PRM58

# Accuracy of text processing tokenisers for automated identification of diseases and interventions in abstracts of studies on humanistic and economic burden of disease.

Rob Challen;[1] Alison Martin;[2] Chris Martin[2]

[1]Terminological Ltd, Exeter, UK; [2]Crystallise Ltd, Essex, UK email: rob@terminological.co.uk

## Objectives

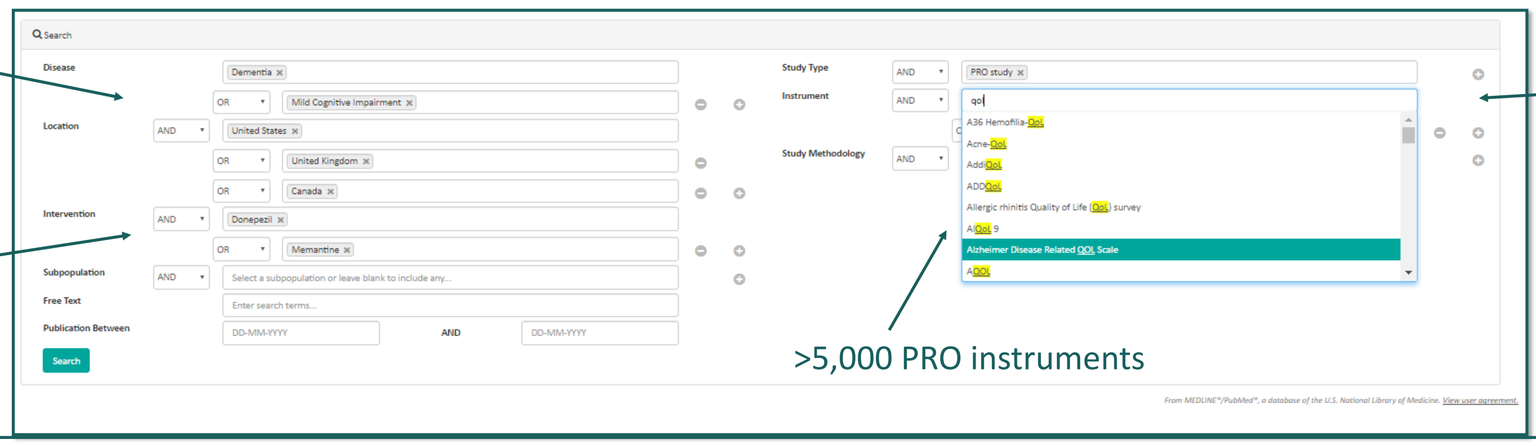
To determine the sensitivity and specificity of software based on text processing analysis to classify diseases and interventions in PubMed abstracts relevant to the humanistic or economic burden of disease.

## The problem

The heoro.com database contains >100,000 abstracts identified by a systematic search of PubMed on the humanistic and economic burden of disease from 2005, and 70,000 abstracts from 1960-2004. We needed to find a quick, affordable and accurate way of indexing these to our ontologies.

>4,700 disease entries

>8,500 intervention entries




>5,000 PRO instruments

4 study types

11 study subtypes

2 study methodologies

## Methods

We manually indexed 10,000 abstracts to detailed ontologies of diseases and interventions, as well as to study types, PRO instruments and geographical setting. The disease and intervention ontologies were developed from MeSH terms and lists of licensed drugs from the US and UK, with new items added when identified from the abstracts. We used this training set to develop tokenisers to facilitate matching the text, MeSH headings and metadata in the abstracts to relevant ontology items.

We then assessed the initial accuracy of the tokeniser matching on a sample of 150 abstracts from the unmoderated set published from 2005, using expert evaluation, prior to further software refinements.

## Results

The tokeniser matching had a sensitivity of 95% for disease ontology items and 85% for intervention ontology items compared with expert assessment. The specificity, defined as matching to any ontology items that appeared in the text, MeSH headings or metadata of each abstract, was 89% for diseases and 91% for interventions. The accuracy of matching was higher for drug terms than for non-pharmaceutical interventions, which tend to be described less consistently.

| | Diseases | | | | Interventions | | | |
|---|---|---|---|---|---|---|---|---|
| | Correctly tagged to disease | Correctly tagged but not focus of study | Tagged incorrectly | Missing relevant disease tag | Correctly tagged to intervention | Correctly tagged but not focus of study | Tagged incorrectly | Missing relevant intervention tag |
| | **269** | **244** | **65** | **27** | **266** | **170** | **42** | **76** |
| Sensitivity | **95%** | | | | **85%** | | | |
| Specificity overall | **89%** | | | | **91%** | | | |
| Specificity for focus of paper | **47%** | | | | **56%** | | | |

## Conclusions

With overall accuracy of around 90%, the initial tokeniser matching compared reasonably to indexing of abstracts by less experienced scientists. Ongoing final expert checking and further software refinement will improve the specificity of the indexing to topics that were the focus of the research. As 90,000 abstracts could be indexed within hours, this method facilitates a streamlined approach to identifying relevant data for health economics and outcomes research.

**Crystallise Ltd.** Unit 21 Thames Enterprise Centre, Thames Industrial Park, East Tilbury, Essex RM18 8RH, UK . Tel +44 (0)1375 488020.

**For a copy of this poster, email**: alison.martin@crystallise.com

heoro.com

www.heoro.com

**Presented at the ISPOR 19th Annual European Congress**

**October 29-November 2 2016 ; Vienna, Austria**