Guidance for the selection of variables and functional form for continuous variables – Why and for whom?

Willi Sauerbrei

for the STRATOS Initiative

Medical Center – University of Freiburg, Germany

http://stratos-initiative.org/

Overview

- The STRATOS initiative – Why?
- Key issues of topic group 2:
  selection of variables and their functional forms
- Comparison of statistical methods: How?

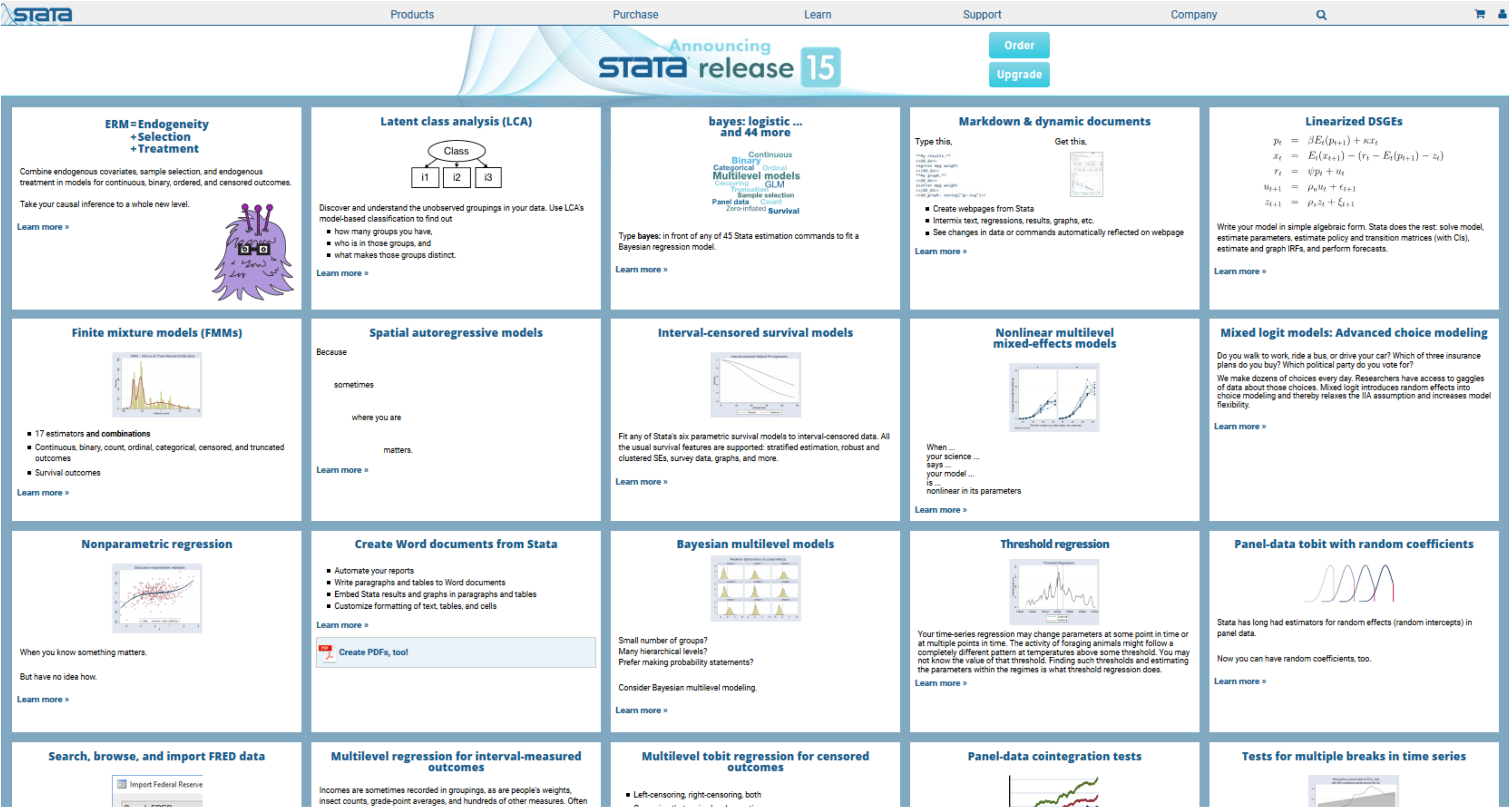
The STRATOS initiative – Why?

Current situation in statistical methodology

- Statistical methodology has seen substantial development
- Computer facilities can be viewed as the cornerstone
- Possible to assess properties and compare complex model building strategies using simulation studies
- Resampling and Bayesian methods allow investigations that were impossible two decades ago
- Wealth of new statistical software packages allows a rapid implementation and verification of new statistical ideas

Software package STATA

new procedures in 2018



Products Purchase Learn Support Company

Announcing Stata release 15

Order

Upgrade

ERM = Endogeneity + Selection + Treatment

Combine endogenous covariates, sample selection, and endogenous treatment in models for continuous, binary, ordered, and censored outcomes.

Take your causal inference to a whole new level.

Learn more »

Latent class analysis (LCA)

Discover and understand the unobserved groupings in your data. Use LCA's model-based classification to find out

- how many groups you have,
- who is in those groups, and
- what makes those groups distinct.

Learn more »

bayes: logistic ... and 44 more

Type **bayes:** in front of any of 45 Stata estimation commands to fit a Bayesian regression model.

Learn more »

Markdown & dynamic documents

Type this,

```
**My results.**
<<dd_do>>
regress mpg weight
<</dd_do>>
**My graph.**
<<dd_do>>
scatter mpg weight
<</dd_do>>
<<dd_graph: saving("gr.svg")>>
```

Get this,

- Create webpages from Stata
- Intermix text, regressions, results, graphs, etc.
- See changes in data or commands automatically reflected on webpage

Learn more »

Linearized DSGEs

$$\begin{aligned} p_t &= \beta E_t(p_{t+1}) + \kappa x_t \\ x_t &= E_t(x_{t+1}) - (r_t - E_t(p_{t+1}) - z_t) \\ r_t &= \psi p_t + u_t \\ u_{t+1} &= \rho_u u_t + \epsilon_{t+1} \\ z_{t+1} &= \rho_z z_t + \xi_{t+1} \end{aligned}$$

Write your model in simple algebraic form. Stata does the rest: solve model, estimate parameters, estimate policy and transition matrices (with CIs), estimate and graph IRFs, and perform forecasts.

Learn more »

Finite mixture models (FMMs)

- 17 estimators **and combinations**
- Continuous, binary, count, ordinal, categorical, censored, and truncated outcomes
- Survival outcomes

Learn more »

Spatial autoregressive models

Because

sometimes

where you are

matters.

Learn more »

Interval-censored survival models

Fit any of Stata's six parametric survival models to interval-censored data. All the usual survival features are supported: stratified estimation, robust and clustered SEs, survey data, graphs, and more.

Learn more »

Nonlinear multilevel mixed-effects models

When ...
your science ...
says ...
your model ...
is ...
nonlinear in its parameters

Learn more »

Mixed logit models: Advanced choice modeling

Do you walk to work, ride a bus, or drive your car? Which of three insurance plans do you buy? Which political party do you vote for?

We make dozens of choices every day. Researchers have access to gaggles of data about those choices. Mixed logit introduces random effects into choice modeling and thereby relaxes the IIA assumption and increases model flexibility.

Learn more »

Nonparametric regression

When you know something matters.

But have no idea how.

Learn more »

Create Word documents from Stata

- Automate your reports
- Write paragraphs and tables to Word documents
- Embed Stata results and graphs in paragraphs and tables
- Customize formatting of text, tables, and cells

Learn more »

Create PDFs, too!

Bayesian multilevel models

Small number of groups?
Many hierarchical levels?
Prefer making probability statements?

Consider Bayesian multilevel modeling.

Learn more »

Threshold regression

Your time-series regression may change parameters at some point in time or at multiple points in time. The activity of foraging animals might follow a completely different pattern at temperatures above some threshold. You may not know the value of that threshold. Finding such thresholds and estimating the parameters within the regimes is what threshold regression does.

Learn more »

Panel-data tobit with random coefficients

Stata has long had estimators for random effects (random intercepts) in panel data.

Now you can have random coefficients, too.

Learn more »

Search, browse, and import FRED data

Import Federal Reserve

Multilevel regression for interval-measured outcomes

Incomes are sometimes recorded in groupings, as are people's weights, insect counts, grade-point averages, and hundreds of other measures. Often

Multilevel tobit regression for censored outcomes

- Left-censoring, right-censoring, both

Panel-data cointegration tests

Tests for multiple breaks in time series

Splines

a brief overview of regression packages in R

Package	Downloads	Vignette	Book	Website	Datasets
quantreg	2001231	X	X		7
mgcv	1438166	X	X		2
survival	1229305	X	X		33
VGAM	297308	X	X	X	50
gbm	271362			X	3
gam	168143		X	X	1
gamlss	78295	X	X	X	29

Perperoglou et al, talk at ISCB 2017, see STRATOS website

Current situation in practical analyses

- Unfortunately, many sensible i**mprovements** are **ignored**

Reasons why improved strategies are ignored

- Overwhelming concern with **theoretical aspects**
- Very **limited guidance** on key issues that are **vital in practice**, discourages analysts from utilizing more sophisticated and possibly more appropriate methods in their analyses

Statistical methodology – problems are well known

The severeness of problems is even discussed in the public press:

The Economist ‘Unreliable research: Trouble at the lab.’ (October 2013):

“Scientists’ grasp of statistics has not kept pace with the development of complex mathematical techniques for crunching data. Some scientists use inappropriate techniques because those are the ones they feel comfortable with; others latch on to new ones without understanding their subtleties. Some just rely on the methods built into their software, even if they don’t understand them.”

The Lancet Research:
Increasing Value, Reducing Waste Series

Comment (Introduction)

How should medical science change?

In 2009, we published a Viewpoint by Iain Chalmers and Paul Glasziou called "Avoidable waste in the production and reporting of research evidence", which made the extraordinary claim that as much as 85% of research investment was wasted.

Kleinert and Horton, 2014

"Although this vast enterprise has led to substantial health improvements, many more gains are possible if the waste and inefficiency in the ways that biomedical research is chosen, designed, done, analysed, regulated, managed, disseminated, and reported can be addressed."

Macleod et al., 2014

Better use of statistical methods

- At least two tasks are essential:
  1. **Experts** in specific methodological areas have to work towards **developing guidance**
  2. An ever-increasing need for **continuing education** at all stages of the career
- For busy applied researchers it is often difficult to follow methodological progress even in their principal application area
  - Reasons are diverse
  - Consequence is that analyses are often deficient
- **Knowledge** gained through research on statistical methodology needs to be **transferred** to the broader community
- Many **analysts** would be **grateful for** an overview on the current **state of the art** and for **practical guidance**

Aims of the initiative

- **Provide evidence supported guidance** for highly relevant issues in the design and analysis of observational studies
- As the **statistical knowledge** of the analyst **varies** substantially, guidance has to keep this background in mind. **Guidance** has to be provided **at several levels**
- For the **start** we will concentrate on **state-of-the-art** guidance and the necessary evidence
- Help to identify questions requiring much more primary research

The overarching long-term aim is to improve key parts of design and statistical analyses of observational studies in practice

Different levels of statistical knowledge

Level 1: Low statistical knowledge

- Most analyses are done by analysts at that level

Level 2: Experienced statistician

- Methodology perhaps slightly below state of the art, but doable by every experienced analyst

Level 3: Expert in a specific area

- To improve statistical models and to adapt them to complex real problems, researches develop new and more complicated approaches. Advantages and usefulness in practice need to be assessed

STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative

Willi Sauerbrei,[a*†] Michal Abrahamowicz,[b]
Douglas G. Altman,[c] Saskia le Cessie,[d] and[‡] James Carpenter[e]
on behalf of the STRATOS initiative

Statistics in Medicine 2014

Roots in **Reporting Guidelines**, co-ordinated by the **EQUATOR** network

2011	ISCB Ottawa, Epidemiology Sub-Comm.	Preliminary ideas
2012	ISCB Bergen	Discussions, SG
2013	ISCB Munich	Initiative launched
2014-16	ISCB	Invited Sessions
2016	**BIRS**	**First general meeting**
2016	IBC Victoria	Invited Session
2016	HEC Munich	Invited Session
2017	IBS-EMR Thessaloniki	Invited Session
2017	ISCB Vigo	Scientific topic
2017	CEN-ISBS Vienna	Invited Session
2017	GMDS Oldenburg	Invited Session
2018	ISCB, RSS, …	Invited Sessions
2019	**BIRS**	**Second general meeting**

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

Topic groups

Topic Group		Chairs
1	**Missing data**	James Carpenter, Kate Lee
2	**Selection of variables and functional forms in multivariable analysis**	Georg Heinze, Aris Perperoglou, Willi Sauerbrei
3	**Initial data analysis**	Marianne Huebner, Saskia le Cessie, Werner Vach
4	**Measurement error and misclassification**	Laurence Freedman, Victor Kipnis
5	**Study design**	Mitchell Gail, Suzanne Cadarette
6	**Evaluating diagnostic tests and prediction models**	Gary Collins, Carl Moons, Ewout Steyerberg
7	**Causal inference**	Els Goetghebeur, Ingeborg Waernbaum
8	**Survival analysis**	Michal Abrahamowicz, Per Kragh Andersen, Terry Therneau
9	**High-dimensional data**	Lisa McShane, Joerg Rahnenfuehrer

Cross-cutting panels

Panel		Chairs and Co-Chairs	
MP	**Membership**	Chairs:	James Carpenter, Willi Sauerbrei
PP	**Publications**	Chairs:	Bianca De Stavola, Stephen Walter
		Co-Chairs:	Mitchell Gail, Petra Macaskill
GP	**Glossary**	Chairs:	Simon Day, Marianne Huebner, Jim Slattery
WP	**Website**	Chairs:	Joerg Rahnenfuehrer, Willi Sauerbrei
RP	**Literature Review**	Chairs:	Gary Collins, Carl Moons
BP	**Bibliography**	Chairs:	to be determined
SP	**Simulation Studies**	Chairs:	Michal Abrahamowicz, Anne-Laure Boulesteix
DP	**Data Sets**	Chairs:	Saskia Le Cessie, Aris Perperoglou
TP	**Knowledge Translation**	Chair:	Suzanne Cadarette
		Co-Chair:	Catherine Quantin
CP	**Contact Organisations**	Chairs:	Willi Sauerbrei
VP	**Visualisation**	Chairs:	Mark Baillie

Necessity of STRATOS illustrated by listing key issues of TG2:

Selection of variables and their functional forms in multivariable analysis

Building multivariable regression models – some preliminaries

- Initial data analysis (TG3)
- ‚Reasonable‘ model class was chosen

. . .

Aim of a model and model complexity

Most important distinction:
„to explain or to predict“ (Shmueli, 2010)

Prediction (TG6)

Here: **TG2**
model for explanation (or descriptive modelling)

Causal inference (TG7)

TG2: Part 1 – Selection of variables

- Central issues:
  - To select or not to select (full model)?
  - Which variables to include?
- A large number of methods proposed (for many decades)
- High-dimensional data triggered the development of further proposals
- Many critical issues

(Traditional) methods for variable selection

Full model

- variance inflation in the case of multicollinearity
  - Wald-statistic

Stepwise procedures ⇒ prespecified (α_{in}, α_{out}) and
actual significance level?

- forward selection (FS)
- stepwise selection (StS)
- backward elimination (BE)

All subset selection ⇒ which criteria?

• C_p	Mallows
• AIC	Akaike Information Criterion
• BIC	Bayes Information Criterion

Bayes variable selection

MORE OR LESS COMPLEX MODELS?

Other procedures

- Bootstrap selection
- Change-in-estimate
- Variable clustering
- Incomplete principal components
- Penalized approaches (selection and shrinkage; Lasso, Garotte, SCAD, …)
    - TG 9: High-dimensional data
- Directed acyclic graph (DAG-) based selections
    - TG 7: Causal inference
-
-

Selection of variables: Statistical prerequisites

Types of models by distribution of error

Assumptions of models

Hypothesis tests: Likelihood ratio, Score, Wald

Model estimation: maximum likelihood

Likelihood and information-theoretic measures

Resampling techniques

Shrinkage

AIC and AICc

Penalized likelihood

Prior knowledge

Bias-variance tradeoff

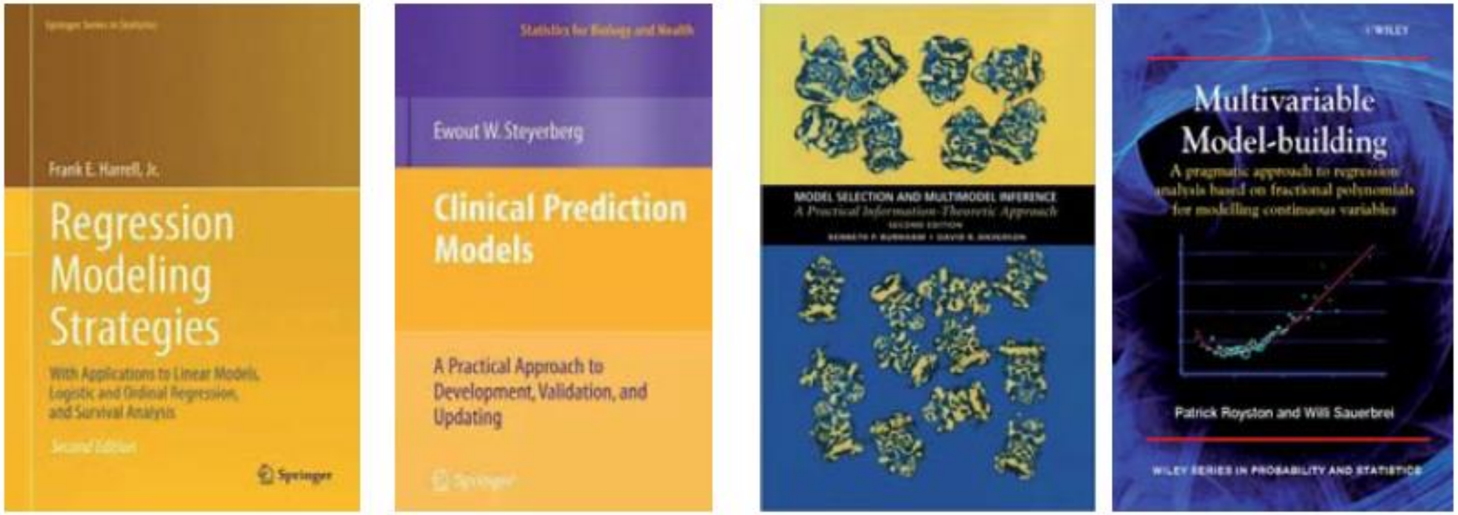
Confounding

Change-in-estimate criterion

Opinions on variable selection

for models with focus on prediction and explanation

☹ ← Variable selection → ☺


Springer Series in Statistics
Frank E. Harrell, Jr.
Regression Modeling Strategies
With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis
Second Edition
Springer
Statistics for Biology and Health
Ewout W. Steyerberg
Clinical Prediction Models
A Practical Approach to Development, Validation, and Updating
Springer
MODEL SELECTION AND MULTIMODEL INFERENCE
A Practical Information-Theoretic Approach
SECOND EDITION
KENNETH P. BURNHAM • DAVID R. ANDERSON
WILEY
Multivariable Model-building
A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables
Patrick Royston and Willi Sauerbrei
WILEY SERIES IN PROBABILITY AND STATISTICS

(Harrell, 2001; Steyerberg, 2009; Burnham & Anderson, 2002, Royston & Sauerbrei, 2008)

Heinze et al., BiomJ, 2018

→ Different philosophies

→ Emphasis on different aims

"Recommendations" from the literature

We do not know any recommendation which is supported by good evidence from theory or meaningful simulation studies

Problem of the practicing statistician:
What to do?

TG 2: Part 2 – selection of functional forms

- Assume linearity
  - Often ok but sometimes wrong. Can lead to wrong conclusions
- Cut-points
  - Many problems known for a long time. Nevertheless still very popular
- ‘Optimal’ cut-points
  - Worse than cutpoints
- Fractional polynomials and Splines
  - Flexible procedures but many open issues
  - More comparisons (simulation studies) needed

TG 2: Part 3 – Combining variable and function selection

Two inter-related questions, common to many multivariable explanatory models

Results of data-dependent selections of independent variables may depend on

- decisions regarding functional forms of both
  1. the variable of interest (X)
  2. other variables, correlated with X

and *vice versa*

For survival data (TG8):

- Effects may vary in time
- Another interrelated issue

TG 2 - State of the art?

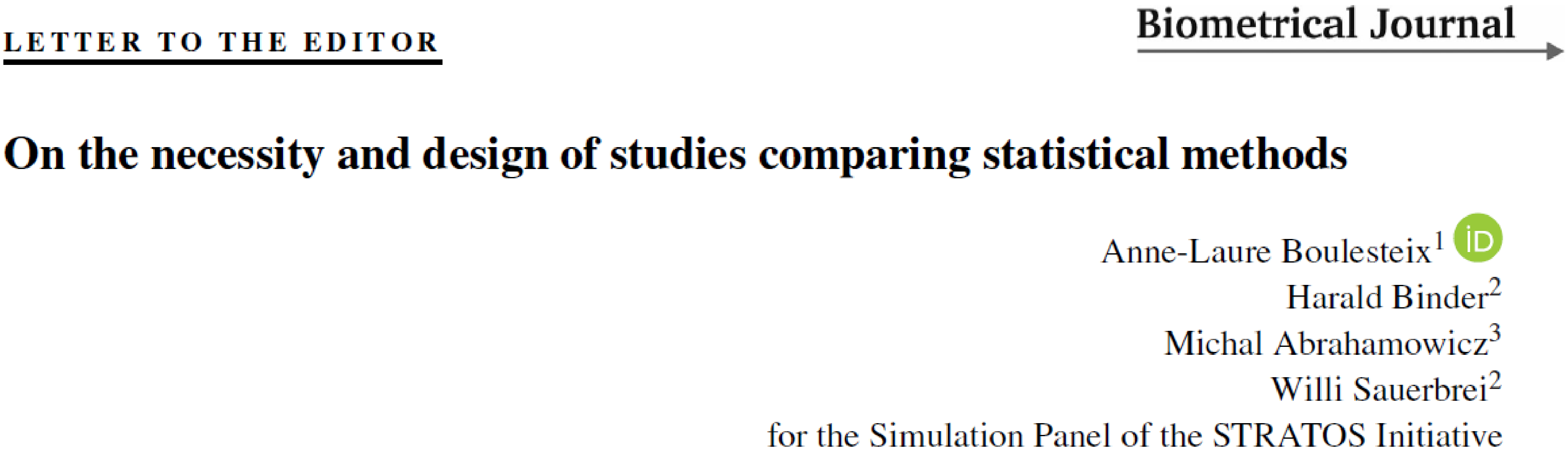
- Which **strategies for variable selection** exist?
  What about their properties?
- Data-dependent modeling introduces bias.
  What about the role of **shrinkage** approaches?
- Comparison of **spline procedures** in a univariate context.
  Which criteria are relevant? Can we derive guidance for practice?
- What about variables with a **'spike-at-zero'**?
- **Multivariable procedures**
  MFP well defined strategy
  Which of the spline based procedures?
- Multivariable procedures and **correction for selection bias**
  How relevant? One step or two step approaches?
  E.g. selection of variables and forms followed by shrinkage
- **Big Data**
  Does it influence properties of procedures and their comparison?
- Evaluation of new approaches for **post-selection inference**
- Role of **validation**

State-of-the-art - EVIDENCE is required!

Much research required!

Comparison of statistical methods

How?



LETTER TO THE EDITOR

Biometrical Journal

On the necessity and design of studies comparing statistical methods

Anne-Laure Boulesteix[1]
Harald Binder[2]
Michal Abrahamowicz[3]
Willi Sauerbrei[2]
for the Simulation Panel of the STRATOS Initiative

http://onlinelibrary.wiley.com/doi/10.1002/bimj.201700129/full

More new methods needed?

“[…]It becomes more and more difficult to get an overview of existing methods, not to mention the overview of their respective performances in different settings.

[…] Moreover, it is well known that studies comparing a suggested new method to existing methods may be (strongly) biased in favor of the new method.

***neutral* comparison studies**

- do not aim to demonstrate the superiority of a particular method
- involve authors who are, as a collective, approximately equally competent on all considered methods.
- may be very time consuming and difficult to both organize and perform“

More (meta)research needed

No consensus on what makes a reliable comparison study

- Which designs are most appropriate?
- What are typical sources of potential biases and how can they be avoided?
- How can the results be interpreted without the tendency for overinterpretation?
- Which mixture of simulated and real data should be used?
- How should real data be selected?
- How should simulated data be generated in a realistic way inspired from real datasets?

… continued

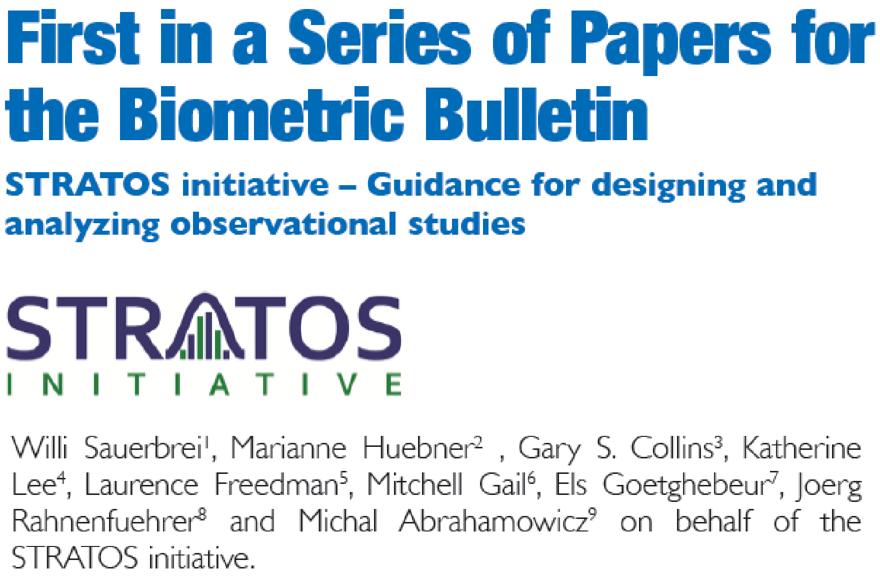
- What parameters and assumptions should be varied across the simulated scenarios?
- What range of sample sizes should be assessed?
- How can we assess the practical relevance of simulation results, which depends on the real-life plausibility of the simulation scenarios?
- How can an acceptable neutrality of the authors team be achieved and how can non-neutrality (the analogon of "conflicts of interest" in clinical research) be disclosed?
- Which "competing methods" should be considered?

We need to recognize that there is **no agreement among experts on the "state-of-the-art" methods for many topics relevant in practice**.

Guidance for whom?
Needed by many stakeholders!!

analysts with different levels of knowledge, teachers, reviewers, journalists, ……

Researchers



First in a Series of Papers for the Biometric Bulletin

STRATOS initiative – Guidance for designing and analyzing observational studies

STRATOS INITIATIVE

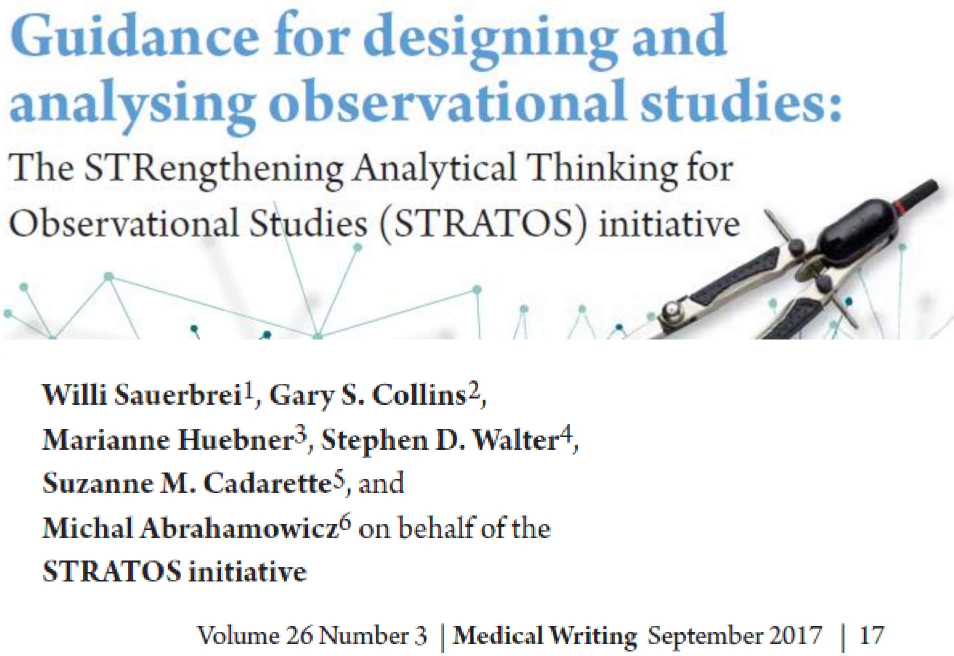
Willi Sauerbrei[1], Marianne Huebner[2] , Gary S. Collins[3], Katherine Lee[4], Laurence Freedman[5], Mitchell Gail[6], Els Goetghebeur[7], Joerg Rahnenfuehrer[8] and Michal Abrahamowicz[9] on behalf of the STRATOS initiative.

➡ Short papers from

- **TG1** – missing data
- **TG4** – measurement error and misclassification
- **TG3** – initial data analysis
- **TG2** – Variable and function selection

have appeared

Consumers



Guidance for designing and analysing observational studies:

The STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative

Willi Sauerbrei[1], **Gary S. Collins**[2], **Marianne Huebner**[3], **Stephen D. Walter**[4], **Suzanne M. Cadarette**[5], and **Michal Abrahamowicz**[6] on behalf of the **STRATOS initiative**

Volume 26 Number 3 | **Medical Writing** September 2017 | 17

Journal of the European Medical Writers Association (EMWA)

Thanks to all members of TG2 !

- Georg Heinze (Austria)
- Aris Perperoglou (U.K.)
- Willi Sauerbrei (Germany)
- Michal Abrahamowicz (Canada)
- Heiko Becher (Germany)
- Harald Binder (Germany)
- Daniela Dunkler (Austria)
- Frank Harrell (U.S.A)
- Geraldine Rauch (Germany)
- Patrick Royston (U.K.)
- Matthias Schmid (Germany)