iSRRS2019
The 15th International Conference on Rapid Response Systems and Medical Emergency Teams
18-19 April | Suntec Singapore Exhibition Convention & Centre
Organised by:
In association with:
6thSG ANZICS
Managed by:
KENES GROUP

Abstract

SG-iSRRS1108

Text Mining of Clinical Progress Notes to Predict Future Onset of Sepsis in Hospitalized Patients

[1]Goh Kim Huat,[2]Adrian Yeow Yong Kwang,[3]Hermione Poh,[3]Li Ke,[3]Joannas Yeow Jie Lin,[3]Gamaliel Tan Yu Heng

[1]Division of Information Technology and Operations Management, Nanyang Technological University, Singapore
[2]School of Business, Singapore University of Social Sciences, Singapore
[3]Medical Informatics, Jurong Health Services, Singapore

Objectives:

Sepsis is associated with a high mortality rate and presents as a complex clinical problem for clinicians. Identifying and treating sepsis early has been shown to improve outcomes. As such there has been much work done in developing sepsis predictors (sniffers) in order to assist clinicians in identifying patients most at risk. However, there is a wide range of accuracy among these sniffers, which have been built using structured data models (reported ROC AUC ranging from 0.64 to 0.90 (Henry et al. 2015; Mani et al. 2014; Thiel et al. 2010)). Our study attempts to develop a sepsis sniffer that can combine structured data models with clinical information gleaned from clinical progress notes. It is a well-established fact that clinical progress notes are a rich source of unstructured data and many clinical applications have used text-mining as a parser or automated coder to extract symptoms and medication administration from these notes. In this study, we used text mining to extract relevant clinical information from free-text clinical progress notes to improve the accuracy of sepsis prediction.

Methods:

Study Period and cohort: April 2015 to February 2018.

We collected clinical progress notes from patients in Ng Teng Fong General Hospital (NTFGH) stored in the hospital's Electronic Medical Record (EMR) system (Epic Systems©). Each clinical progress note in the EMR is treated as a unit of analysis. All clinical progress notes were extracted by staff from the NTFGH Medical Informatics department and de-identified using recommendations published outlined the Personal Data Protection Act (PDPA) as well as the Health Insurance Portability and Accountability Act (HIPAA) section on anonymization of data prior to processing.

Training and Validation Phase (April 2015 to October 2017): Our training and validation data set contained 21,561 medical records of patients that were coded to have developed sepsis (ICD-10 classification) during their hospital stay and 77,316 medical records of randomly selected non-septic hospital patients (total of 98,877 medical records). We obtained the medical records of the septic patients prior to the onset of sepsis during their hospitalization. This sample was used to train and test the predictive model.

Testing Phase (November 2017 to February 2018): Our hold-out sample contained 2,678 medical records of patients that developed sepsis during their hospital stay and 8,815 medical records of non-septic hospital patients during the same time period (total of 11,493 medical records). This hold-out sample was used to test the accuracy of the predictive model developed in the training and validation phase. Patients' septic statuses were masked during the test and only used to test the accuracy of the model.

Data Source: Ng Teng Fong General Hospital EMR database

Data collected: Clinical progress notes, patient demographics, vital sign measures, lab test results, use of vasopressors, all culture results.

Ethics Approval: NHG DSRB Ref: 2018/00455

Estimation Model: We first applied Natural language processing (NLP) to the clinical progress notes to derive textual information (in the form of vectors of topics). We then used the textual information together with other structured predictors of sepsis as independent variables in a logistic regression for early prediction of sepsis among patients.

Results:

In the training and validation phase, the developed predictive model that included textual information from the progress notes has a ROC AUC of 0.9397 (compared to the predictive model without textual information – ROC AUC of 0.9001). The technical specification of this model was subsequently applied to a hold-out sample. The predictive model with textual information when tested on the hold-out sample also produced a similar ROC AUC of 0.9338 (the predictive model without textual information – ROC AUC of 0.8885). The test results achieved a good balance between sensitivity (89%) and specificity (89%).

We also observed that the addition of the textual information from the progress notes to the sepsis prediction model increased the AUC by approximately 4 percent in the training sample and approximately 4.5 percent in the testing sample.

Conclusions:

The improvement in AUC in the sepsis predictor suggests that the addition of topics derived from text-mined progress notes provided additional value to the sepsis predictive model. Specifically, our study shows that textual information help improves both the sensitivity and specificity of a predictive model. More importantly, based on our review of current medical studies, our predictive model achieves significantly better performance when compared to existing sepsis detection models which rely solely on structured variables. Our study also demonstrates the value of adopting text mining to implement machine learning on unstructured data in EMR systems.