

Application of natural language processing methods to extract coded data from administrative data held in the Scottish Prescribing Information System

Nangle, Clifford^{1*}, McTaggart, Stuart¹, MacLeod, Margaret¹, Caldwell, Jackie¹, and Bennie, Marion¹¹Public Health Intelligence, NHS National Services Scotland

Objectives

The Prescribing Information System (PIS) datamart, hosted by NHS National Services Scotland receives around 90 million electronic prescription messages per year from GP practices across Scotland. Prescription messages contain information including drug name, quantity and strength stored as coded, machine readable, data while prescription dose instructions are unstructured free text and difficult to interpret and analyse in volume.

The aim, using Natural Language Processing (NLP), was to extract drug dose amount, unit and frequency metadata from freely typed text in dose instructions to support calculating the intended number of days' treatment. This then allows comparison with actual prescription frequency, treatment adherence and the impact upon prescribing safety and effectiveness.

Approach

An NLP algorithm was developed using the Ciao implementation of Prolog to extract dose amount, unit and frequency metadata from dose instructions held in the PIS datamart for drugs used in the treatment of gastrointestinal, cardiovascular and respiratory disease. Accuracy estimates were obtained by randomly sampling 0.1% of the distinct dose instructions from source records, comparing these with metadata extracted by the algorithm and an iterative approach was used to modify the algorithm to increase accuracy and coverage.

Results

The NLP algorithm was applied to 39,943,465 prescription instructions issued in 2014, consisting of 575,340 distinct dose instructions. For drugs used in the gastrointestinal, cardiovascular and respiratory systems (i.e. chapters 1, 2 and 3 of the British National Formulary (BNF)) the NLP algorithm successfully extracted drug dose amount, unit and frequency metadata from

95.1%, 98.5% and 97.4% of prescriptions respectively. However, instructions containing terms such as 'as directed' or 'as required' reduce the usability of the metadata by making it difficult to calculate the total dose intended for a specific time period as 7.9%, 0.9% and 27.9% of dose instructions contained terms meaning 'as required' while 3.2%, 3.7% and 4.0% contained terms meaning 'as directed', for drugs used in BNF chapters 1, 2 and 3 respectively.

Conclusion

The NLP algorithm developed can extract dose, unit and frequency metadata from text found in prescriptions issued to treat a wide range of conditions and this information may be used to support calculating treatment durations, medicines adherence and cumulative drug exposure. The presence of terms such as 'as required' and 'as directed' has a negative impact on the usability of the metadata and further work is required to determine the level of impact this has on calculating treatment durations and cumulative drug exposure.

*Corresponding Author:

Email Address: c.nangle@nhs.net (C. Nangle)