STRengthening Analytical Thinking for Observational Studies (STRATOS): Recent activities of the Topic Group on Selection of Variables and Functional Forms in Multivariable Analysis (TG2)

Georg Heinze (1), Aris Perperoglou (2), Willi Sauerbrei (3) on behalf of Topic Group 2 of the STRATOS initiative

1 Medical University of Vienna, CeMSIIS, Section for Clinical Biometrics, Spitalgasse 23, 1090 Vienna, Austria

2 School of Mathematics, Statistics and Astrophysics, Newcastle University, UK

3 Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Germany

The Biometric Bulletin introduced Topic Group (TG) 2 of the STRATOS initiative in 2018, where an overview of the problems in the practice of multivariable modelling was given [1]. Since then some time has passed, and hence we would like to provide the readership with an update on recent activities of this TG.

At present, our Topic Group consists of the regular members Georg Heinze, Aris Perperoglou, Willi Sauerbrei (co-chairs), Michal Abrahamowicz, Heiko Becher, Harald Binder, Thomas Cowling, Daniela Dunkler, Rolf Groenwold, Frank Harrell, Nadja Klein, Geraldine Rauch, Patrick Royston, and Matthias Schmid. Thomas Cowling, Rolf Groenwold and Nadja Klein joined after publication of the first report. Michael Kammer, Edwin Kipruto, Kim Luijken and Christine Wallisch have become the first members of the newly established category of 'Early Career Adjunct Members' in the STRATOS initiative and they actively participate in our bimonthly Topic Group teleconferences and work on various related projects.

In 2020, our topic group published a detailed overview of the approaches to multivariable model building, in particular procedures for variable selection and for modelling continuous variables by specifying or selecting flexible functional forms [2]. The project was led by Willi Sauerbrei and Georg Heinze. When we identified the current knowledge on multivariable model building, we had to conclude that many procedures for variable selection have been proposed, but still severe gaps exist concerning knowledge about their performance. Well-designed comparison studies of competing methods are scarce. An evidence based state-of-the-art does not yet exist, and we highlighted seven topics that need more empirical research. Some of them require further research with well-designed and conducted simulation studies, ideally neutral comparison studies. For more details on the relevance of such studies see the BB article from the simulation panel (3). Neutral simulation studies and comparative analyses of real studies are essential tools to gain knowledge about advantages and disadvantages of alternative statistical approaches, and their dependence on the data structure. Such knowledge is required to establish the state-of-the-art needed for developing relevant guidance. Our overview has already received a lot of attention and has been presented at several conferences, e.g. at the ISCB 2019 in Leuven. It has triggered several new research projects by TG2 members on some of the seven issues highlighted in our paper [2], and here we just pick two:

Aris Perperoglou and Matthias Schmid led a review study on implementations of splines in R, where we noticed that the choice of basis functions seems less relevant for the result than the choice of hyperparameters [4]. Regarding issue 2, further investigations are underway, e.g. a comparison of different default settings for hyperparameters in simulation studies that cover a wide range of scenarios likely to be encountered in practice.

Issue 5 is about the role of shrinkage to correct for bias introduced by data dependent modelling. Edwin Kipruto and Willi Sauerbrei are working on this topic, and besides analytical work they have already designed a large simulation study to compare several approaches which combine variable selection and shrinkage. At the RSS 2021 meeting Willi will give a talk about this topic.

In medical research, many data analyses are conducted by analysts with only basic statistical training. Hence, as also already identified in our paper [2], often statistical methods are employed that do not efficiently use the available data or that even lead to serious biases and mis conclusions. For example, categorization of continuous predictors is still in wide use. This type of problem may be caused by a lack of translation of methodological research into guidance for data analysts. To investigate this further, Geraldine Rauch, Christine Wallisch and colleagues are currently reviewing a series of statistical (short) papers and tutorials that were published by medical journals with respect to various aspects of multivariable modelling. The protocol of this study was recently published [5], and a publication with the results will follow soon. Geraldine Rauch will present some of their results at the ISCB 2021.

One of the gaps in knowledge translation that we identified is missing guidance on adequate methods to consider continuous covariates in multivariable models. To fill this gap, the 'Vienna branch' of STRATOS-TG2 is currently working on an interactive tool to support education on modelling with fractional polynomials or splines. Christine Wallisch will present this tool at the ISCB 2021. Rolf Groenwold has initiated a series of short videos about basic issues in statistical modelling. TG2 has completed work on two videos about categorisation and modelling of continuous variables. Some of the other TGs have also started work on short videos and STRATOS intends to release a series of videos before the end of 2021.

Our Topic Group has started a collaboration with Topic Group 3 on Initial Data Analysis, where we developed a principled

approach to data screening before a regression model is estimated. Some intermediate results of this project were presented by Georg Heinze and Marianne Huebner in 2020 at the ISCB and MEMTAB conferences.

Our Topic Group will be represented in some upcoming conferences. Besides ISCB 2021, there will be a session on Multivariable Modelling organized by Aris Perperoglou for the Meeting of the Royal Statistical Society taking place in Manchester in September featuring Willi Sauerbrei and Michal Abrahamowicz from our Topic Group. Georg Heinze, Christine Wallisch and Daniela Dunkler will give a pre-conference workshop on variable selection at the IBS Austro-Swiss Region conference in Salzburg in September 2021. The workshop is based on their review paper published in 2018 in the Biometrical Journal [6].

An overview of current and past activities of our Topic Group and further materials about multivariable modelling can be obtained at our website https://www.stratos-tg2.org.

References

[1] Perperoglou, A., Heinze, G., Sauerbrei, W. (2018) Introducing the Topic Group on Selection of Variables and Functional Forms in Multivariable Analysis (TG2). Biometric Bulletin 35(3).

[2] Sauerbrei, W., Perperoglou, A., Schmid, M., Abrahamowicz, M., Becher, H., Binder, H., Dunkler, D., Harrell Jr, F.E., Royston, P., Heinze, G. for Topic Group 2 of the STRATOS initiative (2020). State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. Diagn Progn Res 4(3). https://doi.org/10.1186/s41512-020-00074-3

[3] Boulesteix AL, Morris T, Sauerbrei W, Abrahamowicz M on behalf of the Simulation Panel (2020): STRengthening Analytical Thinking for Observational Studies (STRATOS): Introducing the Simulation Panel (SP). Biometric Bulletin; 37(2): 11-12.

[4] Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., Schmid, M. for Topic Group 2 of the STRATOS initiative (2019). A review of spline function procedures in R. BMC Med Res Methodol 19(46). https://doi.org/10.1186/s12874-019-0666-3

[5] Bach P, Wallisch C, Klein N, Hafermann L, Sauerbrei W, Steyerberg EW, Heinze G, Rauch G for Topic Group 2 of the STRATOS initiative (2020). Systematic review of education and practical guidance on regression modeling for medical researchers who lack a strong statistical background: Study protocol. PLoS ONE 15(12): e0241427. https://doi.org/10.1371/journal.pone.0241427

[6] Heinze G, Wallisch C, Dunkler D (2018). Variable selection – a review and recommendations for the practicing statistician. Biometrical Journal 60: 431-449. https://doi.org/10.1002/bimj.201700067

Software Corner: R-Ladies Global: promoting gender diversity worldwide in the R Community

Yanina Bellini Saibene and Sheila Saia

R-Ladies Global Team

Having programming skills is becoming increasingly important for many activities both in the public and private spheres, whether they are scientific, educational, academic, or commercial. It's also increasingly important to make processing and analysis as reproducible and repeatable as possible. One of the most common languages of choice for data science and statistical computing is the R language.

The R community, as other programming communities, suffers from underrepresentation of women and other minority genders in every role and area of participation, including as leaders, package developers, conference speakers, conference participants, educators, or users.

As a diversity initiative, the mission of R-Ladies Global is to achieve proportionate representation by encouraging, inspiring, and empowering people of genders currently underrepresented in the R community. The primary function of R-Ladies Global is to support underrepresented-gender R enthusiasts to achieve their programming potential, by building a collaborative global network of R leaders, mentors, learners, and developers to facilitate individual and collective progress worldwide.

R-Ladies Global received funding for the first time in 2016, from the R Consortium (a Linux Foundation Project). The organization is composed of 'chapters', groups hosting events in cities or remotely, the latter for the benefit of everyone, regardless of geographic location and personal circumstances. To date, R-Ladies Global fosters the development of 198 chapters organizing more than 3000 events in 56 countries around the world with more than 70,000 members and over 80,000 followers across the various Twitter accounts.

R-Ladies Global provides the infrastructure for the chapters to work, such as email, Meetup, GitHub, Zoom account, YouTube channel, a guide for organizers, a new chapter mentor program, and the community workspace in Slack.

In addition, R-Ladies Global conducts worldwide activities such as hosting the R-Ladies Global directory, the abstract reviewer network, the Community Slack, the rotating Twitter account @WeAreRLadies, the R-Ladies Global Blog, the campaigns for International Day of Women and meta-meetups (events involving several chapters from different countries), participation as community partners of the largest R conferences, such as useR!, RStudio::conf, and LatinR among other initiatives.