Atencion Primaria 58 (2026) 103363

Atención Primaria

www.elsevier.es/ap

ORIGINAL ARTICLE

Participant selection for lung cancer screening using primary care electronic medical records: The Catalan scenario

Mercè Marzo-Castillejo[a,b,c,*], Juanjo Mascort Roca[a,d,c], Albert Brau Tarrida[a,e,c], Lucia Carrasco Ribelles[a], Mònica Monteagudo Zaragoza[f,g,c], Carolina Guiriguet Capdevila[a,h,c], Josep A. Espinàs Piñol[i,g,j], Olivia Cabrera Godoy[i,g], José Ma Borras Andrés[i,g,j]

[a] *Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain*
[b] *Unitat de Suport a la Recerca, Àmbits d'Atenció Primària Metropolitana Sud i Penedès, Institut Català de la Salut, L'Hospitalet de Llobregat, Spain*
[c] *Grupo de investigación Càncer y Atención Primaria de la IDIAPJGol, Spain*
[d] *Centro de Atención Primaria Florida Sud, Institut Català de la Salut, L'Hospitalet de Llobregat, Barcelona, Spain*
[e] *Centro de Atención Primaria La Mina, Institut Català de la Salut, Sant Adrià del Besòs, Barcelona, Spain*
[f] *Hospital Universitari de Bellvitge Institut Català de la Salut, L'Hospitalet de Llobregat, Barcelona, Spain*
[g] *Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain*
[h] *Primary Care Services Information System (SISAP), Institut Català de la Salut (ICS), Barcelona, Catalonia, Spain*
[i] *Pla director d'Oncologia de Catalunya, Departament de Salut, L'Hospitalet de Llobregat, Barcelona, Spain*
[j] *Facultat de Medicina i Ciències de la Salut, Universitat de Barcelona, Campus Bellvitge, L'Hospitalet de Llobregat, Barcelona, Spain*

Received 14 April 2025; accepted 24 July 2025

KEYWORDS
Lung cancer;
Lung cancer screening;
Primary care;
Electronic health records

Abstract
Objective: To assess the feasibility of using primary care electronic health records (EHRs) and the PLCOm2012noRace lung cancer (LC) risk prediction model to identify high-risk individuals in the Catalan population.
Design: Population-based cohort study.
Site: Catalonia, using data from the Information System for the Improvement of Research in Primary Care (SIDIAP), which covers approximately 80% of the population.
Participants: A total of 1,998,282 individuals aged 55–79 years were initially considered, with data spanning from 2012 to 2023. After applying inclusion and exclusion criteria based on smoking status, 24,294 individuals with complete smoking history were included.
Interventions: Estimation of LC risk using the PLCOm2012noRace model.

* Corresponding author.
E-mail address: mmarzo@idiapjgol.org (M. Marzo-Castillejo).

https://doi.org/10.1016/j.aprim.2025.103363

Main measurements: Variables: age, smoking history, body mass index, educational level, chronic obstructive pulmonary disease, personal history of cancer, and family history of LC. A 6-year risk threshold of ≥2.6% was used to define eligibility for LC screening.
Results: Overall, 18.6% of individuals exceeded the risk threshold, with higher prevalence in men (21.4%) and those aged 60–79 years (23.8%). Current smokers had the highest risk (25.7%), which decreased with time since quitting. On average, high-risk individuals could have been identified 4.29 years before.
Conclusions: The use of EHRs and the PLCOm2012noRace model is a feasible approach to identify individuals at high risk of LC in the Catalan population. However, missing or outdated data, especially regarding smoking intensity, may limit the predictive performance. These findings highlight the need for systematic and timely data collection to support effective risk-based screening strategies.

PALABRAS CLAVE
Cáncer de pulmón;
Cribado cáncer de pulmón;
Atención primaria;
Historia clínica electrónica

Selección de participantes para el cribado del cáncer de pulmón mediante las historias clínicas de atención primaria: el escenario catalán

Resumen
Objetivo: Evaluar la viabilidad de utilizar las historias clínicas electrónicas (HCE) de atención primaria y el modelo de predicción de riesgo de cáncer de pulmón (CP) PLCOm2012noRace para identificar individuos de alto riesgo en la población.
Diseño: Estudio de cohortes de base poblacional.
Emplazamiento: Cataluña, utilizando datos del Sistema de Información para el Desarrollo de la Investigación en Atención Primaria (SIDIAP), que cubre aproximadamente el 80% de la población.
Participantes: Se consideraron 1.998.282 individuos de entre 55 y 79 años (2012-2023). Tras aplicar los criterios de inclusión y exclusión, se incluyeron 24.294 individuos con historia completa de tabaquismo.
Intervenciones: Estimación del riesgo de CP mediante el modelo PLCOm2012noRace.
Mediciones principales: Variables: edad, antecedentes de tabaquismo, índice de masa corporal, nivel educativo, enfermedad pulmonar obstructiva crónica, antecedentes personales de cáncer y antecedentes familiares de CP. Se utilizó un umbral de riesgo a 6 años ≥ 2,6%.
Resultados: El 18,6% de individuos superaron el umbral de riesgo, más en varones (21,4%) entre 60 y 79 años (23,8%). Los fumadores actuales presentaron el mayor riesgo (25,7%), que disminuía con el tiempo desde el abandono del hábito. Se identificaría los individuos de alto riesgo una media de 4,29 años antes.
Conclusiones: El uso de las HCE y el modelo PLCOm2012noRace es factible para identificar los individuos con alto riesgo de CP. La falta o desactualización de datos, especialmente sobre la intensidad del tabaquismo, puede limitar el rendimiento predictivo. Una recolección de datos sistemática y oportuna apoyaría estrategias de cribado basadas en el riesgo.

Introduction

Lung cancer (LC) is a leading cause of cancer-related deaths worldwide, with a poor prognosis due to late detection. In Spain, 29,188 new cases and 22,930 deaths were reported in 2020,[1] with a survival rate of 13.5% for the period 2010–2014.[2] Tobacco use is responsible for around 85% of cases, along with environmental risk factors such as radon and air pollution.[3] Randomized Controlled Trial (RCT) Screening with low-dose computed tomography (LDCT) has been shown to significantly reduce the mortality rate from LC in high-risk individuals.[4–6] However, identifying these high-risk individuals accurately is crucial to ensure the effectiveness and cost-efficiency of screening programs.

The Prostate, Lung, Colorectal, and Ovarian (PLCOm2012) risk prediction model is a clinically validated tool that incorporates additional risk factors beyond age and smoking history to improve risk stratification.[7] The EU-funded 4-IN THE LUNG RUN project, a multicenter randomized controlled trial involving the Catalan Institute of Oncology (ICO), uses the modified PLCOm2012nonrace model for risk stratification, excluding ethnicity as a predictor to improve its global applicability.[8,9] Before implementing large-scale, organized LC screening, it is

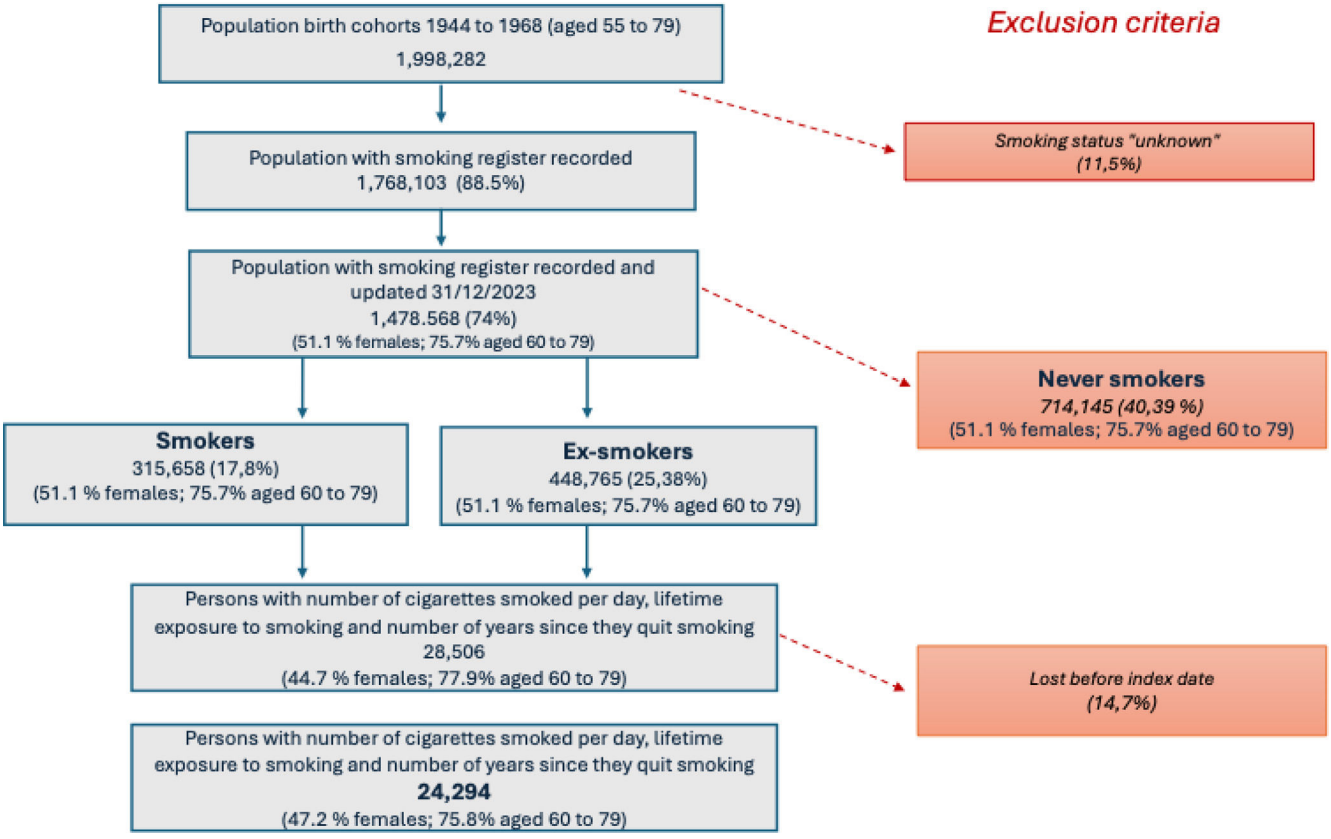

Figure 1 Flowchart through inclusion and exclusion criteria.

essential to explore efficient recruitment strategies to identify high-risk individuals. In the Catalan healthcare system, primary care electronic health records (EHRs) represent a valuable resource. These records contain longitudinal, structured clinical data for most of the population, and can be used to estimate individual LC risk scores.

This study aims to evaluate the performance of the PLCOm2012nonrace model when applied to routine EHR data. The goal is to assess whether this approach could support a population-based screening strategy by identifying high-risk individuals earlier using existing data infrastructure.

Methods

Study design, study population and data source

This population-based cohort study used data from the Information System for the Improvement of Research in Primary Care (SIDIAP), a secondary database source derived from primary care EHRs in Catalonia.[10] SIDIAP routinely collects pseudo-anonymized health data from individuals attending primary care centers, covering approximately 80% of the Catalan population (5,564,292 individuals). The database includes demographic, anthropometric, lifestyle (e.g., smoking), clinical and diagnostic (ICD-10) data, among others.[10] Data from 2012 to 2023 were used for this study. Inclusion and exclusion criteria are shown in Fig. 1. The initial cohort comprised 1,998,282 individuals (born between 1944 and 1968, aged 55–79 years), of whom 1,768,103 had a recorded smoking status and 1,478,568 had an updated record on 31 December 2023. Individuals with unknown smoking status (11.5%) or classified as never smokers (40.39%) were excluded. The final cohort included 315,658 current smokers (17.8%) and 448,765 former smokers (25.38%). Of these, 28,506 individuals had complete data on cigarette consumption, lifetime smoking exposure, and years since quitting. The final study size ultimately included 24,294 individuals aged 55–79 years who were still active at follow-up and had sufficiently detailed smoking histories for LC risk assessment on 31 December 2023.

PLCO2012 nonRace risk score model and variables

We calculated the 6-year LC risk score using the PLCOm2012noRace model,[8] which includes the following variables: (1) age, (2) level of education (categorized as less than high school, high school, post-high school education, some college, college, or postgraduate), (3) body mass index (BMI), (4) presence of chronic obstructive pulmonary disease (COPD), (5) personal history of cancer, (6) family history of LC, (7) smoking status (never, former, or current smoker), (8) average smoking intensity (number of cigarettes per day), (9) smoking duration (years), and (10) time since quitting smoking (set to 0 for current smokers). A score $\geq$2.6 indicates a high risk of LC. Sex was also recorded as an additional variable.

Table 1 Description of the complete study population, and by sex.

Item	Female (n = 11,460; 47.2%)	Male (n = 12,834; 52.8%)	Total (n = 24,294)
1) Age (years)	64.1 [59.7;69.0]	65.8 [60.8;71.3]	65.0 [60.2;70.2]
3) BMI (kg/m^2)	27.2 [24.0;31.1]	28.0 [25.3;31.2]	27.7 [24.7;31.1]
4) COPD (yes)	2841 (24.8%)	4445 (34.6%)	7286 (30.0%)
5) Personal history of cancer (yes)	1171 (10.2%)	2018 (15.7%)	3189 (13.1%)
6) Family history of lung cancer (yes)	98 (0.86%)	68 (0.53%)	166 (0.68%)
7) Current smoking status (smoking)	5446 (47.5%)	5645 (44.0%)	11,091 (45.7%)
8) Smoking intensity (avg. # cigarettes/day)	15.0 [8.00;20.0]	16.0 [8.00;24.0]	15.0 [8.00;22.0]
9) Lifetime smoking duration (years)	15.3 [10.4;21.5]	14.9 [10.1;21.8]	15.1 [10.2;21.6]
10) Years since quitting (years)	0.22 [0.00;3.96]	0.57 [0.00;4.29]	0.43 [0.00;4.15]

BMI: body mass index; COPD: chronic obstructive pulmonary disease.
The description of item 2 (education level) is not provided since the value was equal for every participant.

Table 2 Description of the study population, by age. Table 2. Description of the study population, by age.

Item	55–59 years old (N = 5881; 24.2%)	60–79 years old (N = 18,413; 75.8%)
Sex		
Female	3064 (52.1%)	8396 (45.6%)
Male	2817 (47.9%)	10,017 (54.4%)
1) Age (years)	57.7 [56.4;58.9]	67.4 [63.7;71.8]
3) BMI (kg/m^2)	27.7 [24.5;31.6]	27.7 [24.8;31.0]
4) COPD (Yes)	921 (15.7%)	6365 (34.6%)
5) Personal history of cancer (Yes)	420 (7.14%)	2769 (15.0%)
6) Family history of lung cancer (Yes)	40 (0.68%)	126 (0.68%)
7) Current smoking status (Smoking)	2881 (49.0%)	8210 (44.6%)
8) Smoking intensity (avg. # cigarettes/day)	16.0 [10.0;23.0]	15.0 [8.00;22.0]
9) Lifetime smoking duration (Years)	14.9 [9.79;21.4]	15.1 [10.4;21.8]
10) Years since quitting (Years)	0.11 [0.00;3.86]	0.53 [0.00;4.22]

BMI: body mass index; COPD: chronic obstructive pulmonary disease.
The description of item 2 (education level) is not provided since the value was equal for every participant.

The latest available data for each variable were used in the calculations. If a variable had no recorded value in a participant's history, it was imputed using the mean value reported in the original model.[8] However, this only happened for the variable BMI in five participants (0.02%). Due to registration gaps in SIDIAP, some informed assumptions were made. For instance, as education level is not recorded, all participants were assigned a ''high school'' level based on the most recent regional health survey data for their sex and age group.[11] Regarding smoking intensity, the source data categorize cigarette consumption per day into discrete intervals: 0–10, 11–20, 21–30, and >30 cigarettes/day. As a numerical value is required to calculate the PLCOm2012noRace score, values within the first three intervals were imputed using a uniform distribution, assuming equal probability for any number of cigarettes within the interval. For the last interval (>30 cigarettes/day), imputation was performed using a beta distribution ($\alpha = 1$, $\beta = 3$), where the probability of higher consumption gradually decreases. In addition, a correction was applied if a participant's smoking status was recorded as ''former,'' but more recent data indicated cigarette consumption greater than zero. In such cases, the smoking status was updated to ''current smoker''.

Statistical analysis

In this descriptive study, quantitative variables are summarized using mean and standard deviation (SD) or the median with interquartile range [$Q1$, $Q3$], as appropriate. Data were stratified by sex, age, and years since quitting smoking. Group differences were assessed using the t-test, Mann–Whitney U test, ANOVA, or Kruskal–Wallis's test, depending on the distribution of the data. The time since the score became positive ($\geq$2.6) was calculated to determine how early the model could have detected high LC risk in individuals. The most recent information available in the EHR was used to calculate the score, and the age of this information was described. For individuals positive on December 31, 2023, we traced the onset by recalculating the score each time a new value for any PLCOm2012noRace variable was recorded. All analyses were performed using R v4.3, and the code is publicly available in our GitHub repository.[12]

Results

Population description

A comprehensive description of the entire cohort (24,294 individuals), stratified by sex, age, and time since quitting smoking, is presented in Tables 1–3. Statistically significant differences were observed for all variables when stratified by sex ($p < 0.001$), except for family history of LC ($p = 0.003$) and lifetime smoking duration ($p = 0.25$) (Table 1). When stratified by age, statistically significant differences were found for all variables ($p < 0.001$), except for BMI ($p = 0.25$) and family history of LC ($p = 1.00$) (Table 2). Finally, when the population was stratified by time since quitting smoking, statistically significant differences between the groups were observed for all variables ($p < 0.001$), except for family history of LC ($p = 0.5$) (Table 3).

Individuals with the score $\geq$ 2.6, and time (years) with the score $\geq$ 2.6

Table 4 provides an analysis of the cohort with respect to the PLCOm2012noRace $\geq$ 2.6 risk threshold. In the cohort, 18.6% met the threshold, with a higher proportion of men (21.4%) compared to women (15.6%) ($p < 0.001$). Individuals aged 60–79 years were more likely to meet the risk threshold than those aged 55–59 years (23.8% vs. 2.5%; $p < 0.001$). The prevalence of meeting the PLCOm2012noRace $\geq$ 2.6 threshold was highest among current smokers (25.7%) and gradually decreased with increasing time since smoking cessation, reaching 6.35% among those who had quit for 10 years or more. Among individuals classified as high risk at the end of the study, the PLCOm2012noRace score would have identified them 4.29 years earlier. The time to reach the score of $\geq$2.6 was consistent across sex and quitting-smoking categories; however, as predicted, this consistency was not observed in younger age groups.

Age of the information in the EHR

The age of the most up-to-date information available in the EHR to calculate the different items of the score was analyzed. The age ranged from 1.6 years for BMI (item 3) to 7.5 years for smoking intensity (item 8) (see Fig. 2). Notably, the items with the greatest impact on the score (items 8 and 6) had the longest time since their last update, indicating a potential gap in the timeliness of critical data.

Discussion

In a representative sample of the Catalan population from the SIDIAP database (80% coverage), 89% of nearly 2 million individuals aged 55–79 had a recorded smoking status, and 74% had an updated record on 31 December 2023. Of these, 315,658 were current smokers (17.8%) and 448,765 were former smokers (25.38%), but only 24,294 had a detailed smoking history for LC risk assessment. Using a PLCOm2012noRace threshold of $\geq$2.6, 18.6% were eligible for LDCT screening and the model would have identified high-risk individuals 4.29 years earlier. Model accuracy was

Table 3 Description of the study population, by time since quitting smoking.

Item	Current smoker (n = 11,091; 45.7%)	<1 year (n = 2583; 10.6%)	1–4 years (n = 5782; 23.8%)	5–9 years (n = 4318; 17.8%)	≥10 years (n = 520; 2.1%)
Sex					
Female	5446 (49.1%)	1179 (45.6%)	2705 (46.8%)	1896 (43.9%)	234 (45.0%)
Male	5645 (50.9%)	1404 (54.4%)	3077 (53.2%)	2422 (56.1%)	286 (55.0%)
1) Age (years)	64.6 [59.8;70.0]	65.1 [60.2;69.9]	65.3 [60.6;70.4]	65.6 [60.7;70.9]	65.2 [60.3;70.7]
3) BMI (kg/m^2)	27.0 [24.1;30.4]	27.6 [24.8;31.1]	28.3 [25.4;31.8]	28.3 [25.4;31.8]	28.8 [25.9;31.9]
4) COPD (Yes)	3409 (30.7%)	894 (34.6%)	1710 (29.6%)	1181 (27.4%)	92 (17.7%)
5) Personal history of cancer (Yes)	1341 (12.1%)	380 (14.7%)	767 (13.3%)	631 (14.6%)	70 (13.5%)
6) Family history of lung cancer (Yes)	73 (0.66%)	17 (0.66%)	49 (0.85%)	25 (0.58%)	2 (0.38%)
7) Current smoking status (Smoking)	11,091 (100%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
8) Smoking intensity (avg. # cigarettes/day)	15.0 [8.00;22.0]	15.0 [8.00;22.0]	15.0 [8.00;21.0]	15.0 [8.00;22.0]	16.0 [10.0;23.0]
9) Lifetime smoking duration (Years)	16.6 [12.6;23.3]	16.7 [12.1;23.2]	14.3 [9.84;20.9]	10.4 [6.56;16.2]	7.31 [4.00;12.3]
10) Years since quitting (Years)	0.00 [0.00;0.00]	0.52 [0.24;0.72]	2.76 [1.82;4.06]	6.99 [5.90;8.14]	10.7 [10.3;11.2]

BMI: body mass index; COPD: chronic obstructive pulmonary disease.
The description of item 2 (education level) is not provided since the value was equal for every participant.

Table 4 Score of the PLCOm2012noRace model, proportion of individuals with the score ≥ 2.6, and time (years) with the score ≥ 2.6.

	PLCom2012noRace	PLCom2012noRace ≥ 2.6 (%)	Time (years) with PLCom2012noRace ≥ 2.6
Complete study population (N = 24,294)	0.98 [0.36, 2.08]	4527 (18.6)	4.29 [1.66, 6.65]
Sex			
Female (*N* = 11,460; 47.2%)	0.89 [0.32, 1.85]	1783 (15.6)	4.05 [1.53, 6.24]
Male (*N* = 12,834; 52.8%)	1.07 [0.41, 2.31]	2890 (22.0)	4.54 [1.83, 6.94]
Age (years)			
55–59 (*N* = 5881; 24.2%)	0.59 [0.27, 1.02]	148 (2.52%)	0.91 [0.15, 2.31]
60–79 (*N* = 18,413; 75.8%)	1.23 [0.43, 2.50]	4379 (23.8)	4.40 [1.74, 6.73]
Time since quitting smoking (years)			
Current smoker (*N* = 11,091; 45.7%)	1.31 [0.53,2.66]	2851 (25.7)	3.83 [1.30, 5.88]
<1 year (*N* = 2583; 10.6%)	0.98 [0.41,2.11]	477 (18.5)	4.71 [2.57, 6.83]
1–4 years (*N* = 5782; 23.8%)	0.81 [0.30,1.67]	771 (13.3)	5.12 [3.42, 7.27]
5–9 years (*N* = 4318; 17.8%)	0.66 [0.26,1.36]	395 (9.15)	7.23 [3.11, 8.91]
≥10 years (*N* = 520; 2.1%)	0.52 [0.24,1.04]	33 (6.35)	2.24 [0.13, 10.28]

limited by incomplete EHR data, particularly smoking intensity. Better recorded variables such as BMI and COPD improve risk stratification, reinforcing the need for multifactorial screening criteria beyond age and smoking history.

Smoking prevalence in Catalonia

A comparison between the ESCA survey-based smoking prevalence[11] and our EHR-based cohort suggests an underestimation of smoking rates in the EHR data. ESCA reports a smoking rate of 24.6% among people aged 55–64 years, while the EHR-based cohort reports only 17.8% among people aged 55–64. Although not directly comparable, the discrepancy is likely to reflect missing or outdated smoking records, which could affect LC risk assessment and screening eligibility in routine practice.

Comparison with previous studies

Several studies[13,14] have shown that the identification of individuals at high risk of LC is improved by the inclusion of additional risk factors beyond age and smoking history, which were the criteria used in the National Lung Screening Trial[4] and the NELSON RCTs.[5] The PLCOm2012 model has been shown to be effective in LC risk stratification, with versions such as PLCOm2012NoRace addressing racial bias.[8]

In our study, using the PLCOm2012noRace model with a risk threshold of ≥2.6%, 18.6% of individuals were eligible for screening. When we constructed a simplified version, by setting the variables education level, BMI, COPD, personal history of cancer, family history of LC to the mean value of the population as reported by Ten Haaf K et al.,[13] the score produced significantly lower values than the full version (median: 0.33 vs. 0.98), resulting in fewer individuals being classified as high risk (1.7%). This highlights the importance of multivariable risk criteria. Direct comparisons with other studies are difficult because of differences in thresholds and population characteristics.[15,16] For instance, the Manchester Lung Health Check pilot applied the PLCOm2012 model with a risk threshold of ≥1.51% for LC (6 years) to determine screening eligibility in a socioeconomically deprived cohort. As a result, 56% of participants (*n* = 1430/2541) were classified as high-risk and offered screening.[15] Ongoing research, such as the 4-IN THE LUNG RUN project,[9] will provide further evidence for comparison.

The role of primary care EHRs in supporting population-based screening

Several studies in the United Kingdom[16–18] and the United States[19–22] have demonstrated the effectiveness of using EHRs for LC screening, highlighting their global potential to improve access to screening, particularly for socioeconomically disadvantaged populations.[23] The present study further supports this evidence, showing that primary care EHRs constitute a valuable source of routinely collected clinical data that can be leveraged for risk stratification within a population-based LC screening program.

While data are recorded in the primary care setting, the use of such information to estimate PLCOm2012noRace scores does not necessarily imply that family physicians are responsible for delivering or managing the screening itself. Instead, centralized identification of high-risk individuals—similar to other screening programs like breast or colorectal cancer—can reduce the burden on primary care and preserve the distinction between clinical care and public health interventions.

However, successful implementation requires better integration between primary care and LC screening programs.[24] Key challenges include high workload pressures, data accuracy issues (particularly around smoking status),[23] and a lack of standardized guidelines for identifying eligible patients.[24] Overcoming these barriers requires a structured

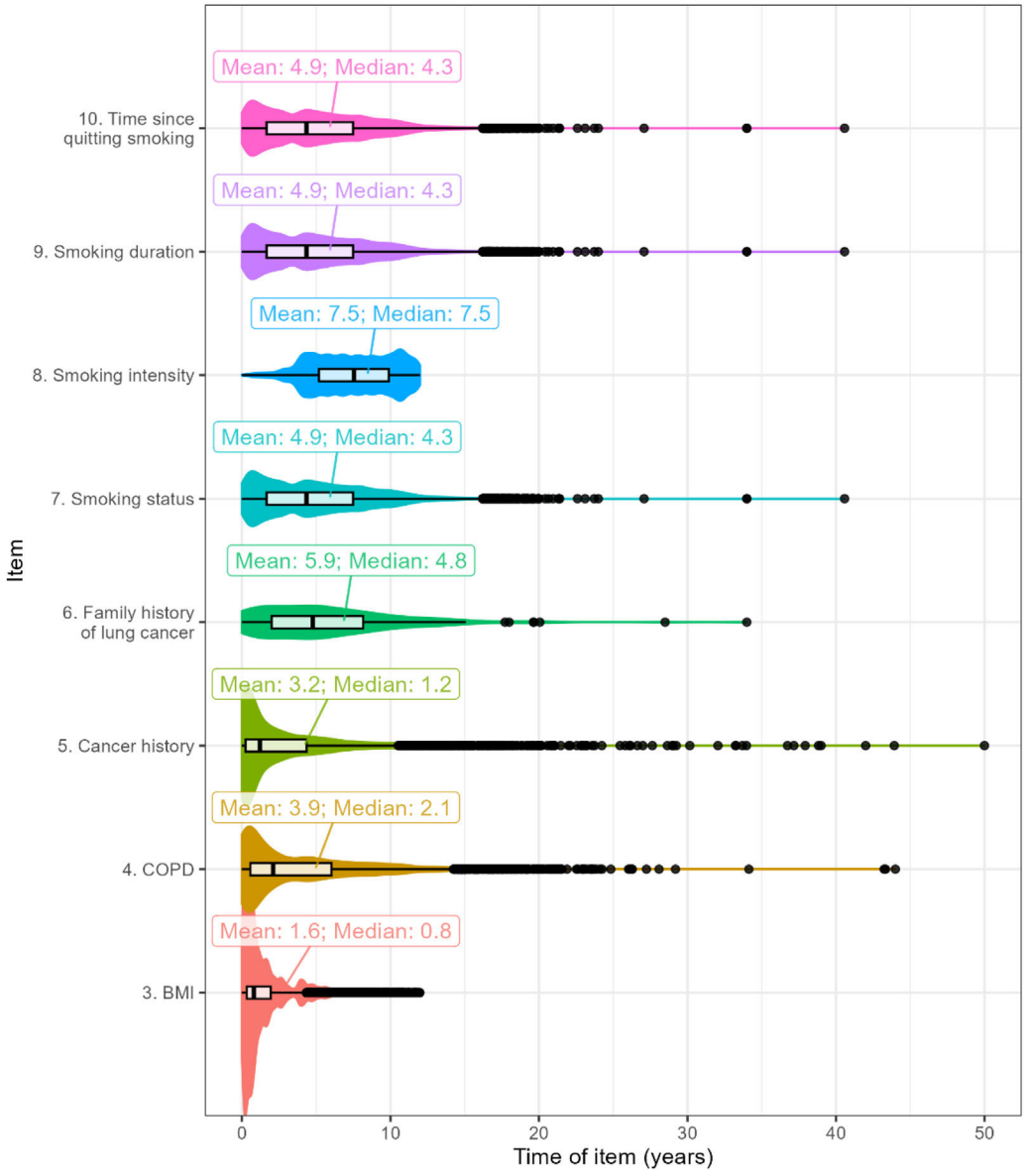

Figure 2 Historical period of the information used for the calculation of each item in the PLCOm2012noRace model. Item 1 (age) is determined at the time of risk calculation, while item 2 (education level) was imputed for all participants.

approach to integrating EHR-based screening models into routine primary care workflows.[24]

Capacity within primary care to support LC screening implementation in Catalonia

Our study suggests that Catalan primary care EHRs could be valuable for identifying individuals at high risk of lung cancer. However, lack of smoking and other relevant data for PLCOm2012noRace model calculation remain a challenge, as smoking history is primarily recorded for tobacco advice and counseling and not for screening (one activity not included in the public list of services). In addition, the COVID-19 pandemic has contributed to delays in updating some information, such as smoking intensity, with an average update interval of 7.5 years.

The European Commission supports LDCT screening in high-risk populations and recommends pilot programs. In the context of the 4-IN THE LUNG RUN and with the aim of identifying the optimal and most cost-effective strategy to invite eligible individuals, the Catalan partner of the project has done this through the primary EHR. Primary care physicians validate the data used to calculate the PLCOm2012NoRace score through a questionnaire inserted into the EHR before referring eligible individuals. This study highlights the role of Catalan primary care in risk-based LC screening and may inform future integrated models.

As gatekeepers and core of the health system, primary care plays a crucial role in identifying high-risk individuals and ensuring equitable and informed participation in screening programs. Their expertise and close relationships with patients enable them to facilitate successful implementation. However, their involvement must remain feasible given

the current workload crisis in primary care.[25] Strong collaboration between primary care and screening programs regarding outcomes, scheduling, and participation is essential for the success of LC screening initiatives.

Strengths and limitations

While EHRs accurately capture smoking status - distinguishing non-smokers, active smokers, and ex-smokers-the accuracy of the most influential variable, smoking intensity, remains limited. Smoking intensity is required as a numerical value for the PLCOm2012noRace model but is often recorded as a discrete category (as part of the Fagerstrom questionnaire to measure dependency on nicotine), reducing its reliability for this new purpose. Despite these limitations, evidence suggests that never-smokers and low-intensity smokers missed by targeted approaches generally have a lower risk of LC,[22] partially mitigating the impact of these inaccuracies on the identification of high-risk individuals.

This study has several strengths, including the use of a large, population-based dataset and the application of the validated PLCOm2012noRace risk prediction model, which incorporates well-documented variables such as BMI and COPD from primary care EHRs. An important strength of this study is its link to the 4-IN THE LUNG RUN project, which uses the same primary care EHRs from which this study was conducted. The evaluation of fieldwork from the European project, where the PLCOm2012NoRace questionnaire is derived from primary care data and missing data are updated during primary care consultations, offers a promising approach to overcome the limitations of primary care smoking registers.

This study is generalizable to the Catalan population served by primary care centers. However, its applicability to other regions with different patterns of access to healthcare requires further evaluation.

Future research directions

Our findings highlight the importance of optimizing EHR data to improve LC screening. Tailoring screening strategies to balance accuracy and resource allocation is essential to ensure sustainability without compromising patient outcomes. Future research should explore the impact of applying different PLCOm2012noRace risk thresholds in clinical practice and how these thresholds align with health system resources, including LDCT scan availability, diagnostic outcomes, treatment capacity at specialist hospital centers, and overall system efficiency.

Collaboration between primary care and national screening programs will be essential to refine these thresholds and support effective implementation. Ongoing projects, such as the 4-IN THE LUNG RUN,[9] will provide critical evidence to develop scalable, resource-adapted screening programs.

In addition, efforts should focus on integrating smoking cessation into LC screening programs, which could further reduce the risk of LC and provide public health benefits. Finally, understanding the role of socioeconomic factors and improving access to screening for underserved populations will be essential to achieving equity in screening programs.

Conclusions

This study demonstrates the feasibility of using primary care EHRs and the PLCOm2012noRace model to identify individuals at high risk of LC in the Catalan population. Although the multivariable risk models improve early detection, their effectiveness is limited by incomplete and outdated data, particularly on smoking history. Improving data quality and implementing strategies to address missing information are essential to optimize risk prediction. Strengthening collaboration between primary care and screening programs will be key to developing a more accurate and effective LC screening model.

What is known about the topic

- Early detection of lung cancer is crucial for improving outcomes in high-risk patients.
- Models such as PLCOm2012 are effective for LC risk stratification but rely on complete and accurate data.
- The use of electronic health records (EHRs) in primary care is promising, but with limitations in data quality, particularly regarding smoking history.

Contributions of this study

- This study demonstrates the potential feasibility of using primary care EHRs to identify individuals at risk of LC in the Catalan population.
- It provides evidence of lack of some smoking data and their impact on the accuracy of risk models, highlighting the need of asking and recording data for this new use.
- Emphasises the importance of integrating risk models into routine clinical practice for the early detection of lung cancer

Ethical considerations

The project was approved by the Research Ethics Committee (CEI) of IDIAPJGol (registration number: 23/107-P).

Funding

This project was made possible with the support of the Departament de Salut de la Generalitat de Catalunya, which provided the necessary funding within the framework of the 2021 call for grants under the Strategic Plan for Research and Innovation in Health (PERIS) 2021–2024, in the category of research projects focused on primary care, with the grant reference SLT021/21/000044.

Conflict of interest

The authors of this study declare that they have no conflicts of interest

Acknowledgements

We thank Anna Berenguera, Josep Basora, and the SIDIAP team at IDIAPJGol for their essential support in carrying out this study. We are grateful to Carmen Cabezas for her thorough review and helpful comments. We also thank Carl Martin Tammemägi for granting full permission to use the PLCOm2012noRace lung cancer risk prediction model.

References

1. Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, et al. Global cancer observatory: cancer today. Lyon, France: International Agency for Research on Cancer; 2024. Available from: https://gco.iarc.who.int/today [accessed 24.2.25].
2. Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Nikšić M, et al. Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37513025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. Lancet. 2018;391:1023–75, http://dx.doi.org/10.1016/S0140-6736(17)33326-3.
3. Turner MC, Andersen ZJ, Baccarelli A, Diver WR, Gapstur SM, Pope CA 3rd, et al. Outdoor air pollution and cancer: an overview of the current evidence and public health recommendations. CA Cancer J Clin. 2020, http://dx.doi.org/10.3322/caac.21632.
4. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365:395–409, http://dx.doi.org/10.1056/NEJMoa1102873.
5. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. N Engl J Med. 2020;382:503–13, http://dx.doi.org/10.1056/NEJMoa1911793.
6. Muriana P, Rossetti F, Novellis P, Veronesi G. Lung cancer screening: the European perspective. Thorac Surg Clin. 2023;33:375–83, http://dx.doi.org/10.1016/j.thorsurg.2023.04.017.
7. Tammemägi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, et al. Selection criteria for lung-cancer screening. N Engl J Med. 2013;368:728–36, http://dx.doi.org/10.1056/NEJMoa1211776.
8. Tammemägi MC, Cina K, Kitts AKB, Koop D, Petereit MA, Sargent M, et al. Sensitivity of US Preventive Services Task Force and PLCOm2012 lung cancer screening eligibility criteria in individuals with lung cancer in South Dakota self-reporting as Indigenous and non-Indigenous. Cancer. 2023;129:3894–904, http://dx.doi.org/10.1002/cncr.34947.
9. van der Aalst C, Vonder M, Hubert J, Moldovanu D, Schmitz A, Delorme S, et al. P1.14-04 European lung cancer screening implementation: 4-IN-THE-LUNG-RUN trial. J Thorac Oncol. 2023;18 Suppl.:S217.
10. Recalde M, Rodríguez C, Burn E, Far M, García D, Carrere-Molina J, et al. Data resource profile: the information system for research in primary care (SIDIAP). Int J Epidemiol. 2022;51:e324–36, http://dx.doi.org/10.1093/ije/dyac068.
11. Health Survey of Catalonia (ESCA), 2023. https://www.idescat.cat/indicadors/?id=aec&n=15798&t=202300 [accessed 24.11.24].
12. https://github.com/IDIAPJGol/CRIPUAP.
13. Ten Haaf K, Jeon J, Tammemägi MC, Han SS, Kong CY, Plevritis SK, et al. Risk prediction models for selection of lung cancer screening candidates: a retrospective validation study. PLoS Med. 2017;14:e1002277. Erratum in: PLoS Med. 2020;17:e1003403. doi:10.1371/journal.pmed.1003403.
14. Tammemägi MC, Ruparel M, Tremblay A, Myers R, Mayo J, Yee J, et al. USPSTF2013 versus PLCOm2012 lung cancer screening eligibility criteria (International Lung Screening Trial): interim analysis of a prospective cohort study. Lancet Oncol. 2022;23:138–48, http://dx.doi.org/10.1016/S1470-2045(21)00590-8.
15. Lebrett MB, Balata H, Evison M, Colligan D, Duerden R, Elton P, et al. Analysis of lung cancer risk model (PLCOM2012 and LLPv2) performance in a community-based lung cancer screening programme. Thorax. 2020;75:661–8, http://dx.doi.org/10.1136/thoraxjnl-2020-214626.
16. Dickson JL, Hall H, Horst C, Tisi S, Verghese P, Worboys S, et al. Utilisation of primary care electronic patient records for identification and targeted invitation of individuals to a lung cancer screening programme. Lung Cancer. 2022;173:94–100, http://dx.doi.org/10.1016/j.lungcan.2022.09.009.
17. Jani BD, Sullivan MK, Hanlon P, Nicholl BI, Lees JS, Brown L, et al. Personalised lung cancer risk stratification and lung cancer screening: do general practice electronic medical records have a role? Br J Cancer. 2023;129:1968–77, http://dx.doi.org/10.1038/s41416-023-02467-9.
18. McCutchan G, Engela-Volker J, Anyanwu P, Brain K, Abel N, Eccles S. Assessing, updating and utilising primary care smoking records for lung cancer screening. BMC Pulm Med. 2023;23:445, http://dx.doi.org/10.1186/s12890-023-02746-4. PMID: 37974137.
19. O'Brien MA, Sullivan F, Carson A, Siddiqui R, Syed S, Paszat L. Piloting electronic screening forms in primary care: findings from a mixed methods study to identify patients eligible for low dose CT lung cancer screening. BMC Fam Pract. 2017;18:95, http://dx.doi.org/10.1186/s12875-017-0666-5.
20. Wang X, Zhang Y, Hao S, Zheng L, Liao J, Ye C, et al. Prediction of the 1-year risk of incident lung cancer: prospective study using electronic health records from the state of Maine. J Med Internet Res. 2019;21:e13260, http://dx.doi.org/10.2196/13260.
21. Reese TJ, Schlechter CR, Kramer H, Kukhareva P, Weir CR, Del Fiol G, et al. Implementing lung cancer screening in primary care: needs assessment and implementation strategy design. Transl Behav Med. 2022;12:187–97, http://dx.doi.org/10.1093/tbm/ibab115.
22. Steinberg MB, Young WJ, Miller Lo EJ, Bover-Manderski MT, Jordan HM, et al. Electronic health record prompt to improve lung cancer screening in primary care. Am J Prev Med. 2023;65:892–5, http://dx.doi.org/10.1016/j.amepre.2023.05.016.
23. Goodley P, Balata H, Alonso A, Brockelsby C, Conroy M, Cooper-Moss N, et al. Invitation strategies and participation in a community-based lung cancer screening programme located in areas of high socioeconomic deprivation. Thorax. 2023;79:58–67, http://dx.doi.org/10.1136/thorax-2023-220001.
24. Patel P, Bradley SH, McCutchan G, Brain K, Redmond P. What should the role of primary care be in lung cancer screening? Br J Gen Pract. 2023;73:340–1, http://dx.doi.org/10.3399/bjgp23X734397.
25. Martin SA, Johansson M, Heath I, Lehman R, Korownyk C. Sacrificing patient care for prevention: distortion of the role of general practice. BMJ. 2025;388:e080811, http://dx.doi.org/10.1136/bmj-2024-080811.