Poster submission to C@UCA, 6-8 juin 2017 Fréjus (France)
https://cauca2017.sciencesconf.org

Statistical learning on large scale biomedical databases to improve diagnosis, follow-up, and treatment of neurological disorders: an MNC3 project.

Luigi Antelmi, Marco Lorenzi, Valeria Manera, Philippe Robert, Nicholas Ayache.

Recently available large-scale health databases (DBs) are a promising resource for the scientific community to advance our understanding of diseases such as neurological disorders. By analysing this kind of heterogeneous information we aim at developing new markers improving the identification of pathological traits in individuals.

In the MNC3 project *(Médecine Numérique: Cerveau, Cognition, Comportement)* of UCA[Jedi] we will investigate different biomedical DBs, such as the UK Biobank (UBK), ADNI, INSIGHT, MEMENTO, carrying extensive information including data from questionnaires, physical measures, sample assays, accelerometry, multimodal imaging, genome-wide genotype, and longitudinal follow-up for a wide range of health related outcomes. For instance, the UKB is a prospective study with over 500,000 participants aged 40–69 years recruited among the general UK population between 2006 and 2010. The current snapshot of the UKB DB consists in ~20 GB of plain text CSV data (10581 variables for 502639 participants), ~3 TB of genetic data (results available for 150000 participants), and ~24 TB of imaging data (imaging available for 5724 participants).

The management and analysis of this kind of large-scale heterogeneous information must address challenges targeting different disciplines: IT (how to manage the data?), statistics (how to analyze the data?), biomedical (clinical question to be addressed?).

The goal of this project is to leverage on advanced statistical techniques for improving our understanding of neurological disorders, and for leading to better diagnostic, follow-up, and treatment instruments. To this end we focus on the development of multivariate statistical methods able to integrate the heterogeneous data and to scale to massive sizes, such as the one of the UKB DB. The methods should also be robust enough to work in the presence of sparse information, as the amount of missing data is usually very large.

Preliminary results demonstrate that with the current technology we are able to manage and query this kind of large-scale data for the extraction of specific clinical information. For example, **Fig. 1** shows that the UKB cohort is characterized by 21 mm^3 per year volume loss in the hippocampus, a brain region heavily affected by brain aging and involved in several neurological disorders.

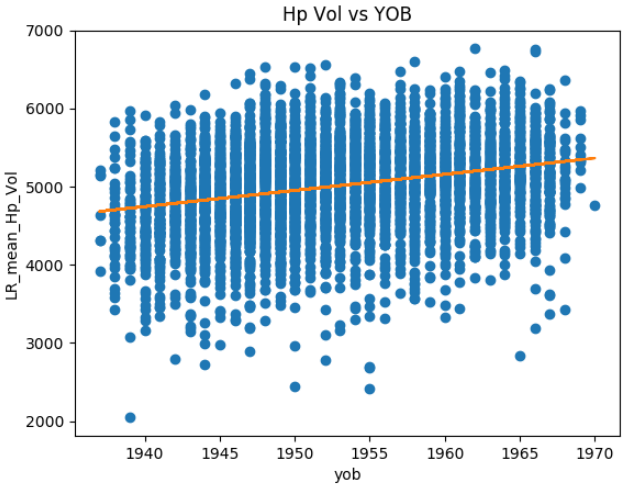

Fig. 1: Linear relation between year of birth (yob) and mean hippocampal volume in the UK-Biobank population who underwent an MRI scan (N=5724). Slope = 20.67 mm^3 / Year.

In the next steps we will model a large panel of variables and identify optimal subspaces where covariances among biomarkers is maximized, so that the contribution of each variable can be exploited at their full diagnostic value. Moreover we will explore unsupervised clustering methods for the automatic stratification of the population by diagnostic subcategories.

This project has the potential of finding new and efficient ways to diagnose, follow-up, and treat neurological diseases, such as Alzheimer's Disease, the most common cause of dementia with the highest burden on the society.

Web Site of MNC3 project: http://www.innovation-alzheimer.fr/mnc3-2/