STRATOS INITIATIVE

Initial data analysis: Making the effort worthwhile

Lara Lusa[1,2]

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

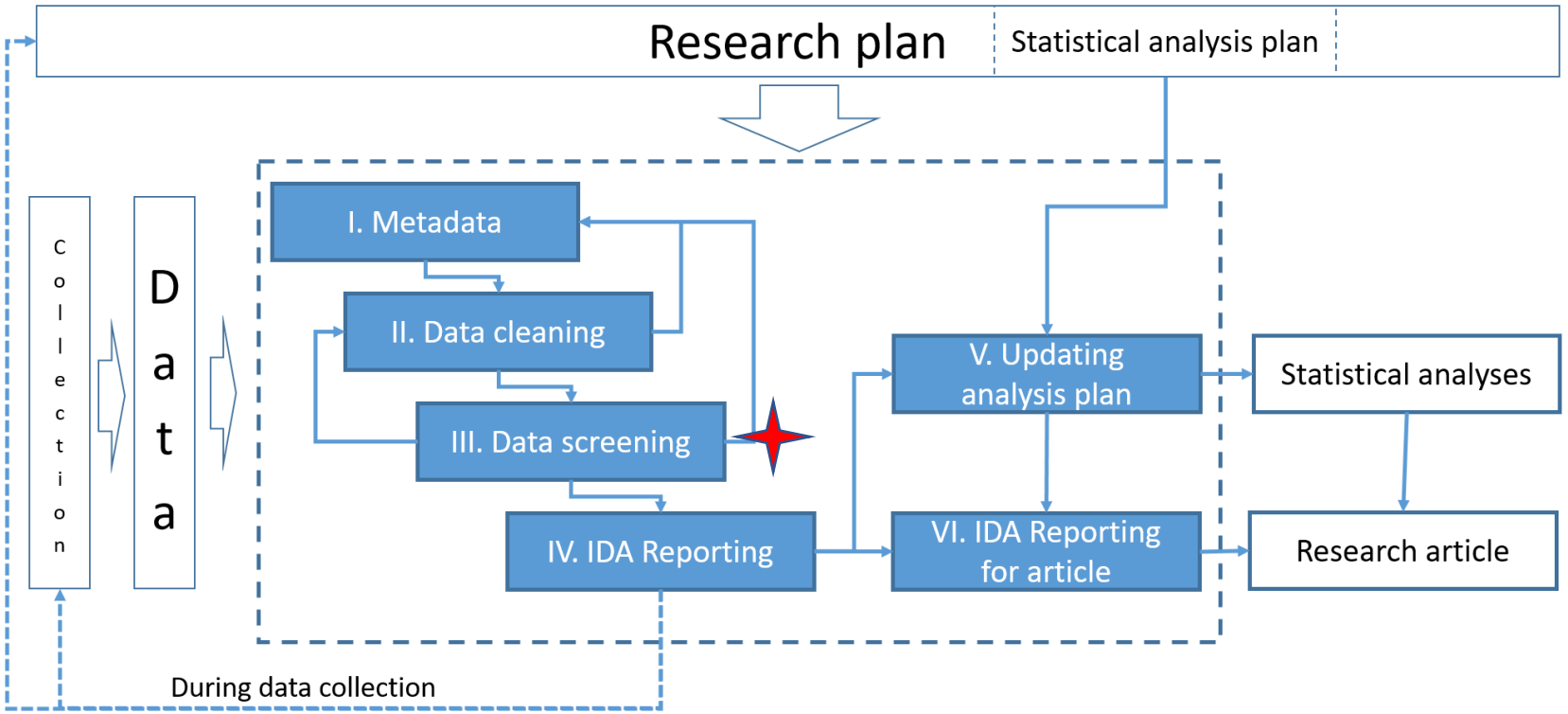
[2]University of Ljubljana, Ljubljana, Slovenia

Carsten Oliver Schmidt, Georg Heinze and Marianne Huebner

on behalf of the Topic Group (TG3) “Initial Data Analysis” of the STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies)

Initial Data Analysis

- AIM: provides analysis-ready data set including reliable knowledge about data properties to answer the research question, the step to check if the observed data correspond to expectations



III. Data screening

Aim: understand data properties

Participation profile (L)

Missing data

Univariate distributions

Multivariate distributions

Longitudinal aspects (L)

Possible consequence: changes in analyses and interpretations

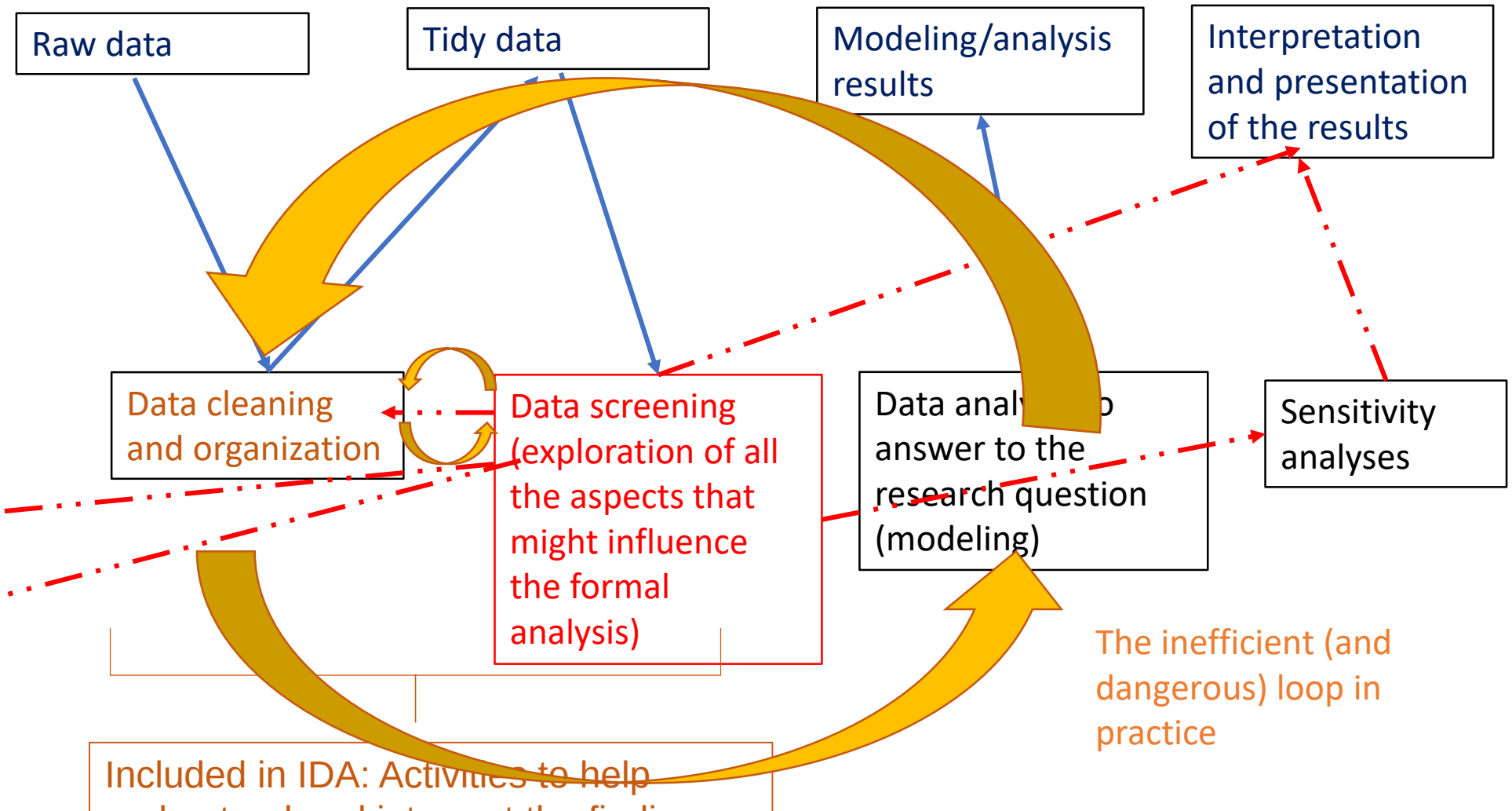
Huebner M, le Cessie S, Schmidt CO, Vach W. A contemporary conceptual framework for initial data analysis. Observational Studies 2018; 4: 171-192. https://doi.org/10.1353/obs.2018.0014

The data pipeline

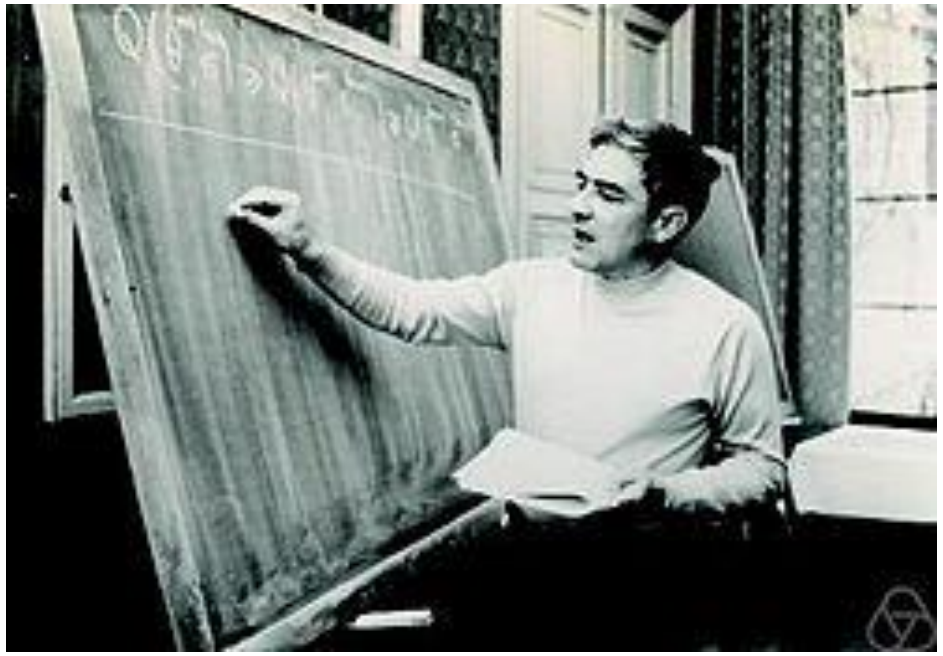
>80% time?

Prerequisites

- Study design
- Definition of the research question
- Background knowledge
- Statistical analysis plan
- Preparation of metadata
- Data collection

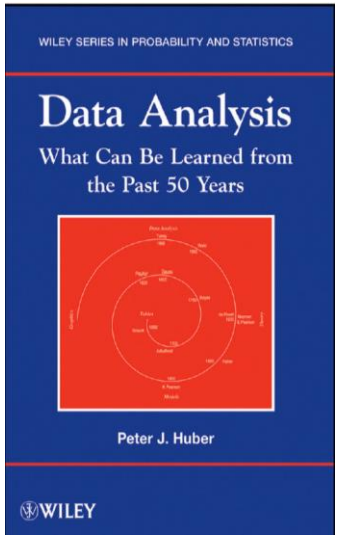


Why not touching the research question?



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 than useless, namely misleading"



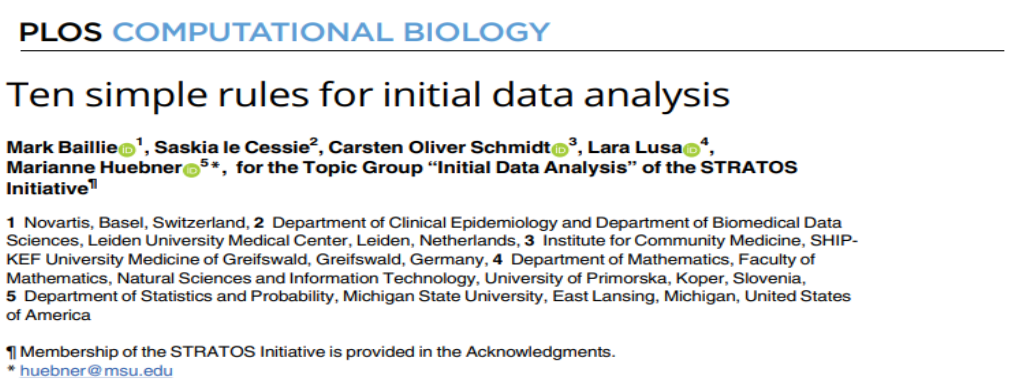
Current situation of IDA

- IDA is often done informally and unstructured
- IDA is usually not included in the (increasingly used) reproducible reports that describe data analyses.
- 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"

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

A (consensus based) checklist IDA for data screening

IDA screening domain: Missing values (predictor and outcome variables)		
Proportion	M1	Provide number and proportion of missing values for each predictor and for the outcome variable; distinguish by reason 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 (structural variables, predictors and outcome)		
Categorical variables	U1	Summarize frequency and proportion for each category or with appropriate plots. If it is considered to collapse rare categories, summarize also frequencies of collapsed categories. Define dummy variables. They should be different for variables with or without an order.
Continuous variables	U2	Inspect distributions with high-resolution histogram, summary of main quantiles (e.g. 1st, 5th, 25th, 50th, 75th, 90th, 99th) and extremes (e.g. 5 highest and 5 lowest values), mean, first four moments (mean, variance/standard deviation, skewness, kurtosis), number of distinct values. Describe the mode of the data and its frequency. Similarly, 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 (pairwise correlations) 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.

Source: https://stratosida.github.io/regression-report/

Context:

descriptive or predictive models

to relate an outcome variable (yes: continuous, counts, binary, no: survival, longitudinal, multivariate)

with a set of predictors (low-dimensional, p<n)

using regression models

To be submitted: Regression without regrets –initial data analysis is an essential prerequisite to multivariable regression

Heinze G., Baillie, M., Lusa L., Sauerbrei W., Schmidt CO, Harrell F., Huebner M. TG2 and TG3 STRATOS

Topic	Item	Features
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).

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.

+ Missing data domain : substantial modifications focused on aspects present in longitudinal studies

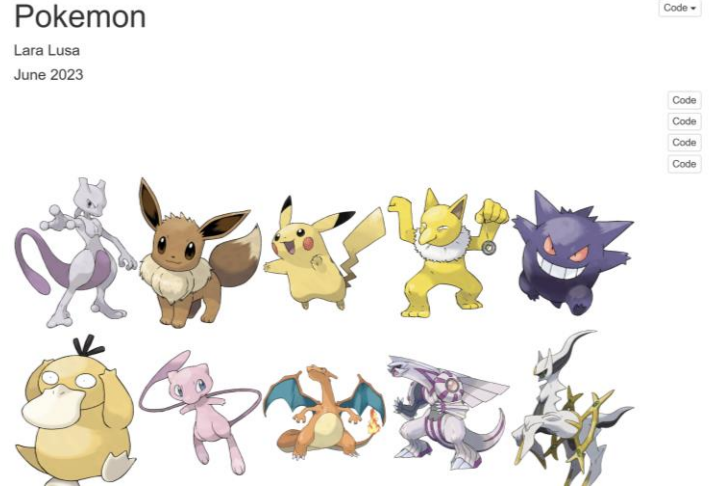
Extension:

Regression models for longitudinal outcomes

To be submitted: Initial data analysis for longitudinal studies to build a solid foundation for reproducible analysis

Lusa L., Huebner M., Schmidt CO, Lee KJ, le Cessie S, Baillie M., Lawrence F, Proust-Lima C.

Pokémon example: research question



- The research aim is
  - to describe the general characteristics of the available data
  - to find predictors of height of the Pokémon species, using
    - a linear regression model;
    - explanatory variables: weight, evolution stage, base stats (total or using the 6 separate stats?), sex percentage, capture rate and happiness;
    - possible additional explanatory variables: Pokémon type (first and second type, or first only?), possibly considering interactions with weight; mythical and/or legendary status of the Pokémon species.
    - (possibly) flexible modeling of non-linear associations between the outcome and the explanatory variables, using restricted cubic splines or transforming the variables.
- The domain expertise suggests that
  - evolution stage is a key predictor of "strength" (not evolving should be treated separately from base)
  - generations might differ systematically

Some examples: IDA for Pokémon

IDA plan: **missing values**
Proportion (M1): describe the number and proportion of missing values

IDA result:

Variable	Missing (count)	Missing (%)
type2	488	48.41
percentage_male	145	14.38
PokemonPhase	0	0.00
generation	0	0.00

Interpretation: 1 out of 7 Pokémon species has missing value in percentage of male sex.

It is missing because the sex is unknown for those species (neither male or female)

Consequence: might not be treated as missing value, might be as a category by itself. See more later

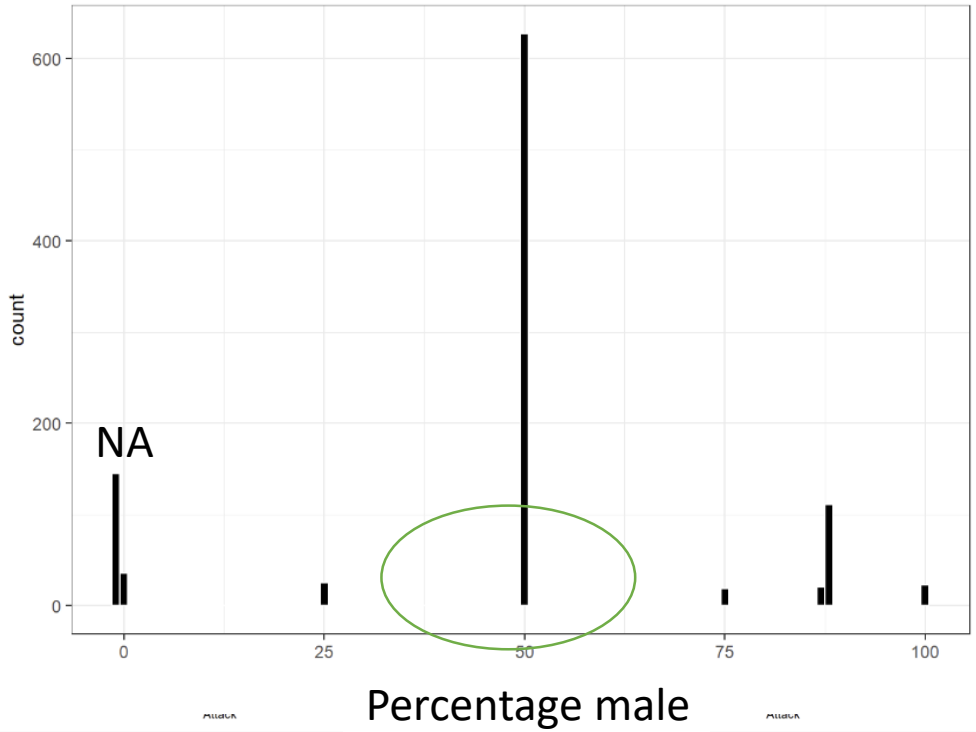
Some examples: IDA for Pokémon

IDA plan: **univariate descriptions**

Summary statistics for categorical and numerical predictors and outcome, overall and stratified by generation (U1/U2)

Sparsity (UE1): plots to identify intervals with

High-resolution histograms - details



Interpretation: highly positively asymmetric distribution of the outcome (and other variables);

extreme values (in outcome and other variables)

Terminal digit preference

Numerical variables with few distinct values

Consequence: It is not sensible to assume a gaussian distribution, log transformation might be advisable, especially for the outcome variable of weight, and height.

There might be influential points, careful interpretation of the associations in the extremes

Categorization of some variables would not remove too much information.

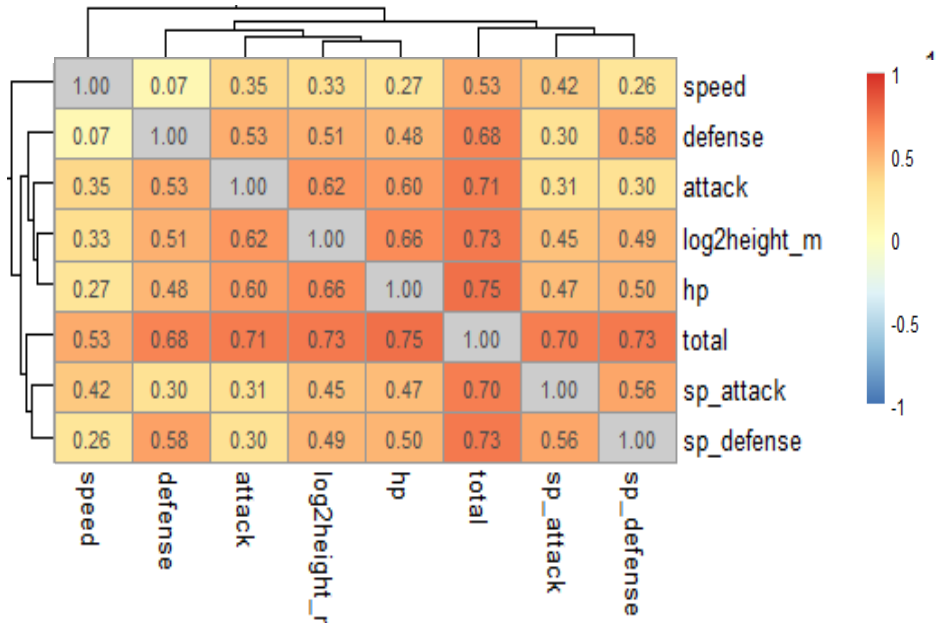
Some examples: IDA for Pokémon

IDA plan: **multivariate descriptions**

Correlation (V2): describe the correlation between explanatory variables

Background knowledge: Total = sum of 6 base stats (HP, defense, attack, speed, speed_defense, speed_attack)

IDA result:



Spearman's correlations

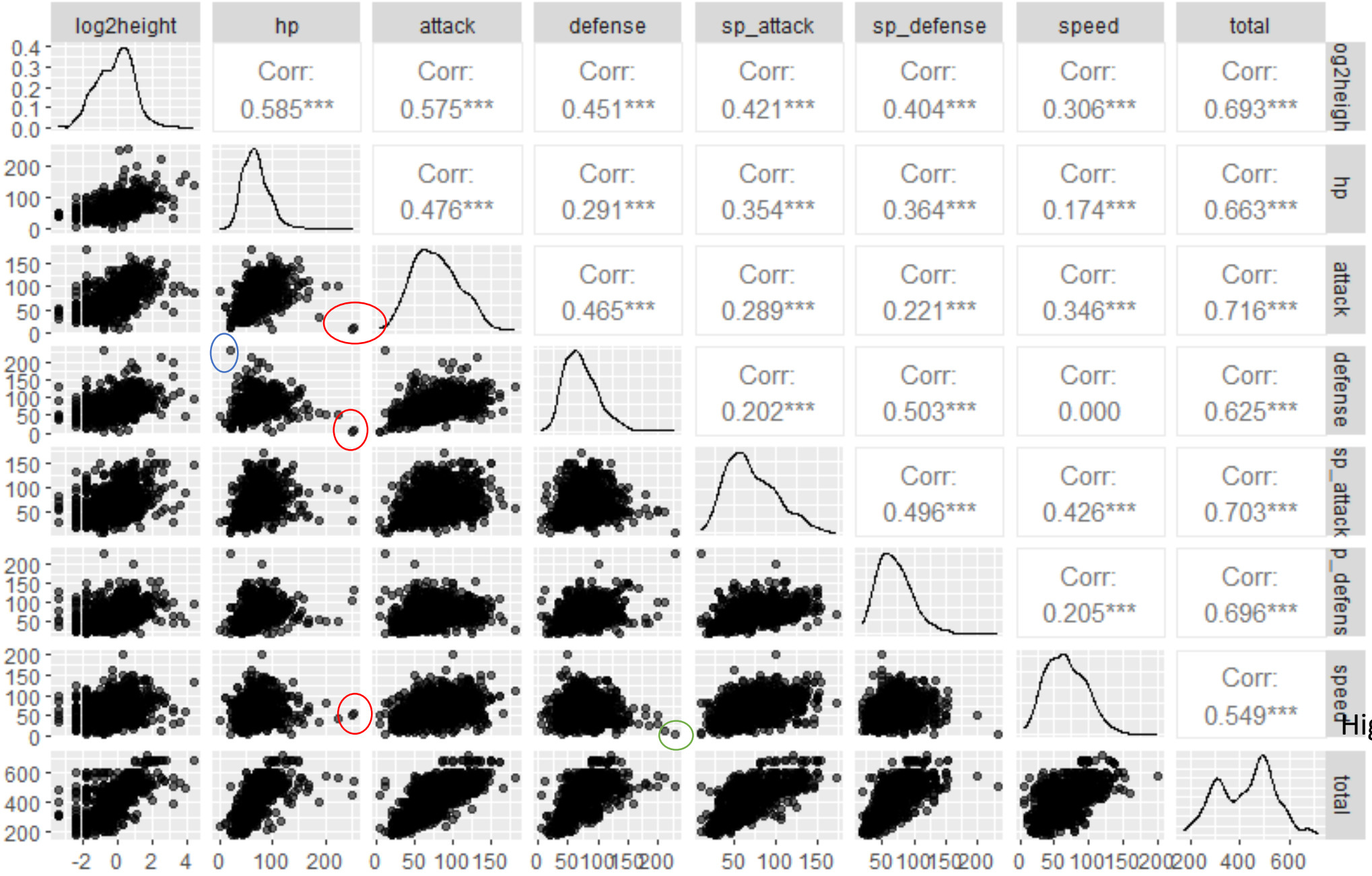
Interpretation: Height is positively correlated to the base stats, base stats are positively correlated.

Speed is the base stat with smallest correlation to total; the other base stats are strongly positively correlated to total (r>0.65)

Consequence:

The use of the 6 base stats might provide additional information compared to total

Large correlations between predictors can inflate standard errors (CI)



Identification of multivariate "outliers"

High HP, low attack, speed, defense

Low HP, high defense

.....

High speed, low defense

Consequence: Can have disproportional impact on results

Pearson's correlation, sometimes lower than Spearman's

References of avaialbe resources

- **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
- **Ten simple rules for IDA:** Baillie M., le Cessie S., Schmidt CO, Lusa L, Huebner M for the Topic Group "Initial Data Analysis" of the STRATOS Initiative, PLOS Computational Biology, 2002 https://doi.org/10.1371/journal.pcbi.1009819
- **Website with IDA report and R code (cross-sectional):** https://stratosida.github.io/regression-regrets/
- **Website with IDA report and R code (longitudinal):** https://github.com/stratosida/longitudinal-report
- **TG3 website**: https://www.stratosida.org
- **Pokémon IDA report** (work in progress): https://www.dropbox.com/scl/fi/niu294tishb74yt1a2sd8/PokemonReport_v5.html?rlkey=lccr2eriyzusgl96qfwnhlifl&dl=0

- What are the next steps?
  - Tutorial papers
  - Reccomandations on how to integrate the IDA data plan in the (statistical) analysis plan

How to efficiently perform IDA? Focus on SAP

Prerequisites

- Study design
- Definition of the research question
- Background knowledge
- Statistical analysis plan
- Preparation of metadata
- Data collection

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)

- Describes (Thomas and Peterson, JAMA 2012)
  - sufficient details to replicate the analysis by independent statisticians
  - the planned study objectives (descriptive and testable)
  - the target population
  - what variables and outcomes will be collected (and possible transformations)
  - which statistical methods will be used to analyze them and how to handle
    - missing data, correlated data, bias, confounding, subgroups, interactions, sensitivity analyses
- "The SAP is to be applied to a clean or validated data set for analysis" (ICH E9)

How to efficiently perform IDA? Focus on SAP

Prerequisites

- Study design
- Definition of the research question
- Background knowledge
- Statistical analysis plan
- Preparation of metadata
- Data collection

- Guidance on how to write SAPs is available also for observational studies
  - DEBATE (Hiemstra et al., 2019)
    - Extending the guidelines from Gamble et al. (2017): Guidelines for the Content of Statistical Analysis Plans in Clinical Trials
  - Stevens et al. (2023): A template for the authoring of statistical analysis plans
  - Yuan et al. (2018): Guide to the statistical analysis plan
  - Thomas and Peterson (2012): The value of statistical analysis plans in observational research

- IDA only partially addressed, with generic statements (DEBATE)
  - Missing data (reporting, assumptions, how to handle)
  - Baseline characteristics (methods to summarize)
  - Time points at which the outcomes are measured
  - Loss to follow-up (timing, reasons, presentation)

- IDA plan can be incorporated in the SAP

How to efficiently perform IDA? Focus on SAP

Prerequisites

- Study design
- Definition of the research question
- Background knowledge
- Statistical analysis plan
- Preparation of metadata
- Data collection

- Results from IDA might indicate the need for a revision of the SAP
  - "While the SAP should be finalized prior to data analysis, authors may make changes to the analytic plan in response to subsequent findings " (TP 2012, focused on SAP for observational studies) – + changes transparently reported
  - "Some revisions to SAP are nearly always necessary; unforeseen issues with the data may indicate alternative statistical methods or unexpected results may require new analysis. Both the SAP and report are revised to reflect changes" (TP 2012)
  - "However, analyses occasionally deviate from those specified. The authors may have discovered an error in their statistical analysis plan, identified something unexpected in the data, found a better way of conducting an analysis, or perhaps learnt something from the data that was worth exploring. In these cases, *The BMJ* requires authors to document any changes clearly and provide a sound rationale for doing so, to ensure full transparency." Isalam et al., BMJ, 2022
- The use of IDA is suggested to handle missing data: (1) devise an analysis plan, (2) examine data for appropriateness, (3) report results (TARMOS, Lee 2021)

- IDA plan can be incorporated in the SAP

IDA plans can be incorporated in SAP

- Nilufer Nourouzpour et. al. Association of anesthesia technique with morbidity and mortality in patients with COVID 19 and surgery for hip fracture: a retrospective population cohort study. Source: ClinicalTrial.gov.
- Missing values – in line with TARMOS guidelines
  - Patients with missing data on key variables will be excluded as described in the inclusion/exclusion criteria.
  - For confounders in modeling if
    1. >10% missing then exclude the confounder from the model
    2. <1% missing then delete case (complete case analysis)
    3. >=1% and <10% missing then multiple imputation
  - The % of missing data for height and weight will be examined in cohort characteristics. If <1% is missing, then BMI will be calculated using these variables.
- Collinearity
  Check for collinearity with variance inflation factor and correlation matrix: if present, combine information from collinear variables if feasible (e.g. new variables is “yes” if “yes” in any of the variables); if not, eliminate the variable that has more missing values or would be less accurately ascertained

Why this helps?

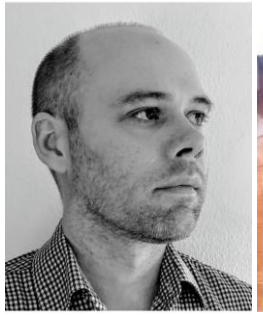
- Learning about data properties (expected or unexpected) before addressing the research question (with modeling) is essential
- Our checklists and tutorials can improve
  - efficiency
  - transparency
  - reproducibility

  Further guidance and examples will hopefully make a more systematic use of IDA more common

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)




- Regression project
  - Georg Heinze (Vienna, Austria)

- Longitudinal data project
  - Cecila Proust Lima (France, TG4), Kate Lee (Australia, TG1)

- Pokémon project
  - Sara Manski (Michigan, USA)

STRATOS INITIATIVE

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

https://www.stratosida.org/

Abstract

Initial data analysis (IDA) is the part of the data pipeline that takes place between the end of data retrieval and the beginning of data analysis that addresses the research question. Systematic IDA and clear reporting of the IDA findings is an important step towards reproducible research. A general framework of IDA for observational studies includes preparation of metadata, data cleaning, data screening, possible updates of pre-planned statistical analyses, and documentation and reporting of IDA results. In this talk we present our proposals on how to efficiently embed the data screening step of IDA in the data analysis process. Our proposals are based on checklists and on reproducible examples that facilitate the planning and performance of data screening. We also discuss how embedding IDA in statistical analysis plans can enhance the quality of planning and reporting of observational studies.

Correlation within evolution chains

- Weigh values are strongly associated within evolution chains
  - Intraclass correlation coefficient = 0.73, as estimated from a mixed effect model for weight, that adjusts for the other predictors
- Profile plots: trends towards increase
- Interactive graphs are useful to visualize changes within evolution chain – see report
- The finding suggests that a mixed effect model, that takes into account the correlation within chain would be appropriate