Predictive Risks of Colorectal Cancer by Machine Learning

Mr. John Mok[1], Mr. PC Ho[1], Mr. Vincent Ng[1], Ms. Vicky Fung[1]

[1]*Information Technology and Health Informatics Division, Hong Kong Hospital Authority, , Hong Kong SAR*

Introduction:

With the advance of data science, clinical applications using machine learning to predict diseases have become possible. This is a proof of concept (POC) study on Hong Kong adult population that we were running machine learning algorithms to predict the risks of colorectal cancer by using age, sex and Complete Blood Count (CBC) laboratory data.

Methodology:

A cohort of de-identified patient selection was conducted for the data preparation and model construction. We extracted and aggregated one year CBC and ten years histopathology data from the Laboratory Information System of a general acute hospital. For preparing the training and testing dataset, the data did not have any histopathology investigation requested before or after individuals CBC results (i.e. cancer-free individuals), the dataset were labelled as negative. Whereas for data had any colorectal cancer results identified one year after the CBC results, the dataset were labelled as positive.

Result:

After the cohort selection, training and testing data were curated and saved as a comma separated variable file for the supervised machine learning. A machine learning software – Weka (Waikato Environment for Knowledge Analysis) was used and several machine learning algorithms were tested. In this study, we found that running 13 features (CBC including WBC, RBC, HGB, HCT, MCV, MCHC, RDW, PLT and MPV, patient's age and sex) and using the Random Forest with Cost Sensitive Classifier could create the best accuracy for data modelling on predictive colorectal cancer – Area Under the Curve (AUC) was 0.814, and its Negative Predictive Value (NPV) was 0.986 which giving us a high confidence that its negative result was true.

Discussion:

In this study, we were grateful of having standardized laboratory data available in the Hong Kong Hospital Authority. However the data manipulation process for the machine learning was laborious that we spent quite some time on the data extraction, cleansing, curation and labelling. Furthermore, because of limited resources, the sample size of positive dataset was not large enough (imbalanced dataset), which affected the accuracy of data modelling. Thus, we would suggest collecting more data from different hospitals in future for the handling of imbalanced dataset issue. Finally, more studies with detailed evaluation and cross-validation are needed before the predictive risks of colorectal cancer can apply in practice.

Conclusion:

This local POC study showed that supervised machine learning could use patient’s age, sex and CBC results to predict the risks of colorectal cancer.