Incorporating natural language processing to improve classification of axial spondyloarthritis using electronic health records

Zhao S[1,2], Hong C[3], Cai T[2,3], Chang X[2], Huang J[2], Ermann J[2,3], Goodson N.J[1], Solomon D.H[2,3], Cai T[3,4], Liao K.P[2,3]

1 Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool, UK,

2 Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, USA

3 Harvard Medical School, Boston, Massachusetts, USA

4Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

Background. Electronic health records (EHR) are increasingly adopted world-wide and offer exciting opportunities for clinical research, particularly for rare diseases and outcomes. However, its use is limited by inaccuracies in the way data, such as diagnosis, is coded.

Aims. To develop algorithms that accurately identify axial spondyloarthritis (axSpA) patients in EHR, and compare the performance of algorithms incorporating free-text data against approaches using only International Classification of Diseases(ICD) codes.

Methods. An enriched cohort of 7,853 eligible patients was created from EHR of two large hospitals in Boston, USA, using automated searches. Key disease concepts from free-text data were extracted using natural language processing (NLP) and combined with ICD codes to develop algorithms. We created both supervised regression-based algorithms - on a training set of 127 axSpA cases and 423 non-cases - and unsupervisedalgorithms (that is, without the need for manually derived labels) to identify patients with high probability of having axSpA. Their performance was compared against classifications using ICD codes only.

Results. NLP extracted four disease concepts of high predictive value: "ankylosing spondylitis", "sacroiliitis", "HLA-B27" and "spondylitis". The unsupervised algorithm, incorporating both the NLP concept and ICD code for AS, identified the greatest number of patients. By setting the probability threshold to attain 80% positive predictive value, it identified 1,509 axSpA patients (mean age 53 years, 71% male). Sensitivity was 0.78, specificity 0.94 and area under the curve (AUC) 0.93. The two supervised algorithms performed similarly but identified fewer patients. All three outperformed traditional approaches using ICD codes alone (AUC 0.80 to 0.87).

Conclusion. Algorithms incorporating free-text data can accurately identify axSpA patients in EHR. Robust algorithms could be developed for diseases with evolving definitions such as axSpA, without needing manual chart-review to create training datasets. Large cohorts identified using these novel methods offer exciting opportunities for future clinical research.