Discontinuation of preventive cardiovascular medication in primary care patients: using Natural Language Processing to disclose medical information in free text for epidemiological analysis

More than 75% of Dutch citizens aged 70 or older use preventive cardiovascular medication (PCVM), even though the benefit- harm balance becomes less favourable with age. The national guideline considers side effects and limited health expectancy valid reasons to discontinue PCVM. The current practice of PCVM discontinuation in primary care may be different. We want to study which older patients stop PCVM and why with routinely collected, electronic patient data from more than 30 general practices. Coded information about diagnoses, medication and laboratory tests is easily accessible and already used in epidemiological research. However, the information in the narrative reports of general practitioners is not easily retrievable. It contains personal and contextual reasons for stopping PCVM, and relevant symptoms and diagnoses that have not been coded. The information would complement the structured data and enhance accuracy of the statistical analyses. The aim of this project is to make the unstructured data available for quantitative research with Natural Language Processing. A data scientist will clean the data, create word embeddings, and classify the reasons for stopping PCVM. Internal validity of the retrieved data will be assessed against gold standard manual extraction, and external validity with data unseen until then. The script and expertise obtained in the current project will be useful for other research in our department as well. In the future, the department would like to add machine learning to find yet unknown factors in the (un)structured data that are relevant to early detection of disease, disease progression, and therapy compliance.