Disambiguation of ICPC codes using free-text and active learning to improve concept mappings

Tom Seinen, Erik van Mulligen, Jan Kors, Katia Verhamme, Peter Rijnbeek
Department of Medical Informatics, Erasmus MC, The Netherlands

Background

The International Classification of Primary Care (ICPC) is a clinical classification system used in general practice (GP) and family care for coding patient data and clinical activity in many countries [1]. The conversion of a GP electronic health record (EHR) database to the OMOP Common Data Model (CDM) requires the mapping of ICPC codes to the OMOP standardized vocabulary and introduces several challenges. One challenge is the inherent ambiguity in certain ICPC codes, e.g. L74 (*Fracture: hand/foot bone*), that results in a one-to-many mapping to the individual sub-classes (*fracture of hand* and *fracture of food*), due to the lack of an equivalent standard code, see Figure 1. Another issue is the high granularity of ICPC concepts while more detailed classifications are desired for observational research, e.g. *endocarditis*, *myocarditis*, and/or *pericarditis* for ICPC coce K70 *Infection of circulatory system*.

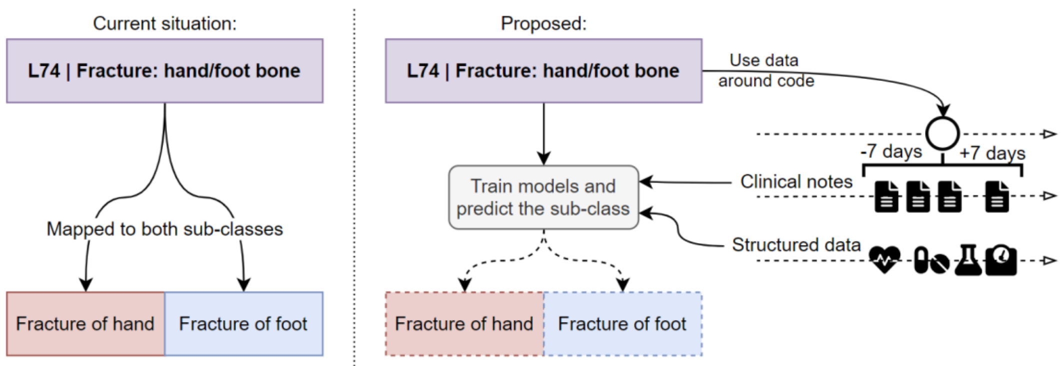

Figure 1. Example ICPC code disambiguation

To address these issues we explored the possibility of the automatic classification of ambiguous ICPC code occurrences in the database into a subset of two or more sub-classes. The goal is to utilize the information available in the structured data and unstructured text around the code occurrence, to train a diagnostic classification model that can distinguish between the different subclasses, see Figure 1. Diagnostic classification models estimate a probability of a patient having a specific disease and have become increasingly popular to aid in clinical decision-making [2]. The use of unstructured data, especially clinical free-text, in prediction modelling research is also increasingly common [3]. Due to the lack of labelled data – the true subclass of each ICPC code occurrence is unknown –, we propose a semi-supervised learning framework that utilizes target rules and active learning to iteratively improve the model using human input. The trained models are evaluated on a manually-annotated validation set.

Methods

Dataset and setting – We used the Integrated Primary Care Information (IPCI) database, which contains observational EHR data from Dutch GPs mapped to the OMOP CDM and includes around 2.8 million patients over a period from 1992 to 2022. For this proof of concept study, we selected one ICPC code that has an ambiguous meaning: L74 (*Fracture: hand/foot bone*).

Feature extraction – Free-text notes and structured data were extracted in a window of 14 days around each code occurrence. The notes were extracted from the OMOP CDM notes table and the text was preprocessed by converting it to lowercase and tokenized creating a bag-of-words feature vector of unigrams and bigrams for each code occurrence. Then the feature matrix was normalized using the Term Frequency-Inverse Document Frequency. Features from the structured data were extracted using the OHDSI Feature Extraction package, including all conditions, medications, measurements, observations, procedures, and demographics within the time window.

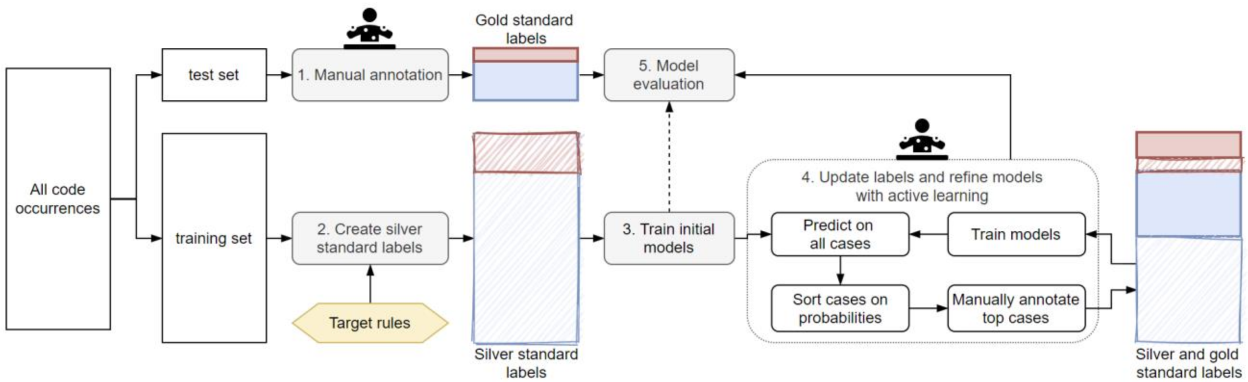

Figure 2. Experimental setup and the active learning framework

Target rules – The code occurrences needed an initial label to allow semi-supervised learning, see Figure 2. Silver standard labels were assigned to the code occurrences using simple target rules that detected the occurrence of sub-class specific words in the text surrounding the code occurrence, i.e. “hand”, “wrist”, and “finger” for the sub-class *Fracture of hand*. The other target rules can be found in Table 1.

Table 1. ICPC codes with their sub-classes, including the target rule words for assigning silver standard labels to code occurrences.

ICPC code	Sub-classes	Target rule words
L74 Fracture: hand/foot bone	Fracture of hand	hand, wrist, finger
	Fracture of foot	foot, ankle, toe

Classification model – A single binary classification model was trained for each code occurrence sub-class, i.e. one binary model for classifying whether the L74 code is a *fracture of hand* or not and another binary model for *fracture of foot*. This allows the possibility of predicting a combination of sub-classes, e.g. indicating the fracture of both a hand and a foot. For all models, we employed regularized logistic regression (LASSO) as the machine learning method.

Active learning – An active learning framework was used to update the silver standard sub-class labels, which the initial classification models used for training, to iteratively improve the models’ performance. The active learning process starts with the initial models, which are used to predict the subclass probabilities for all code occurrences. The code occurrences are then sorted based on their subclass probabilities and the human annotator updates the labels of the code occurrences for which the model is least certain. Using the updated labels the models are retrained and the active learning process is iterated until the evaluation metrics are satisfactory.

Evaluation – We separated a sample of 200 target ICPC code occurrences. These occurrences were manually validated with the sub-class labels, creating a gold standard test set for model evaluation. Another sample of 1000 code occurrences was used for model training. We evaluated the model performance using the areas under the receiver operating curve (AUC) and the precision-recall curve (AUPRC) and the F1 score.

Label combinations – To study the influence of the silver rule-based annotations and the growing number of gold manual annotations on the model performance we tested four different label combinations for training the classification models. The initial model was trained on the silver labels generated by applying the target rules on the training set (1). Refining the model annotates an increasing set of the training set with gold labels. We evaluated these gold labels alone (2), a combination of gold and silver labels (3), and a generalization of the gold labels where a model trained on the gold labels generates predictions for the entire training set on which a new model is trained (4). An increasing number of gold labels, 100, 200, and 300, was evaluated.

Results

The evaluation results of the classification models trained to distinguish fractures of foot and hand for the ICPC code L74 are shown in Table 1. The simple target rules already have a high predictive performance, which is increased if the rules are generalized using the initial classification model. Introducing gold labels during active learning improves the classification performance for both sub-classes even further and increases with the number of annotated observations. Interestingly, the use of only gold labels or the generalization of the gold labels always outperformed the combination of silver and gold labels.

Table 1. Classification results on the test set (200 observations). Background colours indicate the value relative to the other values in the same column: green is high, yellow is low.

	Subclass:	**Fracture of hand**			**Fracture of foot**		
Type:	**Labels used:**	AUC	AUPRC	F1	AUC	AUPRC	F1
Target rules	-	0.87	0.82	0.84	0.91	0.84	0.92
Initial model	1. Silver only	0.91	0.91	0.82	0.95	0.95	0.92
Refined model with	2. Gold only	0.95	0.95	0.88	0.93	0.92	0.9
100 labels	3. Gold and silver	0.92	0.93	0.81	0.96	0.96	0.92
	4. Gold generalized	0.97	0.97	0.89	0.97	0.97	0.92
Refined model with	2. Gold only	0.98	0.98	0.9	0.97	0.96	0.94
200 labels	3. Gold and silver	0.92	0.93	0.81	0.96	0.96	0.92
	4. Gold generalized	0.99	0.99	0.9	0.99	0.98	0.95
Refined model with	2. Gold only	0.98	0.99	0.92	0.99	0.99	0.96
300 labels	3. Gold and silver	0.92	0.93	0.81	0.96	0.96	0.92
	4. Gold generalized	0.99	0.98	0.9	0.99	0.99	0.96

Conclusion

We found that automatic classification of an ambiguous ICPC code occurrence in the database into multiple sub-classes is possible using information from the surrounding structured and unstructured data. This ability to distinguish sub-classes of ambiguous ICPC codes, with only simple target rules and limited labelling, makes it easier to improve concept mappings to the OMOP standardized vocabulary. However, further research is needed into the choice of text representations and machine learning methods and more ICPC codes need to be tested.

References/Citations

1. Soler J-K, Okkes I, Wood M, et al. The coming of age of ICPC: celebrating the 21st birthday of the International Classification of Primary Care. *Family Practice* 2008;25(4):312-17.
2. Hendriksen JM, Geersing G-J, Moons KG, et al. Diagnostic and prognostic prediction models. *Journal of Thrombosis and Haemostasis* 2013;11:129-41.
3. Mujtaba G, Shuib L, Idris N, et al. Clinical text classification research trends: Systematic literature review and open issues. *Expert systems with applications* 2019;116:494-520.