The STRATOS initiative – towards guidance for selection of variables and their functional forms

Willi Sauerbrei[1],
Michal Abrahamowicz[2], Georg Heinze[3], Aris Perperoglou[4]
for TG2 of the STRATOS initiative

[1] Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center University of Freiburg, Germany
[2] Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada
[3] Medical University of Vienna, CeMSIIS, Section for Clinical Biometrics, Austria
[4] School of Mathematics, Statistics and Astrophysics, Newcastle University, UK

Overview

- Introduction of the STRATOS initiative
- Guidance for selection of variables and functional forms
  - 7 methodological issues identified
- Variable selection strategies
- Data dependent model-building introduces biases
  - Combine variable selection and shrinkage
- Selection of functional forms
- Conclusions

General assumption – sample size is ‘acceptable‘

PROBLEMS with Practical Applications of Statistical methods

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

"***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."

NEED for GUIDANCE

- Profusion of new, complex statistical techniques and algorithms
- Unclear which methods are useful in practice, and under what conditions?
- Insufficient awareness and understanding, among practitioners, of both well-established and, especially, new approaches and methods
- For some complex analytical challenges, there is *no* consensus, even among experts, as to the best approach
- Very **limited guidance** on key issues that are **vital in practice** discourages analysts from utilizing possibly more appropriate methods in their real-life applications, thus, reducing the scientific yield of empirical research

STRATOS Initiative: STRengthening Analytical Thinking for Observational Studies

The overarching long-term goal:

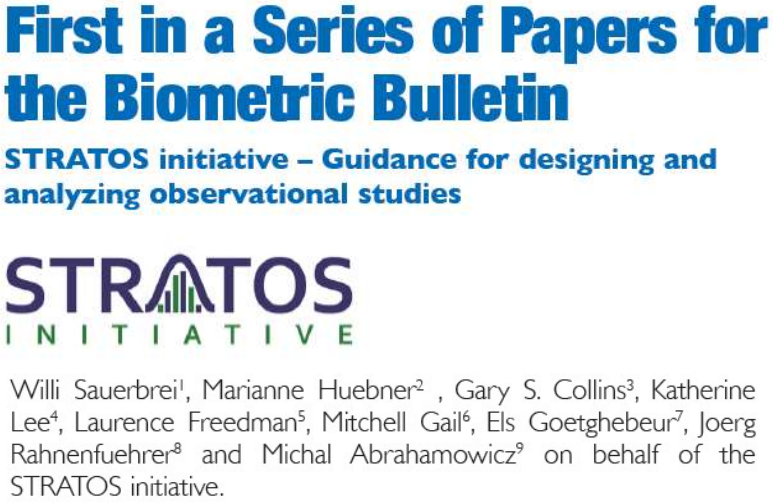
To improve design and statistical analyses of observational studies in practice
by ‘shortening the gap’ between
(i) recent relevant developments in statistical methodology *versus*
(ii) methods applied in real-life observational studies

Specific aims:

- Develop **evidence-supported guidance** for statistical issues of practical importance (*through experience and discussions among experts with different views, and simulations to systematically assess and compare alternative methods*)
- Provide guidance at **several levels of statistical knowledge**
- Start with **state-of-the-art** guidance for issues where there is consensus and necessary evidence
- **Identify and explore complex analytical challenges requiring more primary research** and/or **combining expertise** in different areas of statistical research

Guidance for analysis is needed for many stakeholders (analysts with different levels of knowledge, teachers, reviewers, journalists, ……)

Researchers



First in a Series of Papers for the Biometric Bulletin

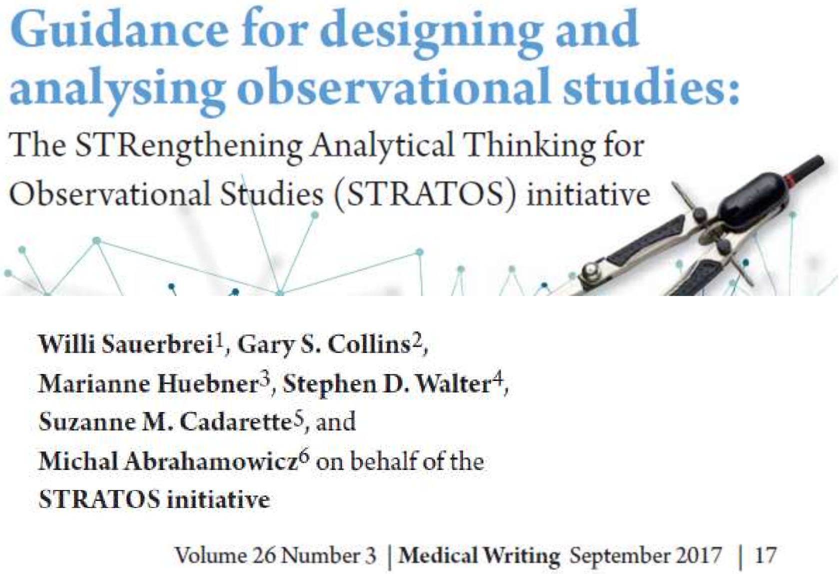
STRATOS initiative – Guidance for designing and analyzing observational studies

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 all TGs and some panels

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)

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

2013: Initiative launched at 44th Int Soc Clin Biostatistics (ISCB) conference

2014: 1st STRATOS paper: Sauerbrei W, Abrahamowicz M, Altman D, le Cessie S, Carpenter J. *STRengthening Analytical Thinking for Observational Studies: The STRATOS initiative, Statistics in Medicine* 2014

2016 & 2019: 2 General meetings, Banff Int Res Station (BIRS), Canada

By 2021: >100 members (from 19 countries on 5 continents)

Invited STRATOS Sessions and Mini-Symposia:

Int Soc Clin Biost (ISCB): 2014, 2015, 2016, 2018, 2019, 2020, 2021
Int Biometric Conf (IBC): 2016, 2020 + Regional IBS meetings: 2017, 2018, 2021
Royal Statistical Soc (RSS): 2018, 2020, 2021
Soc Epi Res (SER): 2021
Other international conferences: HEC 2016, CEN 2018, GMDS 2017

Series in the Biometric Bulletin (16 articles since 2017)

STRATOS Topic Groups (TGs)

Topic Group		Chairs
1	**Missing data**	James Carpenter (UK), Kate Lee (AUS)
2	**Selection of variables and functional forms in multivariable analysis**	Georg Heinze (AUT), Aris Perperoglou (UK), Willi Sauerbrei (GER)
3	**Initial data analysis**	Marianne Huebner (US), Saskia le Cessie(NL), Carsten Oliver Schmidt (GER)
4	**Measurement error and misclassification**	Laurence Freedman (ISR), Victor Kipnis (US)
5	**Study design**	Mitchell Gail (US), Suzanne Cadarette (CAN)
6	**Evaluating diagnostic tests and prediction models**	Ewout Steyerberg (NL), Ben van Calster (NL)
7	**Causal inference**	Els Goetghebeur (BEL), Ingeborg Waernbaum (SWE)
8	**Survival analysis**	Michal Abrahamowicz (CAN), Per Kragh Andersen (DEN), Terry Therneau (US)
9	**High-dimensional data**	Lisa McShane (US), Joerg Rahnenfuehrer (GER), Riccardo de Bin (NOR)

STRATOS Cross-cutting Panels

Panel		Chairs and Co-Chairs	
MP	**Membership**	Chairs:	James Carpenter (UK), Willi Sauerbrei (GER)
PP	**Publications**	Chairs:	Bianca De Stavola (UK), Pam Shaw (US)
		Co-Chairs:	Mitchell Gail (US), Petra Macaskill (AUS)
GP	**Glossary**	Chairs:	Martin Boeker (GER), Marianne Huebner (US)
WP	**Website**	Chairs:	Joerg Rahnenfuehrer (GER), Willi Sauerbrei (GER)
RP	**Literature Review**	Chairs:	Gary Collins (UK), Carl Moons (NL)
BP	**Bibliography**	Chairs:	to be determined
SP	**Simulation Studies**	Chairs:	Michal Abrahamowicz (CAN), Anne-Laure Boulesteix (GER)
DP	**Data Sets**	Chairs:	Saskia Le Cessie (NL), Maarten van Smeden (NL)
TP	**Knowledge Translation**	Chair:	Rolf Groenwold (NL), Maarten van Smeden (NL)
CP	**Contact Organisations**	Chairs:	Willi Sauerbrei (GER)
VP	**Visualisation**	Chairs:	Mark Baillie (SWITZ/CH)

Guidance for selection of variables and functional forms

Building multivariable regression models – some preliminaries

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

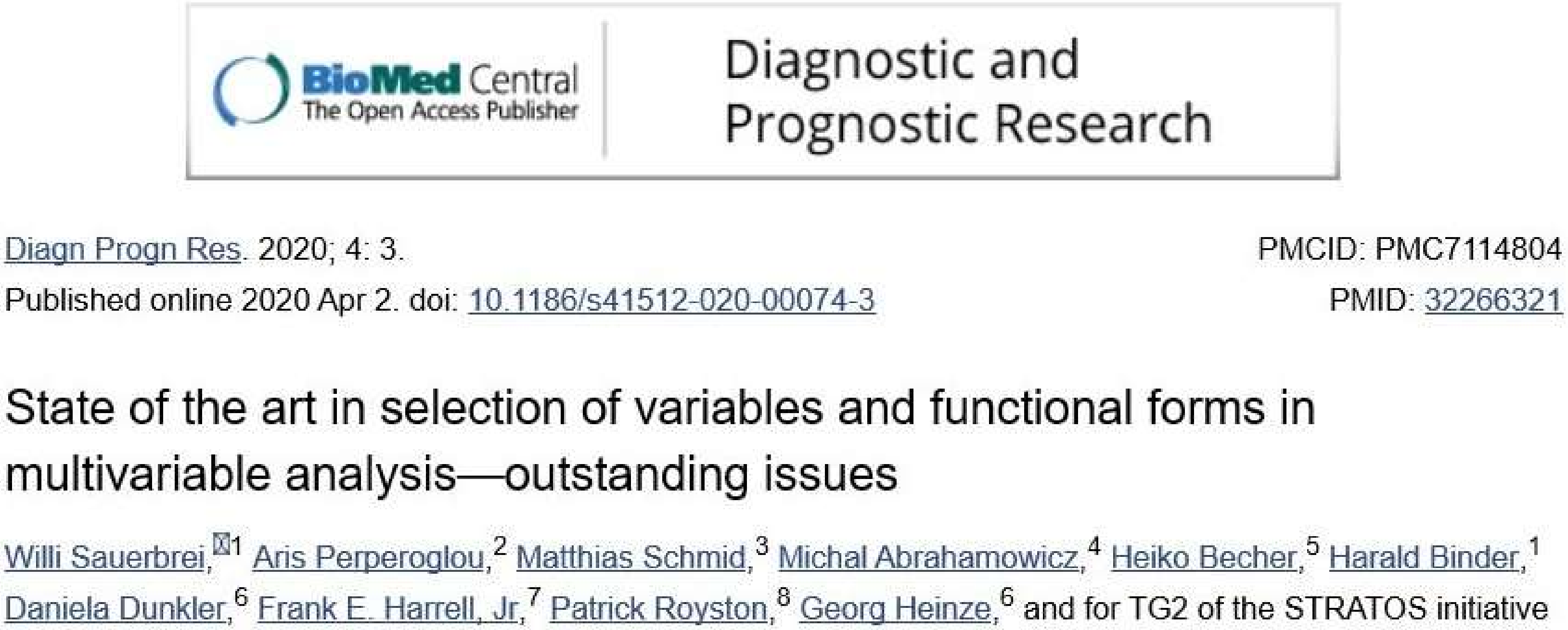
Aim of a model and model complexity

- Most important distinction:

  “to explain or to predict“ (Shmueli, 2010)

- Prediction (TG6)
- Here: **TG2**
  - model for concise description
- Causal inference (TG7)

TG2: Overview paper



BioMed Central The Open Access Publisher | Diagnostic and Prognostic Research

Diagn Progn Res. 2020; 4: 3. PMCID: PMC7114804

Published online 2020 Apr 2. doi: 10.1186/s41512-020-00074-3 PMID: 32266321

State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues

Willi Sauerbrei,✉[1] Aris Perperoglou,[2] Matthias Schmid,[3] Michal Abrahamowicz,[4] Heiko Becher,[5] Harald Binder,[1] Daniela Dunkler,[6] Frank E. Harrell, Jr,[7] Patrick Royston,[8] Georg Heinze,[6] and for TG2 of the STRATOS initiative

- 7 methodological issues identified

Selection of variables and functional forms – outstanding issues

Towards state of the art

1. Investigation and comparison of the properties of **variable selection strategies**
2. Comparison of **spline procedures** in both univariable and multivariable contexts
3. How to model one or more variables with a '**spike-at-zero**'?
4. Comparison of **multivariable procedures for model and function selection**
5. Role **of shrinkage** to correct for bias introduced by data-dependent modelling
6. Evaluation of new approaches for **post-selection inference**
7. Adaptation of procedures for **very large sample sizes** needed?

TG2: Part 1 - selection of variables

- Central issues:
  - Model with focus on prediction (TG6) or description (TG2)?
  - To select or not to select (full model)?
  - Which variables to include?
- A large number of methods proposed (for many decades)
- High-dimensional data (HDD) triggered the development of further proposals
  - HDD - prediction is the main aim (TG9)
- Many critical issues, do we have a ‘state of the art’?

Traditional variable selection strategies

- Full model
  - Variance inflation in case of multicollinearity
- Stepwise procedures
  - Forward Selection (FS)
  - Stepwise Selection (StS)
  - Backward Elimination (BE)
  - Which stopping criteria (AIC, BIC, p-value)?
    - ➢ Has a severe influence on complexity of model selected
- All subset selection
  - which criteria (AIC, BIC)? Or variants of it?

More recent approaches

- Procedures based on ‘change-in-estimate’
- Resampling-based variable selection procedures
- Bayesian approaches
- Modern variable selection strategies
  - Boosting
  - Penalised likelihood
    - Nonnegative garrote
    - Lasso (Extensions: Adaptive Lasso, Relaxed Lasso, etc.)
    - Elastic net
    - Smoothly Clipped Absolute Deviation (SCAD)
    - ……

Data dependent model-building introduces biases

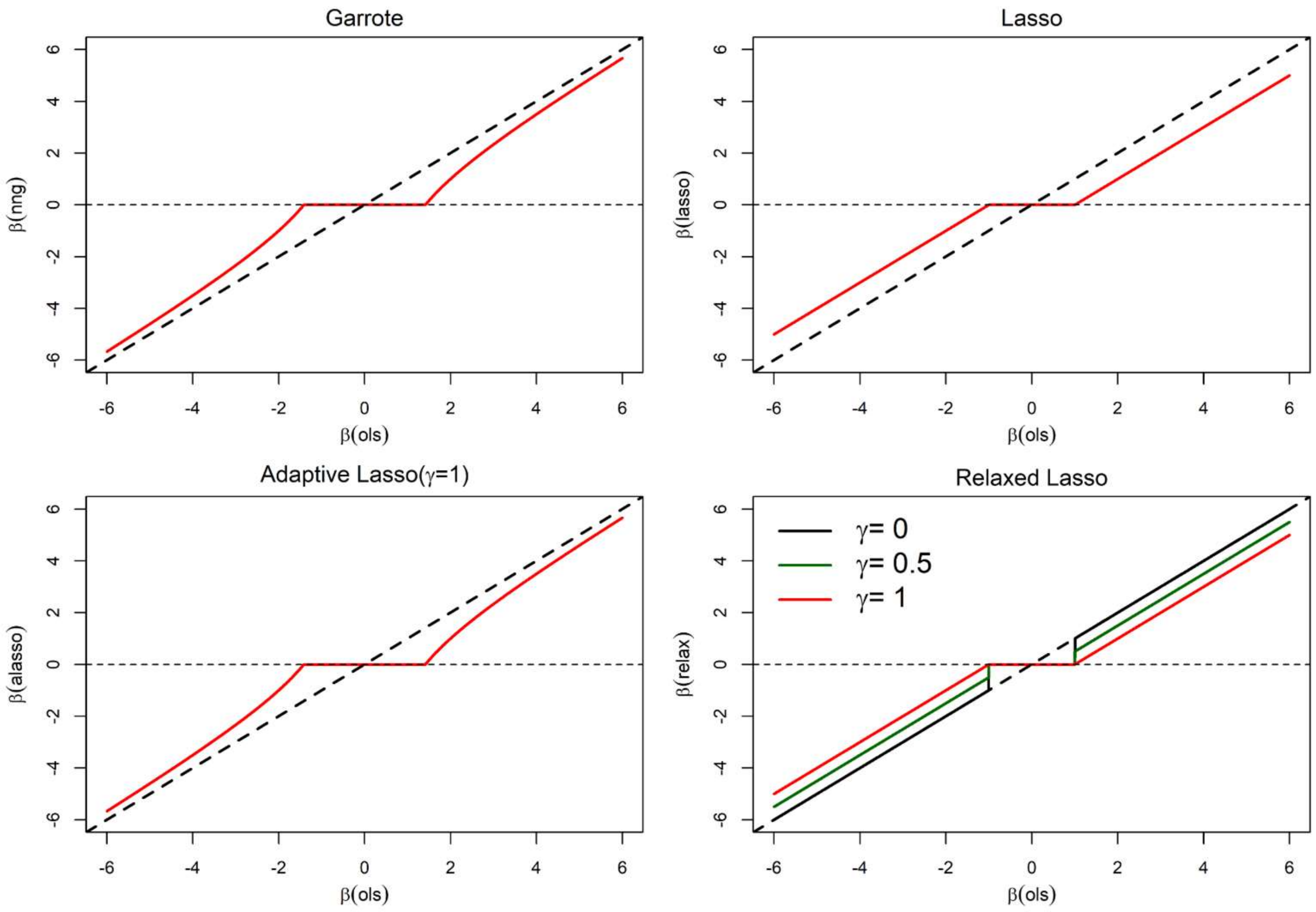
- Bias and the role of shrinkage methods
  - Several modern selection procedures combine variable selection and shrinkage to address bias and reduce MSE.
  - Post-estimation shrinkage (2 step approach) can be used for many types of models.

    Step 1: Select a model

    Step 2: Use leave-one-out (or other resampling technique) to
    estimate parameterwise shrinkage factors

Data dependent model-building introduces biases

Combine variable selection and shrinkage



Method	Effects	
	Large	Small
NNG	Hardly	Severe
Lasso	Equal amount	Equal amount
Alasso	Hardly	Severe
Rlasso	Equal amount	Equal amount
	$\gamma = 1$ Lasso $\gamma = 0$ No shrinkage $\gamma = 0.5$ Less shrinkage than lasso	

Amount of shrinkage

Combine variable selection and shrinkage

- Tuning parameter play a key role
- Lasso is popular for high dimensional data but suffers from overshrinkage of large effects
- Adaptive lasso and relaxed lasso were proposed to correct for overshrinkage
- Many more proposals
- Non negative garotte (NNG) can be used for correlated and high dimensional data

  - ➢ Direct comparisons needed

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

Functional forms:
Models based on cut-points: problems!

- Cut-points are still popular in clinical and epidemiological research
- Use of cut-points in a model gives a step function
- How many cut-points?
- Where should the cut-points be put?
- Biologically implausible step functions are a poor approximation to the true relationship
- Almost always fits the data less well than a suitable continuous function

- Nevertheless, in many areas still the preferred approach!

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**)

Combining variable and function selection

- Multivariable fractional polynomials (MFP)
- Various spline based approaches

Comparison in a large simulation study (Binder et al., 2013) Nevertheless, much more research is needed!

Splines - a brief overview of regression packages in R

Package	Downloads	Vignette	Book	Website	Datasets
quantreg	5099669	X	X		**8**
survival	3511997	X	X		**38**
mgcv	3217720	X	X		**2**
gbm	668984			X	**0**
VGAM	662399	X	X	X	**50**
gam	459497		X	X	**4**
gamlss	**210761**	**X**	**X**	**X**	**43**

Perperoglou et al. (2019)

Conclusion - Selection of variables and functional forms

- **We are far away from ‘state of the art’**

- Many more comparisons are urgently needed!

  - “*Exact distributional results are virtually impossible to obtain, even for simplest of common subset selection algorithms*“

    Picard & Cook, JASA, 1984

→ **Informative simulation studies are needed!**

… Conclusions

- Member of TG2 identified seven issues
- Other experts may have different experiences and preferences … and may raise further issues
- To help deriving evidence-supported guidance, more cooperative and comparative research is needed from experts

Summary – relevance of STRATOS

- Data and data science becomes more and more important
- Answering questions empirically through data analyses often requires the use of complex methodology. It is important to **develop suitable approaches**; needs to be done by **experts (Level 3)**
- **Experienced statisticians (Level 2)** need to be **supported by suitable guidance**. There are (too) many approaches (some are useless) available and suitable comparisons are missing
- **Better simulation studies** are required to assess properties, compare approaches and derive **evidence based guidance for practice**.
- Suitable **educational material** is the key to **improve analyses at a broad level**
- For practically relevant topics we need **greater emphasis on development of Level 1 and 2 guidance**

References

- Sauerbrei W, Abrahamowicz M, Altman DG, le Cessie S and Carpenter J on behalf of the STRATOS initiative. (2014): STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. Statistics in Medicine, 33: 5413-5432, DOI: 10.1002/sim.6265.
- Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M on behalf of TG2 of the STRATOS initiative (2019): A review of spline function procedures in R, BMC Medical Research Methodology, 19:46
- Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, Dunkler D, Harrell FE, Royston P, Heinze G, and TG2 of the STRATOS initiative. (2020): State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. Diagnostic and prognostic research, 4, pp.1-18.
- Heinze G, Perperoglou A, Sauerbrei W on behalf of Topic Group 2 of the STRATOS initiative. (2021): STRengthening Analytical Thinking for Observational Studies (STRATOS): Recent activities of the Topic Group on Selection of Variables and Functional Forms in Multivariable Analysis (TG2). Biometric Bulletin; 38(2):7-8.
- Heinze G, Wallisch C, and Dunkler D. (2018): Variable selection–a review and recommendations for the practicing statistician. Biometrical Journal, 60(3), pp.431-449.
- Binder H, Sauerbrei W, and Royston P. (2013): Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. Statistics in Medicine. 32:2262-2277.
- Desboulets LDD. (2018): A Review on Variable Selection in Regression Analysis. Econometrics, 6(4), 45.
- Lu Z and Lou W. (2021): Bayesian approaches to variable selection: a comparative study from practical perspectives. The International Journal of Biostatistics.

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)
- Thomas Cowling (U.K.)

And the early career adjunct members

- Michael Kammer (Vienna, Austria)
- Edwin Kipruto (Freiburg, Germany)
- Christine Wallisch (Vienna, Austria)

- Daniela Dunkler (Austria)
- Rolf Groenwold (Netherlands)
- Frank Harrell (U.S.A)
- Nadja Klein (Germany)
- Geraldine Rauch (Germany)
- Patrick Royston (U.K.)
- Matthias Schmid (Germany)