Multiple imputation of the big register data

Obstructive sleep apnoea (OSA) and type 2 diabetes (T2D) are commonly co-occurring conditions that significantly increase the risk of cardiovascular consequences such as myocardial infarction, heart failure, and/or death. Despite this, many patients with T2D remain underdiagnosed for OSA, leading to missed treatment opportunities. This is partly due to uncertainty about the long-term health effects of OSA treatment, especially in patients with T2D. Since these questions are challenging to address through randomized clinical trials, where the effects take a long time to materialize, alternative methods are required.

To address these knowledge gaps, we plan to use longitudinal registry data, also known as "real-life data". These data are very big: most datasets contain millions of observations and occupy tens or even hundreds of gigabytes of memory, and contain missing data. The purpose of this project is to develop a statistically motivated methodology for imputing these big data based on Multiple Imputation for Chained Equations (MICE) [1].

Description of Datasets

The data for the project is derived from Swedish national health registries, which offer extensive longitudinal information on diagnoses, treatments and outcomes. The main datasets include:

- **National Diabetes Register (NDR):** Covers the majority of patients with T2D in Sweden and contains detailed information on metabolic control, treatment, and outcomes, with multiple data points per individual over time.
- **Swedish Sleep Apnoea Registry (SESAR):** Provides information on OSA patients, including diagnostic parameters and treatment decisions, but has limited longitudinal follow-up.
- **Cause of Death Register:** Nearly complete coverage of causes of death, though naturally lacking longitudinal follow-up.

Research Questions

1. Which modifications are necessary in the MICE framework to implement it in the Big Data context using Apache Spark and MLLib?
2. What is the quality of prediction of the MICE framework when applied to NDR, SESAR and Cause of Death Register?
3. How does the runtime of the method depend on the number of observations?
4. How does the predictive uncertainty of the values imputed in NDR, SESAR and Cause of Death Register depend on the number of cases included in the analysis?

Prerequisites

Good knowledge of statistics, machine learning and programming. Experience with Big Data analytics in Apache Spark.

Contacts

- The main supervisor for the project is Oleg Sysoev oleg.sysoev@liu.se
- The medical expertise is provided by Fredrik Iredahl, Associate Professor and General Practitioner, and Jonas Agholme, PhD student and Specialist in Internal Medicine.

[1] White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine, 30(4), 377-399.