Application of language model for extracting data from pathology reports

Gyubeom Hwang, MD[1,2], Min-Gyu Kim MD[1,2], Min Ho An MD[1,2], Rae Woong Park MD Ph.D[1]
[1] Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea
[2] Department of Medical Sciences, Graduate School of Ajou University, Suwon, Republic of Korea

Background

The common data model (CDM) provided an opportunity for the analysis stored in data from distributed databases. However, compared to other diseases, there are many obstacles to using a common data model to predict treatment effectiveness and prognosis for cancer patients. One reason for this is that the results of molecular analyses and immunochemical staining assays, which are critical for determining treatment plans and predicting prognosis for cancer patients, are not well transformed into the common data model and instead remain in an unstructured form in pathology reports. With the recent development and success of cancer immunotherapies, it is important to extract the clinical information contained in pathology reports and apply it to cancer research designs.[1] An attempt to transform immunochemistry and molecular test reports into Observational Medical Outcome Partnership (OMOP) CDM for colorectal cancer patients has been made in a previous study.[2] However, the methodology of the study, which used regular expressions, cannot be generalized to other institutions or to other conditions. Recently, large language models(LLMs) have shown good performance on a numerous tasks.[3] In this study, we aimed to explore whether the LLM can be applied to extract clinical information from pathology reports of hepatocellular carcinoma.

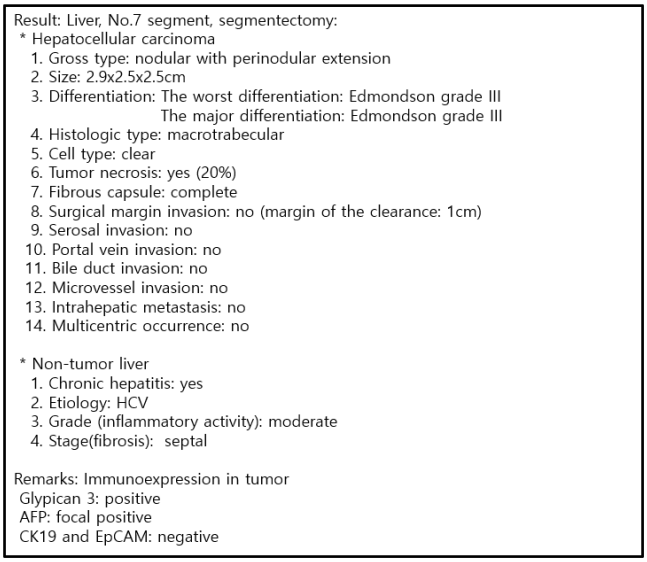
Methods

We used EHR data from Ajou University Medical Center that has been converted to the OMOP-CDM v5.3.1. Patients who had been diagnosed with hepatocellular carcinoma (HCC) and undergone curative hepatectomy as a first-line treatment for HCC within 30 days of diagnosis were selected. All liver specimens obtained from the hepatectomy were pathologically examined, resulting in a pathology report containing the results of immunochemical and molecular testing and these pathology reports were stored in the [NOTE] table. But unlike other notes created during the same admission, the [visit_occurrence_id] column in the pathology notes were empty. Therefore, [note_title] with pathology report written within 2 weeks after surgery were defined as pathology reports following surgery. From the total of 682 pathology reports obtained, 50 were randomly selected for the study. We tested whether the LLM could extract 8 key markers (glutamine synthetase, glypican3, EpCAM, AFP, CK7, CK19, CD34, CD13) from pathology reports. Results were compared with labels from the manual review by a physician. The performance of the model was evaluated by calculating precision and recall. We calculated precision and recall for both excluding and including cases where the model produced an output of N/A. For LLM, we used LLAMA-30b Supercot developed by Meta and fine-tuned it to generate output according to the input and instructions.

Results

Eight immunochemistry datasets were extracted from 50 free-text pathology reports, for a total of 400 datasets for comparison. Through manual review, 225 immunochemistry data had positive or negative records. Out of a total of 400 records, including those reported as N/A by LLM, precision and recall were 0.895 and 0.766, respectively. If we analyze only the results reported as positive or negative by LLM,

precision and recall were 0.976 and 0.991, respectively. When we checked for incorrect results, many of them returned N/A for negative. There were 20 cases that returned N/A for positive cases, and 0 case that returned negative for positive. Many incorrect results were due to minor differences in notation. (ex, 'Glypican3' written as 'Glypican 3' or 'Glypican', 'EpCAM' written as 'Ep CAM'). The correct results were extracted not only as positive or negative, but with additional values. (ex, CD34: Positive with capillarization pattern, CD13: Positive with canaliculi pattern).


Result: Liver, No.7 segment, segmentectomy:
* Hepatocellular carcinoma
1. Gross type: nodular with perinodular extension
2. Size: 2.9x2.5x2.5cm
3. Differentiation: The worst differentiation: Edmondson grade III
The major differentiation: Edmondson grade III
4. Histologic type: macrotrabecular
5. Cell type: clear
6. Tumor necrosis: yes (20%)
7. Fibrous capsule: complete
8. Surgical margin invasion: no (margin of the clearance: 1cm)
9. Serosal invasion: no
10. Portal vein invasion: no
11. Bile duct invasion: no
12. Microvessel invasion: no
13. Intrahepatic metastasis: no
14. Multicentric occurrence: no

* Non-tumor liver
1. Chronic hepatitis: yes
2. Etiology: HCV
3. Grade (inflammatory activity): moderate
4. Stage(fibrosis): septal

Remarks: Immunoexpression in tumor
Glypican 3: positive
AFP: focal positive
CK19 and EpCAM: negative

Figure 1. An example of pathology report from patient

Table 1. Performance of LLM for extracting immunochemistry testing results.

	Reported (+)	Reported (-)	Reported (N/A)
True (+)	**116**	**0**	**20**
True (-)	**1**	**40**	**48**
True (N/A)	**0**	**1**	**174**

Conclusion

We extracted immunochemistry data from free text pathology reports without any preprocessing. LLM performed well even without pre-training. In particular, the accuracy of cases reported by LLM as positive or negative was very high, suggesting that it is possible to use LLM to add additional data. It is noteworthy that LLM showed its feasibility on reports with different formats. This study shows the possible application of LLM for extracting pathology data from different cancer types and different institutions.

References

1. Zhang, Y., & Zhang, Z. (2020). The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. *Cellular & molecular immunology*, *17*(8), 807–821.
2. Ryu, B., Yoon, E., Kim, S., Lee, S., Baek, H., Yi, S., Na, H. Y., Kim, J. W., Baek, R. M., Hwang, H., & Yoo, S. (2020). Transformation of Pathology Reports Into the Common Data Model With Oncology Module:

Use Case for Colon Cancer. *Journal of medical Internet research*, *22*(12), e18526.

3. Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., & Wu, Y. (2022). A large language model for electronic health records. *NPJ digital medicine*, *5*(1), 194.