# STRATOS Update
# TG9: High-dimensional data (HDD)

**TG9 Co-chairs:**

Riccardo De Bin

Lisa McShane

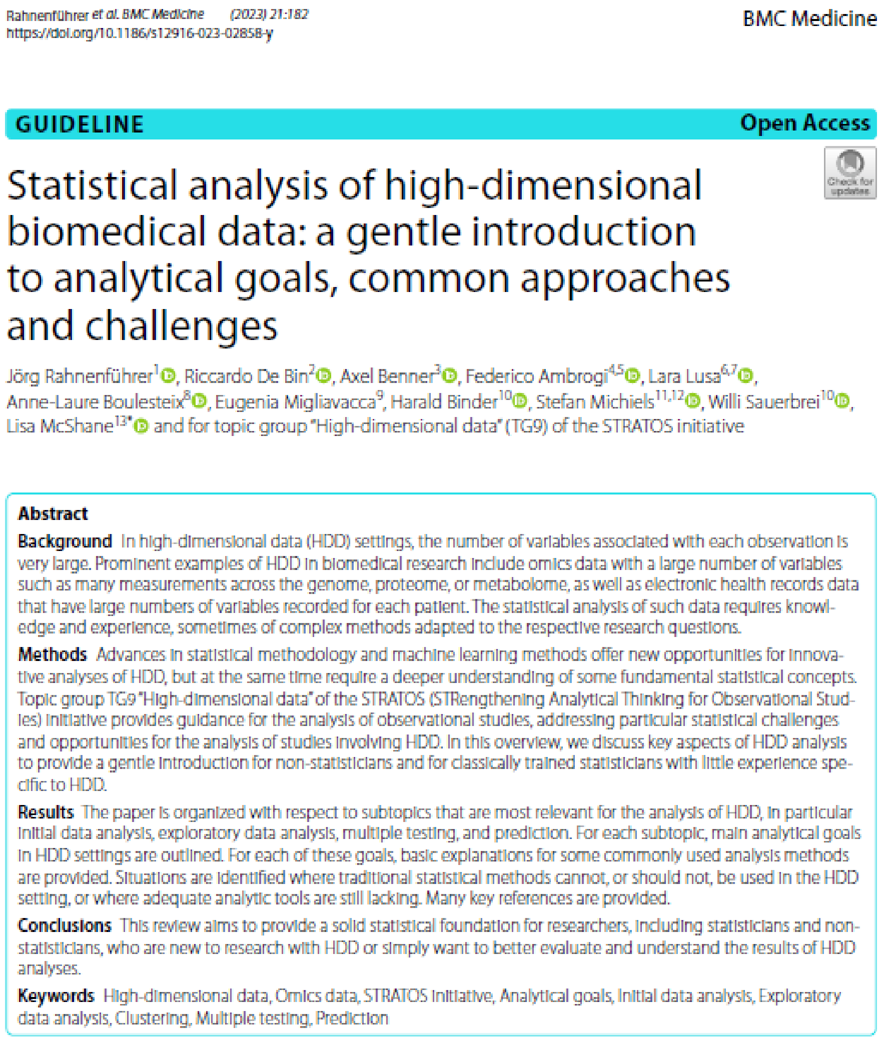
Jorg Rahnenfuhrer (stepping down as a co-chair)

Federico Ambrogi (recently joined as a co-chair)

# TG9 Roster

- Federico Ambrogi, University of Milan
- Axel Benner, DKFZ Heidelberg
- Harald Binder, University of Freiburg
- Anne-Laure Boulesteix, Ludwig Maximilian University of Munich
- Riccardo De Bin, University of Oslo
- Kevin Dobbin, Medical College of Georgia at Augusta
- Roman Hornung, Ludwig Maximilian University of Munich
- Lara Lusa, University of Primorksa and University of Ljubljana
- Lisa McShane, U.S. National Cancer Institute
- Stefan Michiels, Gustave Roussy
- Eugenia Migliavacca, Nestle Research
- Jörg Rahnenführer, TU Dortmund
- Willi Sauerbrei, University of Freiburg
- Nicholas Schreck, DKFZ Heidelberg
- Martin Treppner, University of Freiburg

# STRATOS TG9 high-dimensional data overview paper



Rahnenführer *et al. BMC Medicine* (2023) 21:182
https://doi.org/10.1186/s12916-023-02858-y

BMC Medicine

GUIDELINE Open Access

## Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges

Jörg Rahnenführer[1], Riccardo De Bin[2], Axel Benner[3], Federico Ambrogi[4,5], Lara Lusa[6,7], Anne-Laure Boulesteix[8], Eugenia Migliavacca[9], Harald Binder[10], Stefan Michiels[11,12], Willi Sauerbrei[10], Lisa McShane[13*] and for topic group "High-dimensional data" (TG9) of the STRATOS initiative

**Abstract**

**Background** In high-dimensional data (HDD) settings, the number of variables associated with each observation is very large. Prominent examples of HDD in biomedical research include omics data with a large number of variables such as many measurements across the genome, proteome, or metabolome, as well as electronic health records data that have large numbers of variables recorded for each patient. The statistical analysis of such data requires knowledge and experience, sometimes of complex methods adapted to the respective research questions.

**Methods** Advances in statistical methodology and machine learning methods offer new opportunities for innovative analyses of HDD, but at the same time require a deeper understanding of some fundamental statistical concepts. Topic group TG9 "High-dimensional data" of the STRATOS (STRengthening Analytical Thinking for Observational Studies) initiative provides guidance for the analysis of observational studies, addressing particular statistical challenges and opportunities for the analysis of studies involving HDD. In this overview, we discuss key aspects of HDD analysis to provide a gentle introduction for non-statisticians and for classically trained statisticians with little experience specific to HDD.

**Results** The paper is organized with respect to subtopics that are most relevant for the analysis of HDD, in particular initial data analysis, exploratory data analysis, multiple testing, and prediction. For each subtopic, main analytical goals in HDD settings are outlined. For each of these goals, basic explanations for some commonly used analysis methods are provided. Situations are identified where traditional statistical methods cannot, or should not, be used in the HDD setting, or where adequate analytic tools are still lacking. Many key references are provided.

**Conclusions** This review aims to provide a solid statistical foundation for researchers, including statisticians and non-statisticians, who are new to research with HDD or simply want to better evaluate and understand the results of HDD analyses.

**Keywords** High-dimensional data, Omics data, STRATOS initiative, Analytical goals, Initial data analysis, Exploratory data analysis, Clustering, Multiple testing, Prediction

Rahnenfuhrer et al. *BMC Medicine* 2023;21:182

54 pages
233 refs

TG9 co-chairs:

- Jörg Rahnenfuhrer (TU Dortmund University)
- Lisa McShane (NCI)
- Riccardo De Bin (University of Oslo)

- Recommend best practices & identify common pitfalls for analysis of high-dimensional data (HDD)
- Highlight methods for three main analytic goals: class comparison, clustering/class discovery, prediction/classification
- General advice on design
- Explain

# When is plasmode simulation superior to parametric simulation for comparing classification methods on high-dimensional data?

Marieke Stolte, Nicholas Schreck, Alla Slynko, Maral Saadati, Axel Benner, Jörg Rahnenführer, Andrea Bommert, and for the topic group “High-dimensional data” (TG9) of the STRATOS initiative

- Adequate simulation of complex data structures, e.g. RNAseq data, is challenging but crucial for the evaluation and comparison of methods
- **Parametric simulation** based on only pseudo-random numbers might lead to use oversimplified data
- **Plasmode simulation** (resampling covariates + applying specified outcome-generating model (OGM)) is often claimed to produce more realistic data
- Previous papers of the authors: Discussion of Plasmode simulation (Schreck et al., 2024) and comparison of simulation types for estimating the MSE of least squares in linear regression (Stolte et al., 2024)
- This paper reports on **Comparison of parametric and Plasmode simulation** for estimating classification performance and method ranking

# Results

- Parametric simulation worse than Plasmode if DGP (data generating process) is severely misspecified
- Both simulation types affected equally by wrong OGM (outcome generating model)
- No clear recommendation regarding choice of resampling

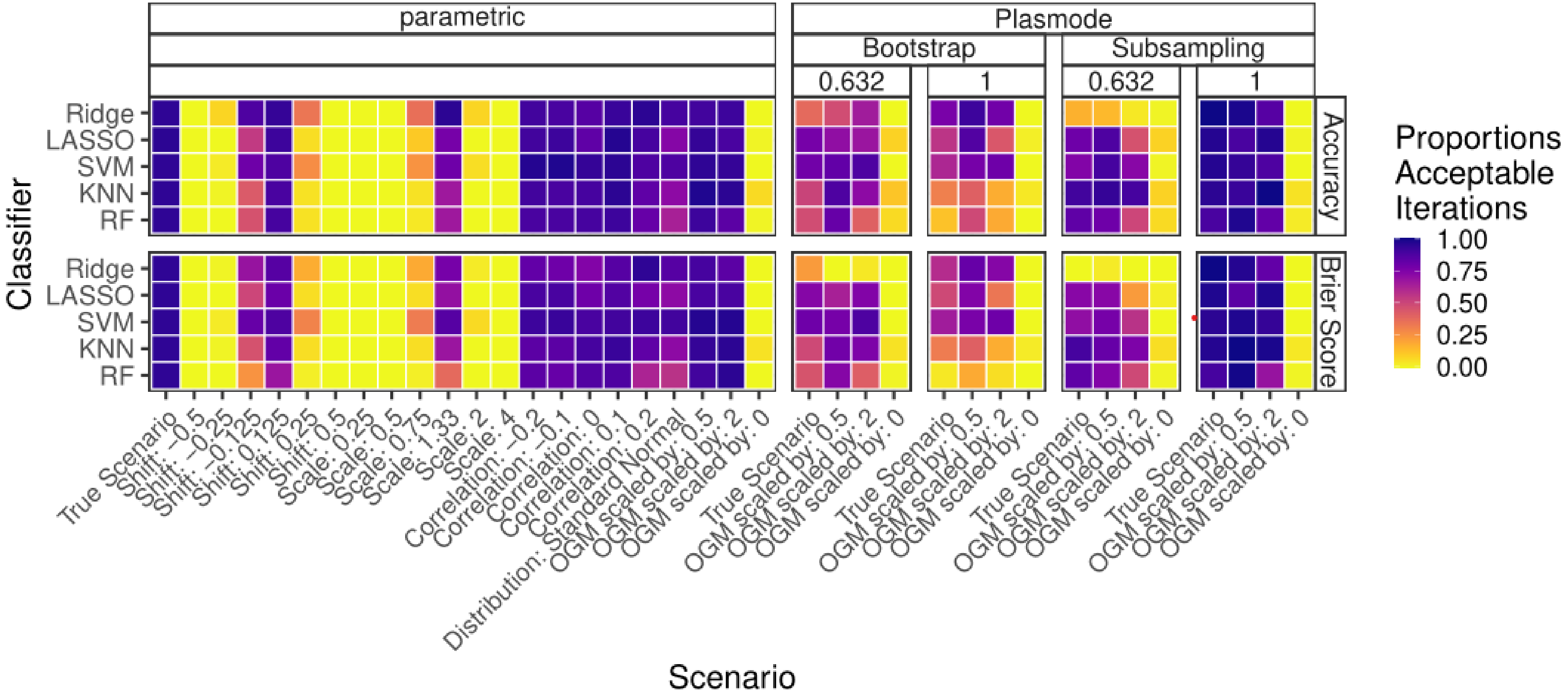


Figure: Proportion of acceptable simulation runs per performance measure for n=100 and p = 50

# A gentle introduction to tuning parameters: their role, importance, and how to choose them

Riccardo De Bin, Roman Hornung, Ilaria Gandin, Lara Lusa, and Stefan Michiels

Abstract

There is much focus in statistics on choosing the best statistical or machine learning method for data analysis. In most cases, however, getting a good result depends less on which method is used than on how well the method is configured to the problem at hand. Specifically, the successful implementation of most methods, from a simple backward elimination in multivariable regression to a complex deep learning algorithm, often depends on the choice of one or several tuning parameters. This paper is an effort made within the STRATOS Initiative (https://stratos-initiative.org/) to describe the importance, role, and influence of tuning parameters. The concepts are presented in simple terms and illustrated by examples to be understandable to the broadest possible audience. But it also touches on debatable points that may be interesting for advanced users: Should we treat the choice of the tuning parameters as a pure optimization problem, or should we use a more theoretically motivated approach? How can we compare methods fairly when their performance depends on the involved tuning parameters? How can we make studies reproducible if the choice of the tuning parameters is stochastic?

Keywords: bootstrapping; cross-validation; hyperparameters; model complexity; training-validation-test sets

# TG9 Sample Size Project

**Part 1:** Identify and summarize statistical methodological papers that that propose approaches to calculate sample size for studies using HDD

- Class comparison
- Prediction
- Clustering (?)

**Part 2:** Conduct a review of applied papers using HDD to record what method the authors used (if any) to determine sample size for this study

- Federico Ambrogi (co-chair)
- Lisa McShane (co-chair)
- Harald Binder
- Kevin Dobbin
- Stefan Michiels
- Eugenia Migliavacca
- Willi Sauerbrei
- Lara Lusa

# Planning sample size with high dimensional data

Background

A research protocol should specify the study design, including planned sample size, which will depend on the primary endpoint, analysis goal, and other key assumptions.

In high-dimensional data settings, traditional sample size calculations break down due to the large number of hypotheses tested or complex modeling or analysis strategies typically employed.

Several approaches for sample size calculation tailored to certain high-dimensional data settings have been proposed in the methodologic literature, although utility and uptake of these methods in practice has not been systematically evaluated.

It is unclear what kind of sample size justification is used by researchers dealing with high-dimensional data

... and if any justification is used at all in published research.

Goal

**Describe the methodologies most used in applied research, distinguishing between:**

Class discovery

Class prediction

**Examples when software is available**

**Recommendations for applied researchers**

**... and for future methodologic work**

# Complexity of identifying methodological studies

Hirt J, Ewald H, Briel M, Schandelmaier S. Searching a methods topic: practical challenges and implications for search design. J Clin Epidemiol. 2024 Feb;166:111201.

- Low sensitivity of traditional methods. Proposed solutions:
  - Restricting searches to journals that regularly publish methodological research
  - Use methods-specific databases
  - Use frequency-based ordering (such as applying the "best match" function in PubMed)
  - Employing machine learning (ML) tools for expansive literature results
- Challenge on the inconsistent, suboptimal, and sometimes missing indexing terms for methods research in databases (could be overcome by searching *both* MEDLINE and Embase, as they have different controlled vocabularies).
- Closely analyzing terms used in key papers, as well as collaborating with experts to identify commonly used terms can help identify all the ways used to describe methods research.
- Backward and forward citation searching for methods research

# Methodological literature search (EMBASE)

- Identified a seed sample of about 30 papers
- **SEARCH STRING: At least one element from “Study design criterion” AND at least one element from (“Study goal or method” OR “Data type”)**
- 239329 papers
- Restricted:
    - Replaced the more generic concept (and consequently the reference terms) of study design criterion with the more specific sample size determination,
    - Focused the search by searching many of the concepts as major focus or, in the free search, only in the title and author keywords fields.
- 11487 papers
- Compared the search to the seed sample.

| Study goal or method | Study design criterion | Data type |
|---|---|---|
| Artificial neural network | Brier score | Big data |
| Artificial intelligence | Calculating number of . . . | Carbohydrate array |
| Boosting | Calculation of number . . . | CGH array |
| Classification | Cases needed . . . | Comparative genomic hybridization/CGH |
| Classification and Regression Trees (CART) | Classification accuracy | ctDNA array |
| Classifier | Design of study(ies) | Deep Sequencing |
| Classifier development | Designing . . . study(ies) | DNA array |
| Cluster discovery/discover clusters/discovering clusters | Determination of number . . . | DNA sequencing |
| Deep learning | Determining number . . . | Electronic health record (EHR) |
| DESeq | Discrimination accuracy/ability | Epigenomic(s) |
| DESeq2 | Estimating number of . . . | Epigenomic(s) |
| Discriminant analysis | Estimation of number . . . | Exposome/exposomic |
| edgeR | False discovery(ies)/False discovery(ies) rate | Gene expression microarray(s) |
| Elastic net | FDR | Gene expression profiling/profiles |
| Find/finding clusters | Individuals needed . . . | Gene panel data |
| Find/finding latent class(es) | Misclassification rate | Gene sequencing |
| Find/finding structure | Number needed . . . | GeneChip |
| Find/finding subtypes | Number of cases | Genome-wide association study (GWAS) |
| Graphical models | Number of individuals | Genomic data |
| Identify cluster(s)/identifying clusters/cluster identification | Number of participants | Genomic studies |
| Identify latent class (s)/identifying latent classes/latent class identification | Number of replicates . . . | Genomic(s) |

# Methodological literature search

## EMBASE

To further refine the search, we planned to inspect in some details some of the papers to understand if they are in fact methods papers and if not, why they were selected.

A total of 180 papers (20 for each study member) were selected at random.

In average, 1-2 of the selected papers out of the 20 for each participant were classified as papers dealing as a main topic or with considerations about sample size planning. Results were reviewed by librarians that refined the string.

A new search string yielded 3951 papers

Final Check with inspection of further 20 papers for each study member.

## PUBMED

Next steps:

- Search on PUBMED
- Deduplication of the results
- Search on Scopus on mathematical/statistical journals
- A first search restricted to
  - o Mathematics - statistics & probability (321 journals)
  - o Decision making - Statistics, probability and uncertainty (188 journals partly overlapping the 321)

yielded 2019 results not retrieved in EMBASE.

# Applied literature search – strategy 1 DRAFT PLANS

| Study title | Journal | Publication Year | Number of variables | Number of subjects | Sample Size calculation (Y/N) | Based on HDD data ? | Sample Size calculation method or justification |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |

- To be included in the review, papers need at least one element from “Data type”
- The review will focus on the 15 journals with the highest impact factor in selected fields (i.e. oncology) in a defined time interval.
- The search will be extended to top 5 (impact factor) for general medicine journals (e.g., NEJM, JAMA, the Lancet, etc.) as well as others, possibly including other biology/biomedical (e.g., Nature, Nature Medicine, Cell, Cancer Cell, etc.) ...
- The percentage of studies dealing with HDD where a sample size calculation was performed will be calculated and method or justification recorded.
- This search is difficult as it involves searching for sample size justification in the article text.
- Scope might be further restricted if the number of publications identified in initial search is too large.