



Using high-volume unstructured GP notes to predict stroke

Anneloes Louwe, Master's Thesis Project

Supervision:

- Hine van Os, dept. Neurology & Epidemiology, LUMC
- Suzan Verberne, Text Mining & Information Retrieval, LIACS

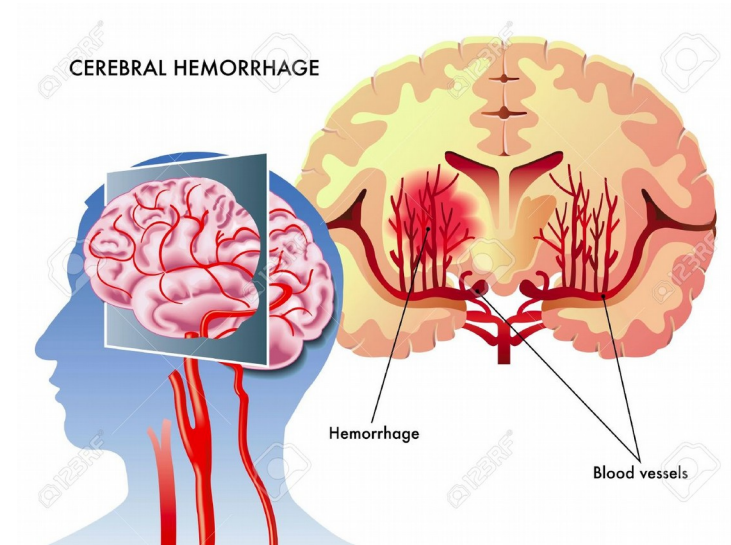
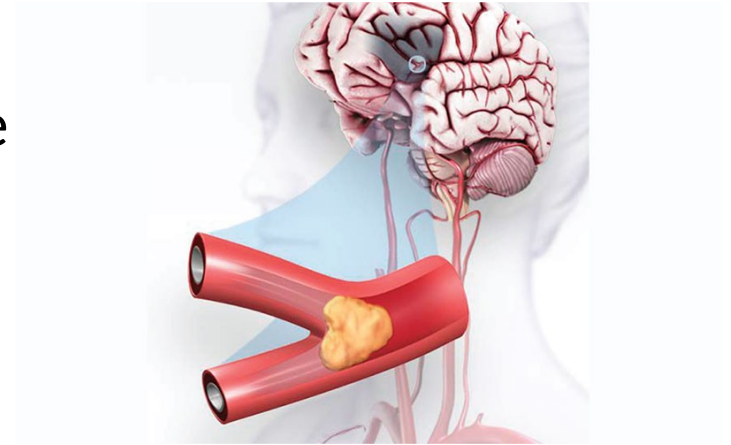


Contents

- Study context and objectives
- Preprocessing of primary care consultation notes
 - Cleaning and tokenization
 - Spelling correction
 - Keyphrase detection
- Feature selection
 - Bag-of-words
 - Topic modeling
- Prediction models

What is stroke?

- Brain infarctions & brain hemorrhage
- NL: 43.000 strokes per year
- 3rd cause of death



Prevention of stroke is key

- Prevention by general practitioner
 - Blood pressure & cholesterol medication
 - Lifestyle change
- Simplistic risk chart, only 5 risk factors
- Need for precision prevention (and thus prediction)!

		Vrouwen									
SBD		Niet-rookster					Rookster				
180	35	38	41	43	44	47	50	>50	>50	>50	
160	28	31	33	35	36	38	41	44	46	48	
140	22	24	26	28	29	31	33	36	38	39	
120	18	19	21	22	23	25	27	29	30	32	
180	14	17	20	24	30	27	32	37	45	>50	
160	10	12	14	17	21	19	22	27	32	39	
140	7	8	10	12	15	14	16	19	23	28	
120	5	6	7	9	11	10	11	14	17	20	
180	10	12	15	18	23	20	23	28	34	42	
160	7	8	11	13	16	14	17	20	24	30	
140	5	6	7	9	12	10	12	14	17	21	
120	4	4	5	7	8	7	8	10	12	15	
180	5	6	8	10	12	10	12	15	18	22	
160	4	4	5	7	9	7	8	10	13	16	
140	3	3	4	5	6	5	6	7	9	11	
120	2	2	3	3	4	4	4	5	6	8	
180	2	3	4	5	6	5	6	7	9	11	
160	2	3	3	3	4	3	4	5	6	8	
140	1	1	2	2	3	2	3	3	4	6	
120	1	1	1	2	2	2	2	2	3	4	
180	1	1	1	1	1	1	1	1	2	2	
160	<1	<1	1	1	1	1	1	1	1	2	
140	<1	<1	<1	1	1	<1	<1	1	1	1	
120	<1	<1	<1	<1	<1	<1	<1	1	1	1	
	4	5	6	7	8	4	5	6	7	8	

Ratio totaal cholesterol/HDL

Aim

- Including free text in a prediction model for stroke
- Identification of novel (women-specific) risk factors

Free text

- Captures patients' narrative
- Supporting evidence
- Uncertainty
- Non-medical information (eg. social problems)

- Diagnosis Descriptions
- SOAP notes
 - S: Subjective
 - O: Objective
 - A: Assessment
 - P: Plan

Data overview

- Pipeline development: ELAN dataset (n = 87000)
- Proof of concept: NEO dataset (n ≈ 6000)
 - Cases (including heart infarctions): 182
 - Controls: 5890
- Main dataset: STIZON dataset (n = 3000000)

Preparation

- ICPC code (re)formatting (e.g. K90.00)
- Grouping SOAP lines

Cleaning and tokenization

- Lowercasing and punctuation removal
- Token removal: Stopwords, numbers, short words, medication specifications (e.g. *100mg* or *100st*), *zorgdomein* codes

Spelling Correction

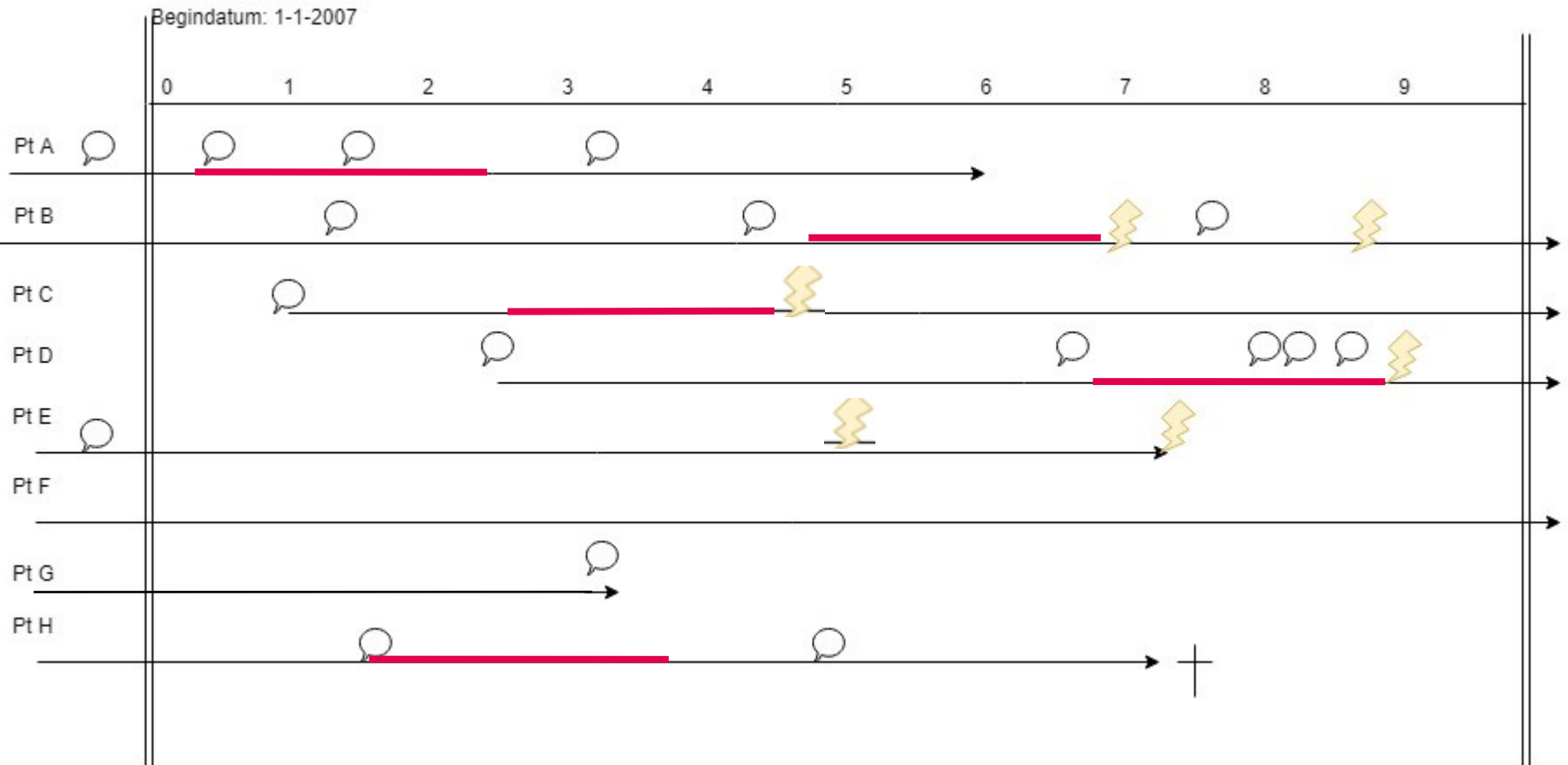
- Vocabulary: Clinspell, ICPC definitions and CoNLL
- Single-character edit identification using Symmetric Delete

Keyphrase Detection

- Kullback–Leibler divergence

$$D_{\text{KL}}(P \parallel Q) = - \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{Q(x)}{P(x)} \right)$$

Cases vs. controls



Feature Selection

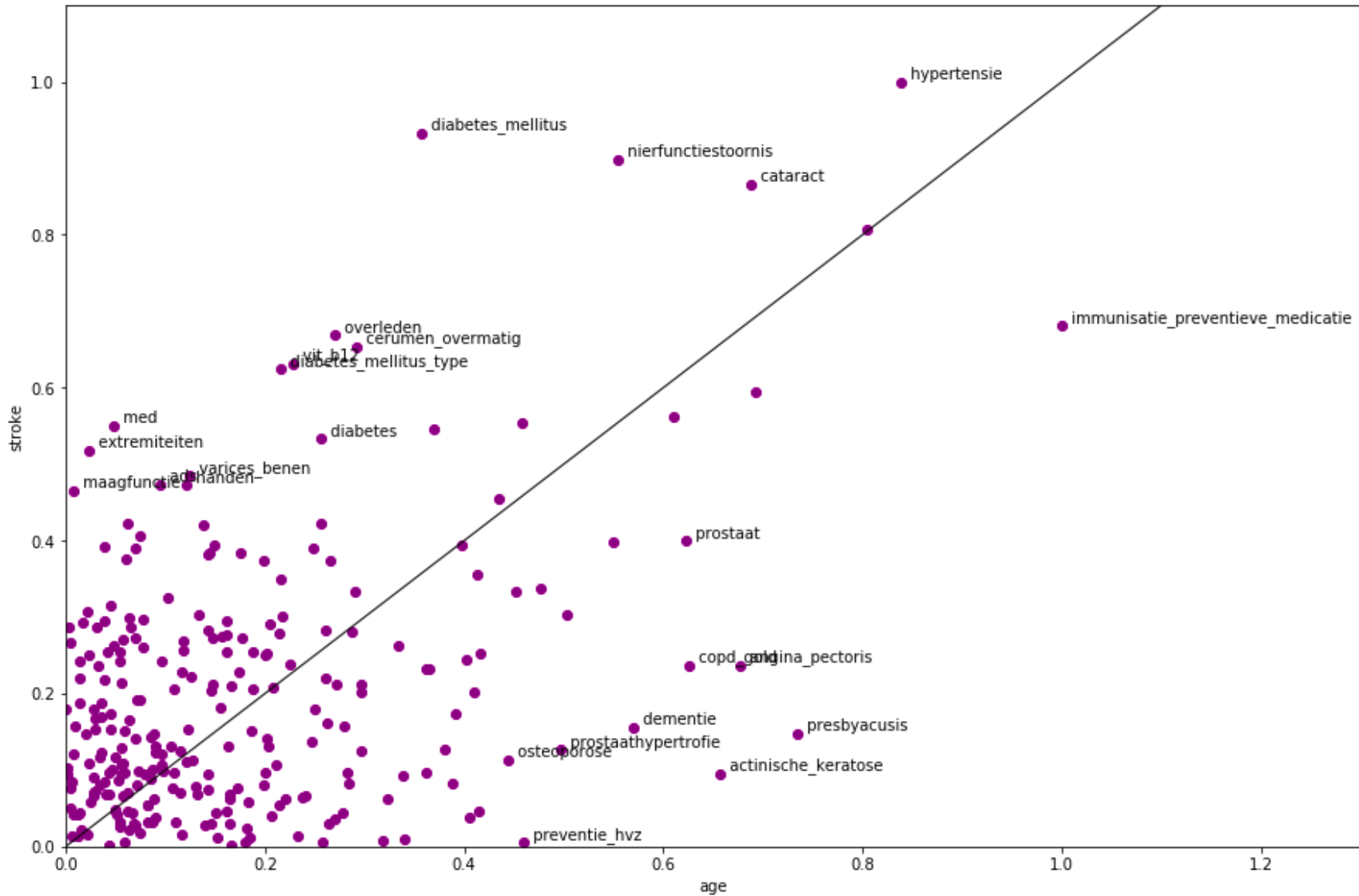
- Unified Medical Language System (ULMS): Medical Concept Extraction
- Bag-of-Words
- Topic Modeling
 - Latent Dirichlet Allocation (LDA)
 - Non-negative Matrix Factorization (NMF)
 - Topic Coherence: Word Embedding model (Word2Vec)

$$\begin{matrix} W \\ \left[\begin{array}{|c|c|c|} \hline & & \\ \hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array} \right] \end{matrix} \times \begin{matrix} H \\ \left[\begin{array}{|c|c|c|c|c|c|} \hline & & & & & \\ \hline & & & & & \\ \hline & & & & & \\ \hline & & & & & \\ \hline \end{array} \right] \end{matrix} \approx \begin{matrix} V \\ \left[\begin{array}{|c|c|c|c|c|} \hline & & & & \\ \hline & & & & \\ \hline & & & & \\ \hline & & & & \\ \hline \end{array} \right] \end{matrix}$$

Models

- Logistic Regression
- Random Forest

Models



Next steps

- STIZON dataset
 - Experimentation
 - Pipeline optimization
- Negation Detection

Thank you!

LUMC Neurologie

- Hendrikus J. H. van Os
- Marieke J. H. Wermer

LUMC PHEG

- Mattijs A. Numans
- Tobias N. Bonten
- Niels H. Chavannes
- Rolf H. H. Groenwold
- Janet Kist
- Michiel Meulenbroek
- Frederike Buechner

Vrije Universiteit

- Mark Hoogendoorn
- Ioannis Pantazis

LIACS

- Matthijs de Leeuw
- Suzan Verberne
- Teddy Etoeharnowo
- Anneloes Louwe

LUMC Statistiek

- Hein Putter
- Erik van Zwet

Turku University (Finland)

- Sepinoud Azimi

