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

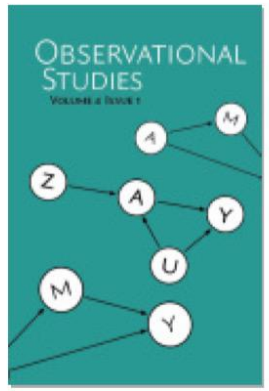
TG3 Papers

1

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

2

RESEARCH ARTICLE

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)

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:

4

RESEARCH ARTICLE

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

3

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

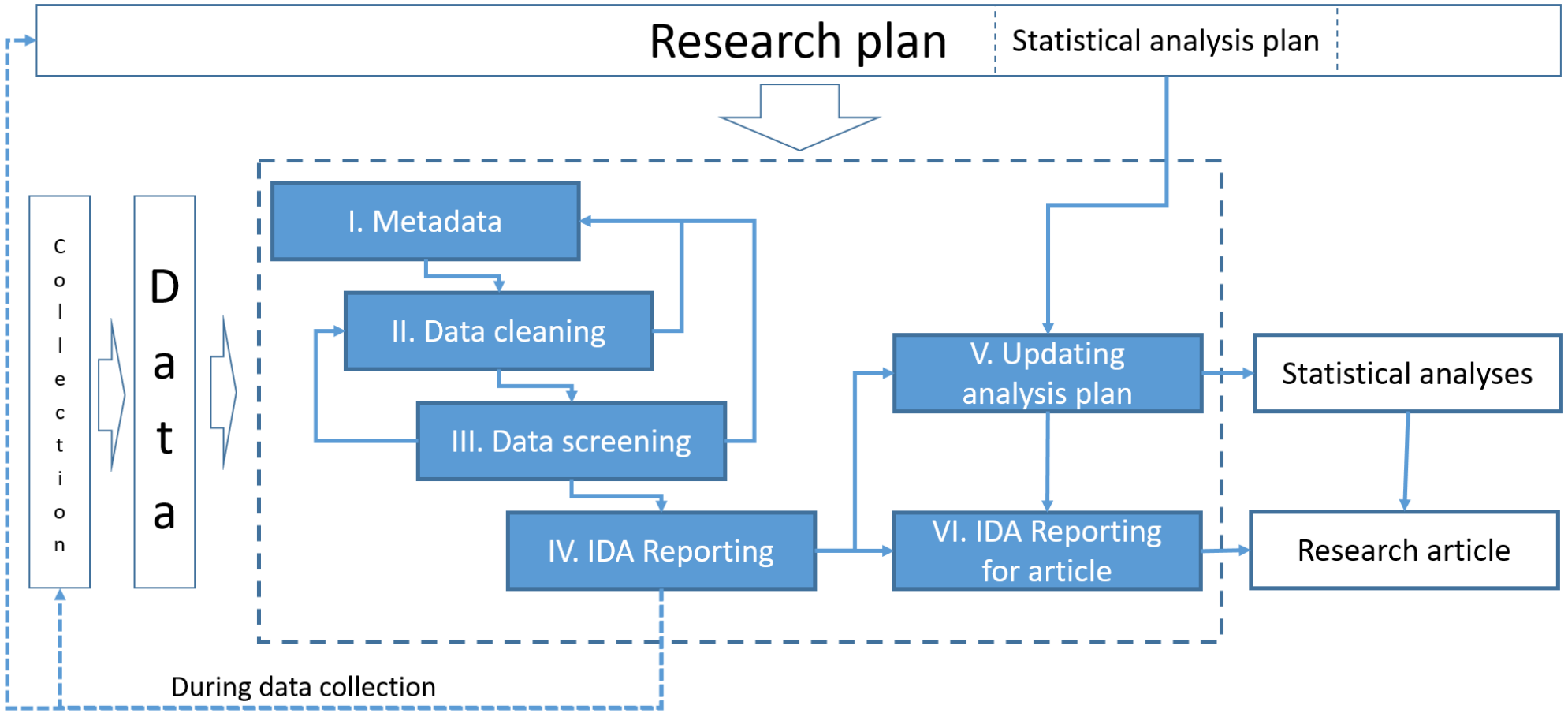
Open Access Article

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]

5

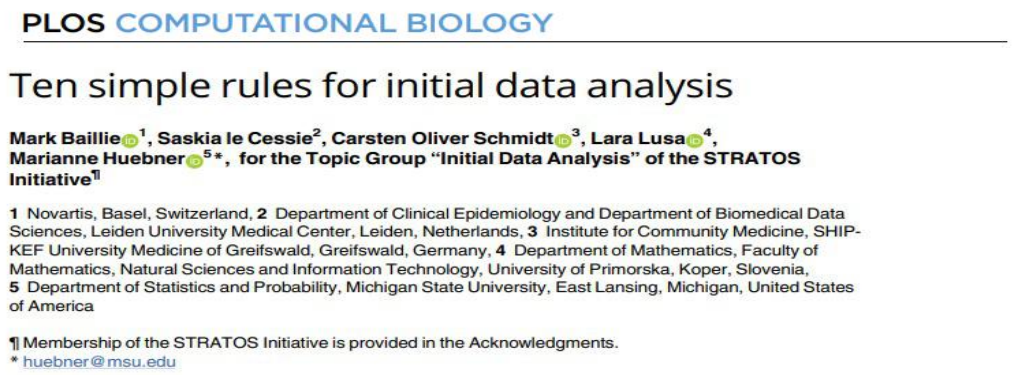
What is Initial Data Analysis?



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

Ten (Simple) Rules of Initial Data Analysis

1. Develop an IDA plan that supports the research objective
2. IDA takes time and resources
3. Make IDA reproducible
4. Context matters: know your data
5. Avoid sneak peeks - IDA does not touch the research question
6. Visualize your data
7. Check for what is missing
8. Communicate the findings and consider the consequences
9. Report IDA findings in research papers
10. Be proactive and rigorous



PLOS COMPUTATIONAL BIOLOGY

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[¶]

1 Novartis, Basel, Switzerland, **2** Department of Clinical Epidemiology and Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands, **3** Institute for Community Medicine, SHIP-KEF University Medicine of Greifswald, Greifswald, Germany, **4** Department of Mathematics, Faculty of Mathematics, Natural Sciences and Information Technology, University of Primorska, Koper, Slovenia, **5** Department of Statistics and Probability, Michigan State University, East Lansing, Michigan, United States of America

¶ Membership of the STRATOS Initiative is provided in the Acknowledgments.
* huebner@msu.edu

The 10 rules are based on extensive experience with research projects, collaborations with domain experts, and discussions among an international group of applied statisticians.

Source: Baillie M, le Cessie S, Schmidt CO, Lusa L, Huebner M, for the Topic Group "Initial Data Analysis" of the STRATOS Initiative (2022) Ten simple rules for initial data analysis. PLoS Comput Biol 18(2): e1009819. https://doi.org/10.1371/journal.pcbi.1009819

IDA in the News

■ FORSCHUNG | 24 11 22

Datenqualität muss in den Fokus!

● **Uneinheitliche Datenstandards, Datenfehler sowie intransparente Wege der Datenaufbereitung und -darstellung sind wesentliche Stolpersteine in den Gesundheits- und Lebenswissenschaften. Ist ein systematischerer und transparenterer Umgang mit Datenqualität möglich?**

Exploratory data analysis

Article Talk

From Wikipedia, the free encyclopedia

In statistics, **exploratory data analysis** (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing. Exploratory data analysis has been promoted by John Tukey since 1970 to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA),[1][2] which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

It's not correct, though -> Glossary/Terminology

DE GRUYTER Open Statistics 2022; 3:39–47

Communication

Werner Vach*

Initial data analysis: A new technology not yet ready to use

> "IDA can only have a positive impact on research quality if findings from a systematic data screening lead to reasonable changes in the final analysis or its interpretation."

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

1. **Regression without regrets (cross-sectional)**

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 checklist for longitudinal data**

Leads: L. Lusa (collaboration with Kate Lee, TG1, and Cecile Proust-Lima, TG4)

Scope: Regression model that uses repeated measurements obtained for individuals

“Generic“ IDA Plan (data screening) for a cross-sectional study

Data screening

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.

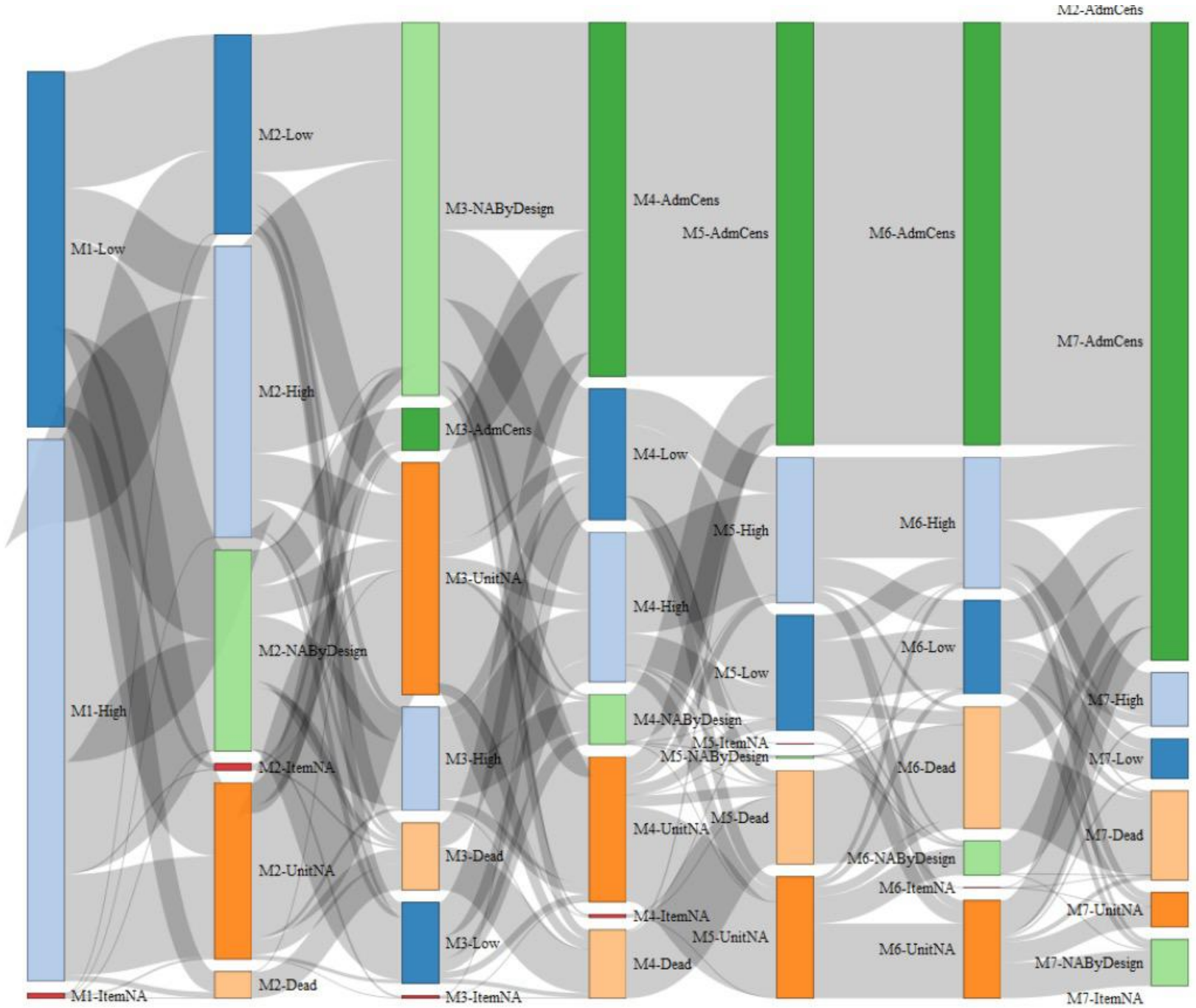
“Generic“ IDA Plan (data screening) for longitudinal studies

Topic	Item	Features
IDA screening domain: Participation profile		
Time frame	P1	Provide number of time points and intervals at which measureme are taken, using the time metric that best reflects the time inclusion in the study (typically time from enrollment, or calen time in studies that involve long enrollment times). Highlight differences between the time of first measurements and follow times.
Time metric	P2	Describe the time metric and corresponding time points speci in the analysis strategy, if different from the time metric descri in P1.
Participants	P3	Provide the by time met
Extensions: Participation Profile		
Other time metrics	PE1	Use different if applicable occasion.
Data collection	PE3	Describe ch

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.
Extensions: Longitudinal aspects		
Cohort/Period effects	LE1	If appropriate, summarize possible cohorts or period effects (for example, age birth cohorts or period cohorts defined by the calendar time/wave of measurement) on the outcome, and on the explanatory variables, to assess if the variation of the outcome can occur because of these effects.

Time metric of data collection process
Time metric of analysis strategy

Time varying variables



IDA in the STROBE Checklist?

IDA data screening elements

- characteristics of study participants
- number of missing participants
- Information about confounders
- summarize follow-up time, report summary measures over time

SAP or consequences of IDA findings

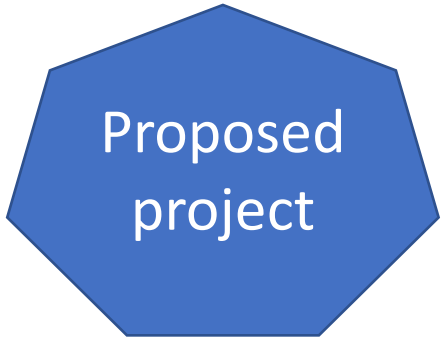
- addressing potential sources of bias, methods for handling missing data
- methods to control for confounding
- methods to examine subgroups and interactions
- sensitivity analyses

IDA in Statistical Analysis Plans?

Generic statements (e.g. DEBATE):

- time points at which the outcomes are measured
- timing of lost to follow-up
- missing data
- description of baseline characteristics and outcomes

What is needed:
Guidance on IDA plan as part of the SAP



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."