Regression without regrets – data screening is needed before modeling

Marianne Huebner[1], Georg Heinze[2], Mark Baillie[3]

[1]Michigan State University, East Lansing, MI, USA; STRATOS-TG3

[2]Medical University of Vienna, Vienna, Austria; STRATOS-TG2

[3]Novartis Pharma AG, Basel, Switzerland; STRATOS-VP

Anecdote

Years I ago a colleague asked me for advice about a modeling problem

They conducted a logistic regression analysis with three variables and 65 observations.

The results table looked strange.

They suspected a “hidden separation problem” and wondered if our correction to it (Heinze&Schemper, StatMed 2002) would help.

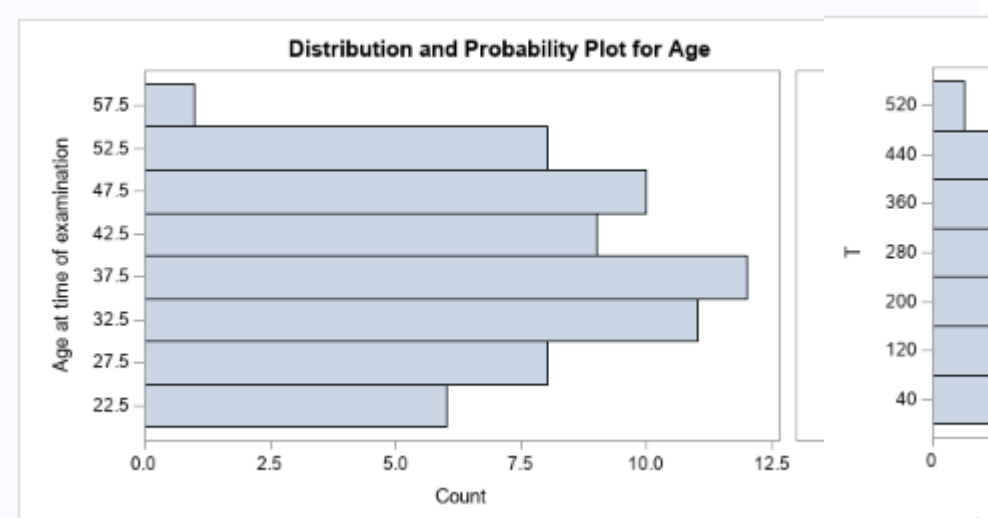
Anecdote

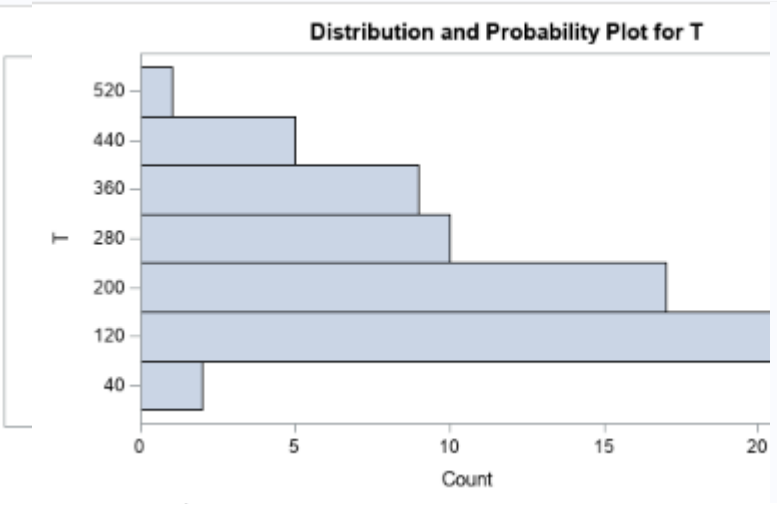
This was their model (maximum likelihood logistic regression):

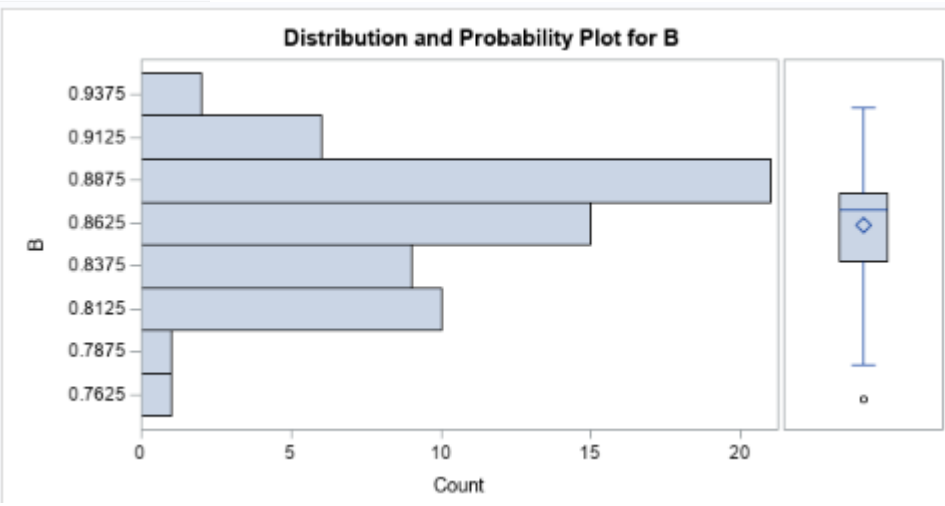
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-26.4933	10.9798	5.8221	0.0158
Age	1	0.00348	0.0353	0.0097	0.9215
T	1	-0.00721	0.00349	4.2754	0.0387
B	1	31.4099	12.2966	6.5248	0.0106

Anecdote

A closer look at the data revealed the following distributions of the independent variables:



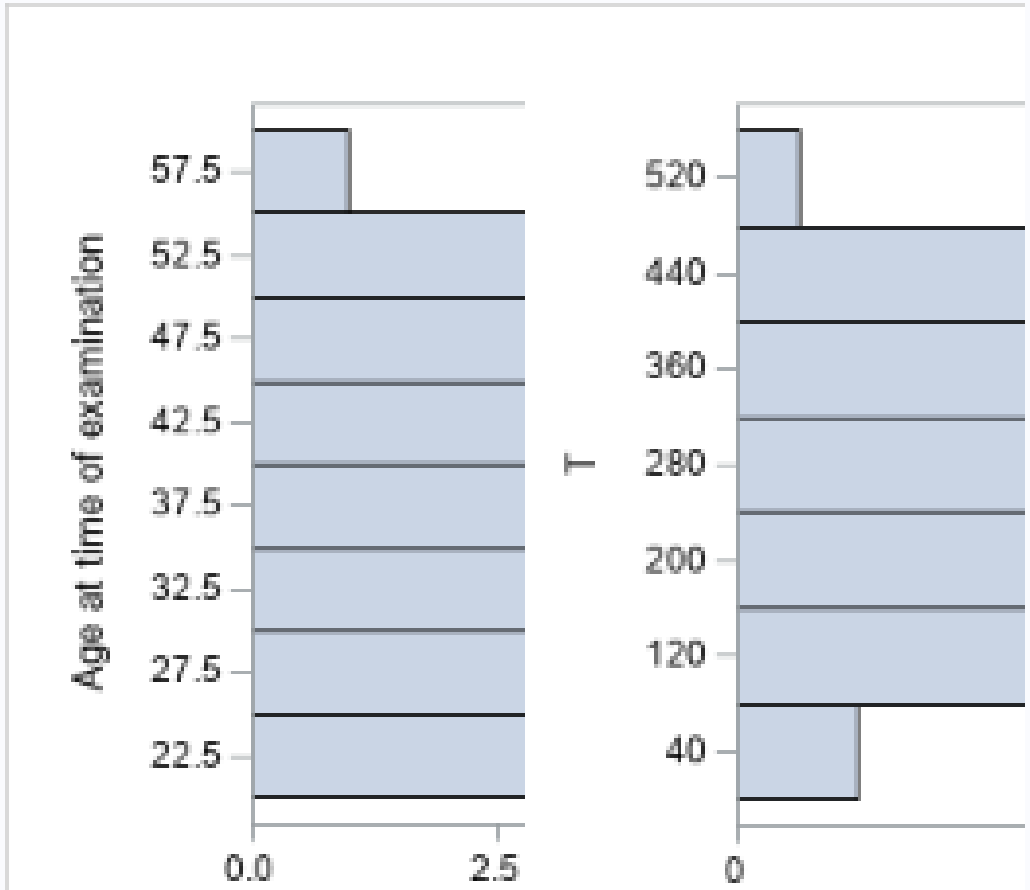


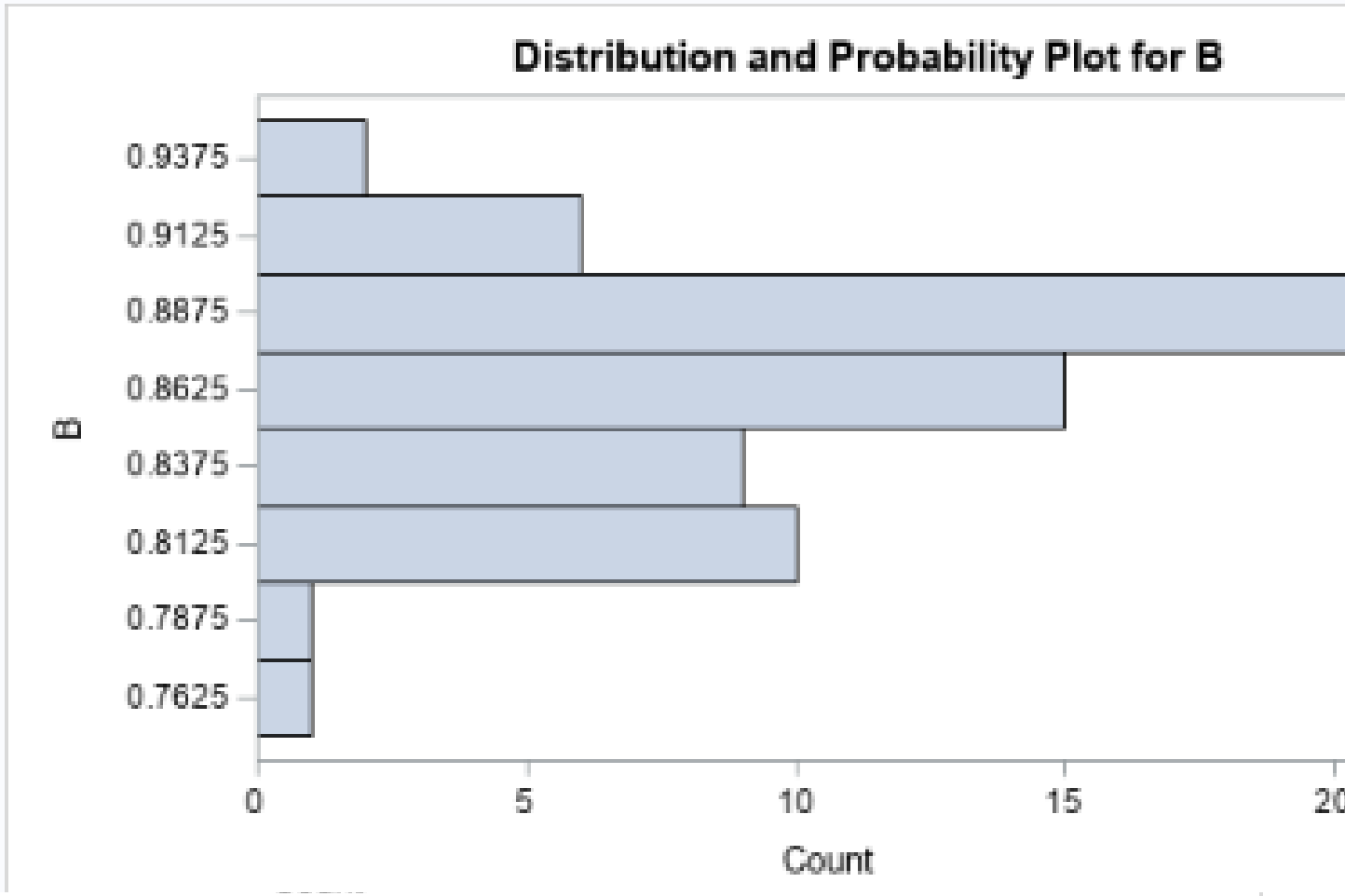


The histograms do not indicate any highly influential points

Correlation coefficients between variables were 0.30, -0.33, -0.24

An even closer look





Explanation and solving the issue

The odds ratio estimates referred to differences of 1 unit in each explanatory variable, but the ranges were very different.

A 1-unit difference means more than the data range for variable B,
but only a little fraction of the ranges for age and T.

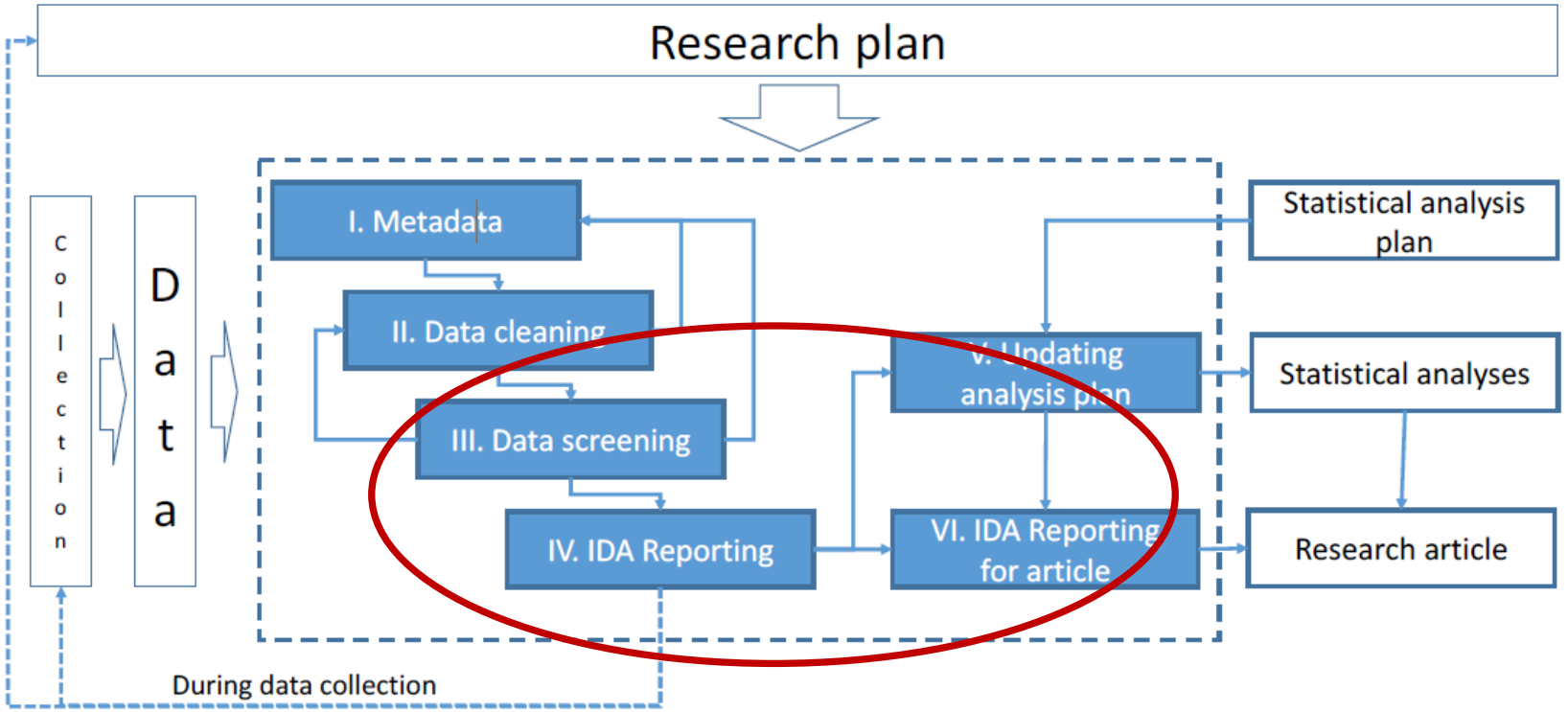
Therefore, odds ratios should be defined for meaningful differences,
e.g., for 10 years of age, 100 units of T and 0.01 units of B.

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

Effect	Unit	Estimate	95% Confidence Limits	
Age	10.0000	1.024	0.526	2.016
T	100.0	0.528	0.259	0.963
B	0.0100	1.324	1.075	1.706

What is Initial Data Analysis (IDA)

Huebner et al (2018) defined a framework of IDA consisting of six steps:



Here we concentrate on some aspects relevant for regression modeling.

Some aims of initial data analysis (IDA)

IDA	Why is this of interest?
Univariate distribution of each variable	• To describe the patient 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
Missing data	• To inform about the relevance of missing information • To decide on a proper strategy to handle missing values

Data screening

- to examine data properties while not touching the research question
- may affect the presentation and the interpretation,
- or **may lead to changes in the statistical analysis plan**

Have you reported IDA in your paper?

TG3 conducted a systematic review of the reporting practice of IDA.
BMC Med Res 2019

- IDA reporting sparse or selective
- Information on IDA can be found in all sections of a paper
- Distinctions between pre-planned and IDA-driven decisions unclear
- Incomplete reporting:
  - uncommented characteristics of participants
  - incomplete information on missingness
  - no information about associations

IDA induces changes to the analysis plan

Table 4 Number of papers with changes of the analysis plan statements by location in the paper

Reasons for change	Number of papers, n (%)	Location in Paper			
		M	R	D	S
Unexpected Values	2 (8%)	2	0	1	0
Heterogeneity	1 (4%)	0	1	0	0
Unexpected confounding	2 (8%)	1	1	2	0
Variable Distribution	4 (16%)	3	1	1	0
Other Data Properties	2 (8%)	2	0	0	0
Missing Data	5 (20%)	4	1	1	0

Abbreviations: *M* Methods, *R* Results, *D* Discussion, *S* Supplement

- Yu et al. exluded from the analyses the "participants from Zhejiang (n=56,813) where heating was rarely reported (0.6%)." [12]

1. Due to variable distributions categories of the variables were grouped, or numerical variables were categorized based on findings from IDA.
   - "Because few women were underweight (1.2%), we combined underweight with normal BMI (normal/underweight) and performed a sensitivity analysis excluding the underweight group." [27]
   - Chow et al. resolved classification problems of patients by using the category with lower value. "If insufficient information was available to distinguish between grades, the lower grade was applied." [23]
   - Gilbert et al. observed that "patients had Hospital Frailty Risk Scores ranging from 0 to 99, but this was heavily skewed to the right" and categorised it using three risk levels [17].

Example for IDA in regression modeling

Scope: continuous or binary outcome -> longitudinal, see presentation by Lara Lusa

Step 1: Specify a statistical analysis plan

Step 2: Perform IDA

Step 3: Evaluate impact of IDA on presentation, interpretation, and analysis plan

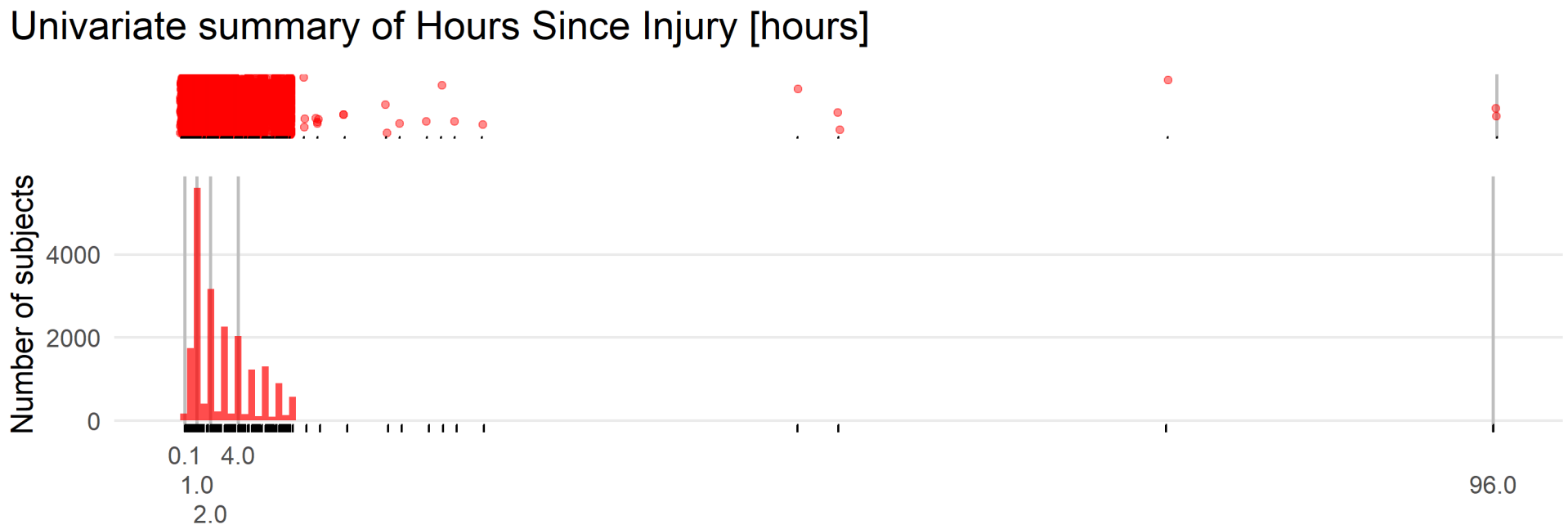
Example: Fit **a prognostic prediction model of early death** after traumatic bleeding similar to the one proposed by Perel et al (BMJ 2012) for the data **CRASH-2 (**Clinical Randomisation of an Antifibrinolytic in Significant Haemorrhage)

Outcome: Early death (binary)

Variables: age, sex, systolic blood pressure, heart rate, respiratory rate, Glasgow coma score, central capillary refill time, hours since injury, type of injury

IDA discoveries

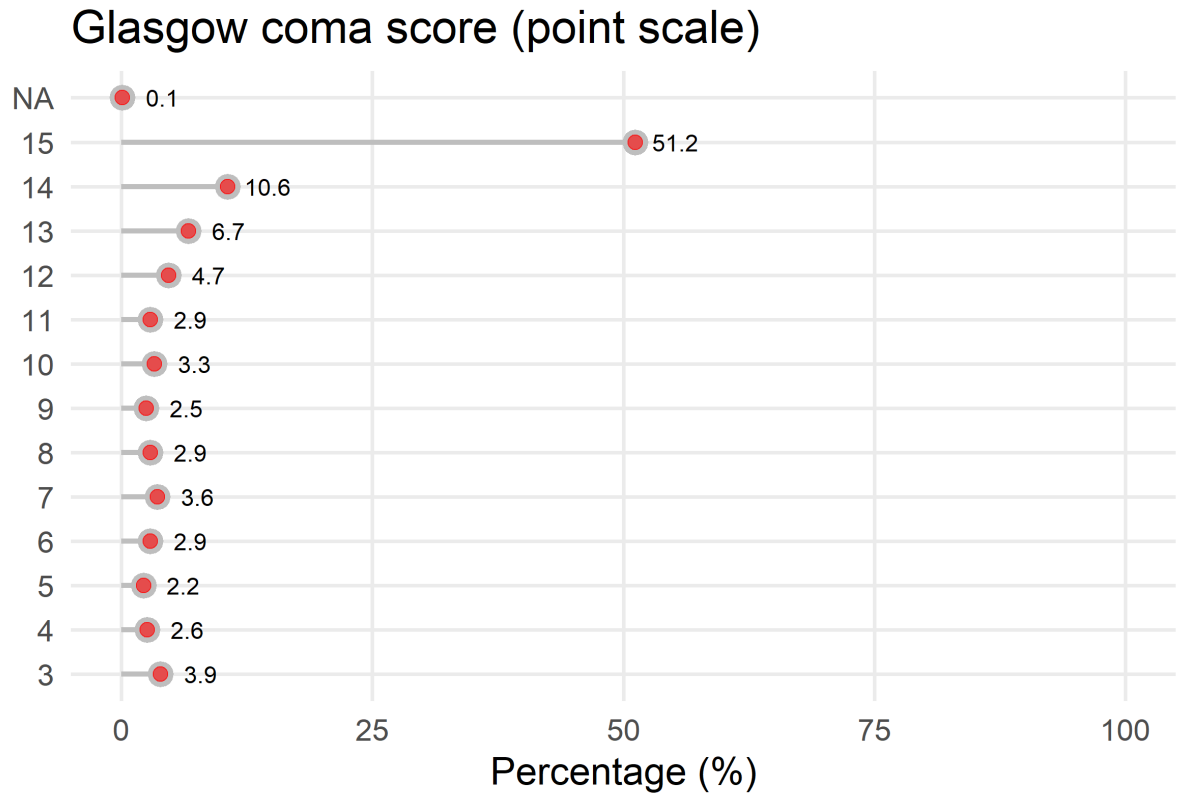
Respiratory rate, central capillary refill time, hours since injury – some highly influential points: keep, drop, winsorize?



IDA discoveries

Digit preferences: measurement error?
More frequent in cases with immediate death?

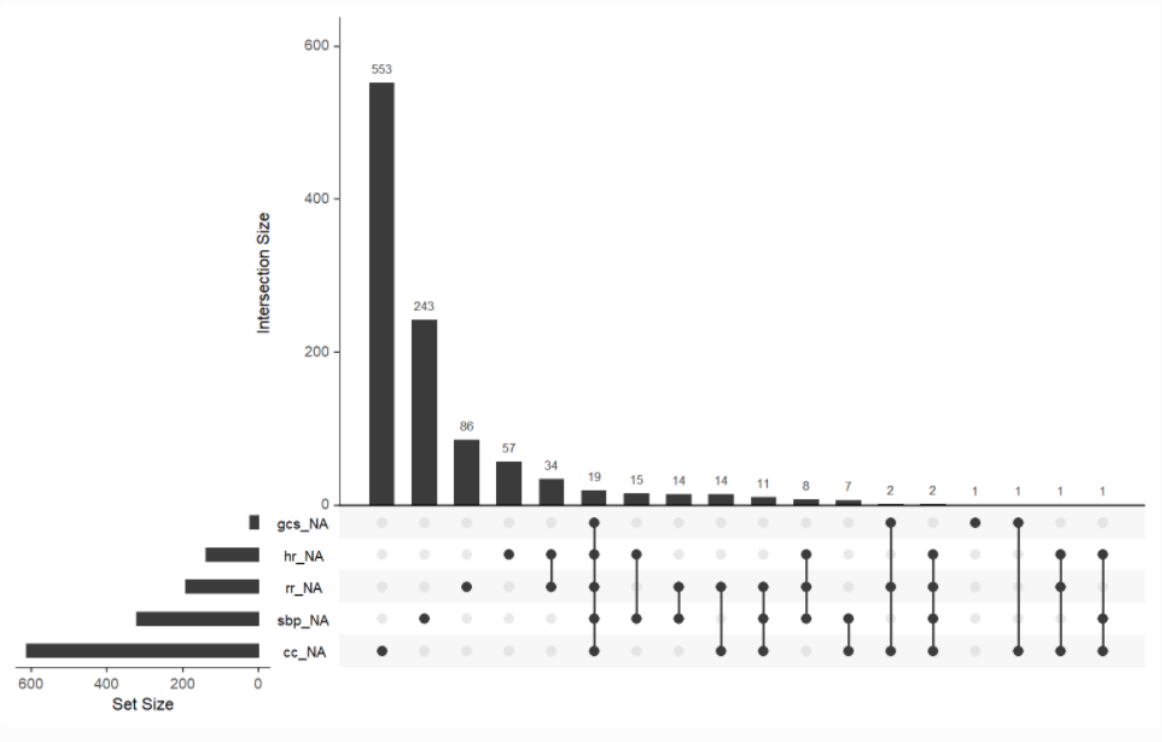
Glasgow coma scale, 15 (highest value) >50%: quadratic form?



NA = missing

IDA discoveries

- Outside inclusion criteria:
  - Age = 1 year: N=1
  - Age < 16 years: N=6
- Missing value patterns:
  - mostly independently missing values, <5%
- Number of cases with event:
  - 3076 (15%)
  - does it justify intended analysis?



IDA reporting in Perel et al. BMJ 2012; 345

- "Type of injury had three categories—penetrating, blunt, or blunt and penetrating—but we analysed it as 'penetrating' or 'blunt and penetrating.'"
- "The [Crash-2] trial included 20 127 trauma patients."
- "[Crash-2] Few data were missing for all the variables." (missingness not reported in paper) → see poster on REMARK profile by Willi Sauerbrei
- "For the validation in the TARN dataset, we did multiple imputations to substitute the missing values of the predictors"
  (missingness patterns not reported in paper)
- "We made the categories by considering clinical and statistical criteria."
  (categorization of age, Glasgow CS, SBP)

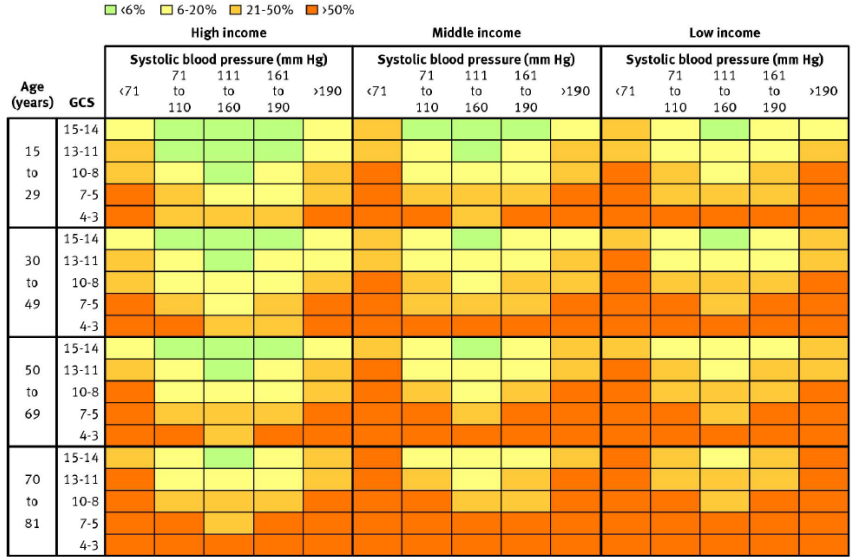


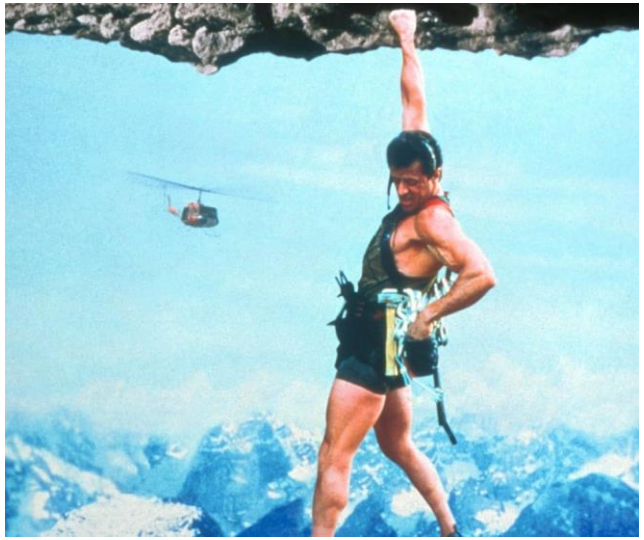
Fig 4 Chart to predict death in trauma patients. GCS=Glasgow coma score

IDA workflow

https://bailliem.github.io/ida-regression-private/

And then...

… we are ready to perform the main analysis.



S. Stallone, 1993, Cliffhanger

(As this is a talk about IDA and not about prediction modeling, it will be reported elsewhere. ☺)

In Summary

IDA is the foundation for 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

Discussion: Does the IDA principle of "Not touching the research question" hold up? (e.g not correlating outcome with independent variables)

- There is MUCH you can do without touching the research question!
- Avoid data snooping with non-transparent impact on results and conclusion
- BUT: It may be needed to for some modeling decisions.

References

- Huebner M, le Cessie S, Schmidt CO, Vach W on behalf of STRATOS-TG3. A contemporary conceptual framework for initial data analysis. Observational Studies 2018; 4: 171-192. Link
- Huebner M, Vach W, le Cessie S, Schmidt C, Lusa L on behalf of STRATOS-TG3. Hidden Analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses. BMC Med Res Meth 2020; 20:61. Link

Data set:

- Perel P, Prieto-Merino D, Shakur H, Clayton T, Lecky F, Bouamra O, Russell R, Faulkner M, Steyerberg EW, Roberts I. Predicting early death in patients with traumatic bleeding: development and validation of prognostic model. BMJ 2012; 345(aug15 1): e5166. http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/crash2.rda