# Applied Multivariate Data Analysis Project Proposal

Emrecan Ozdogan and Kyle Naddeo

March 29, 2021

In this project we will use data available at this website. The data is acquired from 315 patients who had elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous. Data consist of 14 variables:

1. Age
2. Sex
3. Smoking status
4. Quetelet
5. Vitamin use
6. Daily calorie consumption
7. Daily fiber consumption
8. Number of weekly consumed alcoholic drinks
9. Daily cholesterol consumed
10. Daily beta carotene consumed
11. Daily retinol consumed
12. Plasma beta carotene level
13. Plasma retinol level
14. Plasma concentrations

The last two variables (13 and 14), have good variability from subject to subject and are the main targets. We will be examining their relations to the remaining 12 variables in two main groups, physical and dietary. Principal Component Analysis (PCA) and Factor Analysis (FA) will be used to find better/simpler representation of the data. Then, Linear Discriminant Analysis and Quadratic Discriminant Analysis will be used to predict plasma levels based on other 12 variables. Prediction output will be an interval since plasma retinol and plasma beta carotene levels vary between 50 and 1500. Length of the interval will be determined during the simulations according to prediction accuracy.

Plasma beta carotene and plasma retinol levels can be associated with developing certain types of cancer. Finding relation between physical characteristics and dietary habits of the person and plasma levels can be used to design diets to reduce the possibility for people who are at risk of cancer. It can also be used as a precaution to mitigate the chances of developing cancer.