STRengthening Analytical Thinking for Observational Studies (STRATOS):

Overview of methodological issues when analyzing high-dimensional biomedical data

De Bin R., McShane L., Rahnenführer J. on behalf of STRATOS TG9 (2023)

Here we describe accomplishments and current activities of the STRATOS Topic Group 9 (TG9): High-dimensional data (HDD). In the last year, three new members have joined STRATOS TG9. Current members of the group are: Co-chairs Riccardo De Bin (Norway), Lisa McShane (USA) and Jörg Rahnenführer (Germany); Federico Ambrogi (Italy), Axel Benner (Germany), Harald Binder (Germany), Anne-Laure Boulesteix (Germany), Kevin Dobbin (USA), Roman Hornung (Germany), Lara Lusa (Slovenia), Stefan Michiels (France), Eugenia Migliavacca (Switzerland), Willi Sauerbrei (Germany), and Martin Treppner (Germany).

Proliferation of HDD in biomedical research has brought unprecedented opportunities to advance knowledge. In order to exploit the potential of new analysis methods to reveal useful insights from HDD, it is imperative that researchers have access to guidance on the methods available and their proper application.

In May 2023, TG9 published an extensive overview of statistical approaches for high-dimensional biomedical data in BMC Medicine. Biomedical research now frequently involves high-dimensional data generated from observational studies, controlled laboratory experiments, and clinical trials, including omics data and data from electronic health records. In the overview paper, a solid statistical foundation is provided for researchers, including statisticians and non-statisticians, who have limited familiarity with methods for HDD or simply want to better evaluate and understand the results of HDD analyses. The paper introduces basic concepts and useful strategies for design and analysis of studies involving HDD, with primary focus on omics data. For commonly used analytical methods, minimally technical descriptions are provided. The strengths and limitations of competing approaches are discussed, and some gaps in the availability of appropriate analytical methods are identified. In addition, an extensive list comprising 234 key references is provided.

The section "IDA: Initial data analysis and preprocessing" discusses the importance of initially checking the data for technical artifacts such as batch effects or inconsistent values, which can be especially challenging for HDD. Methods for preprocessing and normalization that have been specifically developed for HDD are explained. Understanding data sources and data generation methods is critical for appropriate initial data analysis and subsequent interpretation of analysis results in collaboration with scientists familiar with the often complex technical and biological aspects of the data and processes used to generate them.

The section "EDA: Exploratory data analysis" of the overview deals with methods for identifying interesting data characteristics and gaining insight into the data structure. New graphical displays to visualize data in lower dimensions have been developed for HDD, such as t-SNE and UMAP. These can capture different aspects of the data structure than a standard principal components analysis. Such methods can provide additional biological insight and help to generate scientific hypotheses.

The section "TEST: Identification of informative variables and multiple testing" reviews methods for identifying variables informative for an outcome or phenotype, for performing multiple testing, and for identifying informative groups of variables. For identifying informative variables, methods appropriate for HDD like limma and DESeq2 have been developed. Approaches such as these exploit the extremely large number of variables by sharing information across test statistics applied to single variables. Methods for low-dimensional data often rely on distributional assumptions that are not reasonable for HDD and will produce misleading results. Some classical methods to address multiplicity such as those that control family-wise error rate are often infeasible in the context of extremely large numbers of variables or tests, and alternate approaches based on criteria such as control of false discovery rate are now widely used in HDD settings. On the other hand, the large number of available variables in HDD may even be an advantage for identification of groups of variables associated with an outcome or phenotype of interest, as these groupings may reveal additional biological insights, such as associations with functional gene groups.

A popular goal in biomedical research is the construction of prediction models, but use of extremely large numbers of predictors to build these models presents multiple challenges and risks, if done inappropriately. The construction of prediction models, the assessment of their performance, and their validation are discussed in detail in the section "PRED: Prediction". New methods and software have been developed for HDD-based prediction modelling, in both statistics and machine learning fields. Unfortunately, such models are often subject to the false belief that highly accurate predictions can always be made as long as there is a sufficiently large amount of data. In the biomedical literature, one can find numerous models that have not been correctly constructed and evaluated, and their predictive quality is thus often dramatically overestimated. In the review, all steps in the model construction and evaluation process are discussed, particularly from an HDD perspective.

Beyond TG9's major goal of providing up-to-date guidance on available methods for HDD analysis, current activities include development of HDD-focused guidance for sample size calculation, influence and choice of tuning parameters, and use of plasmode data for simulations. These points are briefly explained below.

A research protocol or plan should specify the study design, including planned sample size, which will depend on the primary endpoint, analysis goal, and other key assumptions. In HDD settings, traditional sample size calculations break down, for example due to a large number of hypotheses being tested, due to complex modeling or analysis strategies employed, or due to requirements for extensive or complex assumptions that are difficult to specify or are unverifiable. Several approaches for sample size calculation tailored to HDD settings have been proposed in the literature, but utility and uptake of these methods in practice has not been systematically evaluated. Often the rationale behind the sample size is not provided for studies using HDD. TG9 has initiated an extensive literature search to identify common practices and discuss pitfalls and challenges for sample size calculations for studies involving HDD. Statistical and Machine Learning models for classification or prediction that utilize HDD often require specification or selection of hyper parameters to optimize model performance. Adjusting hyper parameters to achieve more favorable performance is called hyper parameter tuning and is an essential task for fitting many models to data. Especially for HDD, this tuning process can result in overoptimistic prediction performance estimates. TG9 is developing an overview of approaches for hyper parameter tuning along with guidance about appropriate procedures to follow to avoid overly optimistic assessments of the quality of resulting models.

Simulation studies are especially challenging for HDD, yet they are essential tools needed to perform evaluation and comparison of different methods. The typical approach for simulation studies is to use synthetic data, for which the entire true data generating process is known, which is called "parametric" simulation. Parametric simulations of HDD are usually based on overly simplistic assumptions about the high-dimensional multivariate data distribution; therefore, they produce data lacking in realistic complexity. One alternative is to perform so-called "plasmode simulations" that preserve a realistic data structure by re-sampling covariate data from real-life datasets instead of using pseudo-random numbers. This idea seems particularly promising for HDD, but its usefulness has not yet been properly evaluated in the literature. TG9 plans to pursue research on simulation strategies for HDD and to eventually provide guidance on suitable methods.

TG9 continues to monitor the landscape of new approaches for processing and analysis of HDD and to evaluate existing methods. These efforts will inform ongoing development and updating of guidance materials for both statisticians and non-statisticians.

1. Rahnenführer J., De Bin R., Benner A., Ambrogi F., Lusa L., Boulesteix A.L., Migliavacca E., Binder H., Michiels S., Sauerbrei W., McShane L., for topic group "High-dimensional data" (TG9) of the STRATOS initiative (2023): Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. BMC Medicine. DOI: https://doi.org/10.1186/s12916-023-02858-y

On 7 September, STRATOS has a satellite symposium 'Ten years of the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative – progress and looking to the future' at the CEN2023 Conference (Central European Network of IBS; https://cen2023.github.io/home/).

Everybody can join the STRATOS symposium for free. In advance, a Zoom link will be made available on the STRATOS website https://www.stratos-initiative.org/en/news

Jörg Rahnenführer will speak about the overview paper.