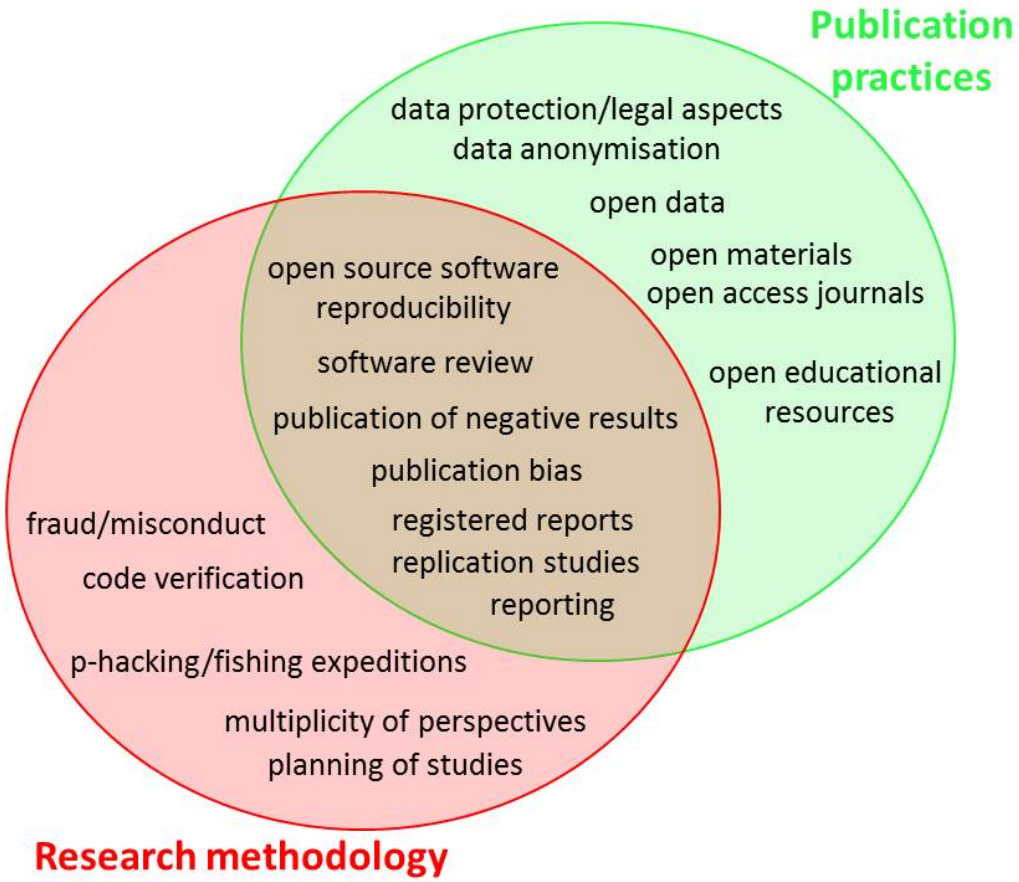
A replication crisis in methodological research? On the design of comparison studies

Anne-Laure Boulesteix
Dept. of Medical Information Processing, Biometry and Epidemiology
LMU Munich, Germany

Virtual ISCB, STRATOS Mini-Symposium
August 27th 2020


Open/Replicable Science



LMU Open Science Center

Role of statisticians in Open/Replicable Science

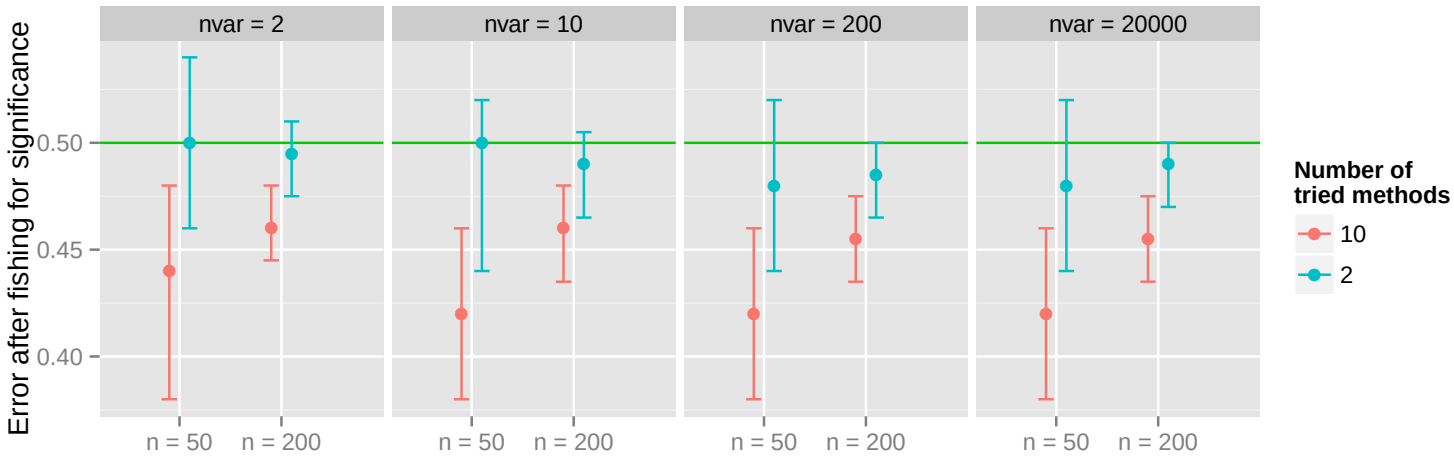
- Applied statisticians:
  - perform statistical planning, write SAP
  - provide/verify codes for reproducibility
  - don't fish for significance
  - design and conduct replication studies
  - correct for publication bias in meta-analysis
- Methodological statistics:
  - provide methods to correct for fishing/publication bias
  - provide methods to design replication studies (Held, 2019)
  - provide methods to cope with the multiplicity of analysis strategies
  - provide methods to ensure confidentiality while sharing
  - provide methods to detect fraud/errors
  - ...

Data analyst's degree of freedom

"Give me information on a single gene and 200 patients, half of them dead, please. I bet that I can show that this gene affects survival ($p<0.05$) even if it does not. One can do analyses: counting or ignoring exact follow-up, censoring at different timepoints, excluding specific causes of death, exploiting subgroup analyses, using dozens of different cut-offs to decide what constitutes inappropriate gene expression, and so forth. Without highly specified a priori hypotheses, there are hundreds of ways to analyse the dullest dataset. "

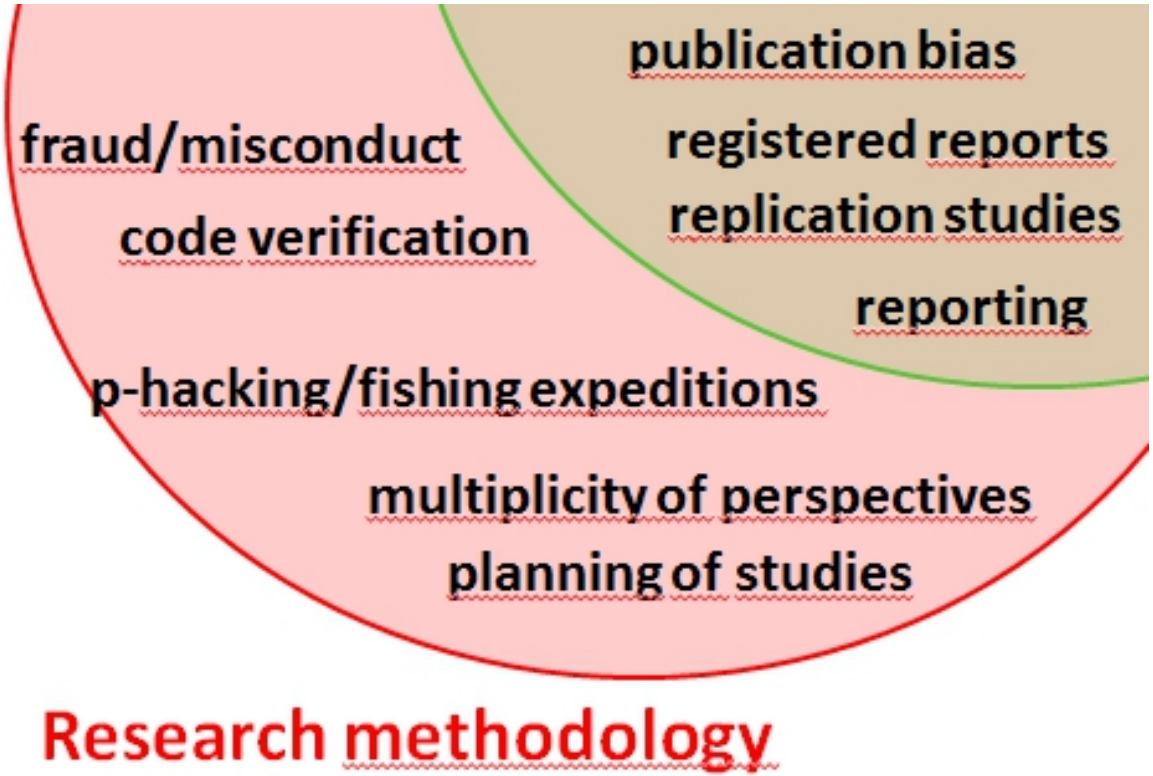
Fishing expeditions in prognostic modelling (high-dimensional data)

- $K = 2$ or $K = 10$ supervised learning algorithms
- sample size $n = 50$ or $n = 200$
- $nvar = 2, 10, 200, 20000$ variables



Boulesteix et al., 2017. In "Ott, Max; Pietsch, Wolfgang; Wernecke, Jörg. Berechenbarkeit der Welt? Philosophie und Wissenschaft im Zeitalter von Big Data. Wiesbaden: Springer VS". pp 155-170.

A replication crisis in statistical *methodological* research?



Replicability vs. reproducibility

- Regression models are easily transportable
- Random forest needs
  - web-based tools
  - or software object
  - or data & code

Survey of 30 research papers:

- presenting Random Forest prediction rules
- with "random forest" in the abstract
- published in PLOS ONE
- in 2014-2015
- in the field "medical and health science"

Boulesteix et al. Biometrical Journal 2019

A replication crisis in statistical *methodological* research?

- no publication of negative results
- strong publication bias
- strong incentive to go for fishing "expeditions"
- poor planning of evaluation studies
- new methods required
- few neutral comparison studies
- no replication studies

Publication bias

- new methods performing worse don't get published
- topic is (almost) taboo in computational science
- well-known in health and social science (Sterling, 1959!)

Publication Bias in Methodological Computational Research

Anne-Laure Boulesteix[1], Veronika Stierle[1] and Alexander Hapfelmeier[2]

[1]Department of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilian University, Munich, Germany. [2]Department of Medical Statistics and Epidemiology, Klinikum rechts der Isar Technical University of Munich, Munich, Germany.

Supplementary Issue: Statistical Systems Theory in Cancer Modeling, Diagnosis, and Therapy

ABSTRACT: The problem of publication bias has long been discussed in research fields such as medicine. There is a consensus that publication bias is a reality and that solutions should be found to reduce it. In methodological computational research, including cancer informatics, publication bias may also be at work. The publication of negative research findings is certainly also a relevant issue, but has attracted very little attention to date. The present paper aims at providing a new formal framework to describe the notion of publication bias in the context of methodological computational research, facilitate and stimulate discussions on this topic, and increase awareness in the scientific community. We report an exemplary pilot study that aims at gaining experiences with the collection and analysis of information on unpublished research efforts with respect to publication bias, and we outline the encountered problems. Based on these experiences, we try to formalize the notion of publication bias.

KEYWORDS: epistemology, publication practice, false research findings, overoptimism

Fishing expeditions in methodological research

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

Methodological research in applied statistics 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	method performance
personalized medicine	meta-learning

Planning of evaluation studies

- decades of research and discussions on appropriate designs of clinical trials
  - sample size
  - inclusion criteria
  - placebo
  - missing values
  - authors' neutrality
  - blinding
  - levels of evidence
  - ...
- almost no research on the design of studies comparing statistical methods...

Boulesteix et al., BMC Med Res Meth 2017

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")?

Boulesteix et al., BMC Med Res Meth 2017

Suggestions to improve research methodology

Received: 15 August 2017 | Revised: 20 October 2017 | Accepted: 22 October 2017

DOI: 10.1002/bimj.201700129

LETTER TO THE EDITOR

Biometrical Journal

On the necessity and design of studies comparing statistical methods

In data analysis sciences in general and in biometrical research particularly, there are strong incentives for presenting work that entails new methods. Many journals require authors to propose new methods as a prerequisite for publication, as this is the most straightforward way to claim the necessary novelty. The development of new methods is also factually often a sine qua non condition to be recruited as a faculty member or to obtain personnel funding from a methods-oriented research agency, not least because it noticeably increases the chance to get published as outlined above. Thus, in statistical research and related methodology-oriented fields such as machine learning or bioinformatics, the well-known adage "publish or perish" could be translated into "propose new methods or perish."

- More neutral comparison studies
- More research on the design of comparison studies:
  - with real data
  - with simulated data

Boulesteix, Binder, Abrahamowicz & Sauerbrei,
for the **STRATOS simulation panel**. Biom J 2018.

STRATOS simulation panel

Level 1 paper:

> Boulesteix, Groenwold, Abrahamowicz, Binder, Briel, Hornung, Morris, Rahnenführer, Sauerbrei, on behalf of the Simulation Panel of the STRATOS initiative. An introduction to statistical simulations in health research (submitted).

Further projects (level 2, lead by M. Abrahamowicz, T. Morris and W. Sauerbrei): design and reporting of simulation studies.

Example study: prognostic modelling with multi-omics

- prognostic modelling with multi-omics data (gene expression, CNV, miRNA, etc.)
- outcome: (censored) survival time
- 18 real cancer datasets from TCGA
- 13 methods based on boosting, random forest, penalized regression
- naive or taking multi-omics structure into account, favoring clinical variables or not
- implemented using 'mlr'

  Herrmann et al., 2020. Briefings in Bioinformatics (in press).

Non-neutrality disclosure

- Myself and my lab (S. Janitza, R. Hornung, P. Probst, R. Couronne) have been involved in methodological research on random forests.
- Myself and my lab (R. De Bin, M. Fuchs, S. Klau, A. Volkmann, R. Hornung) have been involved in methodological research on penalized regression.
- Myself (before 2010) and my lab (R. De Bin, A. Volkmann) have been involved in methodological research on boosting.

Are we **neutral** ("enough")?

Example study: results

- very large variability over datasets!
- different results with c-index and Brier
- 18 is much more than usual but not much...
- no clear winner, poor results...
- blockForest = only method performing (slightly) better than clinical model on average
- over-optimism for blockForest?
- favoring clinical variables is good.
- "best method" depends on criteria (accuracy, transportability, sparsity, etc.)

  Herrmann et al., 2020. Briefings in Bioinformatics (in press).

Thank you!

Thanks to: STRATOS simulation panel, H. Hapfelmeier, M. Herrmann, R. Hornung, M. Jelizarow, V. Jurinovic, S. Lauer, W. Sauerbrei, V. Stierle, R. Wilson

and to the DFG (BO3139/2-2, BO3139/2-3, BO3139/6-1, BO3139/4-2, BO3139/4-3) for funding