Project Description

The goal of this project is to find a methodology that can best discriminate between participants with and without cancer based on exposure to certain chemicals found in pesticides and herbicides.

What motivated me to do this project was that to the best of our knowledge there hasn't been a scientific review of different methodologies to assess the relationship between exposure to the chemicals present in this dataset and Non-Hodgkin Lymphoma.

Previous work has been done detecting single analytes (chemicals) at a time using a "relationship between the analyte and the disease risk modeled..." . However, work has not been done utilizing chemical mixtures (multiple chemicals at a time), to discriminate between positive and negative cancer results in a case-control study as well.

Data Sources

The analytic dataset is from the study "Comparison of pesticide levels in carpet dust and self-reported pest treatment practices in four US sites" (Colt, et.al, 2004), a Non-Hodgkin Lymphoma case-control study in Detroit MI, the state of Iowa, Los Angeles, CA, and Seattle, WA, four areas covered by the Surveillance, Epidemiology, and End Results(SEER) Program of the National Cancer Institute.

Cases and controls were matched on age, sex, race, and area. Participants answered a questionnaire and provided dust samples from their vacuum cleaner, provided they had used their vacuum cleaner in the past year and had owned at least half of their carpets or rugs for at least 5 years.

Laboratory measurements of the 30 analytes contained data subject to limit of detection, and as such missing values were imputed by assigning a value for each missing observation by selecting a value from the assumed distribution using maximum likelihood parameter estimates (Helsel 1990, Moschandreas et al 2001). This imputation was done 10 times, forthisproject I am using the first iteration of the imputation.

Expected challenges with these data include wrangling the data in the dataset, and working with imputed data. Exploratory data analyses show large amounts of collinearity between analytes (which is to be expected) as well as some heavy skewing of data points, warranting a log-transformation (log_{10}) of the analytes. Another challenge is finding a document that presents units of measure and coding on the levels of categorical variables(an example being Locations 1,2, and 3 instead of the site names).

Methods, Techniques and Tools

The analytic dataset is already imputed using the techniques outlined above. Statistical methods for this project includeLogistic Regression, LDA, LASSO, Neural Networks, Conditional Decision Trees, and Random Forests. Each of these methods will be implemented and evaluated using cross validation. Comparisons among the proposed methodologies will be done via cross validated metrics of sensitivity, specificity, PPV, and NPV. Risk Prediction will be explored as well if the models do a poor job of discrimination, but only statisical methods will be appilciable.

The project will be done primarily in R and Rmarkdown as the language of choice. Numerous different libraries will be used for each model of the project as well. All of the work will be saved and shared on my Github repository so that work can be backed-up as well as tracked. I plan to make my work publicly available at the end of the project so that others can use code I have written and make comments and/or advice.

I also plan to make library files for use across the project to keep my code organized and concise. These will range from simple helper methods to extensive model building methods.