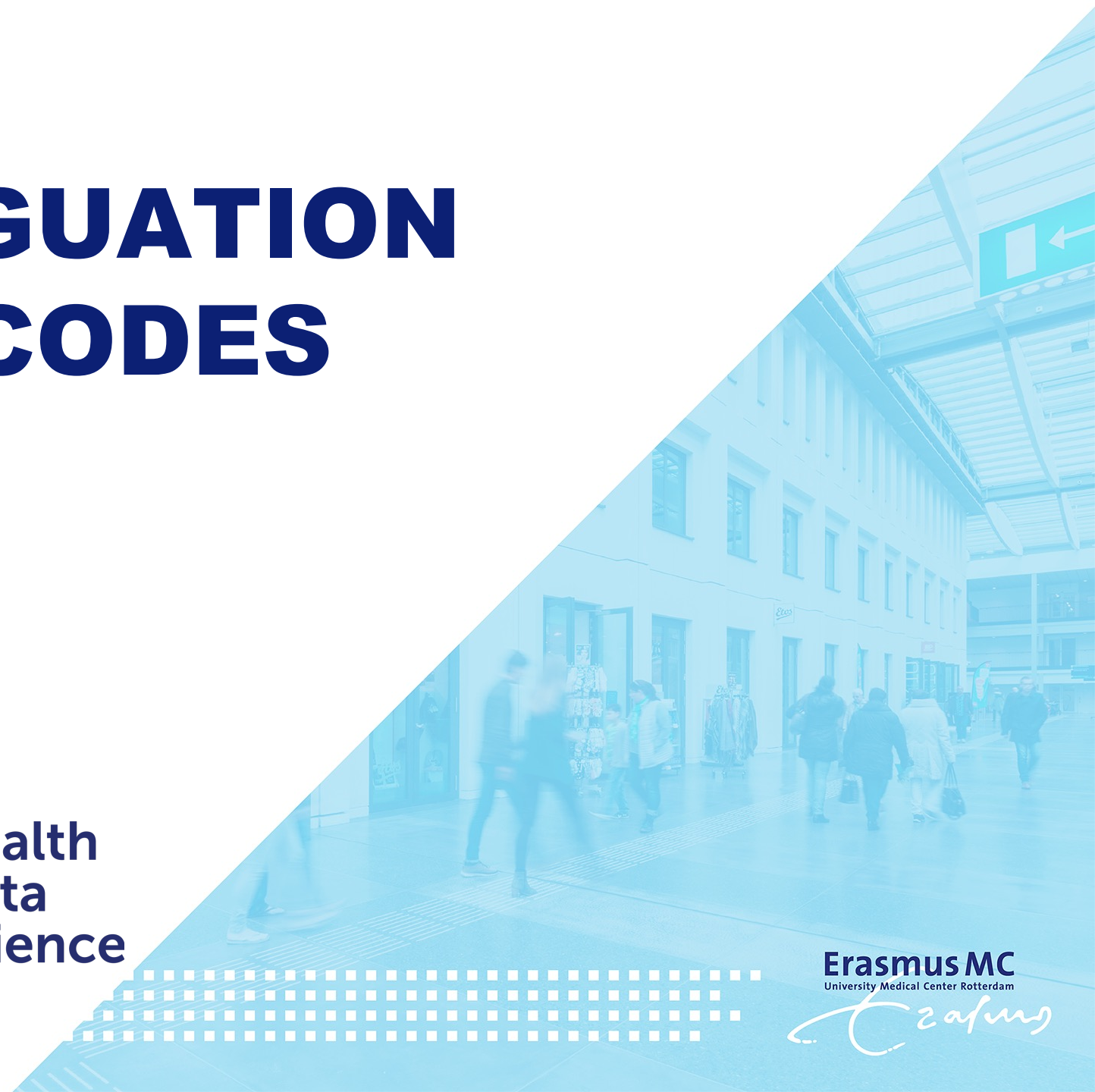
DISAMBIGUATION OF ICPC CODES

Tom Seinen
PhD Student – Erasmus MC



EHDEN
EUROPEAN HEALTH DATA & EVIDENCE NETWORK

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

Health Data Science

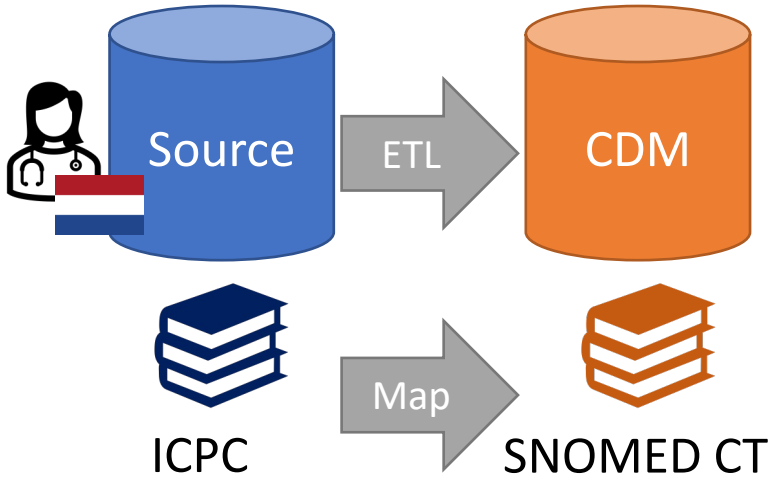
imi innovative medicines initiative

efpia

Erasmus MC
University Medical Center Rotterdam

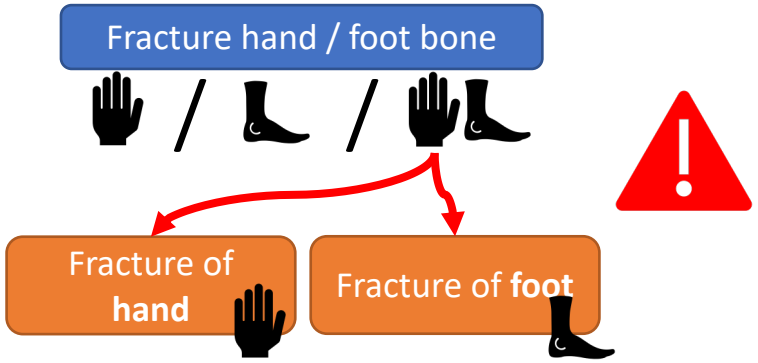
Context & Problem

- **Dutch general practitioner database:**
- 2.5 million patients
- **I**nternational **C**lassification of **P**rimary **C**are (**ICPC**)



Difficult to map codes:

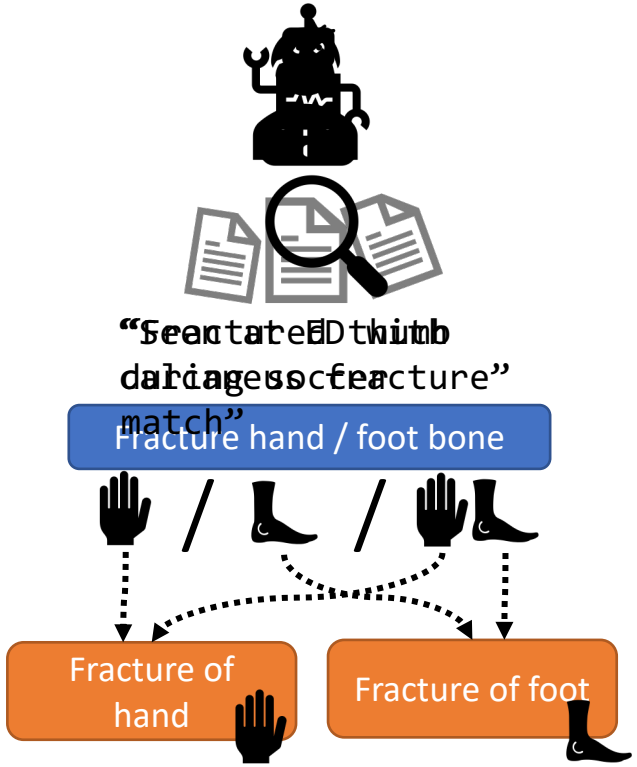
- L74: Fracture hand / foot bone
- K78: Atrial fibrillation / flutter
- D75: Malignant neoplasm colon / rectum



Problem: inaccurate, noisy data

Research question

- **Map single codes manually:**
- Consulting **free-text** clinical notes
- **Around** the registered ICPC code
- **This takes a lot of work:**
- **Thousands** of observations per ICPC code

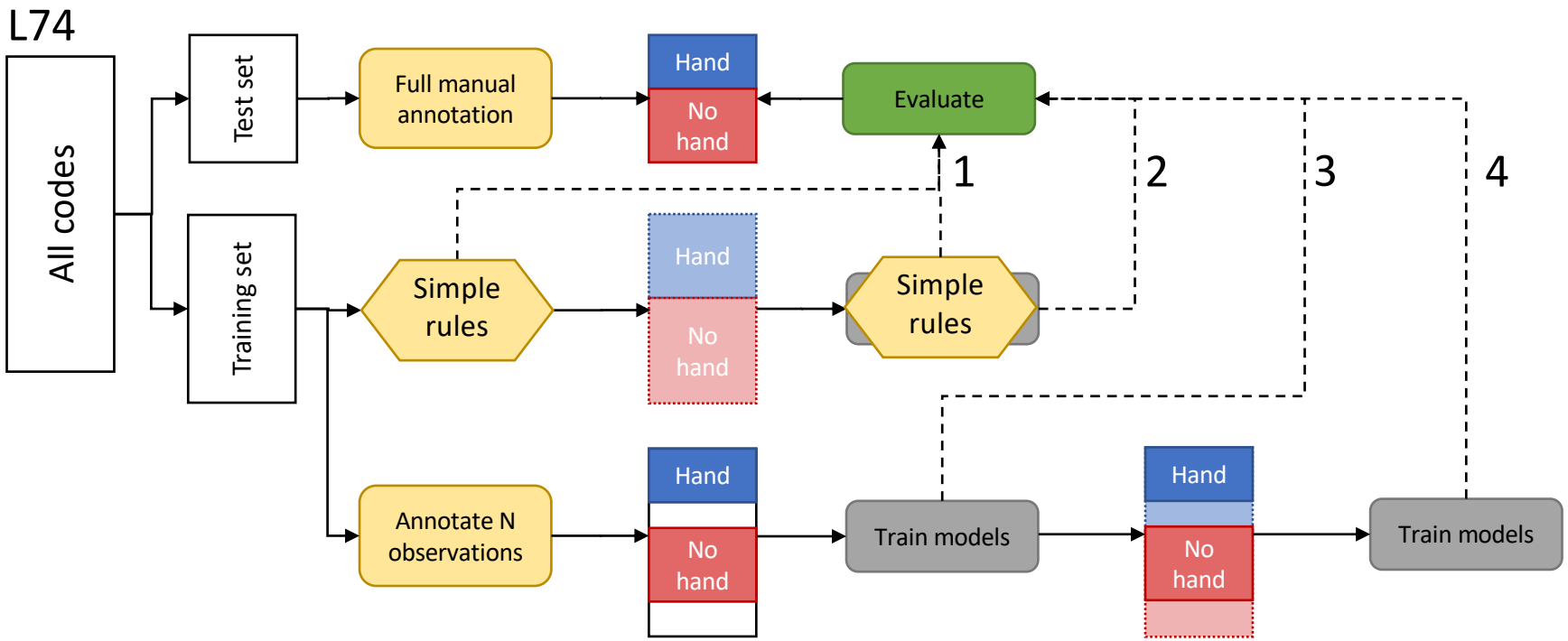


Can we do this automatically?

- **Classification** problem
- **Predict** the right **concepts**
- Given the **surrounding data**

Methods

- Classifier: Lasso
- Features: TFIDF



Simple rules:
“hand”, “wrist”, “finger”
“foot”, “ankle”, “toe”

Results

L74: Fracture hand / foot bone

Performance on test set		Hand		Foot	
	# ann.	AUC	AUPRC	AUC	AUPRC
1. Rules	*NA*	0.87	0.82	0.90	0.83
2. Rules gen.	*NA*	0.91	0.91	0.96	0.95
3. Ann.	100	0.97	0.97	0.97	0.96
4. Ann gen.	100	0.97	0.97	0.96	0.94
3. Ann.	200	**0.98**	0.98	**0.99**	**0.99**
4. Ann gen.	200	0.98	0.98	0.99	0.99
3. Ann.	300	0.98	0.98	0.99	0.99
4. Ann gen.	300	0.98	**0.99**	0.99	0.99

Results

D75: Malignant neoplasm colon / rectum

Performance on test set		**Colon**		**Rectum**	
	# ann.	AUC	AUPRC	AUC	AUPRC
1. Rules	*NA*	0.76	0.71	0.81	0.62
2. Rules gen.	*NA*	0.70	0.74	0.84	0.74
3. Ann.	100	0.72	0.70	0.83	0.68
4. Ann gen.	100	0.75	0.73	0.82	0.66
3. Ann.	200	0.81	0.84	0.85	0.77
4. Ann gen.	200	0.83	0.85	0.87	0.80
3. Ann.	300	**0.85**	**0.87**	**0.90**	**0.84**
4. Ann gen.	300	0.85	0.87	0.89	0.83

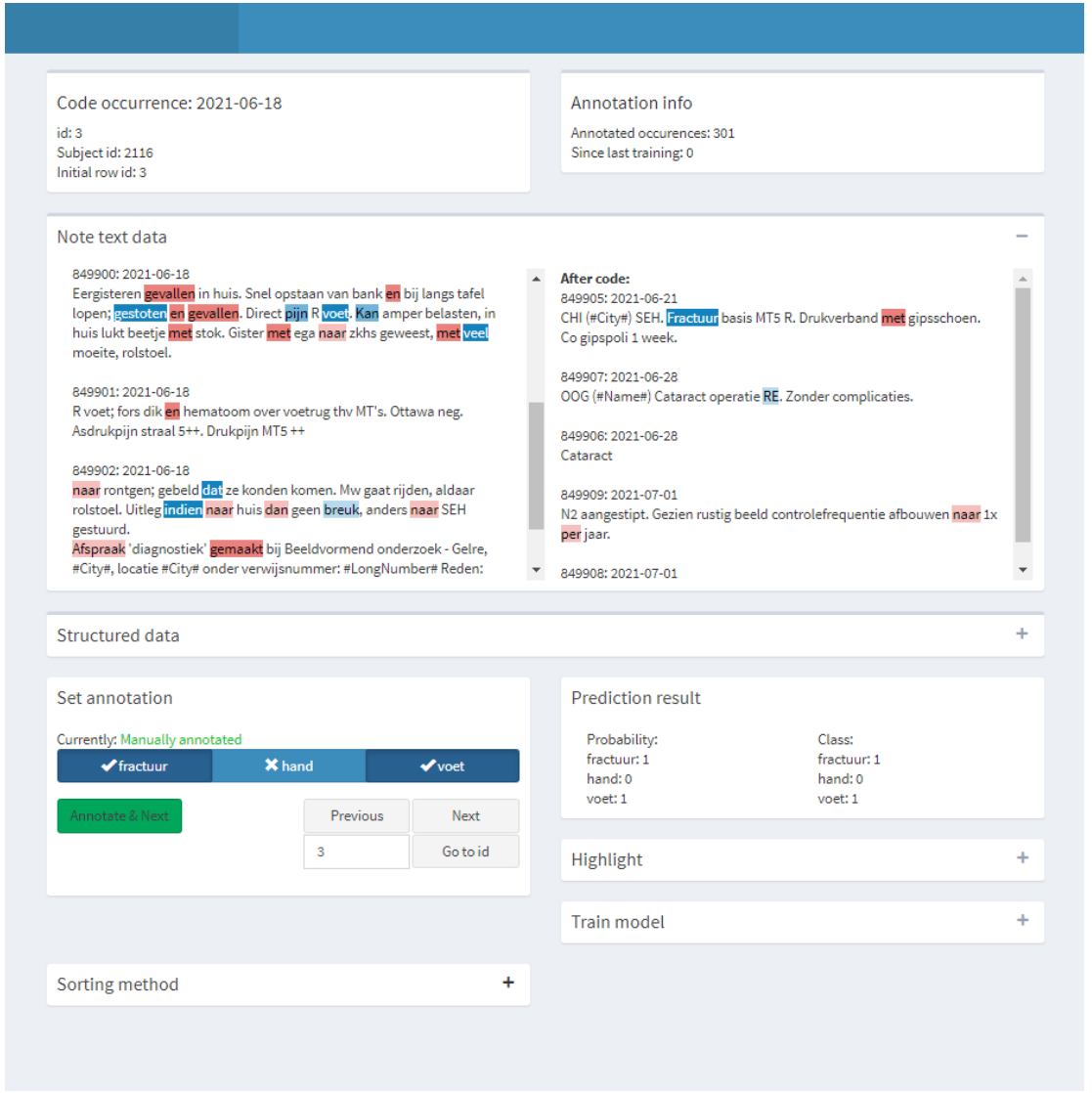
Annotation tool

Lightweight R-shiny application

Runs directly on the **CDM**

Active learning to **sort** observations

Visualize **feature importance**



Conclusion

- **Possible**: Automatically **classify** individual **ICPC codes** into **narrower OMOP concepts**
- **Practical use:** improve concept **mappings**
- **Performance** difference between:
  - ICPC codes
  - Modelling methods
- **More info** & **future steps** or **questions**?
- Meet me at my **poster**

We classified **ambiguous source codes** into narrower **OMOP concepts**