MASTER THESIS PROJECT

Hacking Women's Stroke study – Big Data for Precision Stroke Prevention

Drs. Hendrikus van Os, afdeling Neurologie, LUMC
Dr. Matthijs van Leeuwen, Leiden Institute of Advanced Computer Science (LIACS)
Dr. Mark Hoogendoorn, Department of Computer Science, Vrije Universiteit (VU)
Prof. Dr. Mathijs Numans, afdeling Public Health & Eerstelijns Geneeskunde, LUMC
Prof. Dr. Marieke Wermer, afdeling Neurologie, LUMC

Background

Stroke is one of the leading causes of disability and death in the Netherlands, with an incidence of around 3 per 1000 person-years. [1] In the Netherlands, general practitioners (GPs) currently stratify their patients into risk groups using a simplistic algorithm containing only five traditional risk factors.[2] In women however, evidence on the importance of 'women-specific' risk factors such as migraine, coagulation disturbances is mounting.[3] These factors may work synergistically with each other and with traditional risk factors. Currently, none of these women-specific factors are taken into account for stroke prevention, possibly withholding preventive measures from tens of thousands of women at risk. The advent of artificial intelligence in medicine enables us to create a personalized risk model capturing these recently discovered women-specific risk patterns, and also yet undiscovered patterns. This methodology has recently been proven successful in this type of data environment by our own research group.[4] Within 2.5 years we aim to implement automated software in general practitioner practices to help identify women at increased risk for stroke in time.

Objectives

- What is the best model (machine learning or traditional [logistic or Cox regression]) and best prediction setting (survival vs. non-survival) to predict 10-year risk for stroke?
- Can we extract new information from machine learning models? Can we find patterns in our data that could be the starting point of new, focused research? This underscores the importance of explainable data science, which is essential in medical research.

Data description

We will analyse data from 2 million general practitioner patients with a mean follow-up of 10 years and an estimated incidence of 65.000 new stroke cases over the course of follow-up. We will work with several thousands of potential risk factors (e.g. from free text, medical history, medication use, vital parameters), underscoring the importance of a sound feature selection strategy. Data are readily available.

Student profile

Good knowledge and understanding of machine learning problems and algorithms; programming skills in Python; interest in applying data science in the medical domain.

References

1. Vaartjes I BM, Poos MJJC. Hoe vaak komt een beroerte voor en hoeveel mensen sterven eraan? In: Volksgezondheid toekomst verkenning, nationaal kompas volksgezondheid (2007).
2. NHG. Nhg-standaard cardiovasculair risicomanagement. *(accessed December 20th 2017)*. 2017
3. Schurks M, Rist PM, Bigal ME, Buring JE, Lipton RB, Kurth T. Migraine and cardiovascular disease: Systematic review and meta-analysis. *BMJ*. 2009;339:b3914
4. Kop R, Hoogendoorn M, Teije AT, Buchner FL, Slottje P, Moons LM, et al. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. *Comput. Biol. Med.* 2016;76:30-38