Initial data analysis for longitudinal data
A general framework

Lara Lusa[1,2] and Marianne Huebner[3]

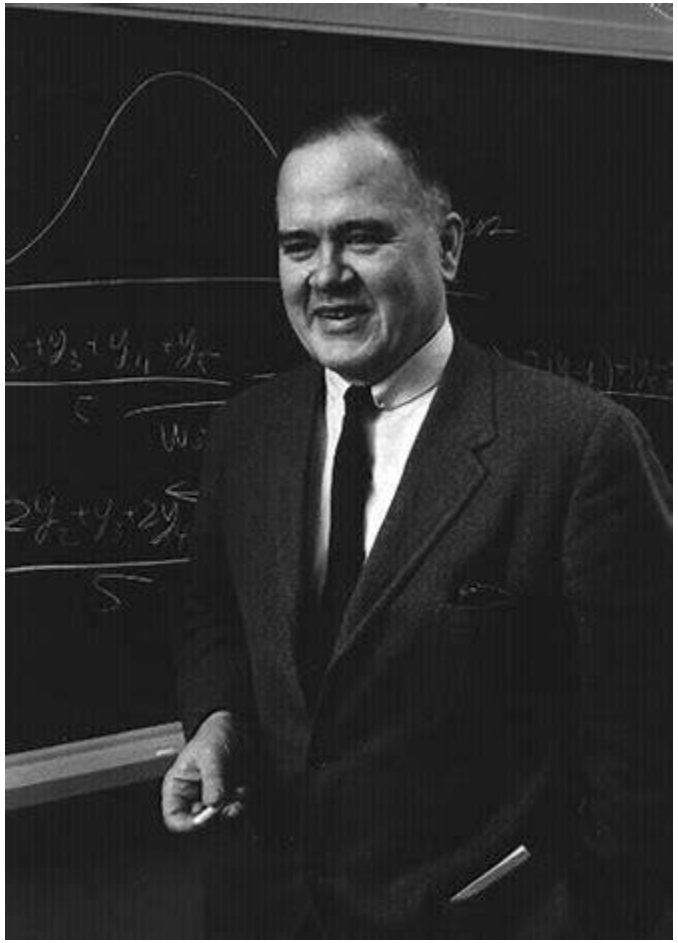
[1]University of Primorska, Koper/Capodistria, Slovenia

[2]University of Ljubljana, Ljubljana, Slovenia

[3]Michigan State University, East Lansing, MI, USA

on behalf of the Topic Group “Initial Data Analysis” of the STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies).

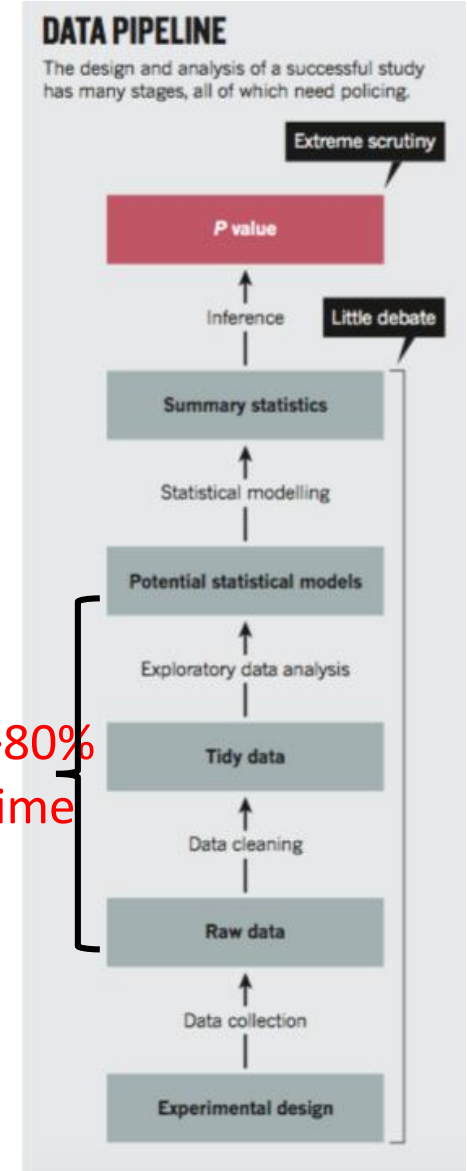
Why Initial Data Analysis?



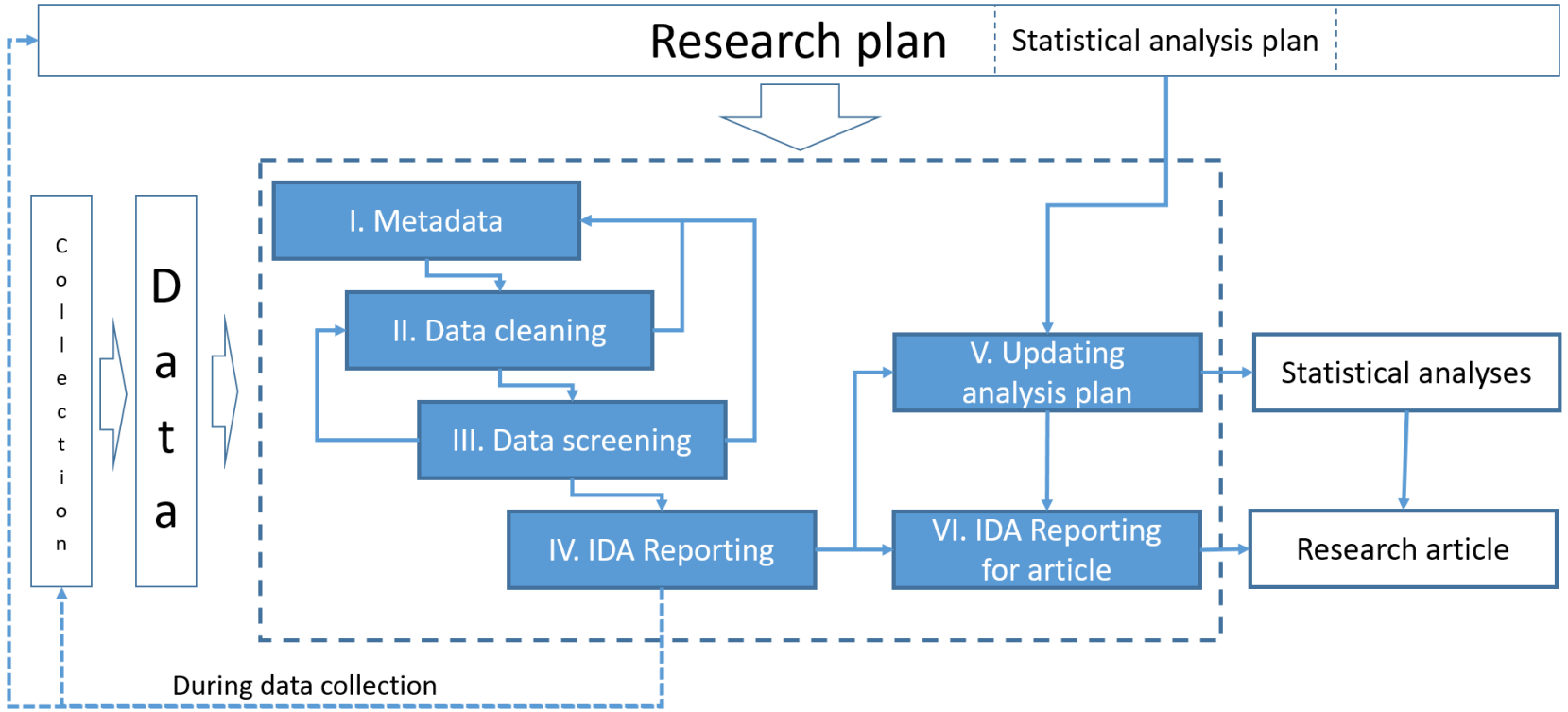
"All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data"

John W. Tukey – Exploratory data analysis, 1977

What is Initial Data Analysis?

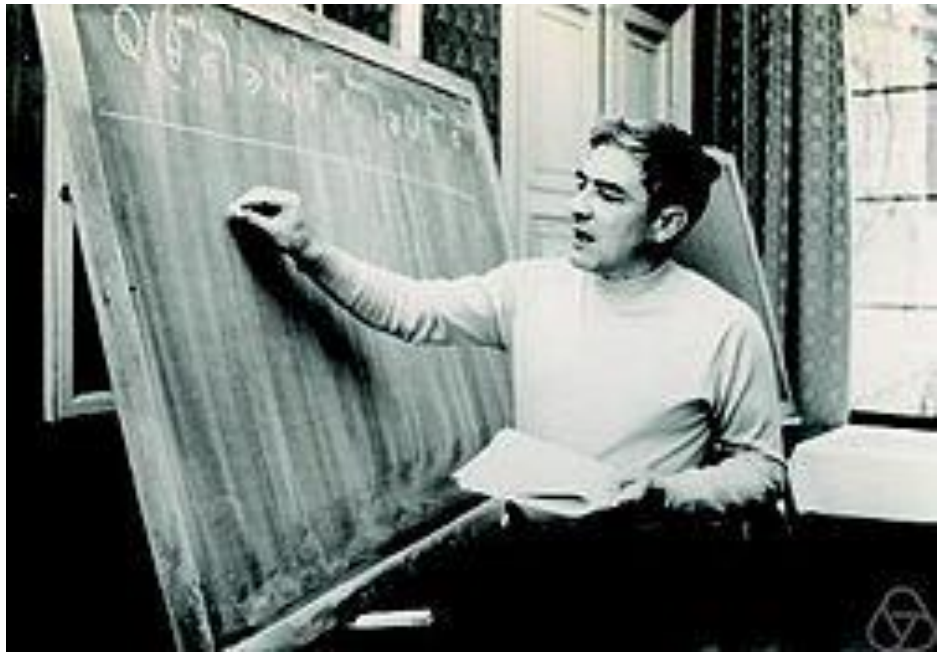


- Part of data pipeline that typically takes place between the end of the data collection and start of statistical analysis



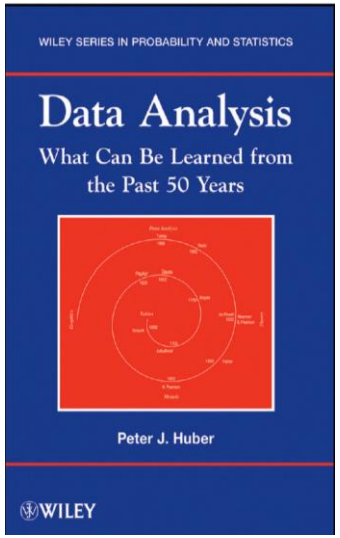
- Key principle: it does not touch the research questions

Why not touching the research question?



Peter J. Huber, Strategic data analysis, 2011

- "It is no longer possible to calculate reliable P-values after one has looked at the data - unless they are based on fresh data, they may be worse that useless, namely misleading"



Current situation of IDA

- IDA is often done informally and unstructured
- The content of IDA is unclear: data cleaning? basic data summaries? exploratory analysis? modeling?
- Often statistical analyses are performed **without**
  - systematically checking for errors in the data,
  - a clear understanding about the underlying features of the data
  - knowledge on the suitability of the intended analyses,
  - knowledge whether the data actually could provide answers to the research questions of interest.
- Top problems: poor data preparation, misinterpretations of results [Rexer, data science survey 2017]

"People have run analyses that resulted in completely false conclusions, which were then used by the business"

Reporting IDA in clinical papers – meta-analysis

Metadata – data cleaning – data screening – refining/updating analysis plan

- Reporting on IDA is sparse or selective.
- Information on IDA can be found in all sections of a paper.
- There are statements that could be interpreted as IDA, but it is not clear whether this was pre-planned.
- Characteristics of participants are listed without comments.
- Associations among variables are not reported.
- Cluster variables were present, but not described nor accounted for in analyses.
- Reporting on missingness is incomplete (decisions about handing of missing data?).
- Are (all) sensitivity analyses preplanned?
- Were transformations preplanned? (e.g categorization, recoding, omissions)
- Changes in modeling strategies? (refinement, extensions, reduction of models, stratification, ...)

Huebner M, Vach W, le Cessie S, Schmidt C, Lusa L. Hidden Analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses. BMC Med Res Meth 2020; 20:61

How should IDA be performed?

- Have an IDA plan, developed based on the statistical analysis plan
- Write a report that is reproducible
  - Follow a systematic process!
  - Use appropriate and effective methods (visualization, summaries, …)
- Be transparent in the reporting
  - IDA methodology (**Methods).**
  - **IDA results (Methods or Results).**
  - Impact of IDA on interpretation **(Discussion)**.
  - Pre-planned decisions **(Methods).**
  - IDA driven alterations of the analysis plan **(Methods)**.
  - Full reporting of missingness **(Results)**.

Generic IDA Plan

0. Metadata and data cleaning

1. IDA domain: missing values

- Number and proportion of missing values for each variable and patterns of missingness

2. IDA domain: univariate distributions

- Categorical variables: frequency and proportion for each category
- Continuous variables: high resolution histogram, quantiles

3. IDA domain: multivariate system of variables

- Matrix/plot of correlation coefficients between all independent variables.
- Visual presentation of the association of each covariate with pivotal covariates.
- Variable clustering or redundancy analysis, if appropriate

IDA for longitudinal data – set up

- Repeated measurements are obtained from the same individual (not time series)
- Time points: preplanned, equally spaced, random/fixed, un/balanced
- Time metric: time since inclusion, time since an event, calendar time, follow up occasion, age, …
- The main research question addressed in the statistical analysis plan (SAP) involves the estimation of a regression model that uses the repeated measurements obtained at individual level.
- In SAP:
  - modelling strategy (model, specification of correlation and variance structure, identification of random variation/components of variation of data)
  - outcome variable
  - covariates: independent variables selected to be included in the regression model
  - "pivotal variables" are independent variables strongly associated with the outcome variable that can be used to structure multivariate IDA.
- Aim: propose a framework that defines an IDA plan for longitudinal studies

DATA retrieval and data management	Actions and why is this of interest?
Preparation of data in different formats* Long (1 row/measurement) and wide (1 row/subject) format	• To facilitate data summarization and modelling • To facilitate visualization • To explore the use of different time metrics • To make the structure of missing data easier to explore
Additional initial data manipulation*	• To harmonize variable definitions across measurements • To link additional information (mortality data, survey weights, ...) • To identify time-fixed and time-varying covarites
DATA cleaning	• To find inconsistencies (between observed data and metadata, in the reporting of values, logical inconsistencies, based on repeated measurements*) • To find mistakes • To provide flags, based on which decisions about corrections will be made

DATA SCREENING	Why is this of interest?
Participation and time frame of the study*	• To summarize the charcteristics of the study over time* • To summarize the participation in the study and drop-outs* • To summarize the characteristics of the time metric and compare the observed and expected data*
Missing data	• To understand the different missing values (Intermittent missingness, death, lost to follow-up, missing by design, missing end of study)* • To inform about the relevance of missing information • To decide on a proper strategy to handle missing values (based on the comparison of the characteristics of participants with complete and missing data, on the preditors of missingness, to evaluate if missingness is predictable based on previous measurements*, ...)
Univariate distribution of each variable	• To describe the population (to whom does the model apply) • To support later decisions in modeling (e.g. collapsing categories, how many df for a variable) • To interpret regression coefficients (do we need to rescale them) • To identify potential robustness issues
Associations between variables	• To learn about bivariate or higher-order distributions (interactions relevant?) • To support later decisions in modeling • To support interpretation of results of modeling • To judge the need for later choices of data reduction methods • To describe the longitudinal trends of the outcome* • To explore the correlation structure between repeated outcomes* • To explore the correlation structure between time-varying covariates* • To explore subject-to-subject changes and variability of the outcome over time*

Survey of Health, Aging, and Retirement in Europe

Data on health and socioeconomic variables of non-institutionalized individuals **aged 50 and older** across **27 European countries and Israel**. 140 000 participants, collected in years 2004 to 2018 in 7 waves. Thousands of questions about demographics, health and socio-economic status.

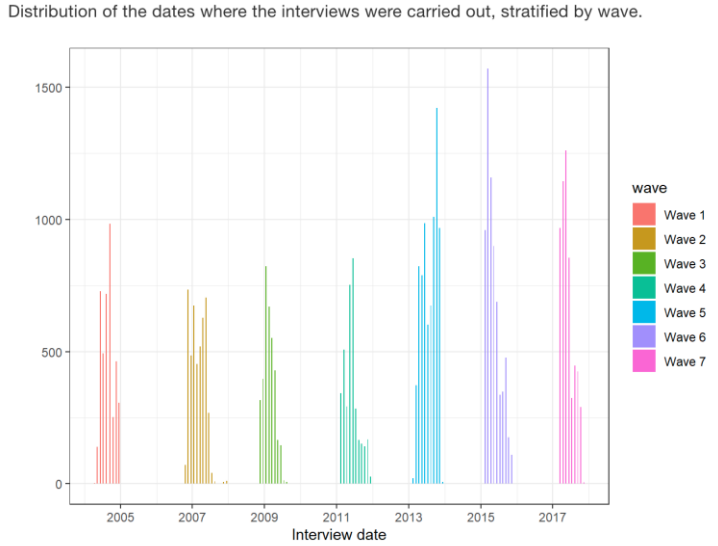
Subset: Denmark and Sweden, 2004 to 2018 (7 waves)
Aim: Investigating age-associated change in max grip strength stratified by gender
Outcome: maximum grip strength
Covariates measured at first interview: gender, height, smoking status
Time-varying covariates: age, weight, physical activity (vigorous or low intensity)
Population characteristics: education level, depression and other comorbidities (cancer, stroke, heart attack, lung disease, cancer)

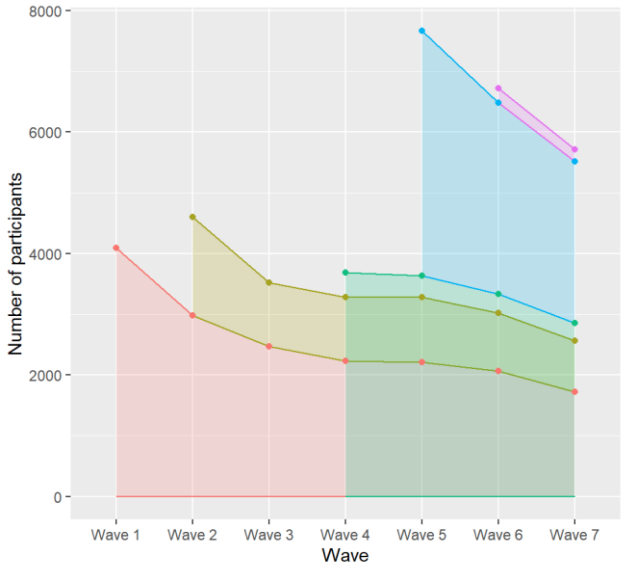
- The complexity of the data makes it a very interesting example, far from the “toy-data” often used in teaching or as examples in methodological papers

Interim Lessons Learned

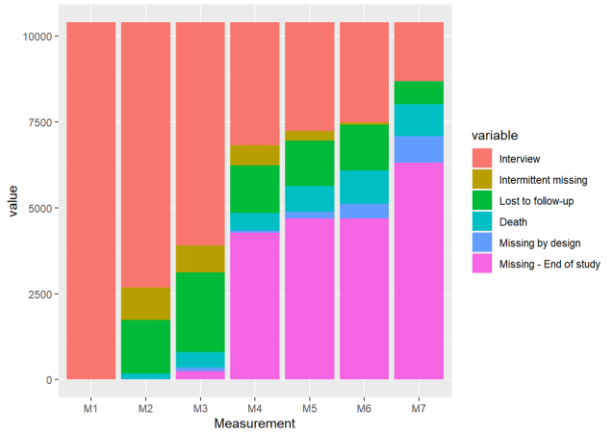
- Graphical displays for longitudinal data with large numbers of participants need to be developed.
- Meta data are essential for understanding data properties and correctly use them in the statistical analysis plan
- Expect a lot of effort for data cleaning despite well documented and pre-cleaned data
  - Inconsistencies between waves, aggregation of data
- Expect to refine the preplanned statistical analysis plan
  - Time metric, correlation matrix, heteroskedastic variances, exclusion/inclusion of variables
- Output: structured, very large, reproducible document containing a very thorough data exploration with consequences on statistical analysis plan

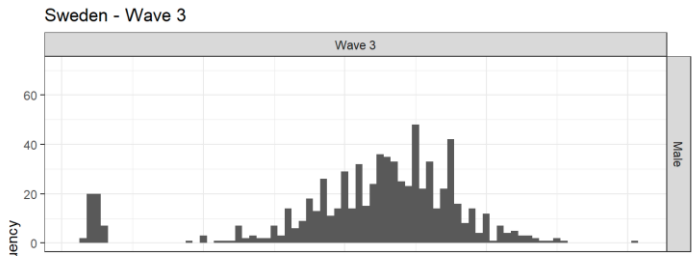
Will be made fully publicly available, can be used by others as a template

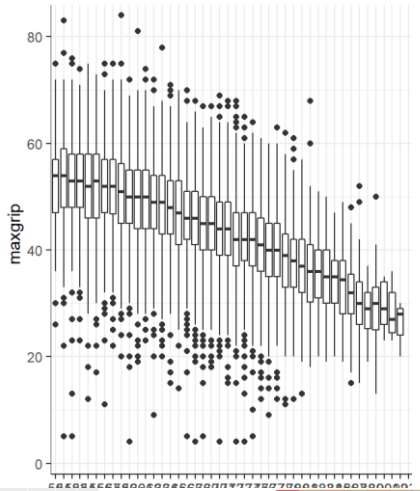


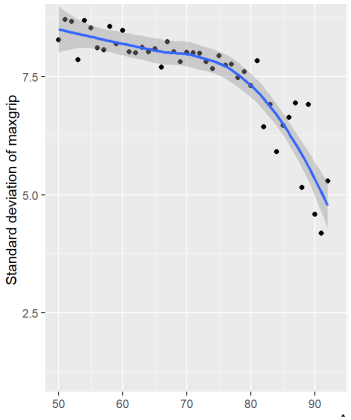


Overall characteristics at baseline.		
	N	*N=10399*
country : Denmark	10399	0.47 $^{4932}/_{10399}$
gender : Female	10399	0.53 $^{5460}/_{10399}$
age_int	10399	55.00 61.00 69.00 62.31 ± 8.47

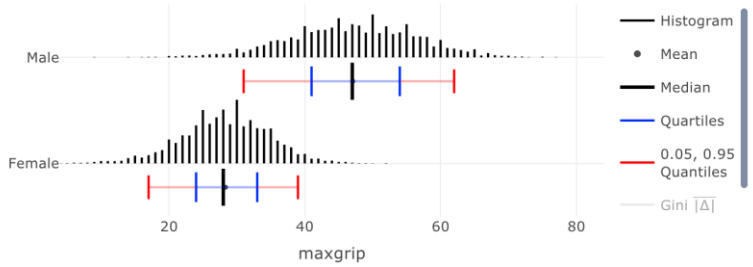


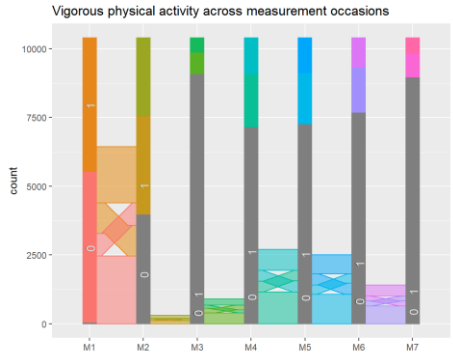


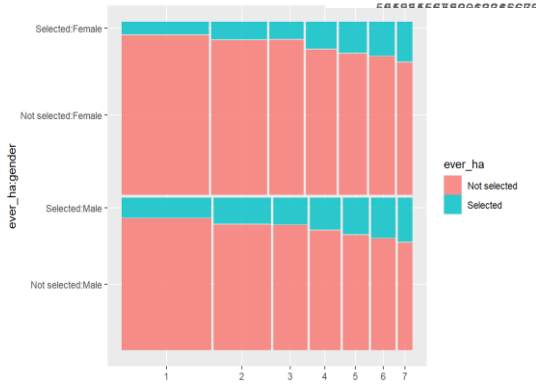


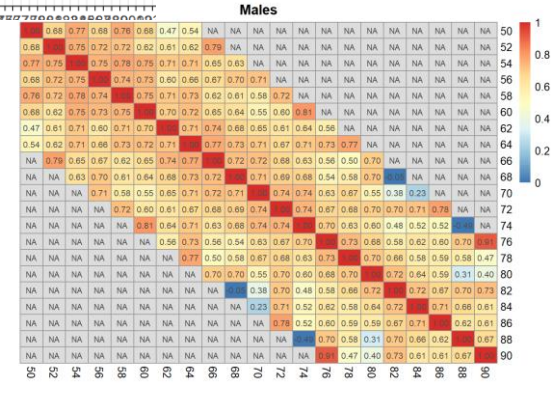


5.4.2.1.1 Grip strength









In Summary

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

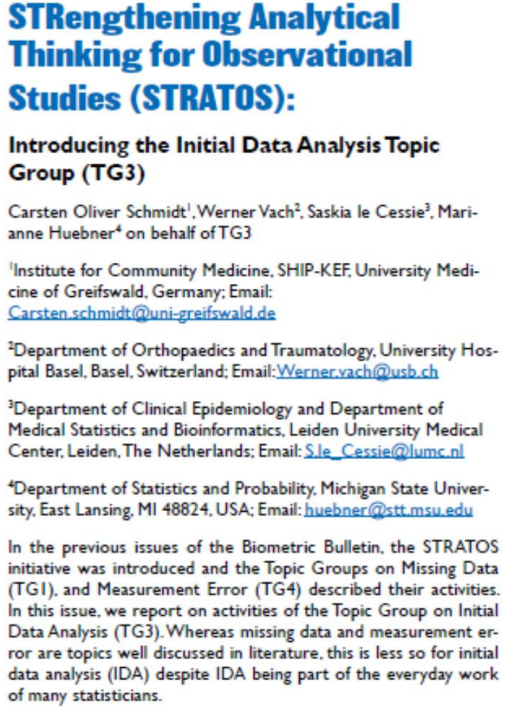
IDA takes time and planning

- BUT: finding problems after modeling takes MORE time and may miss issues (not systematic)
- Help: code and workflow

IDA needs to be reported: Suggestions in Huebner et al, BMC Med Res 2020

Research studies need both:
Statistical analysis plan + IDA plan

TG3 References and conclusions



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

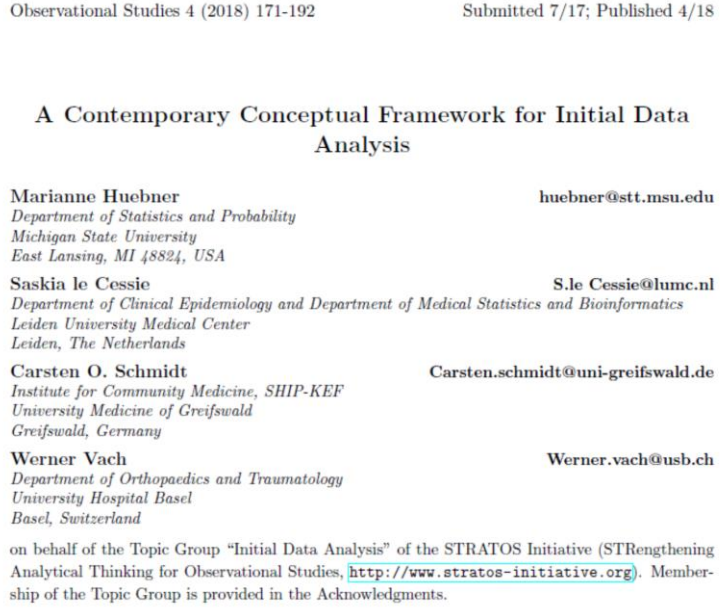
[1]Institute for Community Medicine, SHIP-KEF, University Medicine of Greifswald, Germany; Email: Carsten.schmidt@uni-greifswald.de

[2]Department of Orthopaedics and Traumatology, University Hospital Basel, Basel, Switzerland; Email: Werner.vach@usb.ch

[3]Department of Clinical Epidemiology and Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands; Email: S.le_Cessie@lumc.nl

[4]Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA; Email: huebner@stt.msu.edu

In the previous issues of the Biometric Bulletin, the STRATOS initiative was introduced and the Topic Groups on Missing Data (TG1), and Measurement Error (TG4) described their activities. In this issue, we report on activities of the Topic Group on Initial Data Analysis (TG3). Whereas missing data and measurement error are topics well discussed in literature, this is less so for initial data analysis (IDA) despite IDA being part of the everyday work of many statisticians.



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

A Contemporary Conceptual Framework for Initial Data Analysis

Marianne Huebner huebner@stt.msu.edu
Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824, USA

Saskia le Cessie S.le Cessie@lumc.nl
Department of Clinical Epidemiology and Department of Medical Statistics and Bioinformatics
Leiden University Medical Center
Leiden, The Netherlands

Carsten O. Schmidt Carsten.schmidt@uni-greifswald.de
Institute for Community Medicine, SHIP-KEF
University Medicine of Greifswald
Greifswald, Germany

Werner Vach Werner.vach@usb.ch
Department of Orthopaedics and Traumatology
University Hospital Basel
Basel, Switzerland

on behalf of the Topic Group "Initial Data Analysis" of the STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies, http://www.stratos-initiative.org). Membership of the Topic Group is provided in the Acknowledgments.

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)

"Contributions to correct conclusions made by good data is greater than made by sophisticated analysis" S. Senn

Initial Data Analysis Research Group

- Marianne Huebner, chair, (Michigan, USA)
- Carsten Oliver Schmidt, co-chair, (Greifswald, Germany)
- Saskia le Cessie (Leiden, Netherlands)
- Mark Baillie (Basel, Switzerland)
- Lara Lusa (Slovenia)

- Ackowledgements: Andrej Srakar (SHARE data) and Frank Lawrence (longitudinal modelling)

STRATOS INITIATIVE

https://www.stratos-initiative.org/

https://www.stratosida.org/members