In-Hospital Cancer Mortality Prediction by Multimodal Learning of Non-English Clinical Texts

Shintaro OYAMA\textsuperscript{a}, Taiki FURUKAWA\textsuperscript{b}, Shotaro MISAWA\textsuperscript{c}, Ryuji KANO\textsuperscript{c}, Hirokazu YARIMIZU\textsuperscript{c}, Tomoki TANIGUCHI\textsuperscript{c}, Kohei ONODA\textsuperscript{c}, Kikue SATO\textsuperscript{b} and Yoshimune SHIRATORI\textsuperscript{b}

\textsuperscript{a} Innovative Research Center for Preventive Medical Engineering, Nagoya University, Nagoya, Japan
\textsuperscript{b} Nagoya University Hospital Medical IT Center, Nagoya University, Nagoya, Japan
\textsuperscript{c} FUJIFILM Corporation, Tokyo, Japan

Abstract. Predicting important outcomes in patients with complex medical conditions using multimodal electronic medical records remains a challenge. We trained a machine learning model to predict the inpatient prognosis of cancer patients using EMR data with Japanese clinical text records, which has been considered difficult due to its high context. We confirmed high accuracy of the mortality prediction model using clinical text in addition to other clinical data, suggesting applicability of this method to cancer.

Keywords. Non-English EMR clinical text processing, Inpatient Cancer Mortality Prediction

1. Introduction and Methods

Inpatient outcome and admission prediction has been limited to the use of generalized and simple clinical decision rules (CDRs), which are not sufficiently accurate for individualized prediction, especially for complex pathologies such as those often seen in cancer patients. Using machine learning (ML) techniques that utilizes the multiple variables obtained from electronic medical record (EMR) could solve these problems and may lead to advanced clinical decision support systems. Previous studies have reported mortality prediction of pneumonia patients [1] and general inpatients prediction [2]. However, these studies were limited to specific clinical conditions and did not use the data from medical records, which presented a challenge to accuracy. In particular, medical records entered by physicians and nurses have been considered to have a high importance on supplement subtle changes in a patient’s condition and contribute to accurate prediction [3], but they are mostly unstructured data, which makes them difficult to use. In addition, research on natural language processing in medical records has been conducted in English and other low-content languages, and not enough has been done in high-context languages such as Japanese [4]. In this study, we trained ML model to predict in-hospital mortality of cancer patients using multimodal clinical information.
including Japanese clinical texts, and tested the accuracy of the prediction and the difference in accuracy by cancer tumor. EMR data were extracted for patients admitted to a single tertiary hospital (Nagoya University Hospital) from January 2018 to November 2020. The data characteristics include total admissions: 55,687; mean age: 62.15yo; mortality rate: 1.15%. Inpatients admitted after January 2020 were treated as test data and the rest as training data. Since the number of deaths (205 cases) and survivors (20,158 cases) differed greatly in the training data, we down-sampled survivors to 10 times the number of deaths, as an input feature, we extracted medical statements, tests, procedures, fluid and food intake, and basic patient information as input variables from EMR data up to the second day of admission, and used the Gradient Boosting Decision Trees (GBDT)-based model that we have constructed in a previous study [1]. Test data were selected for cancer types (lung cancer, hepatocellular carcinoma, and colorectal cancer) with more than fifty patients including at least three death cases during the period, and predictive performance was evaluated using the Area Under the ROC Curve (AUROC score) for the three cancer types along with all diseases and cancer types.

### 2. Results, Discussion and Conclusions

Table 1. represents the statistics of our test datasets, and the performance of cancer mortality prediction of our model was competitive to that of all diseases. In addition, mortality of lung cancer was the most difficult to predict.

<table>
<thead>
<tr>
<th>Disease Type</th>
<th>#ADM</th>
<th>Mean days from ADM. to death</th>
<th>Mortality rate</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-Diseases</td>
<td>10,159</td>
<td>12</td>
<td>1.17%</td>
<td>96.50</td>
</tr>
<tr>
<td>All-types of cancer</td>
<td>2,202</td>
<td>12</td>
<td>1.95%</td>
<td>97.07</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>374</td>
<td>16</td>
<td>3.21%</td>
<td>90.24</td>
</tr>
<tr>
<td>Hepatocellular carcinoma</td>
<td>184</td>
<td>6</td>
<td>1.63%</td>
<td>98.90</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>54</td>
<td>17</td>
<td>3.70%</td>
<td>92.31</td>
</tr>
</tbody>
</table>

Predicting inpatient outcomes for complex diseases such as cancer has been considered difficult [2], thus improving accuracy, especially by using unstructured medical text in highly contextual languages such as Japanese, has been sought both for proper medical resource allocation and as a benchmark for hospital quality improvement. We constructed a multimodal machine learning model to predict inpatient mortality from Japanese EMRs and were able to predict inpatient mortality for cancer as well as other diseases. However, this is a preliminary study conducted on a dataset from a single medical institution and needs to be validated with medical information from multiple medical institutions.

### References