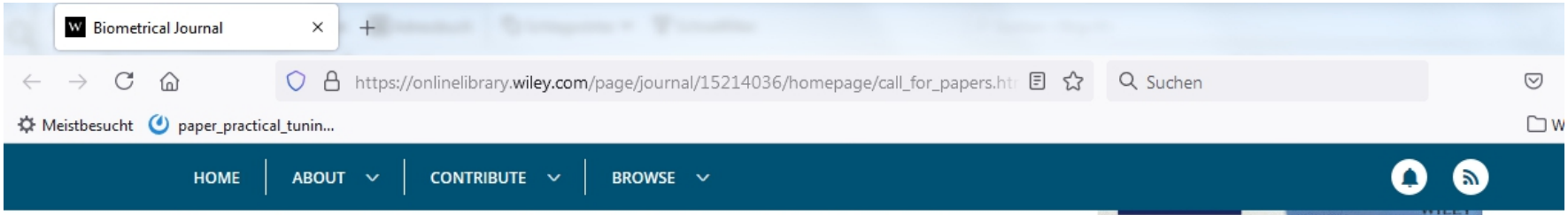
STRATOS Simulation Panel: Overview and discussion on neutral comparison studies

Chairs: Michal Abrahamowicz and Anne-Laure Boulesteix

March 27h, 2023
STRATOS Meeting

Past and present activities

- Boulesteix, H. Binder, M. Abrahamowicz and W. Sauerbrei (Biom J 2018)
- Boulesteix et al. (BMJ Open 2020), including M. Abrahamowicz, T. Morris, W. Sauerbrei, H. Binder, R. Groenwold and J. Rahnenführer
- Heinze et al. (Biom J 2023), including M. Kammer, T. Morris, I. White, ALB
- Abrahamowicz et al. (2023, in revision), see talk by M. Abrahamowicz
- Special issue of Biometrical Journal guest-edited by ALB, W. Sauerbrei, T. Morris, L. Held, D. Edelmann and M. Baillie: "Towards neutral comparison studies in methodological research".



Towards neutral comparison studies in methodological research

Guest editors: Anne-Laure Boulesteix (coordinator), Mark Baillie, Dominic Edelmann, Leonhard Held, Tim Morris, Willi Sauerbrei

Biomedical researchers are frequently faced with an array of methods they might potentially use for the analysis and/or design of studies. It can be difficult to understand the absolute and relative merits of candidate methods beyond one's own particular interests and expertise.
Choosing a method can be difficult even in simple settings but an increase in the volume of data collected, computational power and methods proposed in the literature makes the choice all the more difficult. In this context, it is crucial to provide researchers with evidence-supported guidance derived from appropriately designed studies comparing statistical methods in a neutral way, in particular through well-designed simulation studies.

While neutral comparison studies are an essential cornerstone towards the improvement of this situation, a number of challenges remain with regard to their methodology and acceptance. Numerous difficulties arise when designing, conducting and reporting neutral comparison studies. Practical experience is still scarce and literature on these issues almost inexistent. Furthermore, authors of neutral comparison studies are often faced with incomprehension from a large part of the scientific community which is more interested in the development of 'new' approaches and evaluates the importance of research primarily based on the novelty of the presented methods. Consequently, meaningful comparisons of competing approaches (especially reproducible studies including publicly available code and data) are rarely available and evidence-supported state of the art guidance is largely missing, often resulting in the use of suboptimal methods in practice.

In this context, this special issue intends to publish both:

- well-designed neutral comparison studies of methods (including but not limited to studies arising from community challenges), i.e. comparison studies fulfilling the two following criteria: (i) focused on the comparison of existing methods already described elsewhere rather than on a new prototype method being introduced; (ii) authored by a group of researchers who are (ideally) approximately equally familiar with all the compared methods;
- papers defining, developing, discussing or illustrating concepts related to practical issues and improvement of neutral comparison studies in the context of methodological biometrical research, including but not limited to the design, analysis and presentation of reliable simulation studies, study protocols, study registration and (structured) reporting, replication studies, uncertainty quantification and research synthesis. Papers of this type will provide a lens through which to critically reflect on neutral comparison studies in the future.

Papers addressing parts of one or both aspects are also welcome.

All codes and data should be made available following the reproducibility guidelines of the Biometrical Journal: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201500156

SIGNIFICANCE | ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS | ASA AMERICAN STATISTICAL ASSOCIATION Promoting the Practice and Profession of Statistics | Statistical Society of Australia

Features

A replication crisis in methodological research?

Anne-Laure Boulesteix, Sabine Hoffmann, Alethea Charlton, Heidi Seibold,

First published: 29 September 2020 | https://doi.org/10.1111/1740-9713.01444 | Citations: 1

Read the full text > PDF TOOLS SHARE

Abstract

Statisticians have been keen to critique statistical aspects of the "replication crisis" in other scientific disciplines. But new statistical tools are often published and promoted without any thought to replicability. This needs to change, argue **Anne-Laure Boulesteix, Sabine Hoffmann, Alethea Charlton** and **Heidi Seibold**

ALB/Charlton/Hoffmann/Seibold, Significance 2020.

BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 16 2010, pages 1990–1998
doi:10.1093/bioinformatics/btq323

Gene expression

Advance Access publication June 26, 2010

Over-optimism in bioinformatics: an illustration

Monika Jelizarow[1], Vincent Guillemot[1,2], Arthur Tenenhaus[2], Korbinian Strimmer[3] and Anne-Laure Boulesteix[1,*]

[1]Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany, [2]SUPELEC Sciences des Systèmes (E3S)-Department of Signal Processing and Electronics Systems - 3, rue Joliot Curie, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France and [3]Department of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Härtelstr. 16-18, 04107 Leipzig, Germany

Associate Editor: John Quackenbush

ABSTRACT

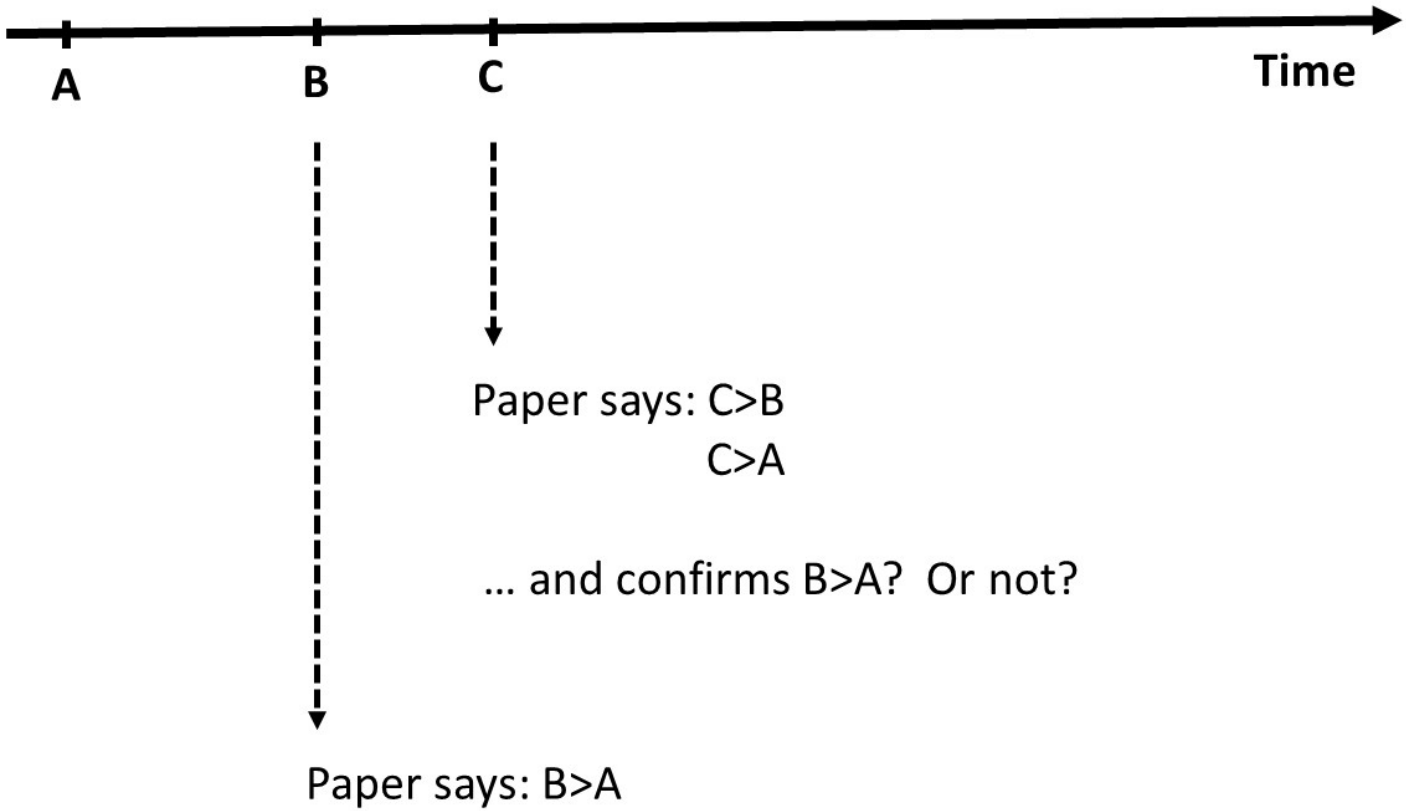
Motivation: In statistical bioinformatics research, different optimization mechanisms potentially lead to 'over-optimism' in published papers. So far, however, a systematic critical study concerning the various sources underlying this over-optimism is lacking.

Results: We present an empirical study on over-optimism using high-dimensional classification as example. Specifically, we consider a 'promising' new classification algorithm, namely linear discriminant analysis incorporating prior knowledge on gene functional groups

it would be wrong to report only favorable datasets without mentioning and/or discussing the other results. This strategy induces an optimistic bias. This aspect of over-optimism is quantitatively investigated in the study by Yousefi *et al.* (2010) and termed as 'optimization of the dataset' in this article.

The second source of over-optimism, which is related to the optimal choice of the dataset mentioned above, is the optimal choice of a particular setting in which the superiority of the new algorithm is more pronounced. For example, researchers could report the results obtained after a particular feature filtering which favors the

"The New Method Performed Better Than Existing Ones"



Buchka/Hapfelmeier/Gardner/Wilson/ALB, Genome Biology 2021.

Methodological Computational Research vs. Clinical Research

Clinical Research	**Methodological Research**
drugs/interventions	methods
improve health outcome	make results of statistical analyses closer to the truth
practitioners	statistical consultants
patients	datasets
trialists	methodological researchers
health outcome personalized medicine	method performance meta-learning

(ALB/Wilson/Hapfelmeier, BMC Med Res Meth 2017)

Design and Interpretation of Experiments

- Decades of research and discussions on appropriate designs of clinical trials
  - sample size
  - inclusion criteria
  - placebo
  - missing values
  - authors' neutrality
  - blinding
  - levels of evidence
  - replication studies
  - reporting
  - publication bias
- Almost no research on the design of empirical studies comparing statistical methods...

New Methods Required...

Imagine that medical journals require authors to present new prototype treatments in all articles, but reject clinical trials because the treatment's principle is not new ("it has been described elsewhere before")?

Towards a phases system in methodological research

Biometrical Journal

RESEARCH ARTICLE | Open Access

Phases of methodological research in biostatistics—Building the evidence base for new methods

Georg Heinze, Anne-Laure Boulesteix, Michael Kammer, Tim P. Morris, Ian R. White, the Simulation Panel of the STRATOS initiative

First published: 03 February 2023 | https://doi.org/10.1002/bimj.202200222 | Citations: 1

ABSTRACT

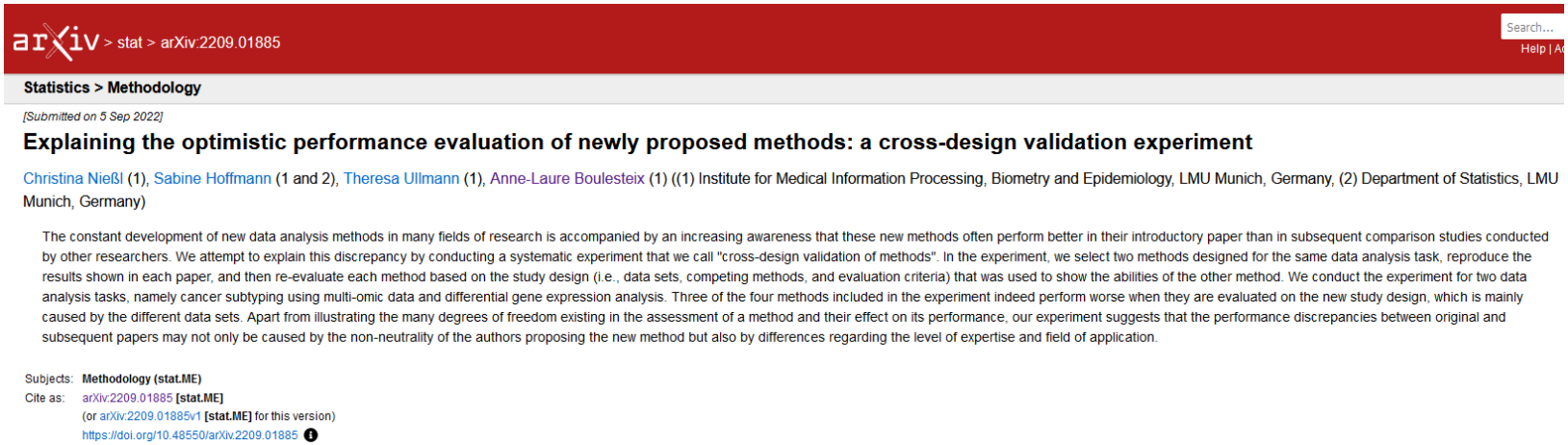
Although new biostatistical methods are published at a very high rate, many of these developments are not trustworthy enough to be adopted by the scientific community. We propose a framework to think about how a piece of methodological work contributes to the evidence base for a method. Similar to the well-known phases of clinical research in drug development, we propose to define four phases of methodological research. These four phases cover (I) proposing a new methodological idea while providing, for example, logical reasoning or proofs, (II) providing empirical evidence, first in a narrow target setting, then (III) in an extended range of settings and for various outcomes, accompanied by appropriate application examples, and (IV) investigations that establish a method as sufficiently well-understood to know when it is preferred over others and when it is not; that is, its pitfalls. We suggest basic definitions of the four phases to provoke thought and discussion rather than devising an unambiguous classification of studies into phases. Too many methodological developments finish before phase III/IV, but we give two examples with references. Our concept rebalances the emphasis to studies in phases III and IV, that is, carefully planned method comparison studies and studies that explore the empirical properties of existing methods in a wider range of problems.

What do we need?

- More **neutral** comparison studies, such as STRATOS studies
- Better acceptance for neutral comparison studies
- More research and efforts on the design of such studies needed
  → STRATOS Simulation Panel
- More transparency and discussions on ethics in methodological (and medical) research
  → STRATOS Open Science Panel, see talk by S. Hoffmann

Cross-design validation

	Performance of method A	Performance of method B
Study design by authors of method A	Shown in original paper	?
Study design by authors of method B	?	Shown in original paper



arXiv > stat > arXiv:2209.01885

Search... Help | A

Statistics > Methodology

[Submitted on 5 Sep 2022]

Explaining the optimistic performance evaluation of newly proposed methods: a cross-design validation experiment

Christina Nießl (1), Sabine Hoffmann (1 and 2), Theresa Ullmann (1), Anne-Laure Boulesteix (1) ((1) Institute for Medical Information Processing, Biometry and Epidemiology, LMU Munich, Germany, (2) Department of Statistics, LMU Munich, Germany)

The constant development of new data analysis methods in many fields of research is accompanied by an increasing awareness that these new methods often perform better in their introductory paper than in subsequent comparison studies conducted by other researchers. We attempt to explain this discrepancy by conducting a systematic experiment that we call "cross-design validation of methods". In the experiment, we select two methods designed for the same data analysis task, reproduce the results shown in each paper, and then re-evaluate each method based on the study design (i.e., data sets, competing methods, and evaluation criteria) that was used to show the abilities of the other method. We conduct the experiment for two data analysis tasks, namely cancer subtyping using multi-omic data and differential gene expression analysis. Three of the four methods included in the experiment indeed perform worse when they are evaluated on the new study design, which is mainly caused by the different data sets. Apart from illustrating the many degrees of freedom existing in the assessment of a method and their effect on its performance, our experiment suggests that the performance discrepancies between original and subsequent papers may not only be caused by the non-neutrality of the authors proposing the new method but also by differences regarding the level of expertise and field of application.

Subjects: **Methodology (stat.ME)**
Cite as: arXiv:2209.01885 **[stat.ME]**
(or arXiv:2209.01885v1 **[stat.ME]** for this version)
https://doi.org/10.48550/arXiv.2209.01885

Nießl/Hoffmann/Ullmann/ALB, Biom J 2023.

Replication

frontiers | Frontiers in Epidemiology

TYPE Perspective
PUBLISHED 14 September 2022
DOI 10.3389/fepid.2022.973470

OPEN ACCESS

EDITED BY
Amand Floriaan Schmidt,
University College London,
United Kingdom

REVIEWED BY
Kellyn F. Arnold,
IQVIA, United Kingdom

*CORRESPONDENCE
Rolf H. H. Groenwold
r.h.h.groenwold@lumc.nl

SPECIALTY SECTION
This article was submitted to Research Methods and Advances in Epidemiology, a section of the journal Frontiers in Epidemiology

RECEIVED 20 June 2022
ACCEPTED 17 August 2022
PUBLISHED 14 September 2022

CITATION
Lohmann A, Astivia OLO, Morris TP and Groenwold RHH (2022) It's time! Ten reasons to start replicating simulation studies. *Front. Epidemiol.* 2:973470. doi: 10.3389/fepid.2022.973470

It's time! Ten reasons to start replicating simulation studies

Anna Lohmann[1], Oscar L. O. Astivia[2], Tim P. Morris[3] and Rolf H. H. Groenwold[1,4]*

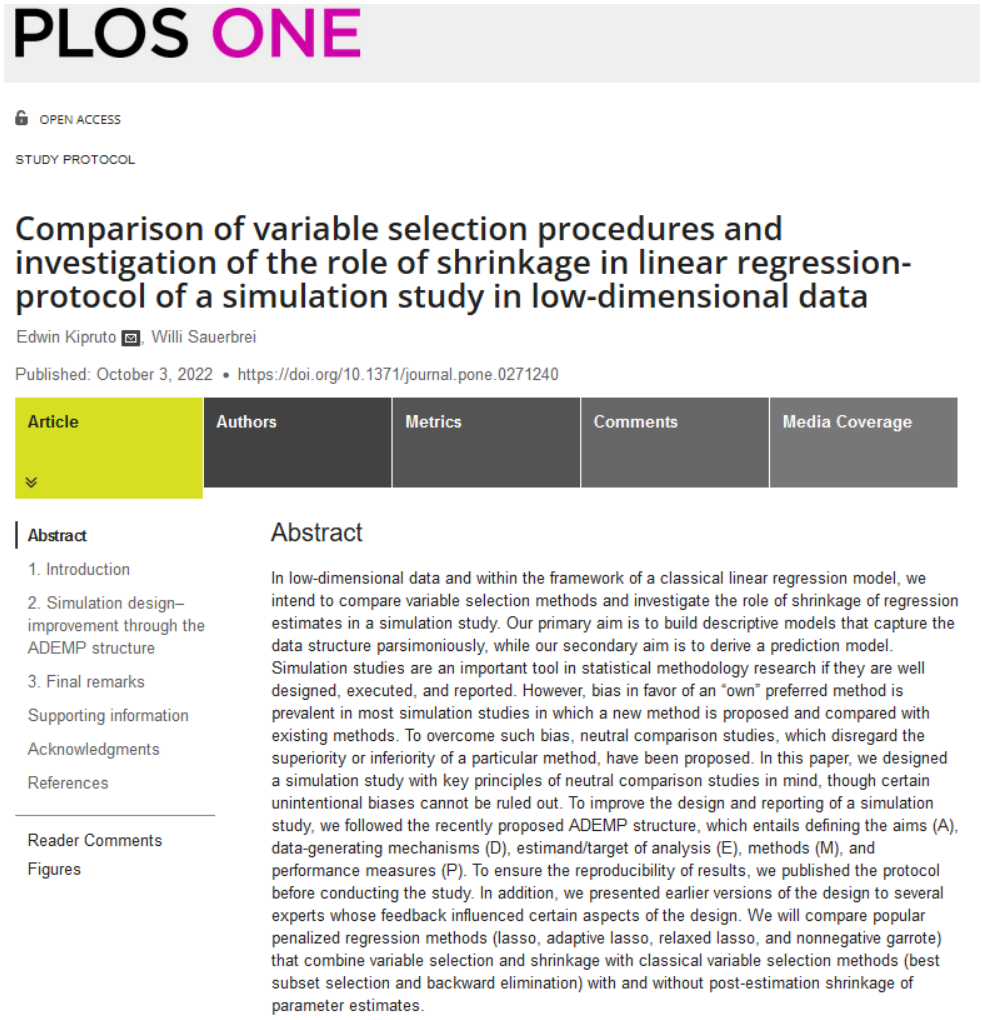
[1]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, Netherlands, [2]College of Education, University of Washington, Seattle, WA, United States, [3]MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, University College London, London, United Kingdom, [4]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands

The quantitative analysis of research data is a core element of empirical research. The performance of statistical methods that are used for analyzing empirical data can be evaluated and compared using computer simulations. A single simulation study can influence the analyses of thousands of empirical studies to follow. With great power comes great responsibility. Here, we argue that this responsibility includes replication of simulation studies to ensure a sound foundation for data analytical decisions. Furthermore, being designed, run, and reported by humans, simulation studies face challenges similar to other experimental empirical research and hence should not be exempt from replication attempts. We highlight that the potential replicability of simulation studies is an opportunity quantitative methodology as a field should pay more attention to.

KEYWORDS

replication, data analysis, research statistics, simulation study, reproduction

Study protocols in methodological research



PLOS ONE

OPEN ACCESS

STUDY PROTOCOL

Comparison of variable selection procedures and investigation of the role of shrinkage in linear regression-protocol of a simulation study in low-dimensional data

Edwin Kipruto, Willi Sauerbrei

Published: October 3, 2022 • https://doi.org/10.1371/journal.pone.0271240

Article | Authors | Metrics | Comments | Media Coverage

Abstract

In low-dimensional data and within the framework of a classical linear regression model, we intend to compare variable selection methods and investigate the role of shrinkage of regression estimates in a simulation study. Our primary aim is to build descriptive models that capture the data structure parsimoniously, while our secondary aim is to derive a prediction model. Simulation studies are an important tool in statistical methodology research if they are well designed, executed, and reported. However, bias in favor of an "own" preferred method is prevalent in most simulation studies in which a new method is proposed and compared with existing methods. To overcome such bias, neutral comparison studies, which disregard the superiority or inferiority of a particular method, have been proposed. In this paper, we designed a simulation study with key principles of neutral comparison studies in mind, though certain unintentional biases cannot be ruled out. To improve the design and reporting of a simulation study, we followed the recently proposed ADEMP structure, which entails defining the aims (A), data-generating mechanisms (D), estimand/target of analysis (E), methods (M), and performance measures (P). To ensure the reproducibility of results, we published the protocol before conducting the study. In addition, we presented earlier versions of the design to several experts whose feedback influenced certain aspects of the design. We will compare popular penalized regression methods (lasso, adaptive lasso, relaxed lasso, and nonnegative garrote) that combine variable selection and shrinkage with classical variable selection methods (best subset selection and backward elimination) with and without post-estimation shrinkage of parameter estimates.

Thanks for attention!
Thanks to collaboration partners and STRATOS!
Thanks to DFG and BMBF for funding!