STRengthening Analytical Thinking for Observational Studies (STRATOS): Introducing the Initial Data Analysis Topic Group (TG3)

Saskia le Cessie[1], Carsten Oliver Schmidt[2], Lara Lusa[3], Mark Baillie[4], Marianne Huebner[5] on behalf of TG3

[1]Department of Clinical Epidemiology and Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands; Email: S.le_Cessie@lumc.nl

[2]Institute for Community Medicine, SHIP-KEF, University Medicine of Greifswald, Germany; Email: Carsten.schmidt@uni-greifswald.de

[3]Department of Mathematics, Faculty of Mathematics, Natural Sciences and Information Technology, University of Primorksa, Koper, Slovenia and Institute of Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana, Slovenia

[4] Advanced Methodology and Data Science, Clinical Development and Analytics. Novartis Pharma AG, Basel, Switzerland

[5]Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA; Email: huebner@msu.edu

The aim of the Topic Group on Initial Data Analysis (TG3) of the STRATOS initiative is to improve awareness of initial data analysis (IDA) as an important part of the research process and to provide guidance on conducting IDA in a systematic and reproducible manner. Researchers need to have a clear understanding about the underlying features and quality of their data to ensure its suitability for the intended statistical models in a statistical analysis plan. Initial Data Analysis primarily consists of all steps performed on the data of a study before the start of those statistical analyses that address research questions and are typically described in the statistical analysis plan. Ideally, IDA should already be performed during ongoing data collections. The Biometric Bulletin introduced TG3 with an overview of the framework for IDA and problems of inadequate handling of IDA in research studies [1]. We will provide here an update of our recent activities here.

Members of this topic group are Mark Baillie (Switzerland), Marianne Huebner (USA), Saskia le Cessie (Netherlands), Lara Lusa (Slovenia), Carsten O. Schmidt (Germany).

In 2018, our topic group published a conceptual framework paper where we discussed the role of IDA in the research process and identified steps in a systematic and reproducible IDA process [2]. We distinguished six steps: setting up the meta-data, data cleaning, data screening, initial reporting of the cleaning and screening findings, if needed, adapting the statistical analysis plan in a transparent way, and reporting results in research papers.

To explore current practices in performing and reporting IDA in research papers, we conducted a literature review on IDA reporting in observational studies [3]. We observed that the reporting of IDA was limited and not systematically described with IDA statements spread throughout the papers. Of the 25 reviewed papers, 40% included a statement about data cleaning, 44% provided information on item missingness and 60% on unit missingness. Based on the findings of the review, we provided a set of recommendations to improve reporting of IDA. This includes describing IDA methodology, reporting missingness, and discussing the impact of IDA findings.

The review motivated to develop a step-by step guide on systematically conducting and reporting of IDA in several examples with publicly accessible data and code. The project "Regression without regrets" is a joint project between TG3 and STRATOS topic group TG2 (Selection of variables and functional forms in multivariable analysis). The focus is to provide explanation and elaboration on conducting IDA in a reproducible manner in the context of regression analyses in a low dimensional setting (3 to 50 explanatory variables). First results have been presented in 2020 at the ISCB and MEMTAB conferences and a video poster is available [4].

Longitudinal studies add to the complexity of conducting IDA. In a joint project between TG3 and Katherine Lee from STRATOS TG1 (Missing data) we develop workflows and propose data visualizations to empower researchers to efficiently work with longitudinal data. All code and data sets for the case studies will be made publicly available. A main conclusion from the current projects is that an IDA plan is needed in advance and should accompany a statistical analysis plan. Recently the results were presented at the European Congress of Mathematics (2021).

TG3 members furthermore collaborated on the development of a framework to describe and asses data quality in observational studies along with R routines to conduct such assessments [5]. The framework distinguishes four dimensions of data quality: compliance with structural and technical requirements on the data (integrity); the availability of data values (completeness); inadmissible, impossible, or uncertain data values or combinations of data values (consistency); unexpected distributions and associations (accuracy). Each dimension forms part of a comprehensive data quality assessment workflow that distinguishes more than 30 indicators. A dedicated web page comprehensively introduces the framework and related tools to visualize distinct aspects of data quality (https://dfg-qa.ship-med.uni-greifswald.de/). Conceptually, the framework has considerable overlap with IDA steps [3] particularly concerning the meta-data, data cleaning, data screening, and initial reporting steps.

An overview of current and past activities of our Topic Group and links to materials can be found at our website https://www.stratosida.org/. More information on the activities of all topic groups is given on the central website of the STRATOS initiative http://www.stratos-initiative.org/.

References:

[1] Schmidt CO, Vach W, le Cessie S, Huebner M. STRATOS: Introducing the Initial Data Analysis Topic Group (TG3). Biometric Bulletin 2018; 35 (2): 10-11.

[2] Huebner M, le Cessie S, Schmidt C, Vach W. A contemporary

conceptual framework for initial data analysis. Obs. Studies 2018; 4: 171-192.

[3] Huebner M, Vach W, le Cessie S, Schmidt CO, Lusa, LL. Hidden analyses: a systematic review of current reporting practice of initial data analyses. BMC Med Res Meth 2020; 20 (1): 1-10.

[4] Heinze G, Huebner M, Baillie M. Video Poster Presentation at Presentation at the International Symposium of Methods for Evaluating Tests and Biomarkers (MEMTAB) 2020. Link: https://mediaspace.msu.edu/media/t/1_1n07g3j0

[5] Schmidt CO, Struckmann S, Enzenbach C, Reineke A, Stausberg J, Damerow S, Huebner M, Schmidt B, Sauerbrei W, Richter A. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Meth 2021; 21(1): 1-15.

Software Corner

R packages for selecting important interactions via regularization

Ryan A. Peterson

Department of Biostatistics and Informatics

Colorado School of Public Health

University of Colorado Anschutz Medical Campus

Have you ever presented null results to disappointed researchers, and then been asked the question "but what about interactions; are any of those significant?" I have heard this question from clinicians and researchers from many fields of science. While usually asked in earnest, **this question is a dangerous one**; the sheer number of interactions can greatly inflate the number of false discoveries in the interactions, resulting in difficult-to-interpret models with many unnecessary interactions. Still, there are times when these expeditions are necessary and fruitful. Thankfully, useful tools are now available to help with the process. This article discusses two regularization-based approaches: Group-Lasso INTERaction-NET (glinternet) and the Sparsity-Ranked Lasso (SRL). The glinternet method implements a hierarchy-preserving selection and estimation procedure, while the SRL is a hierarchy-preferring regularization method which operates under ranked sparsity principles (in short, ranked sparsity methods ensure interactions are treated more skeptically than main effects a priori).

Useful package #1: ranked sparsity methods via sparseR.

While currently in a beta-phase, the **sparseR** package has been designed to make dealing with interactions and polynomials much more analyst-friendly. Building on the **recipes** package, **sparseR** has many built-in tools to facilitate the prepping of a model matrix with interactions and polynomials; these features are presented in the package website located at https://petersonr.github.io/sparseR/. The simplest way to implement the SRL in **sparseR** is via a single call to the sparseR() function, here demonstrated with Fisher's iris data set:

```
(srl <- sparseR(Sepal.Width ~ ., data = iris, k = 1, seed = 1))
Model summary @ min CV:
------------------------------------------------------
 lasso-penalized linear regression with n=150, p=18
 (At lambda=0.0015):
   Nonzero coefficients: 10
   Cross-validation error (deviance): 0.07
   R-squared: 0.62
   Signal-to-noise ratio: 1.64
   Scale estimate (sigma): 0.267

 SR information:
                Vartype Total Selected Saturation Penalty
            Main effect    6        4      0.667    2.45
 Order 1 interaction       12       6      0.500    3.46

Model summary @ CV1se:
------------------------------------------------------
 lasso-penalized linear regression with n=150, p=18
 (At lambda=0.0070):
   Nonzero coefficients: 7
   Cross-validation error (deviance): 0.08
   R-squared: 0.57
   Signal-to-noise ratio: 1.33
   Scale estimate (sigma): 0.285

 SR information:
                Vartype Total Selected Saturation Penalty
            Main effect    6        3      0.500    2.45
 Order 1 interaction       12       4      0.333    3.46

summary(srl, at = "cv1se")

lasso-penalized linear regression with n=150, p=18
At lambda=0.0070:
------------------------------------------------
 Nonzero coefficients          :  7
 Expected nonzero coefficients :  1.38
 Average mfdr (7 features)     :  0.198

                                 Estimate        z  mfdr Selected
Species_setosa                   0.810513  17.9513 < 1e-04        *
Sepal.Length                     0.191210   9.3371 < 1e-04        *
`Petal.Length:Petal.Width`       0.119640   5.0379 < 1e-04        *
`Petal.Width:Species_versicolor` 0.275341   3.1640 0.055680       *
`Sepal.Length:Petal.Length`     -0.052711  -3.2466 0.078121       *
`Sepal.Length:Species_setosa`    0.062782   2.5978 0.251076       *
Species_versicolor              -0.001653  -0.8052 1.000000       *
```

We see (via print and summary functions) that two models are displayed by default corresponding to two "smart" choices for the