Journal of the American Medical Informatics Association, 29(5), 2022, 779–788
https://doi.org/10.1093/jamia/ocac020
Advance Access Publication Date: 15 February 2022
Research and Applications

Research and Applications

Inaccuracies in electronic health records smoking data and a potential approach to address resulting underestimation in determining lung cancer screening eligibility

Polina V. Kukhareva [1], Tanner J. Caverly [2,3,4], Haojia Li[5], Hormuzd A. Katki[6], Li C. Cheung[6], Thomas J. Reese [7], Guilherme Del Fiol[1], Rachel Hess[5,8], David W. Wetter[11], Yue Zhang[5], Teresa Y. Taft[1], Michael C. Flynn[9,10,12], and Kensaku Kawamoto [1]

[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA, [2]Center for Clinical Management Research, Department of Veterans Affairs, Ann Arbor, Michigan, USA, [3]Department of Learning Health Sciences, University of Michigan, Ann Arbor, Michigan, USA, [4]Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA, [5]Department of Population Health Sciences, University of Utah, Salt Lake City, Utah, USA, [6]Division of Cancer Epidemiology & Genetics, National Cancer Institute, Bethesda, Maryland, USA, [7]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA, [8]Department of Internal Medicine, University of Utah, Salt Lake City, Utah, USA, [9]Department of Pediatrics, University of Utah, Salt Lake City, Utah, USA, [10]Community Physicians Group, University of Utah Health, Salt Lake City, Utah, USA, [11]Department of Population Health Sciences and Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah, USA, and [12]Community Physicians Group, University of Utah, Salt Lake City, UT, USA

Corresponding Author: Polina V. Kukhareva, PhD, MPH, Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Suite 108, Salt Lake City, UT 84108, USA; polina.kukhareva@utah.edu

Received 9 November 2021; Revised 28 January 2022; Editorial Decision 31 January 2022; Accepted 1 February 2022

ABSTRACT

Objective: The US Preventive Services Task Force (USPSTF) requires the estimation of lifetime pack-years to determine lung cancer screening eligibility. Leading electronic health record (EHR) vendors calculate pack-years using only the most recently recorded smoking data. The objective was to characterize EHR smoking data issues and to propose an approach to addressing these issues using longitudinal smoking data.

Materials and Methods: In this cross-sectional study, we evaluated 16 874 current or former smokers who met USPSTF age criteria for screening (50–80 years old), had no prior lung cancer diagnosis, and were seen in 2020 at an academic health system using the Epic® EHR. We described and quantified issues in the smoking data. We then estimated how many additional potentially eligible patients could be identified using longitudinal data. The approach was verified through manual review of records from 100 subjects.

Results: Over 80% of evaluated records had inaccuracies, including missing packs-per-day or years-smoked (42.7%), outdated data (25.1%), missing years-quit (17.4%), and a recent change in packs-per-day resulting in inaccurate lifetime pack-years estimation (16.9%). Addressing these issues by using longitudinal data enabled the identification of 49.4% more patients potentially eligible for lung cancer screening ($P < .001$).

Discussion: Missing, outdated, and inaccurate smoking data in the EHR are important barriers to effective lung cancer screening. Data collection and analysis strategies that reflect changes in smoking habits over time could improve the identification of patients eligible for screening.

Conclusion: The use of longitudinal EHR smoking data could improve lung cancer screening.

Key words: lung cancer screening, self-reported smoking history, electronic health records, lung cancer screening eligibility, pack-years

INTRODUCTION

Lung cancer is the leading cause of cancer deaths in the United States.[1] Lung cancer screening with low-dose computed tomography could result in a 20% relative reduction in lung cancer mortality.[2] In March 2021, the US Preventive Services Task Force (USPSTF) recommended that annual low-dose lung cancer screening be offered to patients 50–80 years old with a 20+ pack-year smoking history who are current smokers or quit in the last 15 years.[3] Despite potential benefits, lung cancer screening rates remain at about 5% in the United States among individuals meeting USPSTF screening eligibility criteria.[4]

Electronic health record (EHR) systems could help improve lung cancer screening rates by enabling the collection and use of detailed, discrete data on smoking history. Such data can be used to automate determination of USPSTF eligibility and inform high-quality shared decision-making based on individual-specific benefits and risks from screening.[3,5] In addition, these data can be used to help identify patients who are particularly good candidates for screening due to their individual benefit and risk profile.[6,7]

Key smoking data collected in EHRs include the number of packs smoked per day, years smoked, and smoking quit date.[8–11] For example, the vendors Epic and Cerner, which held a combined 58% ambulatory EHR market share in 2019,[12] use the most recently recorded packs-per-day and years-smoked to calculate lifetime smoking exposure. In the default configuration for Epic, lifetime smoking exposure is calculated using the most recently recorded packs-per-day and years-smoked as "pack-years" (Figure 1). Similarly, Cerner calculates "total pack-years" from the most recent record.[13] Thus, while some EHRs hold over 25 years of longitudinal smoking data, these data are oftentimes not being used to identify patients eligible for lung cancer screening.

Previous studies have reported on the low quality of the most recent smoking data in EHRs.[8,14–18] One retrospective study found 96.2% discordance in pack-year smoking history between the EHR and data obtained through a shared decision-making conversation.[14] In a qualitative study of primary care physicians, some providers expressed lack of trust in EHR smoking data and perceived smoking documentation in the EHR as inaccurate or insufficient for determining whether to order low-dose computed tomographies.[19] Another study evaluated the impact of random error in pack-year estimation.[17] However, these studies did not quantify the potential impact of such data issues nor propose an approach to overcoming such issues using existing longitudinal data.

OBJECTIVE

Our objectives were to address both limitations by (1) characterize EHR smoking data issues, and (2) propose a potential approach to addressing these issues using longitudinal EHR smoking data.

METHODS

This was a cross-sectional study of University of Utah (UU) Health patients 50–80 years old with a history of smoking. This study was approved by the UU Institutional Review Board.

UU Health is comprised of 5 hospitals and 11 community health care centers providing inpatient and ambulatory care across various specialties including primary care, cancer care, and pulmonary care. UU Health has 12 primary care clinics, and all primary care clinics have used the Epic© EHR system since 1999. Smoking data for some patients were electronically collected since 1995.

Patients were included based on the following criteria: (1) in-person or virtual provider visit in a study clinic in 2020; and (2) age 50–80 years on December 31, 2020. Patients were excluded based on the following criteria: (1) had a lung cancer diagnosis before January 1, 2020, (2) had no smoking status recorded, or (3) never smoked according to the EHR data. Lung cancer diagnosis was determined based on Epic diagnosis codes in the problem list, medical history, and visit diagnoses (Supplementary Appendix Box 1).

Figure 1. Smoking history collection form.

Box 1 Logic of the Baseline Approach

The Baseline Approach is as follows:

1. *Pack-years* are calculated by multiplying the *years-smoked* and *packs-per-day* most recently recorded in the EHR.
2. For former smokers, *years-quit* are calculated as the years that have passed since the quit date most recently recorded in the EHR.
3. The following assumptions are used:
   a. The most recently recorded *years-smoked* is accurate.
   b. The most recently recorded *packs-per-day* reflects the average packs-per-day smoked over the individual's entire smoking history as opposed to the packs-per-day currently smoked.
   c. If the most recent record lacks a needed data point, there is no relevant data available. For example, even if it is documented that a patient transitioned from a current smoker to a former smoker 3 years ago, the years-quit is deemed unknown if the field is not populated in the most recent record.

"Never smoker" status was assigned to patients who reported that they never smoked at every visit when the smoking history was recorded. An individual who had previously smoked, but who was not documented as such in the EHR, would have been missed and represent a false negative.

Age, gender, race, ethnicity, and smoking history data were extracted from the EHR on September 2, 2021. Due to the elevated risk of lung cancer among Black patients, the study race was set to Black if at least one race recorded in the EHR was Black.[20]

We used smoking data entered in the EHR in designated structured fields. In the default configuration for Epic used by UU Health, smoking history is usually recorded by the medical assistant or provider in 4 fields: smoking status, packs-per-day, years-smoked, and smoking quit date (Figure 1). Smoking status is a drop down. Years-smoked and packs-per-day fields allow entering free text. Smoking data required some data cleaning before use. For example, we replaced '20+' with 20, and '0.5' with '0.5'. An additional field for smoking start date is available for patients to complete through the patient portal, but the field is usually not populated. Every time smoking history is documented in the EHR, a new smoking history record is created. The most recent smoking data are displayed on the screen (Figure 1). Longitudinal smoking data can be accessed by clicking a link to the "Audit Trail" in the History Tab for "Substance and Sexual Activity."

For screening-eligible patients, we estimated their 10-year risk of developing lung cancer and estimated life-expectancy using equations developed by Bach et al.[21,22] Then, based on the work of Caverly et al[6] and Mazzone et al,[7] we classified patients as being particularly 'high-benefit' patients with regard to lung cancer screening if they had a 10-year risk of developing lung cancer $\geq$5.2% and an estimated life-expectancy >10 years.

Baseline, Longitudinal, and Combined Approaches

Both EHR vendors and external clinical decision support vendors usually use the most recently documented years-smoked, packs-per-day, and years-quit to assess for screening eligibility (Box 1).[19] For example, this Baseline Approach was used by UU Health to generate care reminders and to support shared decision-making for lung cancer screening.[19,23]

To address the issues identified with the smoking data, as well as the presumably inaccurate assumptions used by the Baseline Approach, we developed a Longitudinal Approach using longitudinal smoking data. The main purpose of the Longitudinal Approach was to leverage longitudinal EHR data to identify individuals who may be eligible for lung cancer screening but were deemed to be ineligible using the Baseline Approach. Figure 2 illustrates the Baseline and Longitudinal Approaches. The logic of the Longitudinal Approach is described in Supplementary Box 2 of the Appendix. The main assumptions of the Longitudinal Approach are that (1) current smokers continue to accumulate smoking exposure and that (2) the packs-per-day recorded at a given point in time reflects the packs-per-day the individual was smoking at that time.

To identify as many patients as possible who may be eligible for lung cancer screening based on available EHR data, we combined the Baseline and Longitudinal Approaches into a Combined Approach that identifies patients eligible for screening through either approach.

Quantification of smoking data issues, iterative development of Longitudinal Approach, and manual verification of approach

To characterize smoking data issues and to develop a Longitudinal Approach to address these issues, the following steps were taken.

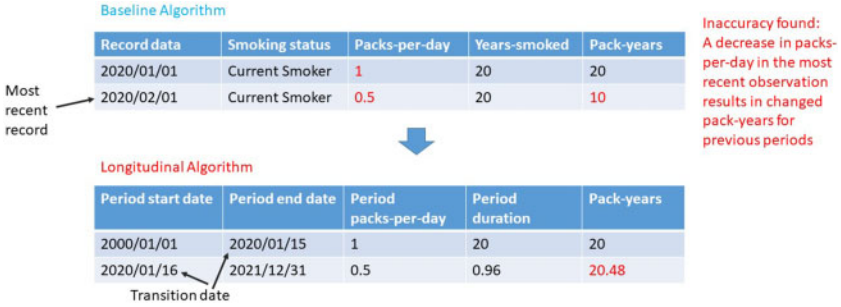

Figure 2. Example of using Baseline and Longitudinal Approaches for patient who smoked 1 pack-per-day for 20 years and then switched to smoking 0.5 packs-per-day.

First, a sample EHR smoking records were reviewed to identify patterns of data issues and an initial Longitudinal Approach was developed to address these issues. Then, the eligibility of study patients based on USPSTF criteria was determined using both the Baseline and Longitudinal Approaches, and patients with discrepancies in eligibility determinations were selected for manual review so as to focus on cases of potential clinical significance. During this review, the patient's longitudinal smoking history was reviewed to identify additional data issues not yet accounted for in the Longitudinal Approach. To facilitate the review, data records that were identical across visits were merged into a single record (eg, records from 10 visits spanning March 1, 2016 to July 15, 2019 all noting that the patient was a 20 years, 1 pack-per-day smoker were collapsed into a single record spanning these dates). Finally, additional identified data issues were addressed by updating the Longitudinal Approach. These last 2 steps of data issue identification and approach enhancement were iteratively repeated until no further data issues were identified following a review of 51 patient record.

Following the development of the Longitudinal Approach addressing all identified issues, the appropriateness of the Longitudinal Approach was manually verified. In this verification process, PVK and KK conducted independent manual review of the smoking-related EHR records of 100 randomly selected patients with available smoking data. The reviewers manually assessed the patient's lifetime tobacco exposure using the Longitudinal Approach and independently determined whether the patient met USPSTF eligibility criteria. Discrepancies between the 2 reviewers were adjudicated through discussion, and all discrepancies between reviewers were identified as resulting from an error in manual application of the approach by one of the reviewers. No new data issues were identified during this process. Finally, the prevalence of each data issue was quantified.

Statistical analysis

All statistical analysis was performed using R version 4.1.0. We used n (%) and mean (standard deviation) to describe nominal and continuous variables. We used bootstrapping to calculate the 95% confidence intervals (CIs) of the relative changes and the absolute differences in average pack-years and years-quit. McNemar's Chi-squared test was used for logical variables and Wilcoxon signed rank test was used for numeric variables to compare algorithm accuracy and patient eligibility for lung cancer screening according to the Baseline and Longitudinal Approaches. We performed a sensitivity analysis in which we removed from the Longitudinal Approach the assumption that the packs-per-day recorded at a given point in time reflects the packs-per-day the individual was smoking at the time. Given that 2020 data could have been affected by the disturbances from coronavirus disease-19 (COVID-19), we repeated the analysis using 2019 data. To visualize the relationship between pack-years estimated by the approaches, we used a local polynomial regression (R ggplot2 package, geom_smooth function, "loess" method). An analysis stratified by race/ethnicity and gender was conducted to ensure fairness.

RESULTS

Patient characteristics

Patient inclusion and exclusion criteria are shown in Figure 3. In total, 16 874 patients met study criteria. About 35% of patients smoked at some point in their life. Patient characteristics for patients meeting inclusion criteria are summarized in Table 1. Among included patients, about 25% of patients were current smokers and 75% were former smokers.

Smoking data issues

Among 16 874 patients meeting eligibility criteria, the EHR data contained 12 types of issues that could cause errors in calculating pack-years and years-quit using the Baseline Approach (Table 2). Over 80% of patient records had at least one such data issue. The 3 most common issues were missing data, stale data, and changes in packs-per-day that affected all the previous years. A given patient could have more than one data issue, including issues that are specific subsets of more general issues.

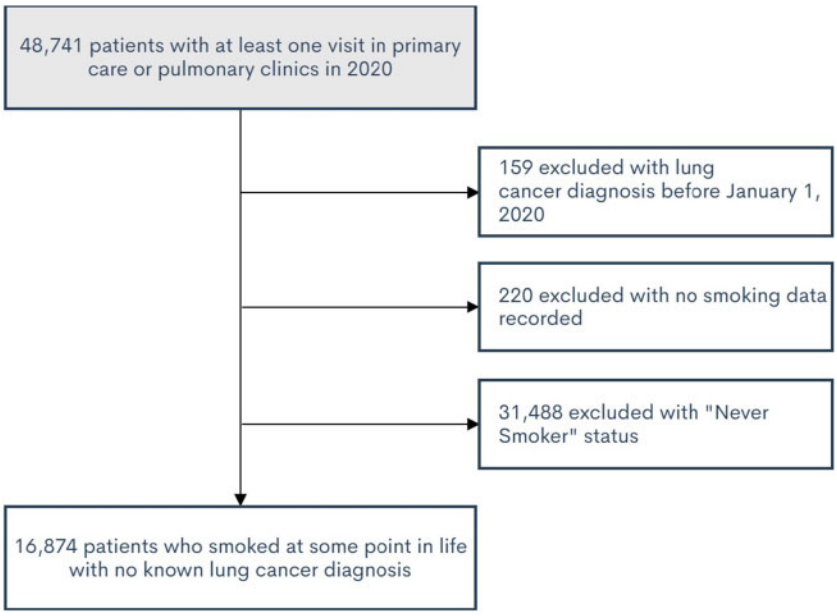

Figure 3. Patient flow through inclusion and exclusion criteria.

Table 1. Patient characteristics

Patient characteristics ($N = 16\,874$)	N (%)
Age	
50–59	5926 (35.1%)
60–69	6438 (38.2%)
70–80	4510 (26.7%)
Female gender	8057 (47.7%)
Race/ethnicity	
Non-Hispanic White	13 230 (78.4%)
Non-Hispanic Black/African American	324 (1.9%)
Hispanic	1822 (10.8%)
Other race/ethnicity[a]	1498 (8.9%)
Smoking status based on last record	
Current smoker	4151 (24.6%)
Former smoker	12 723 (75.4%)

[a]Other race/ethnicity includes non-Hispanic participants with race other than White or Black or those who chose not to disclose race.

Lung cancer screening eligibility determination by Baseline and Longitudinal Approaches

The Longitudinal Approach was verified as being implemented as intended, with the computationally implemented Longitudinal Approach having a sensitivity of 0.97 (95% CI: 0.92, 1) and specificity of 1 (95% CI: 1, 1) with regard to USPSTF eligibility classification in comparison to the manual application of the approach for the 100 randomly selected patients. Interrater reliability for this determination, as calculated using Cohen's kappa, was high (0.91).

Figure 4 presents patient eligibility diagram based on USPSTF 2021 eligibility criteria using Baseline and Longitudinal Approaches. Baseline Approach identified 2228 patients and the Longitudinal Approach identified 3167 patients as eligible for lung cancer screening.

Figure 5 depicts a Venn diagram demonstrating how many patients were identified by just one approach. A total of 2066 individuals were identified as eligible by both approaches. The Longitudinal Approach enabled identification of 1101 screening-eligible individuals in addition to the 2066 individuals identified by both algorithms. These additional records included 271 patients who were missing required data (packs-per-day, years-smoked, and/or years quit) in the most recent observation. The rest were missed by the Baseline Approach due to the other inaccuracies described in Table 2. Conversely, the Baseline Approach identified 162 patients as being eligible whom the Longitudinal Approach did not consider to be eligible. One hundred forty-one (87%) of these 162 patients had a recent increase in packs-per-day. For example, after reporting smoking 0.5 packs-per-day for 30 years, they reported smoking 1 pack-per-day, which would make them eligible according to the Baseline Approach, but not the Longitudinal Approach. Since there are conceivable situations were both algorithms could be correct, it could be reasonable to use both algorithms in a Combined Approach.

Figure 6 depicts the pack-year discrepancies between the 2 algorithms in a scatter plot. The dark vertical lines indicate data patterns related to EHR-recorded pack-years remaining static over multiple years at increments such as 10, 20, and 30 pack-years rather than incrementing upwards with continued tobacco exposure. The regression line shows that the Longitudinal Approach provides higher pack-year values compared to the Baseline Approach in the 0–50 pack-years range, but lower values for the higher pack-year range.

The sensitivity analysis evaluated the impact of removing the Longitudinal Approach's assumption that the packs-per-day recorded at a given point in time reflected the packs-per-day the individual was smoking at the time. Even with this assumption removed, the Longitudinal Approach identified a significantly greater number of patients as being eligible for screening (2758 vs. 2228, $P < .001$). In addition to 2115 individuals identified by both algorithms, the Longitudinal Approach identified 643 additional individuals and the Baseline Approach identified 113 additional individuals.

Additional patients identified using the Combined Approach

Using the Combined Approach, 3329 patients were identified as potentially meeting USPSTF eligibility criteria for lung cancer screening, which was 1101 (49.4% [95% CI: 46%, 53%]) more patients than the 2228 patients identified using the Baseline Approach alone ($P < .001$) (Table 3). This estimate included 399 (40.4% [95% CI: 36%, 45.2%]) more high-benefit patients. Among current smokers with sufficient data to estimate pack-years, the Baseline Approach underestimated a patient's pack-years by an average of 7.8 [95% CI: 6.7, 8.8] years compared to the Longitudinal Approach.

Given that 2020 data could have been affected by the disturbances from COVID-19, we repeated the analysis using 2019 data (Supplementary Appendix Table 1). All the conclusions of the study held for 2019 data.

Fairness analysis

Supplementary Appendix Table 1 shows results of the fairness analysis stratified by race/ethnicity and gender. Using longitudinal data identified significantly more patients across race/ethnicity and gender as potentially eligible for lung cancer screening compared to using the Baseline Approach. Using the longitudinal data might be especially beneficial for Hispanic patients and women. The relative increase in the identification of potentially eligible patients using the Combined Approach was 75.4% (95% CI: 57.5%, 96.1%) for Hispanic patients, 47.1% (95% CI: 43.6%, 50.8%) for White patients, 56.8% (95% CI: 50.8%, 63.4%) for female patients, and 43.7% (95% CI: 39.3%, 48%) for male patients.

DISCUSSION

We quantified issues in EHR smoking data at a single site. Over 80% of patient records had at least one issue, including missing, outdated and inaccurate data. To partially address these issues, we developed an approach that uses longitudinal smoking data. This Longitudinal Approach can help identify patients eligible for lung cancer screening who would be missed by the Baseline Approach, which uses only the most recent EHR data and is the predominant algorithm used by market-leading EHR systems.

Among 16 874 current and former smokers with no prior history of lung cancer, the Combined Approach leveraging both most recent and longitudinal data was able to identify 49.4% (95% CI: 46%, 53%) more patients potentially eligible for lung cancer screening than the Baseline Approach ($P < .001$). This included identifying 40.4% (95% CI: 36%, 45.2%) more high-risk, high-benefit patients ($P < .001$). Screening is particularly important for this high-benefit population and misclassifying such patients as ineligible is particularly disconcerting. If implemented in clinical practice, the Combined Approach could substantially increase the number of individuals who are evaluated for lung cancer screening.

This study has several implications. First, this study underscores the critical need to improve the *collection* of smoking history in the EHR in order to improve lung cancer screening. In our patient popu-

Table 2. Data issues that could affect Baseline Approach

Data issue that could affect Baseline Approach[a]	**Prevalence of issue, $N = 16\,874$**	Illustrative example	How issue is addressed in Longitudinal Approach
1. Insufficient data to calculate pack-years from the most recent smoking record	7204 (42.7%)	While patient is recorded as smoker, no information on packs-per-day or smoking duration is present in the most recent smoking record.	Pack-per-day are calculated from the first available observation. If never available, pack-per-day are imputed as the median value for the sample (0.5 packs-per-day). If years-smoked are missing from the most recent observation, the algorithm uses data from previous observations, with appropriate increase in duration with the passage of time.
2. Insufficient data to calculate pack-years from the longitudinal records	6708 (39.8%)	While patient is recorded as smoker, no information on packs-per-day or smoking duration is present in the longitudinal records.	Longitudinal Approach does not address this issue.
3. Pack-years did not change for over a year when a patient was a current smoker	4235 (25.1%)	Patient remained 20-year smoker for the past 15 years.	Instead of using years-smoked to calculate pack-years, years-smoked is used to calculate the start date of the first smoking period, and pack-years are calculated based on duration of the period.
4. Unknown smoking quit date for a former smoker	2942 (17.4%)	A patient was recorded as a smoker, then was recorded as being a former smoker, but the quit date was not entered.	When smoking date is missing for a former smoker, we use the transition date as the quit date.
5. An increase or decrease in packs-per-day in the most recent observation results in changed pack-years for previous periods	2852 (16.9%)	A patient smoked 1 pack-per-day for 20 years, then cut down to 0.5 packs-per-day, resulting in pack-years decreasing from 20 to 10.	Decreases and increases in packs-per-day do not affect results for previous periods. This assumption was tested in the sensitivity analysis.
6. Pack-years decreased over time (caused by decreased years-smoked or packs-per-day)	2001 (11.9%)	A patient originally said that they smoked for 30 years, then reported later that they smoked for 20 years.	Period-pack-years are added over time to the first period pack-years.
7. Smoking quit date changed to an earlier date	711 (4.2%)	A patient was documented as having quit smoking on January 01, 2010, but the most recent observation overrode the quit date with January 01, 2000.	The latest smoking quit date on record is used to determine how long a patient has been a former smoker.
8. Patients became a 'never smoker' after being a current or former smoker	512 (3%)	A patient was recorded as being smoker for a year, then subsequently was recorded as being a never smoker.	If detailed smoking information is available, it is used to estimate tobacco exposure.
9. New smoking quit date not recorded after patient quit smoking repeatedly	234 (1.4%)	Patient quit smoking on January 01, 1999, then restarted smoking and quit again on January 01, 2009. However, January 01, 1999 was the only quit date recorded.	If the recorded quit date is before the last known smoking date, the last known transition from smoking to not smoking is used to determine the quit date.
10. The duration between recorded smoking start and end dates did not correspond to recorded years-smoked	213 (1.3%)	A patient was documented as having started smoking on January 01, 2000 and having quit smoking on January 01, 2010. However, the smoking duration was recorded as 20 years.	The start date of the first period is calculated as the earliest date across all records indicating when the patient started smoking.
11. Recorded as started smoking before age 5	91 (0.5%)	A patient was recorded as starting smoking when 3 years old.	For patients reported to start smoking before the age of 5, the smoking start date was moved to the 5th birthday.
12. Recorded as smoking over 5 packs per day	36 (0.2%)	A patient was recorded as smoking 10 packs-per-day for 20 years.	Packs-per-day >5 are divided by 20, with an assumption that cigarettes-per-day were mistakenly entered as packs-per-day.
Any of above	13 833 (82%)		

[a]One record can have more than one issue.

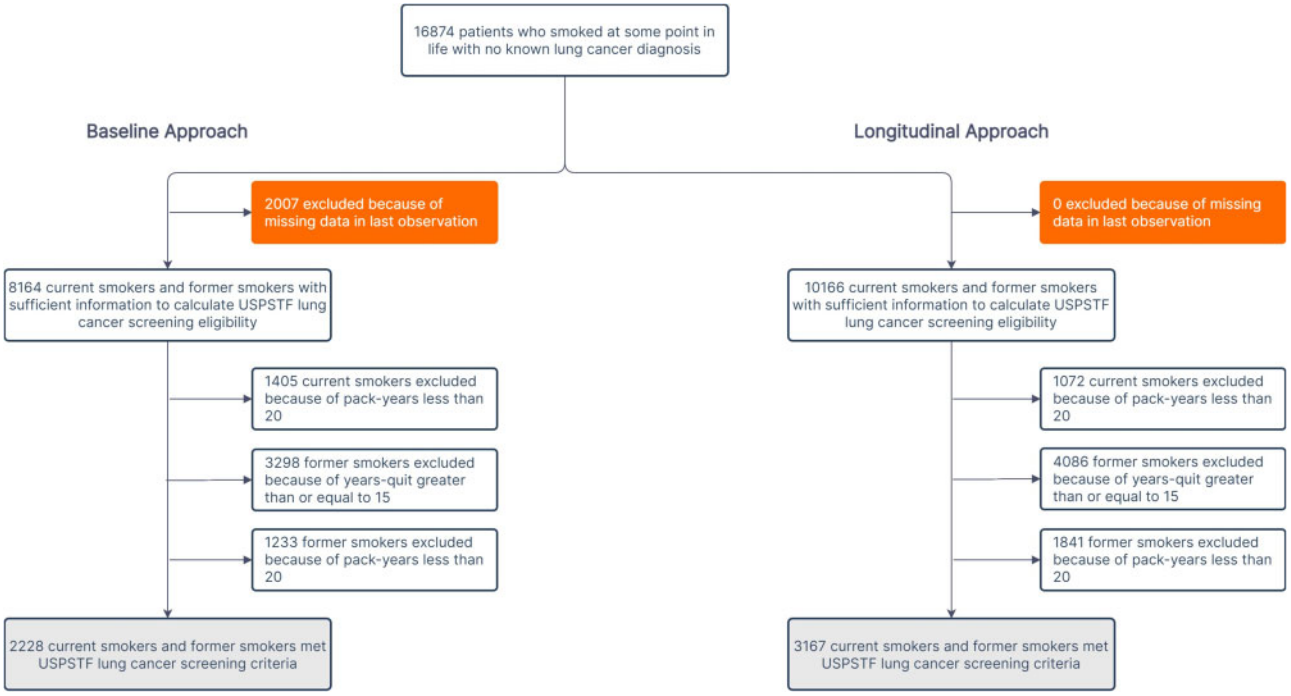

Figure 4. Diagram of patient eligibility for lung cancer screening according to the Baseline and Longitudinal Approaches.

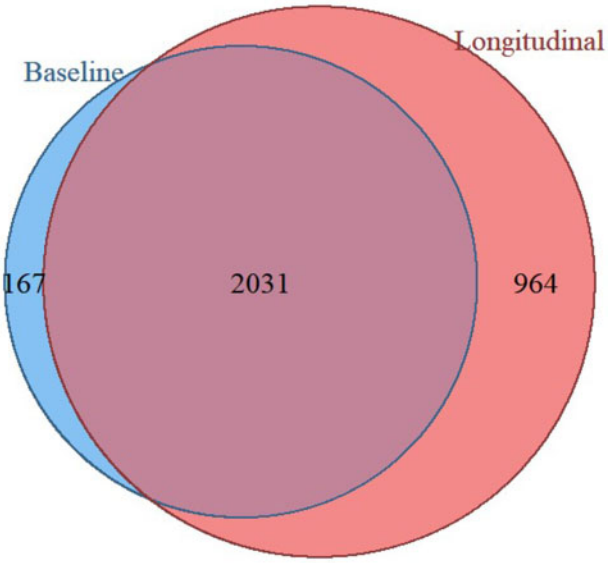

Figure 5. Number of eligible patients identified using Baseline and Longitudinal Approaches.

lation, only about 60% of ever-smokers in the USPSTF-eligible age range had the requisite smoking data recorded in the EHR to determine screening eligibility. Moreover, we found that over 80% of patients with the requisite smoking data had issues in that data. This is consistent with others' findings, such as a study by Modin et al[14] which found that pack-year history calculated using the most recent EHR data were almost always different from pack-year history obtained through clinical interview at a centralized lung cancer screening program. As EHR data are known to be incomplete when it comes to cohort identification,[24] approaches improving data collection should be considered, such as leveraging health information exchanges and directly engaging patients in the collection and review of the needed data.

As a second implication, smoking history should always be *verified* prior to making screening or treatment decisions dependent on smoking history. Clinicians and health IT implementers need to be mindful of the potential impact of missing and inaccurate data when making decisions about lung cancer screening based on EHR smoking data.

As a third implication, the Longitudinal Approach described in this paper add *clinical value* in several ways. As one potential use, the Longitudinal Approach could be used to alert clinicians about patients with significant discrepancy between the results of the Baseline and Longitudinal Approaches, so that the clinicians can verify the smoking history with the patient. For example, when the Baseline Approach estimated higher pack-years than the Longitudinal Approach in our dataset (Figure 6), this often appeared to be a result of cigarettes-per-day being mistakenly entered as pack-per-day by EHR users, which were taken at face value in the Baseline Approach but assumed to be cigarettes-per-day in the Longitudinal Approach (eg, assumed to be 10 cigarettes-per-day as opposed to 10 packs-per-day, which would likely not be possible to achieve). Second, patients eligible for lung cancer screening through *either* the Baseline or Longitudinal Approaches (ie, through the Combined Approach) could be flagged for evaluation for lung cancer screening. This could significantly increase the number of eligible patients who are identified for screening, which is a critical need given that <5% of eligible patients are currently screened in the United States.[4] Third, more individuals with high lifetime smoking exposure could be targeted for smoking cessation outreach, such as through electronic reminders as described by Bar et al.[25] Fourth, the Combined Approach could be used to identify household members with concerning levels of secondhand tobacco exposure. Fifth, the Longitudinal Approach could be used to pre-populate an EHR's smoking history

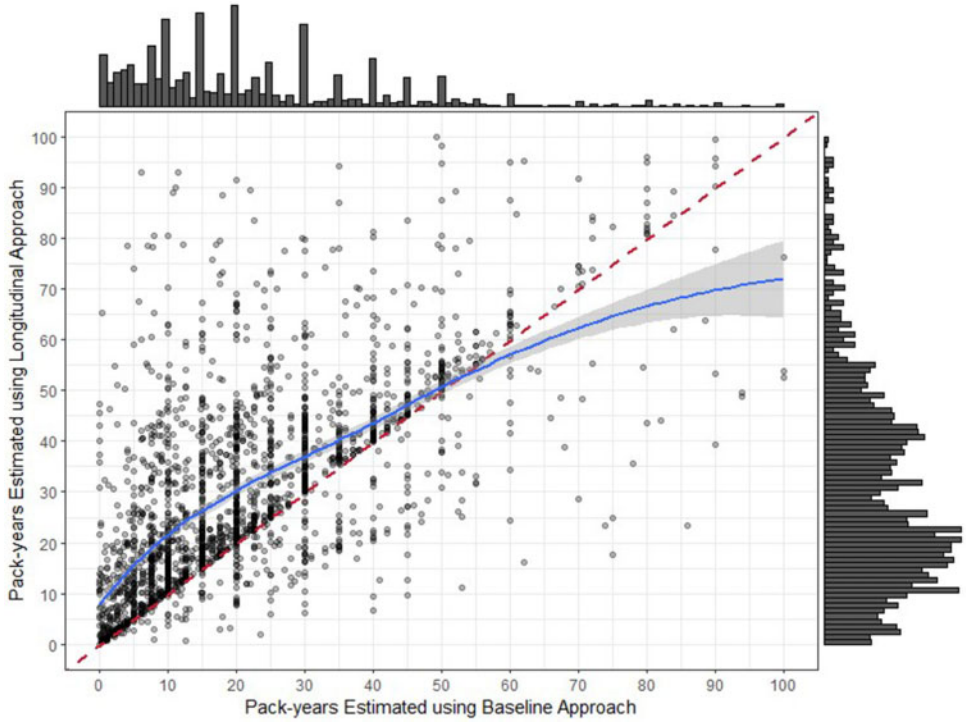

Figure 6. Scatter plot of pack-years estimated using Baseline and Longitudinal Approaches for current smokers. The red line divides the plane in equal parts. The blue line represents the regression line between the pack-years estimated using 2 algorithms fitted by the local polynomial regression model. Forty-seven points are omitted due to pack-years estimated using either of the 2 algorithms larger than 100.

if and when EHR vendors move to a period-based reporting approach (eg, smoked 1 pack-per-day from age 18 to 35, quit from 35 to 40, and then smoked 0.5 packs-per-day since). An important assumption of the Longitudinal Approach was that patients tend to report how much they are currently smoking instead of an average of how much they smoked over their lifetime. A period-based reporting approach would obviate the need for such assumptions, and we are aware of at least one major EHR vendor planning to convert to such an approach.

As a final implication, this study suggests that historical smoking data should ideally be made *accessible* to clinical decision support systems in addition to the most recent smoking data. In particular, it would be ideal if longitudinal smoking data were available from EHRs through their Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) application programming interfaces, so that the data can be used by external apps and Web Services connected to the EHR to identify a patient's eligibility and appropriateness for screening.[19,26]

One of the strengths of this study is that it used a computational approach to quantify issues in EHR smoking data among a large group of patients (ie, all patients with a history of smoking seen at an academic medical center). In comparison, prior studies have only characterized the issue qualitatively or in a much smaller sample of patients,[14,16] or focused on issues arising from random noise rather than bias in the data.[17] As a second strength, this study proposes a computable algorithm to address these issues. Thus, we not only quantify the problem but also provide an actionable approach to addressing the problem. Finally, this study evaluated the accuracy of the Longitudinal Approach in identifying patients considered to be particularly high benefit.[6,7]

This study has several limitations. First, this study did not include a true gold standard for lifetime tobacco exposure. However, due to the evolving nature of smoking habits over time, creating a true gold standard would have required decades-long prospective data collection with frequent questionnaires. When estimating lifetime smoking exposure through a single patient interaction, there are methods to reduce recall bias such as the Timeline Followback.[27] However, it is questionable whether an individual's recollection many decades later is more accurate than what they reported at the time. For example, for a 50-year-old patient who began reporting their smoking habits in the EHR 20 years ago, it is questionable whether their recollection at age 50 of their smoking habits in their 30s or 40s is more accurate than what they reported back when they were that age. Indeed, a 2020 study found that when individuals were prospectively asked to provide a self-reported lifetime smoking history 1 month apart, differences in their recollection just over this 1 month was such that 12% of participants were eligible for lung cancer screening at one but not both assessments.[28] Given the potential inaccuracy of self-recollection for lifetime tobacco exposure, even when done prospectively, we did not seek to establish a gold standard. Instead, we sought to characterize data issues and to develop an approach to help identify as many patients as possible who may be eligible for lung cancer screening based on available EHR data. While use of the Combined Approach may result in a false positive identification of screening-eligible patients as compared to the ground truth, we believe this is a reasonable approach to using EHR data to identify patients for lung cancer screening, as potentially eligible patients can be evaluated by their clinical providers to verify eligibility, whereas patients not flagged as being potentially eligible may easily be overlooked for screening purposes.

Table 3. Patient eligibility for lung cancer screening according to the Baseline and Combined Approaches

	Baseline Approach, $N = 16\,874$	Combined Approach, $N = 16\,874$	Absolute change	Relative change	*P*-value
Sufficient data to calculate pack-years and years-quit	8164 (48.4%)	10 171 (60.3%)	2007 (11.9%)	24.6% (23.3%, 25.9%)	<.001*
Current smoker, sufficient data to calculate pack-years	2592 (15.4%)	2769 (16.4%)	177 (1%)	6.8% (5.8%, 7.9%)	<.001*
Smoked <20 pack-years (did not meet criteria)	1405 (8.3%)	1620 (9.6%)	215 (1.3%)	15.3% (13.2%, 17.5%)	<.001*
Smoked ≥20 pack-years (met criteria)	1187 (7%)	1759 (10.4%)	572 (3.4%)	48.2% (43.4%, 53.2%)	<.001*
Former smoker, sufficient data to calculate pack-years and years-quit	5572 (33%)	7441 (44.1%)	1869 (11.1%)	33.5% (31.6%, 35.6%)	<.001*
Quit ≥15 years ago (did not meet criteria)	3298 (19.5%)	4225 (25%)	927 (5.5%)	28.1% (26%, 30.3%)	<.001*
Smoked <20 pack-years (did not meet criteria)	1233 (7.3%)	2067 (12.2%)	834 (4.9%)	67.6% (61.9%, 73.3%)	<.001*
Smoked ≥20 pack-years (met criteria)	1041 (6.2%)	1577 (9.3%)	536 (3.2%)	51.5% (46.7%, 56.8%)	<.001*
Average pack-years for current smokers with sufficient data for both algorithms[a]	22.7 (27.5)	30.5 (22.2)	7.8 (6.7, 8.8)	34.3% (28.5%, 39.8%)	<.001*
Average pack-years for former smokers with sufficient data for both algorithms[a]	18.1 (23.4)	20.6 (21.5)	2.5 (1.8, 3)	13.8% (9.9%, 17.4%)	<.001*
Average years-quit for former smokers with sufficient data for both algorithms[a]	21.4 (15.3)	20.6 (15.1)	−0.7 (−0.9, −0.6)	−3.3 (−4.3%, −2.8%)	<.001*
Met USPSTF lung cancer screening criteria (combining current and former smokers)	2228 (13.2%)	3329 (19.7%)	1101 (6.5%)	49.4% (46%, 53%)	<.001*
Met USPSTF lung cancer screening criteria, high-benefit population (Bach model)	988 (5.9%)	1387 (8.2%)	399 (2.4%)	40.4% (36%, 45.2%)	<.001*

USPSTF: US Preventive Services Task Force.

[a]Average pack-years and years-quit are calculated for Baseline and Longitudinal Approaches.

*$P < .05$.

As the second limitation, the data analysis was conducted using data from one healthcare system; as such, replication is needed. However, we have no particular reason to believe that our findings are significantly divergent from what would be found in other clinical settings. Of note, our findings are consistent with earlier studies in this area, and the EHR product used at the study site is one of the most widely used EHRs in the United States.[12]

As a third limitation, the study was performed in Utah, where the population is mostly White (78%). However, we have included a fairness analysis, which showed that the Combined Approach would allow identification of more patients potentially eligible for lung cancer screening in White, Black, Hispanic and other populations. The fairness analysis showed that using the longitudinal data would disproportionately benefit women and Hispanics.

As a fourth limitation, we did not explicitly account for the random error inherent in patient-reported measures, including smoking history.[17]

As a fifth limitation, this study did not analyze free text data. However, including free text data in the Longitudinal Approach might have further limited the portability of the approach to other healthcare systems by requiring more adaptation and validation prior to clinical use.

CONCLUSION

This study contributes to the body of research indicating that smoking data recorded in the EHR at a point in time are often inaccurate for assessing longitudinal smoking history. This makes relying on only the most recent smoking history suboptimal. This study showed that a Longitudinal Approach, which leverages longitudinal smoking records in the EHR, can substantially improve upon the Baseline Approach, which uses only the most recent record. Healthcare organizations, EHR vendors, and researchers should consider adopting the Longitudinal Approach.

FUNDING

The work reported in this paper was supported in part by the Agency for Healthcare Research and Quality under Award Number R18HS026198. TJR was supported by the U.S. National Library of Medicine of the National Institutes of Health through grant T15LM007124.

The funding organizations had no role in the conceptualization, design, data collection, analysis, decision to publish, or paper preparation for this case study. The content is solely the responsibility of the authors and does not necessarily represent the official views of the organizations involved.

AUTHOR CONTRIBUTIONS

PVK takes responsibility for the content of the manuscript, including the data and analysis. Each author made substantial contributions to the drafting or substantial revision of the paper. All authors contributed substantially to the study design, data interpretation, and the writing of the manuscript. All authors also approved the paper for submission and agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated, resolved, and documented in the literature. PVK designed and implemented the Longitudinal Approach. PVK, HL, and YZ conducted the statistical analyses.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association Journal* online.

CONFLICT OF INTEREST STATEMENT

Outside of the submitted work, KK reports honoraria, consulting, sponsored research, writing assistance, licensing, or co-development in the past 3 years with McKesson InterQual, Hitachi, Pfizer, Klesis Healthcare, RTI International, Mayo Clinic, the University of California at San Francisco, MD Aware, and the U.S. Office of the National Coordinator for Health IT (via Security Risk Solutions) in the area of health information technology. KK was also an unpaid board member of the non-profit Health Level Seven International health IT standard development organization, he is an unpaid member of the U.S. Health Information Technology Advisory Committee, and he has helped develop a number of health IT tools which may be commercialized to enable wider impact. None of these relationships have direct relevance to the manuscript but are reported in the interest of full disclosure. Other authors have to conflict of interest to report.

DATA AVAILABILITY

The data underlying this article cannot be shared publicly due to UU Health policies regarding the privacy of patients and the protection of sensitive patient health information.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin* 2020; 70 (1): 7–30.
2. Aberle DR, Adams AM, Berg CD, *et al.* Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011; 365 (5): 395–409.
3. US Preventive Services Task Force. Final recommendation statement: lung cancer screening (2021). https://uspreventiveservicestaskforce.org/uspstf/recommendation/lung-cancer-screening. Accessed June 2, 2021.
4. Fedewa SA, Kazerooni EA, Studts JL, *et al.* State variation in low-dose computed tomography scanning for lung cancer screening in the United States. *J Natl Cancer Inst* 2021; 113 (8): 1044–52.
5. Centers for Medicare and Medicaid Services. Decision memo for screening for lung cancer with low dose computed tomography. 2015. https://www.cms.gov/medicare-coverage-database/details/nca-decision-memo.aspx?NCAId=274. Accessed August 29, 2019.
6. Caverly TJ, Cao P, Hayward RA, Meza R. Identifying patients for whom lung cancer screening is preference-sensitive. *Ann Intern Med* 2018; 169 (1): 1–9.
7. Mazzone PJ, Silvestri GA, Patel S, *et al.* Screening for lung cancer: CHEST Guideline and Expert Panel Report. *Chest* 2018; 153 (4): 954–85.
8. Chen LH, Quinn V, Xu L, *et al.* The accuracy and trends of smoking history documentation in electronic medical records in a large managed care organization. *Subst Use Misuse* 2013; 48 (9): 731–42.
9. Garies S, Cummings M, Quan H, *et al.* Methods to improve the quality of smoking records in a primary care EMR database: exploring multiple imputation and pattern-matching algorithms. *BMC Med Inform Decis Mak* 2020; 20 (1): 1–10.
10. Patel N, Miller DP, Snavely AC, *et al.* A comparison of smoking history in the electronic health record with self-report. *Am J Prev Med* 2020; 58 (4): 591–5.
11. Begnaud AL, Joseph AM, Lindgren BR. Randomized electronic promotion of lung cancer screening: a pilot. *JCO Clin Cancer Inform* 2017; 1 (1): 1–6.
12. Top 10 Ambulatory EHR Vendors by 2019 Market Share. https://www.definitivehc.com/blog/top-ambulatory-ehr-systems. Accessed January 18, 2022.
13. Center Cessation Initiative CC, Coordinating Center I. Building Smoking Cessation Electronic Health Record Functionalities and Workflows for the Oncology Setting: A Build Guide for Project Leaders, Clinicians, and Information Technology Personnel (Cerner Version) Cancer Center Cessation Initiative (C3I) Coordinating Center. Published online 2019.
14. Modin HE, Fathi JT, Gilbert CR, *et al.* Pack-year cigarette smoking history for determination of lung cancer screening eligibility: comparison of the electronic medical record versus a shared decision-making conversation. *Ann Am Thorac Soc* 2017; 14 (8): 1320–5.
15. Self TH, Wallace JL, Gray LA, Usery JB, Finch CK, Deaton PR. Are we failing to document adequate smoking histories? A brief review 1999-2009. *Curr Med Res Opin* 2010; 26 (7): 1691–6.
16. Polubriaginof F, Salmasian H, Albert DA, Vawdrey DK. Challenges with collecting smoking status in electronic health records. *AMIA Annu Symp Proc* 2017; 2017: 1392–400.
17. Caverly TJ, Zhang X, Hayward RA, Zhu J, Waljee AK. Effects of random measurement error on lung cancer screening decisions: a retrospective cohort-based microsimulation study. *Chest* 2021; 159 (2): 853–61.
18. Joseph AM, Rothman AJ, Almirall D, *et al.* Lung cancer screening and smoking cessation clinical trials SCALE (Smoking Cessation within the Context of Lung Cancer Screening) collaboration. *Am J Respir Crit Care Med* 2018; 197 (2): 172–82.
19. Reese T, Schlechter C, Kramer H, *et al.* Implementing lung cancer screening in primary care: needs assessment and implementation strategy design. *Transl Behav Med* 2021; ibab115. doi: 10.1093/tbm/ibab115. Epub ahead of print.
20. 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. *J Natl Cancer Inst* 2021; 113 (11): 1590–4.
21. Bach PB, Kattan MW, Thornquist MD, *et al.* Variations in lung cancer risk among smokers. *J Natl Cancer Inst* 2003; 95 (6): 470–8.
22. Bach PB, Elkin EB, Pastorino U, *et al.* Benchmarking lung cancer mortality rates in current and former smokers. *Chest* 2004; 126 (6): 1742–9.
23. Kawamoto K, Kukhareva PV, Weir CR, *et al.* Establishing a multidisciplinary initiative for interoperable electronic health record innovations at an academic medical center. *JAMIA Open* 2021; 4 (3): ooab041.
24. Kharrazi H, Chi W, Chang HY, *et al.* Comparing population-based risk-stratification model performance using demographic, diagnosis and medication data extracted from outpatient electronic health records versus administrative claims. *Med Care* 2017; 55 (8): 789–96.
25. Bae J, Ford EW, Kharrazi HHK, Huerta TR. Electronic medical record reminders and smoking cessation activities in primary care. *Addict Behav* 2018; 77: 203–9.
26. Reese TJ, Schlechter CR, Potter LN, *et al.* Evaluation of revised US Preventive Services Task Force Lung Cancer Screening Guideline among women and racial/ethnic minority populations. *JAMA Netw Open* 2021; 4 (1): e2033769.
27. Sobell LC, Brown J, Leo GI, Sobell MB. The reliability of the Alcohol Timeline Followback when administered by telephone and by computer. *Drug Alcohol Depend* 1996; 42 (1): 49–54.
28. 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 Rep* 2020; 17: 101037.