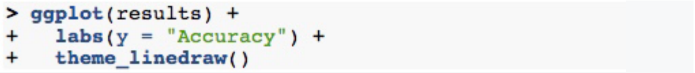
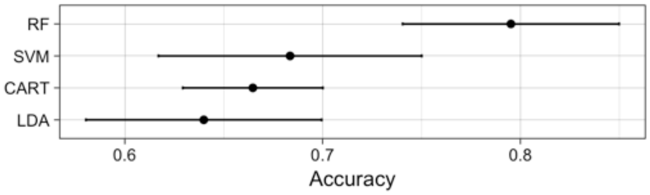
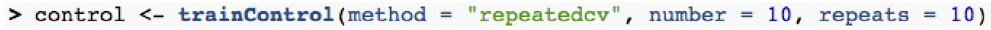
We can collect the resampling results using the resamples() function. When we apply the summary() function to the results, a table of performance measures is printed out. However, we can also visualising these collected results using the ggplot() function, which shows that the random forest tends to perform best and linear discriminant analysis performs worst in terms of overall classification accuracy.



```
> ggplot(results) +
+   labs(y = "Accuracy") +
+   theme_linedraw()
```

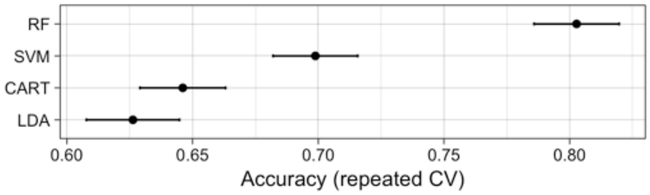


The most fantastic thing about the **caret** package, apart from the huge range of methods that it covers, is how easy it is to assess **out-of-sample** accuracy. In the above example, we have performed one round of 10 fold cross validation, however we know that results can vary from one cross validation iteration to the next, so if instead we wanted to perform repeated 10-fold cross validation, we simply update the trainControl() specification as follows:



```
> control <- trainControl(method = "repeatedcv", number = 10, repeats = 10)
```

If we ran it all again, we would see results averaged over 10 repeats of 10-fold cross validation. This is a good idea because there is inherent randomness in any round of cross validation, so performing the whole procedure a few times (with a new random split each time) and averaging over the different splits leads to more reliable out-of-sample prediction accuracies (as can be seen with the narrower error bars).



For anyone who does predictive modeling, the **caret** package is well worth investigating. It is often the case that biometricians might not know, a priori, which algorithm is going to perform best for a given problem. The caret package makes it easy to switch between different methods and compare methods in a rigorous way. It also facilitates grid search over various parameters to help identify optimal parameter choices. The author, Max Kuhn, has created extensive documentation at this website and also has a book, Applied Predictive Modeling that covers many use cases in detail.

There are a few other packages that do similar things that are also worth checking out that aim to achieve similar functionality:

- MLR: Machine Learning in R (or the future replacement mlr3)
- ClassifyR: A framework for cross-validated classification problems, with applications to differential variability and differential distribution testing.

If there's a package or software tool that makes your life easier and you'd like to share it with the IBS community, please submit an expression of interest to Garth Tarr for an article to appear in the Software Corner!

STRengthening Analytical Thinking for Observational Studies (STRATOS): Introducing the High-dimensional Data topic group (TG9)

McShane L, Rahnenführer J, on behalf of STRATOS TG9 (2019)

This article continues the series describing the STRATOS initiative and its topic groups. In previous issues the topic groups Missing Data (TG1), Measurement Error (TG4), Initial Data Analysis (TG3), Selection of Variables and Functional Forms in Multivariable Analysis (TG2) and Causal inference (TG7) were presented. In this issue, we introduce the STRATOS Topic Group 9 (TG9): High-dimensional data, and we report on current activities. Whereas TGs 1-7 started in 2013, TG9 was launched in 2016.

Members of TG9 are: Co-chairs Lisa McShane (USA) and Jörg Rahnenführer (Germany); Federico Ambrogi (Italy), Axel Benner (Germany), Harald Binder (Germany), Anne-Laure Boulesteix (Germany), Tomasz Burzykowski (Belgium), Riccardo De Bin (Norway), W. Evan Johnson (USA), Lara Lusa (Slovenia), Stefan Michiels (France), Eugenia Migliavacca (Switzerland), Sherri Rose (USA), Willi Sauerbrei (Germany).

The mission of TG9 is to provide guidance for the design, analysis and interpretation of studies involving high-dimensional biological and medical data, such as omics-data and data from electronic health record studies. TG9 will conduct an overview of existing and emerging methods pertinent to high-dimensional biomedical data settings, providing explanations, evaluations, and comparisons, based on analytical arguments and simulation studies. A collection of illustrative examples comprising such data sets together with in-depth evaluation and discussion of applicable statistical and computational approaches are being developed. The examples will be used to reinforce concepts and support specific recommendations for best practices. Didactic material including worked examples will be accumulated in a freely available resource.

Several subtopics within the overall scope of TG9 were identified, many of which overlap with other topic groups. When there is overlap, TG9 will address those aspects of each subtopic that are particularly important for high-dimensional data. Data Pre-processing is a crucial first step in the analysis, e.g., to remove noise representing experimental artifacts. Data Reduction refers to the selection of prototypic observations or variables, or the construction of new summary variables. Exploratory Data Analysis is important for quality control and for elucidating structure in the data. Multiple Testing is a frequent challenge in high-dimensional data analysis, due to the large number of variables and exploratory analyses typically performed. Prediction Modeling is a frequent objective in studies with high-dimensional data that has required novel application of existing methods from statistics and machine learning as well as development of new approaches. Increasing availability of large biomedical databases may introduce new opportunities for use of Comparative Effectiveness and Causal Inference methods which may require adaptations for high-dimensional data. Data Simulation Methods for generating data with complexity characteristic of high-dimensional data are important tools to evaluate and compare analytic methods, and to develop recommendations for analytic methods for which theoretical properties are difficult or even impossible to derive.

The group convened for two workshops with the majority of the group members attending. The workshops took place in Germany, at TU Dortmund University, March 20-23, 2018, and Ludwig-Maximilian Universität München, December 2-7, 2018. During the workshops, the group defined its strategy and prioritized its aims. Initial goals defined were preparation of three papers and development of real data examples that will be useful to illustrate various pre-processing and analysis strategies for different types of high-dimensional biomedical data. The first paper will provide a gentle introduction to common scientific goals and statistical methods at a conceptual level geared toward researchers having little experience with high-dimensional data. The second paper (or possibly set of papers) will offer basic guidance for selection of analysis approaches, with discussion of special considerations for high-dimensional data that motivate consideration of analytic tools beyond traditional statistical methods. The third paper will offer guidance on simulation of high-dimensional data.

In the introductory paper, considerations and challenges for aspects of high-dimensional data analysis including data pre-processing, data reduction, exploratory data analysis, multiple testing, and prediction modeling, will be discussed. For each aspect, examples of research questions along with commonly used analysis methods will be briefly described. Instances in which analysis methods developed for low-dimensional data are inadequate for high-dimensional settings will be highlighted, and key issues requiring further research will be discussed.

The analytical methods paper, or set of papers, will describe some basic analysis approaches that may be considered for data pre-processing and reduction, exploratory analyses such as clustering, multiple testing corrections, and prediction modeling. Each method discussed will be illustrated with real data examples that are publicly accessible and that cover a range of data types. When possible, comparisons of results obtained using multiple analysis approaches will be presented. Computer scripts will be provided to allow researchers the opportunity for hands-on practice implementing the methods.

The paper providing guidance for simulation studies will contain recommendations for planning, conducting and reporting. Simulation studies are an important tool to obtain reproducible and objective evidence about individual and comparative performance of various methods in controlled scenarios. Ideally one would evaluate methods on both real and simulated data. However, high-dimensional data are often generated to address complex research questions. Analysis methods may be correspondingly tailored. Given the wide range of specialized data and analysis approaches, often there will not be a sufficient number of data sets available for evaluation and comparison of methods for every situation. For answering complex research questions, truth may not even be known. Therefore, for high-dimensional data, a heavier reliance on simulated data may be necessary for method performance assessment.

Talks on behalf of TG9 were presented by Axel Benner at the CEN-IBS (Central European Network of the International Biometrics Society) conference (Vienna, Austria, August 2017), by Jörg Rahnenführer, Harald Binder and Axel Benner at the GMDS (German Association for Medical Informatics, Biometry and Epidemiology) conference (Oldenburg, Germany, September 2017), by Lisa McShane, Tomasz Burzykowski, Riccardo de Bin and Willi Sauerbrei at TU Dortmund University (Dortmund, Germany, March 2018) and by Lisa McShane, Willi Sauerbrei and Jörg Rahnenführer at LMU (Munich, Germany, December 2018).

Region News

Australasian Region (AR)

http://www.biometricsociety.org.au/about.html

Introducing IBS-AR President, Alan Welsh

At the Australasian Region's AGM, Alan Welsh was elected President for 2019-20. Alan writes…

I am very pleased and honoured to be the incoming President of the Australasian Region (IBS-AR). IBS-AR is very important in my professional life and I count many friends in the membership, so I look forward to this opportunity to serve IBS-AR and work with the Regional Council to contribute directly to its future development.

I would like to thank the outgoing president Samuel Müller for his leadership and the many other contributions he made to IBS-AR during his presidency. He has done excellent work with the Regional Council to develop IBS-AR and position us for the future. Of course, no President can achieve anything without the help and support of the Regional Council which works very hard for IBS-AR. I would therefore also like to acknowledge and thank the past Regional Council for their contributions.

As usual, a new IBS-AR President starts in a conference year, and planning is already well under way for our regional conference to be held in December in Adelaide. I hope many of you will be able to attend.