Cluster analysis of multi-morbidity in UK Biobank participants

Data Analysis Plan

Introduction

This data analysis plan describes the aims, methodology **and outcomes** that will be used for the cluster analysis of multi-morbidity in UK Biobank participants. These analyses will be carried out by the Barracudas (MSc Health Data Analytics and Machine Learning students).

Aims

The overall objective of this study is to identify subsets of individuals suffering from multi-morbidity who share common environmental and/or biological pathways. To do this we are going to analyse UK Biobank participants.

* We will identify sub-groups of individuals suffering from multi-morbidity
* We will compare cluster analysis methods
* We will attempt to identify possible shared and distinct underlying physio-pathological processes between clusters

Dataset

UK Biobank data.

UK Biobank is a prospective study investigating the contributions of genetic predisposition, and lifestyle and environmental exposures to the development of disease in 500 000 people aged 40-69 in the UK.

The dataset contains numerical, binary and categorical variables.

Using the six disease variables present in the dataset, there are 61 000 UK Biobank participants with multimorbidity.

Exploratory Data Analysis

Dataset overview (Josh-Table 1)

* Numbers of diseases across important epidemiological factors e.g. age, sex, smoking, alcohol

Distributions of diseases across important epidemiological factors