Predicting 30-day readmission through emergency room by using Cross-language Topic Modeling from discharge note in OMOP-CDM

Giup Jang[1], Seng Chan You, MD[2], Dongsu Park[2], Youngmi Yoon, PhD[3], Rae Woong Park, MD, PhD[2,4]

[1]Department of IT Convergence Engineering, Gachon University, Seongnam, Korea;
[2]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea;
[3]Department of Computer Engineering, Gachon University, Seongnam, Korea;
[4]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea

Background

- Although OMOP-CDM contains Note and Note_NLP table, the standardized NLP tool for variable languages other than English has not been developed yet.
- A considerable number of hospitalized patients are readmitted within 30 days, and many of these readmissions are avoidable. Identifying subjects at high risk for readmission is of paramount importance for the timely intervention to reduce socioeconomic burden.
- Though numerous risk prediction models have been proposed until now, most of these models extracted information for the algorithm from only structured fields in electronic health records (EHR) or administrative databases. This limited approach, however, cannot capture more detailed information, such as symptoms and functional status of patients, embedded in the narratives by clinicians.

Purpose

- We aim to develop the cross-language natural language processing (NLP) module for medical free-text in OMOP-CDM by using topic modeling.
- To demonstrate the feasibility, we build prediction model for 30-day readmission through emergency room by combining features from structured clinical data and unstructured free-text in discharge note in OMOP-CDM.

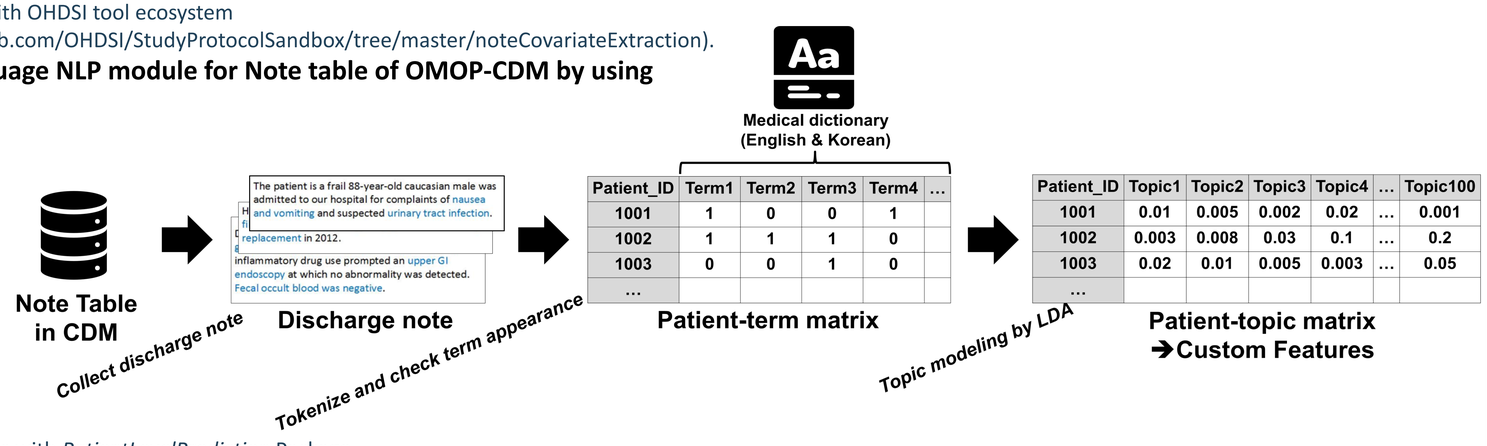
Experiment Settings

- Objective of experiment is building prediction model for readmission through emergency room to validate feasibility and usefulness of proposed NLP process.
- The whole process of experiments is available at OHDSI github (https://github.com/OHDSI/StudyProtocolSandbox/tree/master/PredictionOfRehospitalizationWithNote).
- Database: Ajou university EHR data
- Target cohort at risk: Subjects who admitted to the hospital and stayed 7 days or more from 1st January 2005 to 1st December 2017. Among them, only who had discharge note in the CDM note table were included. The cohort start date of target cohort was set by their discharge date.
- Outcome cohort: Subjects who readmitted through emergency room within 30 days after discharge
- Among 108,541 target subjects at risk, total of 12,236 (11.3%) developed outcome.
- Gender, age group, race, ethnicity, index year, index month and condition within 30 days were used for structured clinical variables.
- Two prediction mode was built. One used only structured clinical variables and the other used combined variables of both structured clinical variables and variables generated by proposed NLP process. The number of topics was set as 100.
- The best hyperparameters were set by two-fold cross-validation. 25% of population were separated for the test.

Method: Overall Process

1. Tokenization of medical free-text based on medical dictionary
   - Tokenization is to divide a sentence into a minimum number of meaningful units (most tokens are separated by spaces).
   - Only the tokens representing medical terms were extracted from the whole tokenized words. Here, we used Korean and English medical dictionary. This process can be applied to medical records in other language by adding the medical dictionary.
2. Topic modeling by latent dirichlet allocation (LDA)
   - Topic modeling is a statistical algorithm for discovering the main topics or themes from vast numbers of unstructured documents.
   - Topic modeling allows a document to have multiple topics and to analyze the characteristics of the document in more detail than common cluster method.
   - LDA is one of the topic modeling algorithm, and it is highly modular and can be easily extended.
3. Extracting features from note (**Fig 1**)
   - Values for each topic estimated from the note by topic modeling were allocated into individual covariates. We developed *noteCoavariateExtraction* function, which is compatible with OHDSI tool ecosystem (https://github.com/OHDSI/StudyProtocolSandbox/tree/master/noteCovariateExtraction).

Fig 1. Cross-language NLP module for Note table of OMOP-CDM by using Topic Modeling



4. Create model settings with *PatientLevelPrediction* Package
   - The model basically used basic setting of gradient boosting machine in *PatientLevelPrediction* package.
5. Fitting the model and evaluating the result with *PatientLevelPrediction* Package

- Acknowledgment : This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea grant number: HI14C3201 and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science, ICT & Future Planning) (NRF-2018R1A2B6006223).
- Conflict of interest : none

Experimental result

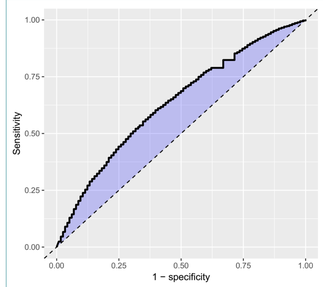


Fig 2. ROC plot

Fig 3. Demographic summary

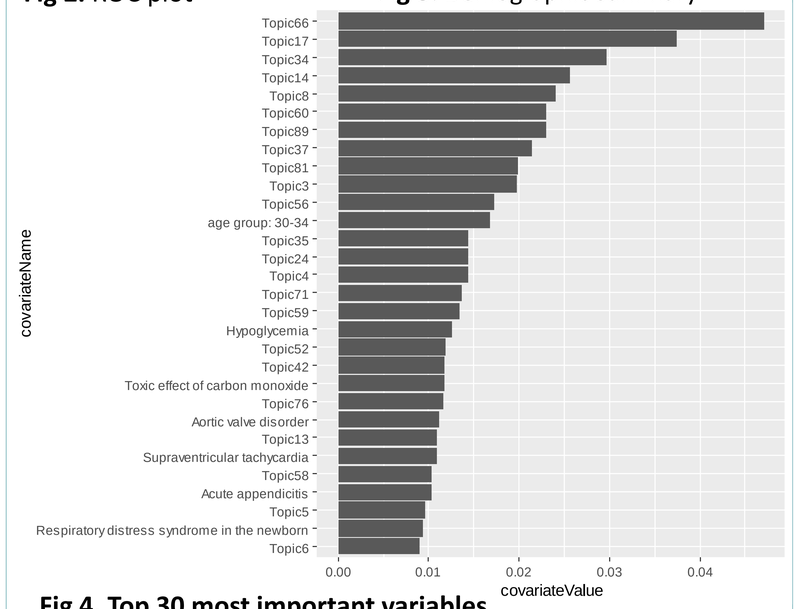


Fig 4. Top 30 most important variables

- The Area under the ROC curve (AUROC) was slightly higher when using variables generated by proposed NLP process than only using structured variables : 0.641 vs. 0.636. Area under precision recall curve was similar between the model using free-text features and model without this: 18.18 vs 17.99, respectively.
- Demographic summary shows that there is no remarkable trends of readmission associated with demographics (**Fig 3**).
- Among the most important top 30 variables, 23 variables (77%) from topics were included (**Fig 4**). Other than variables from the free-text of discharge note, age group 30 to 34, hypoglycemia, toxic effect of carbon monoxide, aortic valve disorder, supraventricular tachycardia, acute appendicitis and respiratory distress syndrome in newborn were important variables.

Conclusion

- We developed a cross-language NLP module for Note table of OMOP-CDM by using topic modeling.
- The feasibility of this cross-language NLP module was demonstrated in the experiment, which predicted 30-day readmission through emergency room.
- Additional module to identify meaning of topics will be added.