EAACI Research Fellowship 2020
Final Report

Project title: **Application of non-linear Machine Learning techniques to understand Allergic Rhinitis**

Research Fellow: **Mario Lovric**

Type of fellowship: Research Fellowship

Host supervisor: Prof. Hans Bisgaard, MD, PhD

Location: Gentofte, Denmark

Duration: 3 months (July – October 2020)

Acknowledgements

I want to thank everybody who has given me support to realize this fellowship. I thank EAACI for providing the fellowship and making it possible in the first place. I thank Prof. Hans Bisgaard for accepting me in COPSAC (COPenhagen Prospective Studies on Asthma in Childhood) and tutoring me together with Ann-Marie Malby Schoos, MD, PhD and Morten A. Rasmussen, PhD all the way. Additionally, I want to thank every employee from COPSAC for their tutoring, the fruitful conversations, and a cozy and friendly working environment. Special thanks go to Prof. Mirjana Turkalj, MD, PhD who gave me the idea to apply for the fellowship as well as staff from the Children's Hospital Srebrnjak who supported my research and encouraged me to take this step.

What questions were addressed and why?

Allergic rhinitis (AR) is an inflammatory disease of the nasal mucosa that affects hundreds of million people worldwide. In contrast to other forms of rhinitis, AR is also associated with allergic sensitization (AS), mainly to inhaled allergens. It is a complex disease, with contributions from environmental and genetic factors. The question we have asked is: "Can the development of Allergic Rhinitis in puberty be predicted and what are the key drivers for it?"

This research question is relevant because it can help pinpoint which factors in the environment families with children at risk for developing allergic rhinitis should be cautious of, and which

factors seem to be safe. This will put down many assumptions that the general population have regarding causes and avoidance of development of allergic rhinitis, that are not evidence based.

What was the nature of the research?

This study is part of the ongoing $COPSAC_{2000}$ prospective mother-child cohort of 411 children born to mothers with asthma. The children were enrolled at age 1 month. The study excluded children born before gestational week 36 and children suffering from any respiratory disorder before enrollment. The children were followed prospectively. The 411 children born were followed from birth to the age of 18. At ages 7, 13 and 18 measurements of allergic sensitization were completed using both skin prick test and sIgE and further evaluation of allergic rhinitis to 8 different aeroallergens was performed.

A common approach in medicine is to evaluate associations between the different clinical variables against the target variables (here Allergic Rhinitis (AR) as a yes/no). The approach I had was to conduct a "data-driven" investigation on the problem, i.e. with prior hypotheses. To do this I had to conduct a thorough data cleaning and processing which I have written in open-source python code so it can be re-used and upgraded later as well. After I obtained clean data and a clear target to model with, I employed non-linear machine learning (ML) techniques to try to predict whether a child will develop AR in puberty or not.

The predictive variables were clinical, genetic, and social; determined until the children's 7^{th} year of age. From a health perspective we tried to predict the outcome in a distant future. The target variable AR is a binary variable which can be solved by multivariate binary classifiers. I have employed two non-linear classifiers, the AdaBoost and Random Forests classifiers since experiments with linear classifiers failed in the preliminary experiments due to a high complexity of the problem. The models had to be optimized by means of cross-validation to increase the predictive power of the models. Sensitivity and specificity were used as quality measures for the predictive models. Furthermore, we used balanced accuracy (mean of specificity and sensitivity) to evaluate how far the models are from random classification.

What was the result?

Approximately 200 variables were fed into the model as predictors based on patient data from their birth to the age of 6 years including parents' habits, social circumstances and IgE. The ML algorithms had a built-in variable selection step to avoid multi-collinearity between the predictive variables and good model performance. The variable selection procedure was strict, and the final models were trained and validated based on up to 10 variables. The preliminary results in prediction of Allergic rhinitis (AR) show that the models largely depend on the child's IgE towards inhalants and mother's and father's immunological profiles. This is a relevant finding as traditional approaches which are often driven by a priori hypotheses tend to include a wide spectrum of variables such as exposure variables which were not observed as important in our models. The models we have developed show a reasonable generalization with the same risk cohort which will be presented by employing validation or test sub-samples which the models haven't seen during training. Further findings will be revealed after additional model training and validations steps and hopefully published in a research paper.

How will the findings impact future research?

One of the aims of this fellowship was to establish future collaboration between COPSAC, Denmark and Children's Hospital Srebrnjak in Croatia. During the fellowship I have noticed a large intersection of the research and medical work done in both institutions driven by the same vision which is - to improve the health of children. Knowing each other's capabilities, databases, and research foci we will aim for co-applying in research projects and work further on scientific exchange.

I learned a lot about the meaning of diverse clinical and biological variables which I can apply in my future research as well. Understanding the meaning and limitations of variables in use is essential for machine learning projects since a huge amount of noisy data can distort the models and lead to irrelevant findings. The clinical knowledge obtained at COPSAC will improve my understanding in how machine learning can support clinical decision making. Furthermore, it is essential to bring communication in heterogeneous research teams to a high level of mutual understanding. My visit in COPSAC improved my understanding of medical terminology and conventions and will surely help me contribute to projects where computer science meets medicine.

It is my personal opinion that researchers who have worked with me during my stay have also profited from working together as this type of investigation is diverting from traditional paradigms

where one has a priori hypotheses, while we worked on a data driven investigation and used all data at hand independent of known or assumed associations.

This fellowship was also an important step in my training to develop my researcher skills, I gained further practice in preparing clinical data and optimized the model training procedure I have built before. Moreover, I had the opportunity to spend three months participating in many scientific meetings which made me broaden my perspective in medical research.

Adaption of the original idea

The original idea (Application of Novel Non-Linear Dimensionality Reduction (NLDR) Techniques to Pattern Recognition in Childhood Asthma) was to find common patterns/phenotypes between Croatian and Danish Asthma cohorts. However, it turned out that the cohorts are different in their definitions, i.e. the Croatian cohort is an observational clinical cohort, and the Danish COPSAC 2000 cohort is a birth cohort. Therefore, we adapted the research course by applying the proposed methods to understand the development of allergic rhinitis in the Danish $COPSAC_{2000}$ cohort.