TG2/TG3 project update: “Regression without regrets: initial data analysis is an essential prerequisite to multivariable regression”

Mark Baillie (on behalf of project team)

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

1. **Regression without regrets (one time point)**

- Leads: G. Heinze, M. Baillie, M. Huebner, a TG2-TG3 collaboration

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

- Georg Heinze, Mark Baillie, Lara Lusa, Willi Sauerbrei, Carsten Oliver Schmidt, Frank Harrell, Marianne Huebner
- on behalf of the Topic Groups “Initial Data Analysis” and “Selection of Variables and Functional Forms in Multivariable Analyses”

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

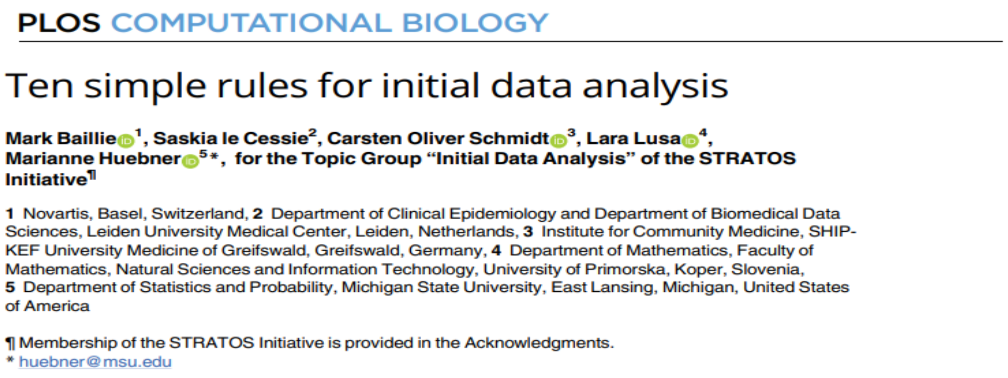
IDA is the foundation for data analysis:
assessing assumptions, analysis strategy decisions, and the interpretation and communication of results

IDA provides necessary context about data properties and structures to avoid pitfalls

- IDA aims to provide reliable knowledge about the data to ensure transparency and integrity of preconditions to conduct appropriate statistical analyses and correct interpretation of the results to answer pre-defined research questions
- IDA enables
  - research transparency and integrity
  - researchers to perform statistical analyses in a responsible manner
  - informed interpretation and communication
  - future researchers (including your future Self) to reliably reuse data and research output
- IDA provides a research team with:
  - transparent and reproducible analysis-ready data
  - reliable information about the data context and its properties
- What IDA is not:
  - IDA is not Exploratory Data Analysis
  - IDA is not an off-the-shelf cookbook

(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¶

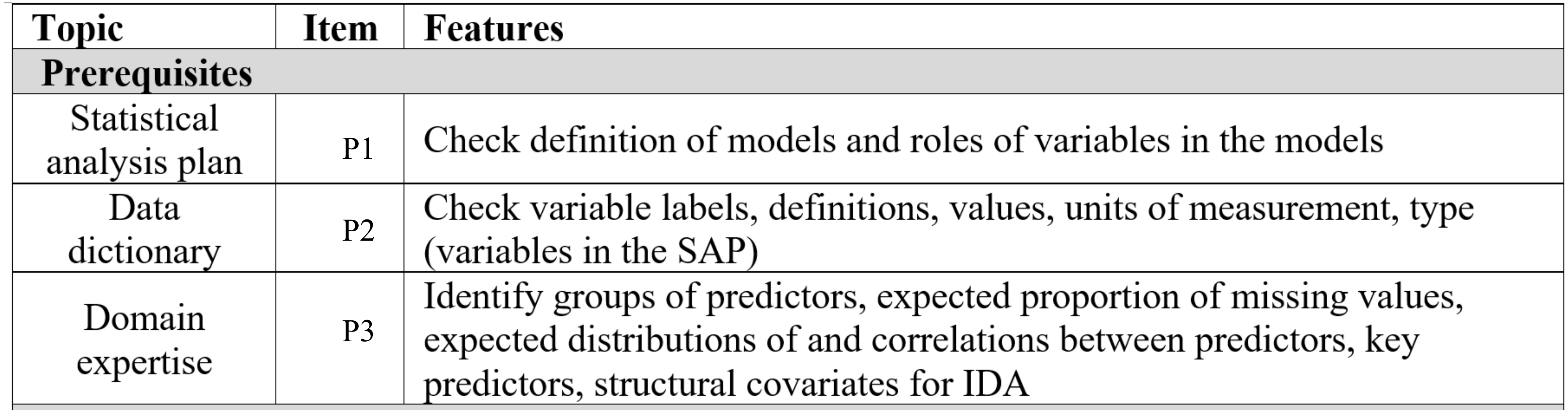
1 Novartis, Basel, Switzerland, **2** Department of Clinical Epidemiology and Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands, **3** Institute for Community Medicine, SHIP-KEF University Medicine of Greifswald, Greifswald, Germany, **4** Department of Mathematics, 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

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

Rule1 : Develop an IDA plan that supports the research objective

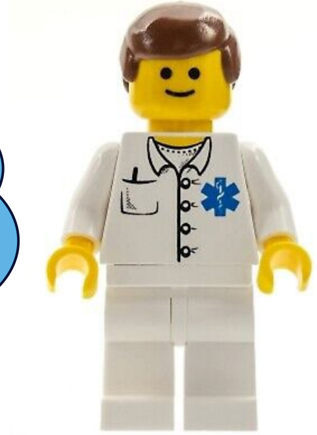
Main aim of Initial Data Analysis (IDA) = provide reliable knowledge about the data to enable responsible statistical analyses and correct interpretation of the results.



Topic	Item	Features
Prerequisites		
Statistical analysis plan	P1	Check definition of models and roles of variables in the models
Data dictionary	P2	Check variable labels, definitions, values, units of measurement, type (variables in the SAP)
Domain expertise	P3	Identify groups of predictors, expected proportion of missing values, expected distributions of and correlations between predictors, key predictors, structural covariates for IDA

How the statistical analysis plan is developed

Research aim!!



Have you got data?

Binary outcome, 50 continuous predictors, 14,691 observations

Logistic regression with fractional polynomials!

Which are your key predictors?

(searches literature) WBC, NEU, AGE, PLT, BUN, CREA

Any other important (groups of) predictors?

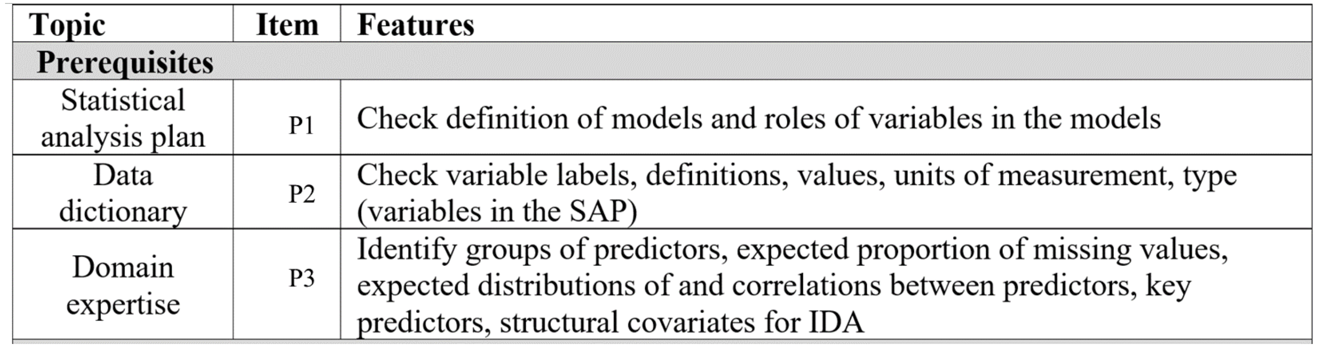
Acute phase reaction indicators
Kidney function indicators

Structural covariates?

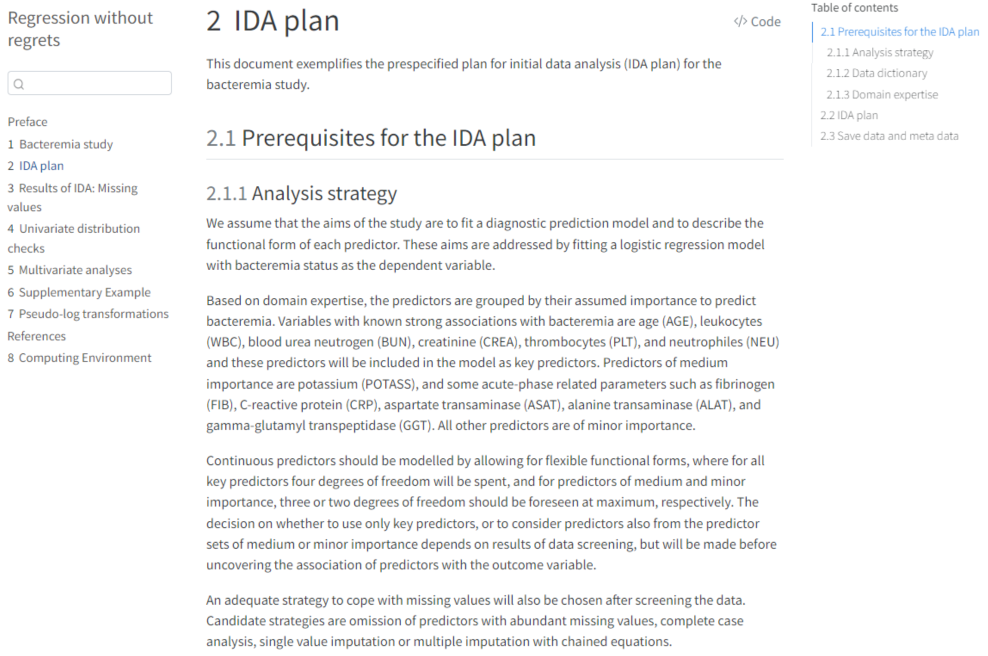
AGE, SEX, WBC

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

- (Statistical) Analysis Strategy
  - We assume that the aims of the study are to fit a diagnostic prediction model and to describe the functional form of each predictor. These aims are addressed by fitting a logistic regression model with bacteremia status as the dependent variable.



Topic	Item	Features
Prerequisites		
Statistical analysis plan	P1	Check definition of models and roles of variables in the models
Data dictionary	P2	Check variable labels, definitions, values, units of measurement, type (variables in the SAP)
Domain expertise	P3	Identify groups of predictors, expected proportion of missing values, expected distributions of and correlations between predictors, key predictors, structural covariates for IDA



2 IDA plan

This document exemplifies the prespecified plan for initial data analysis (IDA plan) for the bacteremia study.

2.1 Prerequisites for the IDA plan

2.1.1 Analysis strategy

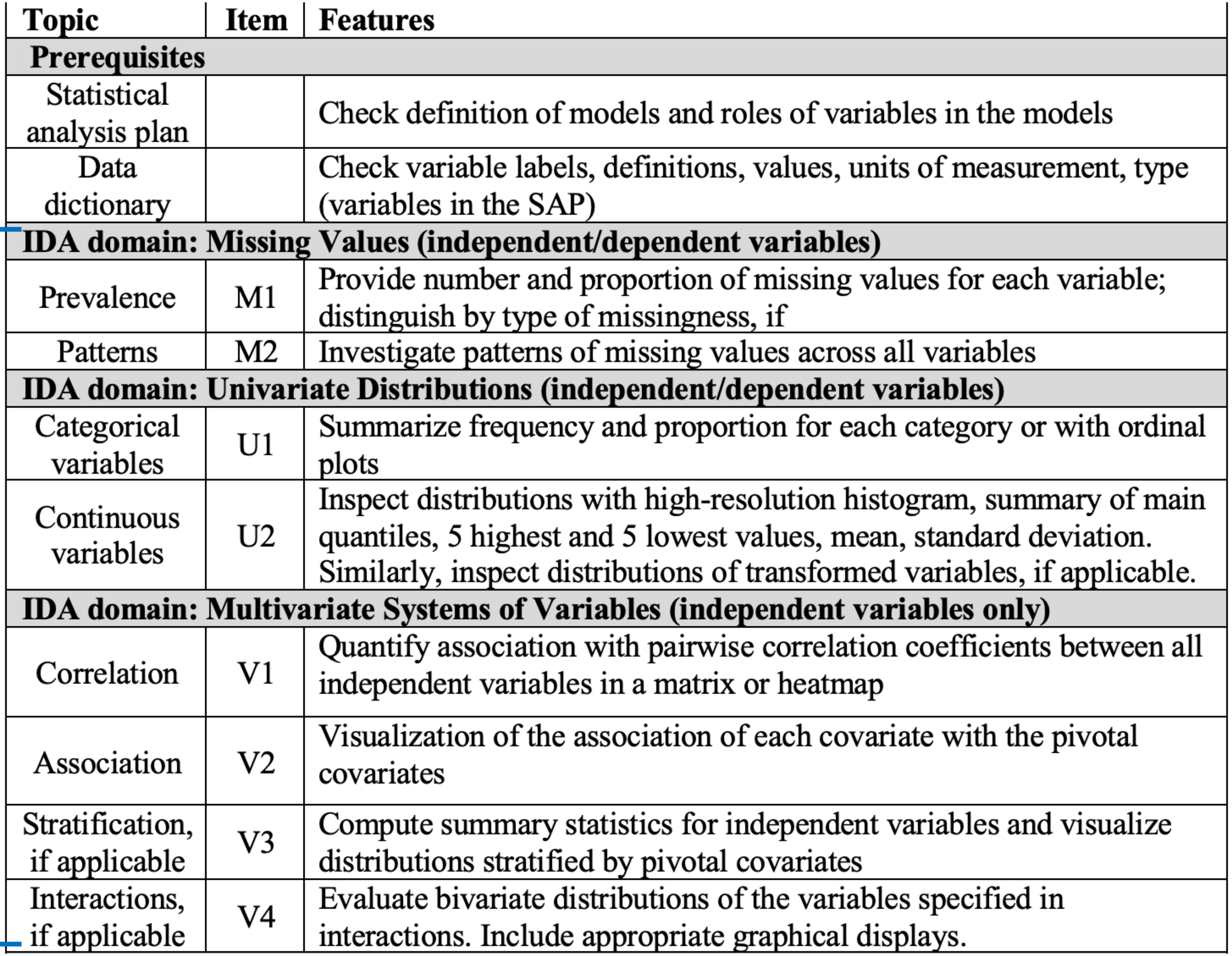
We assume that the aims of the study are to fit a diagnostic prediction model and to describe the functional form of each predictor. These aims are addressed by fitting a logistic regression model with bacteremia status as the dependent variable.

Based on domain expertise, the predictors are grouped by their assumed importance to predict bacteremia. Variables with known strong associations with bacteremia are age (AGE), leukocytes (WBC), blood urea neutrogen (BUN), creatinine (CREA), thrombocytes (PLT), and neutrophiles (NEU) and these predictors will be included in the model as key predictors. Predictors of medium importance are potassium (POTASS), and some acute-phase related parameters such as fibrinogen (FIB), C-reactive protein (CRP), aspartate transaminase (ASAT), alanine transaminase (ALAT), and gamma-glutamyl transpeptidase (GGT). All other predictors are of minor importance.

Continuous predictors should be modelled by allowing for flexible functional forms, where for all key predictors four degrees of freedom will be spent, and for predictors of medium and minor importance, three or two degrees of freedom should be foreseen at maximum, respectively. The decision on whether to use only key predictors, or to consider predictors also from the predictor sets of medium or minor importance depends on results of data screening, but will be made before uncovering the association of predictors with the outcome variable.

An adequate strategy to cope with missing values will also be chosen after screening the data. Candidate strategies are omission of predictors with abundant missing values, complete case analysis, single value imputation or multiple imputation with chained equations.

“Generic“ IDA Plan for a cross-sectional study

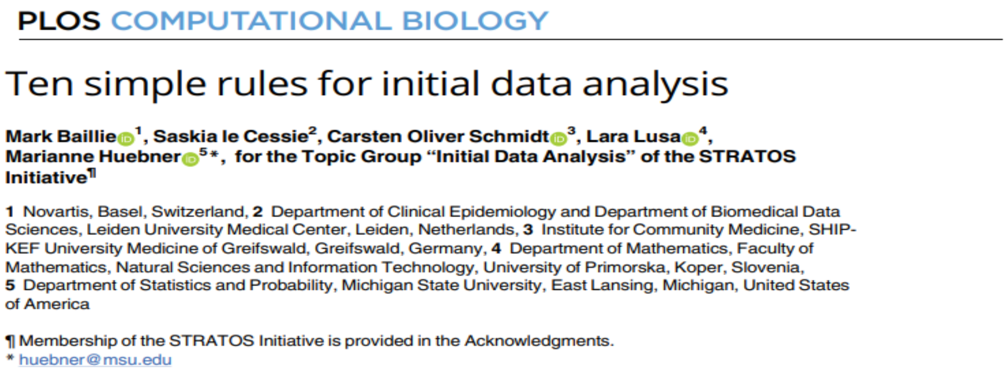


Topic	Item	Features
Prerequisites		
Statistical analysis plan		Check definition of models and roles of variables in the models
Data dictionary		Check variable labels, definitions, values, units of measurement, type (variables in the SAP)
IDA domain: Missing Values (independent/dependent variables)		
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.

Data screening

(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

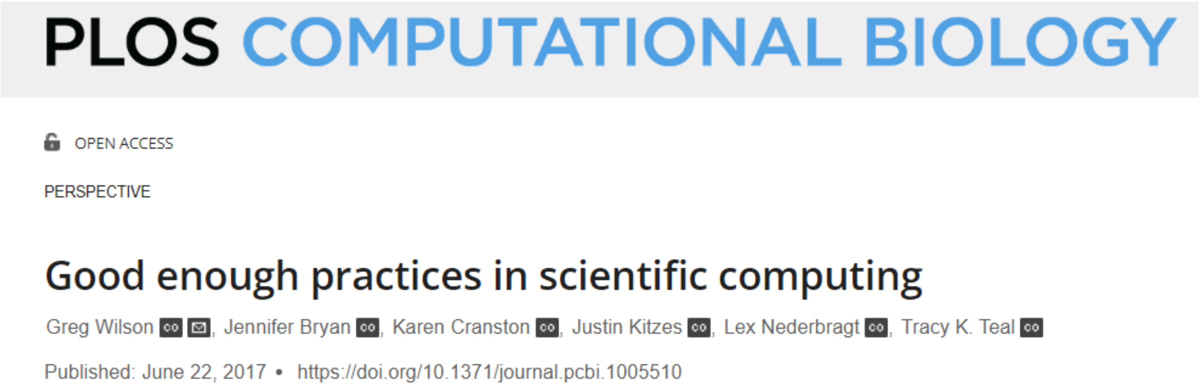
Mark Baillie[1], Saskia le Cessie[2], Carsten Oliver Schmidt[3], Lara Lusa[4], Marianne Huebner[5]*, for the Topic Group "Initial Data Analysis" of the STRATOS Initiative[¶]

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

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

Learnings so far



PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS

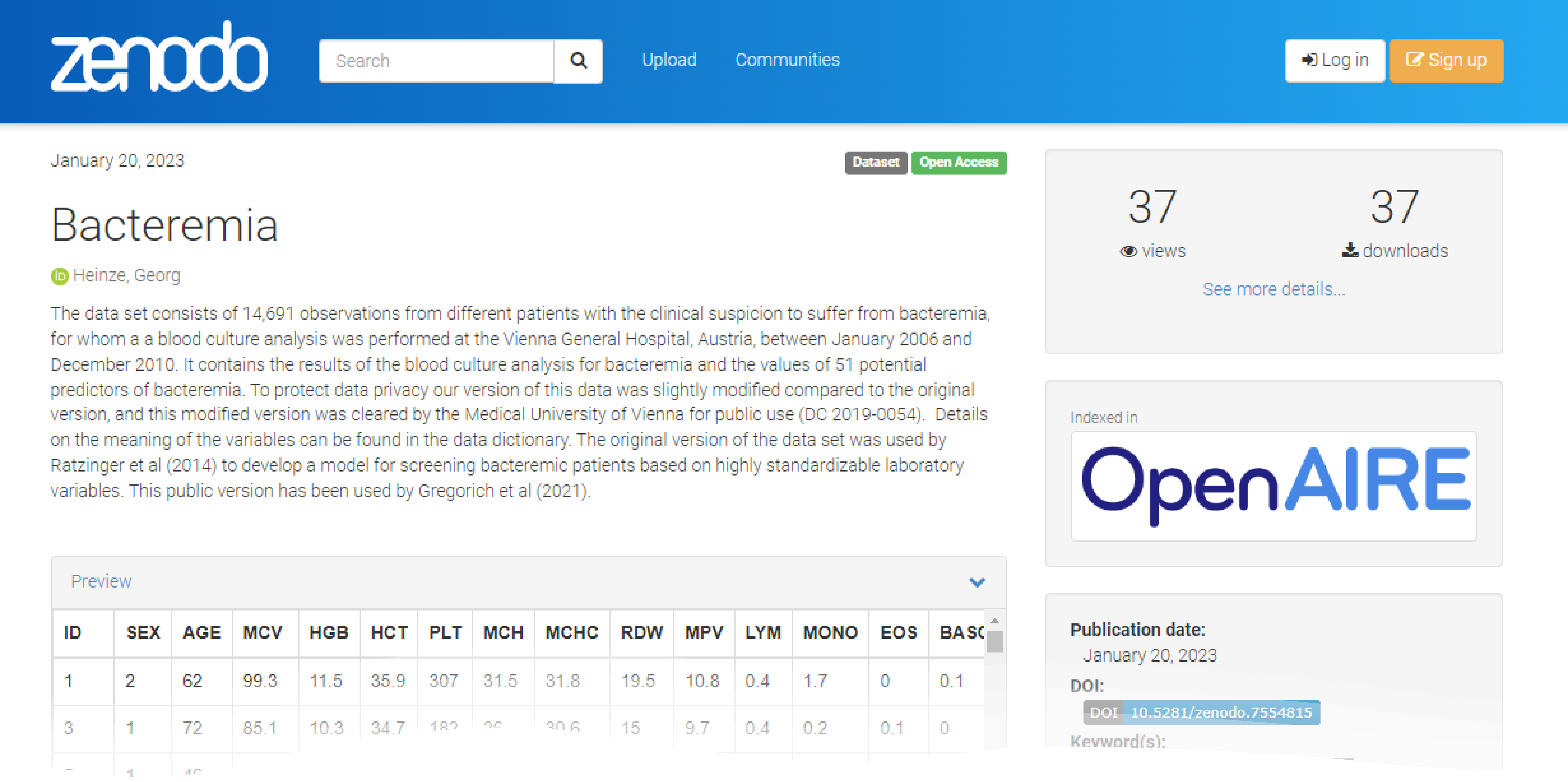
PERSPECTIVE

Good enough practices in scientific computing

Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, Tracy K. Teal

Published: June 22, 2017 • https://doi.org/10.1371/journal.pcbi.1005510

- Data management
  - How to handle development of analysis ready data
- Software
  - Choice and development of software and tools
  - Using existing R packages vs develop own code
- Collaboration
  - Communication / using version code / running on same laptops
- Project organization
  - Place all project materials in the same location i.e. github
- Keeping track of things
  - Use version control
  - Keep track of manuscript changes and how this impacts supplement
- Manuscripts
  - Word vs html



January 20, 2023

Dataset Open Access

Bacteremia

Heinze, Georg

The data set consists of 14,691 observations from different patients with the clinical suspicion to suffer from bacteremia, for whom a a blood culture analysis was performed at the Vienna General Hospital, Austria, between January 2006 and December 2010. It contains the results of the blood culture analysis for bacteremia and the values of 51 potential predictors of bacteremia. To protect data privacy our version of this data was slightly modified compared to the original version, and this modified version was cleared by the Medical University of Vienna for public use (DC 2019-0054). Details on the meaning of the variables can be found in the data dictionary. The original version of the data set was used by Ratzinger et al (2014) to develop a model for screening bacteremic patients based on highly standardizable laboratory variables. This public version has been used by Gregorich et al (2021).

37 views

37 downloads

See more details...

Indexed in

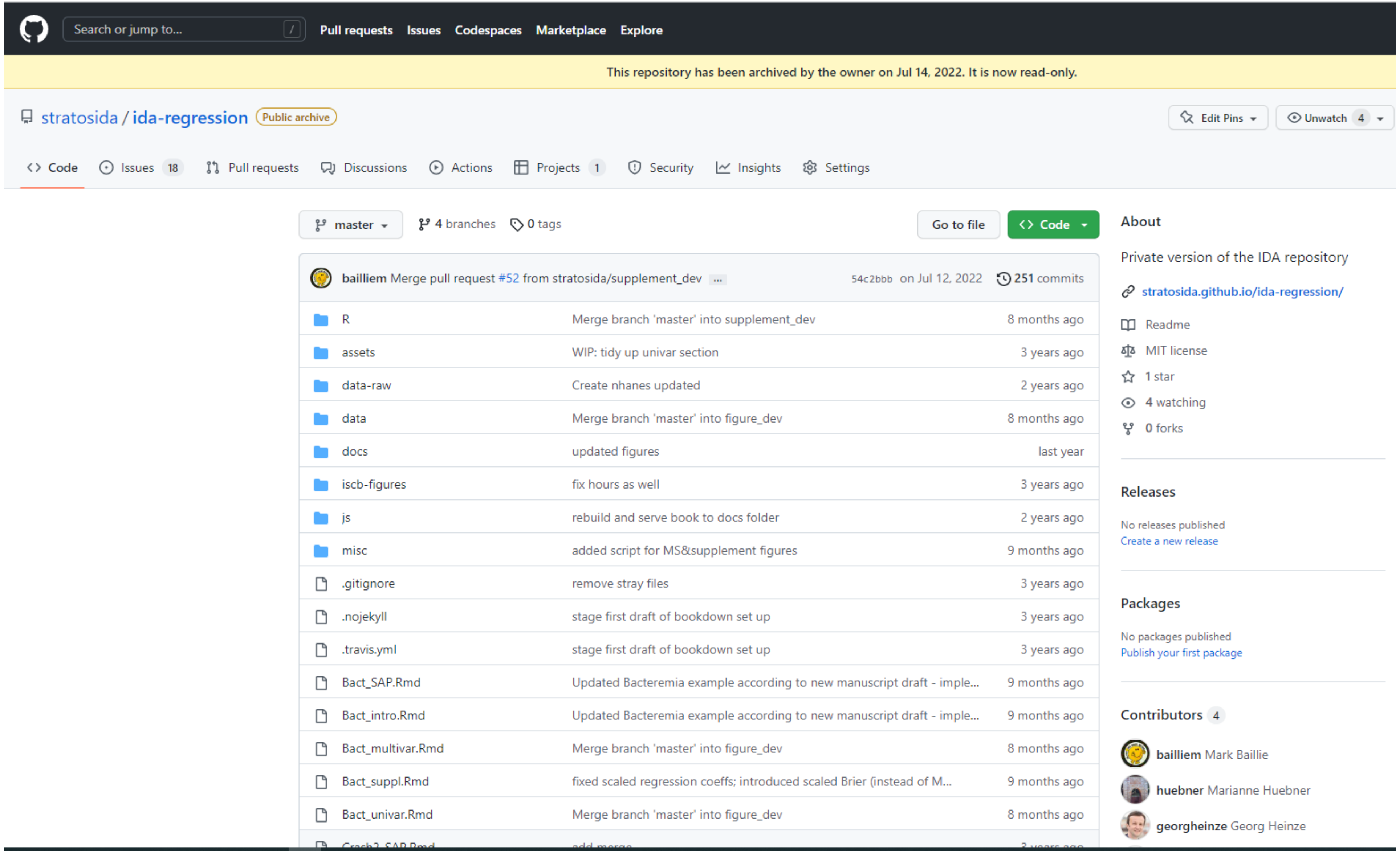
OpenAIRE

Preview

ID	SEX	AGE	MCV	HGB	HCT	PLT	MCH	MCHC	RDW	MPV	LYM	MONO	EOS	BASO
1	2	62	99.3	11.5	35.9	307	31.5	31.8	19.5	10.8	0.4	1.7	0	0.1
3	1	72	85.1	10.3	34.7	182	26	30.6	15	9.7	0.4	0.2	0.1	0
[illegible]	1	[illegible]												

Publication date:
January 20, 2023
DOI:
DOI 10.5281/zenodo.7554815
Keyword(s):

Source: https://doi.org/10.5281/zenodo.7554815



This repository has been archived by the owner on Jul 14, 2022. It is now read-only.

stratosida / ida-regression Public archive

bailliem Merge pull request #52 from stratosida/supplement_dev ... 54c2bbb on Jul 12, 2022 251 commits

R	Merge branch 'master' into supplement_dev	8 months ago
assets	WIP: tidy up univar section	3 years ago
data-raw	Create nhanes updated	2 years ago
data	Merge branch 'master' into figure_dev	8 months ago
docs	updated figures	last year
iscb-figures	fix hours as well	3 years ago
js	rebuild and serve book to docs folder	2 years ago
misc	added script for MS&supplement figures	9 months ago
.gitignore	remove stray files	3 years ago
.nojekyll	stage first draft of bookdown set up	3 years ago
.travis.yml	stage first draft of bookdown set up	3 years ago
Bact_SAP.Rmd	Updated Bacteremia example according to new manuscript draft - imple...	9 months ago
Bact_intro.Rmd	Updated Bacteremia example according to new manuscript draft - imple...	9 months ago
Bact_multivar.Rmd	Merge branch 'master' into figure_dev	8 months ago
Bact_suppl.Rmd	fixed scaled regression coeffs; introduced scaled Brier (instead of M...	9 months ago
Bact_univar.Rmd	Merge branch 'master' into figure_dev	8 months ago
Crash2_SAP.Rmd	add merge	3 years ago

About

Private version of the IDA repository

stratosida.github.io/ida-regression/

Readme

MIT license

1 star

4 watching

0 forks

Releases

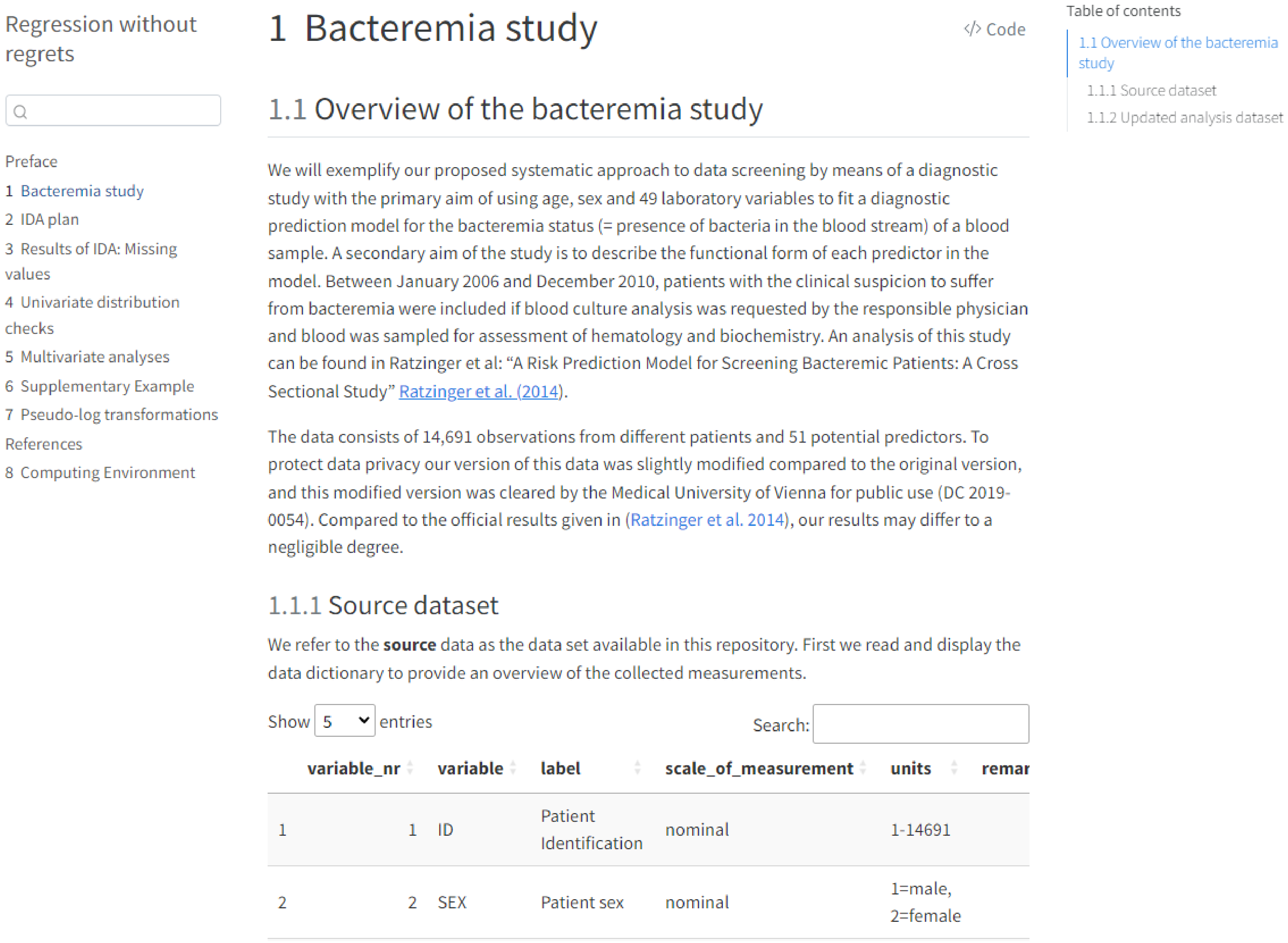
No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 4

bailliem Mark Baillie
huebner Marianne Huebner
georgheinze Georg Heinze



Regression without regrets

1 Bacteremia study

</> Code

1.1 Overview of the bacteremia study

We will exemplify our proposed systematic approach to data screening by means of a diagnostic study with the primary aim of using age, sex and 49 laboratory variables to fit a diagnostic prediction model for the bacteremia status (= presence of bacteria in the blood stream) of a blood sample. A secondary aim of the study is to describe the functional form of each predictor in the model. Between January 2006 and December 2010, patients with the clinical suspicion to suffer from bacteremia were included if blood culture analysis was requested by the responsible physician and blood was sampled for assessment of hematology and biochemistry. An analysis of this study can be found in Ratzinger et al: “A Risk Prediction Model for Screening Bacteremic Patients: A Cross Sectional Study” Ratzinger et al. (2014).

The data consists of 14,691 observations from different patients and 51 potential predictors. To protect data privacy our version of this data was slightly modified compared to the original version, and this modified version was cleared by the Medical University of Vienna for public use (DC 2019-0054). Compared to the official results given in (Ratzinger et al. 2014), our results may differ to a negligible degree.

1.1.1 Source dataset

We refer to the **source** data as the data set available in this repository. First we read and display the data dictionary to provide an overview of the collected measurements.

Show 5 entries

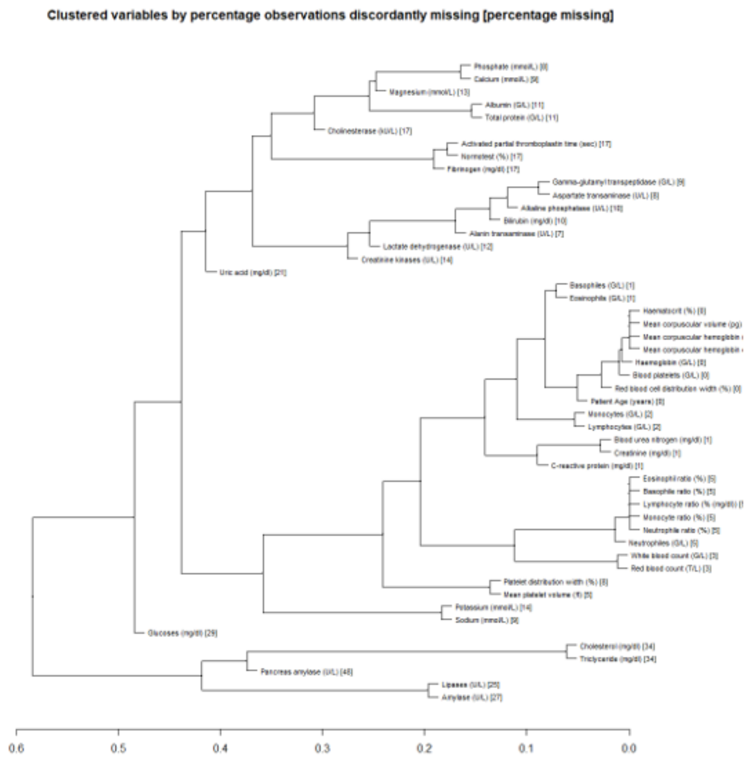
Search:

	variable_nr	variable	label	scale_of_measurement	units	remar
1	1	ID	Patient Identification	nominal	1-14691	
2	2	SEX	Patient sex	nominal	1=male, 2=female	

Table of contents

vertical axis shows the distance between two clusters, which is given by the maximum distance between any element of the first and the second clusters. For example, if two clusters are merged at a height of 25 it means that in 25% of the observations the missingness indicators of the most discordant predictors contained in the two clusters are discordant.

The numbers in brackets are the percentages of missing observations for each predictor.



Clustered variables by percentage observations discordantly missing [by variable percentage missing]

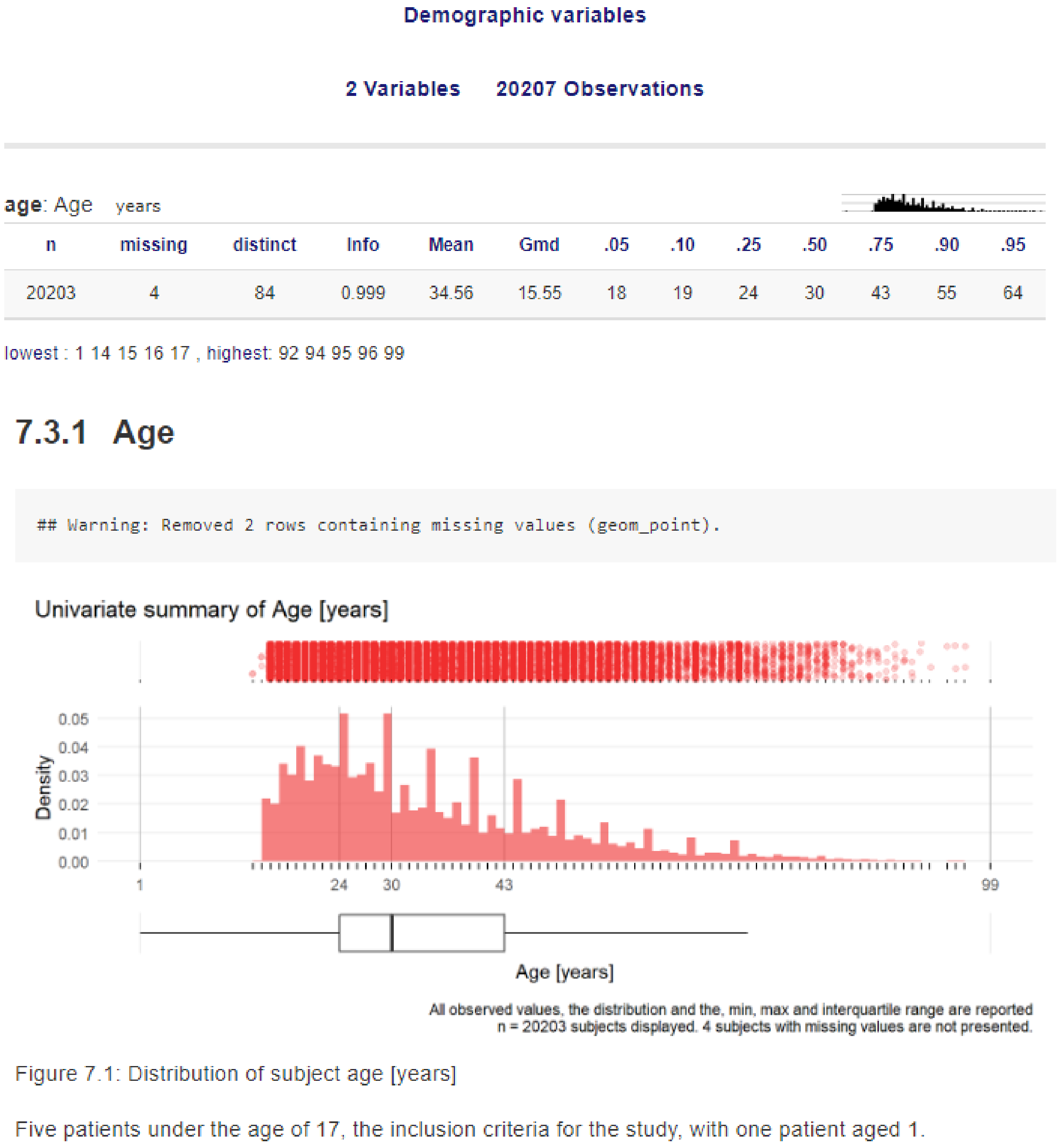
GH **Georg Heinze**
horizontal (in the flipped version)

GH **Georg Heinze**
I would suggest to use only variable names here

Source: https://stratosida.github.io/ida-regression/univariate-distribution-checks.html

Overview first →

Zoom and filter →



Demographic variables

2 Variables **20207 Observations**

age: Age years

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
20203	4	84	0.999	34.56	15.55	18	19	24	30	43	55	64

lowest : 1 14 15 16 17 , highest: 92 94 95 96 99

7.3.1 Age

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

Univariate summary of Age [years]

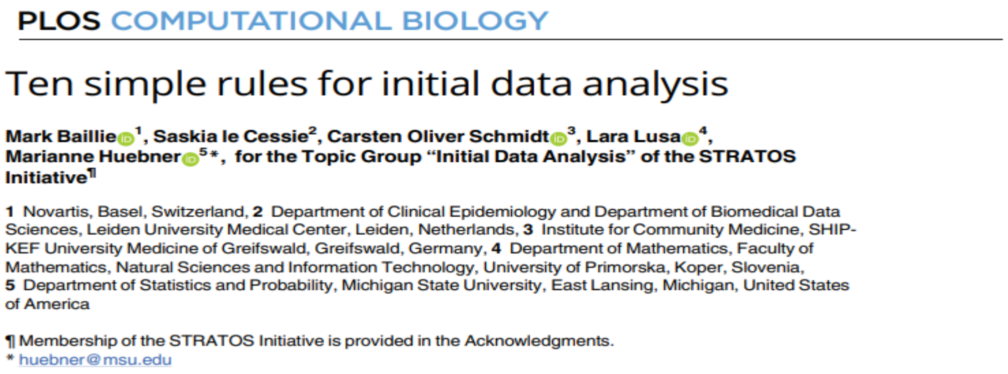
All observed values, the distribution and the, min, max and interquartile range are reported
n = 20203 subjects displayed. 4 subjects with missing values are not presented.

Figure 7.1: Distribution of subject age [years]

Five patients under the age of 17, the inclusion criteria for the study, with one patient aged 1.

(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

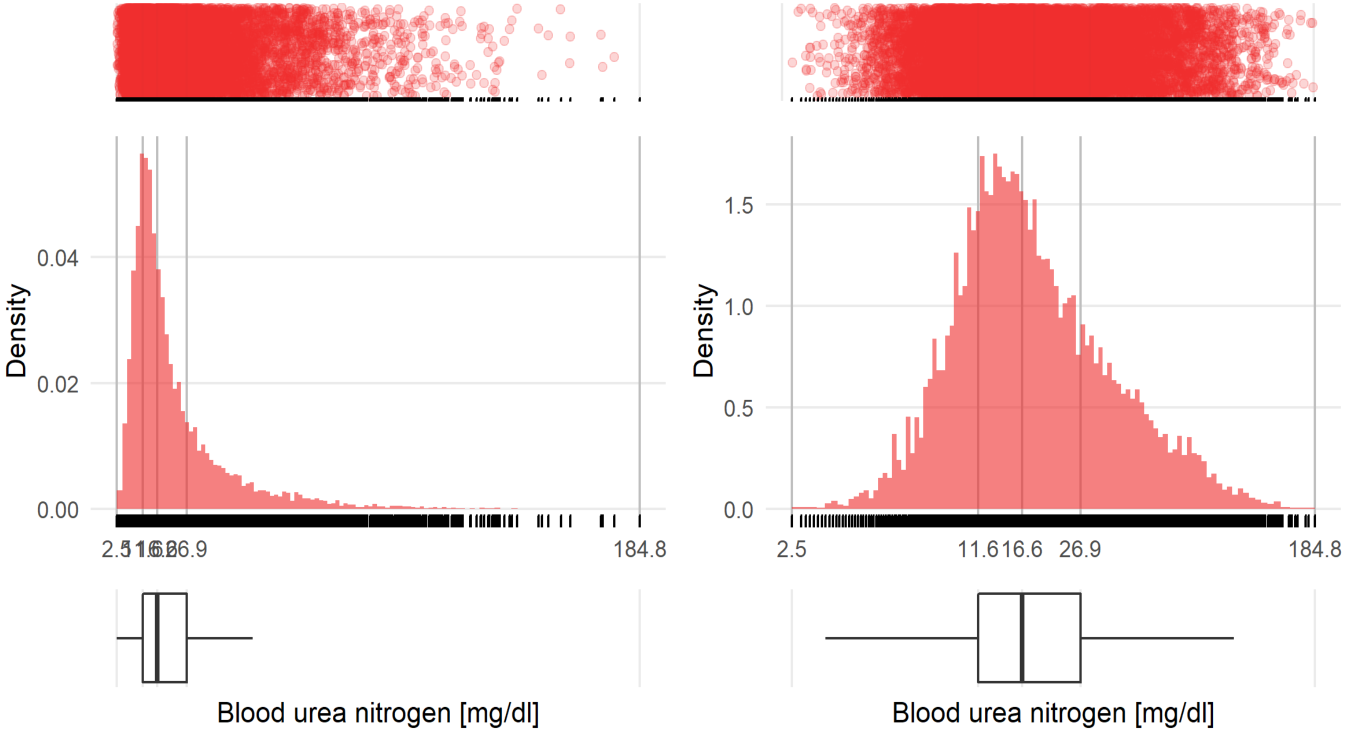
¶ Membership of the STRATOS Initiative is provided in the Acknowledgments.
* huebner@msu.edu

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

Univariate distributions

Univariate summary of Blood urea nitrogen [mg/dl]

original [left] vs. pseudo-log transformed scale [right]



All observed values, the distribution and the, min, max and interquartile range are reported
n = 14519 subjects displayed. 172 subjects with missing values are not presented. Pseudo-log transformation is suggested.

A log transformation stabilizes the distribution of this predictor

But it will change the interpretation of the betas!

Summary: a common and foundational quantitative research task

- IDA enables
  - research transparency and integrity
  - researchers to perform statistical analyses in a responsible manner
  - informed interpretation and communication
  - future researchers (including your future Self) to reliably reuse data and research output
- IDA provides a research team with:
  - transparent and reproducible analysis-ready data
  - reliable information about the data context and its properties