**Part B: Development of random forest methodology for multiple training data sets with varying covariate sets and block-structured covariate data**

Frequently, there are multiple training data sets available for prediction tasks that feature the same target variable but different sets of covariates. Moreover, the test data to which a prediction rule should be applied to in many cases does not feature all covariates that were available for training. In part B1 we treat these two issues by devising an innovative RF variant, which may particularly be applied in the context of “omics” data. In the present grant application, we use the term “omics” to denote molecular biomarkers measured through high-throughput experiments, for example, transcriptomic data, mutation data, copy number variation data, methylation data, etc. Multi-omics data, that is, data for which there are measurements of multiple types of omics biomarkers available for the same patients, are becoming more and more frequent. As a consequence these data have been the focus of much attention in the past years in the statistical learning literature, in particular with respect to prediction modeling where these data are used as covariate data in the prediction of a clinical outcome of interest. Combining these different types of omics data can lead to improved prediction accuracy. For literature on such approaches, see for example our previous work Boulesteix et al. (2017a) and references therein. The construction of prediction models based on multiple blocks (i.e. omics types) of covariates raises, however, challenges from a statistical perspective. In Parts B2 and B3, we aim at addressing these challenges by developing appropriate computational methods based on random forest methodology.

**B1: A RF variant for multiple training data sets with varying covariate sets and test data with missing covariate information**

The issue of missing values has been addressed in the context of random forest in different ways in the literature, see for example Ishwaran et al. (2008), Stekhoven & Bühlmann (2012), Tang & Ishwaran (2017) and Hapfelmeier et al. (2014) and references therein. In this part of the project, we address a very special type of missing values that is common in practice, particularly in the context of multi-omics data. Often, there exist a larger number of training data sets for a specific prediction problem that do, however, not share the same sets of covariates. The covariate matrix of the concatenation of all training data sets contains many missing values, because the values of a specific covariate are missing for each training data set for which this covariate was not made available. In principle, one may treat such a covariate matrix using regular imputation procedures and apply a prediction method to the concatenated data. However, this is not recommendable for the following reasons: First, if the sets of covariates do not strongly overlap across the training data sets, the covariate matrix from all data sets will have very many missing values making imputation techniques unreliable; another reason for the unreliability of the imputation for this application is that the imputation would be performed across different, potentially heterogeneous data sets. Second, learning prediction methods using concatenations of different data sets can lead to bad prediction rules if these data sets are too heterogeneously distributed.

A further common problem in applications is that the test data set to which the prediction rule is to be applied often features only a subset of the covariates that were available for training (and are consequently involved in the RF).

The aim of Part B1 is to adapt the RF algorithm for this context such that it can provide prediction rules which make use of all covariates (including those that are available only for a subset of the training data sets) and provide predictions for test data that do not feature all covariates available for training. In the context of omics data, such a RF variant is particularly useful, because omics data sets are usually of limited size and there are often different training data sets that feature different types of omics data. To cite a simple example, one may have gene expression data from a study ‘A’ and both gene expression data and methylation data from a study ‘B’. Combining studies A and B for training may strongly increase the performance of the resulting prediction rule in comparison to the prediction rules derived using single studies.

**B1: A RF variant for multiple training data sets with varying covariate sets and test data with missing covariate information**

Consider a collection of M data sets that feature different sets of covariates but the same response variable, where the covariate sets may or may not overlap between the data sets. In this section we will talk of “M data sets” for simplicity, but we also address the case of a single data set consisting of several clusters of patients that do not have the same covariate sets (e.g., omics blocks) available. Our idea to derive an RF making use of all data sets is briefly described in the rest of this paragraph. For i = 1,…,ntree: 1) Randomly choose a data set out of the M data sets; 2) Draw a subsample from this data set and construct a tree using this subsample, where for each split sqrt(pm) randomly selected covariates are considered, with pm denoting the number of covariates in the selected data set. It may also make sense to tune the numbers of randomly selected covariates for each split for each data set separately instead of using the default numbers sqrt(p1), …, sqrt(pM), because the values of the optimal numbers are not expected to be the same.

For prediction of test data, one then only uses the subset of covariates included in the test data that are also included in at least one of the training data set. Using this subset of covariates, prediction can be performed as follows: 1) Remove all trees from the forest which use in the first split a covariate that is not available for the new test data; 2) Prune the remaining trees in the following way: Starting from the top node, follow each branch of the tree and “cut” the branch as soon as a covariate is used for splitting that is not available for the new data; 3) Using the forest resulting in 2), obtain predictions as in the case of a standard RF.

We plan to refine and implement this basic idea. Subsequently, we will apply it in real data analyses, illustrating the broad applicability of this approach by considering versatile practically relevant settings.

Given her long-standing experience in (medical) statistics and, in particular, her knowledge of multi-omics data, Anne-Laure Boulesteix will contribute to this project with respect to the choice of realistic application scenarios and with respect to the corresponding analyses. These considerations may also affect the algorithm, where, of course, we will not include any data in the validation of the performance of the algorithm that was used during its conception.