International Journal of Population Data Science

Journal Website: www.ijpds.org

Development and validation of data quality rules in administrative health data using association rule mining

Peng, M[1], Lee, S[1], Kelly, M[2], Linton, C[2], Penchoff, M[2], Sabljakovic, S[2], and Quan, H[1]

[1]University of Calgary

[2]Canadian Institute for Health Information (CIHI)

Introduction

Data quality assessment is a challenging facet for researches using coded administrative health data. Our previous study had demonstrated the potentials of association rule mining to assess data quality. The objective of this study is to develop and validate a set of coding association rules for data quality assessment.

Objectives and Approach

We used the Canadian reabstracted hospital discharge abstract data (DAD) with clinical diagnosis coded in International Classification of Disease – 10th revision, Canada (ICD-10-CA) codes for rule development. The DAD data were divided into 5 age groups. Association rule mining were conducted on reabstracted DAD in each age group to extract ICD-10 coding association rules at the three and four digits levels. The rule strength was assessed using support and confidence. The rules will be reviewed by a panel of 5 physicians and 2 coding specialists to assess their appropriateness from clinical and coding perspectives using a modified Delphi rating

Results

In total, 975 rules at the three digits level and 822 rules at the four digits level were learned from the data. Half of the rules were in the age group of $\geq$65 and no rules were found in the age group of 5 to 19. The interquartile range of rule confidences were 0.112 to 0.425 in the three digits level and 0.073 to 0.222 in the four digits level. Two-thirds of rules had the diagnosis codes related to endocrine and metabolic disorders and diseases of circulatory, respiratory and genitourinary systems. The panel review will be conducted in early April and will have the final set of rules available before the conference.

Conclusion/Implications

This study developed a set of validated ICD-10 coding association rules and creates a useful tool to cost-effectively assess data quality in routinely collected administrative health data.