**Abstract**

**Motivation**: Electronic health records (EHR) are widely used by many hospitals to store and organize patient information; however, the crucial information is usually buried within the extensive descriptive text. To fully exploit the utility of EHR, natural language processing (NLP) may aid doctors to summarize the patient history and status.

**Methods**: Given EHR annotated with coronary artery disease (CAD) risk factors, data was cleaned to unify the structure of every EHR. Three models: rule-based, deep learning and traditional machine learning method were compared for their performance then Naive Bayes algorithm and rule-based algorithm are combined and implemented to group each word in text into categories. Specifically, rule-based algorithm focused on family history while Naive Bayes is applied to the rest of the categories.

**Results**: The evaluation is done on a document level to reflect whether a patient shows any signs of this risk factor. The weighted F1 score of the combined model is 0.916. The result is summarized in a user interface.

**Conclusion**

The final model evaluation is done on the document-level to focus on patient-level statistics. Medication has the best performance due to its unique word bank. CAD is less ideal perhaps due to original tag quality. However, the performance of each category has well supported the document-level risk factor prediction based on patient’s health record.

---

**Methods**

Remove extra spaces and special characters

Remove fully overlapped tags and merge partially overlapped tags

Modify phrases into alliteration; remove stop words

Map the cleaned tags back to original text and using longest match

**Preprocess data**

**Compare models**

**Select models and train**

**User interface**

**Fig.1 General method for modeling and user interface**

**Evaluation**

Performance of three different methods on text-level were tested. Deep learning method is less suitable due to the limited data source and was not used. Among classic machine learning models, Naive Bayes shows better performance and rule-based method is applied to category with few instances.

<table>
<thead>
<tr>
<th></th>
<th>Machine learning</th>
<th>Deep learning</th>
<th>Rule-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted ave.</td>
<td>0.8990</td>
<td>0.6870</td>
<td>0.8990</td>
</tr>
<tr>
<td>Micro ave.</td>
<td>0.8990</td>
<td>0.6870</td>
<td>0.8990</td>
</tr>
<tr>
<td>Macro ave.</td>
<td>0.8990</td>
<td>0.6870</td>
<td>0.8990</td>
</tr>
<tr>
<td>Precision</td>
<td>0.6770</td>
<td>0.7770</td>
<td>0.6770</td>
</tr>
<tr>
<td>Recall</td>
<td>0.6770</td>
<td>0.7770</td>
<td>0.6770</td>
</tr>
<tr>
<td>F1</td>
<td>0.3106</td>
<td>0.3106</td>
<td>0.3106</td>
</tr>
</tbody>
</table>

**Fig.5 Evaluation and Comparison of different models**

The final model evaluation is done on the document-level to focus on patient-level statistics. Medication has the best performance due to its unique word bank. CAD is less ideal perhaps due to original tag quality. However, the performance of each category has well supported the document-level risk factor prediction based on patient’s health record.

**Conclusion**

The performance of risk factor prediction is improved by using a multi-model approach. The development of document-level model and user interface may help summarize patients’ risk factors over time and help clinicians manage patient care with reduced burden. In future work, we would aim to improve performance of identifying these risk factors by incorporating additional rules. We would also want to get clinician input on the utility of this tool and design feedback.