SA13

Approaches to Electronic Health Records Notes Selection: Considerations for Best Practices

Bonnie Bui, Rifky Tkatch

Optum, Eden Prairie, MN, USA

INTRODUCTION

- Electronic health records (EHR) are a unique data source that provide the ability to analyze real world patient data to obtain insight into the patient journey.
- Natural language processing (NLP) is a method of EHR notes analysis that transforms unstructured narrative data to structured data, allowing researchers the ability to systematically analyze provider notes on a large scale.
- EHR notes are unstructured, heterogeneous, and idiosyncratic, with some notes rich in content and others sparse.
- One of the challenges prior to the analysis of EHR notes is optimal note selection that provide the necessary data. However, NLP methods are largely in the domain of clinical concepts extraction.

OBJECTIVE

The objective of this targeted literature review is to examine studies on NLP use in this area to provide recommendations on which methods require advances to aid in optimal note selection.

METHODS

- A targeted literature review was conducted to identify current NLP methodology for clinical concepts extraction in EHR notes. Searches were performed in April 2024 in the PubMed database, limited to human studies published in English.
- Articles were reviewed and themes studied and categorized, with a focus on current techniques and existing challenges that must be addressed to realize the full potential of unstructured notes.

RESULTS

- Various searches were applied, with the goal of finding articles specifically addressing extracting information or concepts from unstructured notes. The final search included the following terms:
  - (“electronic health record” OR “electronic health records” OR “electronic medical record” OR “electronic medical records” OR “electronic patient record” OR “electronic patient records” OR “EHR” OR “EMR” OR “EPR”) AND (“concept extraction” OR “information extraction”) AND (“natural language processing” OR “NLP") AND (“notes selection”) NOT (Review[Publication Type])
- The searches conducted in PubMed identified 18 records for review, with 17 with free full-text.
- Table 1 presents use cases and related challenges identified from this targeted literature review.

RESULTS (cont.)

Table 1. Use Cases and Challenges to EHR Notes Selection and Data Extraction Using NLP

Use Cases	Challenges
Clinical knowledge acquisition and information extraction[8,15]	• Complex, massive, heterogeneous, and noisy[1,7,8] • Extracting relevant clinical information due to unstructured nature of notes[1,5] • Relation and pattern extraction[12] • Capturing semantic relationships (relationship between a verb and its arguments)[11,15,16,17] • Capturing temporal relations[11] • Capturing social determinants of health (SDoH) in EHRs[13] • Capturing social risk factors[2] • Extracting semi-structured text in medical notes[6]
Supplementing codified data with unstructured data[7]	• Combining codified data and unstructured data efficiently[7] • Quantifying the presence, absence, and strength of relationships between different features[7]
Automated note de-identification[4,9]	• Information altered due to overlap between clinical information and PHI information[9]
Measure relationship between clinical features, capture drug side effects and symptoms, and disease phenotyping[7]	• Dependent on quality of training data[7]
Patient identification on the basis of EHR information[10,11,12]	• Moderate accuracy[9] • Compare against ICD code algorithm detection[12]
Rank ordering medical terms for patient comprehension[3]	• Does not need a large number of training examples, but does require a training dataset with a rich set of learning features[3]
Using topic models to classify text[14]	• Selecting the most effective topic model is a non-trivial task[14]

CONCLUSIONS

- The literature on optimal note selection strategy methods that exists is inadequate, lacks clarity, and requires much needed precision. The focus is primarily on use cases of NLP and challenges in information extraction rather than note selection strategies.
- Improvements in EHR note search and selection strategies are needed. Manually reviewing EHRs as a note selection strategy can be time-consuming and inefficient.
- NLP screening allows for large scale note selection using keywords. However, optimal note selection methodology beyond keyword searches is often required to assess treatment over time or to understand sequence of events, especially when notes lack the sufficient detail.
- Current gaps in practice are:
  - (1) limited recognition of relationships between clinical concepts (such as treatment and outcome relationships)
  - (2) difficulties in extraction of temporal information to understand timing of clinical events and/or disease progression
- The complexity, diversity, and noisiness of unstructured note data make efficient extraction of relevant clinical information difficult.
- Areas of opportunities for advances in methodology include focusing on semantic relationships and word embeddings to allow for the extraction of relations and patterns in clinical notes.

SUMMARY

EHR notes can generate valuable real-world data. However, there are gaps in current methods in note selection processes. More research focusing on scalable note selection strategies beyond the challenges of information extraction is needed but made difficult by the noisiness of notes (i.e., heterogeneity in content, verbiage, and formatting; redundancies and duplicated clinical narratives across notes of multiple visits for a single patient). Future research should focus on strategies for parsing out redundancies and irrelevant details and better capturing temporal information to gain insight on relations and patterns of clinical features and symptoms.

Presented at ISPOR 2024, Atlanta, GA, May 5-8, 2024

REFERENCES

1. Assale M, Dui LG, Cina A, Seveso A, Cabitza F. The Revival of the Notes Field: Leveraging the Unstructured Content in Electronic Health Records. *Front Med (Lausanne)*. 2019;6:66. doi:10.3389/fmed.2019.00066
2. Brown JR, Ricket IM, Reeves RM, et al. Information Extraction From Electronic Health Records to Predict Readmission Following Acute Myocardial Infarction: Does Natural Language Processing Using Clinical Notes Improve Prediction of Readmission? *J Am Heart Assoc*. 2022;11(7):e024198. doi:10.1161/JAHA.121.024198
3. Chen J, Jagannatha AN, Fodeh SJ, Yu H. Ranking Medical Terms to Support Expansion of Lay Language Resources for Patient Comprehension of Electronic Health Record Notes: Adapted Distant Supervision Approach. *JMIR Med Inform*. 2017;5(4):e42. doi:10.2196/medinform.8531
4. Deleger L, Molnar K, Savova G, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *J Am Med Inform Assoc*. 2013;20(1):84-94. doi:10.1136/amiajnl-2012-001012
5. Fang A, Hu J, Zhao W, et al. Extracting clinical named entity for pituitary adenomas from Chinese electronic medical records. *BMC Med Inform Decis Mak*. 2022;22(1):72. doi:10.1186/s12911-022-01810-z
6. Finch DK, McCart JA, Luther SL. TagLine: Information Extraction for Semi-Structured Text in Medical Progress Notes. *AMIA Annu Symp Proc*. 2014;2014:534-543.
7. Gan Z, Zhou D, Rush E, et al. ARCH: Large-scale Knowledge Graph via Aggregated Narrative Codified Health Records Analysis. *medRxiv*. Published online May 21, 2023:2023.05.14.23289955. doi:10.1101/2023.05.14.23289955
8. Liu S, Wang L, Ihrke D, et al. Correlating Lab Test Results in Clinical Notes with Structured Lab Data: A Case Study in HbA1c and Glucose. *AMIA Jt Summits Transl Sci Proc*. 2017;2017:221-228.
9. Meystre SM, Ferrández Ó, Friedlin FJ, South BR, Shen S, Samore MH. Text de-identification for privacy protection: A study of its impact on clinical text information content. *Journal of Biomedical Informatics*. 2014;50:142-150. doi:10.1016/j.jbi.2014.01.011
10. Meystre SM, Heider PM, Cates A, et al. Piloting an automated clinical trial eligibility surveillance and provider alert system based on artificial intelligence and standard data models. *BMC Med Res Methodol*. 2023;23(1):88. doi:10.1186/s12874-023-01916-6
11. Ni Y, Kennebeck S, Dexheimer JW, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. *J Am Med Inform Assoc*. 2015;22(1):166-178. doi:10.1136/amiajnl-2014-002887
12. Pan J, Zhang Z, Peters SR, et al. Cerebrovascular disease case identification in inpatient electronic medical record data using natural language processing. *Brain Inform*. 2023;10(1):22. doi:10.1186/s40708-023-00203-w
13. Patra BG, Sharma MM, Vekaria V, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc*. 2021;28(12):2716-2727. doi:10.1093/jamia/ocab170
14. Rijcken E, Kaymak U, Scheepers F, Mosteiro P, Zervanou K, Spruit M. Topic Modeling for Interpretable Text Classification From EHRs. *Front Big Data*. 2022;5:846930. doi:10.3389/fdata.2022.846930
15. Wang Y, Pakhomov S, Melton GB. Predicate Argument Structure Frames for Modeling Information in Operative Notes. *Stud Health Technol Inform*. 2013;192:783-787.
16. Wang Y, Liu S, Afzal N, et al. A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform*. 2018;87:12-20. doi:10.1016/j.jbi.2018.09.008
17. Wu H, Toti G, Morley KI, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J Am Med Inform Assoc*. 2018;25(5):530-537. doi:10.1093/jamia/ocx160

Disclosure: This study was exclusively for research purposes and all authors declare no conflict of interest.

Corresponding author: Bonnie Bui (bonnie.bui@optum.com)

Optum