Leveraging Natural Language Processing Tool to Identify Eligible Lung Cancer Screening Patients in the Electronic Health Record

Siru Liu, PhD[1], Allison B. McCoy, PhD[1], Bryan Steitz, PhD[1], Adam Wright, PhD[1]
[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

Introduction

Lung cancer is the leading cause of cancer mortality in the US, accounting for 25% of all cancer deaths.[1] Lung cancer screening (LCS) has been shown to reduce mortality rates by 20%[2], but only 5% of eligible adults have received LCS nationally.[3] Several gaps exist in the implementation of LCS, particularly in the critical first step of identifying eligible patients. The eligibility criteria rely on smoking history, which is inconsistently documented in electronic health records (EHRs).[4] Structured documentation tends to be incomplete, but additional information may be present in clinical notes. These shortcoming make it hard for clinicians, and especially for clinical decision support (CDS) tools, to recognize patients' full smoking history.[5] Moreover, the 2021 United States Preventive Services Task Force (USPSTF) LCS guideline nearly doubles the number of individuals eligible for LCS to 14.5 million Americans,[6] so efficient strategies for identifying and referring these patients are needed. In this study, we sought to develop a natural language processing (NLP) pipeline to systematically extract smoking information from clinical notes to identify patients eligible for LCS.

Methods

For patients aged 50-80 years who had been seen at least once in the past 3 years in a primary care clinic at Vanderbilt University Medical Center (VUMC), we extracted documented smoking information from our Epic EHR. Structured data included packs-per-day, years-smoked, smoking status, and quit time. We additionally extracted all clinical notes written during outpatient encounters in 2021 for each patient in our cohort. We applied an NLP tool, consisting of a two-layer rule engine, to extract pack-years, packs-per-day, years-smoked, and quit time from the unstructured text in the notes.[7] We tested and optimized the NLP tool's performance on a corpus of VUMC notes for 5000 patients randomly selected from our study cohort. We determined LCS eligibility according to the 2021 USPSTF guideline: adults aged 50 to 80 years with a 20-pack-year smoking history who currently smoke or quit smoking within the last 15 years.[4] We combined the data extracted from clinical notes with structured data from the EHR to consider whether patients met the criteria. We used the maximum value of extracted pack-years from notes as the patient's pack-year. For patients without any pack-years reported in the clinical notes, we used the extracted packs-per-day and years-smoked to fill in missing or zero values in the structured dataset and then calculated pack-years. For missing values of quit time in the structured dataset, we also used the extracted values from the notes. To assess our performance, we compared our NLP approach with Epic's LCS eligibility calculation, which uses the most recent values of packs-per-day and years-smoked stored in structured data to calculate pack-years. We compared patient cohorts identified using each approach and the respective calculated values. We applied Mann-Whitney U tests and Chi-square tests for numerical variables and categorical variables, respectively. To validate the NLP-based approach, we randomly selected 50 newly identified patients and performed a manual chart review. Statistical analyses were performed in R.

Results

We included 102,475 patients were in the final dataset. The mean age was 63.9±8.4 years, 9,780 (9.5%) patients were Black/African American, and 57,986 (56.6%) patients were female; 11,500 (11.2%) were current smokers and 29,451 (28.7%) were in a quit status. Of the 40,951 patients with a history of smoking, 23,941 (58.5%) patients had insufficient smoking data that could be used to assess LCS eligibility using the baseline approach. After considering all structured data, there are still 23,210 (56.7%) of current/past tobacco users without sufficient information to determine their LCS eligibility.

While refining the NLP tool, we identified 63 new patients potentially eligible for LCS. Through manual chart review, we identified 17 individuals for whom pack-years were incorrectly extracted, with some of the extracted information coming from descriptions of the LCS guidelines in the clinical notes (e.g., LCS may be recommended if you are a heavy smoker currently or quit less than 15 years ago and have smoked 30 pack years, talk to your doctor about this test). Therefore, we added rules to exclude LCS guideline descriptions, and the new tool achieved an overall F1 of 0.979 on the testing dataset (i.e., 50 clinical notes).

The baseline approach was able to identify 5,887 patients eligible for the LCS. Using the improved algorithm on the structured data could identify 7,194 patients with LCS eligibility, with an increment of 22.2%. After adding NLP

extracted smoking information from clinical notes in 1 year and 3 years, the number of identified LCS eligible patients were 8,931 (51.7% increment) and 10,231 (73.8%), respectively. Demographic and smoking information for the 5,887 current identified patients and the 4,344 new patients are listed in Table 1. The NLP-based approach identified 589 new Black/African Americans, a significant increase of 119% compared to the baseline approach. In the validation set of 50 newly identified patients, the accuracy rate was 81.2% (41/50). Inconsistencies between the extracted pack-year and structured data are presented in Table 2.

Table 1. Demographic and smoking information of patients eligible for lung cancer screening (*. $P<0.001$).

	Patients identified using structured data only	Newly identified patients using the NLP-based approach
Number of patients	5,887	4,344
Age (mean, std)	64.3 (7.5)	63.6 (7.7)*
Black/African American	495 (8.4%)	589 (13.6%)*
Female	2,701 (45.9%)	2,072 (47.7%)
Private Insurance	2,265 (38.5%)	1,673 (38.5%)
Pack-year (mean, std)	43.8 (26.5)	53.0 (88.0)*
Pack-per-day (mean, std)	1.2 (0.9)	1.6 (2.5)*
Years-smoked	38.6 (13.5)	34.4 (12.8)*
Current Smoker (%)	2,888(49.1%)	2,039 (46.9%)*
Former Smoker (%)	2,999 (50.9%)	2,311 (53.2%)*
Quit Years (mean, std)	6.4 (4.9)	4.6 (4.7)*

Table 2. Inconsistencies between the extracted pack-year and structured data in the validation set (n=50).

NLP Wrong	9 (18%)
Extract wrong value	1 (2%)
LCS guideline	8 (16%)
NLP Correct	41 (81.2%)
No recent years-smoked or packs-per-day stored in the structured dataset	13 (26%)
Decrease in years-smoked	4 (8%)
Decrease in packs-per-day	24 (48%)

Discussion

We developed and tested an NLP-based approach to extract smoking information from clinical notes, which we combined with structured data. This approach can effectively identify patients eligible for LCS in EHRs, supplementing the patient population identified using the baseline approach by 73.8%. The inconsistencies of smoking information shown between the structured dataset and clinical notes are in line with previous research.[8] In addition, the NLP-based approach identified 119% more Black/African Americans, which has a great potential to diminish disparities in the LCS, highlighting the importance of integrating smoking information from the clinical notes into the LCS CDS tools. Previous research has reported that Black/African American are less likely to use self-reported tools[9] (e.g. patient portals) and are more concerned about security/privacy,[10] potentially contributing to the lack of smoking information in structured dataset. In future work, we will develop a workflow-embedded CDS tool for primary care providers that will include an alert and a patient list within the EHR. Overall, we presented a feasible NLP-based approach to identify LCS eligible patients in the EHR. This work provides a solid technical basis for the development of a CDS tool and further implementation into clinical practice to efficiently improve the utilization of LCS and diminish healthcare disparities in LCS.

Acknowledgements: This work was supported by NIH grant: R01AG062499-01.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin*. 2022;72(1):7-33.
2. Nanavaty P, Alvarez MS, Alberts WM. Lung Cancer Screening: Advantages, Controversies, and Applications. *Cancer Control*. 2014;21(1):9-14. doi:10.1177/107327481402100102
3. Fedewa SA, Kazerooni EA, Studts JL, et al. State Variation in Low-Dose Computed Tomography Scanning for Lung Cancer Screening in the United States. *JNCI J Natl Cancer Inst*. 2021;113(8):1044-1052. doi:10.1093/jnci/djaa170
4. US Preventive Services Task Force. Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA*. 2021;325(10):962-970. doi:10.1001/jama.2021.1117
5. Polubriaginof F, Salmasian H, Albert DA, Vawdrey DK. Challenges with Collecting Smoking Status in Electronic Health Records. *AMIA . Annu Symp proceedings AMIA Symp*. 2017;2017:1392-1400. /pmc/articles/PMC5977725/
6. Landy R, Young CD, Skarzynski M, et al. Using Prediction Models to Reduce Persistent Racial and Ethnic Disparities in the Draft 2020 USPSTF Lung Cancer Screening Guidelines. 2021;113(11). /pmc/articles/PMC8562965/
7. Yang X, Yang H, Lyu T, et al. A Natural Language Processing Tool to Extract Quantitative Smoking Status from Clinical Narratives. In: *2020 IEEE International Conference on Healthcare Informatics (ICHI)*. Vol 2020. IEEE; 2020:1-2.
8. Kukhareva P V, Caverly TJ, Li H, et al. Inaccuracies in electronic health records smoking data and a potential approach to address resulting underestimation in determining lung cancer screening eligibility. *J Am Med Informatics Assoc*. 2022;2022(0):1-10. doi:10.1093/jamia/ocac020
9. Turner K, Hong Y-R, Yadav S, Huo J, Mainous AG. Patient portal utilization: before and after stage 2 electronic health record meaningful use. *J Am Med Informatics Assoc*. 2019;26(10):960-967. doi:10.1093/jamia/ocz030
10. Lyles CR, Allen JY, Poole D, Tieu L, Kanter MH, Garrido T. "I Want to Keep the Personal Relationship With My Doctor": Understanding Barriers to Portal Use among African Americans and Latinos. *J Med Internet Res*. 2016;18(10):e263.