International Journal of Population Data Science

Journal Website: www.ijpds.org

Automatic coding of nearly 2 million hospitalisation events to ICD-10 in the China Kadoorie Biobank

Sansome, S[1], Turnbull, I[1], and McDonnell, J[1]

[1]University of Oxford

Introduction

Using linkage to the Chinese National Health Insurance (HI) system, we identified disease outcomes from a prospective cohort study of 512,000 middle-aged Chinese adults. Mandarin free-text diagnosis data were supplied by over 30 different agencies across 10 areas, often without an accompanying International Classification of Diseases 10th revision (ICD-10) code.

Objectives and Approach

To facilitate a genome-wide association study (GWAS) of all our genotyped participants, we needed to code as many of our 2.02 million hospitalisation events as possible. We developed software to assign ICD-10 codes to unique disease descriptions and stored the coded diagnoses in an internal corpus. The software used an interface which allowed clinicians to select and code disease descriptions individually, or collectively using Chinese keywords. All coded disease descriptions were subsequently validated by an independent Mandarin-speaking clinician. All new events with descriptions which matched exactly those already in the corpus were automatically coded to ICD-10.

Results

By the end of 2016, there were 2,021,352 hospitalisation events coded to ICD-10. 436,702 (21.6%) were automatically assigned codes where disease descriptions corresponded to those in the Chinese version of the ICD-10 codebook. A further 1,084,197 (53.6%) were coded by a clinician using our standardisation software; all disease descriptions linked to 200 or more events were included. Finally, a remaining 454,237 (22.5%) events were given the ICD-10 codes supplied by the health insurance agency (after cleaning). In total, 97.7% of all health insurance events were coded to ICD-10. Overall, over 17,000 unique disease descriptions have been clinically classified.

Conclusion/Implications

Automatic coding of hospitalisation events to ICD-10 has enabled our study to investigate a greater range of diseases and use GWAS to detect novel genetic variants. We are now well positioned to test semantic matching and machine learning strategies for coding of the remaining 46,216 (2.3%) uncoded events.

http://dx.doi.org/10.23889/ijpds.v3i4.646