Data Description

The data was collected as part of a clinical case-control study in the field of asthma research and contained 521 observations. The outcome was defined as the presence of asthma

The data set comprises five different data sources with 274 predictor variables in total. Those features originated from a questionnaire (n=44), clinical routine diagnostics (n=16), information on allergen sensitization (n=19), cytokine expression data (n=29), and gene expression data (n=166). We splitted gene expression data in two data sets as the feature space had a block-wise missing structure itself.

The number of observations were:

* Questionaire 521
* Clinical routine diagnostics 516
* Allergen sensitization 472
* Cytokine expression data 149
* Gene expression data l 66
* Gene expression data ll 46

The structure inversely reflected the effort of generating the data. I.e. the less observations the more valuable the measured features.

We conducted data preparation as follows: First, features (especially from the questionnaire) were selected such that they were either continuous or 0/1 distributed and they were reasonable/useful/interpretable as predictor variables. Second, all data sources were cleaned by removing variables with a high proportion of missings (>30%) and imputating missing values using the *missForest*-package. Third, data sources were combined using an unique id variable.