Initial data analysis plans are part of research projects

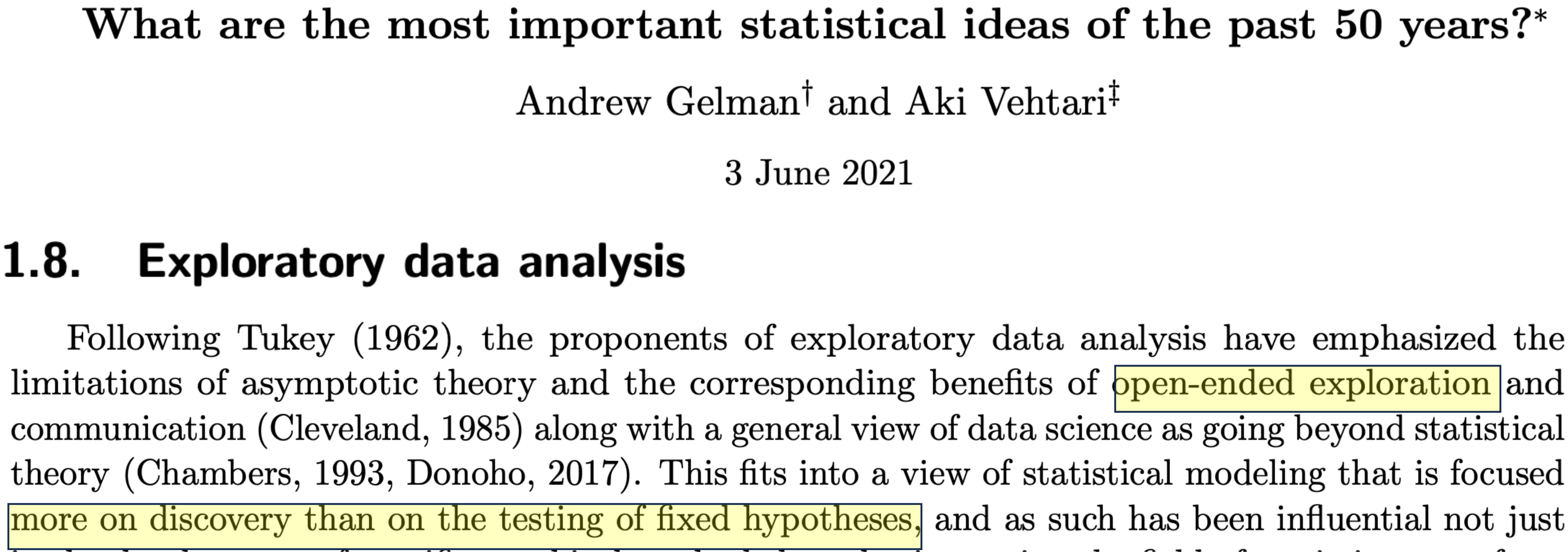
Marianne Huebner (Michigan State University, USA),

Carsten Oliver Schmidt (University Medicine Greifswald, Germany)

Lara Lusa (University of Primorska, Slovenia)

Website: https://www.stratosida.org

STRATOS INITIATIVE



What are the most important statistical ideas of the past 50 years?[*]

Andrew Gelman[†] and Aki Vehtari[‡]

3 June 2021

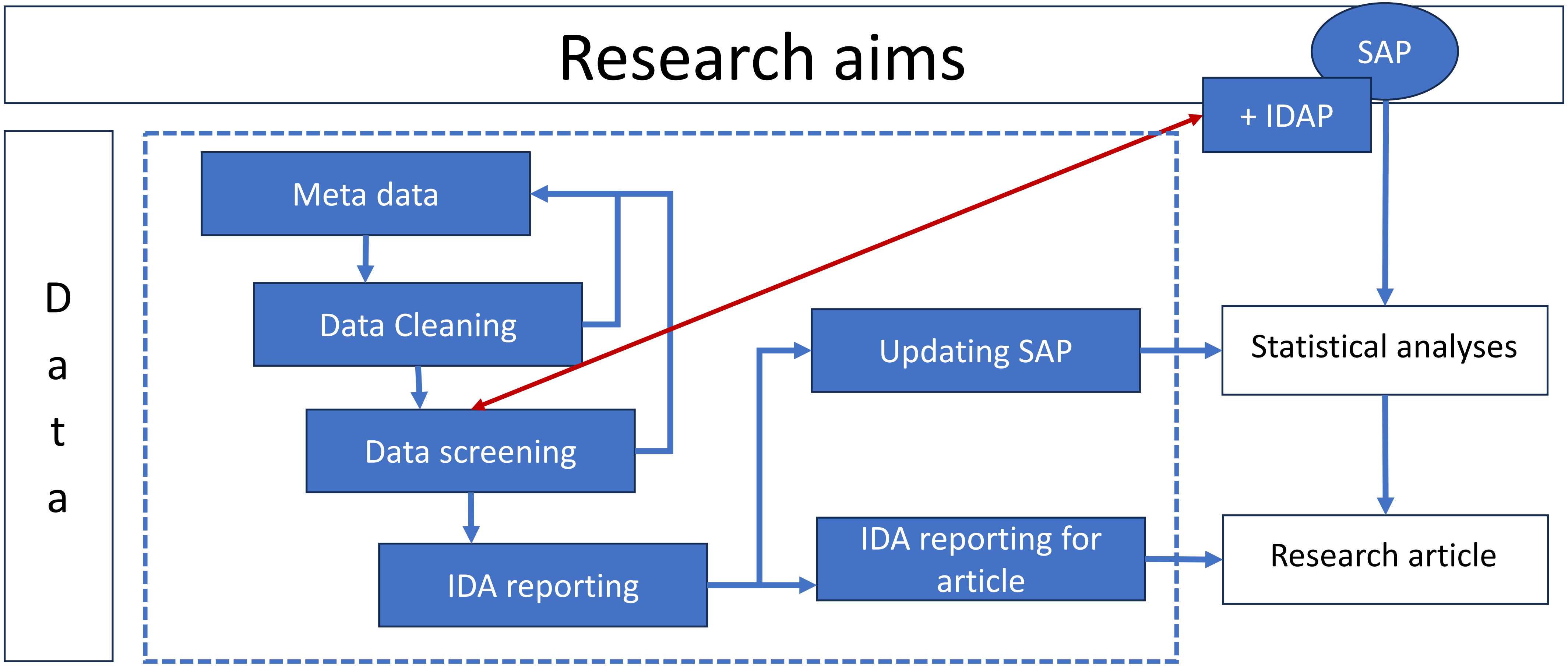
1.8. Exploratory data analysis

Following Tukey (1962), the proponents of exploratory data analysis have emphasized the limitations of asymptotic theory and the corresponding benefits of open-ended exploration and communication (Cleveland, 1985) along with a general view of data science as going beyond statistical theory (Chambers, 1993, Donoho, 2017). This fits into a view of statistical modeling that is focused more on discovery than on the testing of fixed hypotheses, and as such has been influential not just

Initial data analysis ≠ Exploratory data analysis

- Share toolbox (e.g. data visualization)
- IDA orients itself around research aim and statistical analysis plan

Initial Data Analysis Steps



IDA plans are limited to generic statements in statistical analysis plans/manuscripts

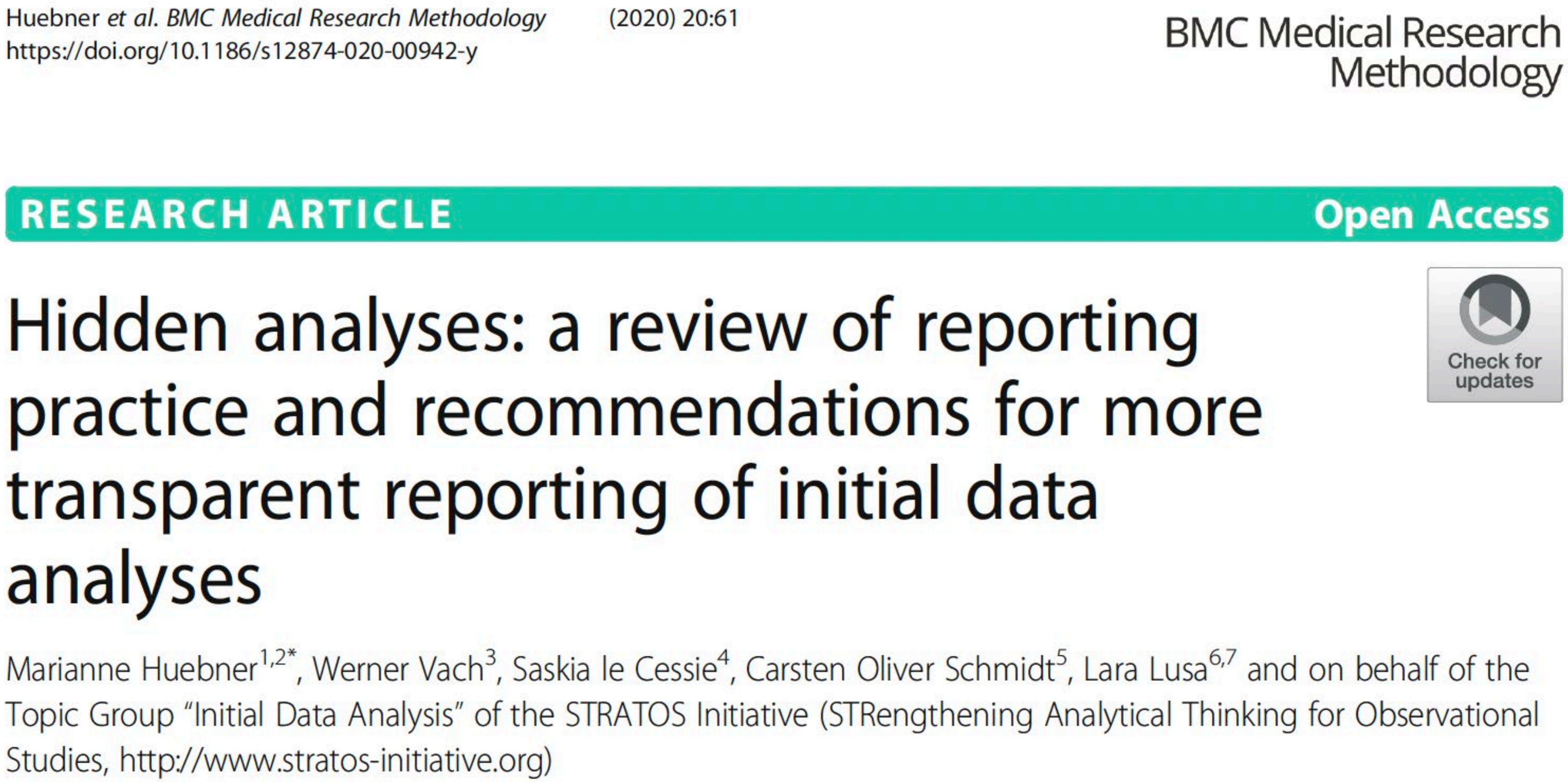
“Means (SDs) or counts and percentages were computed for all continuous or categorical demographic variables. ”

JAMA, August 2023

“Baseline characteristics of the analytical sample were summarised across three lifestyle groups as a percentage for categorical variables and mean and standard deviation for continuous variables. Missing values are summarised in supplementary table 1.”

BMJ, August 2023

clinicaltrials.gov - similar



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

BMC Medical Research Methodology

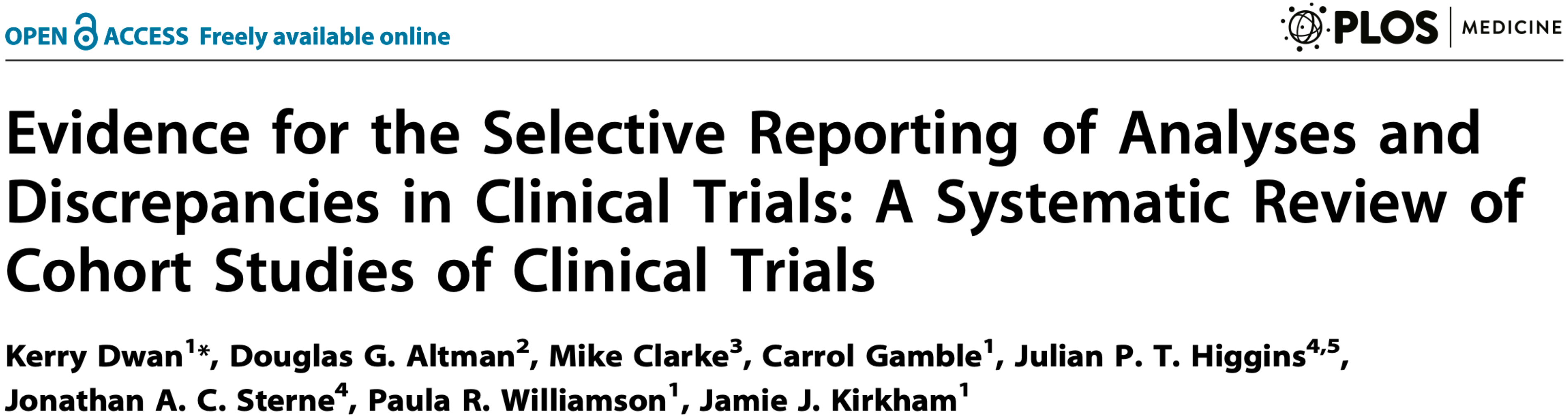
RESEARCH ARTICLE Open Access

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)

IDA statements were reported in the methods, results, discussion, and supplement of papers.

- 40% (of 25 papers) included a statement about data cleaning.
- 44% included statements on item missingness, 60% on unit missingness
- 44% mentioned some changes in the analysis plan



OPEN ACCESS Freely available online

PLOS | MEDICINE

Evidence for the Selective Reporting of Analyses and Discrepancies in Clinical Trials: A Systematic Review of Cohort Studies of Clinical Trials

Kerry Dwan[1]*, Douglas G. Altman[2], Mike Clarke[3], Carrol Gamble[1], Julian P. T. Higgins[4,5], Jonathan A. C. Sterne[4], Paula R. Williamson[1], Jamie J. Kirkham[1]

Discrepancy between protocols and publications:

- statistical analyses
- handling of missing data
- handling of continuous data
- subgroup analyses

Reasons often not discussed

“post hoc decisions about which subgroups to analyze and report may be influenced by the findings of those or related analyses”

Analysis Plans

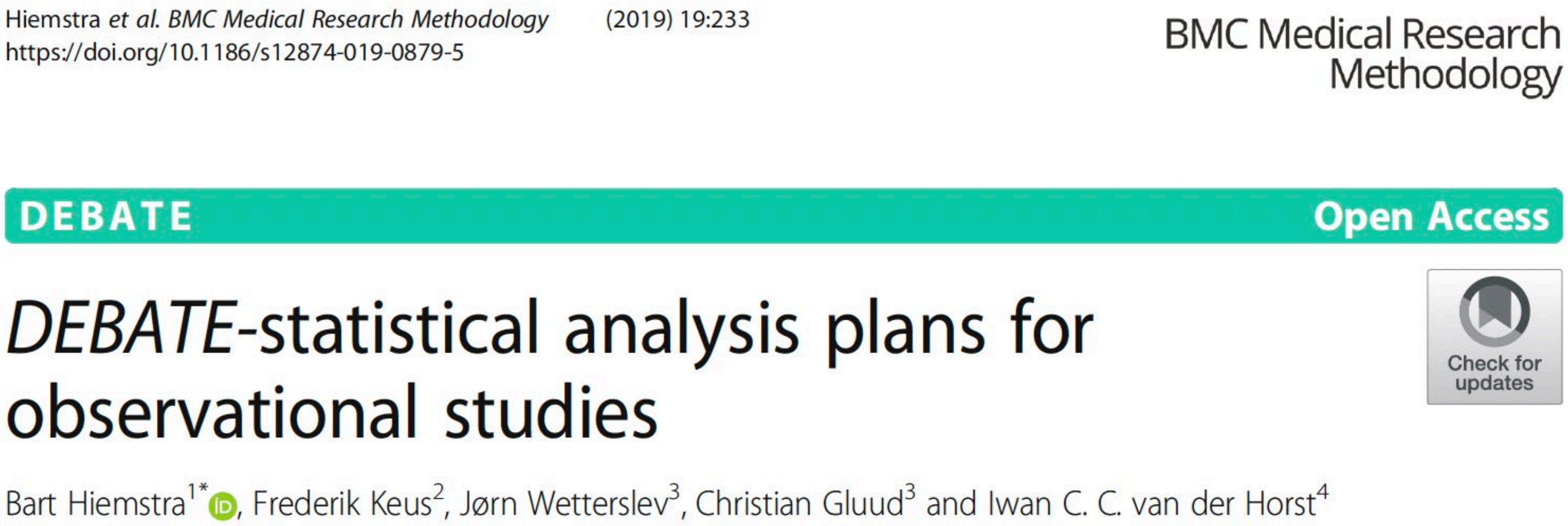
Statistical analysis plan (SAP) = "a document that […] includes detailed procedures for executing the statistical analysis of the primary and secondary variables and other data" (ICH E9)

"The SAP is to be applied to a clean or validated data set for analysis" -

Guidance for SAP is available

Initial data analysis plan (IDAP) – Is this included in the SAP?

What is needed:
Guidance for IDAP as part of the SAP



Hiemstra *et al. BMC Medical Research Methodology* (2019) 19:233
https://doi.org/10.1186/s12874-019-0879-5

BMC Medical Research Methodology

DEBATE Open Access

DEBATE-statistical analysis plans for observational studies

Bart Hiemstra[1*], Frederik Keus[2], Jørn Wetterslev[3], Christian Gluud[3] and Iwan C. C. van der Horst[4]

IDAP in DEBATE? Partially, with generic statements

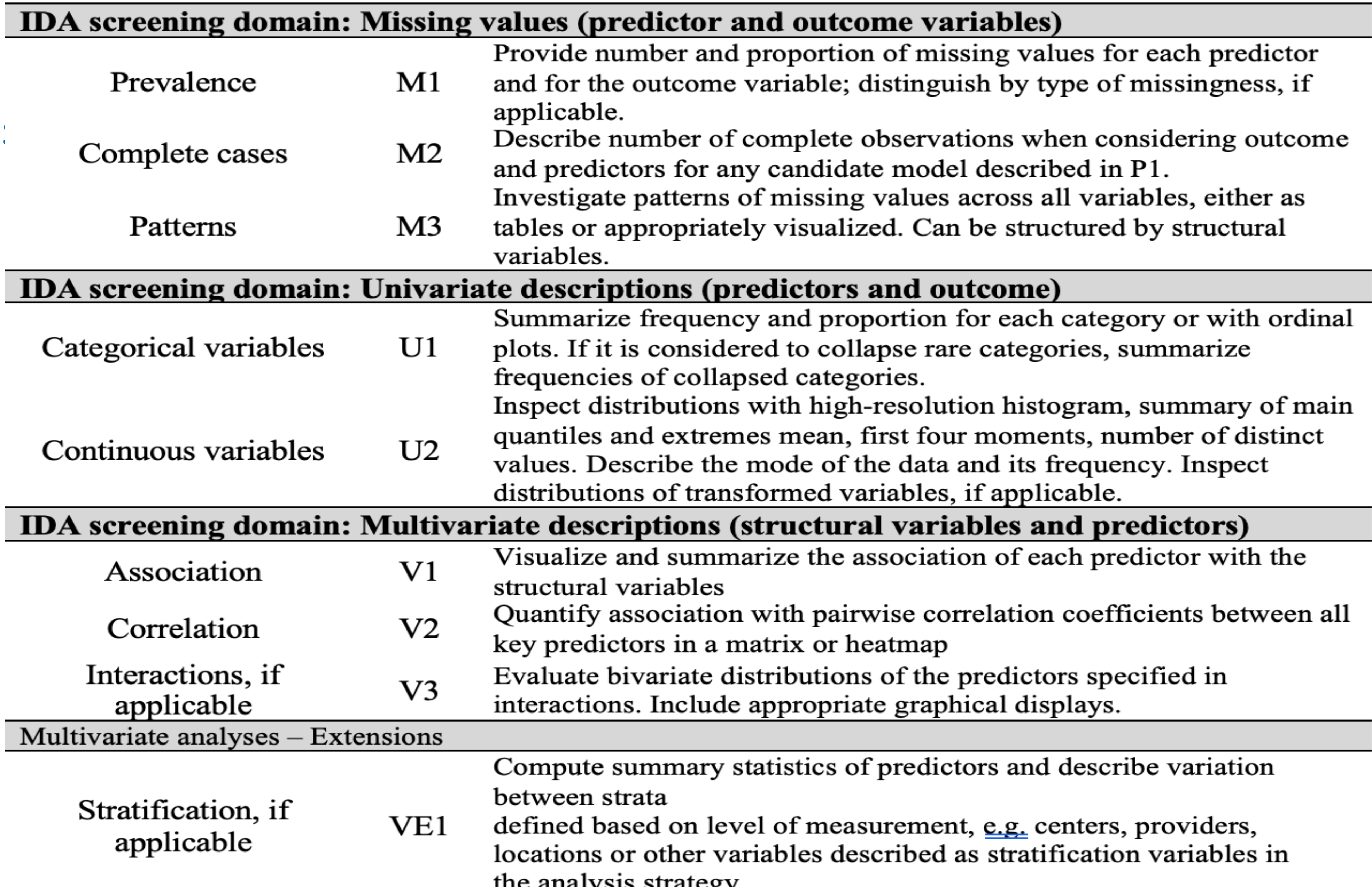
1. Missing data (reporting, assumptions, how to handle)
2. Baseline characteristics (methods to summarize)
3. Time points at which the outcomes are measured
4. Loss to follow-up (timing, reasons, presentation)

IDA Plan (data screening) for cross-sectional studies

Missing values

Univariate descriptions

Multivariate descriptions



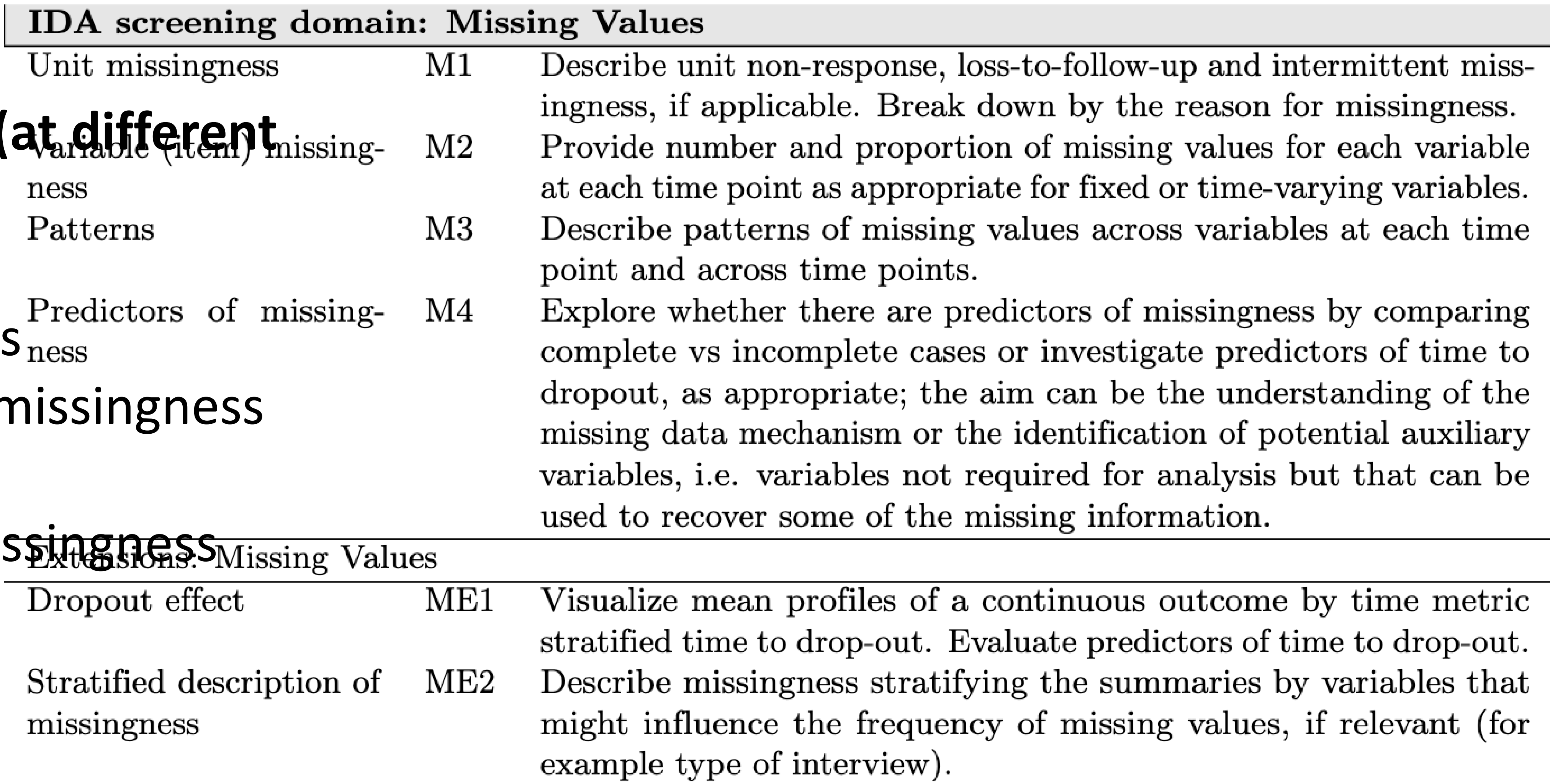
IDA screening domain: Missing values (predictor and outcome variables)		
Prevalence	M1	Provide number and proportion of missing values for each predictor and for the outcome variable; distinguish by type of missingness, if applicable.
Complete cases	M2	Describe number of complete observations when considering outcome and predictors for any candidate model described in P1.
Patterns	M3	Investigate patterns of missing values across all variables, either as tables or appropriately visualized. Can be structured by structural variables.
IDA screening domain: Univariate descriptions (predictors and outcome)		
Categorical variables	U1	Summarize frequency and proportion for each category or with ordinal plots. If it is considered to collapse rare categories, summarize frequencies of collapsed categories.
Continuous variables	U2	Inspect distributions with high-resolution histogram, summary of main quantiles and extremes mean, first four moments, number of distinct values. Describe the mode of the data and its frequency. Inspect distributions of transformed variables, if applicable.
IDA screening domain: Multivariate descriptions (structural variables and predictors)		
Association	V1	Visualize and summarize the association of each predictor with the structural variables
Correlation	V2	Quantify association with pairwise correlation coefficients between all key predictors in a matrix or heatmap
Interactions, if applicable	V3	Evaluate bivariate distributions of the predictors specified in interactions. Include appropriate graphical displays.
Multivariate analyses – Extensions		
Stratification, if applicable	VE1	Compute summary statistics of predictors and describe variation between strata defined based on level of measurement, e.g. centers, providers, locations or other variables described as stratification variables in the analysis strategy

IDA Plan (data screening) for longitudinal studies

Missing values (at different time points)

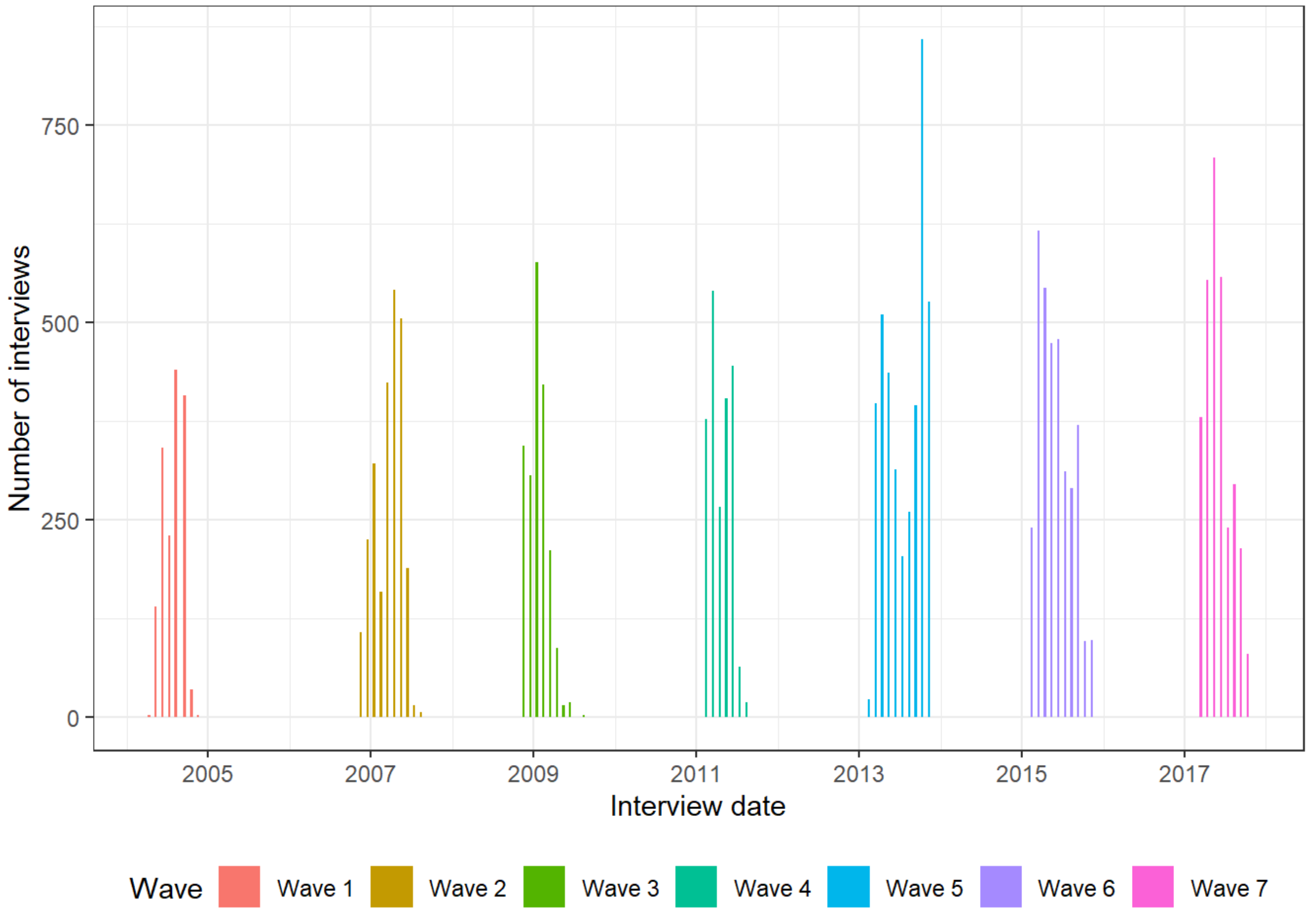
Unit missingness
Variable (item) missingness
Patterns
Predictors of missingness

Dropout effect



IDA screening domain: Missing Values		
Unit missingness	M1	Describe unit non-response, loss-to-follow-up and intermittent missingness, if applicable. Break down by the reason for missingness.
Variable (item) missingness	M2	Provide number and proportion of missing values for each variable at each time point as appropriate for fixed or time-varying variables.
Patterns	M3	Describe patterns of missing values across variables at each time point and across time points.
Predictors of missingness	M4	Explore whether there are predictors of missingness by comparing complete vs incomplete cases or investigate predictors of time to dropout, as appropriate; the aim can be the understanding of the missing data mechanism or the identification of potential auxiliary variables, i.e. variables not required for analysis but that can be used to recover some of the missing information.
Extensions: Missing Values		
Dropout effect	ME1	Visualize mean profiles of a continuous outcome by time metric stratified time to drop-out. Evaluate predictors of time to drop-out.
Stratified description of missingness	ME2	Describe missingness stratifying the summaries by variables that might influence the frequency of missing values, if relevant (for example type of interview).

IDA Plan (data screening) for longitudinal studies



Time metric of data collection process
Time metric of analysis strategy

IDA Plan (data screening) for longitudinal studies

Participation profile

Time frame

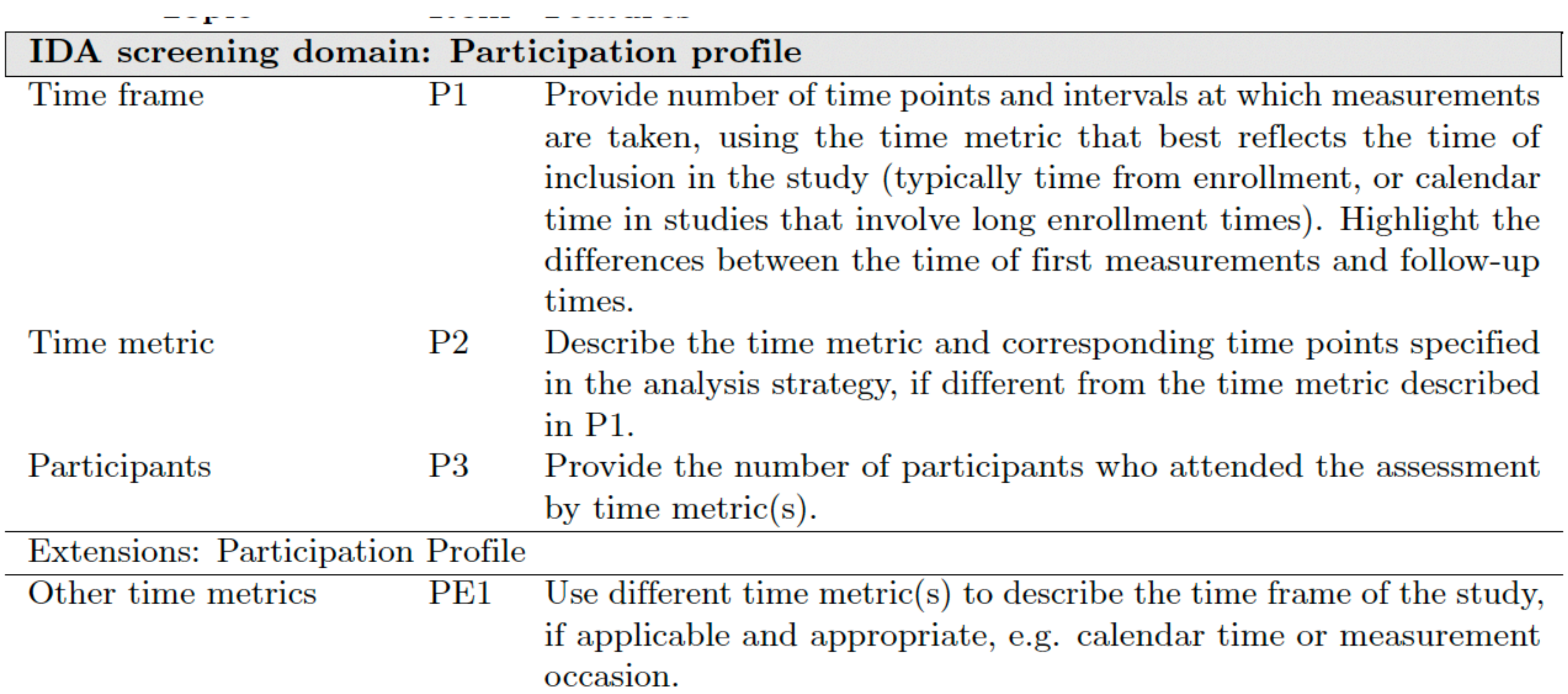
- Number of time points

Time metric

- Time metric of data col
- Time metric of analysis

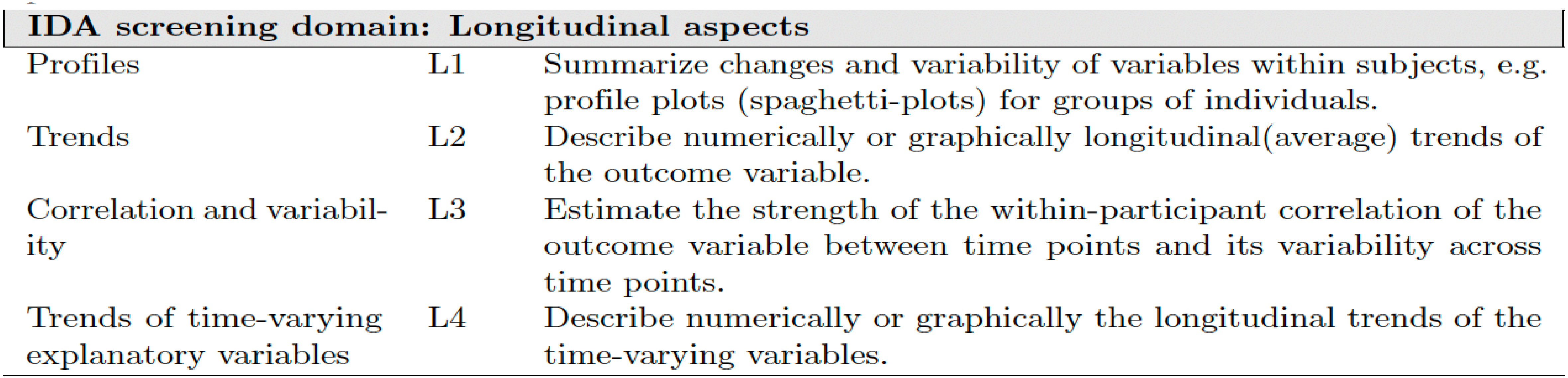
Participants

- Number of participants
  measurement occasion



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 of inclusion in the study (typically time from enrollment, 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).
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 measurement occasion.

IDA Plan (data screening) for longitudinal studies



IDA screening domain: Longitudinal aspects		
Profiles	L1	Summarize changes and variability of variables within subjects, e.g. profile plots (spaghetti-plots) for groups of individuals.
Trends	L2	Describe numerically or graphically longitudinal(average) trends of the outcome variable.
Correlation and variability	L3	Estimate the strength of the within-participant correlation of the outcome variable between time points and its variability across time points.
Trends of time-varying explanatory variables	L4	Describe numerically or graphically the longitudinal trends of the time-varying variables.

Profiles

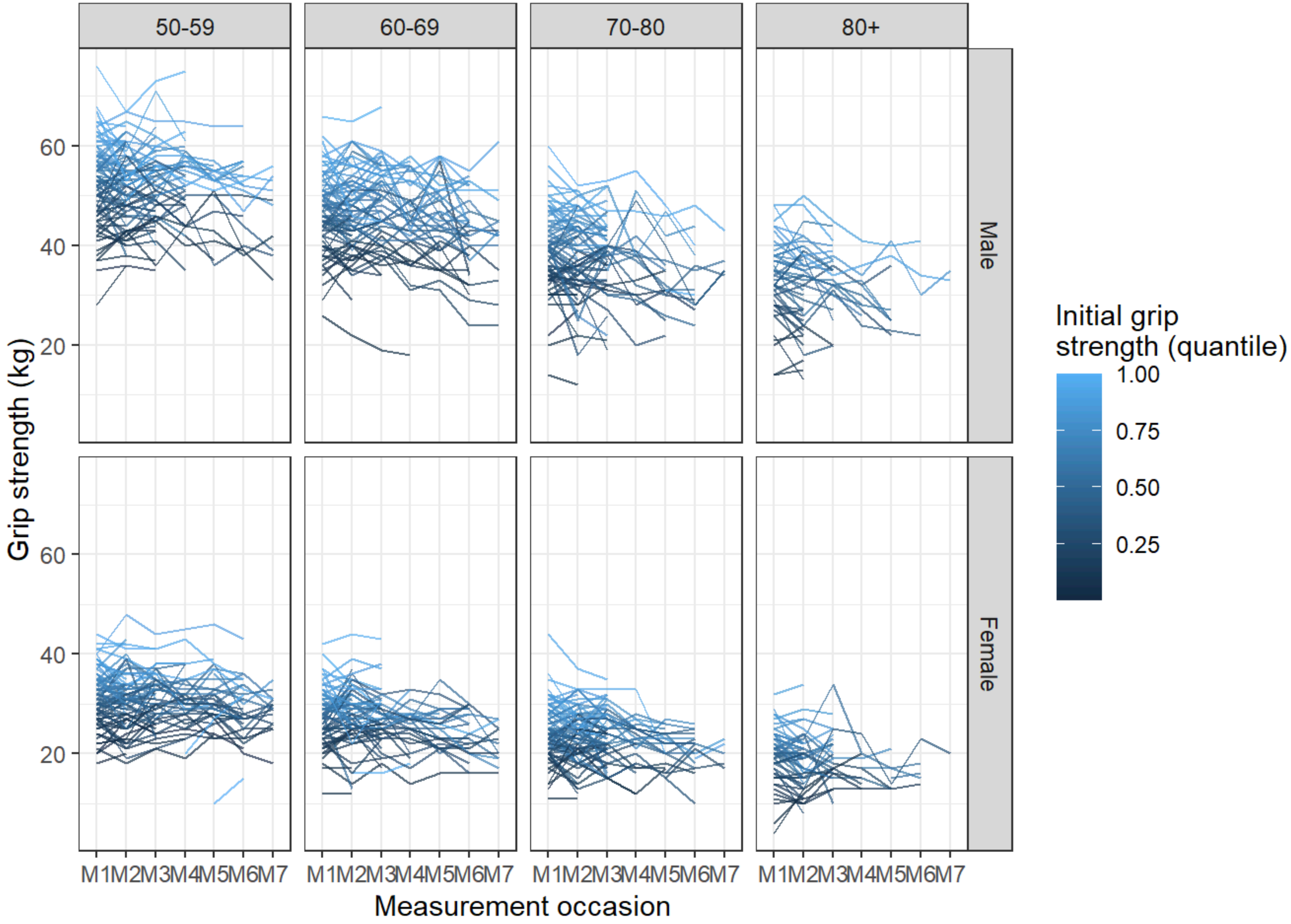
- Changes of variables within subjects

Trends

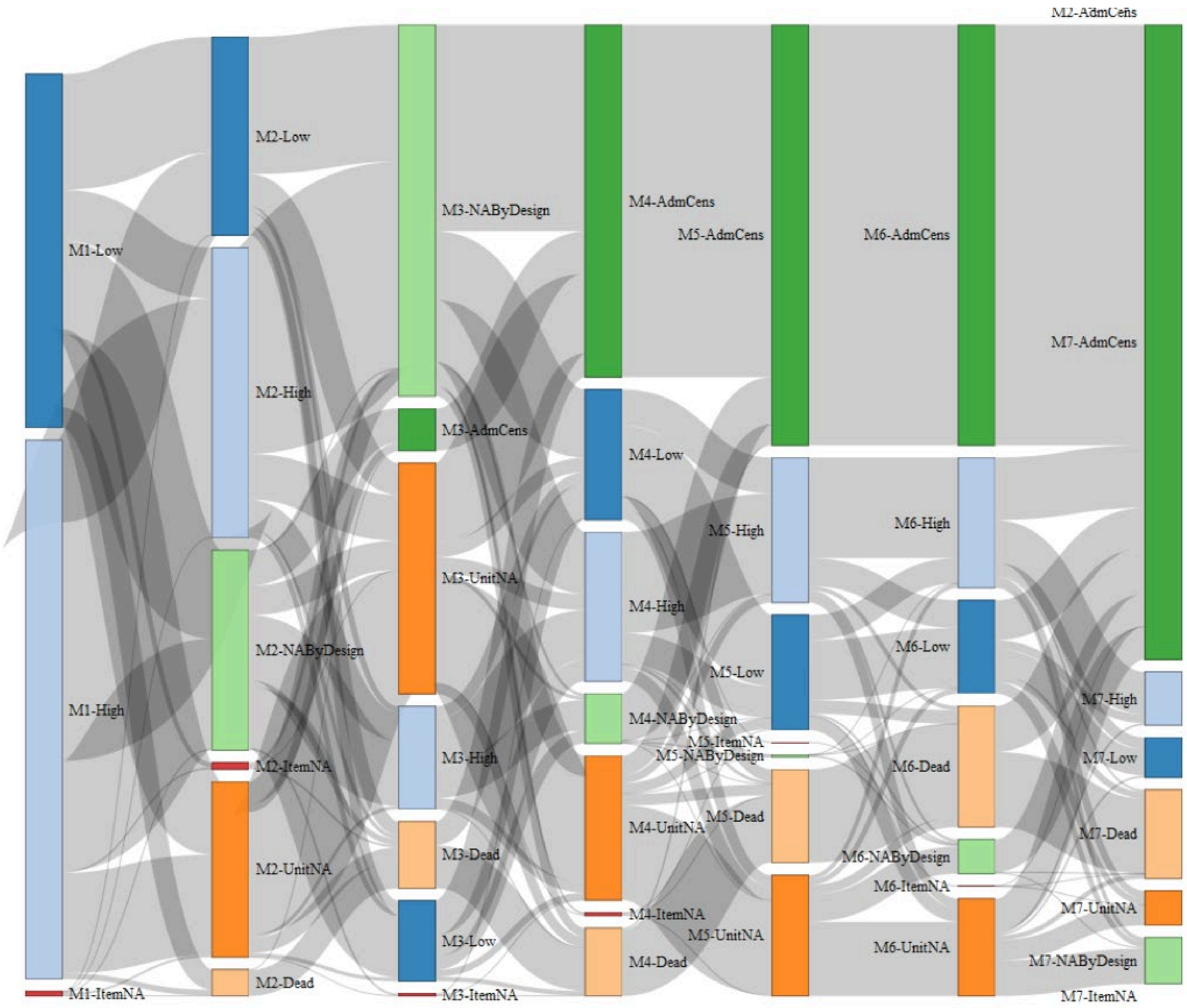
- Longitudinal trends of the outcome variable
- Longitudinal trends of time-varying explanatory variables

Correlations and variability Variability across time points

Profile plots



Time varying variables



In practice: Communication and workflow

1. Project title, investigators
   - Objectives
   - Hypotheses
   - Outcomes, inclusion, exclusion criteria

Meeting with collaborators: What are the objectives and hypotheses? Will the data support objectives? Discuss analysis plan.

2. Initial data analysis (data screening)
   - Missingness
   - Univariate descriptions
   - Multivariate descriptions

Meeting with collaborators: Are data as expected? Any issues to consider in the analysis plan?

3. Statistical Analyses

Meeting with collaborators: Explain/interpret/discuss results

Lessons learned

Propose an IDA plan and a statistical analysis plan for the study protocol

> **“**The data analysis plan consists of two parts. The aim of initial data analysis (IDA) is to examine data properties to ensure transparency and integrity of preconditions to conduct appropriate statistical analyses to answer the research questions. […]”

Clinicians and statisticians learn from each other when discussing IDA report:

- Understanding data content better; learning about expected or unexpected data properties
- Update metadata/database
- “Reality hits” for research aims; confirm suspicions from the protocol phase
- If applicable: time metric/time zero discussed multiple times in the process

What the PI knew (or not) and didn’t tell the statistician

1. Data were collected in two periods with a gap of several months
2. New procedure was introduced and performed together with standard practice procedure for a while (outcome= duration)
3. Physical activity data were collected with GPS: indoors had limited use
4. Unexpected univariate distributions (multi-modal, spikes at zero)
5. IDA findings were more interesting to the investigators than the statistical model

IDA is the foundation for modeling: presentation, checking expectations, interpretation, model decisions.

References

- **IDA framework**: Huebner M, le Cessie S, Schmidt CO, Vach W . A contemporary conceptual framework for initial data analysis. Observational Studies 2018; 4: 171-192. https://doi.org/10.1353/obs.2018.0014
- **Website with IDA report and R code (cross-sectional):** https://stratosida.github.io/regression-regrets/
- **TG3 website**: https://www.stratosida.org