# Mining gold from EMRs: using clinical notes to automate patient screenings

Merijn Beeksma, Madhumita Sushil, Pieter Fivez, Simon Šuster, Florian Kunneman, Ghiath Ghanem, Walter Daelemans, Antal van den Bosch

Difficulties in recruiting patients influences the quality of clinical trials. It time-consuming to screen medical records by hand, and clinical notes, which provide all kinds of nuances and details which cannot be captured in clinical codes, tend to be ignored. Automatic text processing has the potential to drastically decrease the workload, while enabling full exploitation of the available information.

This study focuses on applying natural language processing techniques to clinical notes, to determine which patients fit a set of predefined criteria: 'ALCOHOL-ABUSE', 'DRUG-ABUSE', 'SPEAKS-ENGLISH', 'MAKES-DECISIONS', 'ABDOMINAL' (history of abdominal surgery), 'MAJOR-DIABETES' (at least one major diabetes-related complication), 'ADVANCED-CAD' (two or more signs of cardiovascular disease), 'MI-6MOS' (myocardial infarction in past six months), 'KETO-1YR' (ketoacidosis in past year), 'DIETSUPP-2MOS' (dietary supplements in past 2 months), 'ASP-FOR-MI' (Aspirin to prevent myocardial infarction), 'HBA1C' (HbA1c value between 6.5% and 9.5%), and 'CREATININE' (abnormally elevated level of serum creatinine).

We tokenize the texts, extract dates from the records, and further recognize medical concepts using the CLAMP toolkit (Soysal et al., 2017). We experiment with several machine learning, rule-based, and hybrid approaches for classification. We compare the effectiveness of the proposed techniques against a majority baseline, and with classification results using bag-of-words and bag-of-embedding-clusters.

The hybrid approaches for the modules for SPEAKS-ENGLISH, ALCOHOL-ABUSE, DRUG-ABUSE, and MAKES-DECISIONS all follow a common method. We use regular expressions to first identify features covering both majority and minority classes, and then fit Logistic Regression and Decision Tree classifiers on these features. It was important to downsample the features based on regular expressions to address high class imbalance in the training data.

For MAJOR-DIABETES, we follow an embedding-based approach to account for indirect mentions of complications. We first expand the list of pre-specified major complications with equivalent UMLS (Bodenreider, 2004) terms. Embeddings for these complications are then calculated as their mean word embeddings[1]. If a reference to diabetes is found in text, we check if any non-negated concept of type 'problem' is highly similar to any of the major complications, based on their embedding similarity. If the similarity is higher than a certain threshold, we conclude that the criterion is 'met' for the patient, and 'not met' otherwise.

To detect ABDOMINAL, ADVANCED-CAD and ASP-FOR-MI, we use UMLS and Drugbank (Law et al., 2013) to define and expand the set of relevant terms with their synonyms and frequent misspellings. We then check for the matching terms using regular expressions. In case of CREATININE, HBA1C, MI-6MOS, KETO-1YR, and DIETSUPP-2MOS, we additionally check whether measurement values or dates are mentioned, by using regular expressions to search a small window around the found term for such values. If a measurement value or date is found, we determine whether or not it falls within the required range (e.g. myocardial infarction *in past six months*).

Overall, our hybrid approaches resulted in the highest overall micro-averaged F1 score, 0.85, as compared to 0.83 using the random forest classifier with tf*idf-weighted bag-of-words features. For the classifier, especially the highly unbalanced classed proved to be challenging. In these cases, limiting to the most relevant features is beneficial.
Although the hybrid approach outperformed the random forest classifier for most criteria, the latter reached a higher accuracy on ADVANCED-CAD. The results therefore suggests that a combined approach, relying both on classification results and regular expressions, has the potential to improve the overall results.

[1] trained on PubMed.

**References:**

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl_1), D267-D270.

Law, Vivian, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski et al. "DrugBank 4.0: shedding new light on drug metabolism." Nucleic acids research 42, no. D1 (2013): D1091-D1097.

Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., & Xu, H. (2017). CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association.