Utrecht University

university of groningen

LÄKEMEDELSVERKET
SWEDISH MEDICAL PRODUCTS AGENCY

UPPSALA UNIVERSITET

CBG MEB
MEDICINES EVALUATION BOARD

A Natural Language Processing Approach

Towards Harmonized Communication of Uncertainties Identified During the European Medicine Authorization Process

Stefan Verweij

Data scientist @ Dutch Medicines Evaluation Board
PhD-candidate @ University of Groningen

GOOD MEDICINES USED BETTER

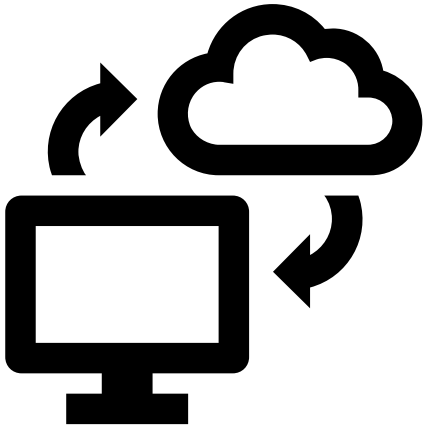
The regulatory setting

- European Medicines Agency (EMA) is responsible for the **decision** of authorization of **medicines**

- Decision is based on extensive details on evidence concering
  Efficacy, **Safety** and **Quality**.

- These decisions are published in the
  European Public Assessment Report (EPAR)
  to ensure transparency

Regulatory work: a paper mill



A dream for NLP

EPAR

CBG MEB
MEDICINES EVALUATION BOARD

Assessment history

Changes since initial authorisation of medicine

Initial marketing authorisation documents

Zynteglo : EPAR - Public assessment report

Adopted

Reference Number: EMA/56140/2020

English (EN) (5.58 MB - PDF)

First published: 03/06/2019 **Last updated:** 14/02/2020

View

Zynteglo : Orphan maintenance assessment report (initial authorisation)

EPAR

CBG MEB
MEDICINES EVALUATION BOARD

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

26 April 2019
EMA/56140/2020/Corr.[1]
Committee for Medicinal Products for Human Use (CHMP)

Assessment report

Zynteglo

International non-proprietary name: betibeglogene autotemcel

Procedure No. EMEA/H/C/003691/0000

nger authorised

143 pages

GOOD MEDICINES
USED BETTER

Decision: Benefit Risk balance

CBG MEB
MEDICINES EVALUATION BOARD

B/R

Evidence often from controlled trials

- Efficacy

- Adverse events
- Toxicity

- What is the disease burden?
- Are there any other medicines on the market already?
- Is it symptomatic or does it cure the disease

Sometimes, uncertainties in the B/R remain

It isn't always feasible to study everything before marketing authorization of the medicinal product

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

26 April 2023
EMA/227054/2023
Committee for Medicinal Products for Human Use (CHMP)

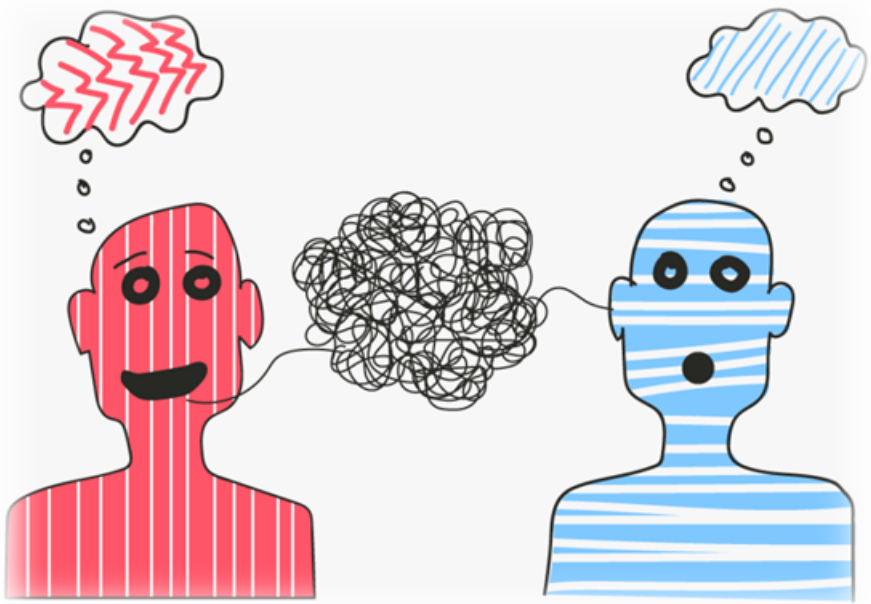
CHMP assessment report

3.5. Uncertainties and limitations about unfavourable effects

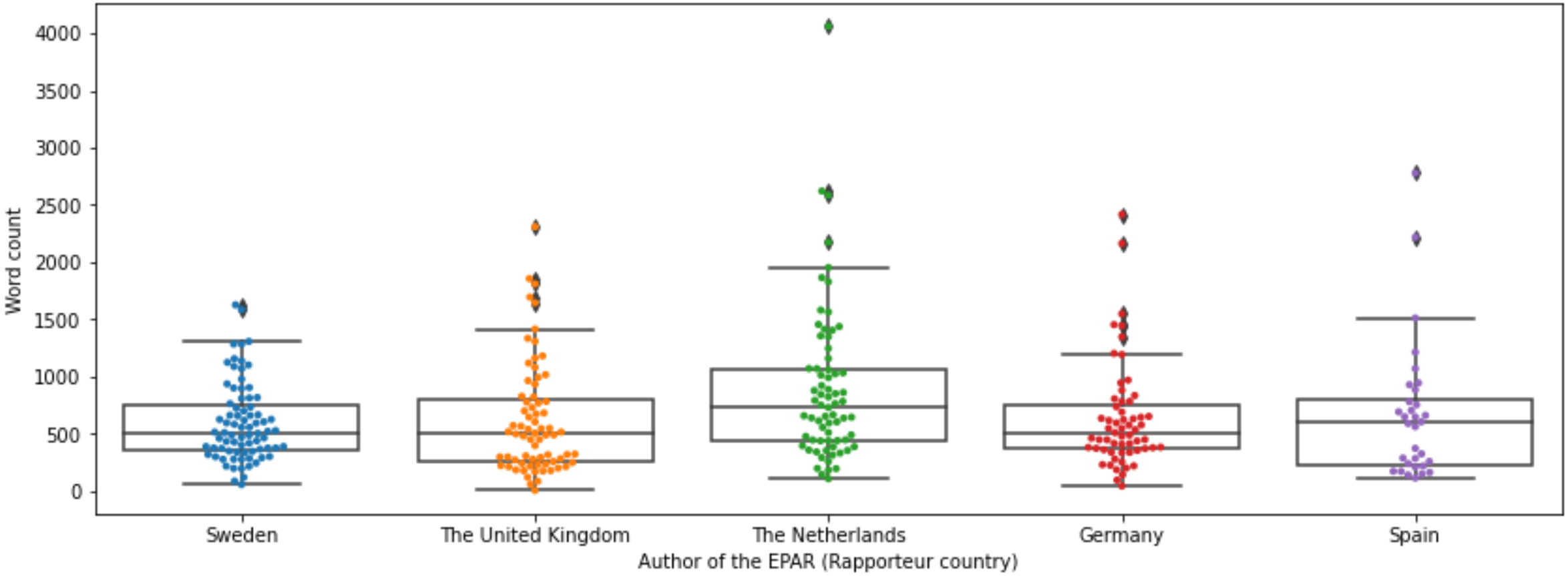
Although 7 months additional safety data has been submitted by the applicant long-term exposure to risankizumab (>18months) is limited. Only 6.6% were exposed for more than 2 years and only 3.4% exposed for more than 3 years in clinical trials. This extent of exposure is insufficient to fully characterize the unfavourable effects particularly those with a long induction period (malignancy) or those that might

The problem

- EPARs are written by **different countries** & **different authors**
- **No common language** or **taxonomy** to describe *uncertainties*
- Differences in **culture** and **linguistic** background



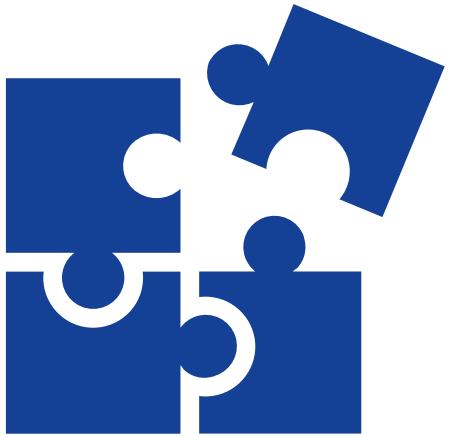
The Dutch used the most number of words to describe uncertainties



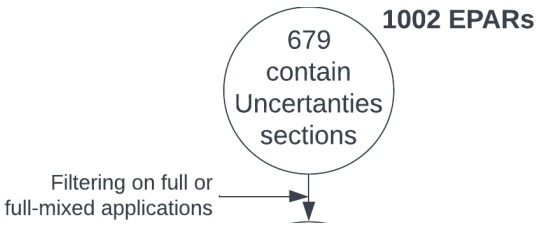
The solution

Cluster uncertainties on their overarching topics

as a first steppingstone towards harmonization



How?

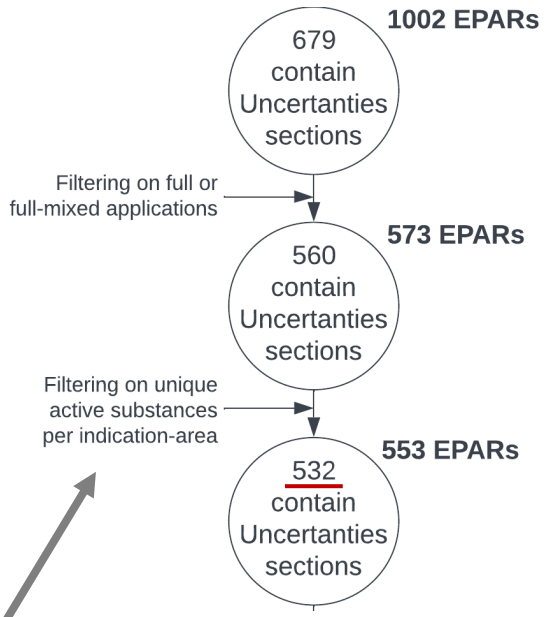


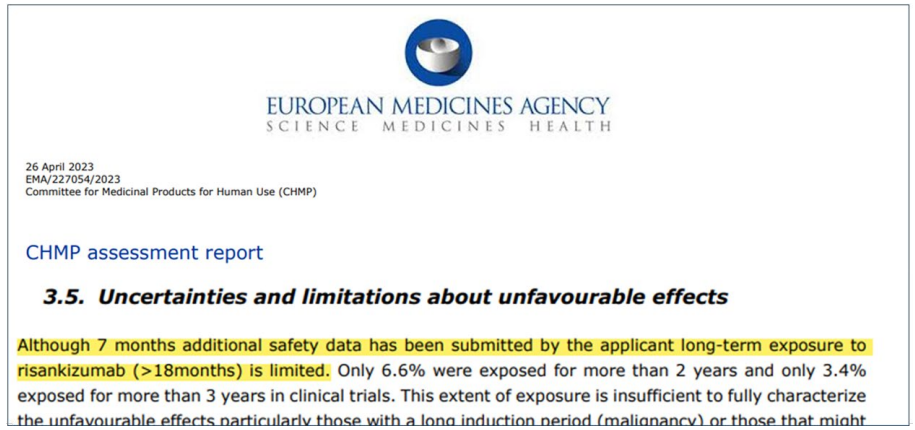
- No biosimilar
- No generic
- No well-established use

Innovative medicines ony

Method from Bergman, E., Sherwood, K., Forslund, M., Arlett, P., & Westman, G. (2022). A natural language processing approach towards harmonisation of European medicinal product information. *Plos one*, *17*(10), e0275386.

How?





EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

26 April 2023
EMA/227054/2023
Committee for Medicinal Products for Human Use (CHMP)

CHMP assessment report

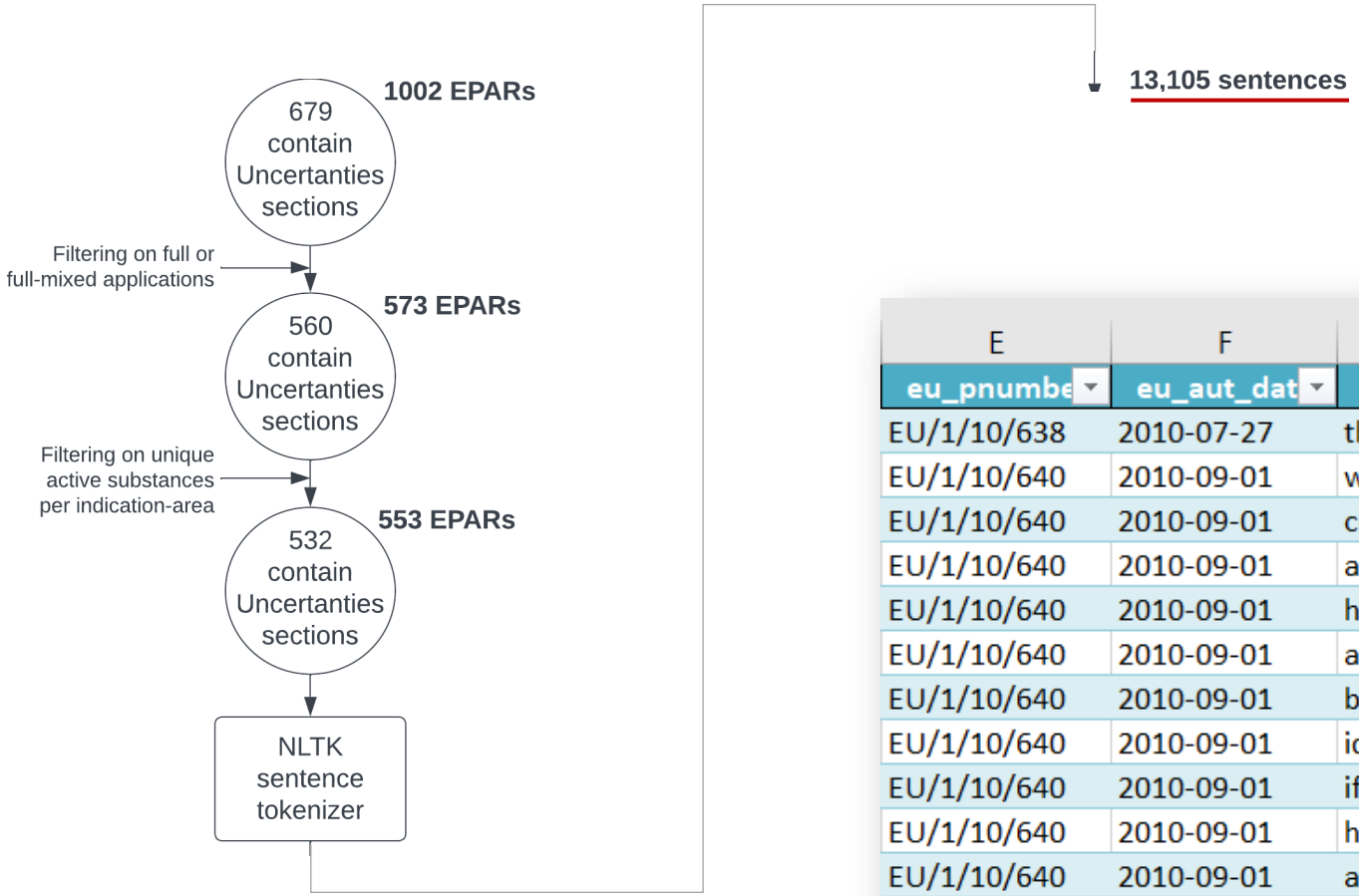
3.5. Uncertainties and limitations about unfavourable effects

Although 7 months additional safety data has been submitted by the applicant long-term exposure to risankizumab (>18months) is limited. Only 6.6% were exposed for more than 2 years and only 3.4% exposed for more than 3 years in clinical trials. This extent of exposure is insufficient to fully characterize the unfavourable effects particularly those with a long induction period (malignancy) or those that might

No commercial duplicates

Method from Bergman, E., Sherwood, K., Forslund, M., Arlett, P., & Westman, G. (2022). A natural language processing approach towards harmonisation of European medicinal product information. *Plos one*, *17*(10), e0275386.

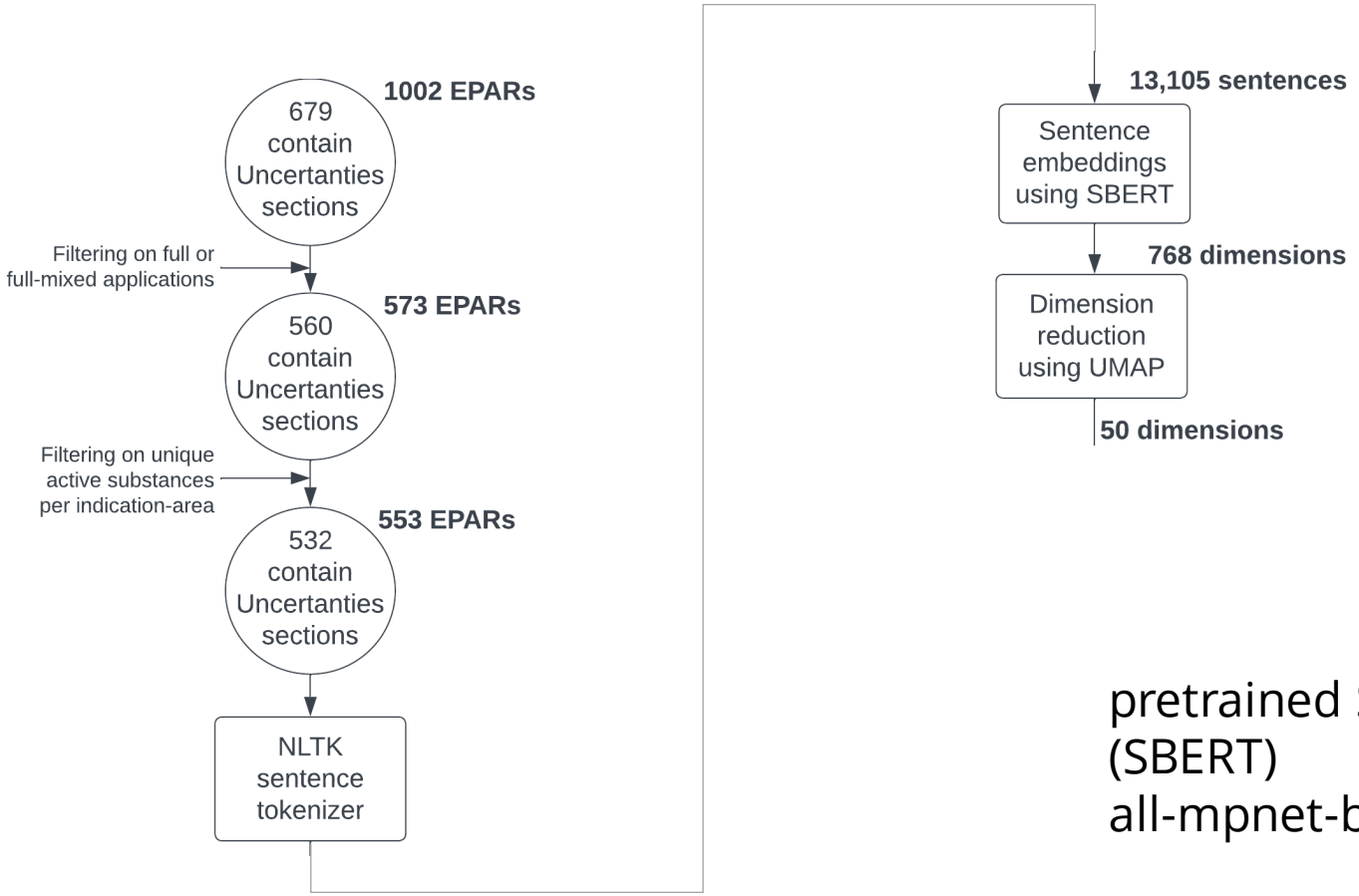
How?



E	F	G
eu_pnumbe	eu_aut_dat	lines
EU/1/10/638	2010-07-27	there is a lack of experience with repeat administ
EU/1/10/640	2010-09-01	with respect to phase ii shortterm trials 041002, 0
EU/1/10/640	2010-09-01	compared with placebo, olanzapine treatment res
EU/1/10/640	2010-09-01	analyses of the primary endpoint using oc and mm
EU/1/10/640	2010-09-01	however, the pooled 95 ci was very close to zero.
EU/1/10/640	2010-09-01	adding in the data from the phase 2 study 041004
EU/1/10/640	2010-09-01	but as this combines the hypothesis generating an
EU/1/10/640	2010-09-01	ideally the phase 3 data alone should provide com
EU/1/10/640	2010-09-01	if short term benefit could be established this wou
EU/1/10/640	2010-09-01	however, in the absence of shortterm benefit hav
EU/1/10/640	2010-09-01	at baseline, the panss total scores were 92.1 for b
EU/1/10/640	2010-09-01	at week 52, the panss total scores were 71.0 and
EU/1/10/640	2010-09-01	however, this treatment difference was not statis
EU/1/10/640	2010-09-01	a statistically significant greater percentage of sul
EU/1/10/640	2010-09-01	additionally, 35.5 of subjects in the asenapine gro
EU/1/10/640	2010-09-01	overall, the chmp concluded that noninferiority o

Method from Bergman, E., Sherwood, K., Forslund, M., Arlett, P., & Westman, G. (2022). A natural language processing approach towards harmonisation of European medicinal product information. *Plos one*, *17*(10), e0275386.

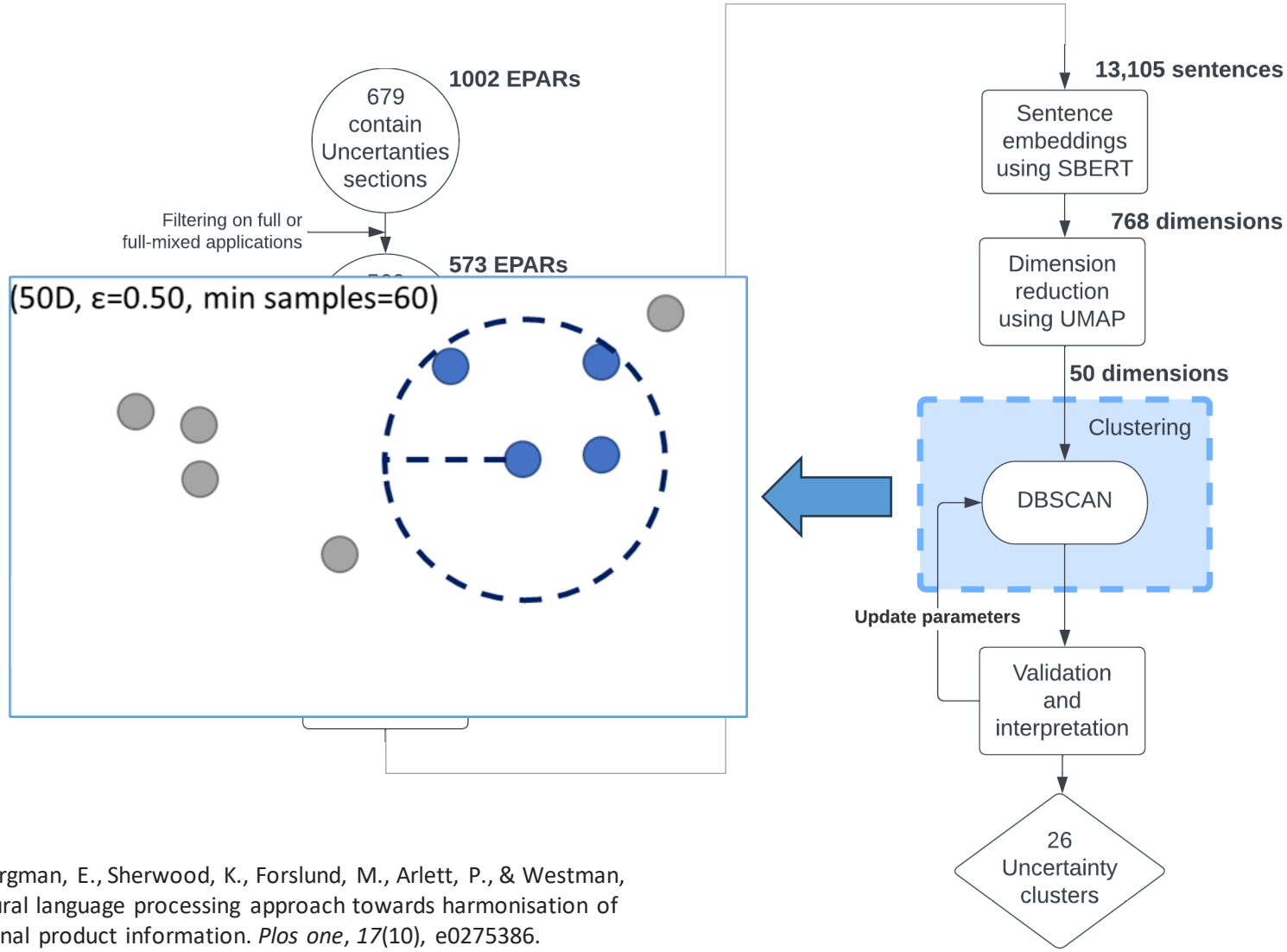
How?



pretrained Sentence-BERT (SBERT)
all-mpnet-base-v2 model

Method from Bergman, E., Sherwood, K., Forslund, M., Arlett, P., & Westman, G. (2022). A natural language processing approach towards harmonisation of European medicinal product information. *Plos one*, *17*(10), e0275386.

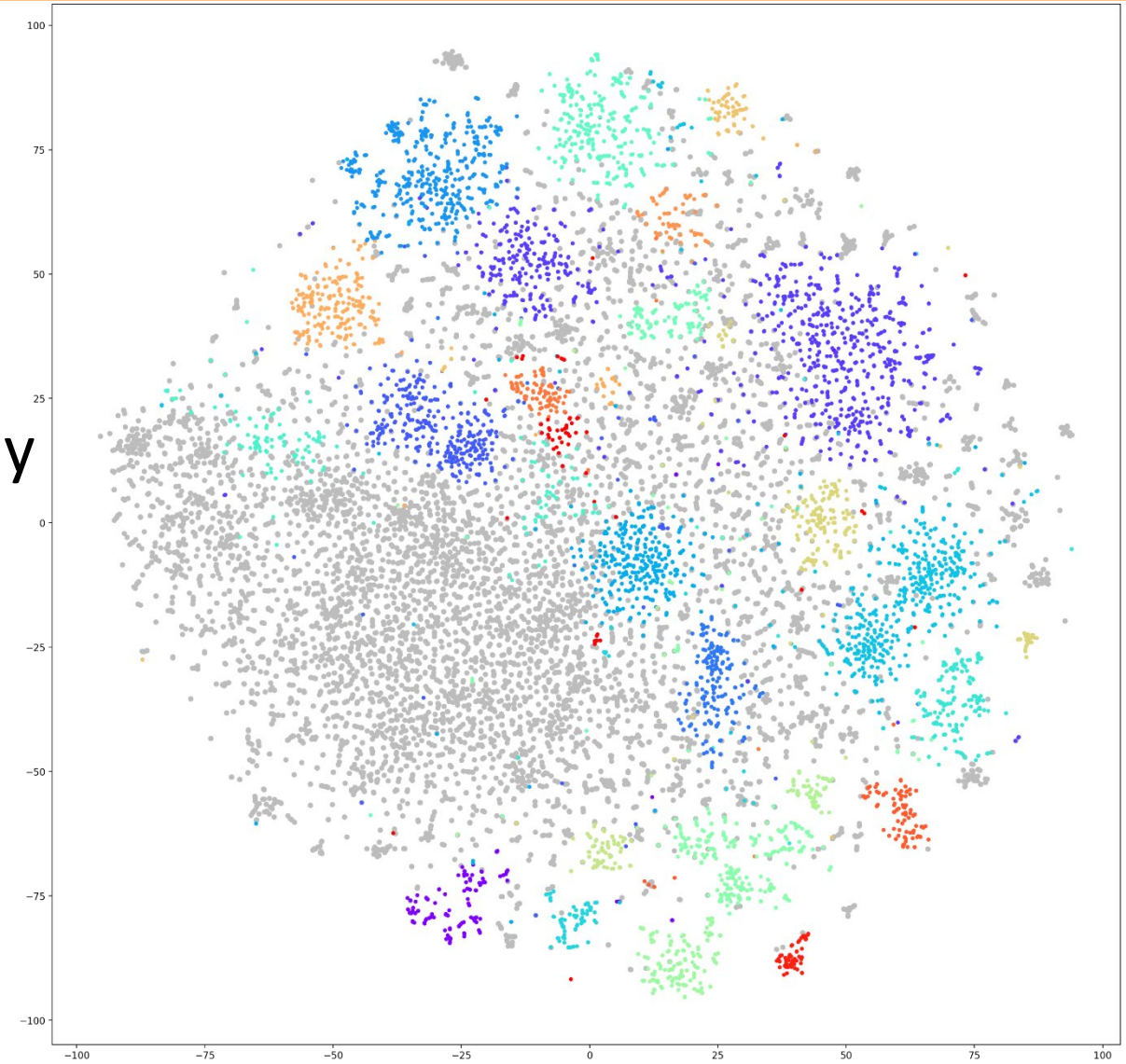
How?



Method from Bergman, E., Sherwood, K., Forslund, M., Arlett, P., & Westman, G. (2022). A natural language processing approach towards harmonisation of European medicinal product information. *Plos one*, *17*(10), e0275386.

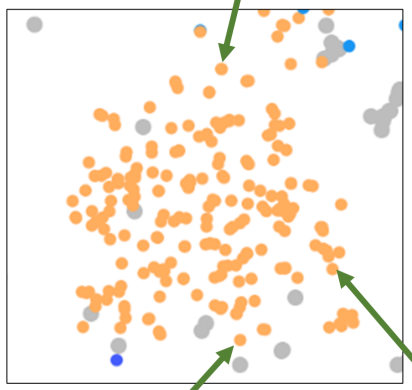
26 clusters

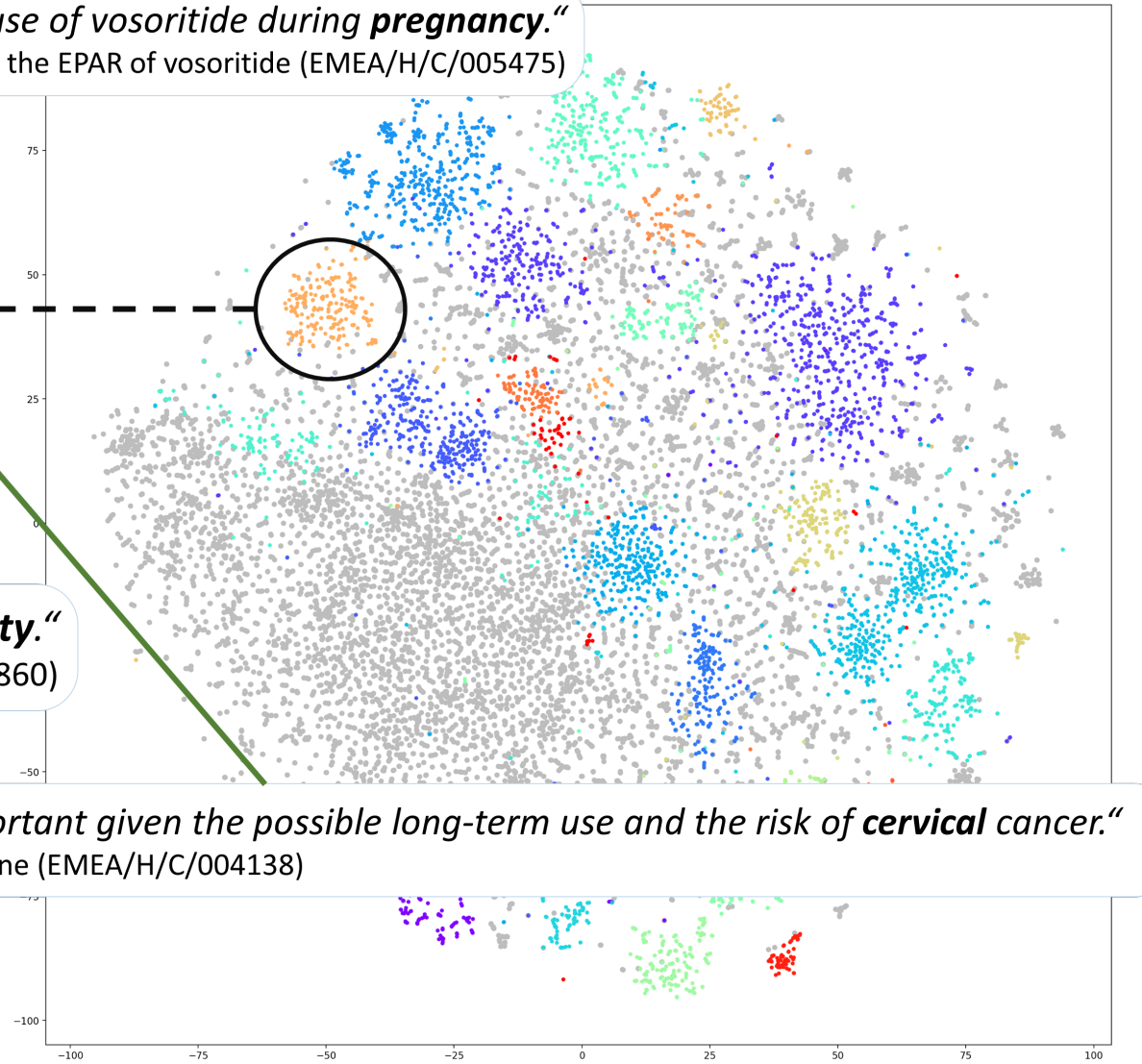
Each cluster consists of semantically similar sentence embeddings



26 clusters

*"There is no data regarding the use of vosoritide during **pregnancy**."*
From the EPAR of vosoritide (EMEA/H/C/005475)

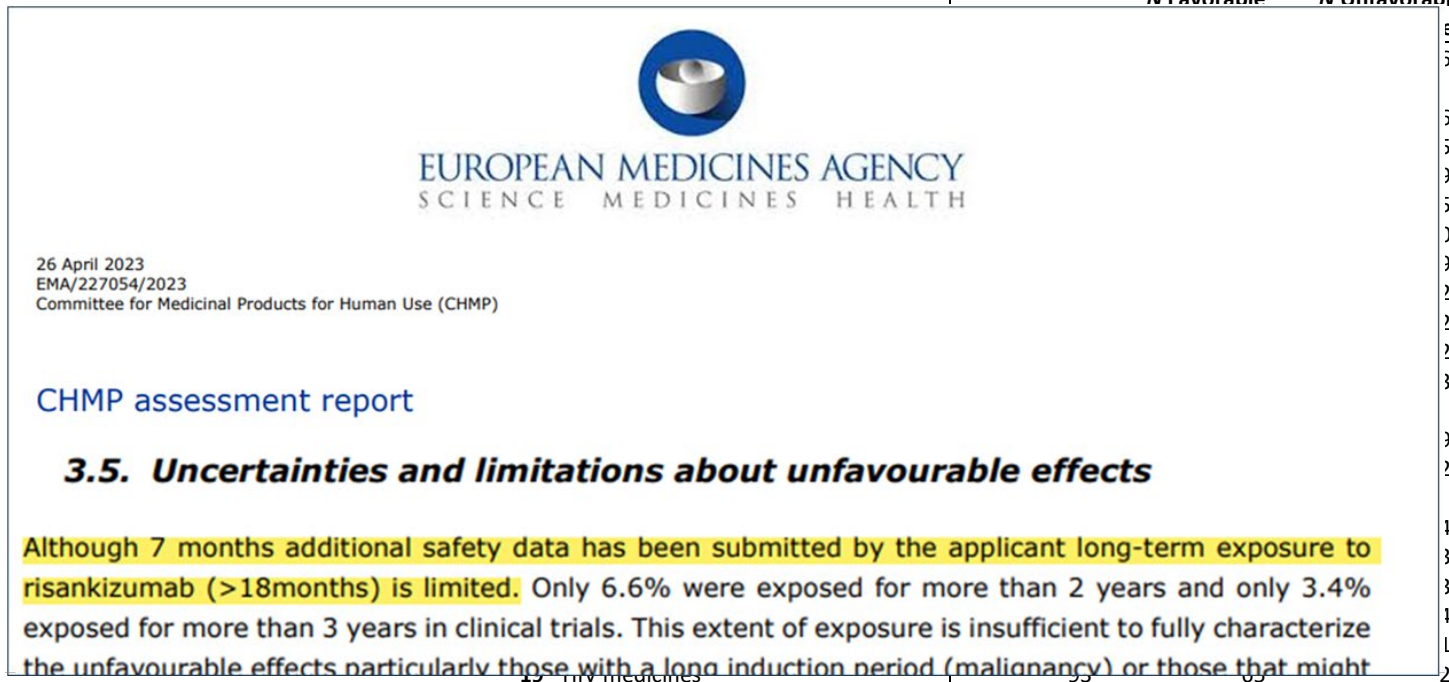




*"Animal studies did not indicate **reproductive toxicity**."*
From the EPAR of mepolizumab (EMEA/H/C/003860)

*"This concern is important given the possible long-term use and the risk of **cervical** cancer."*
From the EPAR of prasterone (EMEA/H/C/004138)

26 clusters



EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

26 April 2023
EMA/227054/2023
Committee for Medicinal Products for Human Use (CHMP)

CHMP assessment report

3.5. Uncertainties and limitations about unfavourable effects

Although 7 months additional safety data has been submitted by the applicant long-term exposure to risankizumab (>18months) is limited. Only 6.6% were exposed for more than 2 years and only 3.4% exposed for more than 3 years in clinical trials. This extent of exposure is insufficient to fully characterize the unfavourable effects particularly those with a long induction period (malignancy) or those that might

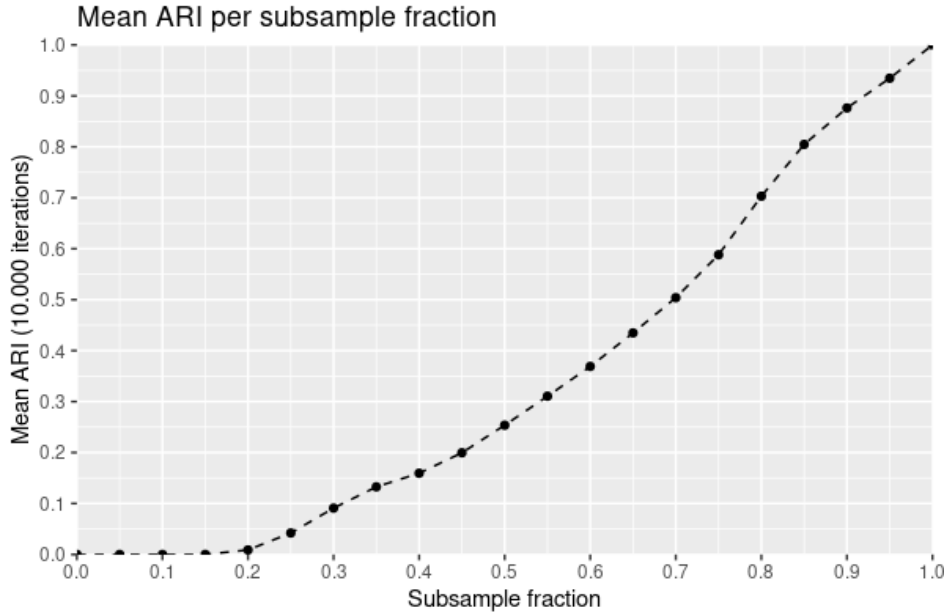
		Cluster information					
			N Favorable	*N Unfavorable*	% Noise	*N Unique innovative medicines*	*N Unique active substances*
				es			
				54	2.4%	25	25
				53	100.0%	504	492
				56	2.0%	254	249
				93	5.9%	185	180
				57	14.0%	95	94
				04	0.2%	49	47
				94	5.9%	158	155
				24	4.3%	195	193
				24	4.5%	15	15
				25	1.5%	25	25
				36	19.4%	130	129
				92	2.9%	69	68
				22	5.0%	28	28
				45	10.4%	55	53
				38	1.6%	38	36
				38	0.0%	7	7
				45	0.0%	18	18
				15	0.0%	51	50
19	HIV medicines	95	65	28	1.1%	8	8
20	Pregnancy and fertility	259	39	220	0.8%	97	95
21	Antibiotics and antimycotics	79	65	14	0.0%	15	15
22	Patient group differences in adverse event occurrence	100	2	98	1.0%	65	64
23	Effects on LDL and LDL-lowering medicines	132	54	78	3.8%	14	14
24	Noise	50	45				
25	Migraine and headache	60	45				
26	Adverse events and injections	65	8				
n.a.	Outliers	4866	2956	1956	100.0%	492	480
	Sum (mean for noise)	13105	6899	6206	3.9%		

>7000 sentences were either noise or outliers

Sensitivity analysis

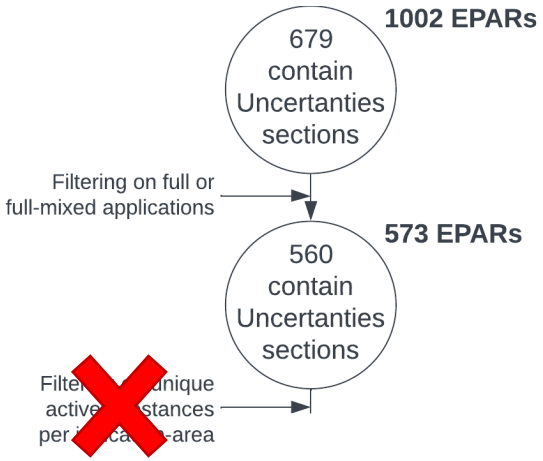
Bootstrap sampling (10,000 iterations)
Mean Adjusted Rand Index (ARI): **0.87** (95% CI: **0.81-0.91**)

Subsampling (sample fraction of 0.9 for 10,000 iterations)
Mean ARI: **0.88** (95% CI: **0.85-0.90**)



But…. Input data affects outcomes!

Secondary analysis



4 new clusters related to *Risk management*

- *Communication of uncertainties*
- *Post-authorization risk management*
- *Conclusion/interpretation of uncertainties*
- *Lacking outcome data*

Future: EMRD

Utrecht University

CBG MEB

European Medicines Regulatory Database

https://emrd.uu.nl

Explore by Name, Code, Brand, Condition, etc. (COVID-19, EU/1/20, L01xx, medisee, impairment)

Search

Human Selection Table - last updated: 20-6-2024

Filter & Sort

LAUNCH Q3/Q4 2024

EU Product Number	Active substance	Brand name (EU, current)	Marketing Authorisation I	Authorisation date (EU)
EU/1/95/001	Follitropin alfa	GONAL-f	Merck Europe B.V.	1995-10-20
EU/1/95/002	Docetaxel	Taxotere	Sanofi Winthrop Industrie	1995-11-27
EU/1/95/003	Interferon beta-1b	Betaferon	Bayer AG	1995-11-30
EU/1/96/004	Toremifene	Fareston	Orion Corporation	1996-02-14
EU/1/96/005	Mycophenolate mofetil	CellCept	Roche Registration GmbH	1996-02-14
EU/1/96/006	Eptacog alfa (activated)	NovoSeven	Novo Nordisk A/S	1996-02-23
EU/1/96/007	Insulin lispro	Humalog	Eli Lilly Nederland B.V.	1996-04-30
EU/1/96/008	follitropin beta	Puregon	N.V. Organon	1996-05-03

Discussion

+ Relatively easy application of NLP

- Lots of outliers & noise (>50%) in clustering
  - E.g., due to multi-sentence reasoning
  - Context of punctuation (*drug A vs.*)
  - …

- **First step towards harmonization (?)**

https://doi.org/10.1002/cpt.3195

Link to the paper →

Utrecht University

university of groningen

LÄKEMEDELSVERKET
SWEDISH MEDICAL PRODUCTS AGENCY

UPPSALA UNIVERSITET

CBG MEB
MEDICINES EVALUATION BOARD

GOOD MEDICINES USED BETTER

UMAP: **U**niform **M**anifold **A**pproximation and **P**rojection

To allow clustering of *nonlinear* geometries in embedding space

That is

Being able to identify **underlying structure in a sentence** of a lower intrinsic dimension