On the use of different strategies to combine clinical and high-throughput omics predictors in a prediction model

Riccardo De Bin

In biomedical research, the goal of a prediction model is to provide a function useful to predict a specific disease outcome. In recent years, a lot of attention has been devoted to taking advantage of the information contained in high-throughput molecular data (omics data), as evidenced by the proliferation of studies which provide prediction models based on gene expressions or other predictors derived from omics data. Nonetheless, in medical practice, several clinical predictors are often available, which have a predictive value well validated in the literature. Boulesteix & Sauerbrei (2011) and De Bin et al. (2014) discussed possible strategies to combine these two kinds of data in a prediction model. It is known, in this regard, that a combined prediction model may outperform both only clinical- and only molecular-based models. The main challenge in deriving a good combined prediction model lies in the ability of fully exploiting both data sources. Clinical data, in particular, are low-dimensional, and the information contained in them risks to get lost among the large number of high-dimensional omics predictors (Binder & Schumacher, 2008). In this talk we review the results of De Bin et al. (2014), showing how it is possible to exploit and adapt some well-known statistical methods (univariate selection, boosting, lasso, etc.) to follow different combining strategies. We also sketch a simulation study whose results may help in providing the practitioner with some guidelines on the choice of the best combining strategy to be used in specific situations (in terms of data correlation structure, relative importance of the data sources, etc.).

References

- Binder, H. & Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics* 9, 14.
- Boulesteix, A. L. & Sauerbrei, W. (2011). Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in Bioinformatics* 12, 215-229.
- De Bin, R., Sauerbrei, W. & Boulesteix, A. L. (2014). Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. *Statistics in Medicine* 33, 5310-5329.