[BIOINFO 2023]

Developing Prompts from Large Language Model for Extracting Clinical Information from Pathology and Ultrasound Reports in Breast Cancer

Hyeon Seok Choi, MD[1,2], Jun Yeong Song, MD[1,2], Kyung Hwan Shin, MD, PhD[1,2,3], Ji Hyun Chang, MD, PhD[1,2], Bum-Sup Jang, MD, PhD[1,2]

[1]Department of Radiation Oncology, Seoul National University Hospital, Seoul, South Korea
[2]Department of Radiation Oncology, Seoul National University College of Medicine, Seoul, South Korea
[3]Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul, South Korea

Correspondence
Bum Sup Jang, MD, PhD
Mailing address: Department of Radiation Oncology, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul, 03080, Korea
Tel: 02-2072-1161
E-mail: bigwiz83@gmail.com

We aimed to evaluate the time and cost of developing prompts using large language model (LLM), tailored to extract clinical factors in breast cancer patients and their accuracy. We collected data from reports of surgical pathology and ultrasound from breast cancer patients who underwent radiotherapy from 2020 to 2022. We extracted the information using the GPT for Sheets™ and Docs™ extension plugin and termed this the 'LLM' method. The time and cost of developing the prompts were assessed and compared with those spent on collecting information with 'Full manual' and 'LLM-assisted manual' methods. To assess accuracy, 60 patients were randomly selected, and the extracted information by 'LLM' method were compared with those collected by 'Full manual' method. Data from 2,931 patients were collected. We developed 12 prompts for Extract function and 12 for Format function to extract and standardize the information. The overall accuracy was 88.3%. For surgery type and lymphovascular invasion, it was 100%. Developing and processing the prompts took 3.5 hours and 15 minutes, respectively. Utilizing the ChatGPT API cost $65.8 and with the estimated wage, the cost was $91.7. 'LLM-assisted manual' and 'LLM' methods were time- and cost-efficient compared to the 'Full manual' method. Developing and facilitating prompts for LLM to derive clinical factors was efficient to extract crucial information from huge medical records. This study demonstrated the potential of the application of natural language processing using LLM model in breast cancer patients. Prompts from current study can be re-used for other researches to collect clinical information.