Title: A machine learning approach for outlier detection in cost data used for RIW regression models in Canada.

Authors: Koffi Kpelitse, Rachel Zhang, Victoria Zhu and Sheril Perry

Introduction:

The Canadian Institute for Health Information (CIHI) uses the acute care Canadian Patient Cost Data (CPCD) to derive resource intensity weights (RIWs) for its acute inpatient and ambulatory case-mix grouping methodologies (Case Mix Groups+ (CMG+) and Comprehensive Ambulatory Classification System (CACS)). RIWs are relative cost weights and are critical for acute care management and planning. CIHI is reviewing the current non-statistical and statistical criteria used to identify outliers in RIW calculation data, exploring alternative approaches that are less likely to disproportionately classify high-cost records as outliers, which impacts the overall RIW values.

Methods:

Three popular unsupervised machine learning (ML) techniques were explored as alternatives to detecting outliers in CPCD data: Isolation Forest (IF), Local Outlier Factor (LOF) and One-Class Support Vector Machine (SVM). These three methods were tested using inpatient care data from fiscal years 2016-17 to 2018-19. For each ML technique, various specifications were tested with different outlier thresholds using various combinations of total cost per patient stay, length of stay and per-diem cost as input features.

For each iteration, the performance of the three techniques is measured by comparing the distributions of inliers and outliers and the exclusion rates for low-cost and high-cost records.

Results:

Out of the three ML approaches explored, SVM proved to best fit the needs of identifying extreme cost data with a reasonable balance between case variation and costs. Both the IF and LOC approaches appeared to under-edit the low-cost records and over-edit the high-cost records.

Models that incorporated patient cost per stay, length of stay and per-diem costs as input features performed the best. Although these features are highly correlated, the SVM classification performed with high tolerance and minimal impacts were identified.

With the SMV method and selected modeling variables, an ideal percentage of cases is determined to be excluded, without compromising the quality of the resulting resource estimates. The details of the extent of cost data conservation for RIW calculation will be presented.

Additionally, the decision boundary created by the SVM method resulted in similar proportion of exclusion rates for the vast majority of the CMGs. The CMGs with high outlier rates are generally associated with low volumes or extreme high-cost cases.

Conclusions:

The SMV approach to outlier detection appears to be an effective way to identify acute patient cost data outliers for cost weight production. This approach not only simplifies the data preparation process, but also provides data-driven evidence to control the volume of data points being excluded without compromising the predictive ability of the regression output. The next step will be to test this approach to outlier detection with the ambulatory acute CPCD data.