Published in final edited form as:

Int J Med Inform. 2023 September ; 177: 105136. doi:10.1016/j.ijmedinf.2023.105136.

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

Siru Liu, PhD[1], **Allison B. McCoy, PhD**[1], **Melinda C. Aldrich, PhD, MPH**[1,2,3], **Kim L. Sandler, MD**[4], **Thomas J. Reese, PharmD, PhD**[1], **Bryan Steitz, PhD**[1], **Jiang Bian, PhD**[5], **Yonghui Wu, PhD**[5], **Elise Russo, MPH**[1], **Adam Wright, PhD**[1]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN;

[2]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN;

[3]Department of Thoracic Surgery, Vanderbilt University Medical Center, Nashville, TN;

[4]Department of Radiology and Radiological Sciences, Vanderbilt University Medical Center, Nashville, TN;

[5]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL

Abstract

Objective: To develop and validate an approach that identifies patients eligible for lung cancer screening (LCS) by combining structured and unstructured smoking data from the electronic health record (EHR).

Methods: We identified patients aged 50–80 years who had at least one encounter in a primary care clinic at Vanderbilt University Medical Center (VUMC) between 2019 and 2022. We fine-tuned an existing natural language processing (NLP) tool to extract quantitative smoking information using clinical notes collected from VUMC. Then, we developed an approach to identify patients who are eligible for LCS by combining smoking information from structured data and clinical narratives. We compared this method with two approaches to identify LCS eligibility only using smoking information from structured EHR. We used 50 patients with a documented history of tobacco use for comparison and validation.

Corresponding Author: Siru Liu, PhD, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Ave #1475, Nashville, TN, 37212, Phone: 615-875-5216, siru.liu@vumc.org.

Author Statement

SL conducted data extraction, algorithm developing, statistical analysis, and drafting the work. AW, AM, MA, KS, TR, BS helped to design experiments. AW, AM, MA, KS, TR, BS, JB, YW, ER revised the drafted manuscript. All authors approved the submitted version.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

COMPETING INTERESTS STATEMENT

The authors do not have conflicts of interest related to this study.

Results: 102,475 patients were included. The NLP-based approach achieved an F1-score of 0.909, and accuracy of 0.96. The baseline approach could identify 5,887 patients. Compared to the baseline approach, the number of identified patients using all structured data and the NLP-based algorithm was 7,194 (22.2%) and 10,231 (73.8%), respectively. The NLP-based approach identified 589 Black/African Americans, a significant increase of 119%.

Conclusion: We present a feasible NLP-based approach to identify LCS eligible patients. It provides a technical basis for the development of clinical decision support tools to potentially improve the utilization of LCS and diminish healthcare disparities.

Keywords

Lung cancer screening; natural language processing; electronic health record

BACKGROUND AND SIGNIFICANCE

Lung cancer is the leading cause of cancer mortality, accounting for 25% of all cancer deaths in the United States.[1] An estimated 236,740 new cases and 130,180 deaths from lung cancer will occur in 2022.[1] Lung cancer screening (LCS) is an effective approach to reducing lung cancer morbidity and mortality. The National Lung Screening Trial reported that LCS with annual low-dose computed tomography in the eligible population reduces mortality by 20% relative to chest radiography,[2] resulting in approximately 20,000 fewer deaths per year.[3] Since the United States Preventive Services Task Force (USPSTF) first recommended LCS in 2013, there has been a significant increase in the incidence of lung cancer detected in the early, localized stage and a reduction in advanced-stage detection.[1]

Rates of LCS among eligible adults are low −5% nationally, and even lower among minorities.[4][5] Recent changes to LCS guidelines increased the number unscreened people from millions to tens of millions in the U.S. In 2021, the USPSTF expanded eligibility for people who currently smoke or quit within 15 years from adults aged 55 to 80 with a 30 pack-year smoking history to those aged 50 to 80 with a 20 pack-year smoking history.[6] This nearly doubled the number of people eligible for LCS to 14.5 million Americans.[3][7] Based on current low LCS rates, approximately 13.7 million Americans who should be screened for lung cancer are not being screened.

Efficient identification of patients who are eligible for LCS from electronic health record (EHR) data and assigning them to primary care physicians in a timely manner may be the first step to improve the situation.[8] A screenshot of a lung cancer screening alert in the EHR system (Epic) at Vanderbilt University Medical Center (VUMC) is shown in Figure 1. Most studies addressing eligibility for LCS use either structured smoking data or smoking data from clinical notes.[9,10] However, eligibility for LCS is based on several quantitative data points, including current smoking status, pack-years, and quit date for former tobacco users. In addition to structured data, smoking history information is often embedded in clinical notes as free text. A previous study analyzing 48,909 inpatient records in 2016 found that more than 98% of patients had smoking information recorded in their clinical notes, with 67.1% of them having more than one note with smoking information.[11] Both structured and unstructured EHR data are essential sources for extracting smoking information. Effective natural language processing (NLP) tools has been widely applied to extract smoking information from clinical notes [12]; for example, a recent study reported a F1 score of 0.96.[13] Furthermore, smoking information typically exists in multiple clinical notes as well as in structured forms and is not standardized, making it difficult for clinical decision support (CDS) tools and clinicians to recognize smoking history and whether the information available in the EHR is accurate.[11] The objective of this study was to develop an approach that automatically combining smoking information from clinic notes with structured smoking data, and evaluate its performance in assessing the eligibility for LCS. This approach has the potential to streamline the identification of LCS-eligible patients and serve as a technological foundation for the future development of a clinical decision support tool.

METHODS

Data Collection

Our study sample included patients aged 50–80 years who had been seen at least once in the past 3 years (July 1, 2019 to July 1, 2022) in a primary care clinic at VUMC. The query date was July 3, 2022. We extracted documented smoking information from VUMC's Epic system for each patient. We extracted the structured data and clinic notes from associated tables in Clarity, which is a relational database for data in the Epic system. Structured data included packs-per-day, years-smoked, smoking status, and quit time. We additionally extracted all clinical notes written about the patients in our cohort during their outpatient and inpatient encounters throughout the study period.

We applied a validated NLP tool, consisting of a two-layer rule engine, to extract smoking information (Appendix 1) from the clinical notes.[15] The NLP tool was originally developed using 200 clinical notes from a cohort of 3,080 patients who received LCS between 2012 and 2019 at the University of Florida Health system in its enterprise data warehouse—the UF Health Integrated Data Repository (IDR). The algorithm first identifies a set of predefined lexicons using regular expressions, which were later used to define high-level rules in its second layer to extract packs-per-day, years-smoked, pack-years, and quit years with a reported overall F1 of 0.963. We randomly sampled 450 clinical notes at VUMC that contained “smoke” or “tobacco” related keywords and manually labeled smoking information. Of these, we used 400 clinical notes to fine-tune the NLP tool. We developed new rules through error analysis and combined them into the NLP tool. We then evaluated the customized NLP tool using a separate random sample of 50 clinical notes from different patients.

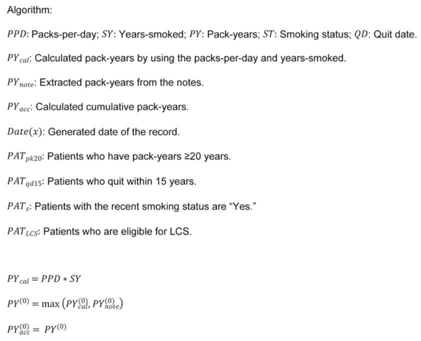
LCS Eligibility Assessment

LCS eligibility was determined 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.[6] We established baseline LCS eligibility using the most recent values of packs-per-day, years-smoked, and quit date stored in structured data. This baseline approach is natively applied in the Epic EHR system and many other EHR vendors.[9] To improve the accuracy and missing values in the most recent structured data, we developed a new algorithm to combine structured data with NLP extracted smoking information from notes to assess the LCS eligibility.

The updated algorithm includes two main parts: pack-years calculation and years-quit calculation. To calculate pack-years, we merged the NLP extracted packs-per-day and years-smoked with the structured data. For each patient, we calculated the pack-years by multiplying packs-per-day and years-smoked generated in the same encounter. We selected the maximum value of the same variable in one encounter. We then combined the extracted pack-years with the calculated pack-years. In addition, we applied a longitudinal approach to calculate cumulative pack-years.[9] The cumulative pack-years was calculated based on the time duration between two records and packs-per-day. For patients with multiple pack-years in the same encounter, we selected the maximum value. If a patient had any record with pack-years ≥20, we placed them with the ≥20 pack-year cohort.

To calculate years-quit, we first converted NLP extracted years-quit (e.g., years since quitting: 35 years) and months-quit (e.g., quit smoking: 3 months) into quit date format using the date of data generation minus the extracted value. Then the converted quit dates were combined with NLP extracted and structured quit dates. In addition, if the quit date was missing, the recorded date for “quit smoking” status for quit smoking transition records (i.e., the smoking status changed from “Yes/Never” to “Quit”) would be used. For each patient, we used the most recent quit date and search query date to calculate years-quit. Finally, we selected patients with years-quit≤15 years.

Next, to determine current or former tobacco use, we used the most recent smoking status (“Yes” or “Quit”) stored in the structured dataset. Thus, the final LCS eligible patients are 1) patients who have pack-years≥20 and the most recent smoking status is “Yes”, and 2) patients who have pack-years≥20, the most recent smoking status is “No” and the years-quit≤15 years.



Algorithm:

PPD: Packs-per-day; SY: Years-smoked; PY: Pack-years; ST: Smoking status; QD: Quit date.

PY_{cal}: Calculated pack-years by using the packs-per-day and years-smoked.

PY_{note}: Extracted pack-years from the notes.

PY_{acc}: Calculated cumulative pack-years.

$Date(x)$: Generated date of the record.

PAT_{pk20}: Patients who have pack-years ≥20 years.

PAT_{qd15}: Patients who quit within 15 years.

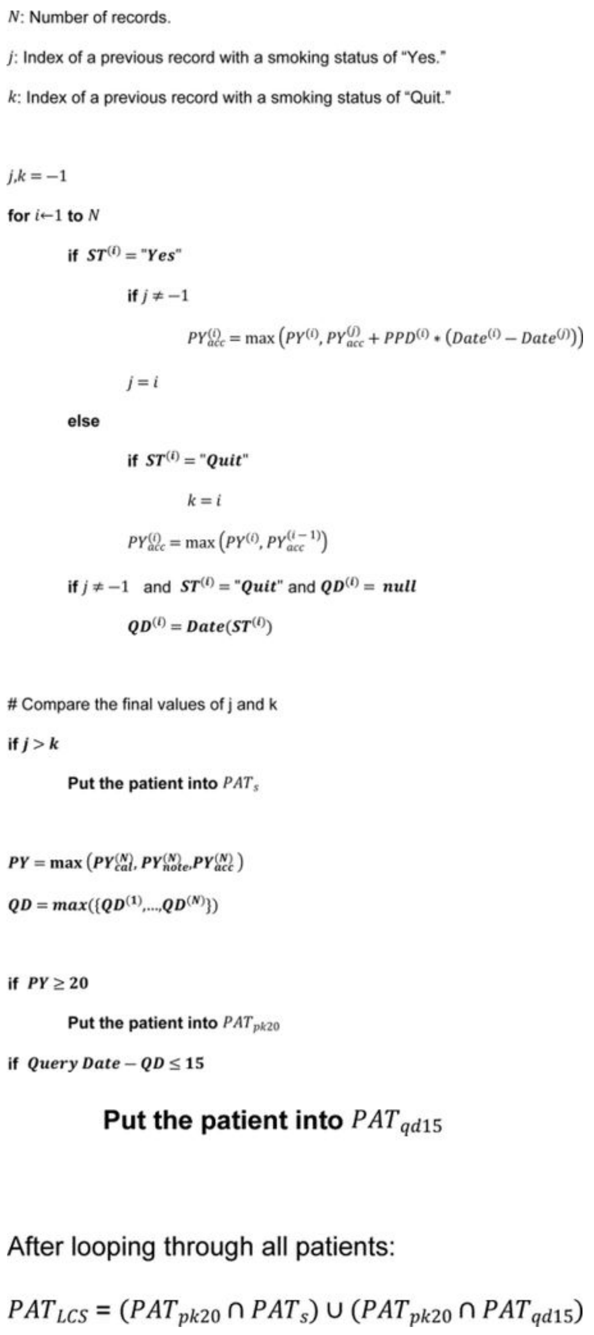
PAT_{s}: Patients with the recent smoking status are “Yes.”

PAT_{LCS}: Patients who are eligible for LCS.

$$PY_{cal} = PPD * SY$$

$$PY^{(0)} = \max\left(PY_{cal}^{(0)}, PY_{note}^{(0)}\right)$$

$$PY_{acc}^{(0)} = PY^{(0)}$$



```
$N$: Number of records.
$j$: Index of a previous record with a smoking status of "Yes."
$k$: Index of a previous record with a smoking status of "Quit."

$j,k = -1$
for $i \leftarrow 1$ to $N$
    if $ST^{(i)} = "Yes"$
        if $j \neq -1$
            $PY_{acc}^{(i)} = \max\left(PY^{(i)}, PY_{acc}^{(j)} + PPD^{(i)} * \left(Date^{(i)} - Date^{(j)}\right)\right)$
        $j = i$
    else
        if $ST^{(i)} = "Quit"$
            $k = i$
        $PY_{acc}^{(i)} = \max\left(PY^{(i)}, PY_{acc}^{(i-1)}\right)$
    if $j \neq -1$ and $ST^{(i)} = "Quit"$ and $QD^{(i)} = null$
        $QD^{(i)} = Date(ST^{(i)})$

# Compare the final values of j and k
if $j > k$
    Put the patient into $PAT_s$

$PY = \max\left(PY_{cal}^{(N)}, PY_{note}^{(N)}, PY_{acc}^{(N)}\right)$
$QD = max(\{QD^{(1)}, \ldots, QD^{(N)}\})$

if $PY \geq 20$
    Put the patient into $PAT_{pk20}$
if $Query\ Date - QD \leq 15$
    Put the patient into $PAT_{qd15}$

After looping through all patients:
$PAT_{LCS} = (PAT_{pk20} \cap PAT_s) \cup (PAT_{pk20} \cap PAT_{qd15})$
```

Evaluation

To validate our NLP-based approach, we randomly selected 50 patients who were current or former tobacco users based on structured data from the patient cohort and manually evaluated their LCS eligibility using smoking related information from Epic EHR. We then applied this NLP-based approach, as well as two previous approaches: the most recent structured data approach (baseline) and the all structured data approach, to our patient cohort. We compared the patients identified by each approach and the relevant smoking information. We applied Mann-Whitney *U* tests and Chi-square tests for numerical variables and categorical variables, respectively. Statistical analyses were performed in R.

RESULTS

We included 102,475 patients 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. We extracted 1,914,701 records with structured smoking data (i.e., packs-per-day, years-smoked, pack-years, smoking status, and quit date). The average number of structured smoking data records for each patient was 19.5; 40,951 patients had a history of smoking, 11,500(11.2%) were currently using tobacco, and 29,451(28.7%) previously smoked cigarettes. These numbers are likely lower than reality due to the scarcity of smoking information stored in the structured data. The total number of clinical notes was 4,521,645 (August 1, 2019 to August 1,2022) from 1,303,592 office visits and hospital encounters. Detailed information about the patient cohort is listed in Table 1.

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. The issues identified in the structured dataset are detailed in Table 2.

The original NLP tool had a reported F1 of 0.946.[15] In our first dataset (i.e., 400 notes), the overall F1 was 0.959. After using the 400 notes to optimize the NLP tool, the new tool achieved an overall F1 of 0.979 on the testing dataset (i.e., 50 clinical notes). The results for each extracted variable were listed in Appendix 2. The lexicons and rules in the updated two-layer rule-engine were listed in Appendix 3. Using the optimized NLP tool, we extracted 1,386,786 records, and after removing duplicates, we included 585,269 records generated from 281,623 clinical notes for 226,204 visits. The NLP-based approach achieved a recall of 0.833, precision of 1, F1-score of 0.909, and accuracy of 0.96. 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% increment), respectively. The number of current and former tobacco users in identified patients were listed in Table 3.

Demographic and smoking information for the 5,887 current identified patients and the 4,344 new patients are listed in Table 4. The NLP-based approach identified 589 new Black/African Americans (average age: 63.7), a significant increase of 119% compared to the baseline approach. The newly identified patients had significantly higher pack-year of 53 and pack-per-day of 1.6. In addition, their average quit years were significantly lower, at 4.6 years.

DISCUSSION

In this study, 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%. Additionally, it identifies a greater proportion of Black/African American patients and more young patients who meet screening guidelines, which may help reduce health disparities in LCS.

Our approach comprehensively analyzed historical smoking information from both structured data and up to 3 years of clinical notes. While previous studies have focused on using historical structured data, our study quantifies the impact of combining smoking information from clinical notes on assessing LCS eligibility and compares it to algorithms that use structured data only.[9] The results indicate that free text smoking information in the EHR plays an important role in assessing eligibility for the LCS. Compared with the baseline approach, our NLP-based approach identified 119% more Black/African Americans who meet screening guidelines. Similarly, this result highlights the importance of integrating smoking information from the clinical notes into LCS CDS tools. Previous research has reported that Black/African Americans are less likely to use self-reporting tools[16] (e.g. patient portals) and are more concerned about security/privacy,[17] potentially contributing to the lack of smoking information in the structured dataset. In addition, the relevant increment for the NLP-based approach in the Hispanic group and females were higher than the average increment, which indicates the NLP-based approach may be more protective of underrepresented groups.

This study has several limitations. First, we developed and tested the NLP-based approach in one healthcare system. However, the NLP tool was originally developed and evaluated using data from the University of Florida. We validated and fine-tuned it using data from VUMC. Second, we used age and smoking history to identify patients who might be eligible for LCS. In practice, clinicians also need to consider patients' current and/or historical diseases and other complex issues. One purpose of our study was to identify more patients who are eligible for LCS. From there, clinicians can provide a shared decision-making process with potential eligible patients to discuss the benefits and harms of LCS. Third, recall bias may exist in the smoking data. Researchers have identified that patients might not be able to accurately recall smoking data over their lifetime.[18] On the other hand, this underscores the urgent need to implement the NLP-based approach into the EHR system in order to notify clinicians of the inaccuracies in the smoking data. Fourth, including all clinical notes might enable identification of a few more patients. However, the demands for computational resources would also increase, while the potential increment might not be significant. After reviewing the smoking information stored in our health care system, we decided to use up to 3 years of clinical notes. Researchers can adjust the number of years used based on their own needs and EHR data. Lastly, a limitation of our algorithm is that we assumed that both the clinical notes and structured data were free from recording errors. However, in cases where a patient has recording errors indicating a pack-year greater than 20, there is a possibility that they might be falsely identified as an LCS eligible patient.

Future work could include implementing an NLP pipeline within the EHR: 1) an optimal state-of-art NLP model to extract smoking information from notes and 2) a local system for running the model in real-time in the EHR. The fine-tuned NLP model will be used to extract quantitative smoking information on all clinical notes and structured smoking data in the EHR database. Then, the NLP pipeline will be connected to our Epic EHR using an event-driven approach. Epic would push new notes to the NLP pipeline, which would extract smoking information, where available, and file it back into Epic as structured data. This work would allow relevant CDS development to further support clinicians' workflow, e.g., notifying the missing/conflicted/expired smoking data. In addition, this work provides a paradigm for extracting information from free text to store and combine with current structured data, which might be used in extracting other information, such as social determinants of health.

CONCLUSION

Overall, we presented a feasible approach to facilitate identifying 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 potentially diminish healthcare disparities in LCS eligibility.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank Dr. Xi Yang for providing the original NLP tool to extract smoking information.

FUNDING STATEMENT

This work was supported by NIH grants: R01AG062499-01 and K99LM014097-01.

DATA AVAILABILITY STATEMENT

The data underlying this article cannot be shared publicly due to patient healthcare data privacy protection requirements.

Abbreviations:

LCS	lung cancer screening
EHR	electronic health record
VUMC	Vanderbilt University Medical Center
NLP	natural language processing
USPSTF	United States Preventive Services Task Force
IDR	Integrated Data Repository

CDS	clinical decision support

REFERENCES

[1]. Siegel RL, Miller KD, Fuchs HE, Jemal A, Cancer statistics, 2022, CA. Cancer J. Clin 72 (2022) 7–33. 10.3322/caac.21708. [PubMed: 35020204]

[2]. Nanavaty P, Alvarez MS, Alberts WM, Lung Cancer Screening: Advantages, Controversies, and Applications, Cancer Control. 21 (2014) 9–14. 10.1177/107327481402100102. [PubMed: 24357736]

[3]. Landy R, Young CD, Skarzynski M, Cheung LC, Berg CD, Rivera MP, Robbins HA, Chaturvedi AK, Katki HA, Using Prediction Models to Reduce Persistent Racial and Ethnic Disparities in the Draft 2020 USPSTF Lung Cancer Screening Guidelines, 113 (2021). /pmc/articles/PMC8562965/ (accessed December 21, 2021).

[4]. Fedewa SA, Kazerooni EA, Studts JL, Smith RA, Bandi P, Sauer AG, Cotter M, Sineshaw HM, Jemal A, Silvestri GA, State Variation in Low-Dose Computed Tomography Scanning for Lung Cancer Screening in the United States, JNCI J. Natl. Cancer Inst 113 (2021) 1044–1052. 10.1093/jnci/djaa170. [PubMed: 33176362]

[5]. Lam ACL, Aggarwal R, Cheung S, Stewart EL, Darling G, Lam S, Xu W, Liu G, Kavanagh J, Predictors of participant nonadherence in lung cancer screening programs: a systematic review and meta-analysis, Lung Cancer. 146 (2020) 134–144. 10.1016/j.lungcan.2020.05.013. [PubMed: 32535225]

[6]. US Preventive Services Task Force, Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement., JAMA. 325 (2021) 962–970. 10.1001/jama.2021.1117. [PubMed: 33687470]

[7]. Reese TJ, Schlechter CR, Potter LN, Kawamoto K, Del Fiol G, Lam CY, Wetter DW, Evaluation of Revised US Preventive Services Task Force Lung Cancer Screening Guideline Among Women and Racial/Ethnic Minority Populations, JAMA Netw. Open 4 (2021) e2033769–e2033769. 10.1001/JAMANETWORKOPEN.2020.33769. [PubMed: 33433600]

[8]. Reese TJ, Schlechter CR, Kramer H, Kukhareva P, Weir CR, Del Fiol G, Caverly T, Hess R, Flynn MC, Taft T, Kawamoto K, Implementing lung cancer screening in primary care: needs assessment and implementation strategy design, Transl. Behav. Med 12 (2022) 187–197. 10.1093/TBM/IBAB115. [PubMed: 34424342]

[9]. V Kukhareva P, Caverly TJ, Li H, Katki HA, Cheung LC, Reese TJ, Del Fiol G, Hess R, Wetter DW, Zhang Y, Taft TY, Flynn MC, Kawamoto K, 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) 1–10. 10.1093/jamia/ocac020.

[10]. Ruckdeschel JC, Parasarathy S, Driscoll C, Identification of Patients Eligible for Lung Cancer Screening by Natural Language Processing, J. Natl. Compr. Cancer Netw 20 (2022) BIO22–029. 10.6004/jnccn.2021.7134.

[11]. 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/ (accessed February 27, 2022).

[12]. Wang L, Ruan X, Yang P, Liu H, Comparison of Three Information Sources for Smoking Information in Electronic Health Records, Cancer Inform. 15 (2016) CIN.S40604. 10.4137/CIN.S40604.

[13]. Onega T, Nutter E, Sargent J, Doherty J, Hassanpour S, Identifying Patient Smoking History for Cessation and Lung Cancer Screening through Mining Electronic Health Records, Cancer Epidemiol. Biomarkers Prev 26 (2017) 437.1–437. 10.1158/1055-9965.EPI-17-0032.

[14]. Fathi JT, White CS, Greenberg GM, Mazzone PJ, Smith RA, Thomson CC, The Integral Role of the Electronic Health Record and Tracking Software in the Implementation of Lung Cancer Screening—A Call to Action to Developers: A White Paper From the National Lung Cancer Roundtable, Chest. 157 (2020) 1674–1679. 10.1016/J.CHEST.2019.12.004. [PubMed: 31877270]

[15]. Yang X, Yang H, Lyu T, Yang S, Guo Y, Bian J, Xu H, Wu Y, A Natural Language Processing Tool to Extract Quantitative Smoking Status from Clinical Narratives, in: 2020 IEEE Int. Conf. Healthc. Informatics, IEEE, 2020: pp. 1–2. 10.1109/ICHI48887.2020.9374369.

[16]. 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 26 (2019) 960–967. 10.1093/jamia/ocz030.

[17]. 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 18 (2016) e263. 10.2196/jmir.5910. [PubMed: 27697748]

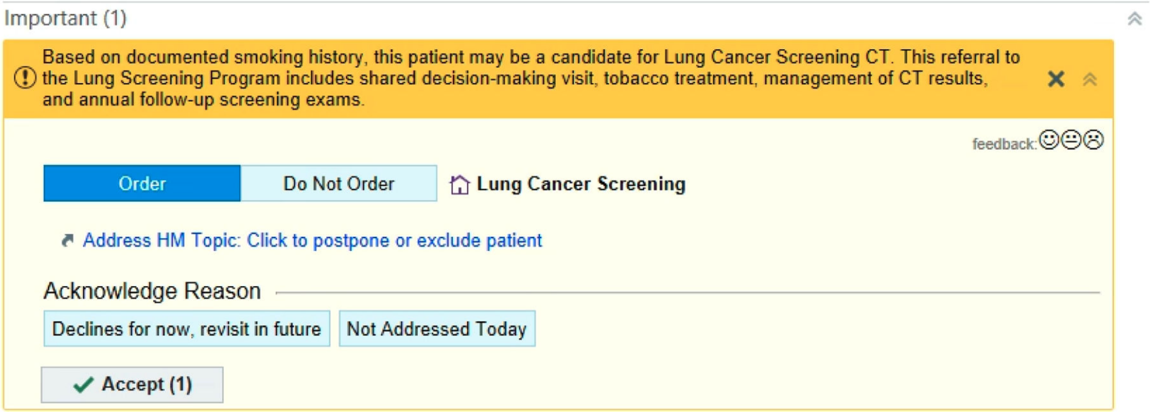
[18]. Volk RJ, Mendoza TR, Hoover DS, Nishi SPE, Choi NJ, Bevers TB, Reliability of self-reported smoking history and its implications for lung cancer screening, Prev. Med. Reports 17 (2020) 101037. 10.1016/j.pmedr.2019.101037.

Highlights

- Adding unstructured data improves lung cancer screening eligibility identification
- Increase in identifying Black/African Americans for lung cancer screening
- Data-driven approach revolutionizes lung cancer screening eligibility assessment

SUMMARY POINTS

- The free text smoking information in the EHR plays an important role in assessing eligibility for the lung cancer screening.
- After adding NLP extracted smoking information from clinical notes in 3 years, the number of identified lung cancer screening eligible patients were 10,231 (73.8% increment).
- This work provides a solid technical basis for the development of a CDS tool and implementation into clinical practice to improve lung cancer screening.



Important (1)

Based on documented smoking history, this patient may be a candidate for Lung Cancer Screening CT. This referral to the Lung Screening Program includes shared decision-making visit, tobacco treatment, management of CT results, and annual follow-up screening exams.

Figure 1.

A screenshot of an alert for lung cancer screening in the EHR system at Vanderbilt University Medical Center.

Table 1.

Patient Cohort.

Age (mean ± std)	69.9 ± 8.4
Black/African American	9,780 (9.5%)
Hispanic	2,073 (2.0%)
Insurance (Private)	51,136 (49.9%)
Female	57,986 (56.6%)
Smoking Status (n=98,294)	
Never	56,744 (55.4%)
Former	29,451 (28.7%)
Yes	11,500 (11.2%)
Other (Passive/Not Asked)	605 (0.6%)
Structured Smoking Data (n=1,914,701)	
Years-smoked (n=21,216)	23.9 ± 16.5
Packs-per-day (n=25,449)	1.0 ± 1.2
Pack Year (n=19,226)	26.9 ± 32.6
Years Since Quit (n=22,050)	22.4 ± 15.6
Notes (n=4,521,645)	
Year	
2021.08–2022.08	1,817,848 (40.2%)
2020.08–2021.08	1,527,950 (33.8%)
2019.08–2020.08	1,175,847 (26%)
Encounter Type	
Hospital Encounter	2,282,976 (50.5%)
Office Visit	2,238,669 (49.5%)
Note Type	
Progress Note	1,820,168 (40.3%)
Assessment & Plan Note	655,911 (14.5%)
Letter	419,086 (9.3%)
Result Encounter Note	190,330 (4.2%)
Patient Instruction	140,459 (3.1%)
Number of notes for each patient (median, [25%, 75%])	22 [8, 51]

Table 2.

Issues identified in the structured dataset.

Issues in the structured dataset	In recent records	In all structured records
Smokers without any value of pack-per-day or years-smoked (%: percentage in current smokers)	5,857 (50.9%)	5,728 (49.8%)
Smokers with a maximum 0 pack-per-day or years-smoked (%: percentage in current smokers)	543 (4.7%)	485 (4.2%)
Past smokers without any value of pack-per-day or years-smoked (%: percentage in past smokers)	13,244 (45.0%)	13,006 (44.2%)
Past smokers with a maximum 0 pack-per-day or years-smoked (%: percentage in past smokers)	2,988 (10.1%)	2,658 (9.0%)
Past smokers without any value of quit date (%: percentage in past smokers)	8,991 (30.5%)	8,889 (30.2%)
Overall number of patients cannot assess the LCS eligibility (%: percentage in current and past smokers)	23,941 (58.5%)	23,210 (56.7%)

Table 3.

Performance of different approaches to assessing LCS eligibility.

	Pack year >=20 and current smoking	Pack year >=20 and quit within 15 years	LCS Eligible	Increment
Structured Data (Baseline)	2,888	2,999	5,887	-
Structured Data + Improved algorithm	3,567	3,627	7,194	22.2%
Structured Data + 3-yr NLP extracted data + Improved algorithm	4,926	5,305	10,231	73.8%

Table 4.

Demographic and smoking information of patients eligible for lung cancer screening within Vanderbilt University Medical Center, 2019–2022.

	Patients identified in the Baseline approach	**Newly identified patients using the NLP-based approach**	**Relative Increment**
Number of patients	5,887	4,344	73.8%
Age (mean, std)	64.3 (7.5)	63.6 (7.7) *	-
Black/African American	495 (8.4%)	**589 (13.6%)** *	**119%**
Hispanic	62 (1.1%)	66 (1.5%)	106.5%
Female	2,701 (45.9%)	2,072 (47.7%)	76.7%
Private Insurance	2,265 (38.5%)	1,673 (38.5%)	73.8%
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%) *	70.6%
Former Smoker (%)	2,999 (50.9%)	2,311 (53.2%) *	77.1%
Quit Years (mean, std)	6.4 (4.9)	4.6 (4.7) *	-

(*. *P*<0.001).