A Novel VTE Prediction Model using Natural language Processing (NLP) and Machine Learning Methods

Nayan Pawar[1], Dr. (Prof). Amol Potgantwar[2], Dr. (Prof). Mangesh Ghonge[3]

[1]Department of Computer Engineering, Sandip Institute of Technology and Research Centre, Nashik, India
[2]HOD, Computer Dept,, Sandip Institute of Technology and Research Centre, Nashik, India
[3]Computer Dept,, Sandip Institute of Technology and Research centre, Nashik, India
[1,2,3]M. E. Computer Engineering. of Sandip Institute of Technology and Research Centre, University of Pune, Maharashtra, India

---

Abstract - Padua linear model is widely used for the risk assessment of venous thromboembolism (VTE), which is a common and preventable complication for inpatients. However, differences of race, genetics and environment between Western and Chinese population limit Padua model' validity in Chinese patients. Extracting VTE risk factors from unstructured medical records in Chinese hospital can help to understand VTE events and develop efficient risk assessment model. In this study, we proposed an ontology-based method to mine VTE risk factors combining natural language processing (NLP) and machine learning (ML) methods. Medical records of 3106 inpatients were processed and terms in multiple ontologies from various sections of records enriched in VTE patients were sorted automatically. Then ML methods were used to estimate terms' importance and terms within admitting diagnosis and progress notes showed better VTE prediction performance than other sections. Finally a novel VTE prediction model was built based on selected terms and showed higher AUC score (0.815) than the Padua model

KEYWORDS: Medical Record, Venous thromboembolism (VTE), Natural Language Processing (NLP), Risk Assessment, Machine Learning (ML).

1. INTRODUCTION

As a common complication for inpatients, venous thromboembolism (VTE) comprising pulmonary embolism (PE) and deep venous thrombosis (DVT) is a preventable cause of death. Since it is closely related to ethnic background and disease spectrum, Chinese differed from the Caucasians in the disease risk assessment, which caused poor performances of the Padua model recommended by American College of Chest Physicians when applied in Chinese inpatients in the Internal Department [1]. Thus, to find the potential VTE risk factor and develop prediction model specified to Chinese inpatients are warranted. Besides, the rapid development of medical informatization and electronic health record (EHR) system allows the accumulation of increasing number of medical records and provides the possibility of investigating diseases in more elaborate and precise methods, compared with traditional approaches with small sample size and fewer variables. Many researchers have studied relationships between Various diseases and risk factors using machine learning (ML) and natural language processing (NLP) methods and showed promising results. Weng, et al. compared predictive validities of multiple ML methods on cardiovascular risk assessment, Casanova, et al. [3] analyzed the Alzheimer's disease risk by regularized logistic regression, and Ferroni, et al. trained the support vector machine to do VTE risk prediction for cancer patients. However, above studies' features are pre-designed and limited, which can hardly take full advantages of quantities of patient information in medical records and discover new knowledge such as other potential variables associated with the disease. Some deep learning models are also proposed to combine medical ontologies to analyze high dimensional and heterogeneous medical records but their results lack of interpretability. In order to help the clinicians explore new candidate VTE risk factors and develop efficient prediction model with certain interpretability from medical records, we propose an ontology-based approach which processes the free-text in medical records carefully, evaluates importance of terms from ontologies automatically and finally constructs the model based on candidate terms. The whole workflow needs no clinician's guidance and the generated results can inspire further medical studies.

II. METHODS

A. Medical records from Chinese hospital :

In this study medical records of 3106 inpatients were collected from Peking Union Medical College Hospital (PUMCH) and every patient had two documents, admission note and progress note, which were both unstructured and had lots of paragraphs consisting of free text. The admit note usually included 11 sections: chief complaint, present history, previous history, personal history, family history, obsterical history, menstrual history, physical examination, laboratory examination, admitting diagnosis and physician's signature, and the progress note had daily description about patient condition. Among them, 224 VTE inpatients were checked and Padua scores were calculated by the clinicians in previous study [7]. All the patients enrolled were over 18 years old with a hospital stay length 72 hours and met the exclusion criteria of receiving anticoagulation treatment (except for anticoagulation for the treatment of VTE diagnosed during hospitalization). DVT was diagnosed by venography or color Doppler ultrasonography. PE was

diagnosed by pulmonary angiography, computed tomographic pulmonary arteriography, MRI or radionuclide lung ventilation-perfusion scans (V/Q scans).

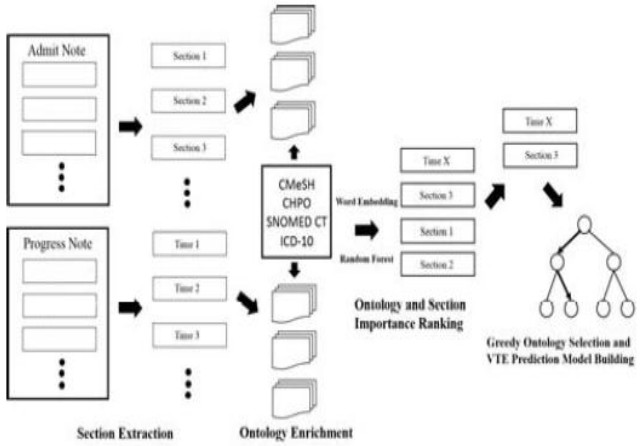

Fig -1: System Architecture

B. Ontology sources :

In order to obtain comprehensive information of patients involving symptoms, diagnoses, drugs, operations and so on, multiple kinds of ontologies were gathered, including the Chinese versions of Medical subject headings (MeSH) [8], Human Phenotype Ontology (HPO) [9], SNOMED [10] and ICD-10 [11]. After preprocessing, 37111 ICD codes, 11903 phenotypic abnormalities from HPO, 55750 MeSH words, and 11652 terms from SNOMED were saved and merged as the final ontology set.

C. Ruled-based section extraction:

The workflow of ontology extraction and risk assessment model establishment was shown in Fig 1 and the first step was parsing sections within medical records. Due to the pattern that sections started with specific titles and white space existed between two sections, we utilized the regular expression to capture the start position of one section, and considering that some sections were removed by clinicians for convenience and orders of them could be changed, a greedy match algorithm was implemented to find the title of section having the minimum distance with current position iteratively. The text between two titles was regarded as a section related to the first title and current position in the document was updated after a section was recognized. For the admit note, 8 sections were parsed excluding the obsterical history, menstrual history and physician's signature. For the progress note, the daily objective description were split via the date.

D. Ontology-based VTE risk assessment model:

Having importance ranking of terms and sections, we proposed an approach to select the section automatically and then build RF model based on picked terms. Just like the greedy feature selection method, every time terms from one section with relatively high AUC (Only) were added to construct new features and RF models were re-trained to try to improve present prediction performance. The selection process stopped when there was no improvement on AUC scores.

III. RESULTS

A. Ontologies from nine kinds of sections

Basic information of terms within distinct sections from non-VTE and VTE patients respectively during the process of 'Automatic Ontology Enrichment' were shown at Table 1, and the number of terms and neighbors were listed. It can be seen that the section 'Present History' had the most terms (974 in non-VTE and 1034 in VTE) and the 'Personal History' was the least (14 in non-VTE and 13 in VTE) in both non-VTE and VTE. As for the neighborhood information, size of neighbors of terms in the 'Progress Note' was far more than other sections (143 in non-VTE and 31 in VTE), and oppositely, the 'Chief Complaint' had the smallest neighbors (13 in non-VTE and 2 in VTE). Generally counts of terms between the non-VTE and the VTE were comparable but terms in the former had more abundant neighbors, which may be resulted from their gap of number of samples.

B. Comparison of ontology and section importance

Top K=100 terms of each section from VTE patients were extracted and RF models were trained based on word vectors of VTE' terms. AUC (Only) and AUC (Exclusion) scores of 9 types of sections were shown at Table 2. Obviously terms from the 'Progress Note' had the highest AUC (Only) score, 0.805 and results excluding them led to the worst AUC (Exclusion) value, 0.630, which verified its key role in VTE prediction. Except the 'Progress Note', AUC (Exclusion) scores of remaining sections were similar, which implied that efficiency of terms of them weren't much different. Only considering the AUC (Only), the second best section was the 'Admitting Diagnosis' with the value 0.690 and terms from the 'Personal History' showed the lowest score, 0.540. One interesting thing was that when we used all terms, the prediction validity was less than the result of terms from single section such as the 'Progress Note', which reflected the importance of ontology and section evaluation.

C. Performance of new VTE risk assessment model

By selecting terms and sections greedily, terms of two sections, the 'Progress Note' and the 'Admitting Diagnosis', were chosen and achieved the best VTE assessment performance. Based on terms from the 'Progress Note' and 'Admitting Diagnosis', a new VTE risk assessment model was built and its AUC score, 0.815, was higher than the traditional Padua model, 0.789 (Table 3). In addition, new model had the superiority of high specificity (0.795) although its sensitivity (0.676) was inferior to the Padua. Compared with the AUC (Only) value (0.805) of the 'Progress Note', supply of terms from the 'Admitting Diagnosis' improved the model predictive validity.

IV. CONCLUSION

In this study, a method of ontology-based VTE risk factors mining and model establishment from medical records is developed and its efficiency is demonstrated on real clinical dataset from PUMCH. Selected terms and sections from medical records help the clinicians discover potential VTE risk factors and RF model built based on these terms improves the performance of VTE prediction. This method is expected to be applied in more diseases and embedded into the EHR system to assist clinical work.

REFERENCES

[1] Barbar, S., Noventa, F., Rossetto, V., Ferrari, A., Brandolin, B., Perlati, M., De, B.E., Tormene, D., Pagnan, A., and Prandoni, P.: 'A risk assessment model for the identification of hospitalized medical patients at risk for venous thromboembolism: the Padua Prediction Score', Journal of Thrombosis & Haemostasis Jth, 2010, 8, (11), pp. 2450-2457

[2] Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M., and Qureshi, N.: 'Can machine-learning improve cardiovascular risk prediction using routine clinical data?', Plos One, 2017, 12, (4), pp. e0174944

[3] Casanova, R., Hsu, F.C., Sink, K.M., Rapp, S.R., Williamson, J.D., Resnick, S.M., and Espeland, M.A.: 'Alzheimer's Disease Risk Assessment Using Large-Scale Machine Learning Methods', Plos One, 2013, 8, (11), pp. e77949

[4] Ferroni, P., Zanzotto, F.M., Scarpato, N., Riondino, S., Nanni, U., Roselli, M., and Guadagni, F.: 'Risk Assessment for Venous Thromboembolism in Chemotherapy-Treated Ambulatory Cancer Patients: A Machine Learning Approach', Medical Decision Making An International Journal of the Society for Medical Decision Making, 2016, 37, (2)

[5] Riccardo, M., Li, L., Kidd, B.A., and Dudley, J.T.: 'Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records', Scientific Reports, 2016, 6, pp. 26094

[6] Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., and Sun, J.: 'GRAM: Graph-based Attention Model for Healthcare Representation Learning', 2016, pp. 787-795

[7] Wang X., Hong X., Li J., Zhao R., Yang Y., Liu S., Sun X., Zhu W., Fan J., and Shi J.: þValue of Padua Risk Assessment Model in Evaluating Venous Thromboembolism of Hospitalized Patients in the Development of Internal Medicineÿ, Meidical Journal of Peking Union Medical College Hospital, 2018, (3)

[8] Lipscomb, C.E.: 'Medical subject headings (MeSH)', Bulletin of the Medical Library Association, 2000, 88, (3), pp. 265

[9] Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., and Campbell, J.: 'The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data', Nucleic acids research, 2013, 42, (D1), pp. D966-D974

[10] Donnelly, K.: 'SNOMED-CT: The advanced terminology and coding system for eHealth', Studies in health technology and informatics, 2006, 121, pp. 279

[11] Trott, P.: 'International classification of diseases for oncology', Journal of clinical pathology, 1977, 30, (8), pp. 782

[12] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: ' Efficient estimation of word representations in vector space', arXiv preprint arXiv:1301.3781, 2013

[13] Rehurek, R., and Sojka, P.: 'Software framework for topic modelling with large corpora', in Editor (Ed.)^(Eds.): 'Book Software framework for topic modelling with large corpora' (Citeseer, 2010, edn.), pp.