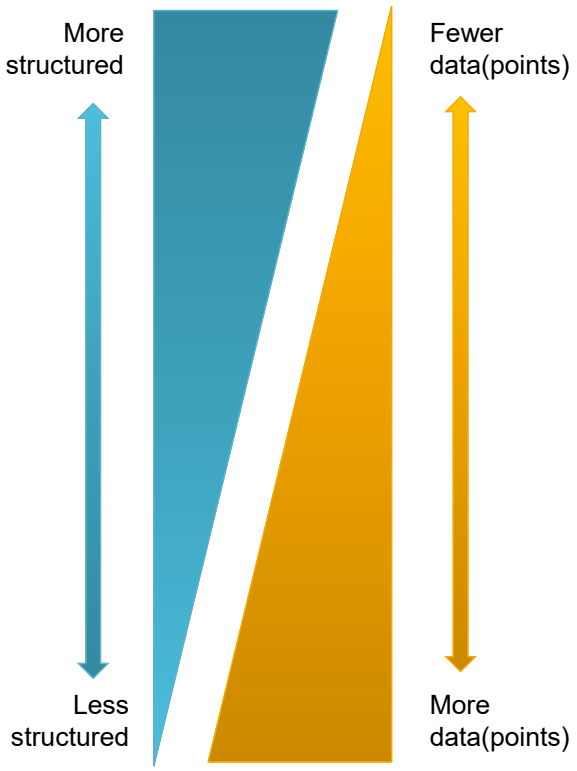
Marijn van Wingerden, assistant professor Computational Psychiatry

Machine Learning approaches to Health Data

Mannheim Center for Data Science
Lecture series "Data Science in Action"

Health Data – rich datasets, underused: data/structure tradeoff

- Health data == wealth of data
  - Electronic Health Records
  - Clinical measurements
  - Patient-reported outcome measures (PROMs)
  - Interactions with healthcare providers (text, speech)
  - Sensoring (activity, smartphone, video/audio,…)



Machine Learning with Health Data

WHY IS IT HARD

- Requires strong domain knowledge to understand
  - "reasonable" values for measurements
- Mix of objective measurements and *objectivistic* coded opinions
  - All kinds of bias
- Missing data
  - MAR vs. MNAR
- Distribution shifts
  - Covid-19

WHY SHOULD YOU DO IT

- The human gold standard is an overworked, underpaid nurse (…, …, …)
  - Decent room for improvement
  - Changing economy etc.
  - Administrative load
- Combinatorial evidence is hard to gauge as a single professional in a web
  - We are looking for high-dimensional outliers
- Decision support mechanisms have very real impact

Today's talk

TOPICS

- Psychiatric symptom clustering
- Longitudinal modelling of PROMs
- Predicting Falls in Elderly Care
  - Work in progress

INTERACTION

- Feel free to interrupt at any moment with questions ☺

Contact: e.j.m.vanwingerden@tilburguniversity.edu

Psychiatric symptom clustering

In collaboration with Erasmus Medical Center, Rotterdam and the Rotterdam Study

Psychiatric symptom clustering

THE PROBLEM

- Psychiatric symptoms are endorsed or assigned
  - DSM-5, what are they really?
  - Depression: 5 or more for 2 weeks
  - Many ways of being depressed

Table 1. Summary of *DSM-5* Diagnostic Criterion A for Major Depression: Symptoms[12]

1. **Depressed mood** (ie, "sad," "empty," "hopeless") for most of the day, nearly every day. Can be from self-reports or reports of others.*
2. **Decreased interest or pleasure** in most activities, most of the day, nearly every day. Can be from self-reports or reports of others.*
3. **Body weight change** (increase or decrease) of > 5% when not dieting.
4. **Insomnia or hypersomnia** nearly every day.
5. **Psychomotor agitation, restlessness, or slowing** of physical movement that is apparent on observation.
6. **Fatigue or loss of energy,** nearly every day.
7. **Feelings of worthlessness or guilt** that is excessive or inappropriate, nearly every day.
8. **Diminished ability to think or concentrate, or indecisiveness,** nearly every day. Can be from self-reports or reports of others.
9. **Recurrent thoughts of death and/or suicide**, with or without a specific plan.

*For diagnosis of major depressive disorder, at least 1 of these symptoms is required. These symptoms must last most of the day, nearly every day, for a minimum of 2 weeks.

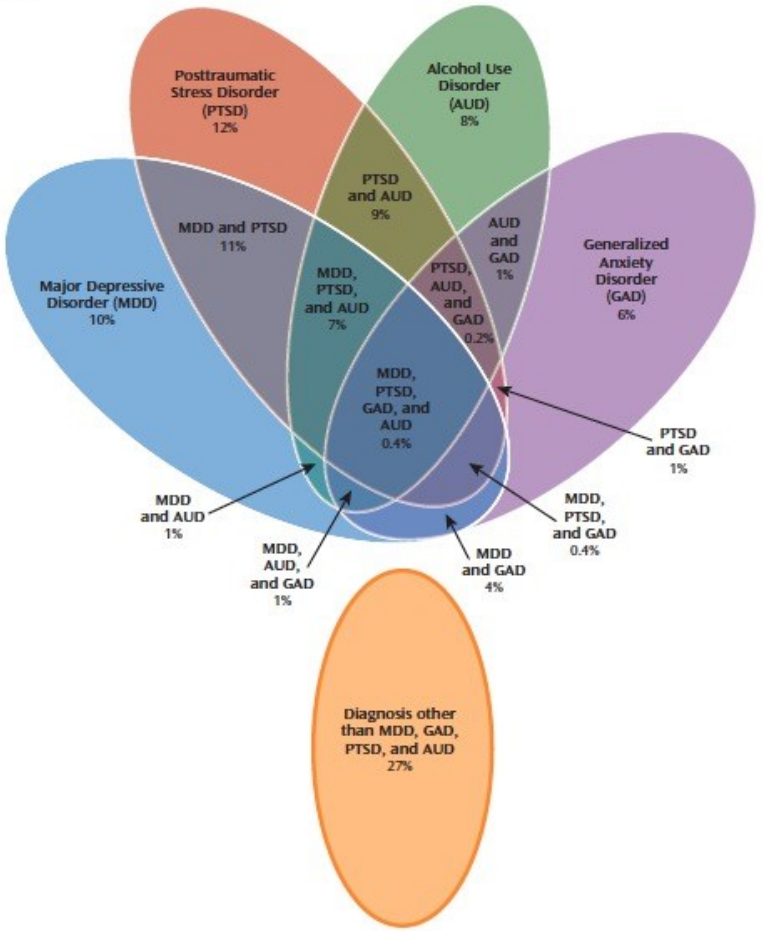
Abbreviation: DSM-5, *Diagnostic and Statistical Manual of Mental Disorders,* Fifth Edition.

Psychiatric symptom clustering

THE PROBLEM

- Psychiatric symptoms are endorsed or assigned
- High co-morbidity
  - Symptoms shared between diagnoses (e.g. Anx/Dep)
  - Treatments may differ with respect to diagnostic boundaries

FIGURE 1. Comorbidity of Major Depressive Disorder, Posttraumatic Stress Disorder, Alcohol Use Disorder, and Generalized Anxiety Disorder[a]



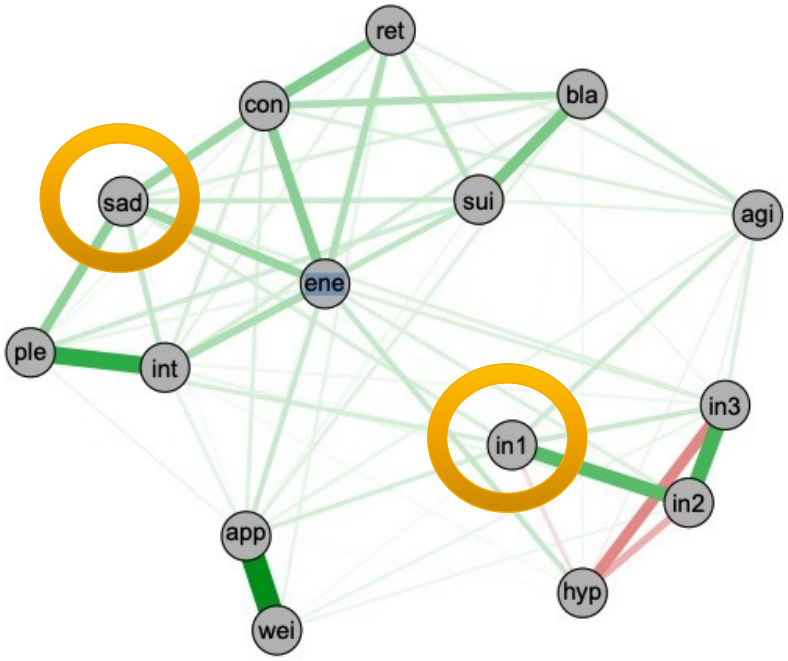
[a] Rates are average weighted percentages from Houston VA/Menninger (N=264).

Regier, Darrel & Narrow, William & Clarke, Diana & Kraemer, Helena & Kuramoto, Satoko & Kuhl, Emily & Kupfer, David. (2012). DSM-5 Field Trials in the United States and Canada, Part II: Test-Retest Reliability of Selected Categorical Diagnoses. The American journal of psychiatry. 170. 10.1176/appi.ajp.2012.12070999.

Psychiatric symptom clustering

THE PROBLEM

- Psychiatric symptoms are endorsed or assigned
- High co-morbidity
- Sub-clinical presentation?
  - Relation between sleep and mood even when not clinically depressed



Fried, Eiko & Epskamp, Sacha & Nesse, Randolph & Tuerlinckx, Francis & Borsboom, Denny. (2015). What are 'good' depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. Journal of Affective Disorders. 10.1016/j.jad.2015.09.005.

Psychiatric symptom clustering

THE PROBLEM

- Psychiatric symptoms are endorsed or assigned
- High co-morbidity
- Sub-clinical presentation?

THE APPROACH

- Use a population-based study
  - Rotterdam Study, N=6602 for discovery, N-=3005 validation

Psychiatric symptom clustering

THE PROBLEM

- Psychiatric symptoms are endorsed or assigned
- High co-morbidity
- Sub-clinical presentation?

THE APPROACH

- Use a population-based study
- Use general scales on mood and sleep
  - CES-D
    - Center for Epidemiological Studies Depression Scale
  - HADS-A
    - Hospital Anxiety & Depression
  - PSQI
    - Pittsburgh Sleep Quality Index
  - 43 symptoms (Likert scales)

Psychiatric symptom clustering

THE PROBLEM

- Psychiatric symptoms are endorsed or assigned
- High co-morbidity
- Sub-clinical presentation?

THE APPROACH

- Use a population-based study
- Use general scales on mood and sleep
- Unsupervised clustering algorithms to find *transdiagnostic* patterns of symptom co-occurrence

Psychiatric symptom clustering

METHOD

- Hierarchical Cluster Analysis (HCA) approach
  - Each symptom starts in its own cluster
  - Merged with maximally similar symptom
    - Jaccard index for similarity – how similar are symptoms across participants (and vice versa)

- a = the number of attributes that equal 1 for both objects i an j
- b = the number of attributes that equal 0 for object i but equal 1 for object j
- c = the number of attributes that equal 1 for object i but equal 0 for object j
- d = the number of attributes that equal 0 for both objects i and j.

Then, Jaccard Similarity for these attributes is calculated by the following equation:

$$J(i,j) = sim(i,j) = \frac{a}{a+b+c}$$

Notice the number of 0 matches is considered unimportant in this computation and is ignored because the items are asymmetric binary attributes.

Psychiatric symptom clustering

METHOD

- Hierarchical Cluster Analysis (HCA) approach
  - Each symptom starts in its own cluster
  - Merged with maximally similar symptom
    - Jaccard index for similarity
    - Ward's Method as a linkage criterion

Ward's method says that the distance between two clusters, A and B, is how much the sum of squares will increase when we merge them:

$$\Delta(A,B) = \sum_{i \in A \cup B} \|\vec{x}_i - \vec{m}_{A \cup B}\|^2 - \sum_{i \in A} \|\vec{x}_i - \vec{m}_A\|^2 - \sum_{i \in B} \|\vec{x}_i - \vec{m}_B\|^2 \quad (2)$$

$$= \frac{n_A n_B}{n_A + n_B} \|\vec{m}_A - \vec{m}_B\|^2 \quad (3)$$

where $\vec{m}_j$ is the center of cluster j, and n_j is the number of points in it. Δ is called the **merging cost** of combining the clusters A and B.

Results

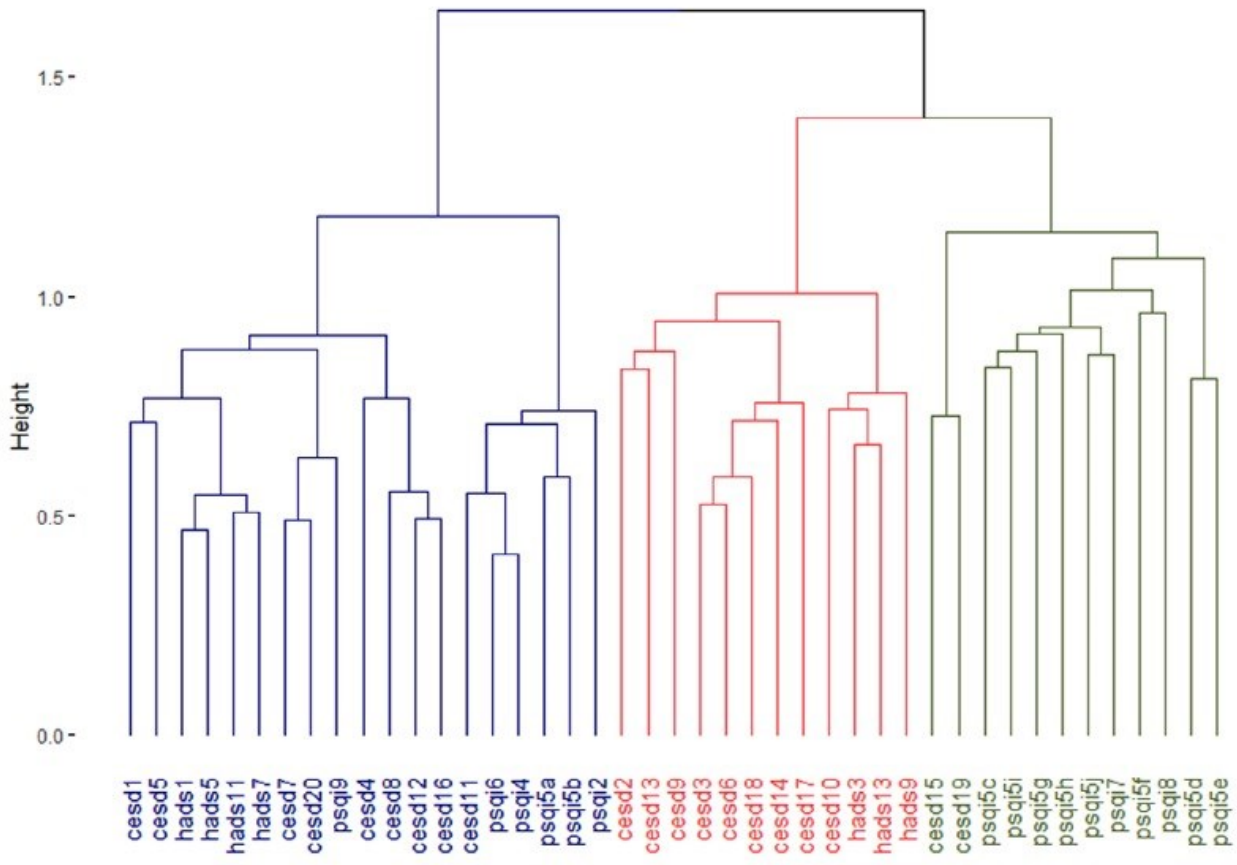


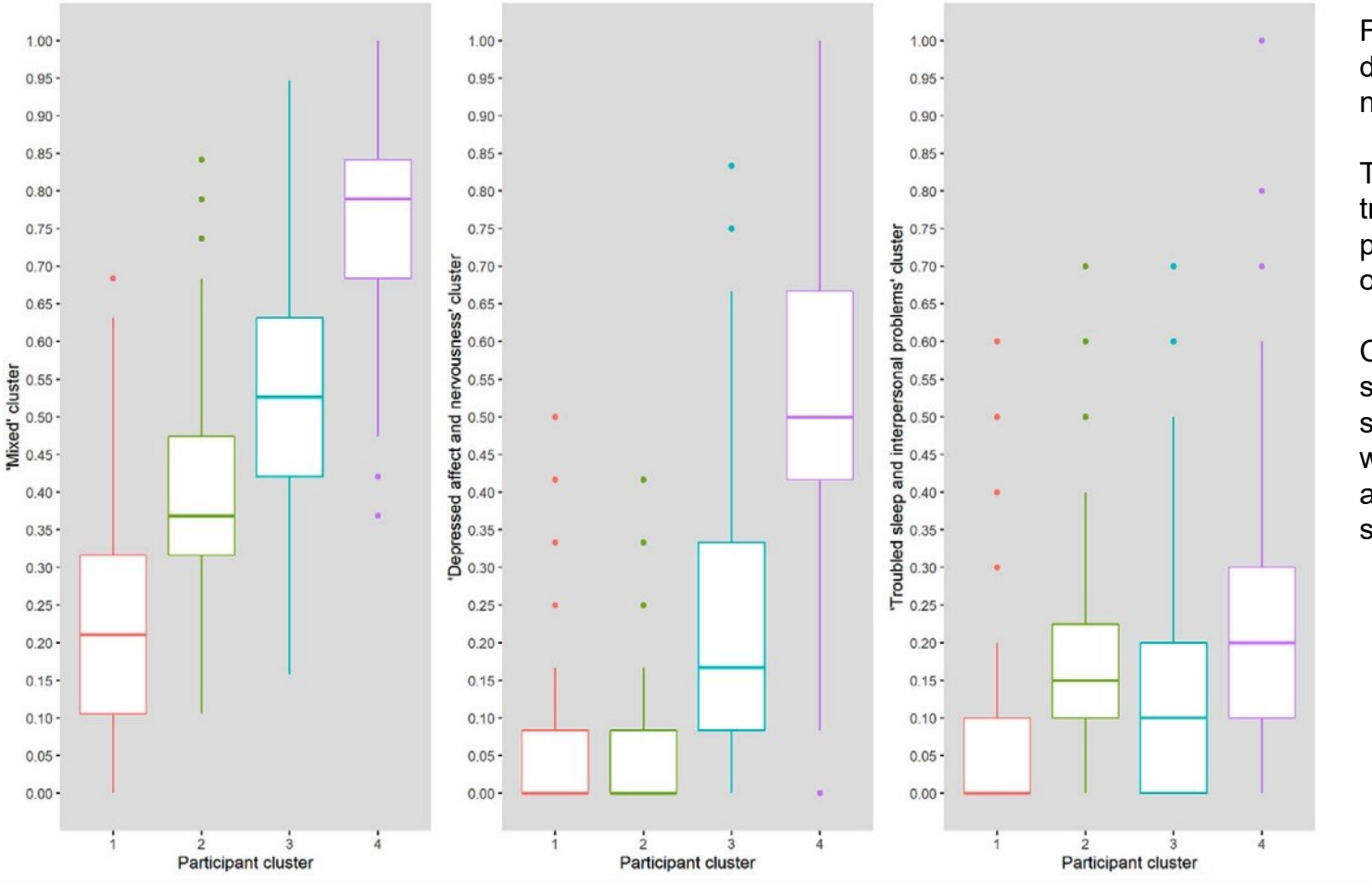
Figure 1. Dendrogram that represents the three-cluster solution of hierarchical clustering analysis on test items. Items that often co-occur cluster at lower levels (*y*-axis) in the dendrogram, while items that less often co-occur cluster only at higher level. The Jaccard index was used as the proximity measure and Ward's method as the linkage criterion. Description of questionnaire items can be found in Table 3. CES-D, Center for Epidemiologic Studies Depression Scale; HADS, Hospital Anxiety and Depression Scale; PSQI, Pittsburgh Sleep Quality Index.

Finding: Symptoms from different questionnaires do not cluster together

This may suggest transdiagnostic/subclinical patterns in co-endorsement of symptoms

Hofman, A., Lier, I., Ikram, M. A., van Wingerden, M., & Luik, A. I. (2023)
Uncovering psychiatric phenotypes using unsupervised machine learning: A data-driven symptoms approach.
European Psychiatry, *66*(1). https://doi.org/10.1192/j.eurpsy.2023.13

Clustering of participants



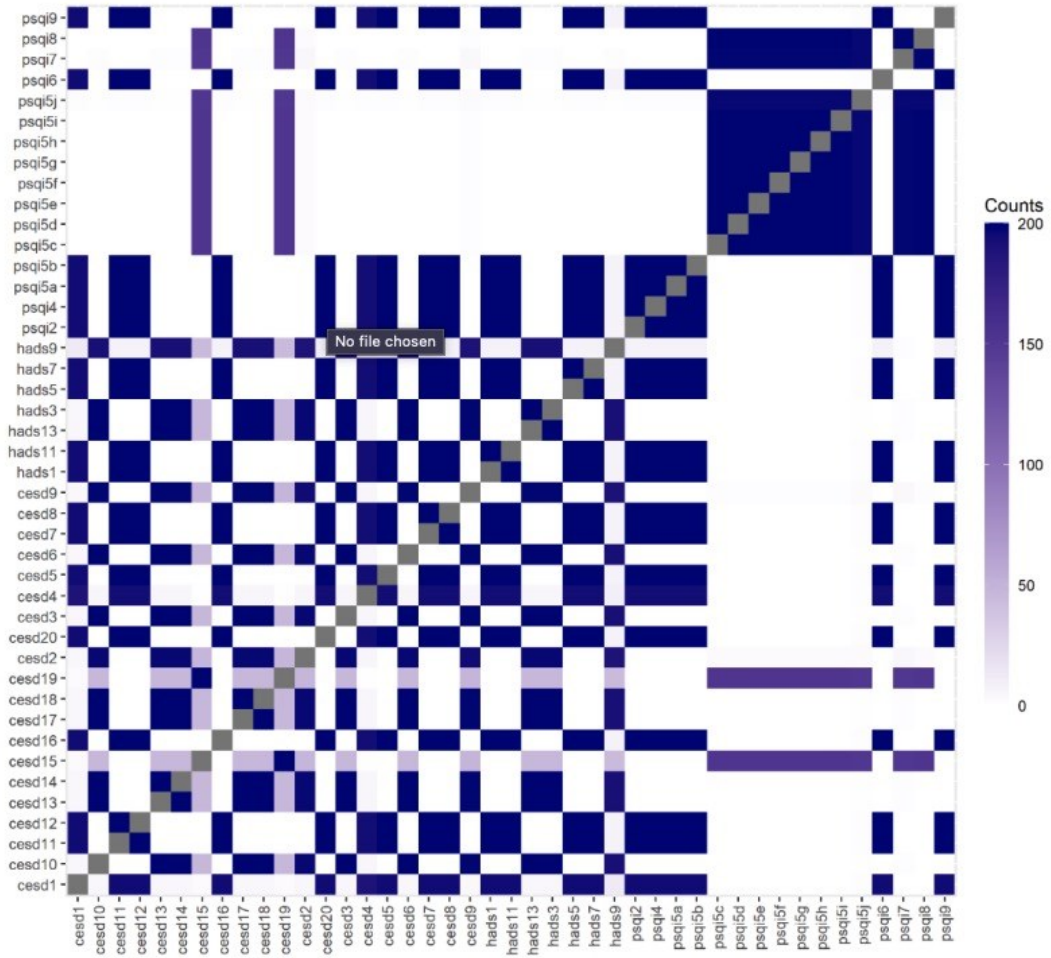
Finding: Symptoms from different questionnaires do not cluster together

This may suggest transdiagnostic/subclinical patterns in co-endorsement of symptoms

Comparing clustering solutions on average silhoutte suggested a 4-way clustering participants across 3 aggregated symptom-cluster scores

Hofman, A., Lier, I., Ikram, M. A., van Wingerden, M., & Luik, A. I. (2023)
Uncovering psychiatric phenotypes using unsupervised machine learning: A data-driven symptoms approach.
European Psychiatry, *66*(1). https://doi.org/10.1192/j.eurpsy.2023.13

Clustering of participants



Appendix, Figure 3. Heatmap that indicates how often two items occurred in the same cluster when randomly resampling and clustering the data using hierarchical clustering analysis 200 times.
Description of questionnaire items can be found in Appendix Table 4.

Finding: Symptoms from different questionnaires do not cluster together

This may suggest transdiagnostic/subclinical patterns in co-endorsement of symptoms

Comparing clustering solutions on average silhoutte suggested a 4-way clustering participants across 3 aggregated symptom-cluster scores

Randomly subsampling the participant database yielded stable symptom clusters (co-occurrence across resamples)

Hofman, A., Lier, I., Ikram, M. A., van Wingerden, M., & Luik, A. I. (2023) Uncovering psychiatric phenotypes using unsupervised machine learning: A data-driven symptoms approach. *European Psychiatry*, *66*(1). https://doi.org/10.1192/j.eurpsy.2023.13

Longitudinal modelling of PROMs

In collaboration with Elizabeth-Tweesteden Hospital in Tilburg and Network Actute Care North-Brabant

Longitudinal modelling of PROMs

THE PROBLEM

- After hospitalisation for a traumatic injury, *subjective* recovery differs quite a lot
  - Patient-reported outcome measures
    - Physical Health
    - Psychological Health
- Prediction in recovery can help set expectations
  - Target interventions early
  - What are sensible clinical outcome groups?

BIOS: BRABANT INJURY OUTCOME SURVEILLANCE

OBJECTIVES

1. to investigate the short-term and long-term HRQoL, functional, psychological and economic outcome after non-fatal trauma;
2. to investigate the risk factors for decreased HRQoL, functional, psychological and economic outcome after non-fatal trauma;
3. to describe the healthcare use, medical costs and productivity loss due to non-fatal trauma;
4. to develop a risk profile for recovery after non-fatal injury in the short and long term;
5. to validate and develop models for predicting non-fatal outcome after trauma;
6. to investigate whether a structural enlargement of the trauma registry with patient-reported outcome measurement does add value.

Stoitsas, K., Bahulikar, S., de Munter, L., de Jongh, M. A. C., Jansen, M. A. C., Jung, M. M., van Wingerden, M., & van Deun, K. (2022). Clustering of trauma patients based on longitudinal data and the application of machine learning to predict recovery. *Scientific Reports*, *12*(1)

Longitudinal modelling of PROMs

THE DATASET

- BIOS: Brabant Injury Outcome Surveillance
  - ICU admitted patients
    - N=4883
    - Six timepoints
  - EQ-5D-3L & VAS (slider)
  - Health Utilities Index 2 &3
  - Hospital Anxiety & Depression Scale
    - HADS-A
    - HADS-D

EUROQOL-5D-3L

Under each heading, please tick the ONE box that best describes your health TODAY.

MOBILITY

- I have no problems in walking about ☑
- I have some problems in walking about ☐
- I am confined to bed ☐

SELF-CARE

- I have no problems with self-care ☑
- I have some problems washing or dressing myself ☐
- I am unable to wash or dress myself ☐

USUAL ACTIVITIES (e.g. work, study, housework, family or leisure activities)

- I have no problems with performing my usual activities ☐
- I have some problems with performing my usual activities ☑
- I am unable to perform my usual activities ☐

PAIN/DISCOMFORT

- I have no pain or discomfort ☐
- I have moderate pain or discomfort ☐
- I have extreme pain or discomfort ☑

ANXIETY/DEPRESSION

- I am not anxious or depressed ☐
- I am moderately anxious or depressed ☑
- I am extremely anxious or depressed ☐

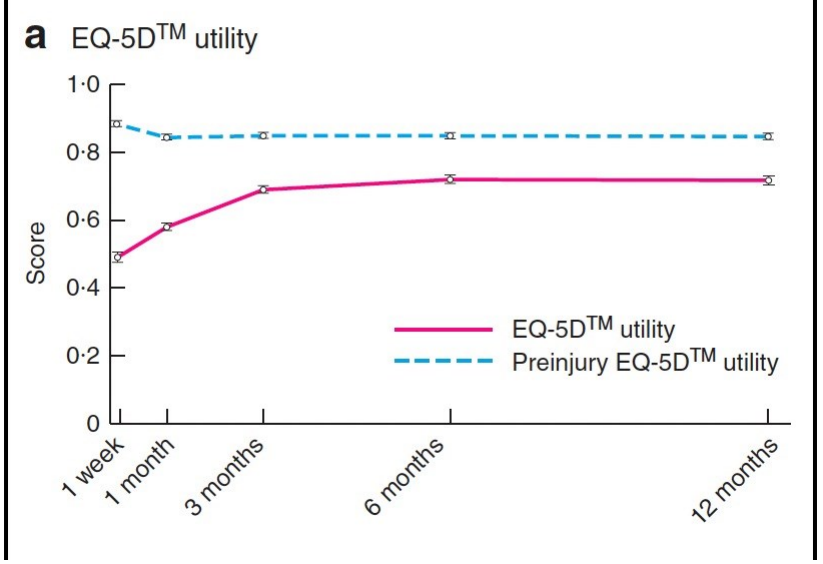
Longitudinal modelling of PROMs

PREVIOUS APPROACH

- Univariate statistical modelling
  - Linear Mixed Models
  - Beta regression weights
  - Best predictors:
    - Pre-injury health
    - Overall Frailty
    - Age, Sex
    - Duration of hospitalization

RECOVERY PATTERNS

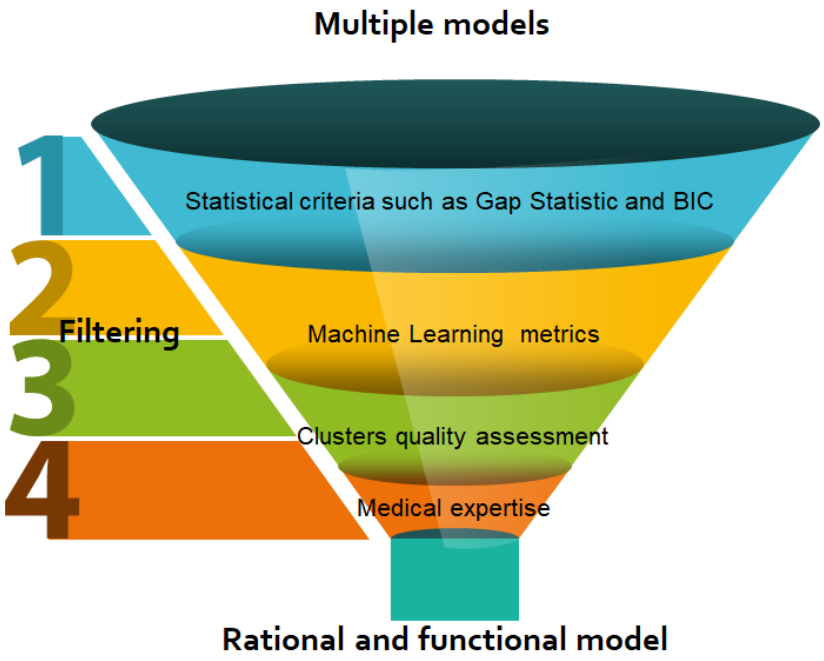
Fig. 2 Mean scores on health status questionnaires during follow-up



de Munter, L., Polinder, S., van de Ree, C. L. P., Kruithof, N., Lansink, K. W. W., Steyerberg, E. W., & de Jongh, M. A. C. (2019). Predicting health status in the first year after trauma. *British Journal of Surgery*, *106*(6), 701–710. https://doi.org/10.1002/bjs.11132

Longitudinal modelling of PROMs

APPROACH

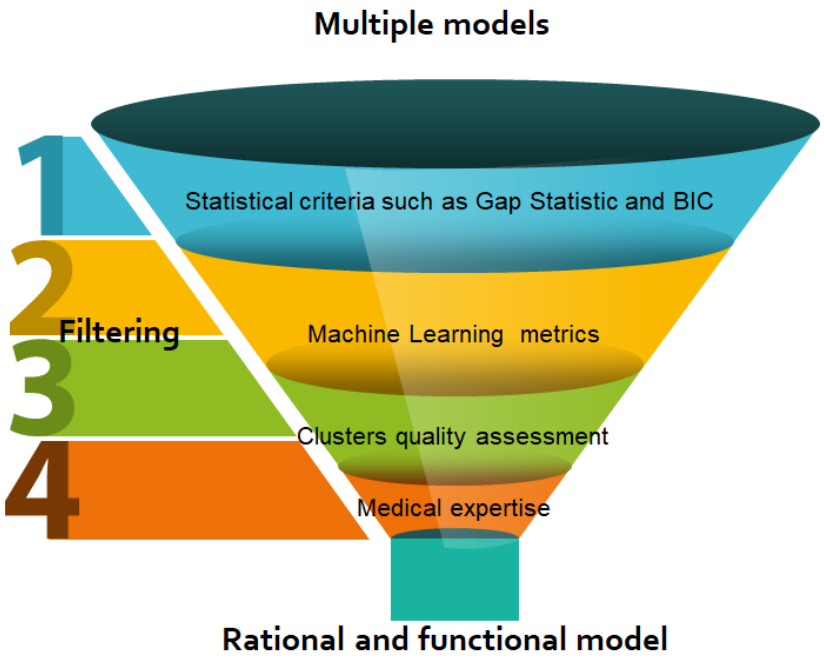


METHODS

- Step 1: Unsupervised longitudinal clustering
  - Create multivariate longitudinal clusters
    - K-ml3D
    - Deepgmm
    - HDclassif
  - Assess *quantitative* cluster quality
    - Gap Statistic
    - BIC

Longitudinal modelling of PROMs

APPROACH



METHODS

- Step 2: Supervised prediction of clusters
  - Assuming the cluster label as ground truth, how well can cluster be predicted?
    - From baseline measures
    - Demographic variables
    - Not used in clustering
  - Compare ML models
    - LR, RF, XGBoost

Longitudinal modelling of PROMs

RESULTS

- Gap statistic for choice of number of clusters per clustering algorithm
- Tuned across clustering hyper-parameters

CLUSTERING

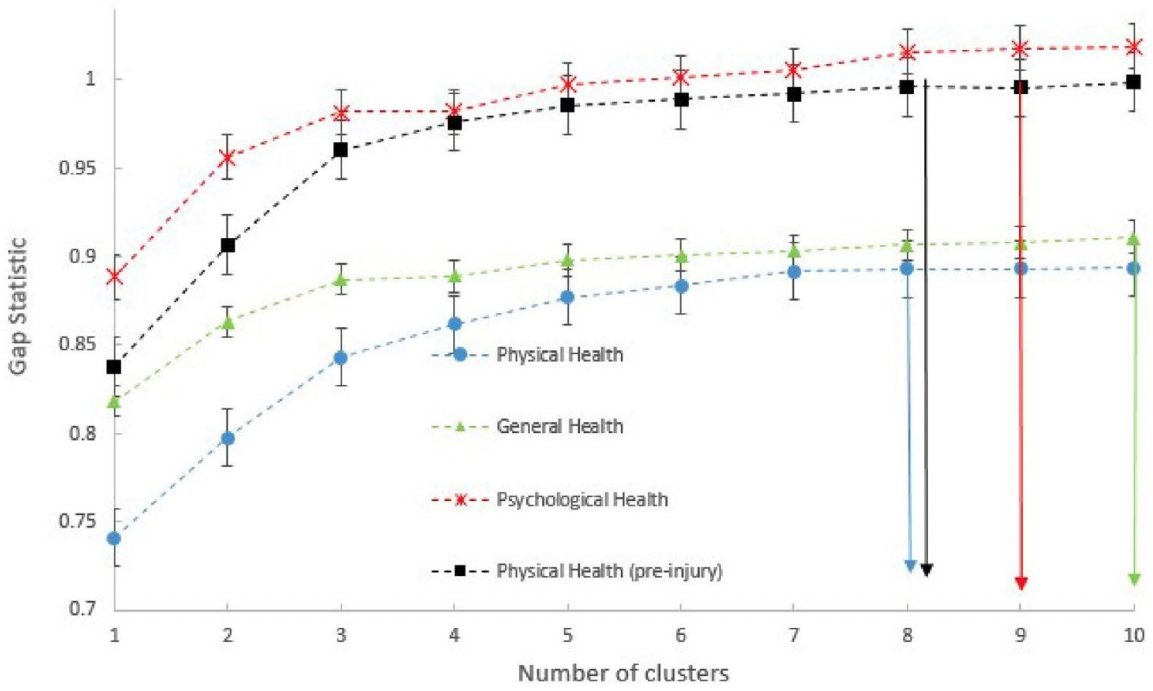
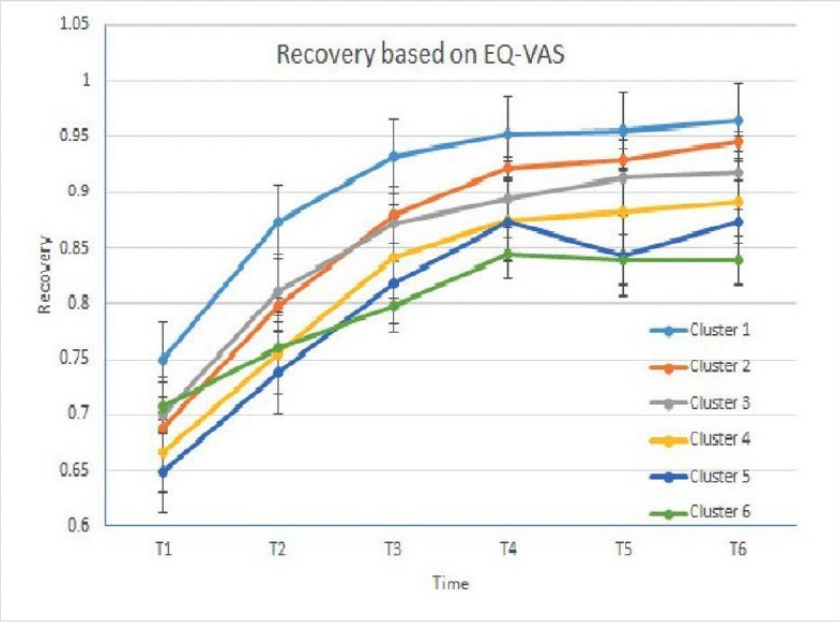


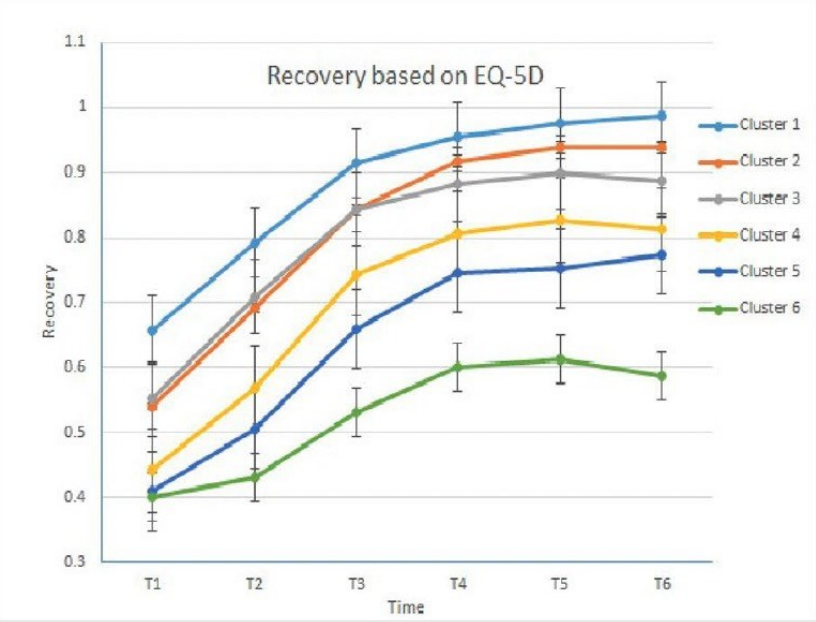
Figure 2. Optimum number of clusters with kml3d for the four different cases of variables and k-means.

Longitudinal modelling of PROMs

RESULTS



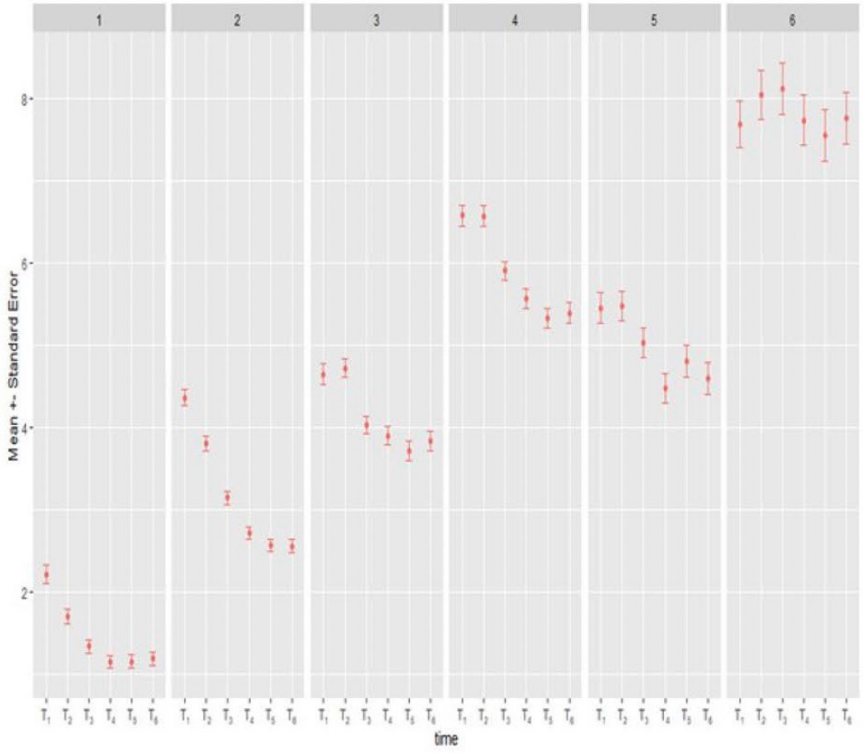
CLUSTERING



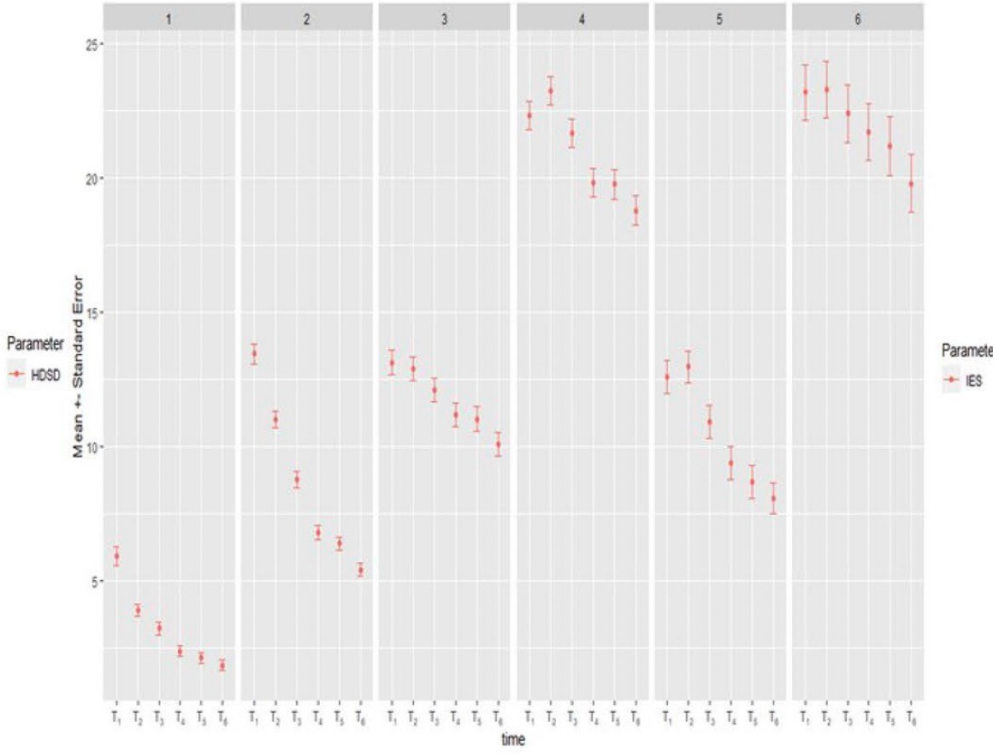
In these graphs, recovery is the EQ5D score normalised by the pre-injury baseline

Longitudinal modelling of PROMs

RESULTS



CLUSTERING



In these graphs, depression & anxiety scores vary from low (cl. 1+2 – with recovery) to high (cluster 6 – some recovery)

Longitudinal modelling of PROMs

RESULTS

MACHINE LEARNING

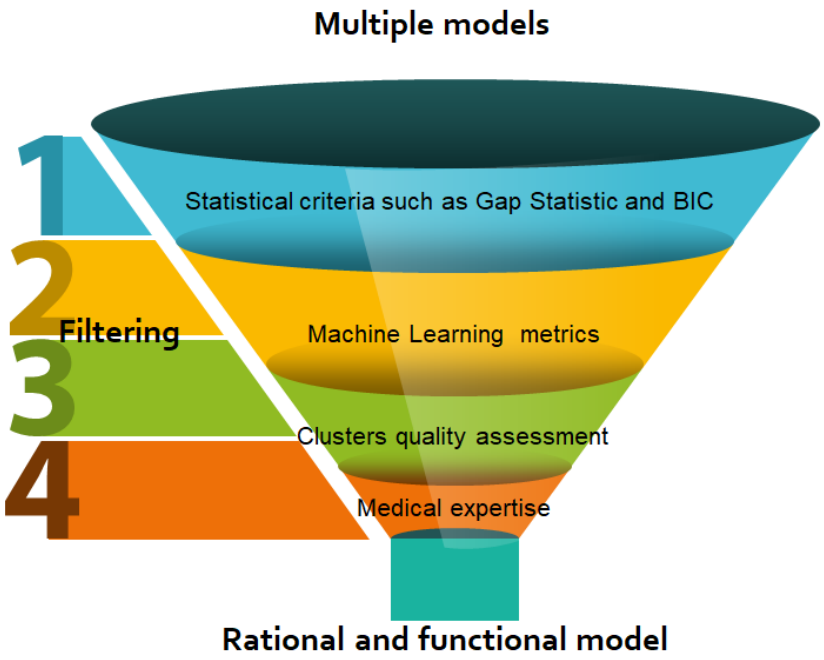
- Algorithms tuned for accuraccy + f1-macro
  - Per Clustering algorithm, per number of clusters

Model	Mean accuracy % (f_1_macro)	95% CI for accuracy	Optimized hyper-parameters
Logistic regression	36.53 (33.33)	[35.60–37.46]	Solver = 'newton-cg', C = 10, penalty = 'l_2'
Logistic regression (under-sampling)	36.11 (34.10)	[34.44–37.79]	Solver = 'newton-cg', C = 10^{-2}, penalty = 'l_2'
Logistic regression (smote)	37.88 (35.53)	[36.83–38.94]	Solver = 'saga', C = 10^{-2}, penalty = 'l_2'
Logistic regression (over-sampling)	37.20 (35.64)	[35.96–38.45]	Solver = 'lib-linear', C = 10^{-2}, penalty = 'l_2'
Random forest	36.98 (35.89)	[34.69–39.27]	Estimators = 200, max depth = 15, min samples split = 5
Random forest (under-sampling)	36.07 (35.12)	[34.24–37.89]	Estimators = 50, max depth = 5, min samples split = 10
Random forest (smote)	54.75 (53. 67)	[52.79–56.70]	Estimators = 500, max depth = 50, min samples split = 2
Random forest (over-sampling)	69.12 (68.71)	[67.81–70.44]	Estimators = 500, max depth = 50, min samples split = 2
XGBClassifier (over-sampling)	68.52 (67.78)	[67.37–69.67]	Estimators = 500, max depth = 10
XGBClassifier (smote)	57.14 (56.23)	[55.95–58.34]	Estimators = 500, max depth = 20

Table 2. Example of comparison models for the classification of six clusters obtained for the case of Physical Health (pre-injury) with HDclassif method.

Longitudinal modelling of PROMs

APPROACH



METHODS

- Steps 3+4: Cluster Quality & Clinical assessment
  - Clusters should diverge on known predictors
    - Age
    - Sex
    - Frailty
    - Hip Fracture
    - Hospital Stay
    - Etc.

Longitudinal modelling of PROMs

Case/method	Cluster	Age		Frailty		Comorbidities		Severity score		Admission days in hospital		Gender		Hip fracture	
		Mean	S.E	Mean	S.E	Mean	S.E	Mean	S.E	Mean	S.E	Male %	Female %	No %	Yes %
General Health	1	58.52	0.69	0.63	0.07	0.60	0.04	5.76	0.19	3.91	0.17	65.47	34.53	83.30	16.70
HDclassif	2	61.78	0.48	1.95	0.12	0.88	0.03	6.32	0.13	5.13	0.12	51.98	48.02	78.80	21.20
Highly sensible (+++)	3	64.48	0.64	2.79	0.16	1.08	0.04	6.33	0.16	5.91	0.19	52.73	47.27	77.00	23.00
	4	65.91	0.57	4.86	0.19	1.44	0.04	7.01	0.16	7.50	0.20	39.70	60.30	72.00	28.00
	5	73.89	0.81	5.72	0.27	1.87	0.07	7.02	0.22	8.17	0.40	35.73	64.27	61.40	38.60
	6	74.72	0.93	7.96	0.28	2.04	0.08	7.76	0.28	9.65	0.54	30.95	69.05	54.80	45.20
Physical Health (pre-injury)	A	63.82	0.61	1.89	0.11	0.97	0.04	6.06	0.15	4.83	0.14	49.49	50.51	78.14	21.86
kml3d	B	58.42	0.63	1.06	0.09	0.54	0.03	5.53	0.16	3.93	0.15	66.21	33.79	86.01	13.99
Medium sensible (++)	C	58.06	0.68	1.46	0.15	0.69	0.03	6.72	0.19	5.31	0.17	57.11	42.89	79.77	20.23
	D	69.27	0.66	3.53	0.18	1.58	0.05	6.51	0.18	6.84	0.22	41.43	58.57	69.47	30.53
	E	63.22	0.74	2.81	0.21	1.19	0.04	7.09	0.22	7.30	0.27	43.37	56.63	76.91	23.09
	F	72.78	0.95	6.02	0.28	1.84	0.07	7.72	0.28	10.00	0.49	27.78	72.22	59.34	40.66
	G	73.64	0.84	7.09	0.24	2.05	0.08	7.31	0.28	8.27	0.37	34.96	65.04	62.18	37.82
	H	78.45	0.82	9.47	0.22	2.38	0.08	7.70	0.28	9.60	0.53	28.21	71.79	50.32	49.68
Psychological health	1	64.21	1.44	3.38	0.42	1.19	0.09	6.90	0.38	5.95	0.35	52.02	47.98	70.71	29.29
Deepgmm	2	64.22	2.15	3.78	0.77	1.37	0.16	8.59	0.87	8.67	0.79	40.24	59.76	76.52	23.48
Poorly sensible (+)	3	64.79	0.30	5.04	0.48	1.37	0.09	5.98	0.30	6.40	0.47	39.47	60.53	73.68	26.32
	4	64.84	0.66	3.62	0.1	1.17	0.02	6.54	0.07	6.22	0.10	48.50	51.40	71.95	28.05
	5	66.70	1.82	3.84	0.59	1.34	0.12	6.76	0.57	6.05	0.63	40.87	59.13	74.41	25.59
	6	67.53	1.54	4.31	0.54	1.39	0.11	6.93	0.42	7.38	0.61	37.50	62.50	69.37	30.63

Table 5. Descriptive statistics for clusters obtained with different methods and that have been evaluated as Highly (+++), Medium (++) and Poorly (+) sensible.

Evaluate predictor values between groups with an MANOVA test

Longitudinal modelling of PROMs

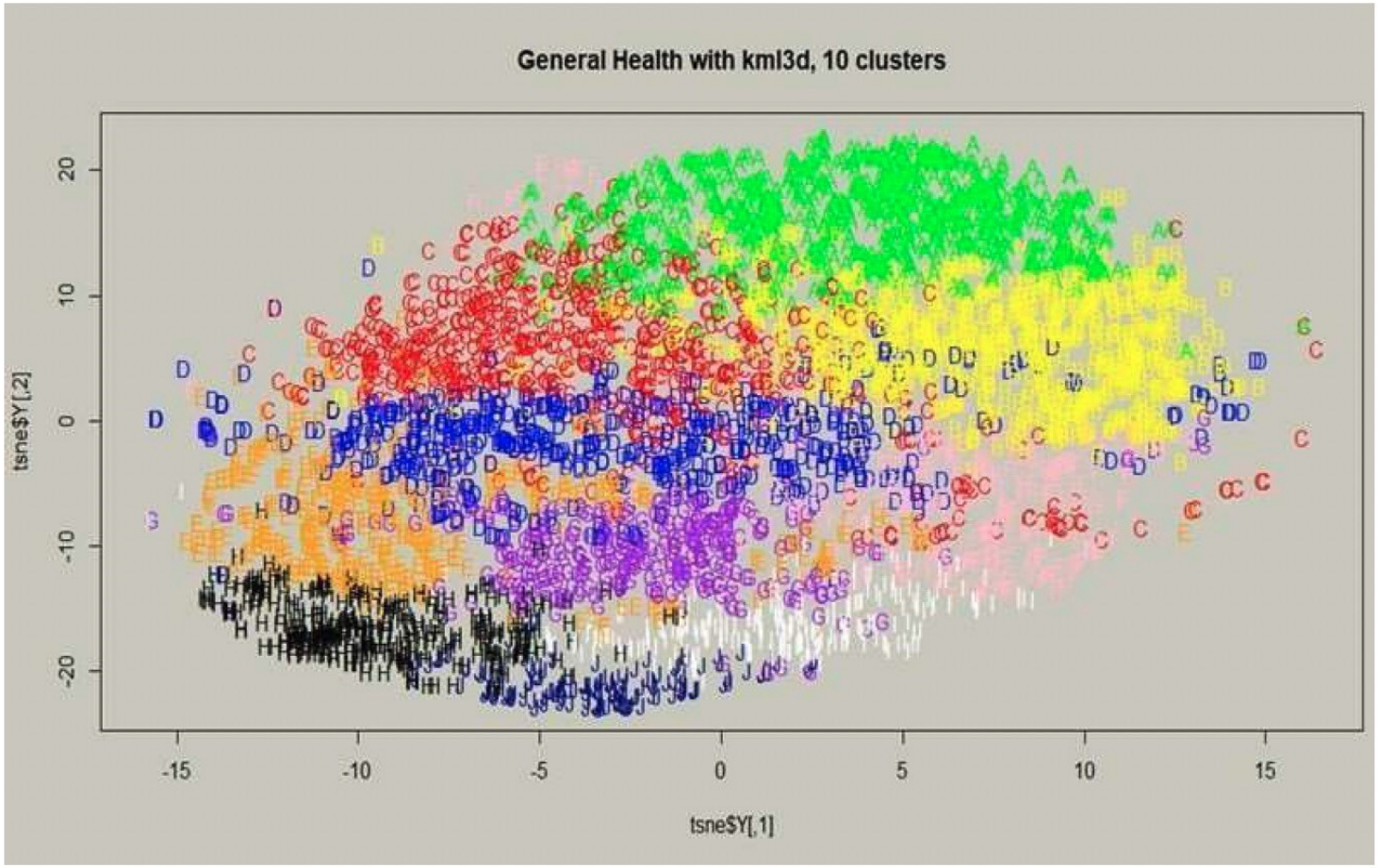
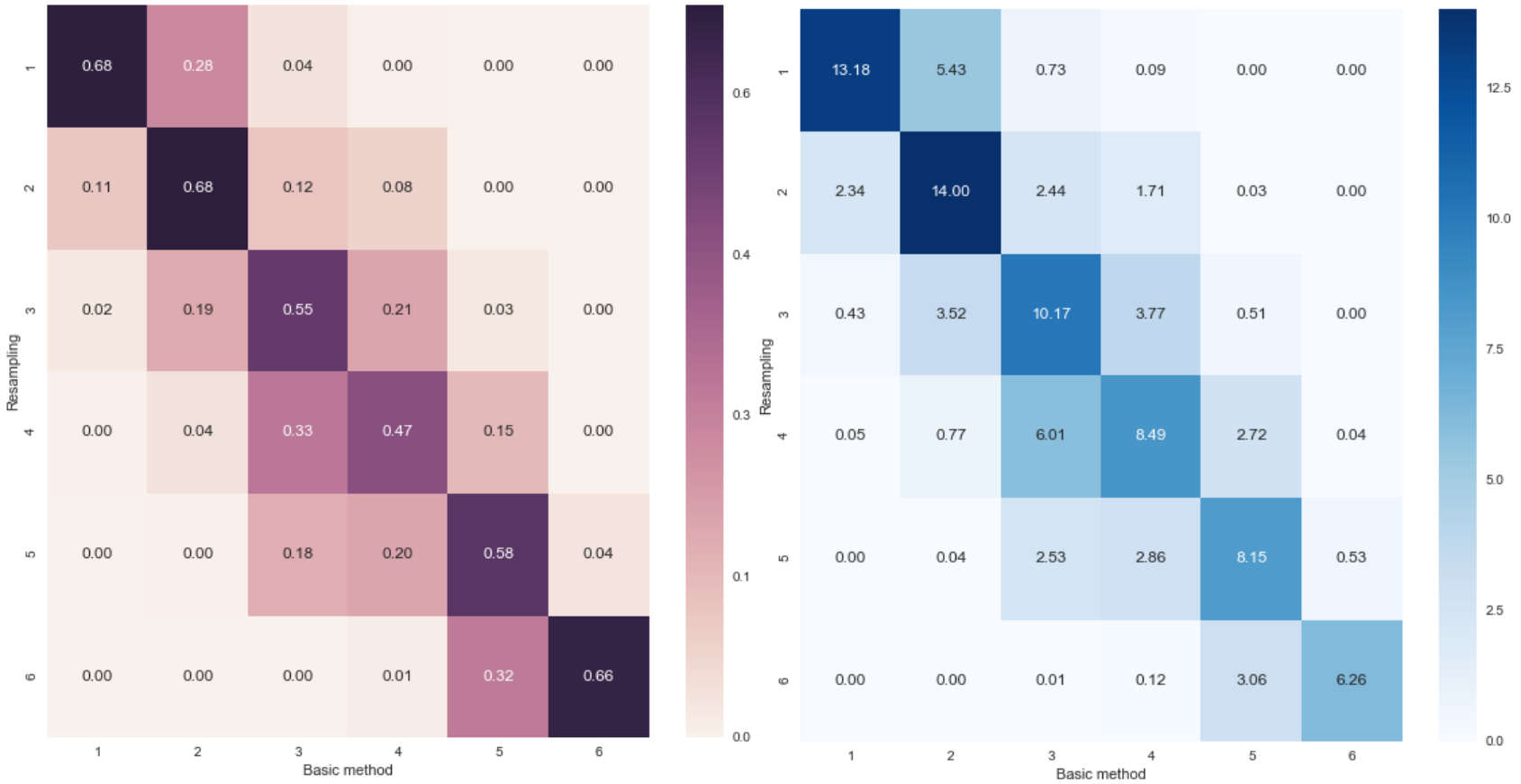


Figure S5. t-distributed stochastic neighbor embedding (t-SNE) graph for the case of General Health with kml3d and 10clusters.

Longitudinal modelling of PROMs



Cluster stability analysis with resampling: how often do participants end up in the same cluster?

Longitudinal modelling of PROMs

RESULTS

MACHINE LEARNING

- Algorithms tuned for accuraccy + f1-macro
  - Selected based on clinical sensibleness

Case	Method	Optimum Nr of clusters	Majority baseline	Accuracy ($f_{1\_}$ macro) RF over-sampling	95% CI for accuracy	Clinical sensibleness
Physical Health	kml3d	8	16.48	51.89 (50.03)	[50.84–52.94]	+++
Psychological Health	kml3d	9	26.15	83.12 (82.63)	[82.36–83.87]	++
General Health	kml3d	10	17.07	68.26 (68.02)	[67.66–68.85]	+++
Physical Health (pre-injury)	kml3d	8	18.08	61.56 (60.74)	[60.32–62.81]	++
Physical Health	HDclassif	7	24.49	69.52 (68.61)	[68.00–71.05]	+++
Psychological Health	HDclassif	10	19.12	70.24 (69.75)	[68.78–71.70]	+
General Health	**HDclassif**	**6**	**30.03**	**73.96 (72.59)**	**[72.89–75.03]**	**+++**
Physical Health (pre-injury)	HDclassif	6	26.07	69.12 (68.64)	[67.81–70.44]	+++
Physical Health	**Deepgmm**	**6**	**45.32**	**91.30 (90.85)**	**[90.50–92.11]**	**+++**
Psychological Health	Deepgmm	6	84.70	99.96 (98.67)	[99.94–99.98]	+
General Health	Deepgmm	6	62.13	98.20 (97.87)	[97.87–98.54]	++
Physical Health (pre-injury)	Deepgmm	6	61.02	94.78 (93.55)	[94.37–95.18]	+

Table 3. Summary of classification results and quality assessment for clusters obtained from longitudinal data. Best models based on classification metrics and clinical sensibleness are in bold.

RF prediction of cluster membership should work well if known predictors are linked to sensible outcome clusters

Resampling shows that clusters are stabel longitudinal entities

F1-macro used for model comparison

Majority baseline addresses whether clusters are balanced

When models make mistakes, confusion is usually limited to 1 severity level up or down

Predicting Falls in Elderly Care

A case study in secure data sharing and merging across elderly care institutions

Predicting Falls in Elderly Care

THE PROBLEM

- Until recently, pertinent information on clients-at-risk was qualitative
  - Many care professionals
  - Possibility for missing combinatorial evidence
  - Time pressure / Triage
- Fall risks are diverse
  - Bed
  - Bathroom
  - Medication

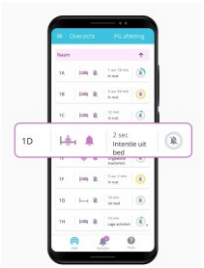
THE CONSEQUENCE

- Hip fractures mortality is very high
  - During first year after hip fractures surgery 15% - 36%
  - In elderly (65+), 3–4 times higher than expected in the general population

Observation: underused data potential

VVTs continuously record qualitative data regarding 'fall incidents' that are potentially of value in predicting and/or preventing future fall incidents.

Entrepreneurs continuously record quantitative data regarding 'fall risk-fall incidents' that are potentially of value in predicting and/or preventing future fall incidents.



Dataproject Fall Prevention

GOALS

- Use data from innovative devices in combination with regular care data
  - Prevent or improve care/increase health
  - Fewer fall incidents & lower staff & financial burden.
- Blueprint for sharing sensitive health data
  - GDPR-proof

METHOD

- Hybrid machine learning model
  - Supervised: to predict whether a fall incident will take place yes/no (% -> risk score)
  - Combines hand-chosen features + AI-generated featured (text analysis with LLMs)

Dataproject valpreventie

- Goals of participating VVT organizations:
  - How data can work for them in the future?
  - To allow clients to move more freely
  - To reduce the number of falls with a hip or pelvic fracture by 5%

Aspirational

- Board-level concerns:
  - reality check governance/staffing
  - invest in data quality (?)
  - break down barriers

Realistic

Concerns/Barriers

- Interest is high from multiple layers and areas of expertise within each organization
  - Care content: practitioners; Innovation: policy advisors; ICT: analysis and analytics
  - Consensus: qualitative data (description incidents and free text fields) and sensor systems (hip airbag, bedsense, etc.) yield much potential for improvement of care.
- GDPR objections:
  - Sharing of data for research purposes not without 'approval' GDPR expert.
  - Board responsibility in case of data breach
  - Unknown and unloved: how can data be shared GDPR-proof?
- Time
  - Unfamiliarity and unfamiliarity with how to spend time, including from the IT departmnet, to share data
  - Legal issues regarding contracts / data ownership
  - Dealing with contracted third parties involved in IT infrastructure (e.g. SAAS)

Fall Prevention: data sharing principles

- Care institutions own the data (on behalf of their clients)
  - They have data processing agreements with technological partners
  - Data should be accessible to care institution for further processing/sharing
  - Data can be combined at the level of care institution before anonymization
    - Share only fully anonymized data (initially)
- Data sharing principles:
  - Raw data never leaves the VVT organization for the purpose of this project.
  - Parties will not have access to raw data without data processing agreement (Data Management Plan).
  - Only fully anonymized data shared to knowledge partners (Tilburg University/Fontys university of applied sciences).

We start with completely anonymised data

INPUT (AT CARE ORGANIZATION)

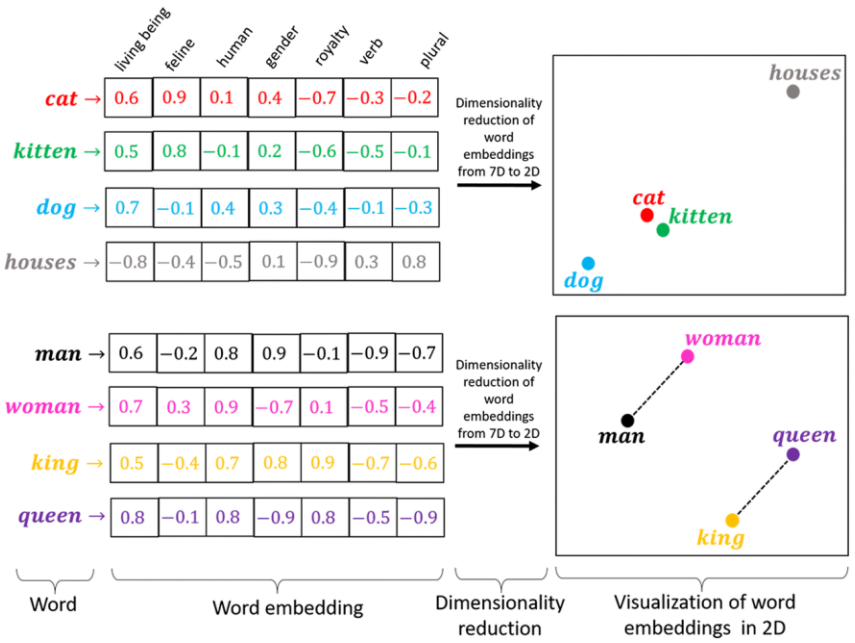
- Linked data files with unique identifiers
  - Optional: include technology partners (Momo, Wolk etc.)
- Combined and anonymised
  - Removal protected words
  - Removal unique identifiers
  - Binning of unique characteristics
    - Ages, zip-code, DBC, etc.
- Script adapted to each VVT org

OUTPUT (FOR ML MODELS)

- Harmonised numerical representation of text data
- Outcome label for prediction models
  - Is this a fall incident report (yes/no) – automated detection
  - Will this client experience a fall incident in the next xx days
    - Requires labeled cases
    - Requires same data structure between care organisations

Ways to deal with "free-text data": natural language processing (Large Language Models)

- Intuition: we transcode the words to a numeric representation
  - At the data source
- Protected words can be excluded
  - Names, organizational identifiers
- The model only sees the numeric representation
  - Conceptually related words have similar embeddings
    - "val", "viel", "gevallen"



Example from Word2Vec embedding
Currently using BERT

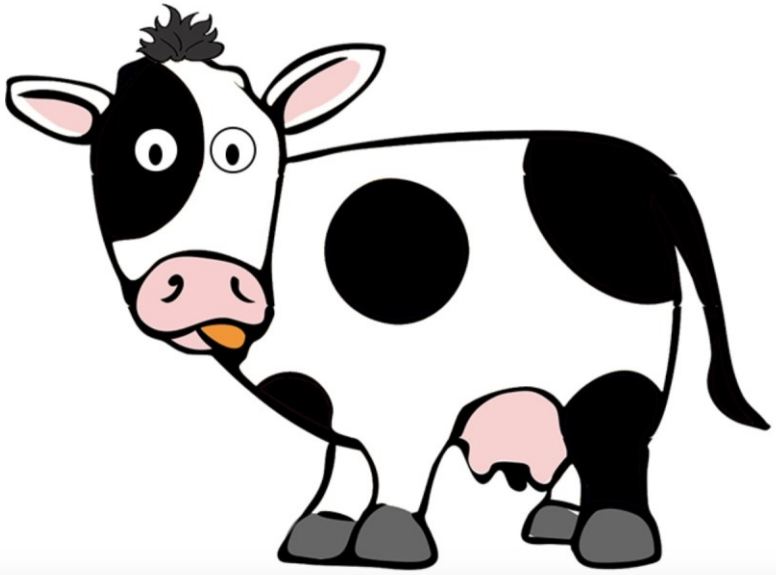
BERTje: Dutch LLM (Large Language Model)

BERTje: A Dutch BERT model

Wietse de Vries • Andreas van Cranenburgh • Arianna Bisazza • Tommaso Caselli • Gertjan van Noord • Malvina Nissim

Model description

BERTje is a Dutch pre-trained BERT model developed at the University of Groningen.



Downloads last month
75,390

LLMs are trained by trying to predict a masked word

Hosted inference API

Fill-Mask

Mask token: [MASK]

Mw heeft conditioneel erg in[illegible]everd na haar CVA en opname in Damast. Traplopen gaat n[illegible]el, maar dient gemonitord te worden. Mw loopt niet meer [MASK] naar buiten. Heeft een rollator. en een PAS-alarm. Fysio is betrokken. De linkerkant van mw lijkt minder aandacht te hebben: voetplaatsing links is minder, spullen aan de linkerkant lijkt mw minder goed te zien.

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 5.268 s

terug	0.113
mee	0.062
helemaal	0.052
zelfstandig	0.040
recht	0.030

</> JSON Output Maximize

Spaces using GroNLP/bert-base-dutch-cased 7

Onder de motorkap van het model

FREE TEXT INPUT

'Mw heeft conditioneel erg ingeleverd na haar CVA en opname in Damast. Traplopen gaat nog wel, maar dient gemonitord te worden. Mw loopt niet meer **zelfstandig** naar buiten. Heeft een rollator. en een PAS-alarm. Fysio is betrokken. De linkerkant van mw lijkt minder aandacht te hebben: voetplaatsing links is minder, spullen aan de linkerkant lijkt mw minder goed te zien.

EMBEDDING OF THIS TEXT

(tensor([[[-0.7453, 0.3119, 1.2152, ..., 0.5540, 1.7434, 0.2024], [0.6120, 1.0086, 0.8587, ..., 0.2685, 0.3518, -0.6152], [0.7964, -0.3194, 0.7142, ..., 0.5721, -0.3139, -0.4117], ..., [-1.2376, 0.0883, 1.0412, ..., 0.3173, 0.0867, 1.1322], [-0.4769, -0.2500, -0.1664, ..., 0.2237, -0.1251, 0.0473], [-0.3798, -0.2515, -0.9949, ..., 1.3126, 0.4502, 0.6987]]]), tensor([[[-0.6799, -0.1637, 0.8963, ..., -0.1299, 1.6992, -0.2568], [1.0078, 1.1362, 0.7193, ..., 0.1212, -0.7654, -0.9073], [0.6861, -0.6270, 0.6657, ..., 0.1701, -0.3719, -0.6057], ..., [-0.5104, -0.1291, 0.5278, ..., -0.1978, -0.4018, 0.2476], [-0.3430, -0.3793, 0.2438, ..., -0.0353, -0.1616, -0.3710], [-0.2114, -0.0578, -0.4720, ..., 0.3360, 0.1924, 0.3138]]]), tensor([[[-6.6567e-01, -1.3152e-02, 7.1449e-01, ..., 2.6461e-01, 1.0560e+00, -5.5462e-01], [1.1666e+00, 1.5905e+00, 9.6009e-01, ..., 2.3620e-01, -8.2783e-01, -7.5930e-01], [7.6011e-01, -6.8159e-01, 7.6649e-01, ..., -7.1756e-02, -3.4518e-01, -4.0810e-01], ..., [-5.4589e-01, 4.3929e-01, 6.9808e-01, ..., -7.1502e-02, -7.1228e-02, 3.4497e-01], [-1.1461e-03, -4.0201e-01, 1.9189e-01, ..., 6.6027e-02, -4.1093e-02, -1.9744e-01], [1.2044e-01, 6.0430e-02, -1.2672e-01, ..., 2.4313e-01, -3.4186e-02, 8.0941e-02]]]), tensor([[[-0.4142, -0.2211, 0.5400, ..., 0.3761, 0.6140, -0.5144], [0.9146, 0.8400, 0.7304, ..., -0.0113, -0.2229, -0.4556], [0.9352, -0.3508, 0.9446, ..., 0.5668, -0.3999, -0.2121], ..., [-0.2960, 0.0732, 1.3859, ..., 0.3576, 0.3850, 0.2180], [0.0598, -0.4895, 0.4806, ..., 0.2585, 0.1896, -0.0444], [0.0198, -0.1742, -0.4478, ..., 0.1739, 0.0183, 0.3441]]]), tensor([[[-0.5538, -0.4022, 0.4387, ..., 0.4716, 0.5005, -0.4182], [0.1738, 0.4300, 0.8461, ..., 0.0282, 0.4323, -0.0118], [0.9688, -0.2894, 1.0744, ..., -0.0185, -0.1000, 0.1121], ..., [-0.0370, -0.0163, 1.3415, ..., 0.0680, 0.3925, 0.2905], [0.2620, -0.4416, 0.3802, ..., 0.1005, 0.2902, 0.0374], [0.0946, -0.2485, -0.4516, ..., 0.1807, 0.1451, 0.5536]]]), tensor([[[-0.5347, -0.3462, 0.6396, ..., 0.2445, 0.5632, -0.3742], [0.3895, 0.2678, 0.5365, ..., 0.0878, 0.0842, -0.0127], [0.4665, -0.2619, 1.5913, ..., 0.3120, 0.2740, 0.2927], ..., [-0.0687, -0.1805, 1.4676, ..., -0.2361, 0.3784, 0.0525], [0.2494, -0.4532, 0.1437, ..., -0.1023, 0.0742, 0.0784], [0.0517, -0.3246, -0.3403, ..., -0.0335, -0.1094, 0.3346]]]), tensor([[[-0.6970, -0.4306, 0.5621, ..., 0.3357, 0.5068, -0.3939], [0.3714, 0.2149, 0.3471, ..., -0.3191, 0.1400, -0.3044], [0.3886, -0.1555, 1.1147, ..., 0.2614, -0.0140, 0.6039], ..., [-0.3017, 0.0307, 1.2923, ..., -0.1196, 0.2522, -0.3017], [0.1703, -0.4838, -0.0319, ..., 0.0060, 0.1567, -0.0997], [-0.0827, -0.2825, -0.2450, ..., 0.0568, 0.0197, 0.3032]]]), tensor([[[-6.5551e-01, -3.1119e-01, 4.6724e-01, ..., 2.2278e-01, 3.9724e-01, -5.4669e-01], [6.3339e-01, 2.0060e-01, -2.6382e-01, ..., -1.9159e-02, -3.7794e-01, -2.4338e-01], [7.0683e-02, 2.3454e-01, 1.0452e+00, ..., -1.4925e-01, -2.3614e-01, 1.9369e-01], ..., [2.4605e-01, 6.1788e-04, 7.9603e-01, ..., -3.4970e-01, 3.1079e-01, -8.9779e-01], [-7.2897e-02, -2.4169e-01, -1.6887e-01, ..., 3.7206e-01, -9.9561e-02, -6.8134e-02], [-3.1988e-01, -3.2249e-01, -2.6230e-01, ..., 1.3556e-01, -3.4116e-02, 3.7821e-01]]]), tensor([[[-0.5452, -0.4349, 0.5776, ..., 0.1616, 0.0929, -0.4662], [1.0067, -0.0716, -0.3914, ..., 0.6914, -0.3050, -0.2833], [-0.3331, 0.0894, 0.6061, ..., -0.1637, -0.1837, 0.3857], ..., [0.0867, 0.0418, 1.0209, ..., -0.1391, 0.3836, -1.1330], [-0.0204, -0.1747, 0.0603, ..., 0.3766, 0.1290, -0.1035], [-0.2391, -0.2346, 0.0272, ..., 0.1989, 0.1316, 0.2529]]]), tensor([[[-0.2554, -0.7003, 0.6565, ..., 0.2290, 0.1936, -0.2808], [0.7636, -0.2991, -0.1245, ..., 0.6955, -0.5123, -0.1927], [-0.3194, -0.1029, 0.6832, ..., 0.4494, -0.2578, 0.0902], ..., [0.5714, -0.2095, 0.7303, ..., -0.2527, 0.1716, -0.7688], [0.0226, 0.0560, 0.0099, ..., 0.2366, 0.0284, -0.0035], [0.0523, -0.2752, 0.2031, ..., 0.2754, 0.2008, 0.1414]]]), tensor([[[-0.1345, -0.6517, 0.5974, ..., 0.1250, 0.0838, -0.2675], [0.9010, -0.0325, -0.4924, ..., 0.7549, -0.2886, -0.2095], [-0.1929, 0.1229, -0.0873, ..., 0.9852, -0.2818, 0.4708], ..., [1.1784, -0.7571, 1.1082, ..., -0.0929, -0.1834, -0.4198], [0.0896, 0.1353, 0.0409, ..., 0.1628, 0.0836, 0.0699], [0.1322, 0.0889, 0.1223, ..., 0.2258, 0.1343, 0.0196]]]), tensor([[[0.1813, -0.3751, 0.3770, ..., -0.1302, 0.0922, -0.4154], [0.3620, -0.0534, -0.7343, ..., 0.1011, -0.7042, -0.4897], [-0.1454, 0.1026, -0.1521, ..., 0.2790, -0.5408, -0.1365], ..., [1.2305, -0.0750, 0.9459, ..., -0.1514, 0.4930, -0.2138], [0.0575, 0.0791, 0.0185, ..., 0.0741, 0.0306, -0.0125], [0.0693, 0.0784, 0.0257, ..., 0.0846, 0.0318, -0.0365]]]), tensor([[[0.1239, -0.4247, 0.1897, ..., -0.2236, 0.2189, -0.4692], [1.1674, -0.4315, -0.9943, ..., -0.1113, 0.0480, -0.2320], [0.8771, -0.6990, -0.5200, ..., -0.1372, 0.0061, -0.0201], ..., [1.5614, -0.3155, 0.6422, ..., 0.1137, 0.8893, -0.1498], [0.2553, -0.6581, 0.4373, ..., -0.2505, 0.2505, 0.1159], [0.2926, -0.6715, 0.4440, ..., -0.2204, 0.2420, 0.0796]]]))

Thebe sample data (ca. 5000 rows)

Predict

Probleemgebied PUUR	Actiebeschrijving	Werkinstructies	OmschrijvingAssesmentgebied	InterventieActieGeldig Van	InterventieActieGeldigTm	Client
Neuro/musculaire/skelet functie	Valpreventie.	numerical representation	numerical representation	03/03/2021	01/01/2200	1
Neuro/musculaire/skelet functie	Valpreventie/omgevingsgevaren monitoren	numerical representation	numerical representation	02/06/2021	01/01/2200	2
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevaren	numerical representation	numerical representation	16/11/2021	01/01/2200	3
Neuro/musculaire/skelet functie	valpreventie	numerical representation	numerical representation	25/01/2022	01/01/2200	4
Neuro/musculaire/skelet functie	Valpreventie	numerical representation	numerical representation	18/01/2021	01/01/2200	5
Neuro/musculaire/skelet functie	Valpreventie	numerical representation	numerical representation	14/05/2021	01/01/2200	6
Neuro/musculaire/skelet functie	valpreventie	numerical representation	numerical representation	18/02/2021	01/01/2200	7
Neuro/musculaire/skelet functie	valpreventie	numerical representation	numerical representation	09/02/2021	17/02/2021	7
Neuro/musculaire/skelet functie	Valpreventie, mobiliteit,	numerical representation	numerical representation	20/01/2021	01/01/2200	9
Neuro/musculaire/skelet functie	aandacht voor valpreventie mevr is 10/12 verhuisd, indie dingen in de nieuwe woning niet handig/onveilig zijn gra melden bij zoco/wv	numerical representation	numerical representation	09/12/2021	01/01/2200	10
Neuro/musculaire/skelet functie	Aandacht voor valpreventie. Gebruik hulpmiddelen e.d.	numerical representation	numerical representation	14/01/2021	01/01/2200	11
Neuro/musculaire/skelet functie	a veilige verplaatsingstechnieken/lichaamsmotoriek b energiebesparing c valpreventie	numerical representation	numerical representation	22/10/2021	01/01/2200	12
Fysieke activiteit	valpreventie	numerical representation	numerical representation	14/10/2021	01/01/2200	12
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevare	numerical representation	numerical representation	21/07/2021	01/01/2200	14
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevaren b mobiliteit	numerical representation	numerical representation	01/06/2021	20/07/2021	14
Neuro/musculaire/skelet functie	- stimuleren behoud mobiliteit - stimuleren gebruik hulpmiddelen; valpreventie	numerical representation	numerical representation	25/05/2021	01/01/2200	16
Neuro/musculaire/skelet functie	- adviseren m.b.t. valpreventie bij verminderd evenwicht	numerical representation	numerical representation	01/03/2023	01/01/2200	17
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevaren	numerical representation	numerical representation	30/08/2021	05/09/2021	18
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevaren	numerical representation	numerical representation	10/03/2022	01/01/2200	19
Neuro/musculaire/skelet functie	Advies geven mbt Valpreventie.	numerical representation	numerical representation	07/01/2022	01/01/2200	20
Neuro/musculaire/skelet functie	Monitoren op dragen pols alarm Doel: veiligheid bij valpreventie	numerical representation	numerical representation	23/05/2022	01/01/2200	21
Neuro/musculaire/skelet functie	Valpreventie b mobiliteit	numerical representation	numerical representation	04/04/2023	01/01/2200	22
Neuro/musculaire/skelet functie	Valpreventie b mobiliteit	numerical representation	numerical representation	24/03/2023	03/04/2023	22
Neuro/musculaire/skelet functie	Valpreventie b mobiliteit	numerical representation	numerical representation	04/02/2023	23/03/2023	22
Neuro/musculaire/skelet functie	Valpreventie	numerical representation	numerical representation	11/01/2023	03/02/2023	22
Neuro/musculaire/skelet functie	- Monitoren mobiliteit - adviseren m.b.t. valpreventie.	numerical representation	numerical representation	25/01/2023	01/01/2200	26

This is the outcome (not yet labeled)

These columns were completely anonymised, we only have the BERT embedding

Thebe sample data

Probleemgebied PUUR	Actiebeschrijving	Werkinstructies	OmschrijvingAssesmentgebied	InterventieActieGeldig Van	InterventieActieGeldigTm	Client
Neuro/musculaire/skelet functie	Valpreventie.	numerical representation	numerical representation	03/03/2021	01/01/2200	1
Neuro/musculaire/skelet functie	Valpreventie/omgevingsgevaren monitoren	num		06/2021	01/01/2200	2
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevaren	num		11/2021	01/01/2200	3
Neuro/musculaire/skelet functie	valpreventie	num		01/2022	01/01/2200	4
Neuro/musculaire/skelet functie	Valpreventie	num		01/2021	01/01/2200	5
Neuro/musculaire/skelet functie	Valpreventie	num		05/2021	01/01/2200	6
Neuro/musculaire/skelet functie	valpreventie	num		02/2021	01/01/2200	7
Neuro/musculaire/skelet functie	valpreventie	num		02/2021	17/02/2021	7
Neuro/musculaire/skelet functie	Valpreventie, mobiliteit,	num		01/2021	01/01/2200	9
Neuro/musculaire/skelet functie	aandacht voor valpreventie mevr is 10/12 verhuisd, indien dingen in de nieuwe woning niet handig/onveilig zijn graag melden bij zoco/wv	num		12/2021	01/01/2200	10
Neuro/musculaire/skelet functie	Aandacht voor valpreventie. Gebruik hulpmiddelen e.d.	num		01/2021	01/01/2200	11
Neuro/musculaire/skelet functie	a veilige verplaatsingstechnieken/lichaamsmotoriek b energiebesparing c valpreventie	num		10/2021	01/01/2200	12
Fysieke activiteit	valpreventie	num		10/2021	01/01/2200	12
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevare	num		07/2021	01/01/2200	14
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevaren b mobiliteit	num		06/2021	20/07/2021	14
Neuro/musculaire/skelet functie	- stimuleren behoud mobiliteit - stimuleren gebruik hulpmiddelen; valpreventie	num		05/2021	01/01/2200	16
Neuro/musculaire/skelet functie	- adviseren m.b.t. valpreventie bij verminderd evenwicht	num		03/2023	01/01/2200	17
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevaren	num		08/2021	05/09/2021	18
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevaren	num		03/2022	01/01/2200	19
Neuro/musculaire/skelet functie	Advies geven mbt Valpreventie.	num		01/2022	01/01/2200	20
Neuro/musculaire/skelet functie	Monitoren op dragen pols alarm Doel: veiligheid bij valpreventie	num		05/2022	01/01/2200	21
Neuro/musculaire/skelet functie	Valpreventie b mobiliteit	num		04/2023	01/01/2200	22
Neuro/musculaire/skelet functie	Valpreventie b mobiliteit	num		03/2023	03/04/2023	22
Neuro/musculaire/skelet functie	Valpreventie b mobiliteit	num		02/2023	23/03/2023	22
Neuro/musculaire/skelet functie	Valpreventie	num		01/2023	03/02/2023	22
Neuro/musculaire/skelet functie	- Monitoren mobiliteit - adviseren m.b.t. valpreventie.	num		01/2023	01/01/2200	26

How to define outcome labels from "Actiebeschrijving"

1. Look for keyword/ letter combination (e.g. "val").

➔Also find "bevallen", "revalidatie"

➔Problem: different ways of registration

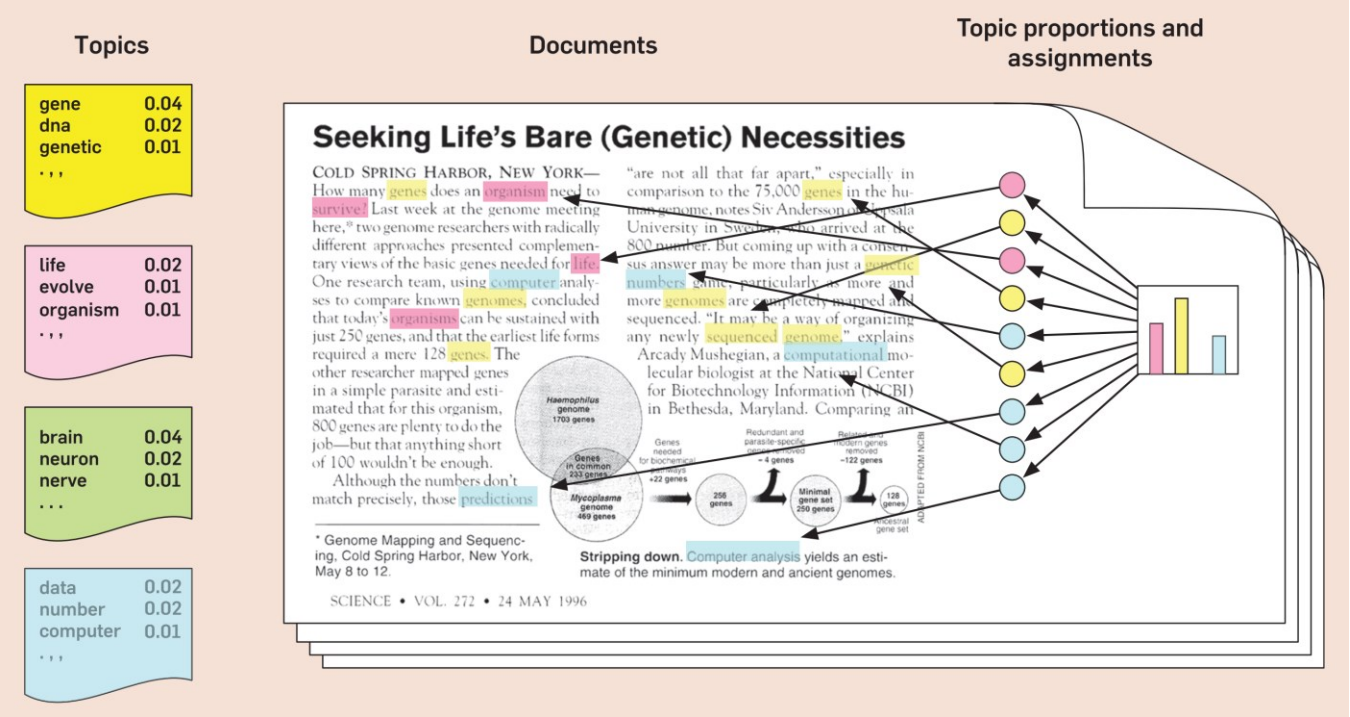
Clusters of free-text content: Topic Modeling

Medication

Disturbed Sleep

Risk of Falling

Etc…



Thebe sample data

Probleemgebied PUUR	Actiebeschrijving
Neuro/musculaire/skelet functie	Valpreventie.
Neuro/musculaire/skelet functie	Valpreventie/omgevingsgevaren monitoren
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevaren
Neuro/musculaire/skelet functie	valpreventie
Neuro/musculaire/skelet functie	Valpreventie
Neuro/musculaire/skelet functie	Valpreventie
Neuro/musculaire/skelet functie	valpreventie
Neuro/musculaire/skelet functie	valpreventie
Neuro/musculaire/skelet functie	Valpreventie, mobiliteit,
Neuro/musculaire/skelet functie	aandacht voor valpreventie mevr is 10/12 verhuisd, indien dingen in de nieuwe woning niet handig/onveilig zijn graag melden bij zoco/wv
Neuro/musculaire/skelet functie	Aandacht voor valpreventie. Gebruik hulpmiddelen e.d.
Neuro/musculaire/skelet functie	a veilige verplaatsingstechnieken/lichaamsmotoriek b energiebesparing c valpreventie
Fysieke activiteit	valpreventie
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevare
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevaren b mobiliteit
Neuro/musculaire/skelet functie	- stimuleren behoud mobiliteit - stimuleren gebruik hulpmiddelen; valpreventie
Neuro/musculaire/skelet functie	- adviseren m.b.t. valpreventie bij verminderd evenwicht
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevaren
Neuro/musculaire/skelet functie	a valpreventie/omgevingsgevaren
Neuro/musculaire/skelet functie	Advies geven mbt Valpreventie.
Neuro/musculaire/skelet functie	Monitoren op dragen pols alarm Doel: veiligheid bij valpreventie
Neuro/musculaire/skelet functie	Valpreventie b mobiliteit
Neuro/musculaire/skelet functie	Valpreventie b mobiliteit
Neuro/musculaire/skelet functie	Valpreventie b mobiliteit
Neuro/musculaire/skelet functie	Valpreventie
Neuro/musculaire/skelet functie	- Monitoren mobiliteit - adviseren m.b.t. valpreventie.

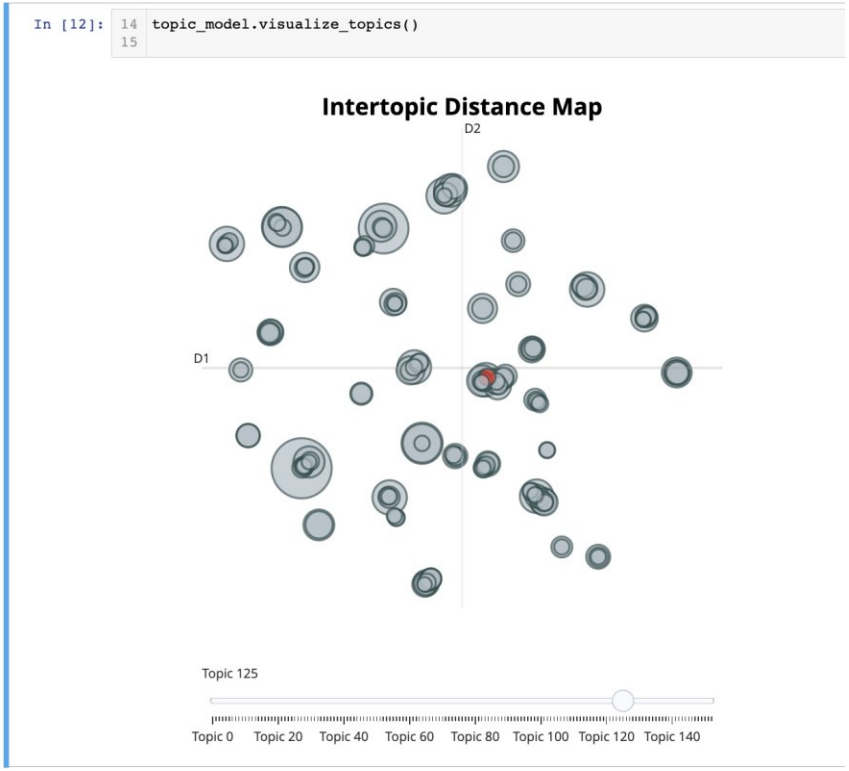
How to define outcome labels from "Actiebeschrijving"

1. Look for keyword/ letter combination (e.g. "val").
➔Also find "bevallen", "revalidatie"

2. Look in "semantic space" based on keyword ("val" as single word) -> which other words are used in the "actiebeschrijving" that are semantically related?
➔Circumvent specific terms
➔Build a "common dictionary" over VVT organizations

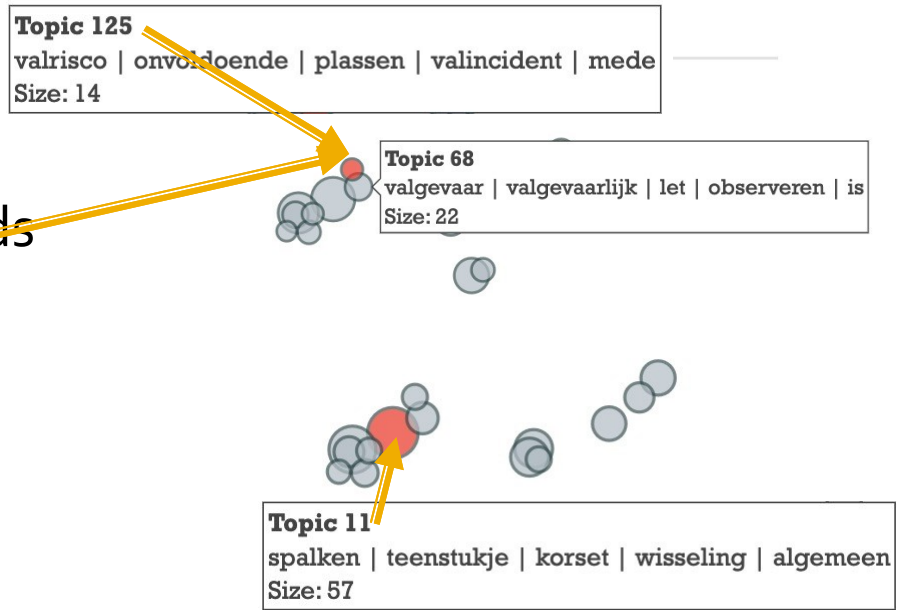
Approach

- BERT topic modelling
  - Ca 150 "topics"
  - Collection of related words
  - Topic 125:
    - valrisco
    - onvoldoende
    - plassen
    - valincident
    - mede
    - valrisico
    - incident
    - inzicht
    - doordat
    - ziekteproces



Approach

- BERT topic modelling
  - Ca 150 "topics"
  - Collection of related words
  - Topic 125:
    - valrisco
    - onvoldoende
    - plassen
    - valincident
    - mede
    - valrisico
    - incident
    - inzicht
    - doordat
    - ziekteproces



Interactive map of "Semantic Space"
Words ordered in topics clustered by similarity in **meaning** rather than actual words used

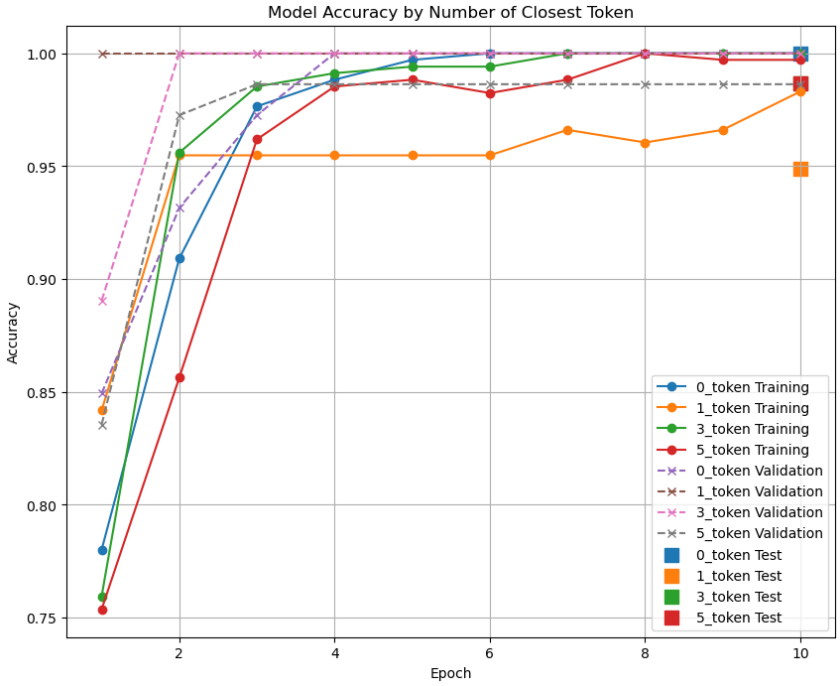
Approach

- BERT topic modelling
  - Ca 150 "topics"
  - Collection of related words
  - Topic 125:
    - valrisco
    - onvoldoende
    - plassen
    - valincident
    - mede
    - valrisico
    - incident
    - inzicht
    - doordat
    - ziekteproces

- Generate list of top-related words based on cue ("gevallen")
- 'valgevaar', 'verslechtering', 'verplaatsingstechnieken', 'lichaamsmotoriek', 'avondzorg', 'behoefte', '**valpreventie**', 'preventie', 'tek', 'nekkraag', 'nah', 'overig', 'beenprothese', 'orthese', 'liner', 'fentanylpleister', 'dgn', 'beademing', '**valrisco**', 'aanvragen', 'ergotherapeut', 'svt', 'pac', 'intrathecale', 'verwijzen', 'samenwerking', 'ondersteunend', 'materiaal', 'behoefde', 'levensfase', 'allerlei', 'sling', 'losmaken'
- Use these words to label cases for "increased fall risk"
- Train classifier

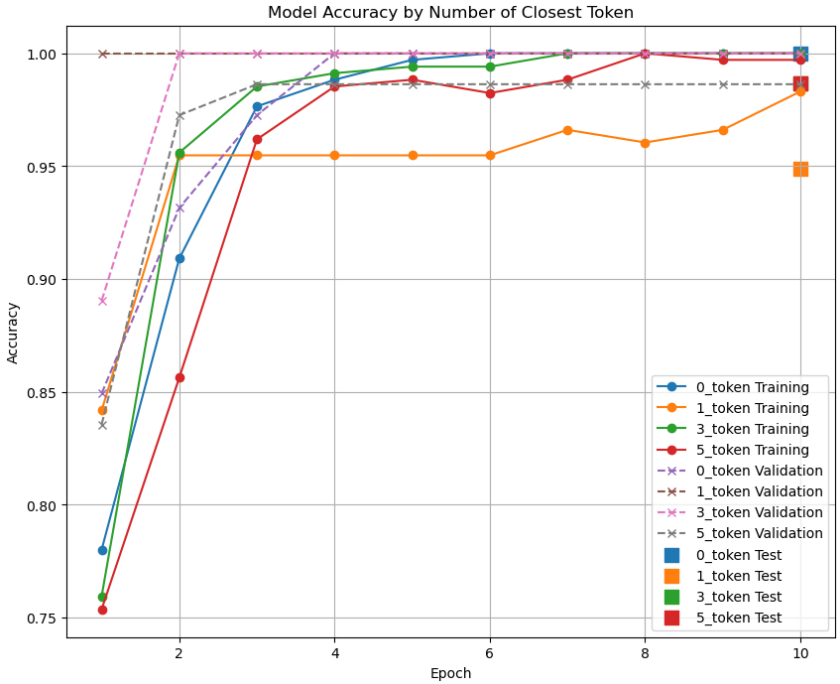
Results

- Sample dataset with 244 cases that are fall-related
  - Selected for having words close to "val"
  - Supplemented by 244 cases that are very *unlikely* to be fall-related
  - Train classifier to separate cases from non-cased
    - Accuracy crosses 95% correct
    - Easy, example task
  - This means that the free-text fields contain adequate information to separate fall-reports from non-fall reports



Results

- PROBLEM: BERTje embeddings can be reversed when original model is at hand
  - Not unlikely, not that many models exist yet, but possible to "brute force"
- Solution: change the input by substituting semantically close words before export of embedding
  - 5-token (in the graphs) means: the substituted word is randomly chosen from the top-5 related words.
  - Reconstructed sentence now sufficiently different from input
  - Classifier still works (on toy data)



Method uses FAISS
https://github.com/facebookresearch/faiss

Similarity search in high-dimensional vector space (e.g. LLM embeddings)

Conclusions

- When dealing with health data, involve domain experts early on
  - Understanding the actual question
  - Understanding the data
  - Understanding possible implementations
    - End-user needs

"we should do something with AI and data science…"