UCLA
UCLA Previously Published Works

Title

Toward a Computable Phenotype for Determining Eligibility of Lung Cancer Screening Using Electronic Health Records.

Permalink
https://escholarship.org/uc/item/7mk703tv

Authors
Yang, Shuang
Huang, Yu
Lou, Xiwei
et al.

Publication Date
2025

DOI
10.1200/CCI.24.00139

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Toward a Computable Phenotype for Determining Eligibility of Lung Cancer Screening Using Electronic Health Records

Shuang Yang, MS[1]; Yu Huang, PhD[1]; Xiwei Lou, MS[1]; Tianchen Lyu, MS[1,2]; Ruoqi Wei, PhD[1]; Hiren J. Mehta, MD[3]; Yonghui Wu, PhD[1,2]; Michelle Alvarado, PhD[4]; Ramzi G. Salloum, PhD[1]; Dejana Braithwaite, PhD[5,6]; Jinhai Huo, MD, MsPH, PhD[7]; Ya-Chen Tina Shih, PhD[8]; Yi Guo, PhD[1,2]; and Jiang Bian, PhD[1,2]

DOI https://doi.org/10.1200/CCI.24.00139

ABSTRACT

PURPOSE Lung cancer screening (LCS) has the potential to reduce mortality and detect lung cancer at its early stages, but the high false-positive rate associated with low-dose computed tomography (LDCT) for LCS acts as a barrier to its widespread adoption. This study aims to develop computable phenotype (CP) algorithms on the basis of electronic health records (EHRs) to identify individual's eligibility for LCS, thereby enhancing LCS utilization in real-world settings.

MATERIALS AND METHODS The study cohort included 5,778 individuals who underwent LDCT for LCS from 2012 to 2022, as recorded in the University of Florida Health Integrated Data Repository. CP rules derived from LCS guidelines were used to identify potential candidates, incorporating both structured EHR and clinical notes analyzed via natural language processing. We then conducted manual reviews of 453 randomly selected charts to refine and validate these rules, assessing CP performance using metrics, for example, F1 score, specificity, and sensitivity.

RESULTS We developed an optimal CP rule that integrates both structured and unstructured data, adhering to the US Preventive Services Task Force 2013 and 2020 guidelines. This rule focuses on age (55-80 years for 2013 and 50-80 years for 2020), smoking status (current, former, and others), and pack-years (≥30 for 2013 and ≥20 for 2020), achieving F1 scores of 0.75 and 0.84 for the respective guidelines. Including unstructured data improved the F1 score performance by up to 9.2% for 2013 and 12.9% for 2020, compared with using structured data alone.

CONCLUSION Our findings underscore the critical need for improved documentation of smoking information in EHRs, demonstrate the value of artificial intelligence techniques in enhancing CP performance, and confirm the effectiveness of EHR-based CP in identifying LCS-eligible individuals. This supports its potential to aid clinical decision making and optimize patient care.

ACCOMPANYING CONTENT

Data Sharing Statement

Data Supplement

Accepted December 5, 2024
Published January 16, 2025

JCO Clin Cancer Inform
9:e2400139

© 2025 by American Society of Clinical Oncology

Creative Commons Attribution Non-Commercial No Derivatives 4.0 License

INTRODUCTION

Lung cancer is the leading cause of cancer death in the United States, exceeding the total number of deaths from breast, prostate, colorectal, and leukemia cancers combined.[1] Early detection plays a crucial role in reducing lung cancer mortality.[2] Results from the US National Lung Screening Trial (NLST) have demonstrated that using low-dose computed tomography (LDCT) for screening reduced lung cancer mortality by 20%.[3] Since then, numerous professional societies and medical associations, such as the US Preventive Services Task Force (USPSTF), Centers for Medicare and Medicaid Services (CMS), American Cancer Society (ACS), National Comprehensive Cancer Network (NCCN), and ASCO, have issued guidelines recommending LDCT for lung cancer screening (LCS) in high-risk individuals.[4-11] The eligibility criteria for LCS with LDCT outlined in those guidelines are derived from the NLST study, primarily focusing on two aspects: age and smoking history. However, as shown in the Data Supplement (Table S1), there are slight variations in these criteria across the different guidelines. For example, the age criteria in the 2013 USPSTF guideline was from 55 to 80 years, while it was 55 to 77 years in the CMS decision memo. With more evidence on LCS benefits, the USPSTF has updated their guideline in 2020, expanding the eligibility criteria to include more at-risk populations. The new 2020 USPSTF guideline lowered the initial screening age from 55 to 50 years and the cumulative

CONTEXT

Key Objective

This study develops and validates computable phenotype (CP) algorithms onthe basis of electronic health records (EHRs) from University of Florida Health Integrated Data Repository to identify individual's eligibility for lung cancer screening (LCS) in real-world settings.

Knowledge Generated

The proposed CP algorithms achieve high F1 scores of 0.75 and 0.84 by integrating both structured and unstructured EHRs, in accordance with the 2013 and 2020 US Preventive Services Task Force guidelines, respectively. Our results highlight the importance of artificial intelligence (AI) techniques in enhancing CP performance and validate the effectiveness of EHR-based CP in identifying eligible individuals for LCS.

Relevance (*J.L. Warner*)

While the single-institutional nature of the study may limit generalizability, this large well-conducted study demonstrates that AI can be used with good performance to identify individuals eligible for LCS. Future efforts should focus on external validation with diverse national and international data sets.*

*Relevance section written by *JCO Clinical Cancer Informatics* Editor-in-Chief Jeremy L. Warner, MD, MS, FAMIA, FASCO.

smoking intake requirement from 30 pack-years to 20 pack-years.[9] After this change, both NCCN[10] and CMS[5,11] have updated their recommendations to align with the current USPSTF guideline.

Nevertheless, the high false-positive rate (around 23.3%) of LDCT in the original NLST remains a significant concern,[3] and the issue is even more pronounced in real-world settings. For example, a Veterans Health Affairs study reported a much higher false-positive rate of 58.2%.[12] False positive from LCS can lead to unnecessary diagnostic invasive procedures, including biopsies and surgeries, resulting in postprocedural complications and substantial health care costs. Real-world studies have reported compilation rates of invasive procedures ranging from 16.6% to 30.6%,[13-15] significantly higher than the 9.8% reported in the NLST. These compilations lead to financial burdens ranging from $6,320 in US dollars (USD) for minor complications to $56,845 (USD) for major ones.[14] Hence, the high false-positive rate poses a barrier to the widespread adoption of LCS. Accurately identifying individuals who meet the LCS eligibility criteria in real-world settings has the potential to increase LCS uptake.[16] Computable phenotype (CP)[17] is an increasingly used method that uses real-world data (RWD),[18,19] such as electronic health records (EHRs) to identify individuals with given traits through a set of executable algorithms. Leveraging EHRs, CP algorithms can automate and more precisely define the at-risk patients eligible for cancer screening. For example, Petrik et al validated an EHR-based algorithm for identifying eligible patients in colorectal cancer screening.[34] Leder Macek et al evaluated a rule-based algorithm to automatically determine patient eligibility and adherence of cancer screening.[35] Liu et al used a novel natural language processing (NLP)–based approach to identify patients eligible for LCS using EHR data from Vanderbilt University Medical Center.[36] EHRs provide an electronic representation of a patient's medical history, containing valuable clinical information of individuals such as diagnoses, procedures, medications, vitals, and laboratory tests. The widespread adoption of EHRs and the creation of clinical research networks (eg, the National Patient-Centered Clinical Research Network) with large collections of EHRs (along with linkages to other RWD sources) have facilitated the availability of large-scale, longitudinal clinical data available for research purposes.

Challenges exist in using EHR data to identify eligible individuals for LCS because it requires fine-grained smoking status and history (ie, pack-years), which may not be captured in structured EHRs. Smoking information is increasingly documented in EHRs[20] because of its significance as a risk factor for multiple diseases and the need for appropriate presurgical counseling[21] and assessing readiness to quit smoking.[22] However, this information is often only documented in clinical narratives rather than in structured data. EHR data also face quality issues such as discrepancies over time and variations across different sources within the system.[23,24] For example, a study shows that among 47,849 unique individuals, 67.1% had more than one note documenting smoking status, and 54.5% had implausible changes in documentation (39.1% of which involved conflicting smoking status).[23] Addressing these challenges to accurately extract patient information from EHRs is critical for developing a LCS CP. Clinical NLP is essential for achieving these goals as over 80% of the clinical information in EHRs is documented in clinical narratives.[25]

To accurately identify eligible individuals for LCS in a high-throughput way, it depends on our ability to narrow down proper and computable search criteria to identify *true* cases. This study aimed to develop robust EHR-based CP rules to accurately identify at-risk populations eligible for LCS. We first analyzed existing LCS guidelines[4-8,26] and selected the USPSTF guidelines as our primary criteria. Using the broadest criteria, we developed the initial CP rules. Compared with the USPSTF guidelines, our CP includes more comprehensive criteria to address complex scenarios in real-world settings (eg, accounting for missing or unknown smoking status), preventing the exclusion of at-risk individuals and improving the accuracy of the CP. Leveraging our past work on clinical NLP,[27] we developed a tool to extract fine-grained smoking information from clinical narratives[28] and designed rules to address the data quality issues. We then conducted manual chart reviews to establish the *ground truth* and iteratively validated and revised the CP rules until achieving reasonably high performance.

MATERIALS AND METHODS

In this study, we used RWD from the University of Florida (UF) Health Integrated Data Repository (IDR).[29] Detailed description of data source and study design are provided in the Data Supplement. Figure 1 illustrates our overall workflow for developing the LCS CP. We began by initializing 11 CP rules (Data Supplement) to identify eligible LCS individuals, focusing only on age, smoking status, and smoking history according to LCS guidelines. We then used a two-step iterative process to refine and validate these CP rules.

Step 1—Identify Individuals Potentially Eligible for LCS Using the Current CP Rules, Leveraging Both Structured and Unstructured EHR Data

For individuals who received LDCT in the UF IDR, we retrieved structured EHR data (eg, encounters, diagnoses, and procedures) and unstructured clinical narratives (eg, progress notes, discharge summaries, radiology reports, and pathology reports). Each individual's data before the index date were used to develop the CP. A detailed description of the eligible rules is presented in the Data Supplement (Table S2).

Step 2: Validate the Cohort Through Manual Chart Reviews of Randomly Selected Samples to Assess the CP Performance and Refine the CP Rules

To focus the analysis on individuals without recent cancer history, we excluded those diagnosed with lung cancer within the year preceding their index date, as well as those diagnosed with any type of cancer (malignant or benign) within the 5 years before the index date. To ensure precision of CP algorithms, following Buderer,[30] we calculated a minimal sample size of 292 for manual chart review, assuming an anticipated 0.95 sensitivity and specificity, and a disease prevalence of 6.27%. We systematically generated various combinations of CP rules and used them to identify potential LCS-eligible individuals. On the basis of the distribution of each rule combination, we randomly selected 453 samples for chart review. Detailed description of chart review is provided in the Data Supplement.

RESULTS

In the UF Health IDR cohort of 5,778 individuals who underwent LDCT, 2,130 with a history of cancer were excluded, leaving 3,648 potentially eligible for LCS.

We generated all possible combinations of the 11 initial rules, focusing on four criteria: age, smoking status, pack-years, and quit-years (detailed in Step 1 in the Methods section above), and then applied them to our data set, retaining only the combinations where there are patients met the criteria and excluding those with zero patients. This process results in 27 combinations of rules with patients fulfilling the criteria. Samples were randomly selected from each of the 27 combinations for manual chart review. Table 1 displays the top 10 combinations of individual CP rules on the basis of the USPSTF 2013 and 2020 guidelines, along with the number of individuals meeting each rule combination. For rules with a high number of hits, only a limited number of subsets was reviewed. For instance, of the 1,316/1,417 hits for the first rule, only 109/128 were reviewed for 2013 and 2020 guidelines, respectively. A statistically significant sample size is sufficient for accurate analysis. Beyond a certain point, additional samples offer only marginal improvements to the performance metrics. Of the 453 charts reviewed, the top 10 rule combinations from the 2013/2020 guidelines (363/398 individuals) are presented in Table 1.

Using the results of manual chart review as gold-standard labels (yes, no, and maybe), we systematically generated various permutations of individual CP rules, considering two scenarios: (1) rule combinations using only structured data, and (2) rule combinations using both structured and unstructured data. We also conducted three subgroup analyses on the basis of the gold-standard labels: (1) yes versus maybe/no, (2) yes versus no, and (3) yes/maybe versus no. Table 2 presents the final CP rules under these three subgroups across both scenarios, on the basis of the 2013 and 2020 guidelines. Performance metrics, including positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity, and F1 scores were reported in Table 2.

The most effective CP rules for identifying LCS eligibility on the basis of the 2013 and 2020 USPSTF guidelines were:

- Age between 55 and 80 years (2013)/age between 50 and 80 years (2020) AND smoking status were current or other AND maximum pack-years ≥30 (2013)/≥20 (2020).

OR

- Age between 55 and 80 years (2013)/age between 50 and 80 years (2020) AND smoking status was former AND

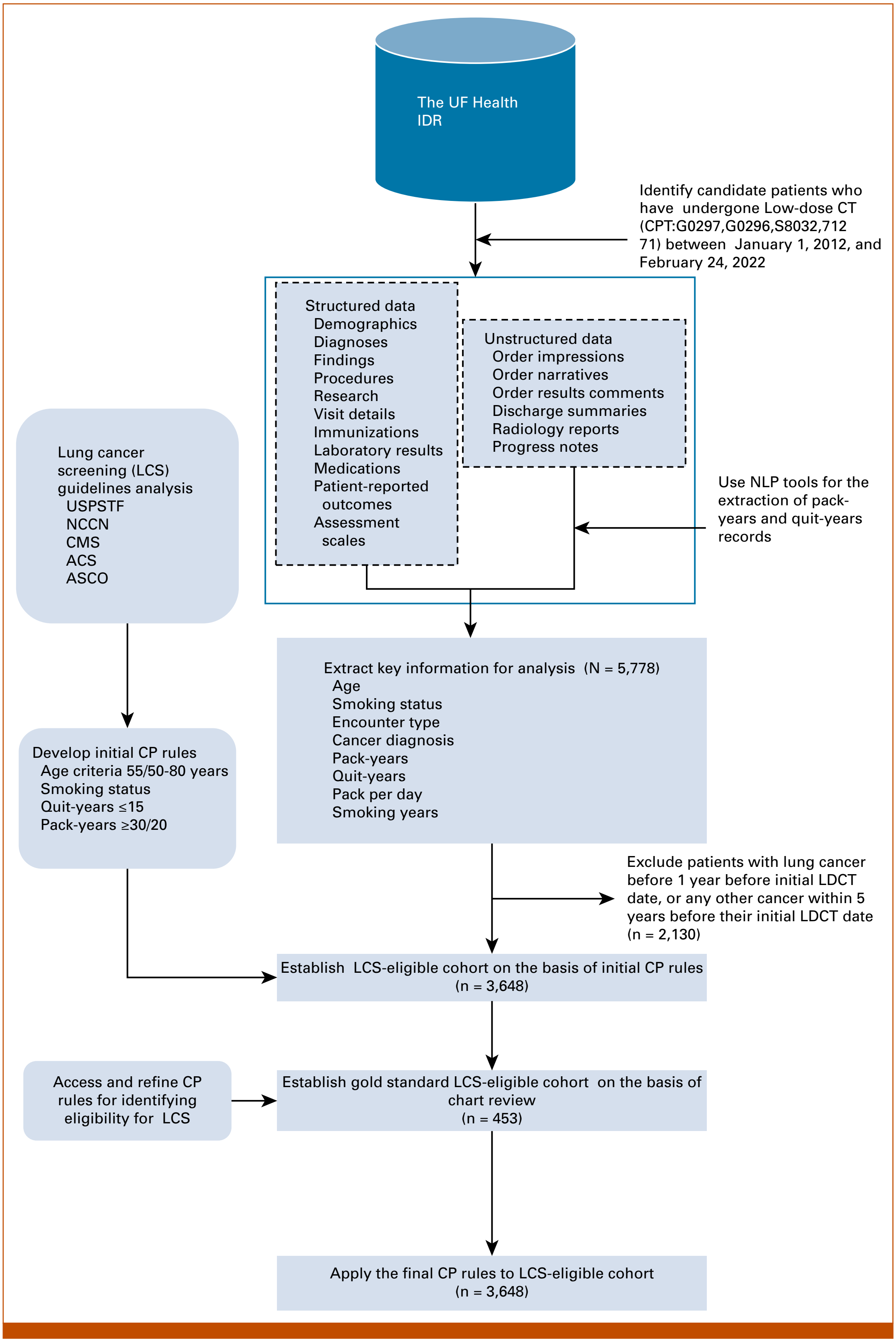
The UF Health
IDR
Identify candidate patients who have undergone Low-dose CT (CPT:G0297,G0296,S8032,71271) between January 1, 2012, and February 24, 2022
Structured data
Demographics
Diagnoses
Findings
Procedures
Research
Visit details
Immunizations
Laboratory results
Medications
Patient-reported outcomes
Assessment scales
Unstructured data
Order impressions
Order narratives
Order results comments
Discharge summaries
Radiology reports
Progress notes
Use NLP tools for the extraction of pack-years and quit-years records
Lung cancer screening (LCS) guidelines analysis
USPSTF
NCCN
CMS
ACS
ASCO
Extract key information for analysis (N = 5,778)
Age
Smoking status
Encounter type
Cancer diagnosis
Pack-years
Quit-years
Pack per day
Smoking years
Develop initial CP rules
Age criteria 55/50-80 years
Smoking status
Quit-years ≤15
Pack-years ≥30/20
Exclude patients with lung cancer before 1 year before initial LDCT date, or any other cancer within 5 years before their initial LDCT date (n = 2,130)
Establish LCS-eligible cohort on the basis of initial CP rules (n = 3,648)
Access and refine CP rules for identifying eligibility for LCS
Establish gold standard LCS-eligible cohort on the basis of chart review (n = 453)
Apply the final CP rules to LCS-eligible cohort (n = 3,648)

FIG 1. Workflow of determining eligibility of LCS on the basis of both structured and unstructured data. ACS, American Cancer Society; CMS, Centers for Medicare & Medicaid Services; CP, computable phenotyping; CT, computed tomography; IDR, Integrated Data Repository; LCS, lung cancer screening; LDCT, low-dose computed tomography; NCCN, National Comprehensive Cancer Network; NLP, natural language processing; USPSTF, US Preventive Services Task Force.

TABLE 1. Combinations of Individual Computable Phenotype Rules With Corresponding Sample Sizes and Manual Chart Review Selections

Age, Years	Smoking Status				Quit-Years		Pack-Years		No. of Patients[a]	No. of Patients Selected[b]
55-80 or 50-80[c]	Current Smoker	Former Smoker	Other Smoker	Unknown Smoker	Smallest ≤15	Unknown	Max ≥30 or 20[c]	Unknown	2013 Guideline/2020 Guideline	2013 Guideline/2020 Guideline
+	+					+	+		1,316/1,417	109/128
+		+			+		+		929/979	116/126
+	+					+		+	473/485	35/36
+		+			+			+	248/256	32/32
+		+				+		+	82/82	28/28
+				+				+	74/75	18/19
+		+				+	+		51/53	13/15
+			+				+		18/20	8/9
+			+					+	6/7	3/3
+							+		5/6	1/2

NOTE. + means the individuals must fulfill this rule. We did so because we wanted to evaluate the performance of each individual rule.

Abbreviation: USPSTF, US Preventive Services Task Force.

[a] Only the top 10 rule combinations on the basis of the 2013 USPSTF guideline/2020 USPSTF guideline are shown here.

[b] We randomly selected individuals for chart review on the basis of the distribution of the total number of individuals who met each rule combination. Overall, we reviewed 453 individuals' charts. Only 363/398 individuals are shown in this table as the table only lists the top 10 rule combinations.

[c] The number preceding the slash corresponds to the 2013 USPSTF guideline, while the number following the slash corresponds to the 2020 USPSTF guideline.

smallest quit-years ≤15 or unknown AND maximum pack-years ≥30 (2013)/≥20 (2020).

OR

- Age between 55 and 80 years (2013)/age between 50 and 80 years (2020) AND smoking status unknown

As shown in Table 2, CP rules using both structured and unstructured data outperformed those using only structured data, achieving higher F1 scores (0.71-0.75 *v* 0.65-0.70 and 0.79-0.84 *v* 0.70-0.75 for the 2013 and 2020 guidelines), NPV (0.55-0.73 *v* 0.53-0.61 and 0.47-0.76 *v* 0.41-0.62), and sensitivity (0.69-0.82 *v* 0.53-0.61 and 0.75-0.85 *v* 0.58-0.65). For the USPSTF 2013 guideline, the inclusion of unstructured data led to a 30.19% increase in sensitivity and a 9.23% increase in the F1 score. Similarly, for the USPSTF 2020 guideline, sensitivity increases by 29.31%, and the F1 score improves by 12.86%. However, the combined data sources exhibited lower PPV (0.63-0.72 *v* 0.72-0.84 and 0.75-0.84 *v* 0.80-0.90) and specificity (0.58-0.60 *v* 0.8-0.84 and 0.61-0.62 *v* 0.77-0.82). Specifically, there was a 14.29% decrease in the PPV and a 30.95% decrease in specificity for the 2013 guideline, and a 6.67% decrease in PPV and a 25.61% decrease for 2020 guideline. CP rules on the basis of the USPSTF 2020 guideline have higher F1 scores (0.70-0.84 *v* 0.65-0.75), PPV (0.75-0.89 *v* 0.63-0.84), and sensitivity (0.65-0.85 *v* 0.53-0.82) compared with those on the basis of the USPSTF 2013 guideline. In subgroup analyses, categorizing maybe as no yields the better F1 scores (0.66-0.80 *v* 0.65-0.79), NPV (0.62-0.80 *v* 0.41-0.53), and sensitivity (0.61-0.85 *v* 0.53-0.75) compared with categorizing maybe as yes. Detailed subset CP rules are listed in the Data Supplement (Table S3), and the rules were selected on the basis of comprehensive evaluation of PPV, NPV, and F1 scores.

Figure 2 illustrates a decision tree that applies our CP rules to determine LCS eligibility. It starts by evaluating cancer history: those diagnosed with lung cancer within the past year or any type of cancer within the past 5 years before the index date are ineligible, as they are considered as cancer survivors requiring different screening practices. Next, individuals age 55-80 or 50-80 years are eligible according to the 2013 guideline or 2020 guideline. Smoking status is reassessed on the basis of the latest available information, including quit-years and other relevant smoking information. For former smokers, eligibility depends on whether they quit in the past 15 years and the number of pack-years smoked. For other smoking status, eligibility is determined by the number of pack-years smoked—>20 (2020 guideline) or >30 (2013 guideline). The tree leads to several end points indicating LCS eligibility, with detailed notes explaining the decision process.

We applied the CP rules from the decision tree to determine LCS eligibility in our UF Health LDCT cohort and assess the adequacy of their LCS procedures. Among 3,648 individuals without a cancer history, 2,389 (65.49%) and 2,544 (69.74%) were deemed eligible for LCS under the USPSTF 2013 and 2020 guidelines, respectively. The Data

TABLE 2. Performance of the Final Computable Phenotypes for Determining Lung Cancer Screening Eligibility

	Performance				
Comparison Group	PPV	NPV	Sensitivity	Specificity	F1 Score
USPSTF 2013 guideline					
Structured data only					
Yes/maybe *v* no	**0.84**	0.53	0.53	**0.84**	0.65
Yes *v* maybe/no	0.72	0.72	0.61	0.8	0.66
Yes *v* no	0.82	0.65	0.61	**0.84**	0.70
Structured data and unstructured data					
Yes/maybe *v* no	0.72	0.55	0.69	0.58	0.71
Yes *v* maybe/no	0.63	**0.80**	**0.82**	0.60	0.71
Yes *v* no	0.70	0.73	**0.82**	0.59	**0.75**
USPSTF 2020 guideline					
Structured data only					
Yes/maybe *v* no	**0.90**	0.41	0.58	**0.82**	0.70
Yes *v* maybe/no	0.80	0.62	0.65	0.77	0.72
Yes *v* no	0.89	0.52	0.65	**0.82**	0.75
Structured data and unstructured data					
Yes/maybe *v* no	0.84	0.47	0.75	0.61	0.79
Yes *v* maybe/no	0.75	**0.76**	**0.85**	0.62	0.80
Yes *v* no	0.83	0.66	**0.85**	0.61	**0.84**

NOTE. Values in bold indicate the best value per metric. Multiple bold numbers per column signify that there is no statistical difference among the top-performing values for the given metric.
Abbreviations: NPV, negative predictive value; PPV, positive predictive value; USPSTF, US Preventive Services Task Force.

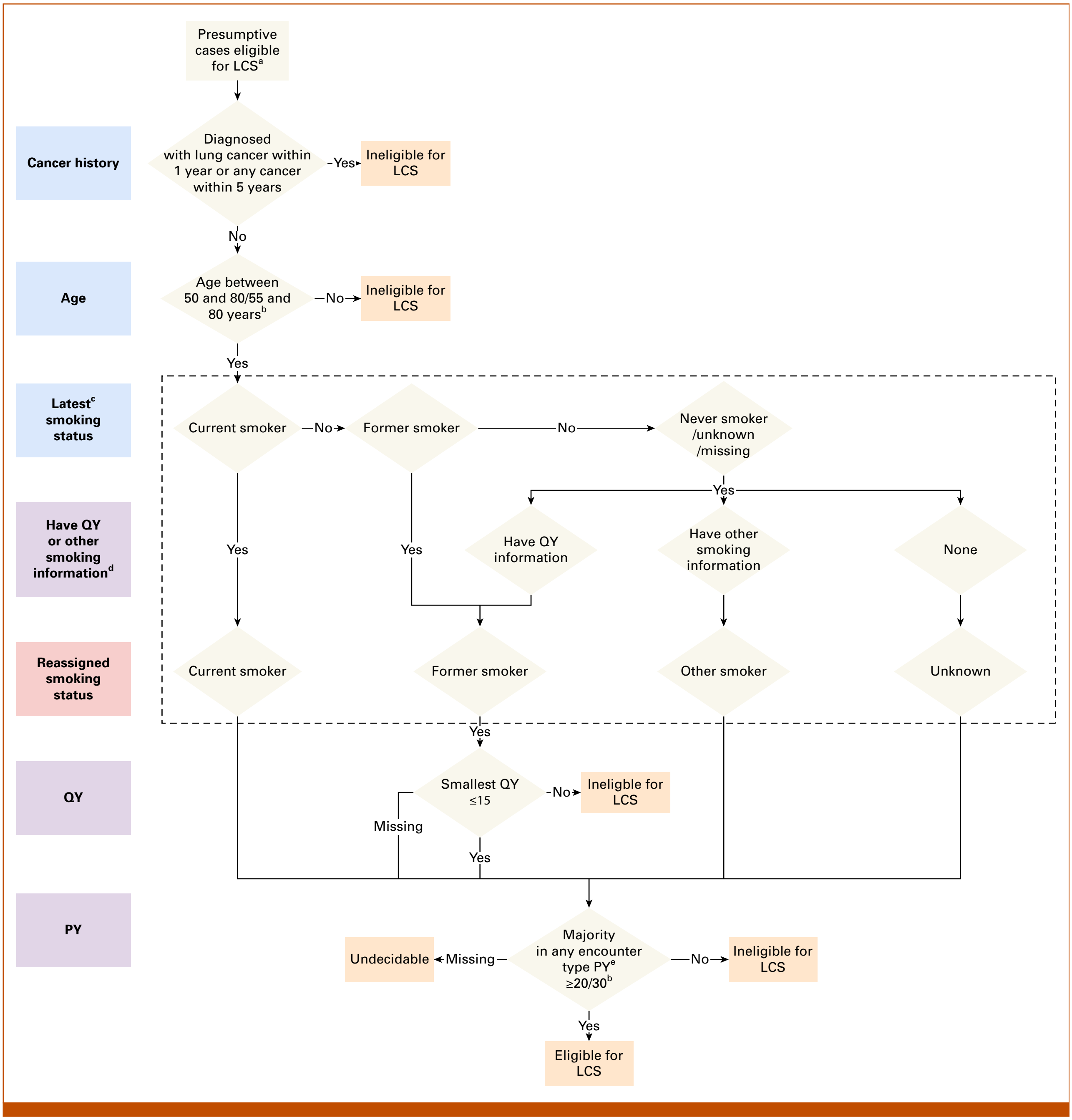
Presumptive cases eligible for LCS[a]
Cancer history
Diagnosed with lung cancer within 1 year or any cancer within 5 years
Yes
Ineligible for LCS
No
Age
Age between 50 and 80/55 and 80 years[b]
No
Ineligible for LCS
Yes
Latest[c] smoking status
Current smoker
No
Former smoker
No
Never smoker /unknown /missing
Yes
Have QY or other smoking information[d]
Yes
Yes
Have QY information
Have other smoking information
None
Reassigned smoking status
Current smoker
Former smoker
Other smoker
Unknown
Yes
QY
Smallest QY ≤15
No
Ineligible for LCS
Missing
Yes
PY
Undecidable
Missing
Majority in any encounter type PY[e] ≥20/30[b]
No
Ineligible for LCS
Yes
Eligible for LCS

FIG 2. The LCS eligibility decision tree. Blue rectangle: information comes from structured data only. Purple rectangle: information comes from both structured data and unstructured data. Pink rectangle: reassigned information. Light white diamond: the rules. Light orange rectangle: the final decision of eligibility of LCS. [a]LCS refers to lung cancer screening. The decision-making process is informed by patient data available before the index date. [b]The USPSTF revised their lung cancer screening guidelines in 2020, reducing the minimum age for screening from 55 to 50 years and the smoking history criterion from 30 to 20 pack-years. [c]Latest pertains to the most recent record before the index date. [d]Other smoking information included (1) smoking start date, tobacco pack per day, tobacco used years, any one of them has value; (2) have current smoker, former smoker, or smoker, current status unknown in records before index date. [e]Majority pack-year refers to the predominant pack-year value obtained from each note date. For determining the majority, each patient encounter counts only unique pack-year records, regardless of the data source. LCS, lung cancer screening; PY, pack-year; QY, quit-year.

Supplement (Fig S1) shows the cohort ascertainment diagram. Table 3 details the demographics of the cohort by eligibility identified by CP rules. Most eligible individuals were age 55-74 years at their first LDCT (90.38% and 92.20% under 2013 and 2020 guidelines, respectively). Non-Hispanic Whites were the majority among eligible individuals (72.00% and 71.03%, respectively), followed by non-Hispanic Blacks, Hispanics, and others. A slight

TABLE 3. Demographics of the Individuals Who Had Taken Low-Dose Computed Tomography in UF Health

		CP Algorithm on the Basis of the USPSTF (2013)[a] Criteria		CP Algorithm on the Basis of the USPSTF (2020)[a] Criteria	
Characteristic	Overall, N = 3,648 (100%)	Eligible (n = 2,389; 65.49%)	Ineligible (n = 1,259; 34.51%)	Eligible (n = 2,544; 69.74%)	Ineligible (n = 1,104; 30.26%)
Age, years, No. (%)[b]					
<55	62 (1.7)	2 (0.08)	60 (4.77)	23 (0.9)	39 (3.53)
55-64	1,625 (44.54)	1,117 (46.76)	508 (40.35)	1,188 (46.7)	437 (39.58)
65-74	1,589 (43.56)	1,042 (43.62)	547 (43.45)	1,099 (43.2)	490 (44.38)
75-80	322 (8.83)	206 (8.62)	116 (9.21)	212 (8.33)	110 (9.96)
>80	50 (1.37)	22 (0.92)	28 (2.22)	22 (0.86)	28 (2.54)
Race/ethnicity, No. (%)					
Hispanic	90 (2.47)	64 (2.68)	26 (2.07)	69 (2.71)	21 (1.9)
NHW	2,497 (68.45)	1,720 (72)	777(61.72)	1,807 (71.03)	690 (62.5)
NHB	915 (25.08)	515 (21.56)	400 (31.77)	569 (22.37)	346 (31.34)
Other	99 (2.71)	61 (2.55)	38 (3.02)	68 (2.67)	31 (2.81)
Unknown	47 (1.29)	29 (1.21)	18 (1.43)	31 (1.22)	16 (1.45)
Sex, No. (%)					
Female	1,765 (48.38)	1,110 (46.46)	655 (52.03)	1,194 (46.93)	571 (51.72)
Male	1,883 (51.62)	1,279 (53.54)	604 (47.97)	1,350 (53.07)	533 (48.28)

Abbreviations: CP, computable phenotype; NHB, non-Hispanic Black; NHW, non-Hispanic White; USPSTF, US Preventive Services Task Force.

[a] The original USPSTF lung cancer screening guideline was released in 2013, and the updated guideline was released in 2020.

[b] Age at the first low-dose computed tomography.

gender imbalance was observed, with males comprising 53.54% under the 2013 criteria and 53.07% under the 2020 criteria.

DISCUSSION

In this study, we used data from UF health IDR to develop and validate two sets of CP rules for assessing eligibility for LCS on the basis of 2013 and 2020 USPSTF guidelines. The CP rules achieved high F1 scores of 0.75 and 0.84, respectively, using both structured and unstructured data. Applying these CP rules to our cohort, we identified 65.49% and 69.74% of individuals as eligible for LCS under the 2013 and 2020 guidelines, respectively. These findings demonstrate their robustness of CPs in effectively assessing LCS eligibility and highlight the benefits of integrating diverse data sources.

Our results demonstrated that incorporating unstructured data improves F1 score performance by up to 9.23% and 12.86% under the 2013 and 2020 guidelines, respectively, compared with using structured data alone. Sensitivity increased by approximately 30%, but this came at the cost of up to 31% decrease in PPV or specificity. The incomplete nature of structured data, often missing critical details such as pack-years, can lead to lower F1 scores and sensitivity because of the omission of potential positives. Although integrating unstructured data enhances the algorithm's ability to identify eligible individuals for LDCT screening, it also inevitably leads to a higher incidence of false positives. There is a tradeoff between increased sensitivity and decreased specificity and PPV. Although higher sensitivity is generally preferred for screening tools to avoid missing opportunities for life-saving interventions, scenario specific considerations must also be taken into account. In the case of LCS, higher false-positive rates would lead to unnecessary invasive procedures, postprocedural complications, and increased health care costs. A more comprehensive tool is needed to illustrate both the benefits and harms of LCS, enabling patients and their physicians to make informed decision about screening. Further analysis revealed that in subgroup assessments, treating maybe as no yielded better F1 scores than treating maybe as yes. This suggests that classifying uncertain cases as likely true negatives enhances CP rule performance by reducing false negatives. Additionally, the 2020 USPSTF guideline outperformed the 2013 version, showing an improvement of up to 12.70% in F1 score. This enhancement is attributed to relaxed eligibility criteria, specifically the reduction in age from 55 to 50 years and the decrease in required pack-years from 30 to 20, which expanded detection capabilities and broadened inclusion criteria. Despite these modifications, the demographic analysis of cohorts derived from both CP sets showed consistent distributions across age, sex, and race/ethnicity, indicating that the updated guidelines successfully expanded eligibility without introducing demographic biases.

We further compared our study with the existing literature by Ruckdeschel et al.[33] They developed an NLP tool to extract smoking information and applied age and smoking criteria from clinical notes to identify LDCT eligibility on the basis of the USPSTF 2020 guideline. Their method was evaluated on MIMIC-III database and achieved an F1 score of 0.88.[33] Although their results were comparable with ours, our study offers two advantages. First, we used EHRs from a general cohort at UF Health, whereas the MIMIC-III data set is limited to patients in critical care units, which is not the primary target population of LCS. Second, we integrated both structured and unstructured EHRs for a comprehensive analysis. As a result, our CP algorithm has been rigorously evaluated in real-world settings, demonstrating its potential to accurately identify true cases for LDCT screening.

We explored the reasons why individuals were categorized as ineligible for LCS, according to their cancer history, age, smoking status, quit-years, and pack-years. We manually reviewed the clinical notes of 35 individuals who are either ineligible or maybe eligible for LCS and summarized the reasons for undergoing LDCT in the Data Supplement (Table S4). The clinical notes included an indication section, explaining the clinicians' rationale for recommending LDCT. The results showed that a significant portion (57.14%) underwent LDCT because of a personal history of tobacco use, presenting hazards to health. This finding emphasizes that these individuals sought to mitigate potential risks linked to their smoking history, despite not meeting the LCS eligibility criteria. This review highlighted that many individuals actively taken LDCT to reduce potential health risks associated with their smoking history, although they did not meet the eligibility criteria for LCS.

We assessed adherence to LDCT screening recommendations by examining how well individuals followed clinician advice. Adherence was defined as having a follow-up computed tomography (CT)/LDCT within 1 month before or 2 months after the recommended timeframe. We reviewed and extracted relevant information for 25 LCS-eligible and 10 LCS-ineligible individuals. An example clinician follow-up recommendation noted was management: 3-month LDCT. Positron emission tomography/CT may be used. Among the eligible individuals, although 16 had follow-up scans, only seven followed the clinician recommendations accurately, resulting in eight instances of adherence (one individual adhered twice, while six adhered once). By contrast, LDCT screening is inappropriate for ineligible individuals, who should not pursue follow-up LCS according to guidelines. Yet, we observed that 2 of 10 ineligible individuals still underwent follow-up scans. Figure 3A illustrates the timelines of a 65-year-old LCS-eligible individual (over 45 pack-years, current smoker) who underwent LDCT four times but failed to adhere to the recommended time frame for the first three scans, and a 70-year-old LCS-ineligible individual (under 20 pack-years, former smoker, over

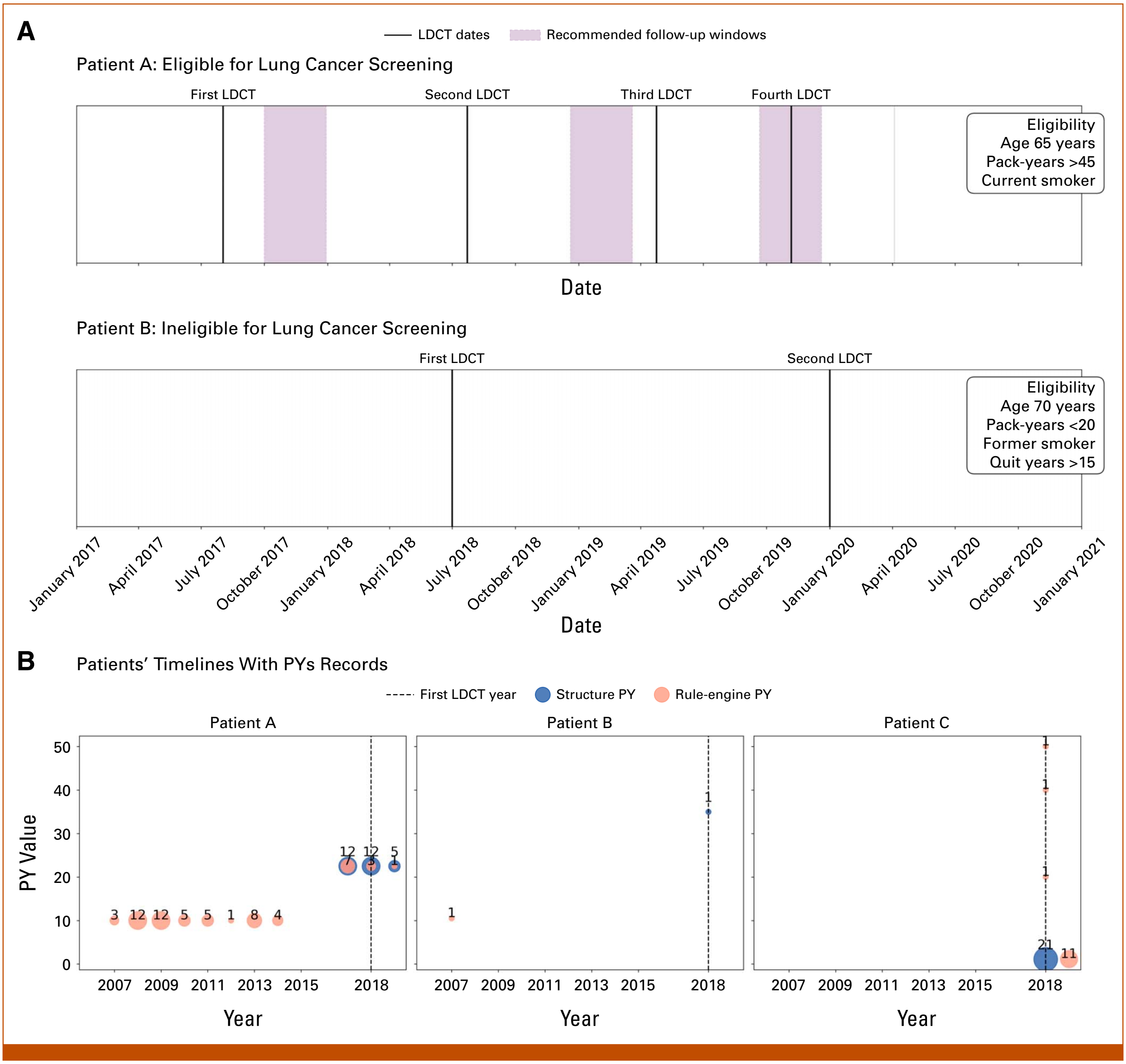

FIG 3. Patient's timelines with LCS or pack-years records. LCS, lung cancer screening; LDCT, low-dose computed tomography; PY, pack-year.

15 quit-years) who inappropriately underwent repeat LDCT. These results indicate a trend of nonadherence among LCS-eligible individuals and inappropriate LDCT use among ineligibles, indicating a deviation from LCS recommendation and follow-up guidelines and inefficient use of health care resources. This analysis underscored the need to perform a causal study on why eligible individuals frequently disregard medical advice, which could enhance adherence and optimize health care resource use.

Our study has several limitations. First, our CP rules do not consider comorbidities, which may affect individuals' eligibility for LCS. Second, LCS eligibility criteria evolve, and our rule-based CP may not fully adapt to these changes. Third, considering both structured and unstructured data reduces specificity compared with those using structured data alone. This is due to NLP tools extracting more information but also introducing errors, which enhances sensitivity but also increases the false-positive rate. Fourth, data quality issues such as data inconsistency affect the accuracy of our CP rules. Figure 3B highlights timelines with inconsistent pack-years records. Patient A's records show fluctuating pack-years, with structured data indicating 10 pack-years from 2007 to 2014, which then changes to 22.5 pack-years in both structured and unstructured data since 2017. Patients B and C exhibit missing and inconsistent pack-year data, leading to questions about data reliability. Patient B's records switch from 10 pack-years in 2007 to 35 in 2018, while Patient C has inconsistent records ranging from 1.12 to 40 pack-years in 2018.

To conclude, we developed a set of CPs on the basis of the USPSTF 2013 and 2020 guidelines that effectively identify individuals eligible for LDCT LCS. These CPs have shown potential as a reliable tool for accurately assessing eligibility for LDCT screening. We also provided a demonstration algorithm illustrating the application of these CPs in clinical settings.

AFFILIATIONS

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL
[2]Cancer Informatics Shared Resource, University of Florida Health Cancer Center, Gainesville, FL
[3]Division of Pulmonary, Critical Care, and Sleep Medicine, University of Florida College of Medicine, Gainesville, FL
[4]Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL
[5]Department of Epidemiology, College of Public Health and Health Professions, University of Florida, Gainesville, FL
[6]Department of Aging and Geriatric Research, College of Medicine, University of Florida, Gainesville, FL
[7]WW HEOR-US Markets, Bristol Myers Squibb, Lawrenceville, NJ
[8]Section of Cancer Economics and Policy, Department of Health Services Research, The University of Texas MD Anderson Cancer Center, Houston, TX

CORRESPONDING AUTHOR

Jiang Bian, PhD; e-mail: bianjiang@ufl.edu.

EQUAL CONTRIBUTION

S.Y. and Y.H. contributed equally to this work.

SUPPORT

Supported by the National Cancer Institute (NCI) grants R01CA246418 and 3R01CA246418-02S1. Y.G. and J.B. were funded in part by the NCI grants R01CA284646, R21CA245858-01, 3R21CA245858-01 A1S1, and 1R21CA253394-01A1, National Institute on Aging grant 1R01AG080624-01, 5R21AG068717-02, and Centers for Disease Control and Prevention grant U18DP006512.

DATA SHARING STATEMENT

A data sharing statement provided by the authors is available with this article at DOI https://doi.org/10.1200/CCI.24.00139.

AUTHOR CONTRIBUTIONS

Conception and design: Yu Huang, Ruoqi Wei, Hiren J. Mehta, Michelle Alvarado, Dejana Braithwaite, Jinhai Huo, Ya-Chen Tina Shih, Yi Guo, Jiang Bian
Financial support: Jiang Bian
Administrative support: Jiang Bian
Collection and assembly of data: Ruoqi Wei
Data analysis and interpretation: Shuang Yang, Yu Huang, Xiwei Lou, Tianchen Lyu, Ruoqi Wei, Hiren J. Mehta, Yonghui Wu, Ramzi G. Salloum, Jinhai Huo, Ya-Chen Tina Shih
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.
Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Jinhai Huo
Employment: Janssen, Bristol Myers Squibb
Stock and Other Ownership Interests: Johnson & Johnson/Janssen

Ya-Chen Tina Shih
Consulting or Advisory Role: Sanofi

Yi Guo
Consulting or Advisory Role: Merck

No other potential conflicts of interest were reported.

REFERENCES

1. Siegel RL, Miller KD, Jemal A: Cancer statistics, 2020. CA Cancer J Clin 70:7-30, 2020
2. Sharma D, Newman TG, Aronow WS: Lung cancer screening: History, current perspectives, and future directions. Arch Med Sci 11:1033-1043, 2015
3. National Lung Screening Trial Research Team; Aberle DR, Adams AM, et al: Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 365:395-409, 2011
4. Moyer VA; US Preventive Services Task Force: Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 160:330-338, 2014
5. Screening for Lung Cancer with Low Dose Computed Tomography (LDCT). 2024. https://www.cms.gov/medicare-coverage-database/view/ncacal-decision-memo.aspx?proposed=N&ncaid=304
6. Wender R, Fontham ETH, Barrera E Jr, et al: American Cancer Society lung cancer screening guidelines. CA Cancer J Clin 63:107-117, 2013
7. Wood DE, Eapen GA, Ettinger DS, et al: Lung cancer screening. J Natl Compr Canc Netw 10:240-265, 2012
8. Begnaud A, Hall T, Allen T: Lung cancer screening with low-dose CT: Implementation amid changing public policy at one health care system. Am Soc Clin Oncol Educ Book 35:e468-e475, 2016
9. US Preventive Services Task Force; Krist AH, Davidson KW, et al: Screening for lung cancer: US Preventive Services Task Force recommendation statement. JAMA 325:962-970, 2021
10. Wood DE, Kazerooni EA, Aberle D, et al: NCCN Guidelines® Insights: Lung cancer screening, version 1.2022. J Natl Compr Canc Netw 20:754-764, 2022
11. Reference deleted
12. Caverly TJ, Fagerlin A, Wiener RS, et al: Comparison of observed harms and expected mortality benefit for persons in the Veterans Health Affairs lung cancer screening demonstration project. JAMA Intern Med 178:426-428, 2018
13. Yang S, Shih YCT, Huo J, et al: Procedural complications associated with invasive diagnostic procedures after lung cancer screening with low-dose computed tomography. Lung Cancer 165:141-144, 2022
14. Huo J, Xu Y, Sheu T, et al: Complication rates and downstream medical costs associated with invasive diagnostic procedures for lung abnormalities in the community setting. JAMA Intern Med 179:324-332, 2019

15. Rendle KA, Saia CA, Vachani A, et al: Rates of downstream procedures and complications associated with lung cancer screening in routine clinical practice: A retrospective cohort study. Ann Intern Med 177:18-28, 2024
16. Benzaquen J, Boutros J, Marquette C, et al: Lung cancer screening, towards a multidimensional approach: Why and how? Cancers 11:212, 2019
17. Mo H, Thompson WK, Rasmussen LV, et al: Desiderata for computable representations of electronic health records-driven phenotype algorithms. J Am Med Inform Assoc 22:1220-1230, 2015
18. Sherman RE, Anderson SA, Dal Pan GJ, et al: Real-world evidence—What is it and what can it tell us? N Engl J Med 375:2293-2297, 2016
19. Real-World Evidence. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
20. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine: Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2. Washington, DC, National Academies Press, 2015
21. Bottorff JL, Seaton CL, Lamont S: Patients' awareness of the surgical risks of smoking: Implications for supporting smoking cessation. Can Fam Physician 61:e562-e569, 2015
22. Chaiton M, Diemert L, Cohen JE, et al: Estimating the number of quit attempts it takes to quit smoking successfully in a longitudinal cohort of smokers. BMJ Open 6:e011045, 2016
23. Polubriaginof F, Salmasian H, Albert DA, et al: Challenges with collecting smoking status in electronic health records. AMIA Annu Symp Proc 2017:1392-1400, 2017
24. Wang Y, Chen ES, Pakhomov S, et al: Investigating longitudinal tobacco use information from social history and clinical notes in the electronic health record. AMIA Annu Symp Proc 2016: 1209-1218, 2016
25. Savova GK, Kipper-Schuler KC, Hurdle JF, et al: Extracting information from textual documents in the electronic health record: A review of recent research. Yearb Med Inform 17:128-144, 2008
26. US. Preventive Services Task Force Issues Final: Recommendation Statement on Screening for Lung Cancer. https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/lung-cancer-screening
27. Yang X, Bian J, Hogan WR, et al: Clinical concept extraction using transformers. J Am Med Inform Assoc 27:1935-1942, 2020
28. Yang X, Yang H, Lyu T, et al: A natural language processing tool to extract quantitative smoking status from clinical narratives. medRxiv 10.1101/2020.10.30.20223511
29. University of Florida Clinical and Translational Science Institute: Integrated Data Repository. https://www.ctsi.ufl.edu/category/data-informatics-and-it/integrated-data-repository-idr/
30. Buderer NM: Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. Acad Emerg Med 3:895-900, 1996
31. Reference deleted
32. Reference deleted
33. Ruckdeschel JC, Riley M, Parsatharathy S, et al: Unstructured data are superior to structured data for eliciting quantitative smoking history from the electronic health record. JCO Clin Cancer Inform 10.1200/CCI.22.00155
34. Petrik AF, Green BB, Vollmer WM, et al: The validation of electronic health records in accurately identifying patients eligible for colorectal cancer screening in safety net clinics. Fam Pract 33: 639-643, 2016
35. Leder Macek AJ, Kirschenbaum JD, Ricklan SJ, et al: Validation of rule-based algorithms to determine colorectal, breast, and cervical cancer screening status using electronic health record data from an urban healthcare system in New York City. Prev Med Rep 24:101599, 2021
36. Liu S, McCoy AB, Aldrich MC, et al: Leveraging natural language processing to identify eligible lung cancer screening patients with the electronic health record. Int J Med Inform 177:105136, 2023