Deep learning approach to predict brain metastases development in lung cancer patients from EHR data

Zhao Li, MS¹, Ping Zhu, MMed¹,², Yujia Zhou, MS¹, Degui Zhi, PhD¹, Hua Xu, PhD¹, Yoshua Esquenazi, MD², W. Jim Zheng, PhD¹

¹School of Biomedical Informatics, ²Vivian L. Smith Department of Neurosurgery, ³Department of Epidemiology, Human Genetics, and Environmental Sciences, The University of Texas Health Science Center at Houston, Houston, Texas, United States

Introduction

Brain metastasis (BM) develops when cancer cells migrate from their primary site to the brain. The most common origins of BM include lung cancer, breast cancer, and melanoma¹. Early detection of BM will allow non-invasive therapies such as radiosurgery and thereby reduce neurological morbidity. In this study, we took a data drive approach and employed RETAIN to utilize the historical EHR data of lung cancer patients to predict risk of BM².

Materials and Methods

Cerner Health Facts® is a de-identified EHR database that contains over 70 million unique patients from over 600 clinical organizations in the United States³. In this study, we identified 203,504 lung cancer patients whose diagnoses were made before 2016 using ICD-9 code ‘162.*’ and ICD-10 code ‘c34.*’. Among these patients, we identified 26,923 BM patients with ICD-9 code ‘198.3’ and ICD-10 code ‘C79.3’.

Table 1. AUC on the separated test set with different models.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RETAIN – Demographic, diagnosis, medication and surgical</td>
<td>0.797</td>
</tr>
<tr>
<td>RETAIN – Lab test and clinical event</td>
<td>0.688</td>
</tr>
<tr>
<td>RETAIN – All data</td>
<td><strong>0.825</strong></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.756</td>
</tr>
<tr>
<td>Logistic Regression using embedding</td>
<td>0.788</td>
</tr>
</tbody>
</table>

Results and Conclusions

We employed a deep learning model, RETAIN², to predict the brain metastasis development from lung cancer patients using a large EHR dataset. A list of brain metastasis related diagnoses and medications were automatically identified and manually reviewed by a clinical expert to accurately define cases and controls.

After a series of well-designed case-control matching criteria, the overall performance using RETAIN with all features including demographic, diagnosis, medication, surgical, lab test and clinical event reaches to 82.5% for AUC as shown in Table 1. A list of high contribution factors was identified by averaging the contribution weights from RETAIN model, which indicates metastasis has already observed in other organs like bone and liver, and these patients are in their late stage of lung cancer.

Discussions

This is the first deep learning application to predict incident BM in lung cancer patients using EHR data. A main challenge for this study is to design good criteria for constructing a high-quality cohort. A limitation of this study is the absence of many important information in Cerner EHR database.

Acknowledgement This work is supported by the CPRIT grant RP170668 (WJZ). We would like to acknowledge the use of “Cerner Health Facts” and the assistance provided by UTHealth SBMI Data Service team to extract data.

References