Topzorg voor uw levenskwaliteit

zgt

UNIVERSITY OF TWENTE.

Comorbidity identification in clinical documents with weak supervision.

Sylvain Brouwer

Supervisors:

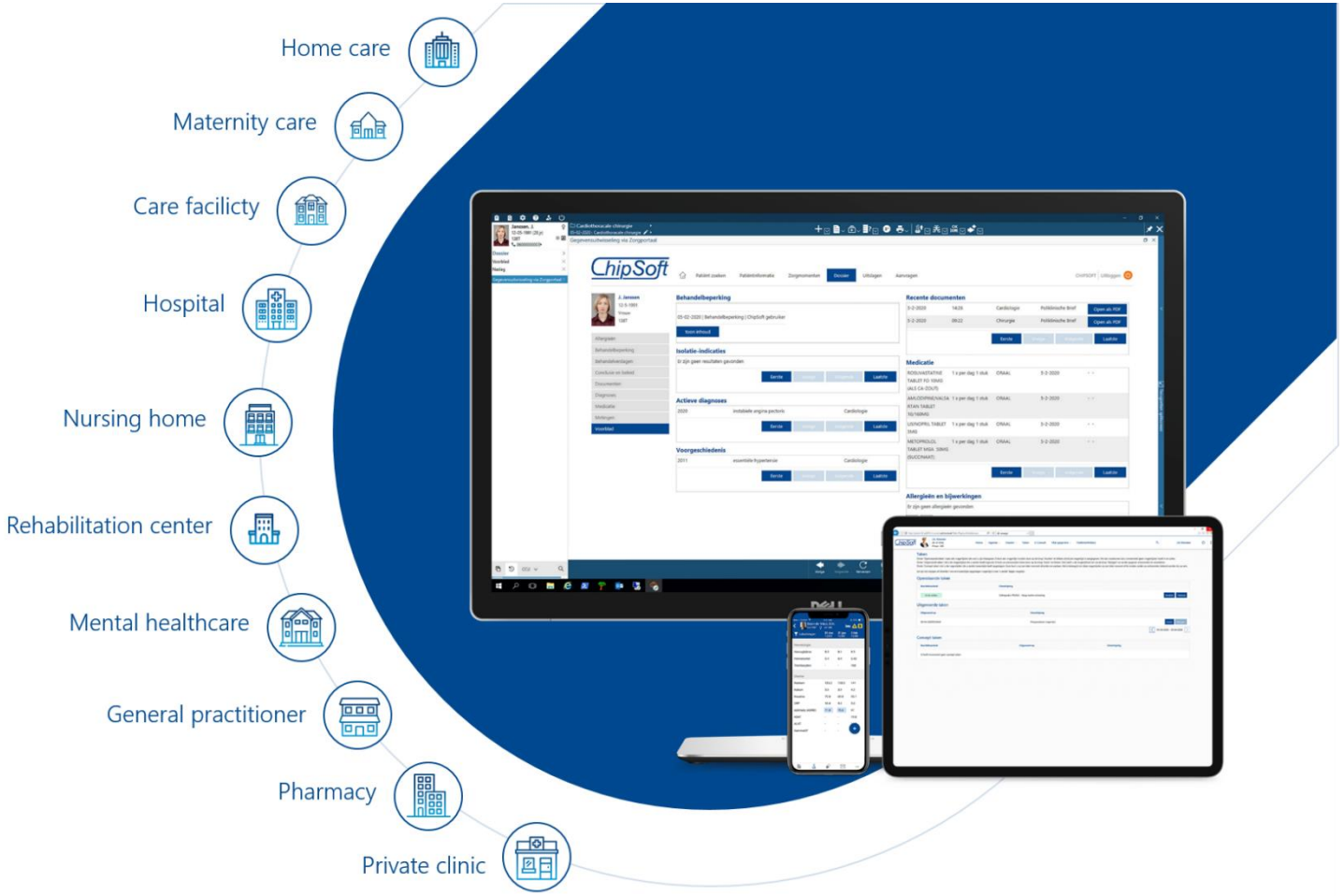
Maurice van Keulen

Johannes H. Hegeman

Jeroen Geerdink

/// Patient data: the EHR

- Electronic Health Record
- Record and access patient data
- 20% of data is structured
  - Lab measurements
  - Medication lists
  - List of diagnoses
- 80% of data is unstructured
  - Images
  - Documents



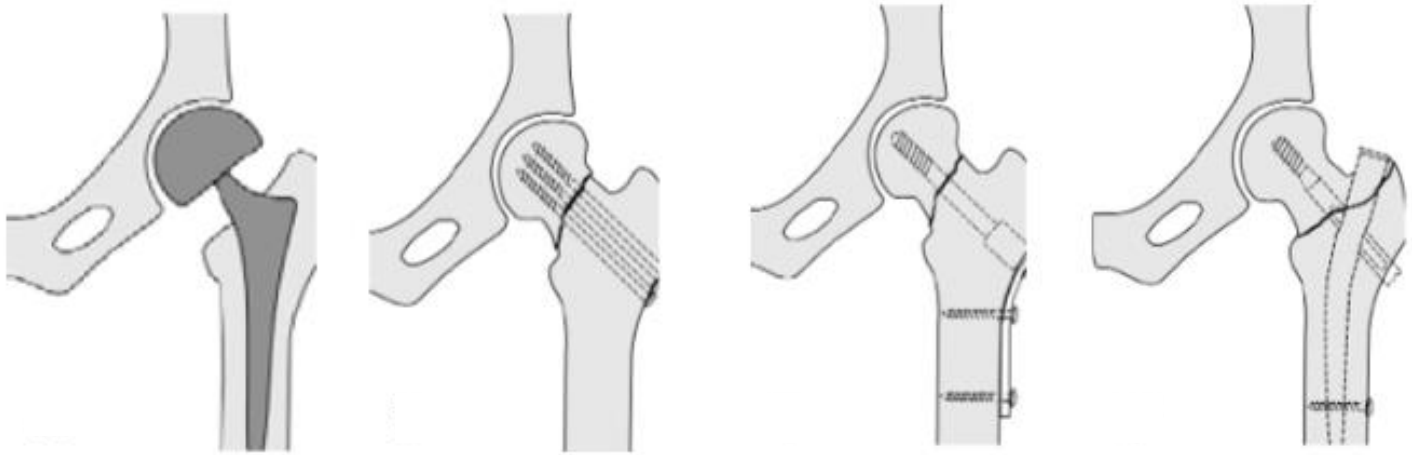
/// Comorbidity: Definition

Presence of additional chronic diseases concurrently with an index condition in one individual.[1]

[1] Valderas et al. (2009) Defining Comorbidity: Implications for Understanding Health and Health Services

/// Comorbidity: Definition

Presence of additional chronic diseases concurrently with an ***index condition*** in one individual.[1]



[1] Valderas et al. (2009) Defining Comorbidity: Implications for Understanding Health and Health Services

/// Relevant Conditions: Charlson Index

Weight	Condition
1	Peripheral vascular disease
	Dementia
	Myocardial infarction
	Chronic pulmonary disease
	Mild liver disease
	Congestive heart failure
	Peptic ulcer disease
	Cerebrovascular disease
	Diabetes, without chronic complications
	Rheumatic disease
2	Hemiplegia
	Renal disease
	Malignancy, except skin neoplasms
	Diabetes, with chronic complications
3	Moderate/severe liver disease
6	Metastatic solid tumor
	AIDS/HIV

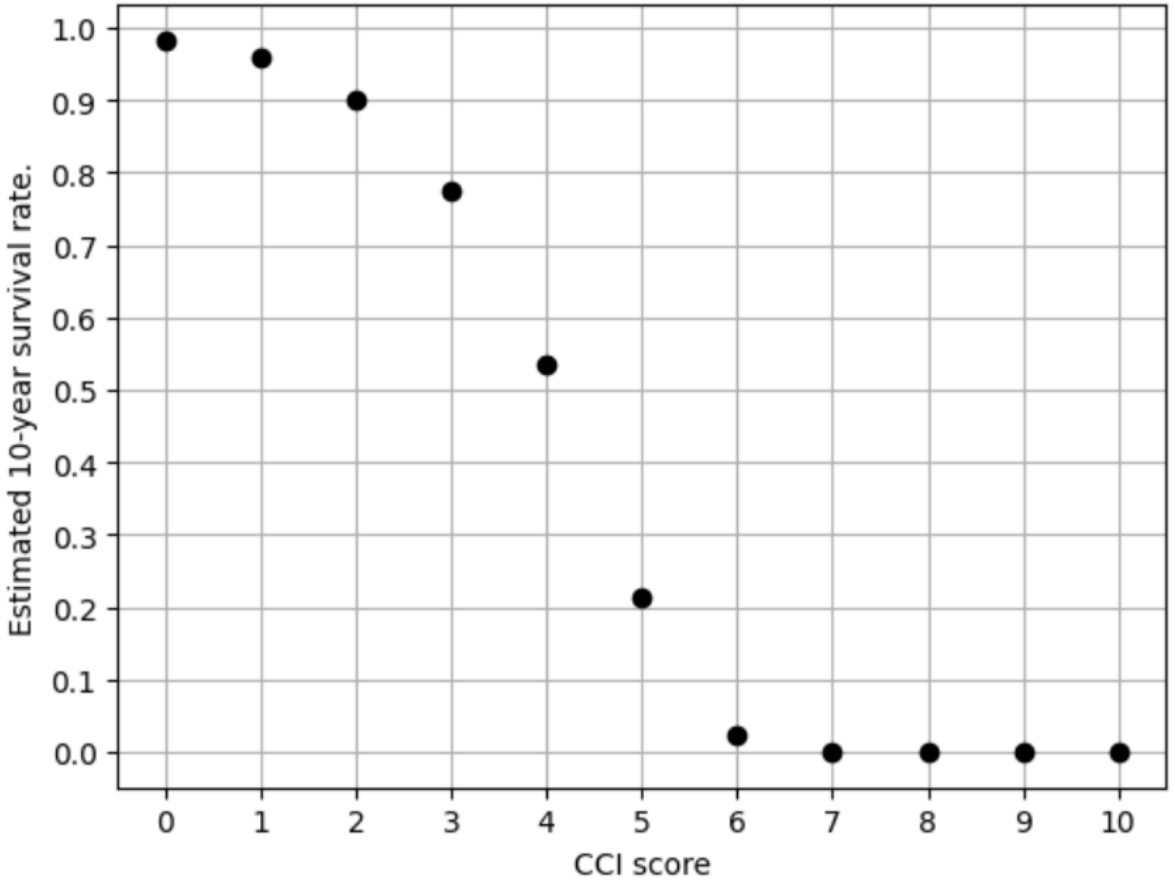


FIGURE 3.1: Estimated 10-year survival rate for CCI scores.

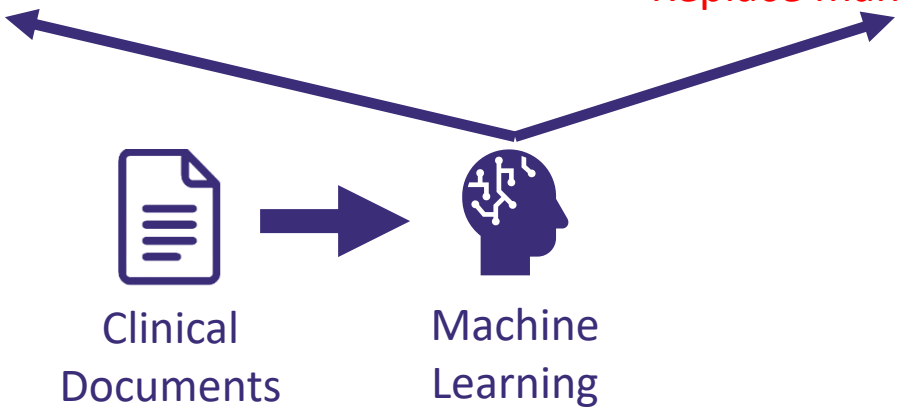
/// Motivation

Clinical Practice

- Clinicians would like a comprehensive overview of patient comorbidity.
- Comorbidities are buried in texts, not available immediately.
- Complete the overview.

Research

- Comorbidities are important inputs for research and predictive models.
- Manual extraction of comorbidities from the EHR is a time-consuming task for large patient cohorts.
- Replace manual annotation.



/// Classify at a document level

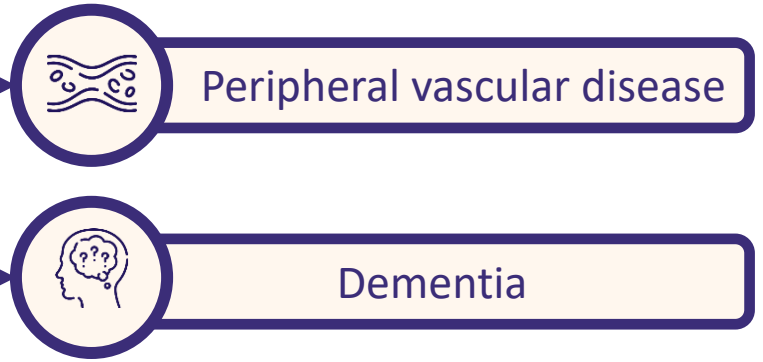
complaint:
potential collum fracture r after fall

anamnesis:
heteroanamnesis due to **dementia**.
patient fell out of bed this morning, was no longer able to mobilize afterwards.

medical history:
hypertension, osteoporosis, **dvt**
2010 – **claudicatio intermittens**
2002 – knee fracture

lab: …

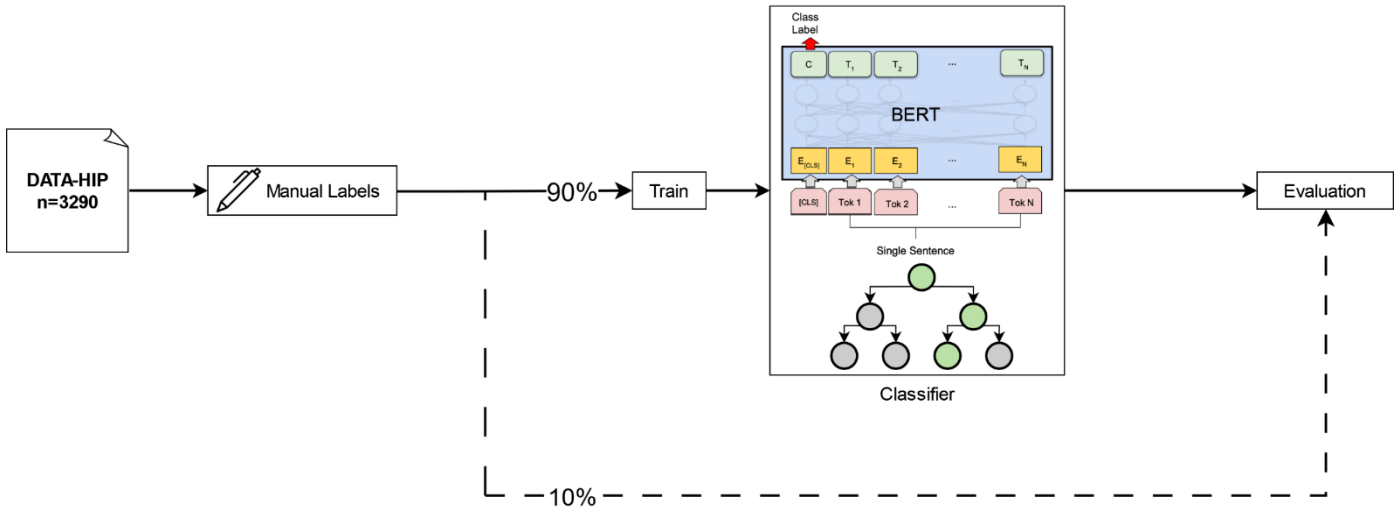
conclusion/therapy: …



/// Baseline: Full supervision

- 3290 documents
  - Hip fracture patients
  - Hand-labeled
  - Age ≥ 70

- 4 Considered models:
  - Naïve Bayes
  - Gradient Boosted Trees
  - Random Forest
  - Transformers (BERT / RoBERTa)



/// Dataset: Class Imbalance

Topzorg voor uw levenskwaliteit

zgt

UNIVERSITY OF TWENTE.

TABLE 5.3: Occurrence rates of CCI categories in **DATA-HIP**

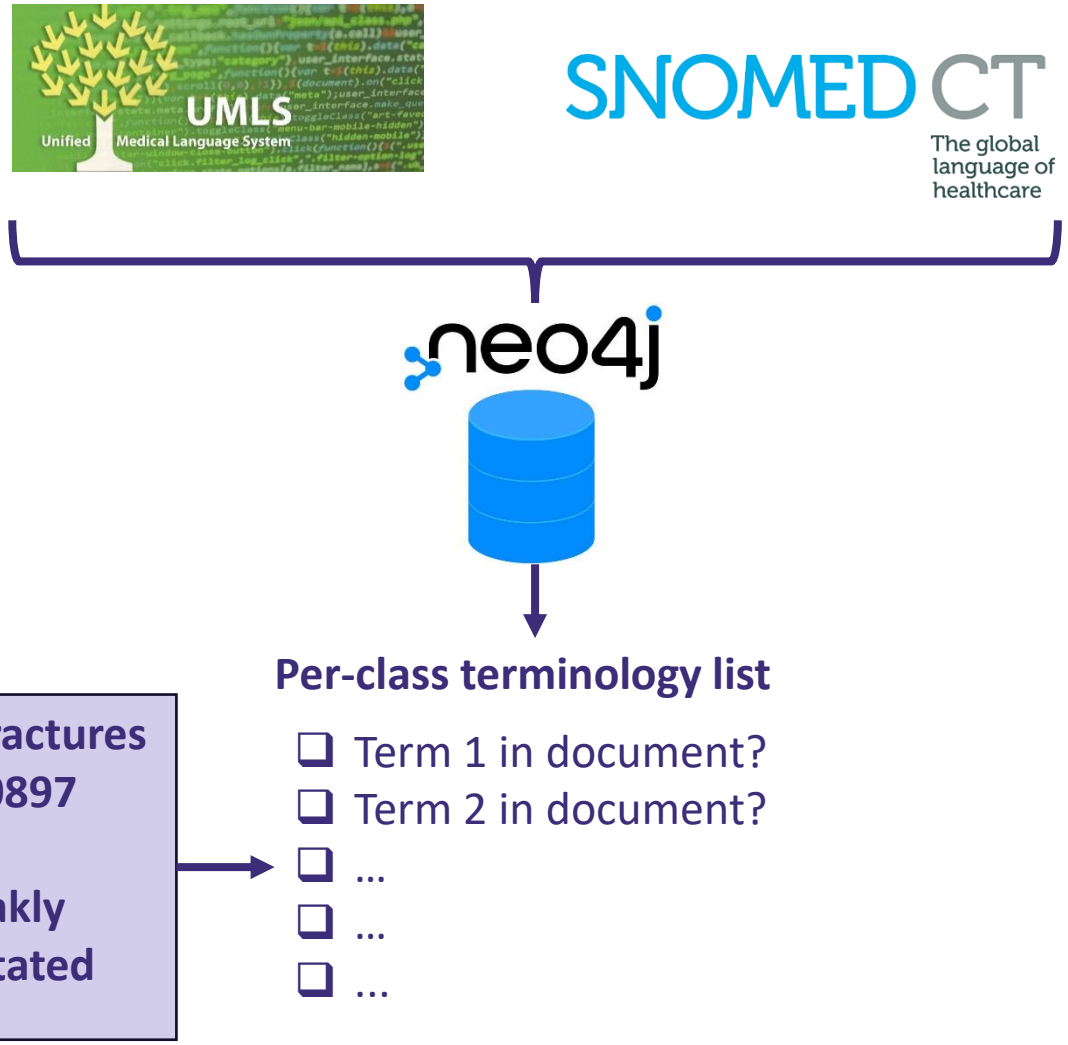
Category	Occurrence rate
Cerebrovascular disease	0.188
Dementia	0.170
Congestive heart failure	0.153
Diabetes, without chronic complications	0.147
Malignancy, except skin neoplasms	0.146
Chronic pulmonary disease	0.136
Peripheral vascular disease	0.121
Renal disease	0.089
Rheumatic disease	0.086
Myocardial infarction	0.078
Diabetes, with chronic complications	0.047
Hemiplegia / paraplegia	0.024
Metastatic solid tumor	0.020
Peptic ulcer disease	0.020
Mild liver disease	0.009
Moderate / severe liver disease	0.003
AIDS / HIV	0.000

/// How can we generate enough examples of rare conditions?

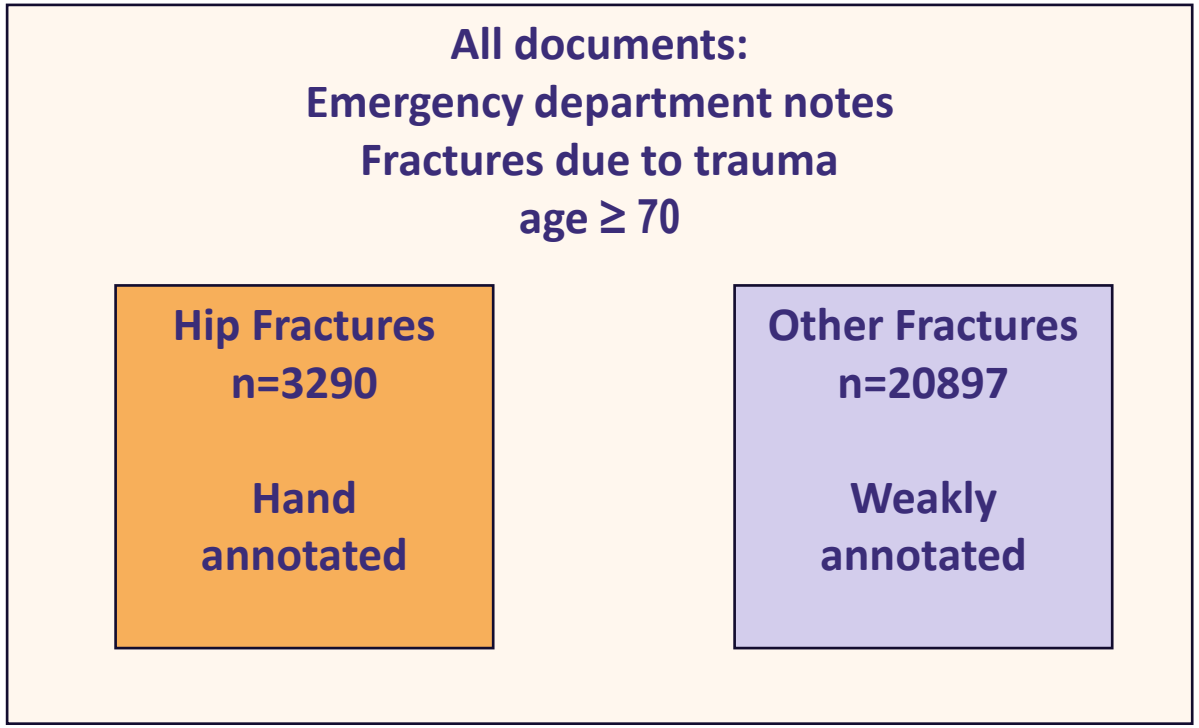
1. Aggregate terminologies onto SNOMED CT
2. Retrieve relevant terms for comorbidities from SNOMED
3. Check for occurrences of terms from retrieved list in unlabeled documents

Difficulties:

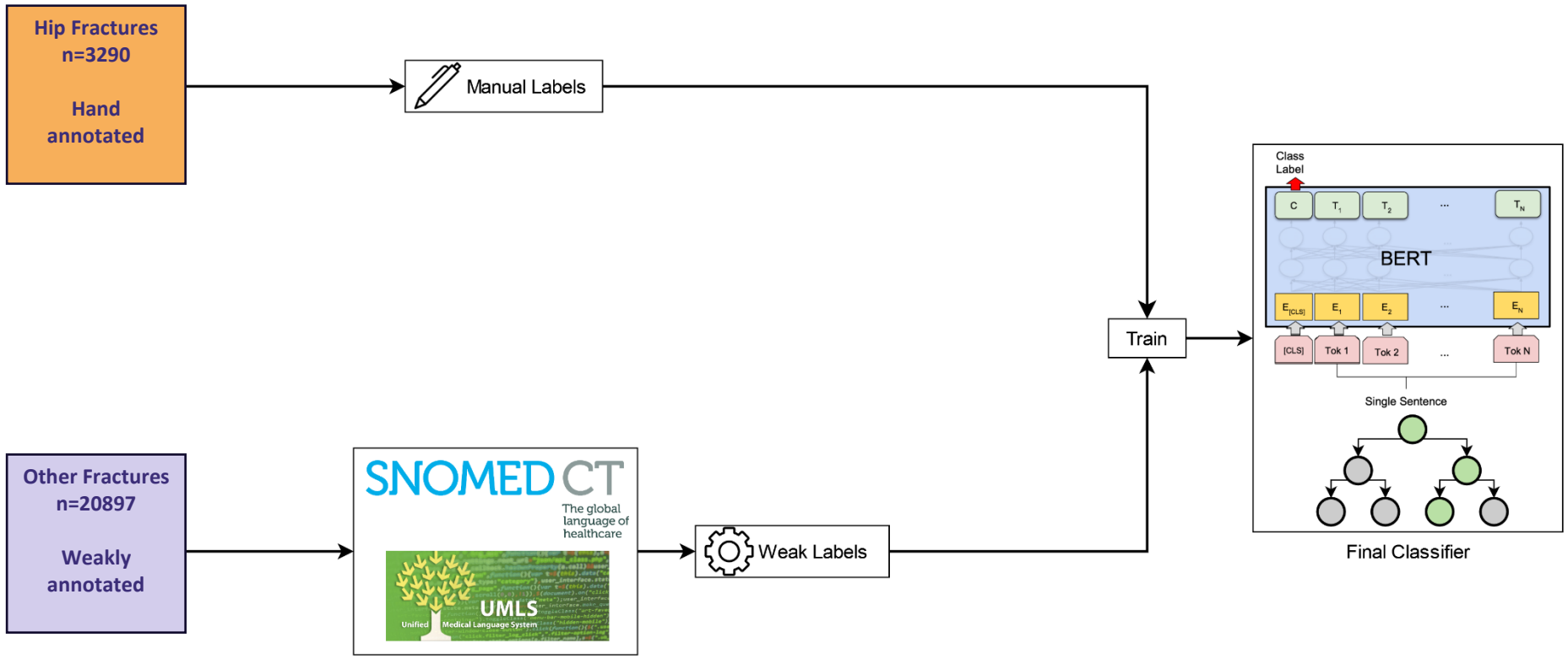
- Negations
- Misspellings
- Ambiguous abbreviations



/// Dataset: Augmented



/// Weak Supervision Pipeline



/// Problem: Mismatch in language

- Clinicians often use terms or phases that can not be found in medical terminologies like SNOMED CT.

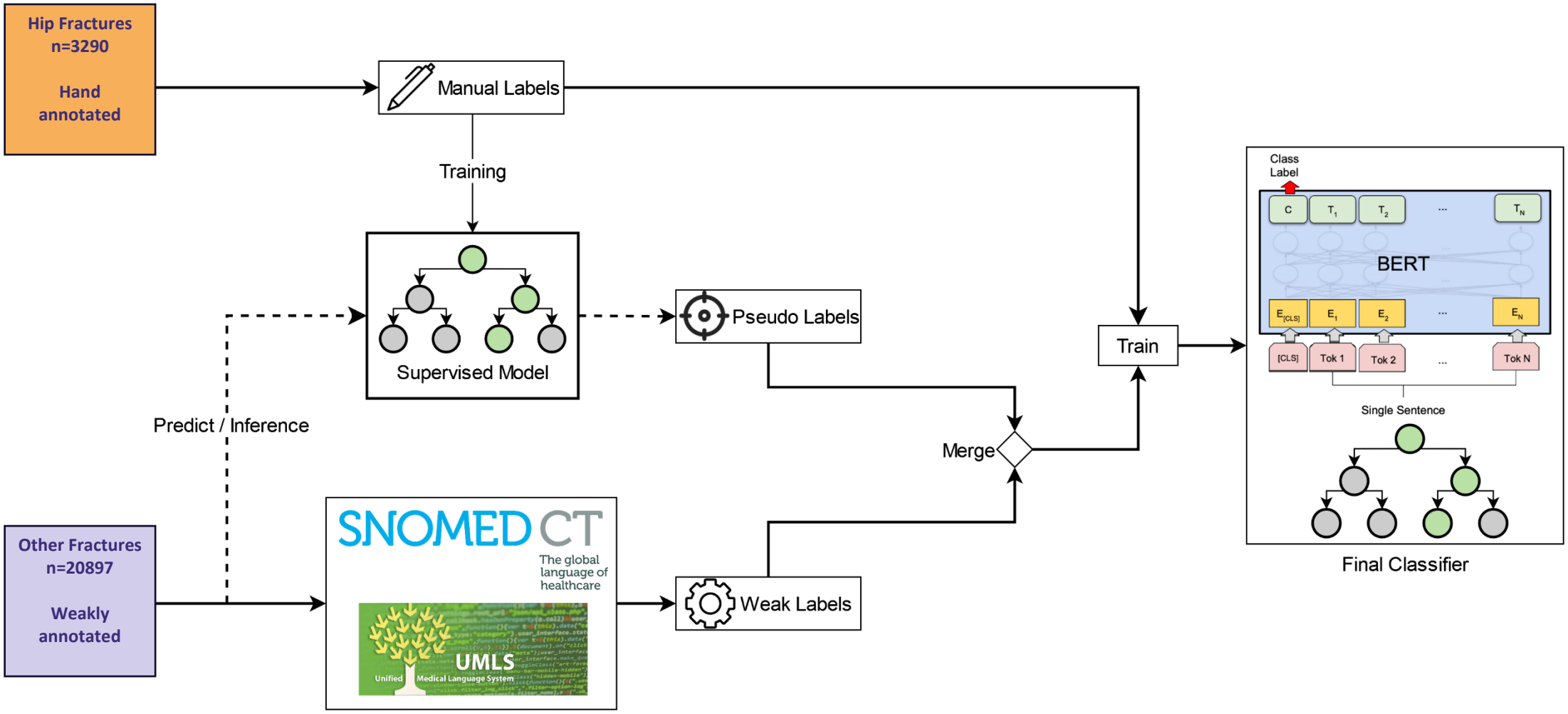
“hemibeeld” instead of “hemiplegie” / “hemiparese”

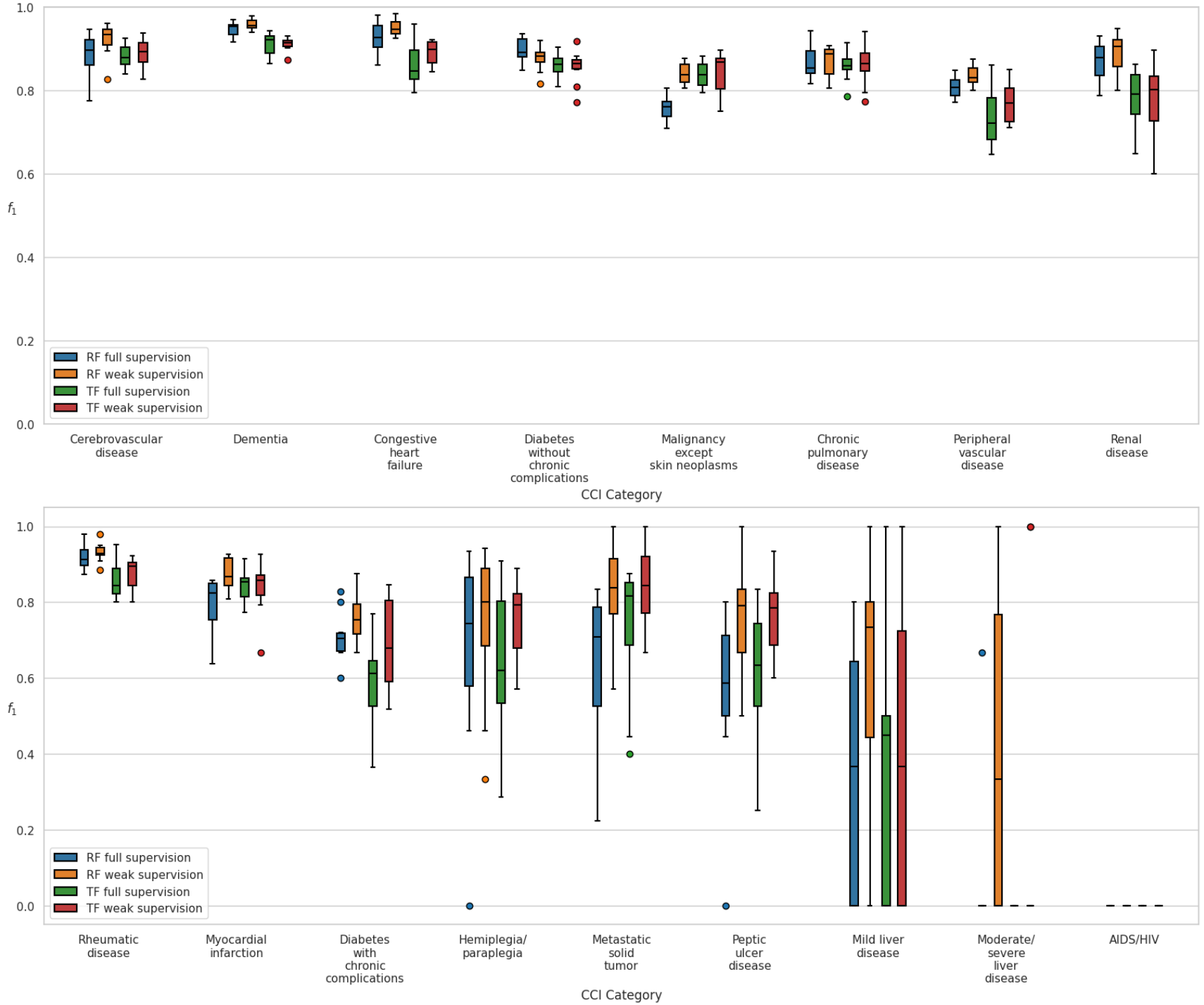
“diabetes met voetafwijking” instead of “diabetische voet”

Our solution:
Pseudo-labeling:

1. *Train a supervised classifier based on hand-annotated data.*
2. *Have supervised classifier predict labels for unannotated data.*
3. *Augment keyword-based weak labels with predicted (pseudo-) labels.*

/// Weak Supervision + Pseudo-labeling





Topzorg voor uw levenskwaliteit

zgt

UNIVERSITY OF TWENTE.

- Improvements in f1 score: 0.05-0.35 for <5% categories.
- Best classification accuracy: Random Forest - **75%**
- **92%** of documents were within 1 CCI point

/// Takeaways

- Random Forests + Weak supervision performed best.
  - Classification accuracy of **75%**. (71% w/o weak supervision)
  - Within 1 point of the correct CCI score in **92%** of test cases. (89% w/o weak supervision)

- Weak supervision with terminologies is effective at generating samples at low cost but care should be taken to bridge the language gap between terminologies and practice.
  - Small amount of hand-labeled data.
  - **Pseudo labeling.**
  - Maintain list of nonstandard vocabulary.
  - **Disambiguation of abbreviations.**

Thank You!

sylvainbrouwer@gmail.com

https://github.com/SylvainBrouwer/

VOORUITSTREVEND

VERBINDEND

MET OPRECHTE AANDACHT

Attributions

Template:

- Hospital Group Twente (ZGT)

Images:

- https://www.istockphoto.com/nl
- https://www.chipsoft.com

Icons:

- Vitaly Gorbachev @ https://www.flaticon.com/authors/vitaly-gorbachev
- https://www.freepik.com/
- https://www.cleanpng.com/