Comorbidity identification in clinical documents with medical terminology-based weak supervision

Sylvain Brouwer[1,2], Maurice van Keulen[1], Jeroen Geerdink[2], Johannes H. Hegeman[1,2]

[1]University of Twente [2]Hospital Group Twente (ZGT)

1: Introduction

Knowledge of patient comorbidities is crucial for effective healthcare decision-making and predictive modeling. Information regarding comorbidities is often buried in unstructured text in EHRs, posing a challenge for data extraction. Natural language processing and machine learning may offer solutions for extracting comorbidity data from unstructured text.

2: Dataset

All used documents are emergency department intake notes of elderly patients (age ≥ 70) with fractures due to trauma. The data is divided into two sets: one containing all hip fractures (n=3290), the other all other fracture types (n=20897). The hip fracture set was hand-annotated for conditions in the Charlson Comorbidity Index (CCI) and is used for training and evaluation. The second set was not annotated, as it is used only in weakly supervised training.

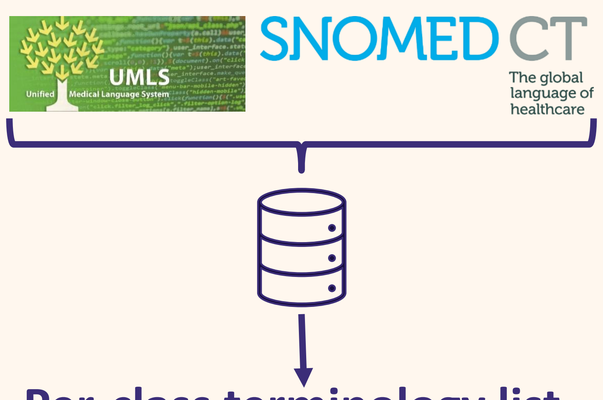
Category	Occurrence rate	CCI points
Cerebrovascular disease	0.188	1
Dementia	0.170	1
Congestive heart failure	0.153	1
Diabetes, without chronic complications	0.147	1
Malignancy, except skin neoplasms	0.146	2
Chronic pulmonary disease	0.136	1
Peripheral vascular disease	0.121	1
Renal disease	0.089	2
Rheumatic disease	0.086	1
Myocardial infarction	0.078	1
Diabetes, with chronic complications	0.047	2
Hemiplegia / paraplegia	0.024	2
Metastatic solid tumor	0.020	6
Peptic ulcer disease	0.020	1
Mild liver disease	0.009	1
Moderate / severe liver disease	0.003	3
AIDS / HIV	0.000	6

Table: Occurrence rates of CCI conditions in the hip fracture dataset.

3: Methods

- Four Machine Learning models were evaluated in a fully supervised setting: Naïve Bayes (NB), Gradient Boosting (GB), Random Forests (RF) and a RoBERTa-based transformer. (TF)
- The two best performing models (RF & TF) were reevaluated after additional data samples were generated with weak supervision

1. Medical terminologies are mapped onto SNOMED CT in a graph database.
2. The DB is queried for relevant terminology for each CCI condition.
3. Resulting terminology lists are used in keyword-based labeling of documents.



- Keyword-based labeling alone proved insufficient for creating quality labeled samples due to the differences between the language used in medical terminologies and clinical practice.
- To bridge the language gap, we augmented the resulting weak labels with pseudo-labels generated by a supervised RF.
- An actively learned classifier was used to disambiguate medical abbreviations during weak label generation.

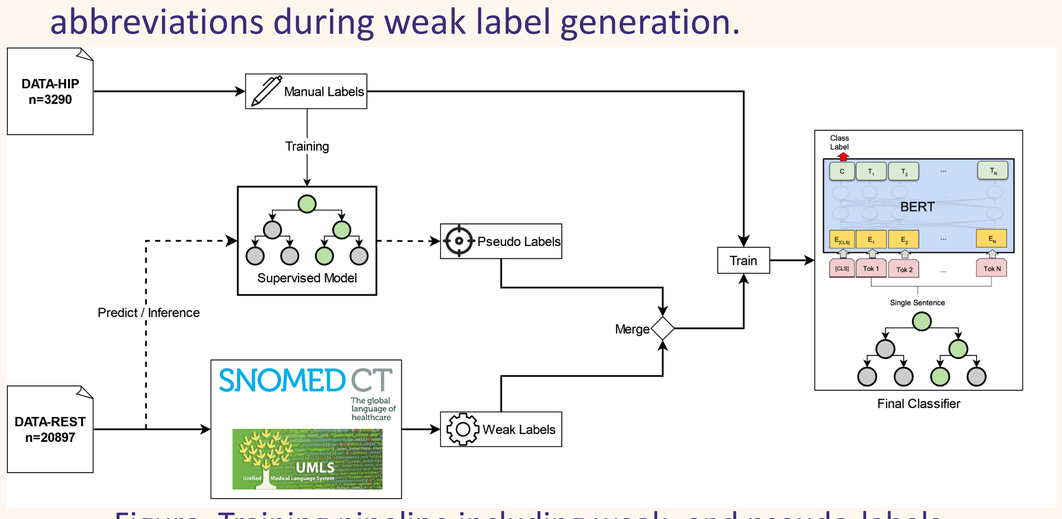


Figure: Training pipeline including weak- and pseudo-labels.

4: Results

Out of the evaluated models, Random Forest performed the best. We found that under the weak supervision scheme, RF was able to predict the correct CCI score in **75%** of test cases and was within 1 point of the correct CCI score in **92%** of test cases. In a per-class evaluation we observe that performance drops significantly with class occurrence rate. This effect was reduced somewhat by the inclusion of additional data through weak supervision.

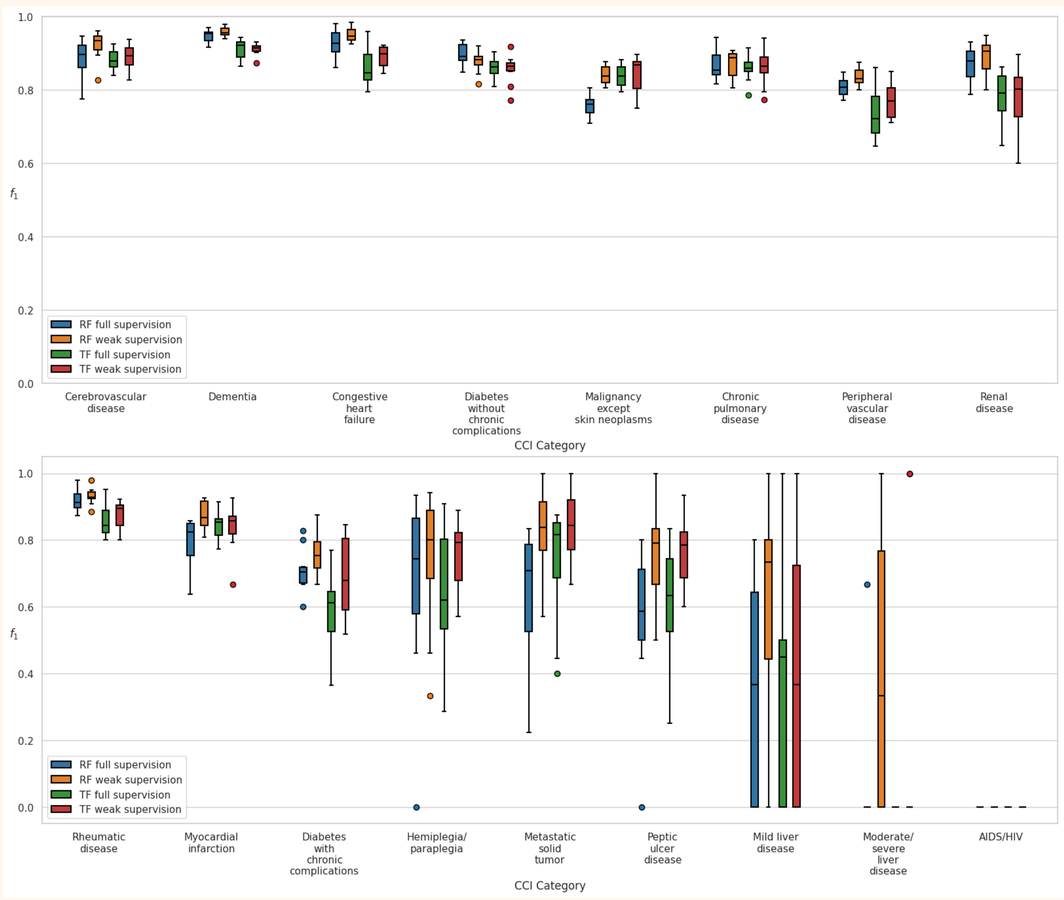


Figure: Per-class F1 scores for RF and TF in a fully and weakly supervised setting.

5: Conclusion

Machine learning can aid in identifying classes of comorbid conditions in clinical notes, for common classes with sufficient positive samples we found good performance (F1 > 0.80). Predicting document labels is a significantly more difficult due to errors cascading over classes and poorer performance for rare CCI categories, resulting in an accuracy of 75%. Weak supervision is effective at generating data samples at a low time cost, resulting in increased performance for rare classes. Medical terminologies were valuable in facilitating weak labeling, but care should be taken to bridge the language gap between these systems and clinical practice.

UNIVERSITY OF TWENTE.

Topzorg voor uw levenskwaliteit zgt