International Journal of Population Data Science

Journal Website: www.ijpds.org

Development of data-driven framework for automatically identifying patient cohorts from linked electronic health records

Fernández-Gutiérrez, Fabiola[1], Kennedy, Jonathan[2], Cooksey, Roxanne[2], Atkinson, Mark[2], Choy, Ernest[3], Brophy, Sinead[2], and Zhou, Shang-Ming[2*]

[1]The University of Manchester

[2]University of Glasgow

[3]Cardiff University

Objectives

1. To develop a fully data-driven framework for automatically identifying patients with a condition from routine electronic primary care records.

2. to identify informative codes (risk factors) of arthropathy conditions in primary care records that can accurately predict a diagnosis of the conditions in secondary care records.

Approach

This study linked routine primary and secondary care records in Wales, UK held in the SAIL (Secured Anonymised Information Linkage) databank, in which the secondary care records were used as golden standard. As such, we proposed to use machine learning techniques to extract patient information and identify cohorts with a condition from the large and high-dimensional linked dataset using the following phases: data preparation, performed in the machine learning context fashion; pre-selection of initial features, ranking and selecting features into a meaningful subset by using feature selection methods; and identification algorithm development which incorporates mechanisms of tackling the imbalanced nature of the data. This data-driven framework was then validated on an independent dataset, and compared with existing algorithm which had been developed using expert clinician knowledge for arthropathy conditions.

Results

Rheumatoid arthritis (RA) and ankylosing spondylitis (AS) were used to demonstrate the feasibility of this framework. Linking primary care records with the secondary care rheumatology clinical system, we collected 9,657 patients with 1,484 RA patients and 204 AS patients. The proposed framework identified various compact subsets of informative features (risk factors) from 43,100 potential Read codes. Applying to an independent test data, this framework achieved the classification accuracy and positive predictive values (PPVs) of 86.19% and 88.46% respectively for RA and 99.23% and 97.75% respectively for AS, which are comparable with the performance of clinical knowledge-based method - the accuracy of 85.85%, the PPV of 85.28% for RA and the accuracy of 97.86% , the PPV of 95.65% for AS.

Conclusion

The proposed data-driven framework provides a rapid and cost-effective way of reliably identifying patients with a medical condition from primary care data. It performed as well as the clinically derived algorithm. This framework does not intend to substitute clinical expertise, instead it provides an decision support tool for clinicians during their decision process, in particular selection of patients for clinical trials.

*Corresponding Author:
Email Address: s.zhou@swansea.ac.uk (S. Zhou)

http://dx.doi.org/10.23889/ijpds.v1i1.86
August 2016 © The Authors. Open Access under CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en)