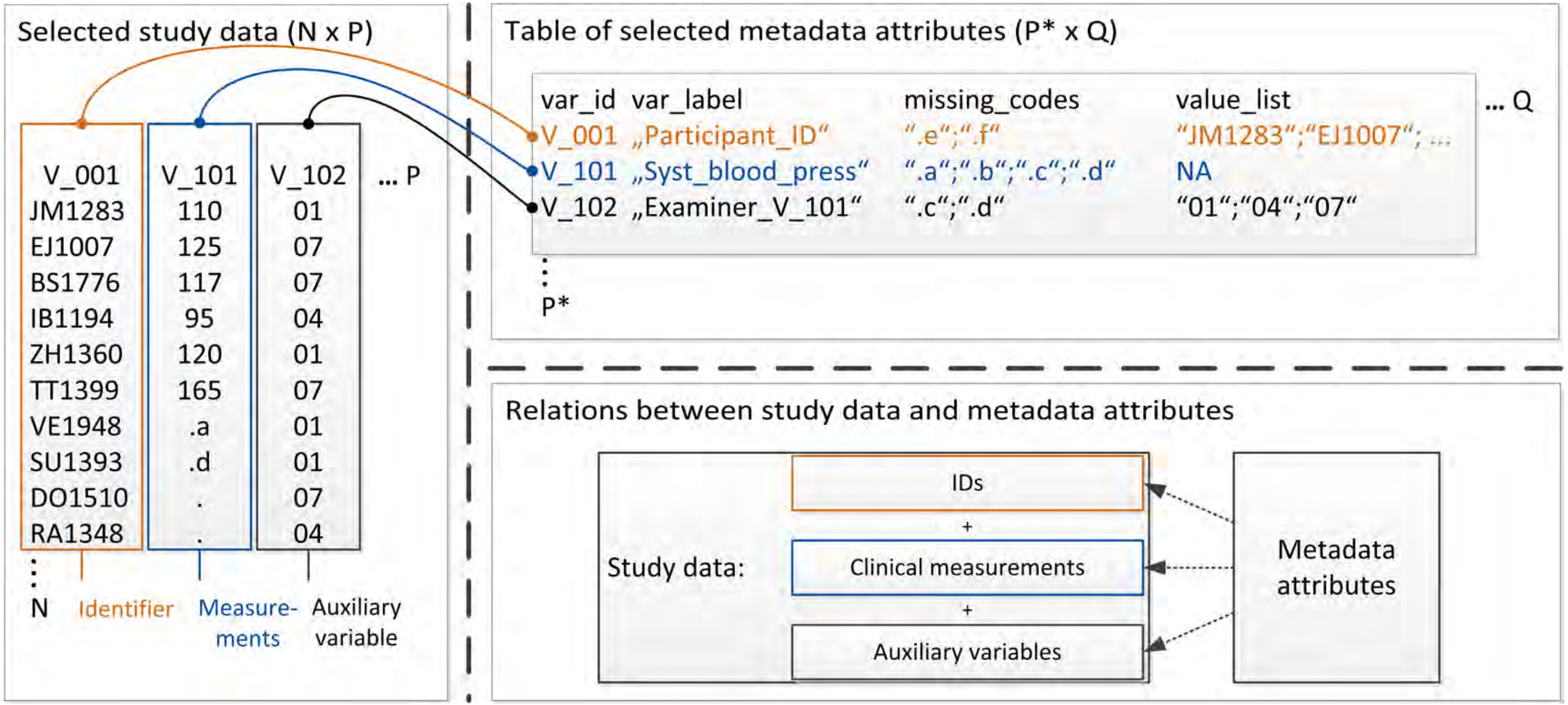
Input
Metadata

Data Quality
Assessments
IDA

Output
Assessment Results

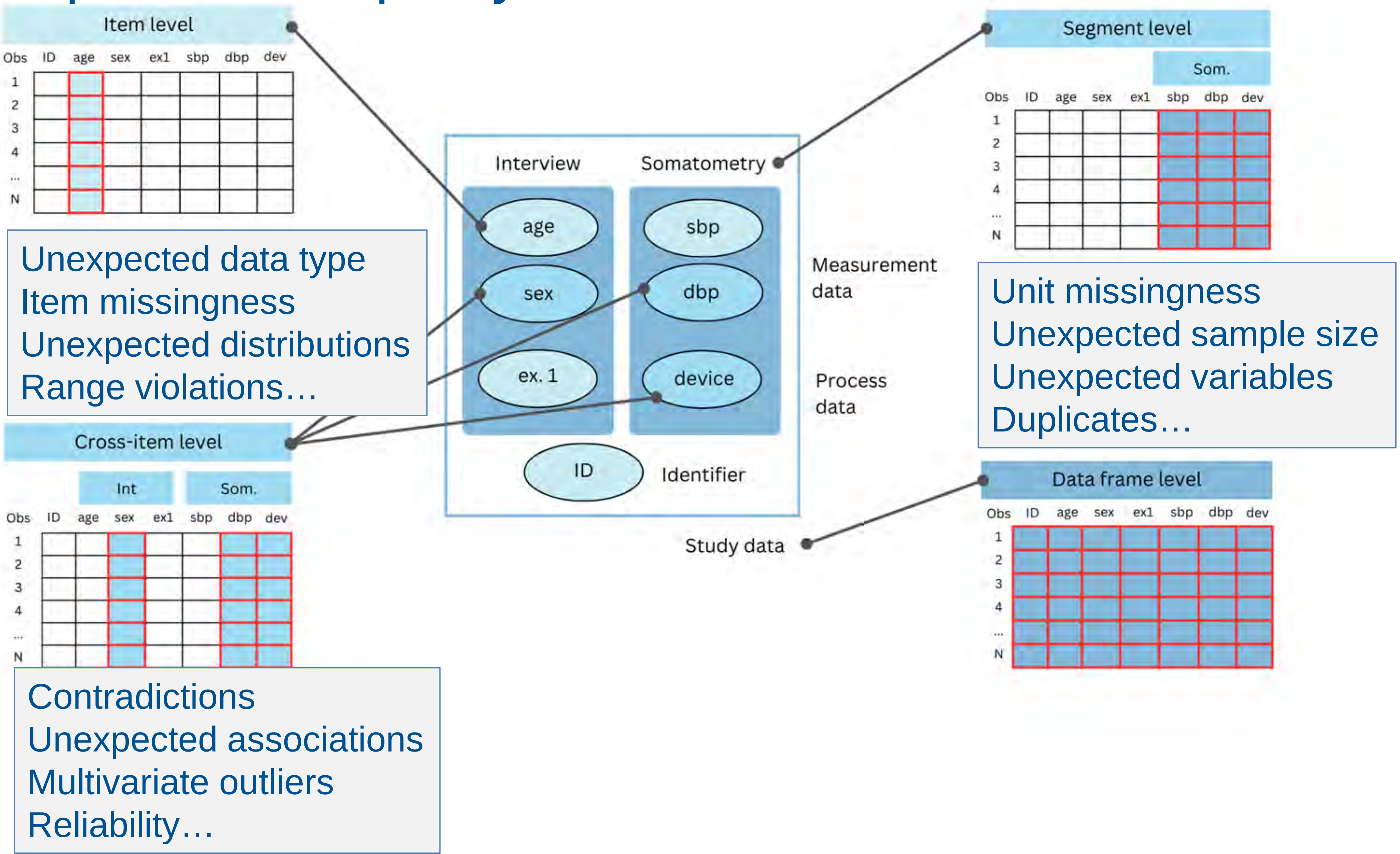
Machine Readibility

Information perspective

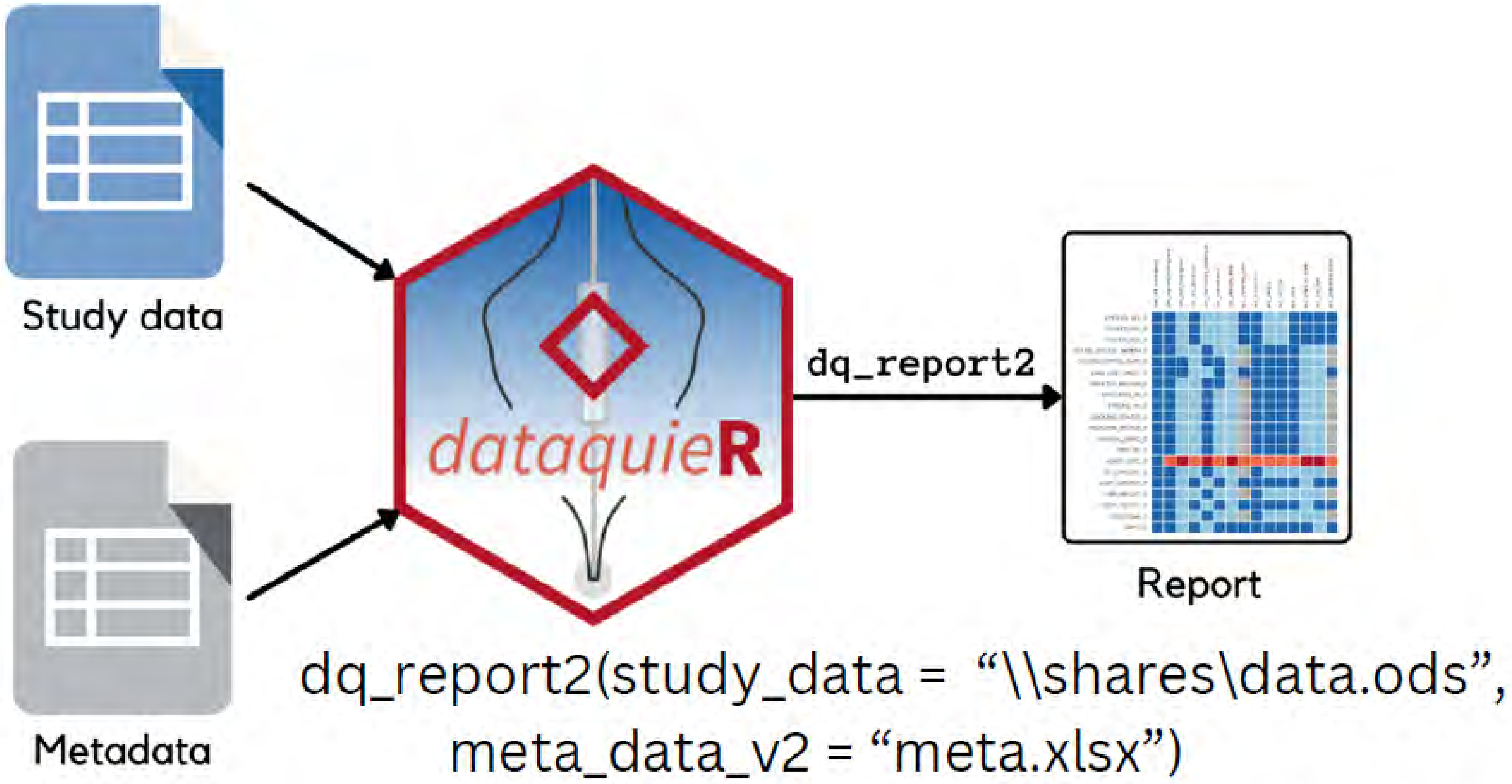


Richter et al. 2019

Example 2: Data quality assessments – Plan

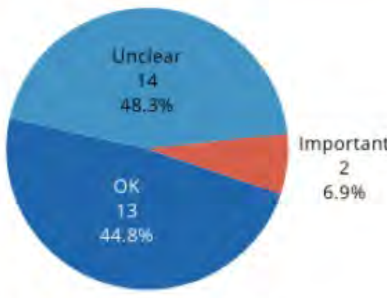


Example 2: Data quality assessments – Check

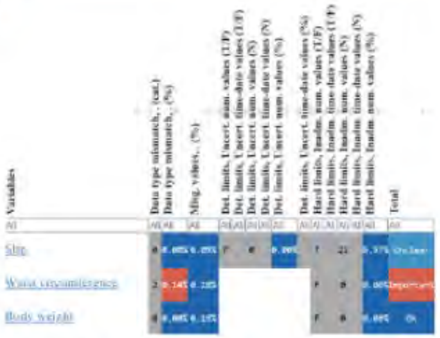


Example 2: Data quality assessments – Check

1. Data quality overview

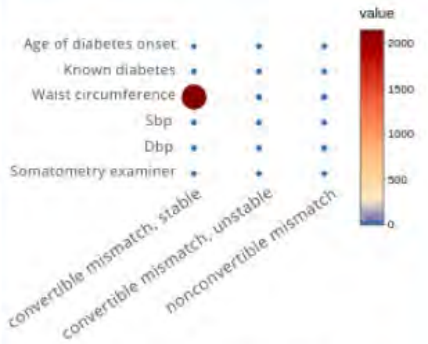


Percentage of variables per quality categories



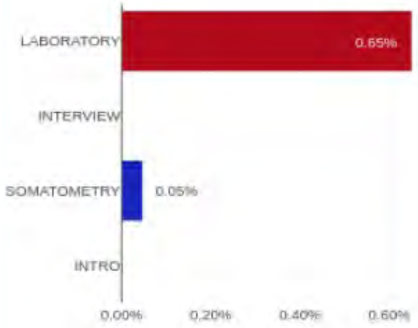
Targeted quality indicators, potential issues, and applicability problems

2. Integrity



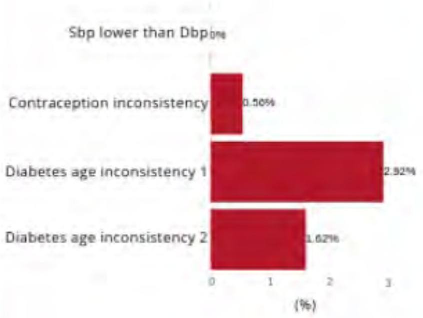
Data type mismatch

3. Completeness



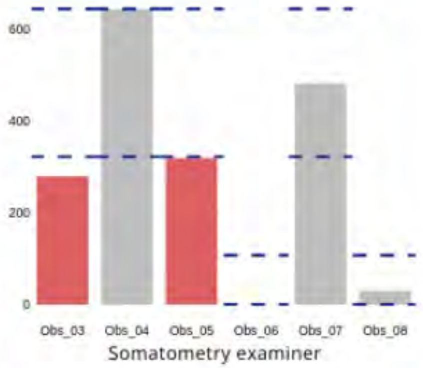
Missing values per segment

4. Consistency



Contradictions

5. Accuracy



Unexpected proportions

See a sample report

Conclusion

Proper IDA (including DQ assessments)

1. … is the foundation for correct modeling
by providing comprehensive knowledge about data properties and issues

2. … takes time
yet, finding problems after modeling takes MORE time

3. … requires appropriate information management
to ensure a comprehensive coverage of all potential aspects

4. … needs to be reported
to ensure transparent and sustainable sciences

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

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

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

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-initiati

'Initial data analysis for longitudinal studies to build a solid foundation for reproducible analysis

Lara Lusa[1,2*], Marianne Huebner [3,4], Carsten O. Schmidt[5], Katherine J. Lee[6,7], Saskia le Cessie[8,9], Mark Baillie[10], Frank Lawrence[4], Cécile Proust-Lima[11], on behalf of TG3 of the STRATOS Initiative ¶

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

Research Square

Regression without regrets – initial data analysis is an essential prerequisite to multivariable regression

Georg Heinze (georg.heinze@meduniwien.ac.a
Medical University of Vienna
Mark Baillie
Novartis (Switzerland)
Lara Lusa
University of Primorska
Willi Sauerbrei
University of Freiburg
Carsten Oliver Schmidt
University Medicine of Greifswald
Frank E. Harrell
Vanderbilt Un
Marianne Huet

A Contemporary Conceptual Framework for Initial Data Analysis

ner@stt.msu.edu
Cessie@lumc.nl
Bioinformatics
uni-greifswald.de

STRATOS INITIATIVE TG3

PLOS COMPUTATIONAL BIOLOGY

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

EDITORIAL

Ten simple rules for initial data analysis

Mark Baillie[1], Saskia le Cessie[2], Carsten Oliver Schmidt[3], Lara Lusa[4], Marianne Huebner[5*], for the Topic Group "Initial Data Analysis" of the STRATOS Initiative¶