International Journal of Population Data Science

Journal Website: www.ijpds.org

Application of EMR data-based LLM-based Disease Identification Framework to improve ICD data-based risk adjustment algorithms

Seungwon Lee[1], Elliot Martin[1], Kiarash Riazi[2], Noopur Swadas[3], Cathy Eastwood[2], Jinli Yao[4], Robin Walker[5], Danielle Southern[3], Bing Li[6], and Jeffrey Bakal[6]

[1]Provincial Research Data Services, Health Shared Services, Calgary, Canada; Centre for Health Informatics, University of Calgary, Calgary, Canada

[2]Centre for Health Informatics, University of Calgary, Calgary, Canada; Department of Community Health Sciences, University of Calgary, Calgary, Canada

[3]Centre for Health Informatics, University of Calgary, Calgary, Canada

[4]Department of Community Health Sciences, University of Calgary, Calgary, Canada; Centre for Health Informatics, University of Calgary, Calgary, Canada

[5]Department of Community Health Sciences, University of Calgary, Calgary, Canada; Applied Research & Innovation,Primary Care Alberta, Calgary, Canada

[6]Provincial Research Data Services, Health Shared Services, Calgary, Canada

Risk adjustment is essential for comparisons across populations by integrating individual health status and demographic factors to evaluate healthcare outcomes. We hypothesized that risk adjustment based on Large Language Models (LLMs) for case definitions, using electronic medical records (EMR) data, would outperform international classification of diseases (ICD) based algorithms in predicting inpatient mortality. Study Design and Methods A retrospective chart review cohort (n=10,659) consisting of randomly selected patients aged 18 years or older who were discharged from acute care settings was used. The chart review data were deterministically linked to an ICD database and an EMR database. We developed and applied an LLM-based framework (i.e., Phi-4) to EMR data for large-scale disease identification and compared it with ICD-based algorithms. To predict inpatient mortality, we calculated C-statistics using logistic regression. Among the cohort, 717 patients experienced inpatient mortality. Across all disease categories, LLM-based identification outperformed the ICD-data-based method. The ICD data-based risk adjustment algorithm achieved a C-statistic of 0.66 (95% CI 0.64 to 0.68) for in-hospital mortality, while the EMR data-based algorithm achieved a 0.76 (95% CI: 0.74 to 0.78) C-statistic. The chart review had a C-statistic of 0.71 (95% CI: 0.68 to 0.73) Effective risk adjustment for predicting health outcomes requires accurate information on patient comorbidity and demographic profiles. EMR data-based case definitions can account for disease severity (e.g., disease subtypes) and other variables (e.g., social determinants of health) that may not be readily available with historical ICD-based methods.

https://doi.org/10.23889/ijpds.v11i5.3481