International Journal of Population Data Science (2018) 3:3:351

International Journal of Population Data Science

Journal Website: www.ijpds.org

Improving the Coding Completeness of Hypertension in Inpatient Administrative Health Data Using Machine Learning Methods

D'Souza, A[1], Liang , Z[1], Williamson, T[1], Smith, T[2], Quan, H[1], and Peng, M[1]

[1]University of Calgary

[2]Faculty of Computing & Mathematical Sciences, The University of Waikato

Introduction

The Discharge Abstract Database (DAD) associates ICD-10-CA diagnosis codes with inpatient care episodes at acute-care facilities. Codes are assigned by human coders, based on chart review. Coding guidelines stipulate mandatory coding of major and fatal conditions but only optional coding of secondary conditions, which results in undercoding for many conditions.

Objectives and Approach

This research evaluates machine learning approaches for identifying and completing records with missing codes, to improve data quality. The Alberta Hospital DAD for 2013-14 was used in this study. We assumed that the existing ICD-10-CA codes in the DAD are correct, and used them as training examples. Several ML classifiers, including logistic regression and random forest, were used to develop models to assess the coding probability, using existing codes and demographic information. 3300 chart-review records were used as the reference standard. We focused on hypertension-related codes. Validity of raw diagnosis codes in the DAD was used as the baseline.

Results

A record is deemed to have a missing hypertension diagnosis code if the predicted probability is high, but without the diagnosis codes having been assigned by the coders. In the baseline, the original hypertension codes have high PPV (ranging from 0.902 for the age group 35-54 to 1.000 for the age group 18-34) but low sensitivity (ranging from 0.200 for the age group 18-34 to 0.565 for the age group 75+). The most successful models that we have tested so far have provided improvements of 2-6% in the sensitivity, while maintaining the PPV. More improvement is generally seen for the younger age groups. Initial experiments indicate greater improvements in sensitivity may be possible for other conditions, such as peptic ulcer disease and cerebrovascular disease.

Conclusion/Implications

Machine learning approaches can be useful and cost-effective for improving data quality in DAD. While the improvements in sensitivity relative to the baseline are modest at present, further experiments with different models and feature sets are warranted. Experiments with other conditions may also be fruitful.

http://dx.doi.org/10.23889/ijpds.v3i4.946