Mining Social Determinants of Health for Heart Failure Patient 30-Day Readmission via Large Language Model

Mingchen SHAO[a], Youjeong KANG[a], Xiao HU[a], Hyunjung Gloria KWAK[a]
Carl YANG[a] and Jiaying LU✉ [a,1]
[a]*Emory University, Atlanta, Georgia, US*

Abstract. Heart Failure (HF) is one of the most cause of hospital readmission for older adults. While Social Determinants of Health (SDOH) play critical roles in health outcomes, they are often underrepresented in structured electronic health records (EHR) and hidden in unstructured clinical notes. This study leverages advanced large language models (LLM) to extract SDOHs from clinical text and uses logistic regression to analyze their association with HF readmissions. By identifying key SDOHs (e.g. tobacco usage, limited transportation) linked to readmission risk, this work also offers actionable insights for reducing readmissions and improving patient care.

Keywords. Heart Failure, Social Determinant of Health, Large Language Models

1. Introduction

The lifetime risk of heart failure (HF) has increased to 24% in the US, and approximately 13% of patients require readmission within 30 days following hospital discharge [1]. Such readmissions are not only associated with increased risk of mortality but also increase the burden on healthcare system. While medical factors for readmission (e.g., comorbidities) recorded in structured EHR have been widely studied, the impact of SDOH remains underexplored [2]. On the other hand, unstructured clinical notes (e.g., admission notes, discharge summaries) contain valuable SDOH-related information [3], though their free-text nature makes extraction difficult. To address this, we propose to leverage LLM, which is capable of processing and understanding domain-specific text [4], to systemically extract SDOHs from discharge notes. We further identify key SDOHs linked to readmission risk, providing actionable insights to guide interventions and reduce readmission rates.

2. Methods

This study uses the MIMIC-III database [5]. HF patients are identified by ICD-9 code, and each data sample includes two consecutive ICU admissions, yielding 3604 notes from 2065 unique patients. To mine SDOH, we input a patient's note into a LLM with

[1] This research is partially supported by NIH NIDDK: K25DK135913 and NHLBI: K23HL148545.

SDOH-specific prompts to extract SDOHs that can be either spans from or paraphrases of the original text. We use zero-shot prompting with an open source LLM, Llama-3.1-8B, to protect the privacy of patient health information. The prompt strategy and format are manually optimized to achieve best performance. The prompt template is “Can you extract the patient’s [SDOH] from the given discharge note? Please choose from candidate list, or ‘unspecified’ if not mentioned in the note”. Overall, we cover two categories of SDOHs: (1) charted in EHR: *Gender, Age, Ethnicity, Language, Marital Status, Insurance*; (2) non-charted: *Alcohol, Tobacco, Drug, Transportation, Housing, Parental, Employment, Social Support*. After LLM-based mining, we apply logistic regression to identify the key SDOHs contributing to readmission.

3. Results

We evaluate the accuracy of the charted SDOH attributes against the corresponding EHR data and the uncharted SDOH attributes against human annotated dataset [2, 4]. For categorical attributes, we use accuracy as the metric, while for the continuous variable “Age”, we use the mean absolute error (MAE). The mining approach demonstrates strong overall performance, particularly for patient gender (99.80%), marital status (70.21%), alcohol (73.02%), and tobacco use (81.15%), as well as a low MAE for age (3.08). Logistic regression identifies several significant factors associated with higher readmission risk, including older age, unspecified ethnicity, Medicare insurance, past tobacco use, transportation barriers, and limited social support, which align with known clinical risk factors.

4. Conclusion

Our training-free LLM-based approach enables SDOH mining from unstructured clinical notes. The effectiveness of our mining method is justified by the great performance over annotated data. In the future, we plan to (1) extend our LLM-based mining approach to cover more aspects of SDOHs from various types of clinical notes, and (2) explore advanced causal relationship mining to identify novel SDOH that directly contributes to HF 30-day rehospitalization. Observing the limited numbers of notes containing certain SDOHs, we advocate for clinicians to document these factors more consistently to enhance HF care intervention after hospital discharge.

References

1. Bozkurt B, Ahmad T, Alexander K, Baker WL, Bosak K, Breathett K, et al. HF STATS 2024: Heart Failure Epidemiology and Outcomes Statistics An Updated 2024 Report from the Heart Failure Society of America. J Card Fail. 2025;31(1):66-116.
2. Ahsan H, Ohnuki E, Mitra A, Yu H. MIMIC-SBDH: A Dataset for Social and Behavioral Determinants of Health. Proc Mach Learn Res. 2021;149:391-413.
3. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. J Am Med Inform Assoc. 2020;27(11):1764-73.
4. Guevara M, Chen S, Thomas S, Chaunzwa TL, Franco I, Kann BH, et al. Large language models to identify social determinants of health in electronic health records. NPJ Digit Med. 2024;7(1):6.
5. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.