



Initial Data Analysis in Practice: Between Pragmatism, Pitfalls, and Principles

From STRATOS TG3 to SAPI

Carsten O. Schmidt
on behalf of TG3

Pitfalls.....

- Treating the delivered dataset as already “analysis-ready”
- Starting modelling before the data structure and provenance are understood
- Making data-driven changes without documenting when and why they were made
- Reporting only the final clean dataset and concealing the decisions that produced it
- Relying on manual spreadsheet edits that cannot be reproduced
- Using different dataset versions for different analyses
- Allowing outcome information to influence cleaning, transformations, or coding decisions
- Changing eligibility criteria after observing the resulting sample
- Excluding observations because they appear inconvenient or influential
- Removing “outliers” without distinguishing data errors from valid extreme values
- Correcting implausible values without retaining the original values and an audit trail
- Applying automatic cleaning rules without considering variable meaning or measurement context
- Describing missingness only variable by variable and overlooking multivariable patterns
-

Research



Cite this article: Stefan AM, Schönbrodt FD. 2023 Big little lies: a compendium and simulation of *p*-hacking strategies. *R. Soc. Op. Sci.* **10**: 220346. <https://doi.org/10.1098/rsos.220346>

Received: 17 March 2022
Accepted: 11 January 2023

Subject Category:
Psychology and cognitive neuroscience

Subject Areas:
psychology/statistics/computer modelling and simulation

Keywords:
questionable research practices, simulation, error rates, significance, false-positive rate, Shiny app, *p*-curve

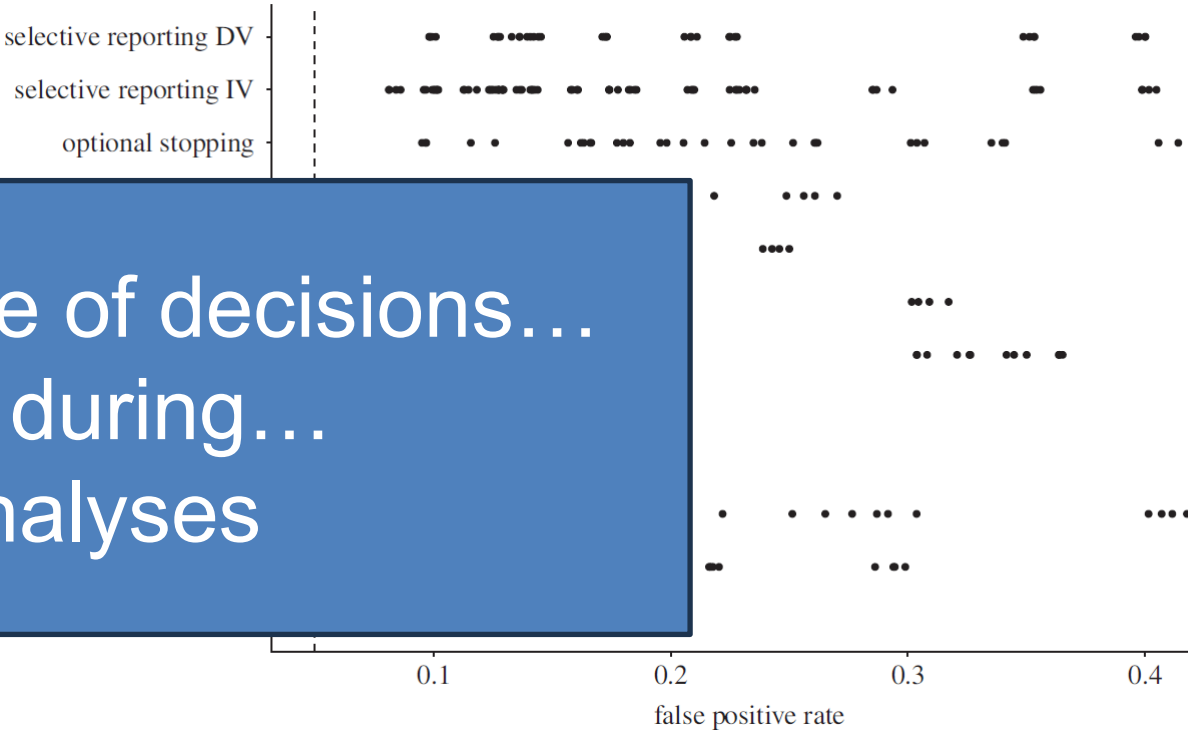
Big little lies: a compendium and simulation of *p*-hacking strategies

Angelika M. Stefan^{1,2} and Felix D. Schönbrodt³

¹Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands
²Department of Psychology, Universität der Bundeswehr München, München, Germany

..we have a multiverse of decisions...
...before AND during...
Statistical Analyses

control their level of severity, and demonstrate their impact on false-positive rates using simulation studies. We also use our simulation results to evaluate several approaches that have been proposed to mitigate the influence of questionable research practices. Our results show that investigating *p*-hacking at the level of strategies can provide a better understanding of the process of *p*-hacking, as well as a broader basis for developing effective countermeasures. By making our analyses available through a Shiny app and R package, we facilitate future meta-scientific research aimed at investigating the ramifications of *p*-hacking across multiple strategies, and we hope to start a broader discussion about different manifestations of *p*-hacking in practice.





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
- Is NO exploratory data analysis

The foundation matters.....



STRATOS TG3: work programme...creating principles...

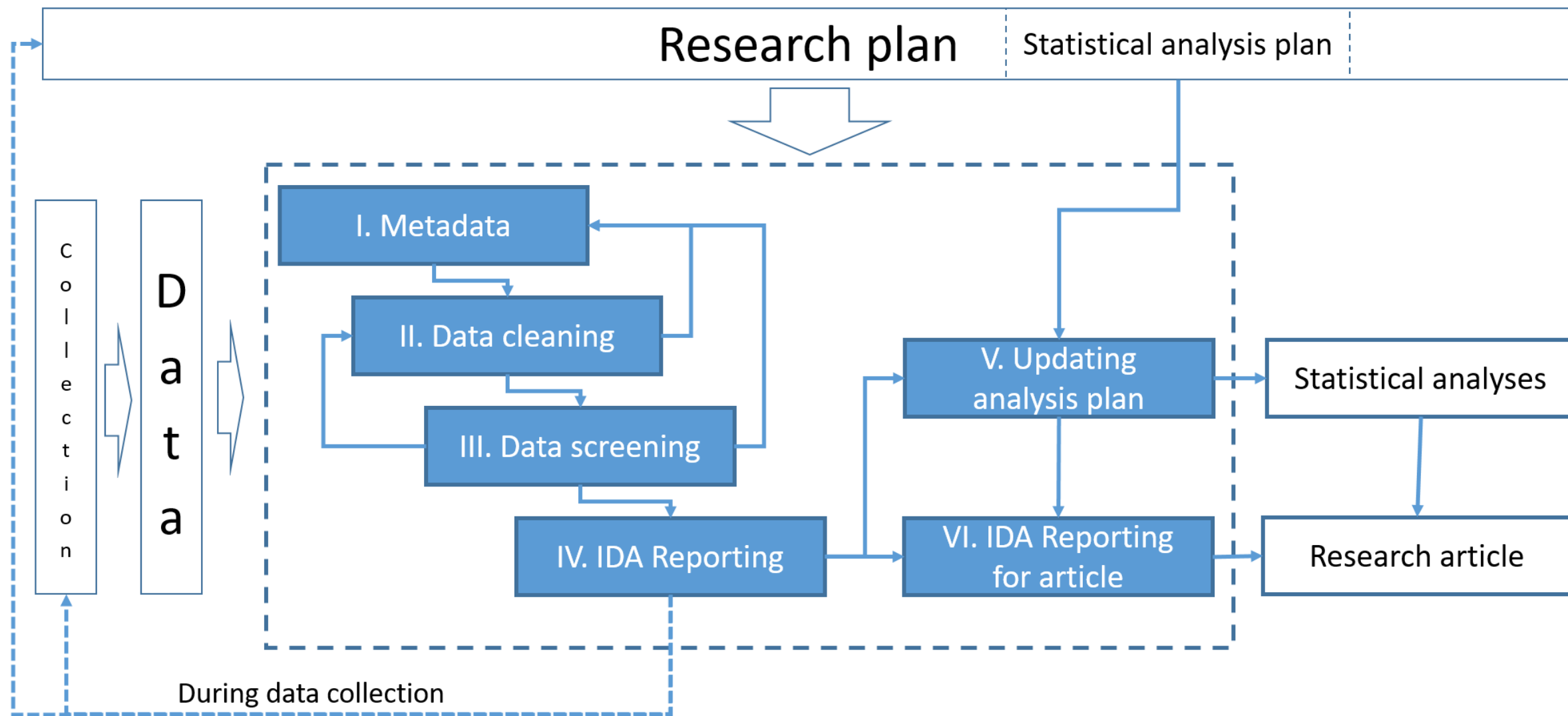
1.

IDA framework

Concepts;
Workflows;
Definitions



Initial Data Analysis Workflow



STRATOS TG3: work programme develop pragmatic solutions

1.

IDA framework

Concepts;
Workflows;
Definitions

2.

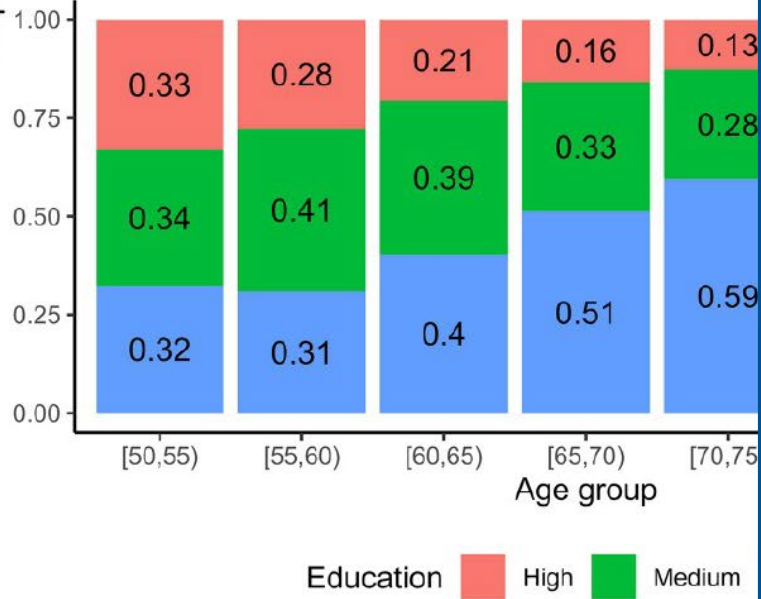
Applications

Regression without
regrets;
longitudinal IDA;
worked examples



Report – Comprehensively summarize findings

General population



Lusa et al. 2024

- 1 Preface
- 2 Initial data analysis and data screening checklist for longitudinal data
- 3 SHARE data description
- 4 Data example (from research question to IDA plan)
- 5 Data screening
 - 5.1 Participation profile
 - 5.2 Missing values
 - 5.2.1 Non-enrollment (M1)
 - 5.2.2 Drop-out (M2) and intermittent missingness (M3)
 - 5.2.3 Variable missingness (item missingness, M4)
 - 5.2.4 Patterns (M5)
 - 5.2.5 Comparison of non-enrolled and target population (ME1)
 - 5.2.6 Probability of loss to follow-up and death (ME2)
 - 5.2.7 Dropout effect on outcome (ME3)
 - 5.3 Univariate descriptions
 - 5.4 Multivariate description of data
 - 5.5 Longitudinal aspects

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	165.00 172.00 171.82 ± 9.04	163.00 169.00 169.34 ± 8.80	165.00 172.00 171.66 ± 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 / ₉₆₉	0.41 ¹⁹⁵ / ₄₇₂
High		0.45 ¹²¹² / ₂₆₇₈	0.23 ²²³ / ₉₆₉	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.

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 Vach³, Saskia le Cessie⁴, Carsten Oliver Schmidt⁵, 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>)



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

Georg Heinze (✉ georg.heinze@meduniwien.ac.at)
Medical University of Vienna

Mark Baillie
Novartis (Switzerland)

Lara Lusa
University of Primorska

Willi Sauerbrei
University of Freiburg

Carsten Oliver Schmidt

PLOS ONE

RESEARCH ARTICLE

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

Lara Lusa^{1,2*}, Cécile Proust-Lima³, Carsten O. Schmidt⁴, Katherine J. Lee^{5,6}, Saskia le Cessie^{7,8}, Mark Baillie⁹, Frank Lawrence¹⁰, Marianne Huebner^{10,11}, on behalf of TG3 of the STRATOS Initiative¹¹

¹ Department of Mathematics, Faculty of Mathematics, Natural Sciences and Informatics Technology

Observational Studies 4 (2018) 171-192

Submitted 7/17; Published 4/18

A Contemporary Conceptual Framework for Initial Data Analysis

Marianne Huebner
Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824, USA

huebner@stt.msu.edu

Saskia le Cessie
Department of Clinical Epidemiology and Department of Medical Statistics and Bioinformatics
Leiden University Medical Center
Leiden, The Netherlands

S.le.Cessie@lumc.nl



vald.de

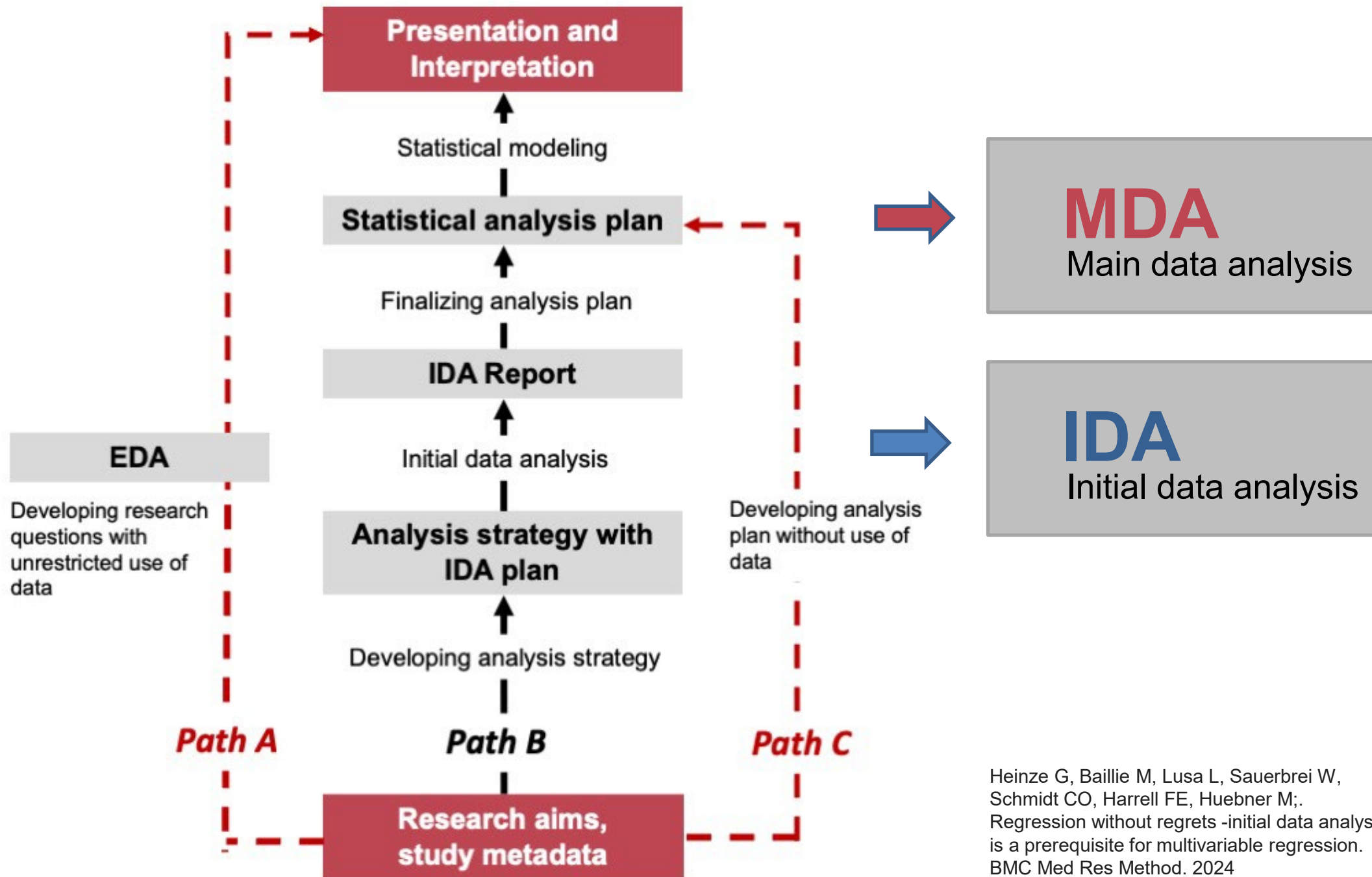
usb.ch

EDITORIAL

Ten simple rules for initial data analysis

Mark Baillie¹, Saskia le Cessie², Carsten Oliver Schmidt³, Lara Lusa⁴, Marianne Huebner^{5*}, for the Topic Group "Initial Data Analysis" of the STRATOS Initiative¹¹

¹ Novartis, Basel, Switzerland, ² Department of Clinical Epidemiology and Department of Biomedical



STRATOS TG3: work programme ... integrate principles beyond IDA

1.

IDA framework

Concepts;
Workflows;
Definitions

2.

Applications

Regression without
regrets;
longitudinal IDA;
worked examples

4.

SAPI

Integrating
IDA into
SAP for
observational
studies

Overview: Statistical Analysis Plan with Initial Data Analysis

Objective:

To develop a checklist for an extended statistical analysis plan for observational studies that integrates demands for MDA and IDA and records deviations from initial plans in one document (SAPI).

Overview: Statistical Analysis Plan with Initial Data Analysis

Scope

- Observational studies
- Descriptive, predictive and causal objectives
- Existing or newly collected data

Development

- Literature reviews and IDA framework
- Delphi exercise across STRATOS
- Consensus meetings
- Piloted in 19 research projects

Core innovation

- 27-item checklist + expanded checklist
- Links protocol, SAP, IDA and reporting



SAPI checklist for statistical analysis plans for observational studies

ADMINISTRATIVE INFORMATION

- 1.1 Project title
- 1.2 Project documents
- 1.3 Ethics approval
- 1.4 Names and contact

PROJECT BACKGROUND

- 2.1 Research aims and objectives
- 2.2 Target population

DATA

- 3.1 Data sources
- 3.2 Design
- 3.3 Eligibility Criteria
- 3.3 Data sets

VARIABLES

- 4.1 Variables used in MDA
- 4.2 Other variables

MAIN DATA ANALYSIS (MDA)

- 5.1 Description of observation units
- 5.2 Main data analysis methods
- 5.3 Assumptions and diagnostics
- 5.4 Sample size
- 5.5 Software

INITIAL DATA ANALYSIS (IDA)

- 6.1 Data preprocessing
- 6.2 Unit missingness
- 6.3 Unit profile
- 6.4 Item missingness
- 6.5 Univariable descriptions
- 6.6 Multivariable descriptions

EVALUATION AND UPDATES

- 7.1 Evaluating the SAPI
- 7.2 Updating the SAPI

SUPPLEMENT

- 8.1 Key references
- 8.2 Sustainable handling of outputs



SAPI checklist for statistical analysis plans for observational studies

ADMINISTRATIVE INFORMATION

- 1.1 Project title
- 1.2 Project documents
- 1.3 Ethics approval
- 1.4 Names and contact

PROJECT BACKGROUND

- 2.1 Research aims and objectives
- 2.2 Target population

DATA

- 3.1 Data sources
- 3.2 Design
- 3.3 Eligibility Criteria
- 3.3 Data sets

VARIABLES

- 4.1 Variables used in MDA
- 4.2 Other variables

DESIGN AND DATA

Data sources	3.1	Describe the sources of data (e.g., prospective cohort study, routinely collected electronic health records, registries, surveys)
Design	3.2	State the study design for this analysis project (e.g., cross-sectional, cohort, case-control, case-crossover, self-controlled case series)
Eligibility criteria	3.3	Specify eligibility criteria for the research project based on information available at baseline
Data sets	3.4	Describe how the data are provided for this analysis project (e.g., the format and content of datasets)



SAPI checklist for statistical analysis plans for observational studies

MAIN DATA ANALYSIS (MDA)

- 5.1 Description of observation units
- 5.2 Main data analysis methods
- 5.3 Assumptions and diagnostics
- 5.4 Sample size
- 5.5 Software

INITIAL DATA ANALYSIS (IDA)

- 6.1 Data preprocessing
- 6.2 Unit missingness
- 6.3 Unit profile
- 6.4 Item missingness
- 6.5 Univariable descriptions
- 6.6 Multivariable descriptions

METHODS: INITIAL DATA ANALYSIS (IDA)

Data preprocessing	6.1	Describe any methods for preparing data for the analyses
Unit missingness	6.2	Describe any methods to identify the extent of unit missingness
Unit profile	6.3	Describe any methods to summarize the temporal or structural pattern of observations for each observation unit
Item missingness	6.4	Describe methods to examine item missingness (i.e., missing values in variables)
Univariable descriptions	6.5	Describe the methods to summarize the distribution of each variable used in the analyses with numerical or graphical summaries
Multivariable descriptions	6.6	Describe any multivariable descriptive statistics and graphical summaries

SAPI checklist for statistical analysis plans for observational studies

EVALUATION AND UPDATES

7.1 Evaluating the SAPI

7.2 Updating the SAPI

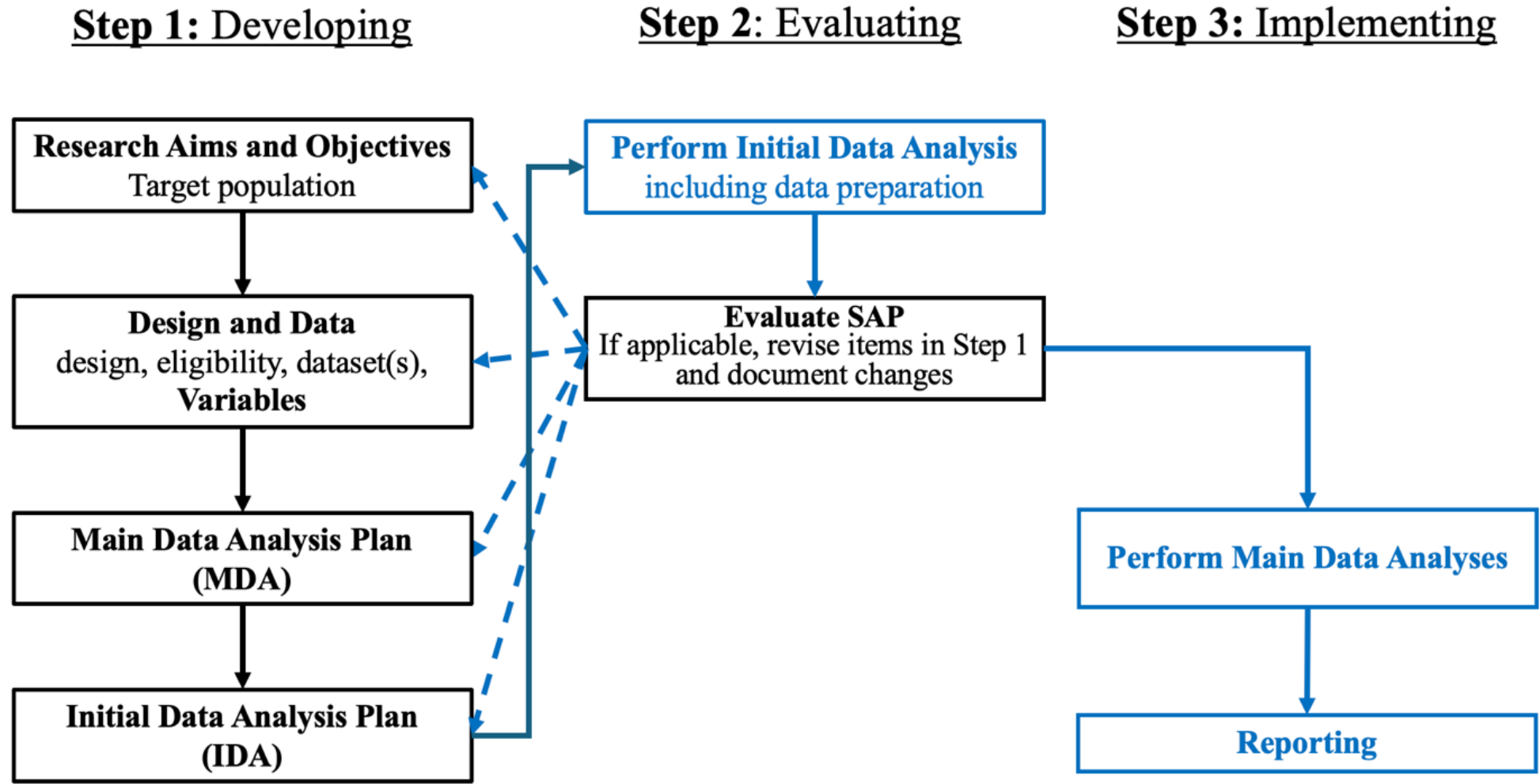
SUPPLEMENT

8.1 Key references

8.2 Sustainable handling of outputs

EVALUATION AND UPDATES		
Evaluating the SAPI	7.1	Indicate if an update of the SAPI is needed after IDA
Updating the SAPI, if applicable	7.2	List all SAPI sections affected by updates and state the reasons for the updates

Iterative process of developing an analysis plan



Scan for “hidden analysis”
and reporting



Conclusion

- TG3
develops concepts and guidance related to determine the suitability of the data for the intended main data analysis.
Aims to make the workflow preceeding the main MDA coherent and transparent
- SAPI is the bridge between statistical analysis planning and reproducible IDA.
 - It structures data understanding before MDA
 - It makes pre-MDA amendments auditable
 - It separates legitimate data preparation from data-driven modelling
- Further work will target the implementation and use of our concepts