Leveraging Multi-Word Concepts to Predict Acute Kidney Injury in Intensive Care

Lorenzo BRANCATOA, Iacer CALIXTOb,c,
Ameen ABU-HANNAb,c and Iacopo VAGLIANOb,c,1

a University of Pavia, Pavia, Italy
b Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands
c Amsterdam Public Health Research Institute, Amsterdam, the Netherlands

ORCiD ID: Lorenzo Brancato https://orcid.org/0000-0002-9357-9175, Iacer Calixto https://orcid.org/0000-0001-6244-7906, Ameen Abu-Hanna https://orcid.org/0000-0003-4324-7954, Iacopo Vagliano https://orcid.org/0000-0002-3066-9464

Abstract. Acute kidney injury (AKI) is an abrupt decrease in kidney function widespread in intensive care. Many AKI prediction models have been proposed, but only few exploit clinical notes and medical terminologies. Previously, we developed and internally validated a model to predict AKI using clinical notes enriched with single-word concepts from medical knowledge graphs. However, an analysis of the impact of using multi-word concepts is lacking. In this study, we compare the use of only the clinical notes as input to prediction to the use of clinical notes retrofitted with both single-word and multi-word concepts. Our results show that 1) retrofitting single-word concepts improved word representations and improved the performance of the prediction model; 2) retrofitting multi-word concepts further improves both results, albeit slightly. Although the improvement with multi-word concepts was small, due to the small number of multi-word concepts that could be annotated, multi-word concepts have proven to be beneficial.

Keywords. Clinical Prediction, Natural Language Processing, Knowledge Graphs.

1. Introduction

Acute kidney injury (AKI) is an abrupt decrease in kidney function, which can affect up to 50% of patients admitted to the intensive care unit (ICU) [1]. Various prediction models have been proposed to predict AKI, typically using clinical variables, such as vitals and laboratory measurements [2]. Few studies used the rich information contained in clinical notes, possibly because it is not well-known how to represent such information in AKI prediction models.

Knowledge graphs may enrich the representation of clinical notes in prediction models. Medical knowledge graphs include the Unified Medical Language System (UMLS) [3]. In a previous study [4], we showed how enriching the representation of clinical notes with single-word concepts from knowledge graphs in AKI prediction models can improve predictive performance. However, the prevalence of multi-word concepts and the impact of using them have not yet been investigated.

1 Corresponding Author: I. Vagliano; Dept. of Medical Informatics, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands; E-mail: i.vagliano@amsterdamumc.nl.
In this study, we compare the use of only clinical notes to the use of clinical notes enriched with both single-word (e.g. Hearth) and multi-word concepts (e.g. Hearth structure). We investigate whether enriched representations of clinical notes can improve the predictive performance of AKI prediction models. We develop and internally validate a model to predict AKI within the first 48 hours of ICU stays using clinical notes.

2. Methods

The publicly-available Medical Information Mart for Intensive Care III (MIMIC-III) dataset was used [5]. AKI was defined according to the KDIGO guidelines [6]. We included patients over 18 years at the time of ICU admission, who have at least one measurement of serum creatinine or urine output, and whose length of stay in the ICU was at least 48 hours, otherwise the patient was discarded.

UMLS concepts were recognized and normalized through ScispaCy [7]. Synonyms were retrieved with PyMedTermino [8]. We only use the synonymy relation across words defined in the UMLS. We pre-trained 100-dimensional word embeddings with GloVe (Global Vectors) [9]. We used retrofitting to generate a refined representation of clinical notes [10].

Our AKI prediction model is based on Long Short Term Memory (LSTM) networks [11]. The model’s input is clinical notes’ embeddings. Three models were trained: one using only clinical notes; one using clinical notes retrofitted with single-word synonyms; and one using clinical notes retrofitted with both single-word synonyms and multi-word synonyms. The model’s architecture is shown in Figure 1. The embedding layer acts as a lookup table to return the word embeddings learned as parameters by the model during training. The bidirectional LSTM layer is followed by a dropout layer and a linear combination layer, which combines the bidirectional output. The final layer returns the probability of a patient having or developing AKI.

The dataset was randomly split into 80% training, 10% validation, and 10% test sets. Discrimination was measured with the area under the receiver operating curve (AUROC). The interplay between the positive predictive value and sensitivity was measured by the area under the precision-recall curve (AUPRC). Calibration was assessed with calibration curves. The accuracy of the predicted probabilities was measured by the Brier score. Because of the small number of annotated medical concepts in the notes, we evaluated predictive performance in test sets where we selected only patients whose proportion of medical concepts with respect to the number of words in their notes was above 20% and 21%, respectively. We also evaluate predictive performance in test sets where we selected only patients whose proportion of multi-word...
medical concepts was above 10%, 11%, and 12%, respectively. The selected thresholds were based on the distribution of concepts per patient.

Similarly to Albi et al. [12], we relied on the t-distributed stochastic neighbor embedding (t-SNE) [13] to qualitatively assess the learned embeddings.

3. Results and Discussion

The final dataset consisted of 46,985 ICU stays of 33,795 patients. GloVe embeddings resulted in a vocabulary of 84,879 words and 27,000 (single- and multi-word) synonyms were identified. Table 1 outlines the predictive performance of the models. Using multi-word synonyms achieved the best results in all the test sets. The improvement in AUROC when using multi-word synonyms compared to using only clinical notes in the full test set was about 1%. Using single-word synonyms performed similarly to multi-word synonyms in the full test set. When considering only patients with over 21% multi-word concepts, using multi-word synonyms yielded an increment of 3.3% in AUROC compared to using only clinical notes, and an increment of 1.5% with respect to using single-word synonyms. Using multi-word synonyms resulted in an increase of 6% in AUROC compared to using only clinical notes, and an increase of 1.3% with respect to using only single-word synonyms when considering only patients with over 12% multi-word concepts. In patients with over 12% multi-word concepts, the overall performance was lower because of the small number of patients included. Further thresholds showed a similar trend. The calibration curves of the models and the word embedding visualizations for the selected medical entities are available separately.

Table 1 Model’s discrimination with various inputs.

<table>
<thead>
<tr>
<th>Test set</th>
<th>Measure</th>
<th>Clinical notes</th>
<th>Clinical notes + single-word synonyms</th>
<th>Clinical notes + multi-word synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full (n = 4710)</td>
<td>AUROC</td>
<td>0.790</td>
<td>0.797</td>
<td>0.799</td>
</tr>
<tr>
<td></td>
<td>AUPRC</td>
<td>0.894</td>
<td>0.897</td>
<td>0.898</td>
</tr>
<tr>
<td></td>
<td>Brier score</td>
<td>0.168</td>
<td>0.165</td>
<td>0.164</td>
</tr>
<tr>
<td>≥ 21% synonyms (n = 112)</td>
<td>AUROC</td>
<td>0.819</td>
<td>0.837</td>
<td>0.852</td>
</tr>
<tr>
<td></td>
<td>AUPRC</td>
<td>0.879</td>
<td>0.893</td>
<td>0.904</td>
</tr>
<tr>
<td></td>
<td>Brier score</td>
<td>0.175</td>
<td>0.172</td>
<td>0.167</td>
</tr>
<tr>
<td>≥ 12% multiword synonyms (n = 78)</td>
<td>AUROC</td>
<td>0.629</td>
<td>0.674</td>
<td>0.687</td>
</tr>
<tr>
<td></td>
<td>AUPRC</td>
<td>0.787</td>
<td>0.814</td>
<td>0.831</td>
</tr>
<tr>
<td></td>
<td>Brier score</td>
<td>0.221</td>
<td>0.210</td>
<td>0.200</td>
</tr>
</tbody>
</table>

Our results show good performances (AUROC of 0.8). All models are well-calibrated. The improvement of retrofitted word embeddings is larger when considering patients with a higher proportion of medical concepts with respect to the number of words in their notes. The t-SNE shows a qualitative improvement in the retrofitted embeddings.

The small improvement in performance when using multi-word concepts might be due to the limited number of multi-word synonyms in the notes, which are only 6.4% of the total number of words used. Furthermore, the average proportion per patient of single-word concepts versus the total number of words is 10.1%, and 10.2% for multi-word concepts. Thus, limited information is added by multi-word concepts.

2 https://osf.io/bz8uk, last access 21/04/2023
Our study has some limitations. First, the MIMIC III database includes US patients; thus our results may not generalize to other populations. Second, we performed a simple train/validation/test split, which does not consider the variability of train/validation/test sets. Third, we did not perform full hyperparameter tuning but relied on a set of parameters for our models pre-selected in preliminary experiments. As a strength, our study relies on a public dataset to encourage reproducibility and our code is available at bitbucket.org/aumc-kik/ml-cn-kg-4-aki-prediction-ext.

4. Conclusions

Early recognition of AKI is essential for the effective treatment of this disease in the ICU. We showed that enriching the representation of clinical notes with medical concepts from knowledge graphs can improve predictive performance. Our models yielded good results. Although the further improvement with multi-word concepts was small, due to the low number of multi-word concepts represented, our results suggest that multi-word concepts provide additional predictive value.

References