[BIOINFO 2024]

Assessing GPT4's capabilities for the synthetic EMR generation and diagnosis of rare diseases

Yoonbee Kim[1], Bumsuk Kim[1], Hyekyoung Lee[1], Hyunjin Shin[1,*]

[1] *MOGAM Institute for Biomedical Research*
**Corresponding author: hyunjinshin@mogam.re.kr*

According to the definition of the Korea Disease Control and Prevention Agency, rare diseases are defined as those with a prevalence of less than 20,000 people or those for which the prevalence is unknown due to difficulty in diagnosis. Since there are many different rare diseases that even experts have not experienced or even accessed through data, accurate diagnosis of diseases often takes a long time through multiple medical institutions. Recently, large language models (LLMs) have shown emergent abilities in various fields including biomedicine and health. LLMs can be utilized by non-experts for research purposes through simple prompt with only a very small amount of data unlike traditional machine learning techniques. In this study, we design novel prompts for GPT-4 to generate synthetic electronic medical record (EMR) data for rare diseases and to diagnose the diseases. We generate synthetic and real structured EMR data using three distinct approaches via LLMs. First, we compile 5-7 structured EMR examples for each target rare disease by searching case reports on Google Scholar. Second, we randomly select between 50% and 100% of phenotypes from the disease-phenotype relationships in the Human Phenotype Ontology (HPO) database to generate structured synthetic EMR data for each target disease. Finally, we generate synthetic EMR data using phenotypes and diagnosed diseases from the Peking Union Medical College Hospital (PUMCH) dataset in RareBench, which included EMRs from 75 rare disease patients. We evaluate GPT-4's performance in diagnosing rare diseases using these synthetic and real EMR datasets. Our results show that LLMs can help the diagnosis of rare diseases by addressing the challenge of limited real-world data through the application of LLMs.