STRATOS INITIATIVE

Universitätsmedizin GREIFSWALD

Integrating initial data analysis into statistical analysis plans

Carsten O. Schmidt
on behalf of TG3

ISCB 2025, Basel

Scan for IDA resources: TG3

What is Initial Data Analysis (IDA)?

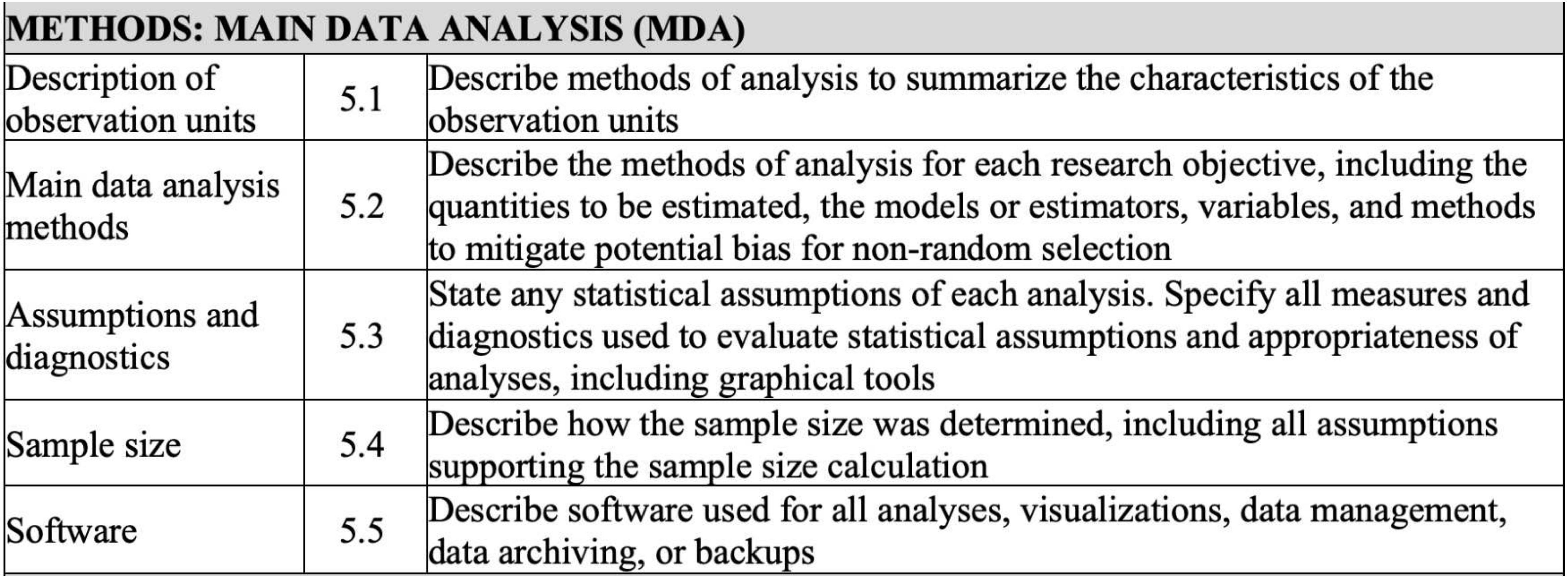
IDA = systematic process to provide reliable knowledge about the data to determine the suitability of the data for the main data analysis

- Aligned with the research aims and the main data analysis
- Does NOT include hypothesis generating activities
- Does NOT include assessing associations between predictors and outcomes

IDA framework: metadata setup; data cleaning; data screening; initial data reporting; documenting and reporting IDA.

The foundation matters……

Statistical Analysis Plan for Observational Studies

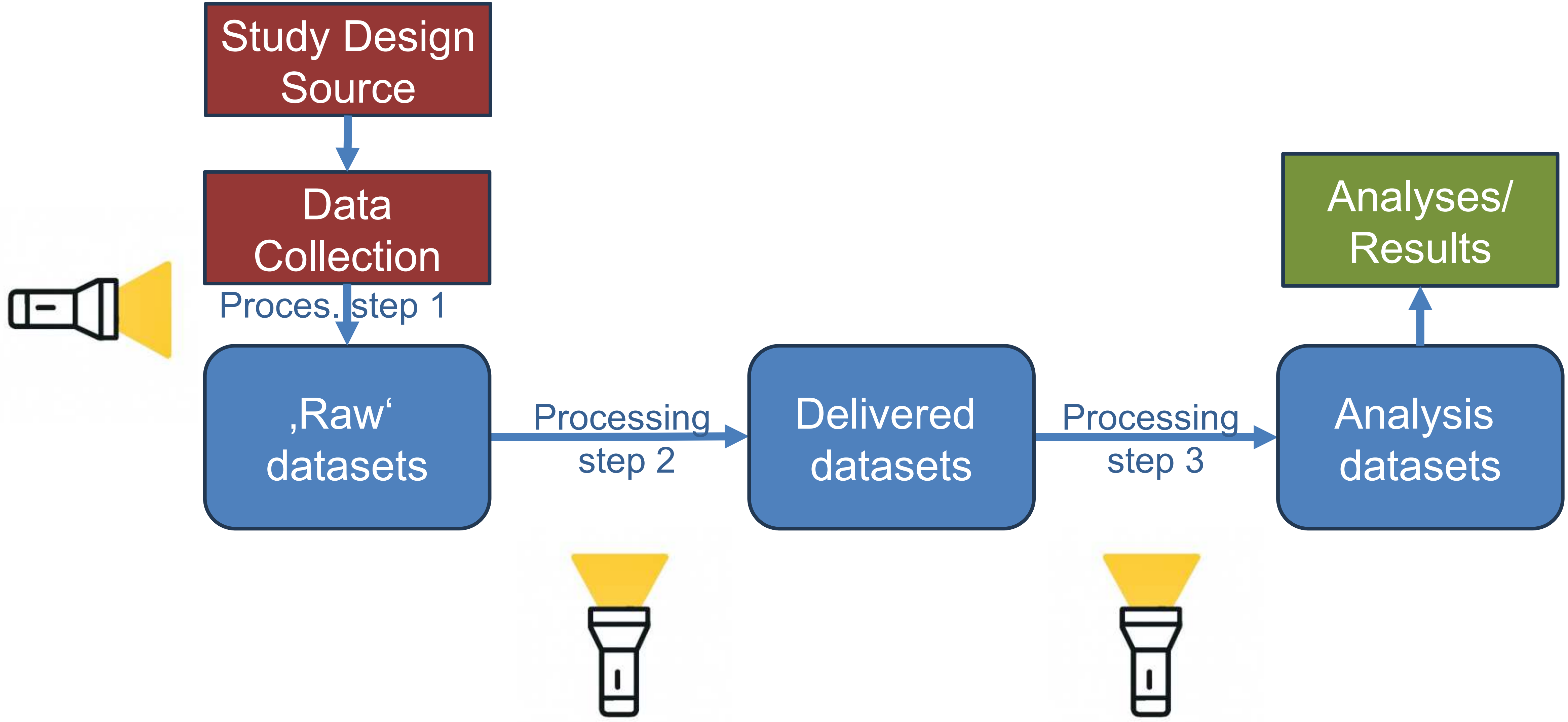


METHODS: MAIN DATA ANALYSIS (MDA)		
Description of observation units	5.1	Describe methods of analysis to summarize the characteristics of the observation units
Main data analysis methods	5.2	Describe the methods of analysis for each research objective, including the quantities to be estimated, the models or estimators, variables, and methods to mitigate potential bias for non-random selection
Assumptions and diagnostics	5.3	State any statistical assumptions of each analysis. Specify all measures and diagnostics used to evaluate statistical assumptions and appropriateness of analyses, including graphical tools
Sample size	5.4	Describe how the sample size was determined, including all assumptions supporting the sample size calculation
Software	5.5	Describe software used for all analyses, visualizations, data management, data archiving, or backups

Assumes “appropriate” dataset →

- Changes mid-analysis
- Ad-hoc decisions – non-transparent
- Time consuming – repeating analyses

Under standing data structures and processing steps



Study Design Source

Data Collection

Proces. step 1

‚Raw‘ datasets

Processing step 2

Delivered datasets

Processing step 3

Analysis datasets

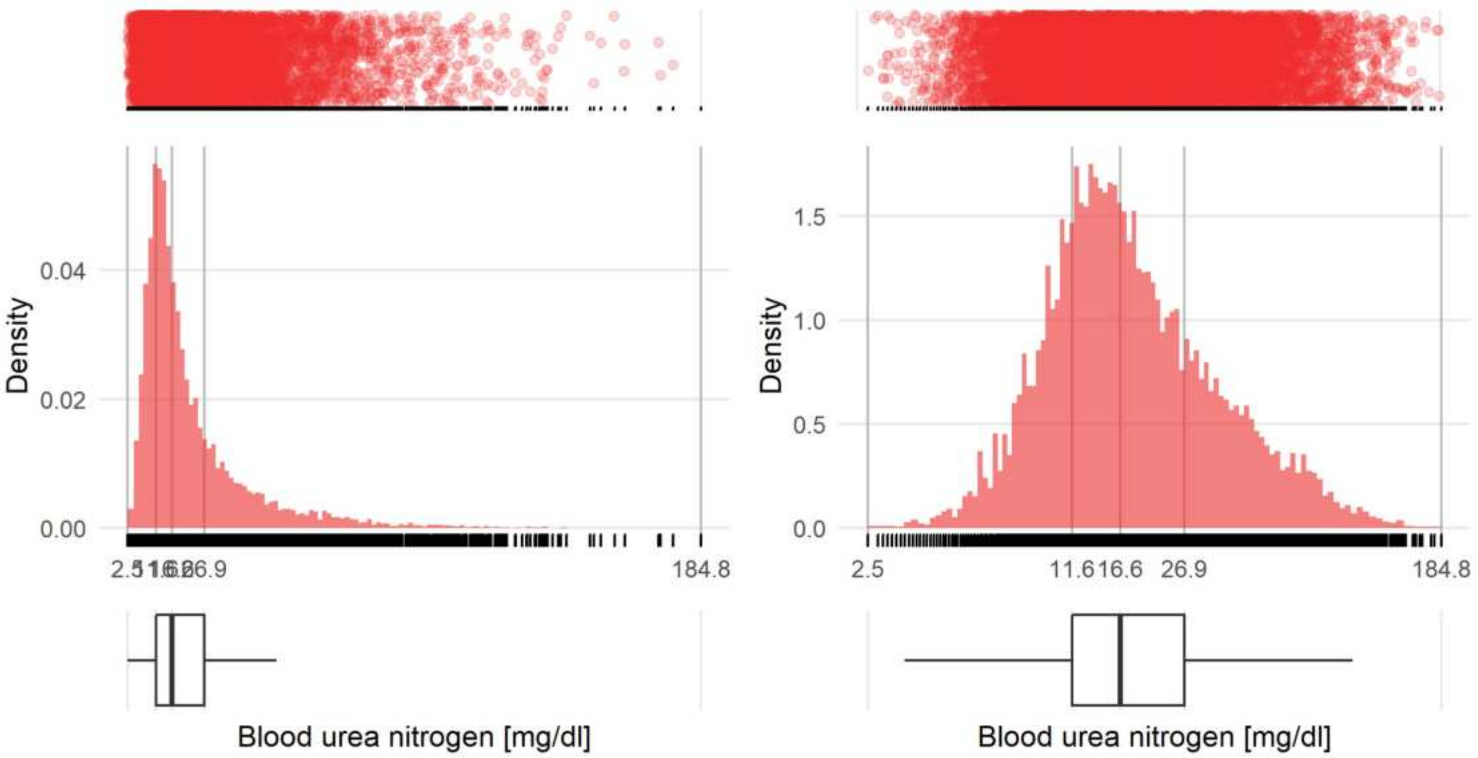
Analyses/ Results

Scan for worked example

Example: Univariable distributions

Univariate summary of Blood urea nitrogen [mg/dl]

original [left] vs. pseudo-log transformed scale [right]



All observed values, the distribution and the, min, max and interquartile range are reported
n = 14519 subjects displayed. 172 subjects with missing values are not presented. Pseudo-log transformation is suggested.

Log transformation to stabilize the distribution of a predictor

Scan for worked example

Example: IDA in longitudinal studies

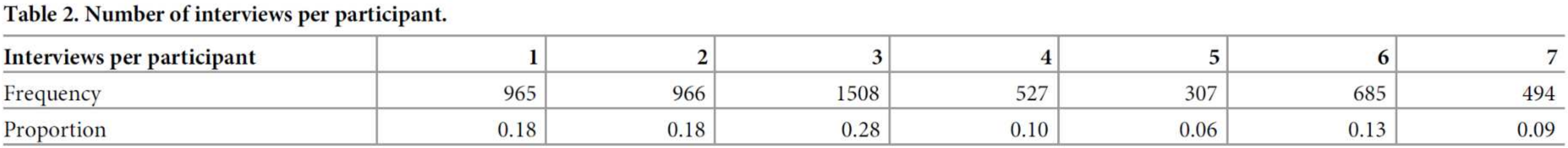


Table 2. Number of interviews per participant.

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

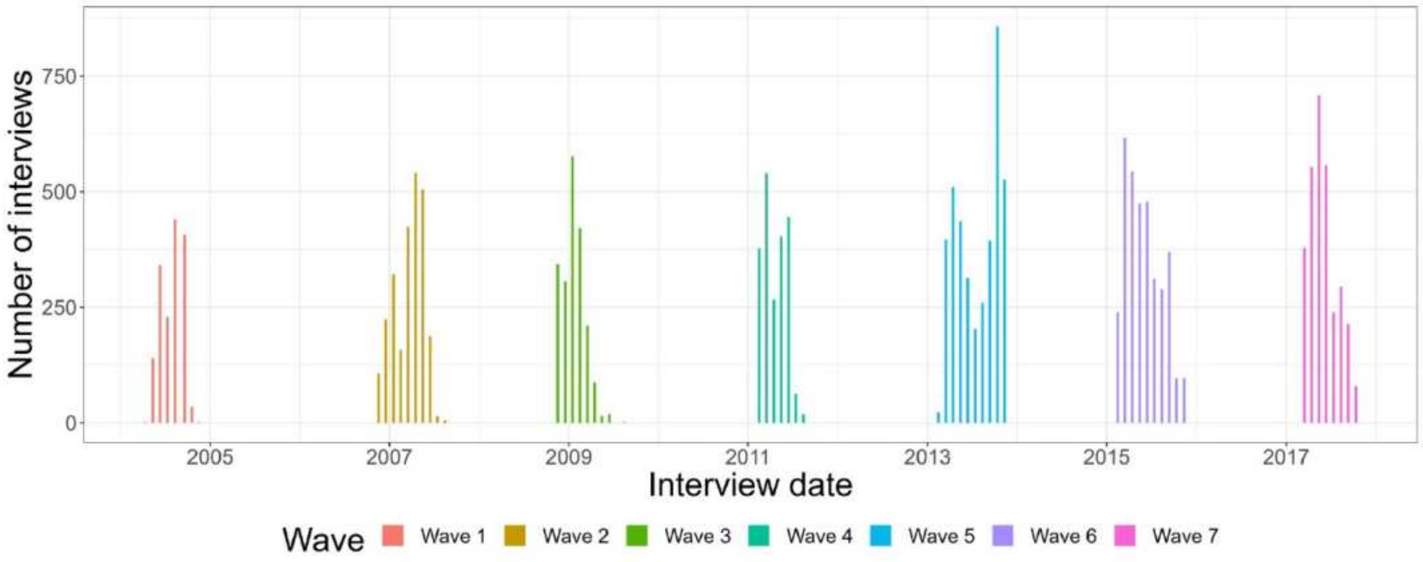
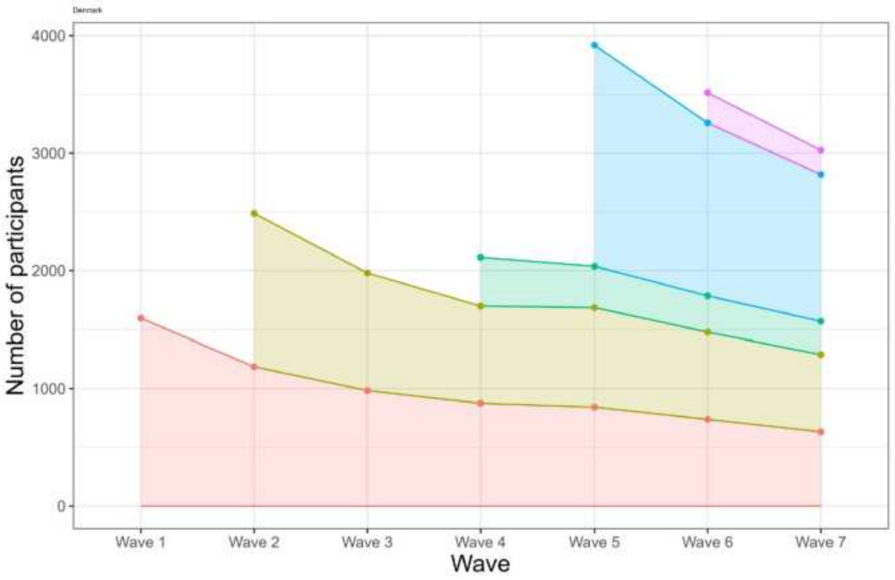


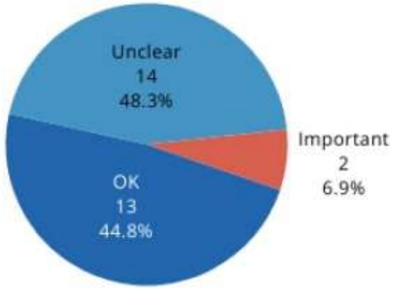
Fig 1. Distribution of the number of interviews carried out in Denmark in the SHARE study in time.



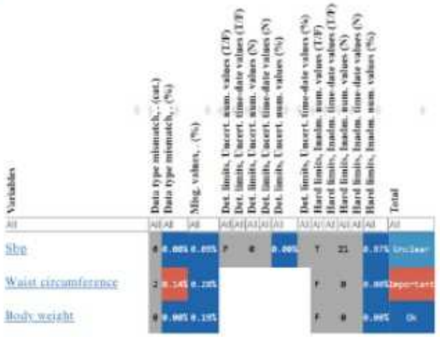
Scan for R package dataquier

Data quality assessment – Report

1. Data quality overview



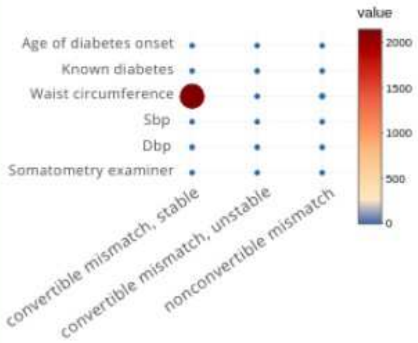
Percentage of variables per quality categories



Variables	Data type mismatch, : (cat.)	Data type mismatch, : (%)	Mlsg. values, : (%)	Det. limits, Uncert. num. values (T/F)	Det. limits, Uncert. time-date values (T/F)	Det. limits, Uncert. num. values (N)	Det. limits, Uncert. time-date values (N)	Det. limits, Uncert. num. values (%)	Det. limits, Uncert. time-date values (%)	Hard limits, Inadm. num. values (T/F)	Hard limits, Inadm. time-date values (T/F)	Hard limits, Inadm. num. values (N)	Hard limits, Inadm. time-date values (N)	Hard limits, Inadm. num. values (%)	Total
All	All	All	All	All	All	All	All	All	All	All	All	All	All	All	All
Sbp	0	0.00%	0.05%	F		0		0.00%		T		21		0.37%	Unclear
Waist circumference	2	0.14%	0.28%							F		0		0.00%	Important
Body weight	0	0.00%	0.15%							F		0		0.00%	Ok

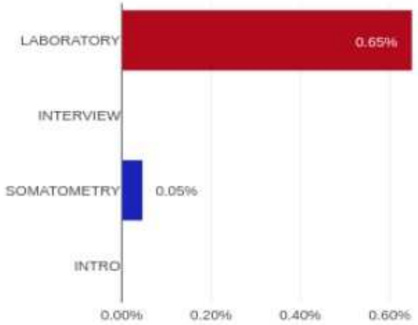
Targeted quality indicators, potential issues, and applicability problems

2. Integrity



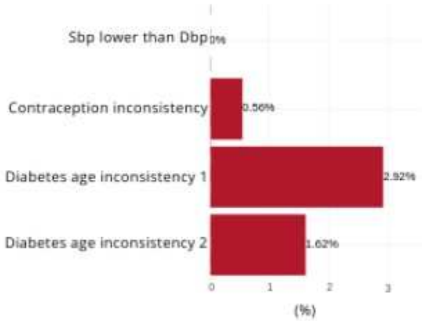
Data type mismatch

3. Completeness



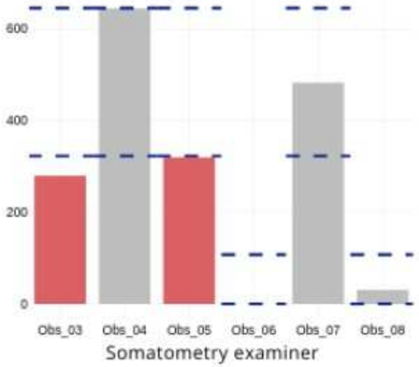
Missing values per segment

4. Consistency



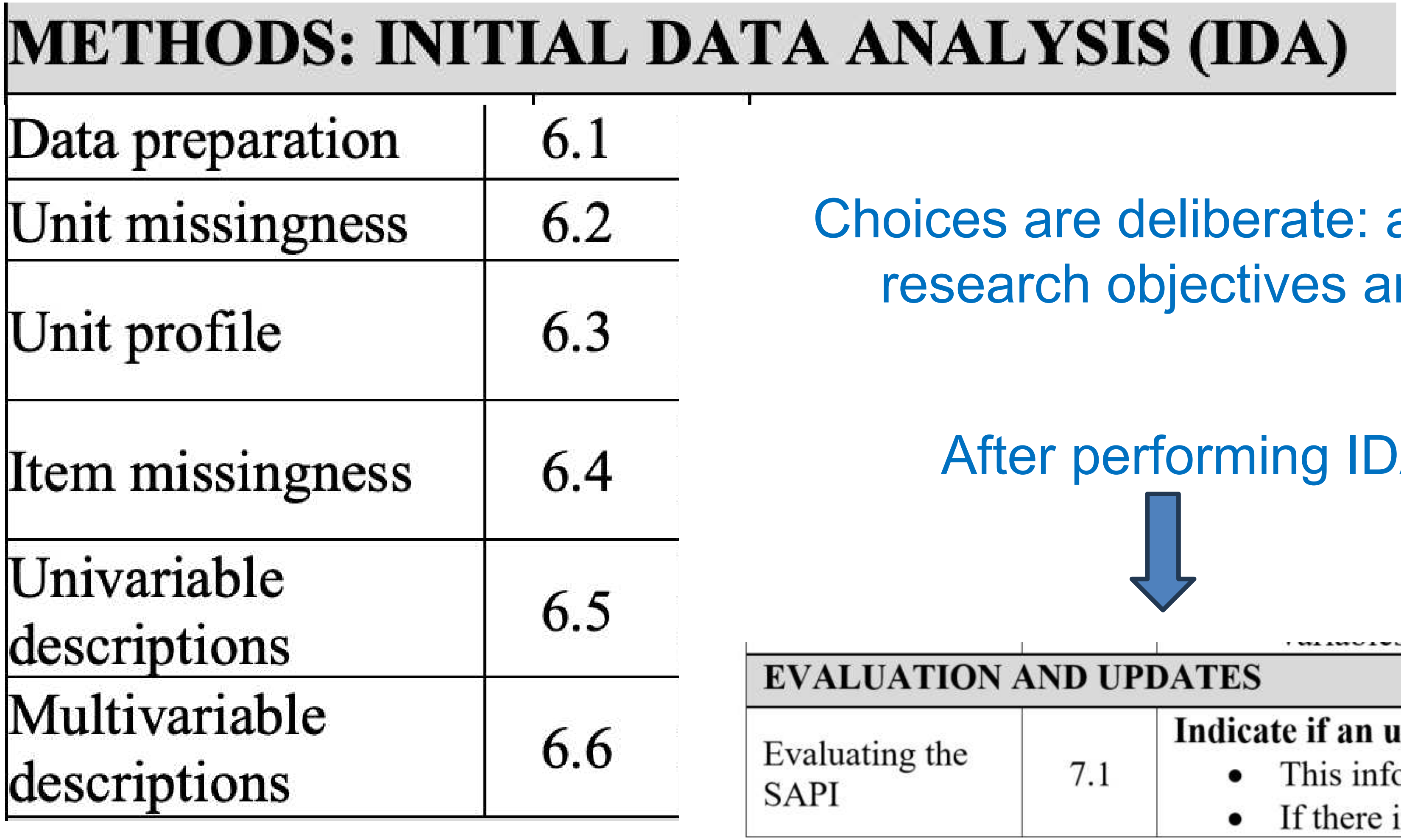
Contradictions

5. Accuracy



Unexpected proportions

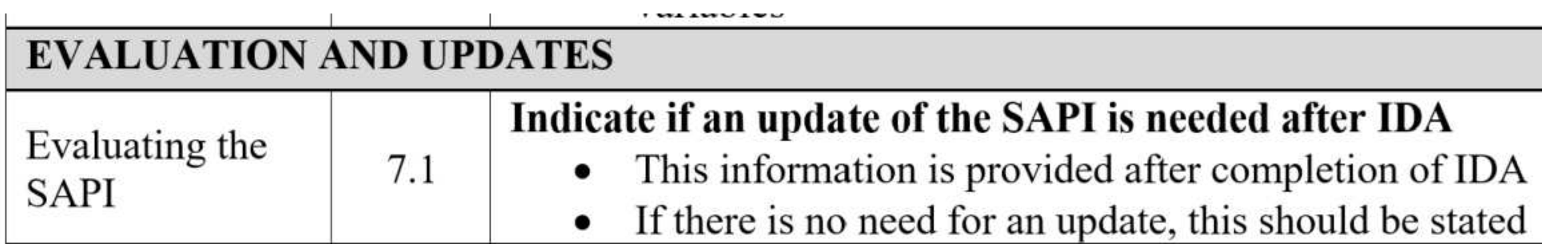
SAPI: Statistical Analysis Plan with **IDA**



METHODS: INITIAL DATA ANALYSIS (IDA)	
Data preparation	6.1
Unit missingness	6.2
Unit profile	6.3
Item missingness	6.4
Univariable descriptions	6.5
Multivariable descriptions	6.6

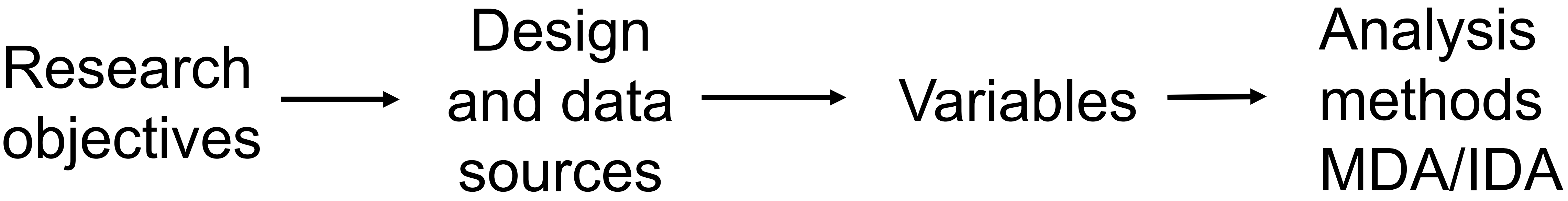
Choices are deliberate: aligned with research objectives and MDA.

After performing IDA



EVALUATION AND UPDATES		
Evaluating the SAPI	7.1	**Indicate if an update of the SAPI is needed after IDA** • This information is provided after completion of IDA • If there is no need for an update, this should be stated

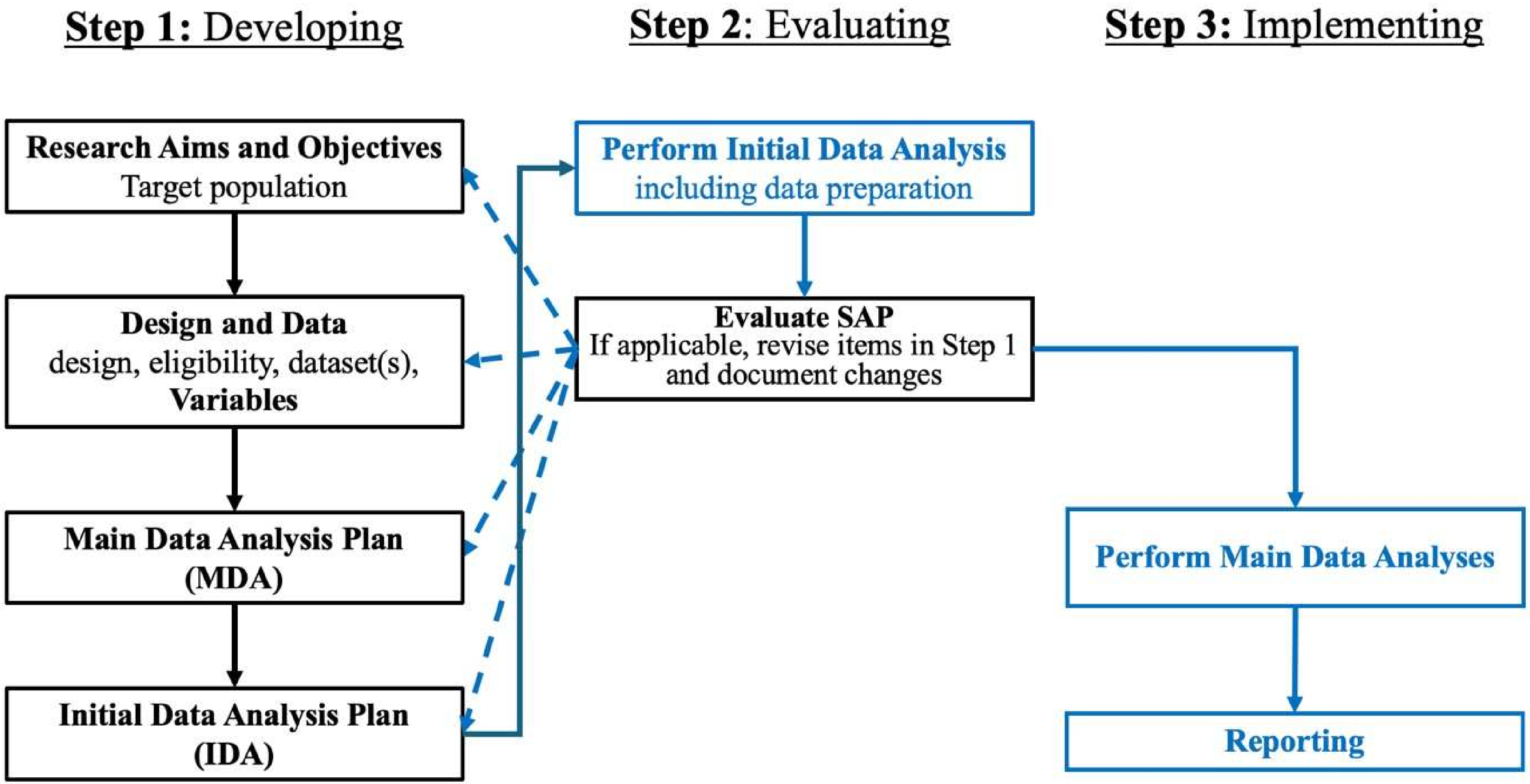
Roadmap for Statistical Analysis Plan for Observational Studies (SAPI)



Research objectives → Design and data sources → Variables → Analysis methods MDA/IDA

SAP guideline developed via a (international) consensus process with researchers, analysts, editors/reviewers, instructors/mentors.

Iterative process of developing an analysis plan



Step 1: Developing

Research Aims and Objectives
Target population

Design and Data
design, eligibility, dataset(s), **Variables**

Main Data Analysis Plan (MDA)

Initial Data Analysis Plan (IDA)

Step 2: Evaluating

Perform Initial Data Analysis
including data preparation

Evaluate SAP
If applicable, revise items in Step 1 and document changes

Step 3: Implementing

Perform Main Data Analyses

Reporting

Scan for "hidden analysis" and reporting

Organizing knowledge about data: Metadata

Variable names

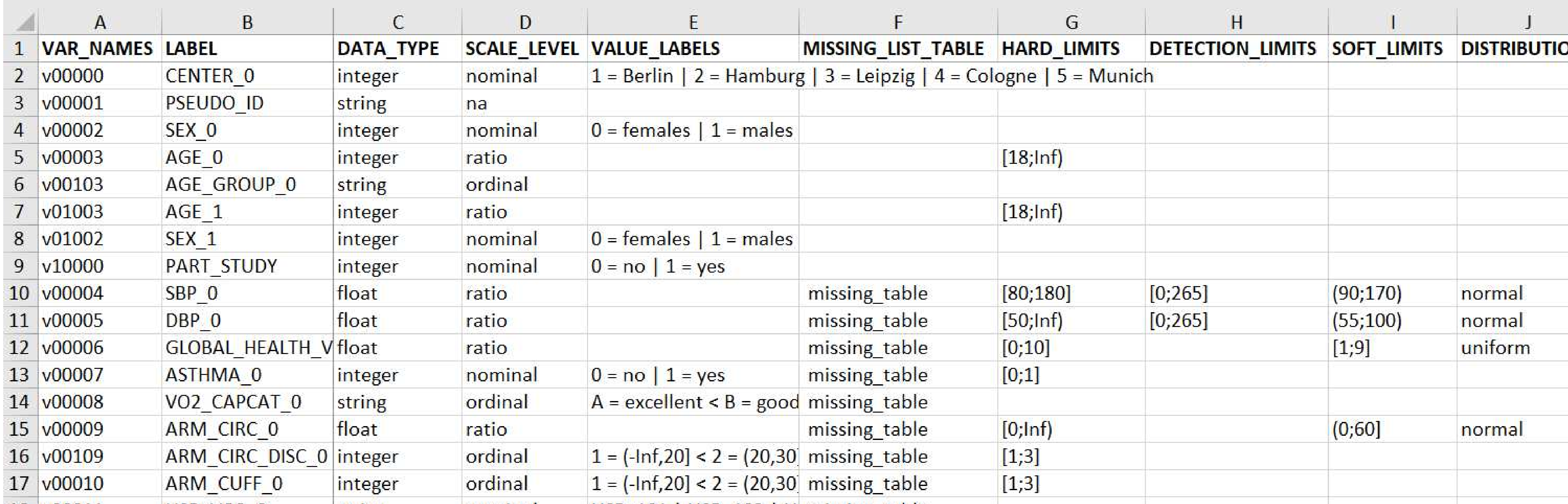
Label

Type (integer, string, date,..)

Values (categories)

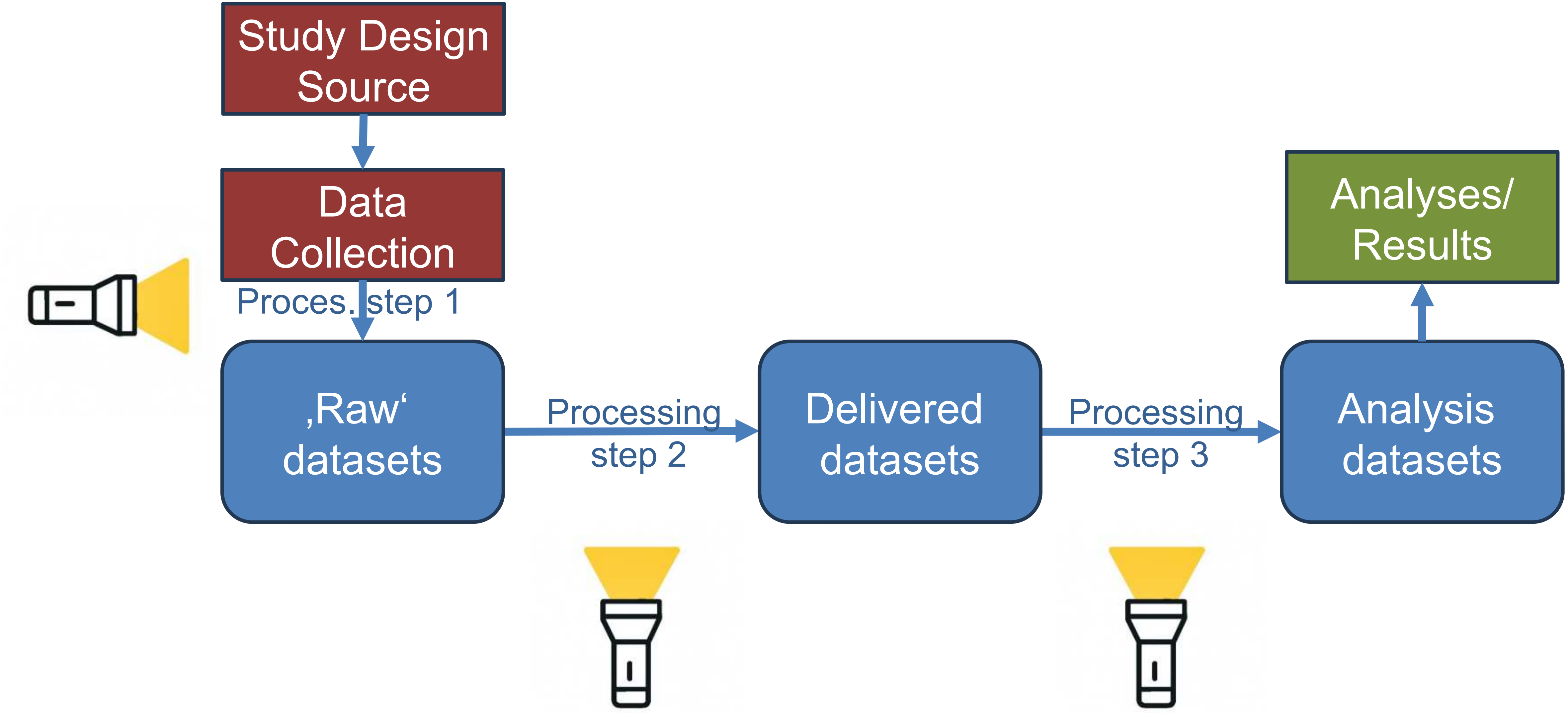
Range (continuous)

Expectations (distribution, missingness,…)



	A	B	C	D	E	F	G	H	I	J
1	VAR_NAMES	LABEL	DATA_TYPE	SCALE_LEVEL	VALUE_LABELS	MISSING_LIST_TABLE	HARD_LIMITS	DETECTION_LIMITS	SOFT_LIMITS	DISTRIBUTIO
2	v00000	CENTER_0	integer	nominal	1 = Berlin | 2 = Hamburg | 3 = Leipzig | 4 = Cologne | 5 = Munich					
3	v00001	PSEUDO_ID	string	na						
4	v00002	SEX_0	integer	nominal	0 = females | 1 = males					
5	v00003	AGE_0	integer	ratio			[18;Inf)			
6	v00103	AGE_GROUP_0	string	ordinal						
7	v01003	AGE_1	integer	ratio			[18;Inf)			
8	v01002	SEX_1	integer	nominal	0 = females | 1 = males					
9	v10000	PART_STUDY	integer	nominal	0 = no | 1 = yes					
10	v00004	SBP_0	float	ratio		missing_table	[80;180]	[0;265]	(90;170)	normal
11	v00005	DBP_0	float	ratio		missing_table	[50;Inf)	[0;265]	(55;100)	normal
12	v00006	GLOBAL_HEALTH_V	float	ratio		missing_table	[0;10]		[1;9]	uniform
13	v00007	ASTHMA_0	integer	nominal	0 = no | 1 = yes	missing_table	[0;1]			
14	v00008	VO2_CAPCAT_0	string	ordinal	A = excellent < B = good	missing_table				
15	v00009	ARM_CIRC_0	float	ratio		missing_table	[0;Inf)		(0;60]	normal
16	v00109	ARM_CIRC_DISC_0	integer	ordinal	1 = (-Inf,20] < 2 = (20,30]	missing_table	[1;3]			
17	v00010	ARM_CUFF_0	integer	ordinal	1 = (-Inf,20] < 2 = (20,30]	missing_table	[1;3]			

One challenge - what granularity is needed?



IDA….

1. … is the foundation for correct statistical analyses
2. … should be included in statistical analysis plans to
   - better structure comprehensive IDA in the full analysis workflow
   - better estimate needed resources
   - better oversee the needed meta-information

STRATOS INITIATIVE TG3