# Recalibrating Risk Prediction Models by Synthesizing Data Sources: Adapting the Lung Cancer PLCO Model for Taiwan

Li-Hsin Chien[1], Tzu-Yu Chen[1], Chung-Hsing Chen[2], Kuan-Yu Chen[3], Chin-Fu Hsiao[1,4], Gee-Chen Chang[5,6,7,8], Ying-Huang Tsai[9,10], Wu-Chou Su[11], Ming-Shyan Huang[12], Yuh-Min Chen[13,14], Chih-Yi Chen[15,16], Sheng-Kai Liang[17,18], Chung-Yu Chen[19], Chih-Liang Wang[20], Hsiao-Han Hung[2], Hsin-Fang Jiang[1], Jia-Wei Hu[1], Nathaniel Rothman[21], Qing Lan[21], Tsang-Wu Liu[2], Chien-Jen Chen[22], Pan-Chyr Yang[3], I-Shou Chang[2], and Chao A. Hsiung[1]

## ABSTRACT

**Background:** Methods synthesizing multiple data sources without prospective datasets have been proposed for absolute risk model development. This study proposed methods for adapting risk models for another population without prospective cohorts, which would help alleviate the health disparities caused by advances in absolute risk models. To exemplify, we adapted the lung cancer risk model $PLCO_{M2012}$, well studied in the west, for Taiwan.

**Methods:** Using Taiwanese multiple data sources, we formed an age-matched case–control study of ever-smokers (AMCCSE), estimated the number of ever-smoking lung cancer patients in 2011–2016 (NESLP2011), and synthesized a dataset resembling the population of cancer-free ever-smokers in 2010 regarding the $PLCO_{M2012}$ risk factors (SPES2010). The AMCCSE was used to estimate the overall calibration slope, and the requirement that NESLP2011 equals the estimated total risk of individuals in SPES2010 was used to handle the calibration-in-the-large problem.

**Results:** The adapted model PLCOT-1 (PLCOT-2) had an AUC of 0.78 (0.75). They had high performance in calibration and clinical usefulness on subgroups of SPES2010 defined by age and smoking experience. Selecting the same number of individuals for low-dose computed tomography screening using PLCOT-1 (PLCOT-2) would have identified approximately 6% (8%) more lung cancers than the US Preventive Services Task Forces 2021 criteria. Smokers having 40+ pack-years had an average PLCOT-1 (PLCOT-2) risk of 3.8% (2.6%).

**Conclusions:** The adapted PLCOT models had high predictive performance.

**Impact:** The PLCOT models could be used to design lung cancer screening programs in Taiwan. The methods could be applicable to other cancer models.

## Introduction

Absolute risk models estimate disease risk in an upcoming time interval based on known risk factors for a healthy individual in a population, accounting for competing causes of death (1, 2). Absolute risk models have important clinical and public health applications (3, 4). Strategies to develop, validate, and update absolute risk models have been important research topics in recent decades (2, 5, 6). Although prospective cohorts are ideal for their development, validation, and updating, the required sample size would be large and follow-up periods long if the disease incidence rate is low, such as cancers at specific sites. Methods that synthesize multiple data sources without using prospective datasets have been proposed for model development following the seminal contribution of Gail and colleagues (1, 7–11).

Indeed, both Gail and colleagues and Costantino and colleagues combined estimates of relative risks associated with certain risk factors

[1]Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan. [2]National Institute of Cancer Research, National Health Research Institutes, Zhunan, Taiwan. [3]Department of Internal Medicine, National Taiwan University Hospital and College of Medicine, National Taiwan University, Taipei, Taiwan. [4]Taiwan Lung Cancer Tissue/Specimen Information Resource Center, National Health Research Institutes, Zhunan, Taiwan. [5]School of Medicine and Institute of Medicine, Chung Shan Medical University, Taichung, Taiwan. [6]Division of Pulmonary Medicine, Department of Internal Medicine, Chung Shan Medical University Hospital, Taichung, Taiwan. [7]Institute of Biomedical Sciences, National Chung Hsing University, Taichung, Taiwan. [8]Division of Chest Medicine, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan. [9]Department of Respiratory Therapy, Chang Gung University, Taoyuan, Taiwan. [10]Department of Pulmonary and Critical Care, Xiamen Chang Gung Hospital, Xiamen, China. [11]Department of Oncology, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan. [12]Department of Internal Medicine, E-Da Cancer Hospital, School of Medicine, I-Shou University, Kaohsiung, Taiwan. [13]School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan. [14]Department of Chest Medicine, Taipei Veterans General Hospital, Taipei, Taiwan. [15]Institute of Medicine, Chung Shan Medical University Hospital, Taichung, Taiwan. [16]Division of Thoracic Surgery, Department of Surgery, Chung Shan Medical University Hospital, Taichung, Taiwan. [17]Department of Internal Medicine, National Taiwan University Hospital Hsinchu Branch, Hsinchu, Taiwan. [18]Department of Medicine, National Taiwan University Cancer Center, Taipei, Taiwan. [19]Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, National Taiwan University Hospital Yunlin Branch, Yunlin, Taiwan. [20]Department of Pulmonary and Critical Care, Chang Gung Memorial Hospital, Taoyuan, Taiwan. [21]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland. [22]Genomics Research Center, Academia Sinica, Taipei, Taiwan.

I.-S. Chang and C.A. Hsiung contributed equally to this article.

**Corresponding Authors:** Chao A. Hsiung, 35 Keyan Road, Zhunan, Miaoli County 35053, Taiwan. Phone: 372-06166, ext. 36120; Fax: 375-86467; E-mail: hsiung@nhri.org.tw; and I-Shou Chang, 35 Keyan Road, Zhunan, Miaoli County 35053, Taiwan. Phone: 372-06166, ext. 36130; E-mail: ischang@nhri.org.tw

Cancer Epidemiol Biomarkers Prev 2022;31:2208–18

**doi:** 10.1158/1055-9965.EPI-22-0281

and estimates of the baseline hazard and attributable risk to obtain estimates of the probability of developing breast cancer, using competing risk models. The modification by Costantino and colleagues used age-specific invasive breast cancer rates and attributable risk estimates from the Surveillance, Epidemiology, and End Results rather than from the Breast Cancer Detection Demonstration Project (1, 8). Recognizing a growing demand to develop and apply models for absolute risk prediction, the iCARE package builds competing risk models by synthesizing multiple data sources containing information on relative risks, the distribution of risk factors in the population, and age-specific incidence rates (11).

Chien and colleagues developed logistic regression models for predicting lung cancer occurrence in the upcoming 6 years among never-smoking Taiwanese females, based on an age-matched case–control study (AMCCS) and the age-specific 6-year lung cancer incidence rates (ASSIR) for never-smoking females in Taiwan (10). The AMCCS was used to estimate the effects of risk factors other than age and the intercept; given these effect estimates, they used the ASSIR and risk factor distributions among the controls to estimate the age effect and intercept. The AMCCS was obtained from a case–control study of lung cancer; ASSIR, accounting for competing causes of death, was estimated using the Taiwan Cancer Registry (TCR), the Taiwan Cause of Death Database (TCOD), age-specific population size, never-smoking rates in the female population and in female patients with lung cancer, and the Taiwan life table.

Because validating and updating risk models are essential toward better risk prediction models (12–14) and except for discrimination, are currently carried out using prospective cohorts, we aimed to propose methods for validating or adapting absolute risk models for another population by synthesizing multiple data sources, when no suitable prospective cohorts are available. This would help alleviate the disparities due to risk models, with or without incorporating polygenic risk scores (15, 16).

To make the presentation concrete, we exemplified the methods by adapting the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial 2012 model ($PLCO_{M2012}$) for Taiwan. It is a logistic regression model that estimates the probability of a smoker developing lung cancer in a 6-year period, using age, ethnicity, education, body mass index (BMI), chronic obstructive pulmonary disease (COPD), family history of lung cancer, personal history of cancer, smoking status, average number of cigarettes smoked per day, years smoked, and quit time (17). It was used to guide the selection of participants for low-dose computed tomography (LDCT) lung cancer screening trials (18) and was prospectively validated in the United States, Germany, Australia, Canada, U.K., Brazil, and Poland (18–28). It was also used to assess the clinical utility of polygenic risk scores for risk stratification regarding LDCT screening (29). The National Comprehensive Cancer Network 2018 guidelines approve selection based on $PLCO_{M2012}$ risk (30). However, the validation or adaptation of the $PLCO_{M2012}$ model has not been reported in Asia.

To take advantage of the excellent performance of the $PLCO_{M2012}$ model in the west and to reduce the possibility of overfitting, we considered two parsimonious approaches to adapting: Approach 1 updated only the intercept to deal with the calibration-in-the-large problem and Approach 2 derived the overall calibration slope and then handled the calibration-in-the-large problem (12, 13). For these, we formed population datasets and derived summary statistics or information for Taiwan and then used them to recalibrate and assess the risk model. We (i) constructed an age-matched case–control study of ever-smokers (AMCCSE), (ii) estimated the number of ever-smoking lung cancer patients diagnosed in 2011–2016 (NESLP2011), and (iii) synthesized a dataset resembling the population of cancer-free ever-smokers in Taiwan at the end of 2010 (SPES2010) with respect to the risk factors in the $PLCO_{M2012}$ model. In this paper, a person is cancer-free at a time point in her/his life if she/he has never been diagnosed with any cancer before that time point, and a person is a cancer survivor at a time point in her/his life if she/he has been diagnosed with certain cancer before that time point.

For Approach 1, we changed the intercept by requiring the resulting total risk of individuals in the SPES2010 equal NESLP2011. For Approach 2, we decomposed its linear predictor into two parts: the intercept and age term and the remaining term. We first used AMCCSE to conduct the "calibration slope" step that recalibrated the remaining term. Given the calibration slope, we then performed the calibration-in-the-large step that recalibrated the intercept and the age term by requiring the resulting total risk of individuals in the SPES2010 equal NESLP2011. In either approach, we assessed the performance of the adapted model in terms of discrimination, subgroup calibration, and clinical usefulness (12) using SPES2010 and other datasets.

## Materials and Methods

### AMCCSE

The ever-smoking patients with lung cancer in the AMCCSE were collected from the case–control component of the Taiwan Genetic Epidemiology Study of Lung Adenocarcinoma (GELAC) and the Taiwan Lung Cancer Pharmacogenomics Study (LCPG). The ever-smoking healthy controls were from the Taiwan Biobank (RRID: SCR_010557) and the case–control component of the GELAC. Limiting to the age range 50 to 74, we formed a total of 798 age-matched groups, where each group had exactly one case and one to five age-matched healthy controls, involving a total of 3,508 controls. **Figure 1A** presents the procedure formatting the AMCCSE, including the inclusion and exclusion criteria applied to the Taiwan Biobank. Supplementary Materials and Methods Texts S1–S4 have the details. Inclusion and exclusion criteria for individuals from the GELAC and LCPG were described in earlier publications (10, 31–33). Although blinding was not applied to this study, individuals in the Taiwan Biobank were deidentified before being provided to us.

### SPES2010

We used the Taiwan Biobank, consisting of cancer-free individuals at recruitment, to construct the dataset SPES2010, which consequently included only cancer-free individuals. We first determined the age- and sex-specific numbers of cancer-free ever-smokers in the SPES2010. It was constructed on the basis of (D1) the age- and sex-specific Taiwanese population size at the end of 2010 using Monthly Bulletin of Interior Statistics (MBIS) from the Taiwan Ministry of the Interior (34); (D2) estimates of age- and sex-specific numbers of cancer survivors (cancer prevalence) in Taiwan at the end of 2010 using the linkage of the TCR, TCOD, National Health Insurance Research Database (NHIRD); (D3) the age- and sex-specific smoking rate for the year 2010 using the Taiwan Adult Smoking Behavior Survey (ASBS); (D4) the age- and sex-specific number of ever-smokers in the Taiwan Biobank having information on all the risk factors in the $PLCO_{M2012}$ model. The procedures leading to the estimates in datasets D2 and D3 are given below in the section on data sources. These datasets are included in Supplementary Tables S1–S3. For each age and sex, Supplementary Table S1 reports the smoking rates (percentage of ever-smokers) in the population; Supplementary Table S2 reports the cancer prevalence at the end of 2010.

Because a person was either cancer-free or a cancer survivor, we used D1 and D2 (Supplementary Table S2) to obtain age- and sex-specific cancer-free population sizes by subtraction. Assuming the age- and sex-specific smoking rates in the cancer-free population approximated those in the general population (Supplementary Table S1), we report in Supplementary Table S3 the estimated age- and sex-specific numbers of cancer-free ever-smokers; Supplementary Table S3A for females, Supplementary Table S3B for males, and Supplementary Table S3C for females and males combined. They determined the age- and sex-specific population size of the SPES2010. Given an individual in SPES2010, we assigned to this individual the risk-factor profile of an ever-smoker randomly selected from the Taiwan Biobank having the same age and sex and without missing information on the risk factors in the $PLCO_{M2012}$ model. This suggests that the age- and sex-specific distribution of these risk factors in the SPES2010 resembled those in the Taiwan Biobank. **Figure 1B** outlines the above procedure.

### NESLP2011

According to the TCR and TCR Long Form (TCRLF), approximately 83% of the lung cancer patients diagnosed in 2011–2016 reported whether they were ever-smokers or never-smokers; see Supplementary Table S4A. On the basis of the age-, sex-, and calendar year–specific smoking rates among patients with lung cancer derived from the TCRLF, we estimated the age-specific numbers of ever-smoking lung cancer patients in the TCR for 2011–2016 (Supplementary Table S4B). These were used to estimate, among those aged 50 to 74 at the beginning of 2011, the number of ever-smoking lung cancer patients diagnosed in 2011–2016. Supplementary Table S4B was also used to estimate the age-specific 6-year lung cancer incidence rates among ever-smokers (ASSIRE; Supplementary Table S4C).

### Adapting the $PLCO_{M2012}$ model

Here, we only explain Approach 2, because Approach 1 is similar to the second step of Approach 2. Because AMCCSE was suitable for modifying the effects for all the risk factors other than age and the intercept and because we preferred a parsimonious approach, we decomposed the linear predictor of the $PLCO_{M2012}$ model (17, 35) into two components: its weighted sum of the intercept and the age effect, $-4.532506+0.0778868$ (Age - 62), is called the intercept-age factor. The remaining part of the linear predictor is called the non–intercept-age factor, which is a weighted sum of the effects of the other risk factors.

We fitted a logistic regression model with the intercept-age factor and the non–intercept-age factor only. Treating the former as the matching variable, we first fitted the logistic regression model, using a conditional likelihood approach (36), to the AMCCSE to obtain the OR of the non–intercept-age factor. Given the OR of the non–intercept-age factor from the first step, the second step estimated the OR of the intercept-age factor by requiring the resulting total risk of individuals aged 50 to 74 in the SPES2010 to be equal to NESLP2011.

The adapted model is called PLCOT-1 if Approach 1 is used and PLCOT-2 if Approach 2. Note that in this adaptation, what we need from SPES2010 was the distribution of the risk factors in the population and the population size. Details are in Supplementary Materials and Methods Text 5.

All computations are carried out using R language. The conditional likelihood approach to logistic regression was implemented using the clogit function in R, which was also used to obtain the 95% confidence interval (CI).

### Assessing discrimination, calibration, and clinical usefulness for the PLCOT models

We assessed discrimination for PLCOT-1 by computing the area under the receiver operating characteristic curve (AUC) using all the cases and controls from the AMCCSE, which was not used in the PLCOT-1 adaptation.

We assessed discrimination for PLCOT-2 by bootstrap. On the basis of the AMCCSE, we obtained bootstrapping optimism-corrected discrimination in terms of AUC. Here one bootstrap sample was a set of age-matched case–control groups sampled from the AMCCSE with replacement and having the same sample size as that of AMCCSE. A total of 1,000 bootstrap samples were used and the correction method is detailed in Section 5.3.4, Steyerberg (6).

Because the TCR and TCRLF during 2011–2016 are follow-up data of the Taiwanese population at the end of 2010, comparison of SPES2010 and the TCR and TCRLF provided opportunities for assessing the calibration and clinical usefulness of the PLCOT models in terms of subgroups defined by age and smoking experiences. We considered sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for risk-based criteria as well as those defined by age and smoking experiences.

Consider, for example, the condition that individuals smoked ≥ 20 pack-years, smoked within the past 15 years, and were aged 50 to 74 at the beginning of 2011. We estimated the number of lung cancer patients diagnosed in 2011–2016 satisfied this condition using the information in the TCR and TCRLF. Because this information in the TCRLF pertained to time at cancer diagnosis, correction was properly made; details are provided in Supplementary Materials and Methods Text S6. We also estimated the total PLCOT risk of individuals satisfying the condition in the SPES2010. The ratio of the former (observed) to the latter (predicted) provided the calibration assessment on this subgroup. The predicted number was also used to study clinical usefulness. For example, we report sensitivity to be the ratio of the predicted number to NESLP2011, assuming the number of cases developed during 2011–2016 among the SPES2010 equaled NESLP2011. We considered here the predicted, rather than the observed, because we wanted to compare the performance of simplified criteria with the corresponding PLCOT risk-based criteria.

The following simplified criteria appeared in the literature. The 2013 US Preventive Services Task Force (USPSTF) criterion is smoking ≥ 30 pack-years, smoking within the past 15 years, and aged 55 to 80 (USPSTF13; ref. 37). The 2021 USPSTF criterion is smoking ≥ 20 pack-years, smoking within the past 15 years, and aged 50 to 80 (USPSTF21; refs. 38, 39). Another criterion was studied in the Nederlands-Leuvens Longkanker Screenings Onderzoek (NELSON) trial (40). One met this criterion if one was aged 50 to 74, smoked at least 10 cigarettes per day for at least 30 years or 15 cigarettes per day for at least 25 years, and smoked within the past 10 years. In this study we considered USPSTF21, NELSON, 40–10 (smoked 40 or more pack-years and smoked in the past 10 years), and 30–15 (smoked ≥ 30 pack-years and smoked in the past 15 years).

### Data sources

The TCR, launched in 1979, is a population-based cancer registry collecting information on newly diagnosed cancers at all hospitals in Taiwan with 50 or more beds. Its quality has been improving and was recently reviewed (41, 42). The TCRLF has included smoking information on patients with cancer since 2011. This study considered only lung cancers that were the first invasive cancers in patients.

The TCOD includes cause-of-death information for individuals in Taiwan since 1971. It is maintained by the Department of Statistics,

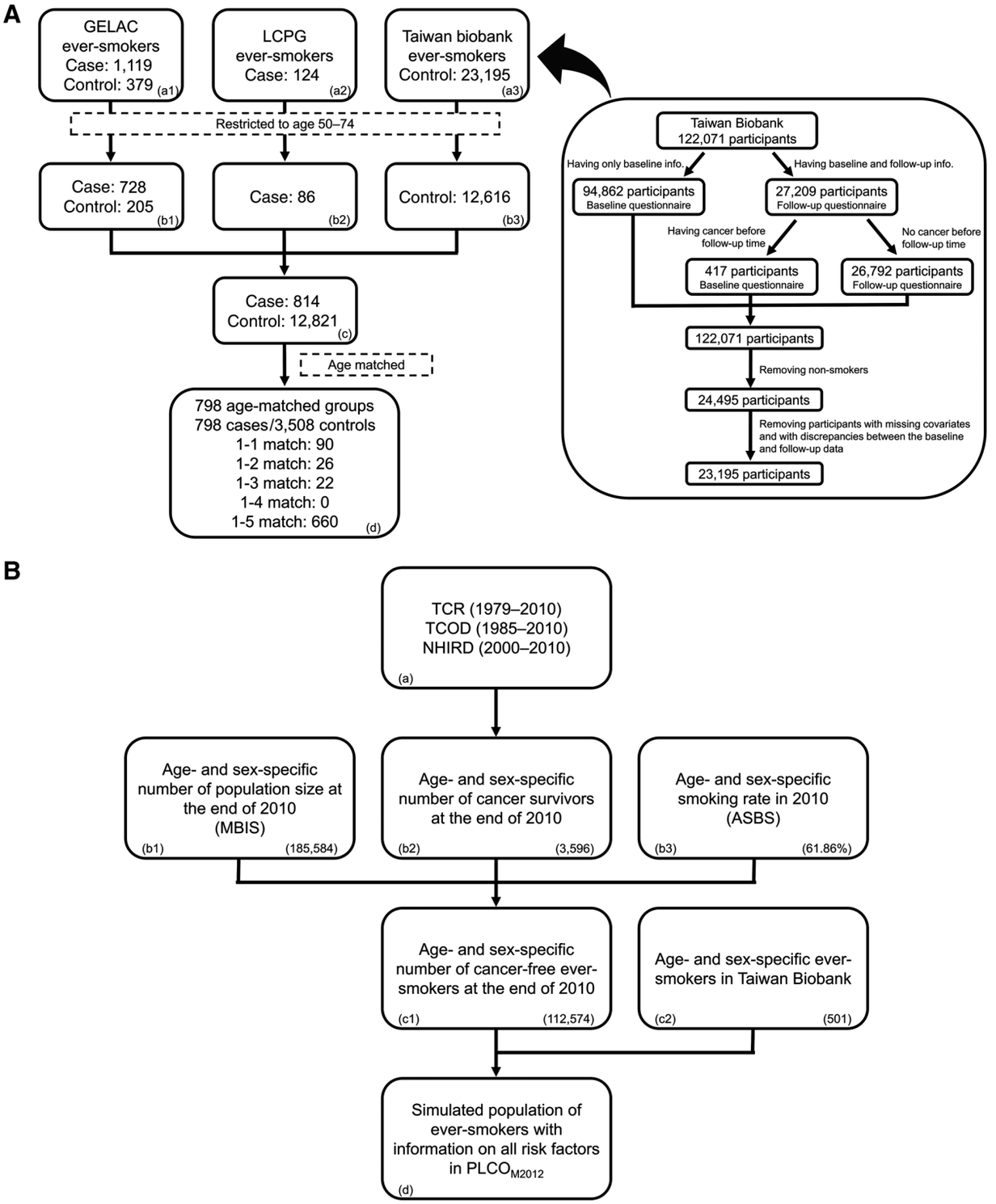


**Figure 1.**

**A,** The procedures to build AMCCSE. The right panel describes the procedure to select healthy ever-smokers from the Taiwan Biobank for use as controls in the AMCCSE; see also box (a3) in the left panel. The left panel gives the procedures resulting in the AMCCSE. The matching process from box (c) to box (d) is detailed in Supplementary Materials and Methods, Text S4. **B,** The procedures to build SPES2010. The numbers at the right-bottom corner in the boxes (b1), (b2), (b3), (c1), and (c2) refer to those for age 50 and male sex. For example, using the TCR, TCOD, and NHIRD shown in Box (a), we obtained age- and sex-specific number of cancer survivors in Box (b2); that for age 50 and male sex was 3,596. Box (c1) shows that the SPES2010 included 112,574 men having age 50, and Box (c2) indicates that their risk-factor profiles were assigned randomly based on 501 ever-smoking men of the same age from the Taiwan Biobank.

Taiwan Ministry of Health and Welfare; its quality has been previously described (43). The TCOD adopted the national identification card number (NICN) in 1985. During 1985–2018, there were 4,398,359 unique death records that included individuals' NICN, sex, birth date, death date, and cause of death. The original TCOD contains 4,405,868 records for this period; thus, < 0.2% of the data were excluded during data cleaning.

Taiwan's NHIRD was based on the administrative database of the National Health Insurance Program, which started in 1995 and has a coverage higher than 99% of its population of 23+ million. The NHIRD has been shown to be a valuable research resource (44).

Taiwan Biobank, started in 2008, is an ongoing community-based cohort of Taiwanese participants aged 30 to 70 who are cancer-free at enrollment and have information from basic physical examinations, questionnaires, and blood samples taken at enrollment. Information was also collected during follow-up appointments. This study used all the Taiwan Biobank data provided to us by November 2020, including a total of 122,071 participants; among them, 27,209 had follow-up data. Among those with follow-up data, 417 participants had a cancer diagnosis. To include participants older than 70 years, we used follow-up data for those who were cancer-free at the follow-up appointment. **Figure 1A** includes the data-cleaning process; details are provided in the Supplementary Materials and Methods. Chien and colleagues contains additional information (10).

Starting in 2007, the Taiwan ASBS reports the sample size and the proportion of ever-smokers for each survey by year, age, and sex. Using these, we obtained the number of ever-smokers in each survey and then calculated the "locally averaged" age-specific smoking rate for each year and sex. For example, the smoking rate for males aged 50 in 2010 was the proportion of ever-smokers among the male samples aged 49 to 51 in 2009–2011. Supplementary Table S1 reports the age- and sex-specific smoking rates for 2010.

Using the linkage of the TCR for 1979–2016, TCOD for 1985–2018, NHIRD for 2000–2017, we estimated the number of cancer survivors at the end of 2010. A cancer survivor is one included in the TCR for 1979–2010, not in the TCOD for 1985–2010, and in the NHIRD 2000–2010. Supplementary Table S2A presents the age- and sex-specific numbers of cancer survivors. Supplementary Table S2B reports the number of cancer survivors whose diagnoses were in the years between 1979 and 1988. Supplementary Table S2B suggests that the underestimation of cancer prevalence due to diagnoses before 1979 is likely minimal.

Cases in the GELAC were Han Chinese aged 18 or older with incident lung cancer diagnosed during 2000–2015 in Taiwan. No limitations on sex, smoking status, histology, or stage were imposed. The controls in the GELAC study were recruited from the health examination centers. The GELAC has been used to study lung cancer in never-smoking females and ever-smokers (10, 32, 33, 45). Supplementary Materials and Methods provides more information.

The LCPG recruited from health records late-stage lung cancer patients for whom epidermal growth factor receptor mutation statuses were available during 2015–2017 (10). More information about the LCPG is provided in Supplementary Materials and Methods. The structured questionnaires were administered to the GELAC and LCPG participants.

This study was approved by the institutional review board of the National Health Research Institutes in Taiwan (RRID: SCR_000335) and conforms to the Declaration of Helsinki provisions. All the datasets used in this study were provided to us after deidentification except GELAC and LCPG. All study subjects in the GELAC and LCPG provided signed informed consent prior to the commencement of this study.

### Data availability

The linkage of TCR, TCOD, and NHIRD can be performed and used for research upon approval of the Data Science Center, MOHW, Taiwan. The Taiwan Biobank dataset can be used for research upon approval of the Taiwan Biobank (https://taiwanview.twbiobank.org.tw/index). ASBS can be freely downloaded from the Health Promotion Administration, MOHW, Taiwan. Age-, year- and sex-specific population sizes can be freely downloaded from MBIS, Ministry of Interior, Taiwan. The use of datasets for the GELAC and LCPG studies need the approval of the NHRI IRB.

## Results

### the AMCCSE and SPES2010 datasets and other summary statistics

Using the Taiwan Biobank, GELAC, and LCPG, we followed the procedures in **Fig. 1A** to form the AMCCSE. Using the MBIS, ASBS, TCR, TCOD, and NHIRD, we followed the procedures in **Fig. 1B** to form the SPES2010. **Table 1** presents the characteristics of the AMCCSE and SPES2010, in view of the risk factors in the $PLCO_{M2012}$ model. Supplementary Table S5 presents the smoking-related characteristics of these data sources. **Table 1** shows that for these risk factors, their distributions among the AMCCSE cases were different from those among the controls, confirming that these were indeed risk factors for lung cancer among the ever-smokers in Taiwan. The characteristics of the SPES2010 shows that there were approximately 1,562,798 cancer-free ever-smokers aged 50 to 74 in Taiwan, and among them, more than 94% were males. These cancer-free ever-smokers accounted for approximately 27% of the Taiwanese population aged 50 to 74, which was approximately 5,765,938, according to Supplementary Table S5C. A comparison of **Table 1** with the Supplementary Table S6 in Chien and colleagues (10) suggests that COPD and family history of lung cancer were more prevalent among ever-smokers than those among never-smoking females.

It follows from Supplementary Table S4B that among those aged 50 to 74 at the beginning of 2011, the number of ever-smoking patients with lung cancer diagnosed in 2011–2016 (NESLP2011) was estimated to be 17,374. Combined with the age-specific cancer-free ever-smokers at the end of 2010 reported in Supplementary Table S3C, we report in Supplementary Table S4C the ASSIRE.

Supplementary Table S5A indicates that, according to the TCRLF dataset, approximately 45% of the patients with lung cancer in Taiwan were ever-smokers (46); that only 36% in the GELAC and LCPG studies were ever-smokers probably reflects their recruitment criteria. The difference in the age and sex distribution reported in Supplementary Table S6 and Supplementary Table S5A suggests that some selection bias exists in the Taiwan Biobank.

### The Taiwan adapted PLCOT models

By requiring that NESLP2011, being 17,374, equals the estimated total risk of individuals in SPES2010, we obtained the adapted model PLCOT-1, whose beta coefficients were the same as those of $PLCO_{M2012}$ except for the intercept. **Table 2** presents these coefficients.

By fitting a logistic regression model with the intercept-age factor and the non–intercept-age factor defined by $PLCO_{M2012}$ to the AMCCSE, we first obtained the OR 0.514 for the non–intercept-age risk factor; given this, we obtained the OR 0.859 for the intercept-age factor by requiring that NESLP2011, being 17,374, equals the

**Table 1.** Characteristics of the AMCCSE and the SPES2010.

| | AMCCSE | | | | |
|---|---|---|---|---|---|
| **Variable[a]** | **Case ($n$ = 798)** | **Control ($n$ = 3,508)** | **OR** | ***P*** | **SPES2010 ($n$ = 1,562,798)** |
| Age | 62.45 (7.07) | 61.07 (6.40) | | | 58.47 (6.66) |
| Gender | | | 0.91 | 4.8E-01 | |
| Female | 72 (9.02) | 317 (9.04) | | | 91,226 (5.84) |
| Male | 726 (90.98) | 3,191 (90.96) | | | 1,471,572 (94.16) |
| BMI | 24.09 (3.50) | 25.29 (3.32) | 0.90 | 1.5E-16 | 25.43 (3.34) |
| Education[b] | | | 0.57 | 4.1E-61 | |
| Level 1 | 534 (66.92) | 859 (24.49) | | | 260,937 (16.70) |
| Level 2 | 163 (20.43) | 1,142 (32.55) | | | 525,214 (33.61) |
| Level 3 | 0 (0) | 0 (0) | | | 0 (0) |
| Level 4 | 0 (0) | 0 (0) | | | 0 (0) |
| Level 5 | 93 (11.65) | 1,326 (37.80) | | | 663,790 (42.47) |
| Level 6 | 8 (1.00) | 181 (5.16) | | | 112,857 (7.22) |
| COPD | | | 1.55 | 4.4E-02 | |
| No | 765 (95.86) | 3,431 (97.81) | | | 1,526,156 (97.66) |
| Yes | 33 (4.14) | 77 (2.19) | | | 36,642 (2.34) |
| Family history | | | 2.23 | 3.8E-11 | |
| No | 677 (84.84) | 3,238 (92.30) | | | 1,428,700 (91.42) |
| Yes | 121 (15.16) | 270 (7.70) | | | 134,098 (8.58) |
| Smoking status | | | 1.27 | 4.3E-03 | |
| Current | 332 (41.60) | 1,324 (37.74) | | | 581,406 (37.20) |
| Former | 466 (58.40) | 2,184 (62.26) | | | 981,392 (62.80) |
| Duration of smoking | 37.19 (12.19) | 23.98 (15.24) | 1.07 | 2.4E-68 | 22.33 (14.66) |
| Smoking intensity[c] | 25.90 (15.21) | 17.90 (13.97) | 1.03 | 5.5E-39 | 17.45 (13.84) |
| Smoking quit time | 4.15 (8.28) | 9.92 (12.01) | 0.94 | 3.9E-36 | 10.17 (11.93) |

[a]Age, BMI, duration of smoking, smoking intensity, and smoking quit time are summarized in mean (sd); gender, education, COPD, family history, and smoking status are summarized in no. (%).
[b]Education was measured in six ordinal levels: less than high-school graduate (1), high-school graduate (2), some training after high school (3), some college (4), college graduate (5), and postgraduate or professional degree (6).
[c]Smoking intensity (the average number of cigarettes smoked per day).

**Table 2.** The coefficients of the linear predictor for the PLCOT risk models[a].

| | PLCOT-1 | | PLCOT-2 | |
|---|---|---|---|---|
| **Variable** | **Beta coefficient** | **OR** | **Beta coefficient** | **OR** |
| Constant | −3.83644550 | | −3.89173311 | |
| Age[b] | 0.07788680 | 1.081 | 0.06687573 | 1.069 |
| Asian | −0.46658500 | 0.627 | −0.23984514 | 0.787 |
| Education[b,c] | −0.08127440 | 0.922 | −0.04177860 | 0.959 |
| BMI[b] | −0.02741940 | 0.973 | −0.01409477 | 0.986 |
| COPD | 0.35530630 | 1.427 | 0.18264301 | 1.200 |
| Personal history of cancer | 0.45899710 | 1.582 | 0.23594462 | 1.266 |
| Family history of cancer | 0.58718500 | 1.799 | 0.30183882 | 1.352 |
| Smoking status | 0.25974310 | 1.297 | 0.13351934 | 1.143 |
| Duration of smoking[b] | 0.03173210 | 1.032 | 0.01631169 | 1.016 |
| Smoking intensity[d] | −1.82260600 | | −0.93689935 | |
| Smoking quit time[b] | −0.03085720 | 0.97 | −0.01586195 | 0.984 |

[a]The PLCOT risk is computed in the same way as the $PLCO_{M2012}$ risk except using different beta coefficients. Supplementary Materials and Methods Text S5 provide the details. Briefly, for categorical variables, multiply the variable or the level beta coefficient of the variable by 1 if the factor is present and by 0 if it is absent. For continuous variables other than smoking intensity, subtract the centering value from the person's value and multiply the difference by the beta coefficient of the variable. For smoking intensity, calculate the contribution of the variable to the model by dividing by 10, exponentiating by the power −1, centering by subtracting 0.4021541613, and multiplying this number by the beta coefficient of the variable. Add together all the previously calculated beta-coefficient products and the model constant. This sum is called the linear predictor of PLCOT, $LP_{PLCOT}$. The risk is $e^{LP_{PLCOT}}/1+e^{LP_{PLCOT}}$.
[b]Age was centered on 62 years, education on level 4, BMI on 27, duration of smoking on 27 years, and smoking quit time on 10 years.
[c]Education was measured in six ordinal levels: less than high-school graduate (1), high-school graduate (2), some training after high school (3), some college (4), college graduate (5), and postgraduate or professional degree (6).
[d]Smoking intensity (the average number of cigarettes smoked per day) had a nonlinear association with lung cancer, and this variable was transformed. See Supplementary Materials and Methods Text S5.

**Table 3.** Calibration of the PLCOT models on 4 subgroups of SPES2010 defined by age and smoking experience; clinical usefulness of these 4 simplified LDCT lung cancer screening criteria and that of the 4 PLCOT risk-based criteria selecting the same numbers of SPES2010 as these 4 simplified criteria.

**Calibration Assessment**

| | | 40–10[a] | 30–15[a] | USPSTF21[b] | NELSON[c] |
|---|---|---|---|---|---|
| PLCOT-1 | Predicted | 6,730 | 10,470 | 12,870 | 12,094 |
| | Observed | 8,167 | 11,245 | 13,801 | 12,750 |
| | Observed/Predicted | 1.21 | 1.07 | 1.07 | 1.05 |
| PLCOT-2 | Predicted | 4,471 | 7,708 | 10,194 | 9,527 |
| | Observed | 8,167 | 11,245 | 13,801 | 12,750 |
| | Observed/Predicted | 1.83 | 1.46 | 1.35 | 1.34 |

**Sensitivity, Specificity, PPV and NPV**

| | PLCOT-1 | | | | PLCOT-2 | | |
|---|---|---|---|---|---|---|---|
| Criteria | With cancer ($N$ = 17,374) | No cancer ($N$ = 1,545,424) | Predictivity | Criteria | With cancer ($N$ = 17,374) | No cancer ($N$ = 1,545,424) | Predictivity |
| 40–10[a] | | | | 40–10[a] | | | |
| Positive | 6,730 | 158,631 | PPV, 4.1% | Positive | 4,471 | 160,890 | PPV, 2.7% |
| Negative | 10,644 | 1,386,793 | NPV, 99.2% | Negative | 12,903 | 1,384,534 | NPV, 99.1% |
| | Sen.[d], 38.7% | 1-Spe.[e], 10.3% | | | Sen.[d], 25.7% | 1-Spe.[e], 10.4% | |
| PLCOT | | | | PLCOT | | | |
| ≥0.0266 | 8,193 | 157,168 | PPV, 5.0% | ≥0.0231 | 5,734 | 159,627 | PPV, 3.5% |
| <0.0266 | 9,181 | 1,388,256 | NPV, 99.3% | <0.0231 | 11,640 | 1,385,797 | NPV, 99.2% |
| | Sen.[d], 47.2% | 1-Spe.[e], 10.2% | | | Sen.[d], 33.0% | 1-Spe.[e], 10.3% | |
| 30–15[a] | | | | 30–15[a] | | | |
| Positive | 10,470 | 342,493 | PPV, 3.0% | Positive | 7,708 | 345,255 | PPV, 2.2% |
| Negative | 6,904 | 1,202,931 | NPV, 99.4% | Negative | 9,666 | 1,200,169 | NPV, 99.2% |
| | Sen.[d], 60.3% | 1-Spe.[e], 22.2% | | | Sen.[d], 44.4% | 1-Spe.[e], 22.3% | |
| PLCOT | | | | PLCOT | | | |
| ≥0.0150 | 11,912 | 341,051 | PPV, 3.4% | ≥0.0158 | 9,280 | 343,683 | PPV, 2.6% |
| <0.0150 | 5,462 | 1,204,373 | NPV, 99.5% | <0.0158 | 8,094 | 1,201,741 | NPV, 99.3% |
| | Sen.[d], 68.6% | 1-Spe.[e], 22.1% | | | Sen.[d], 53.4% | 1-Spe.[e], 22.2% | |
| NELSON[b] | | | | NELSON[b] | | | |
| Positive | 12,094 | 491,102 | PPV, 2.4% | Positive | 9,527 | 493,669 | PPV, 1.9% |
| Negative | 5,280 | 1,054,322 | NPV, 99.5% | Negative | 7,847 | 1,051,755 | NPV, 99.3% |
| | Sen.[d], 69.6% | 1-Spe.[e], 31.8% | | | Sen.[d], 54.8% | 1-Spe.[e], 31.9% | |
| PLCOT | | | | PLCOT | | | |
| ≥0.0106 | 13,811 | 489,385 | PPV, 2.7% | ≥0.0127 | 11,397 | 491,799 | PPV, 2.3% |
| <0.0106 | 3,563 | 1,056,039 | NPV, 99.7% | <0.0127 | 5,977 | 1,053,625 | NPV, 99.4% |
| | Sen.[d], 79.5% | 1-Spe.[e], 31.7% | | | Sen.[d], 65.6% | 1-Spe.[e], 31.8% | |
| USPSTF21[c] | | | | USPSTF21[c] | | | |
| Positive | 12,870 | 505,369 | PPV, 2.5% | Positive | 10,194 | 508,045 | PPV, 2.0% |
| Negative | 4,504 | 1,040,055 | NPV, 99.6% | Negative | 7,180 | 1,037,379 | NPV, 99.3% |
| | Sen.[d], 74.1% | 1-Spe[e]., 32.7% | | | Sen.[d], 58.7% | 1-Spe.[e], 32.9% | |
| PLCOT | | | | PLCOT | | | |
| ≥0.0102 | 13,968 | 504,271 | PPV, 2.7% | ≥0.0124 | 11,585 | 506,654 | PPV, 2.2% |
| <0.0102 | 3,406 | 1,041,153 | NPV, 99.7% | <0.0124 | 5,789 | 1,038,770 | NPV, 99.4% |
| | Sen.[d], 80.4% | 1-Spe.[e], 32.6% | | | Sen.[d], 66.7% | 1-Spe.[e], 32.8% | |

[a]40–10 means having smoked 40 pack-years and smoked in the past 10 years; 30–15 means having smoked 30 pack-years and smoked in the past 15 years.
[b]USPSTF21 restricted to SPES2010 means having smoked 20 pack-years and smoked in the past 15 years.
[c]One met this criterion if one was aged 50 to 74, smoked at least 10 cigarettes per day for at least 30 years or 15 cigarettes per day for at least 25 years, and smoked within the past 10 years.
[d]Sen. = Sensitivity.
[e]1-Spe. = 1-specificity.

estimated total risk of individuals in SPES2010. This resulted in the adapted PLCOT-2 model. **Table 2** also includes the beta coefficients of PLCOT-2. The OR of the non–intercept-age factor had 95% CI (0.443–0.585). Details are in Supplementary Materials and Methods Text S5.

### Performance of the PLCOT models

The AUC for PLCOT-1 was 0.7776, and the bootstrapping optimism-corrected AUC of PLCOT-2 was 0.7549, and its apparent AUC was 0.7553.

**Table 3** evaluates the PLCOT models using SPES2010 and NESLP2011. It reports the calibration assessment for both PLCOT-1 and PLCOT-2 on subgroups defined by 40–10, 30–15, USPSTF21, and NELSON and shows that PLCOT-1 had excellent subgroup calibration and that PLCOT-2 had good calibration.

For each of the PLCOT models, **Table 3** also compares the sensitivity, specificity, PPV, and NPV of the four simplified LDCT lung cancer screening criteria with those of the four PLCOT risk-based criteria selecting the same number of individuals for screening. They

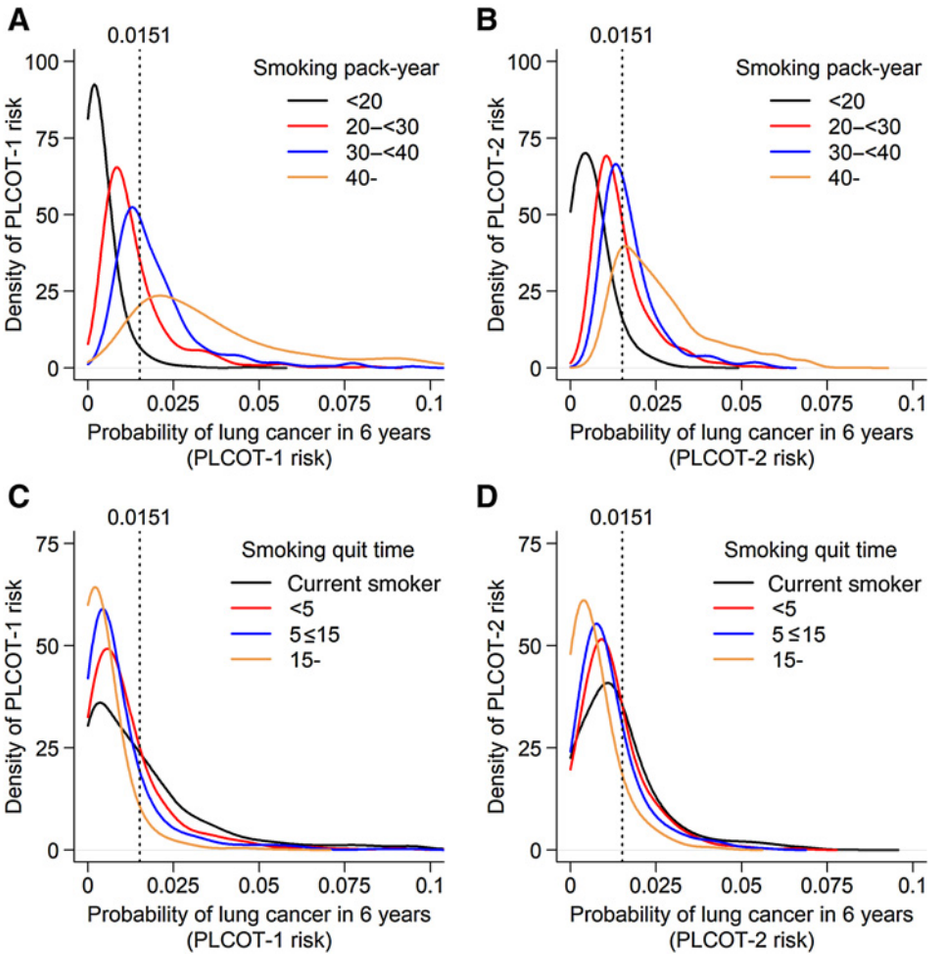


**Figure 2.**
The distributions of PLCOT-1 risks (**A**) and PLCOT-2 risks (**B**) in SPES2010 according to smoking pack-year, and the distribution of PLCOT-1 risks (**C**) and PLCOT-2 risks (**D**) in SPES2010 according to smoking quit time.

show that the corresponding PLCOT risk-based criteria performed better than their simplified criteria counterparts in terms of these measurements. Consider USPSTF21, for example. A total of 518,239 (33.1%) individuals among all the 1,562,798 individuals in the SPES2010 satisfied the USPSTF21 criteria. The PLCOT-2 risk threshold 0.0124 would select the same number of individuals from SPES2010. The latter had a sensitivity of 66.7% and a PPV of 2.2%, while the former had a sensitivity of 58.7% and a PPV of 2.0%. Thus, the PLCOT model risk-based criteria would have potentially identified approximately 8% more cancers than the USPSTF21 criteria if the same number of individuals had been selected for screening.

To help communicate sensitivity and PPV, we present in Supplementary Figures S1 and S2 the Lorenz curves for the PLCOT-based risk distribution on the SPES2010, which plot predicted total lung cancer incidence against the number of individuals at highest risk (6).

**Figure 2** presents the densities of risk among subgroups of SPES2010 by smoking levels, **Figs. 2A** and **B** regarding pack-years smoked and **Figs. 2C** and **D** regarding quit time. While **Fig. 2** shows that both PLCOT models vary with smoking experiences, they suggest that PLCOT-1 risks vary more, in line with the fact that the OR of the non–intercept-age factor is less than 1. Indeed, for PLCOT-2, the mean 6-year risk was 2.7% for those who smoked 40+ pack-years, 1.7% between 30 and 40 pack-years, 1.5% between 20 and 30 pack-years, 0.6% less than 20 pack-years and the corresponding mean risks for PLCOT-1 were 3.8%, 2.0%, 1.4%, and 0.3%, respectively. Similar results hold for quit time.

For the clinically relevant 6-year lung cancer risk threshold of 0.0151 (35), **Fig. 2A** and **B** show that the vast majority of those who smoked 40+ pack-years had risks higher than 0.0151 and very few of those who smoked less than 20 pack-years had risks above this threshold. **Figure 2C** and **D** suggest that the proportion of high-risk individuals decreased persistently with the number of cessation years.

**Figure 3** presents studies about smoking cessation effects on lung cancer risk. **Figure 3A** and **B** compare the densities of the PLCOT risks of lung cancer among 60-year-old current smokers in the SPES2010 who had smoked more than 10 years with more than 20 cigarettes per day with those of the same people if they had quit smoking when they were 50 years of age. **Figure 3C** and **D** compare the age-specific means of PLCOT risk of lung cancer for each age from 50 to 74 among the 50-year-old current smokers in SPES2010 who smoked at least 20 cigarettes per day with those among the same people if they had quit smoking at age 50. **Figure 3A** and **B** suggest a great risk reduction in terms of the 0.0151 threshold. Indeed, nearly no one had risks less than the threshold for either model; however, if they had quit smoking at 50, then 64% (55%) of them had risk less than 0.0151 under PLCOT-1 (PLCOT-2). **Figure 3C** and **D** suggest that the reduction in lung cancer risk increased considerably with quitting time, with PLCOT-1 reporting a much larger reduction.

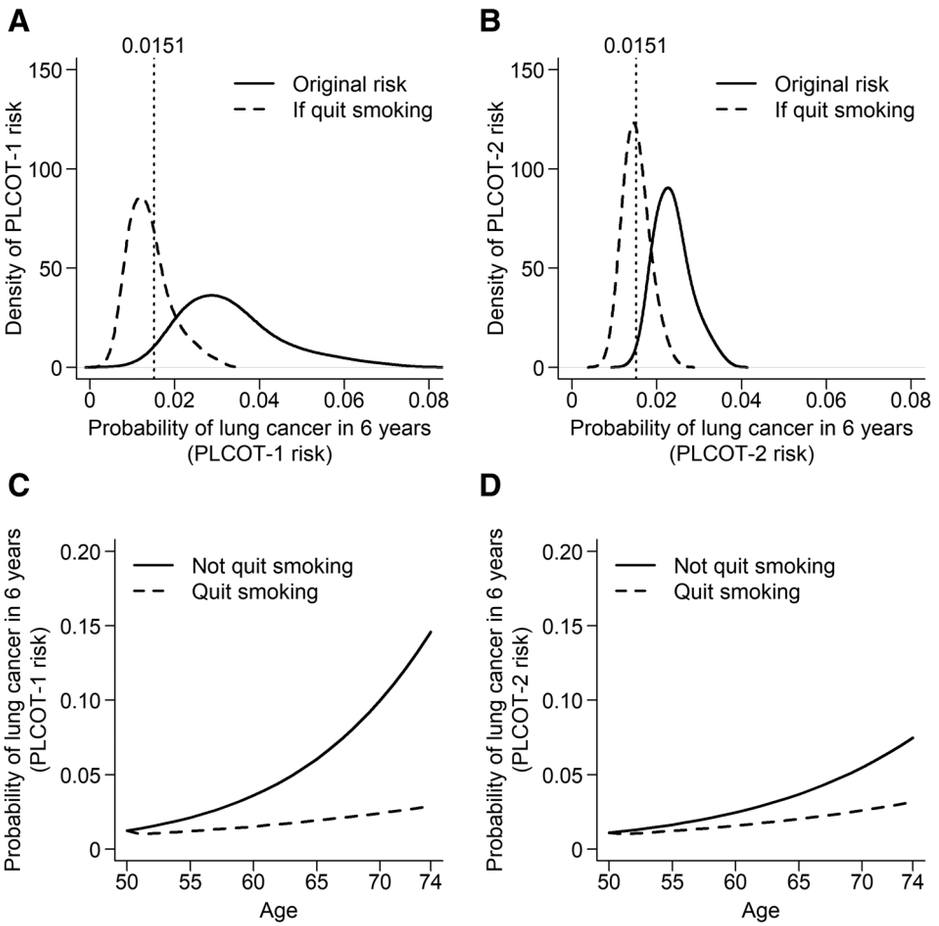


**Figure 3.**

The distributions of PLCOT-1 (**A**) risks and PLCOT-2 risks (**B**) of lung cancer among SPES2010 current smokers who were 60-year-old and had smoked more than 10 years with average number cigarettes smoked per day more than 20 (solid lines) and those among the same people if they had quit smoking when they were 50 years of age (dashed lines). The age-specific means of PLCOT-1 risks (**C**) and PLCOT-2 risks (**D**) of lung cancer for each age from 50 to 74 among the 50-year-old current smokers in SPES2010 who smoked at least 20 cigarettes per day (solid lines) and those among the same people if they had quit smoking at age 50 (dashed lines).

## Discussion

This paper presented methods for adapting risk prediction models without prospective cohorts. Having formed an AMCCSE, estimated the number of ever-smoking lung cancer patients in 2011–2016, and prepared the dataset resembling the population of cancer-free ever-smokers at the end of 2010, we exemplified the methods by adapting the $PLCO_{M2012}$ model for Taiwan use. The PLCOT-1 model had an AUC of 0.78 and excellent performance in terms of subgroup calibration and clinical usefulness. The PLCOT-2 model had a bootstrapping optimism-corrected AUC of 0.75 and quite good performance in terms of subgroup calibration and clinical usefulness. Using these models, we reported risk distributions according to smoking exposure levels and described the effects of smoking cessation on risk reduction. To the best of our knowledge, the PLCOT models represent the first attempts to recalibrate or validate the $PLCO_{M2012}$ model in Asia.

In line with the literature that risk model updating methods range from simply updating the intercept to reestimating for each risk factor (5, 6), we synthesized datasets so that we could conduct updating at roughly three levels of sophistication. At the basic level, Approach 1 used SPES2010 and NESLP2011 to modify only the intercept to handle the calibration-in-the-large problem, resulting in PLCOT-1. At the second level, Approach 2 used the additional dataset AMCCSE to consider both the overall calibration slope and calibration-in-the-large problem, resulting in PLCOT-2. At the next level, we could replace the calibration slope step in Approach 2 by reestimating the effect of each risk factor other than the intercept and age, using the same AMCCSE. Following the approach in Chien and colleagues (10), we could also replace the calibration-in-the-large step in Approach 2 by reestimating the age effect and intercept using the age-specific 6-year lung cancer incidence rates among cancer-free ever-smokers, reported in Supplementary Table S4C. However, extensive model revision requires larger datasets in general. Indeed, we considered several such extensive model revisions and found poor subgroup calibration.

Although PLCOT-1 performed a little better than PLCOT-2 in terms of discrimination, subgroup calibration, sensitivity, specificity, PPV, and NPV, we presented both models in this paper because we intended to exemplify the methodology more fully, and we think future studies using prospective cohorts would give more conclusive comparisons. Indeed, we expect to conduct a validation study when more follow-up data are collected prospectively from the Taiwan Biobank. Because the cases used in adapting the PLCOT models do not overlap with the cases to be developed in the Taiwan Biobank and because the Taiwan Biobank has a large dataset, we could use datasets nonoverlapping with those used for adaptation to give an independent validation study and obtain a more conclusive comparison of PLCOT-1 and PLCOT-2; in particular, we will pay attention to the

calibration and discrimination around or above the risk thresholds relevant to the LDCT lung cancer screening thresholds. Note that only 6 ever-smokers aged 50 to 74 in the Taiwan Biobank developed lung cancer based on the current follow-up dataset.

Because the smoking rate among males has been decreasing in Taiwan, especially since the 2009 implementation of the Tobacco Hazards Prevention Act (47), constantly updating risk models involving smoking exposure is especially desirable.

This study takes advantage of datasets and information from multiple sources in Taiwan, which have been shown to be valuable research resources. A particular strength of this study is that the TCR from 2011 to 2016 provided smoking status information on approximately 83% of the lung cancer patients in the TCR and that among the ever-smokers, 82% had their number of pack-years smoked and quit time available; see Tables S4A and S5A.

This paper considered an age-matched case–control study because the GELAC was initially a frequency matched design. However, it seems possible to extend the current methods to the situation where cases and controls are independently sampled from their respective populations.

The methods described in this paper could adapt other well-studied absolute risk models for different populations, which would help alleviate health disparities due to the lack of risk prediction models (15). For example, the Gail model for breast cancer among Asian American females (1, 8, 48), lung cancer risk models studied in Ten Haaf and colleagues (21) and Katki and colleagues (22), and more recent breast and lung cancer risk models that incorporated polygenic risk scores (29, 49) could be validated or adapted for Taiwan use following the same methods of this paper.

Although this paper provided information useful for designing lung cancer screening programs in Taiwan, it implicitly assumed the efficacy reported in the US National Lung Screening Trial that a reduction of 20% in lung cancer mortality was observed in the LDCT screening arm (50). Eventually, an efficacy study of LDCT screening in Taiwan is desirable, and the results of this paper might be useful in such a study.

This paper presented methodologies for recalibrating or adapting risk models by synthesizing data sources without prospective cohorts and offered preliminary information useful for policy-making on designing lung cancer screening programs in Taiwan. Further studies regarding implementation, such as cost-effectiveness, are warranted.

## Authors' Disclosures

K.Y. Chen reports personal fees from AstraZeneca, Roche, Boehringer Ingelheim, Eli Lilly, Pfizer, Novartis, Merck Sharp & Dohme, Chugai Pharma; and personal fees from Takeda outside the submitted work. H.H. Hung reports grants from Ministry of Science and Technology during the conduct of the study. C.J. Chen reports grants from Ministry of Science and Technology during the conduct of the study. I.S. Chang reports grants from Ministry of Health and Welfare, Taiwan during the conduct of the study. C.A. Hsiung reports grants from Department of Health Taiwan, Ministry of Health and Welfare Taiwan; and grants from Ministry of Science and Technology Taiwan during the conduct of the study. No disclosures were reported by the other authors.

## Disclaimer

The funders had no role in the study design, data collection and analysis, decision to publish, or manuscript preparation.

## Authors' Contributions

**L.-H. Chien:** Conceptualization, data curation, formal analysis, investigation, methodology, writing–original draft, writing–review and editing. **T.-Y. Chen:** Data curation, formal analysis, investigation, writing–review and editing. **C.-H. Chen:** Data curation, formal analysis, investigation, writing–review and editing. **K.-Y. Chen:** Data curation, investigation, writing–review and editing. **C.-F. Hsiao:** Data curation, investigation, writing–review and editing. **G.-C. Chang:** Data curation, investigation, writing–review and editing. **Y.-H. Tsai:** Data curation, investigation, writing–review and editing. **W.-C. Su:** Data curation, investigation, writing–review and editing. **M.-S. Huang:** Data curation, investigation, writing–review and editing. **Y.-M. Chen:** Data curation, investigation, writing–review and editing. **C.-Y. Chen:** Data curation, investigation, writing–review and editing. **S.-K. Liang:** Data curation, investigation, writing–review and editing. **C.-Y. Chen:** Data curation, investigation, writing–review and editing. **C.-L. Wang:** Data curation, investigation, writing–review and editing. **H.-H. Hung:** Data curation, formal analysis, writing–review and editing. **H.-F. Jiang:** Data curation, formal analysis, writing–review and editing. **J.-W. Hu:** Data curation, writing–review and editing. **N. Rothman:** Conceptualization, writing–review and editing. **Q. Lan:** Conceptualization, writing–review and editing. **T.-W. Liu:** Conceptualization, resources, data curation, funding acquisition, investigation, project administration, writing–review and editing. **C.-J. Chen:** Conceptualization, resources, data curation, funding acquisition, investigation, project administration, writing–review and editing. **P.-C. Yang:** Conceptualization, resources, data curation, funding acquisition, investigation, project administration, writing–review and editing. **I.-S. Chang:** Conceptualization, resources, data curation, formal analysis, supervision, funding acquisition, investigation, methodology, writing–original draft, project administration, writing–review and editing. **C.A. Hsiung:** Conceptualization, resources, data curation, formal analysis, supervision, funding acquisition, `investigation, methodology, writing–original draft, project administration, writing–review and editing.

## Acknowledgments

Part of the data analyzed in this study was provided by and analyzed on site in the Health and Welfare Data Science Center, MOHW, Taiwan. This study was supported by the MOHW (Project grants DOH95-TD-G-111–015 (to C.A. Hsiung), DOH101-TD-PB-111-TM015 (C.A. Hsiung), DOH102-TD-PB-111-TM024 (C.A. Hsiung), MOHW103-TDU-PB-211–144003 (to C.A. Hsiung), MOHW106-TDU-B-212–144013 (to I.S. Chang), MOHW107-TDU-B-212–114026 (to I.S. Chang)), NHRI (Project grants 95A1-BSAP01–002 (to C.A. Hsiung), NHRI-PH-110-GP-04 (to C.A. Hsiung), NHRI-PH-110-GP-01 (to C.A. Hsiung), and the Ministry of Science and Technology (MOST 103–2325-B-400–023 (to C.A. Hsiung), MOST 104–2325-B-400–012 (to C.A. Hsiung), MOST 105–2325-B-400–010 (to C.A. Hsiung), MOST 109–2740-B-400–002 (to C.A. Hsiung).

The authors thank Ms. Chia-Yu Chen and Fang-Yu Tsai for technical assistance.

The publication costs of this article were defrayed in part by the payment of publication fees. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

## Note

Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Received May 17, 2022; revised July 20, 2022; accepted September 20, 2022; published first September 21, 2022.

## References

1. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81:1879–86.
2. Pfeiffer RM, Gail MH. Absolute risk: methods and applications in clinical management and public health. First ed: Chapman and Hall/CRC; 2017.225 p.
3. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med 2008;358:2796–803.
4. Gail MH. Personalized estimates of breast cancer risk in clinical practice and public health. Stat Med 2011;30:1090–104.
5. Harrell FE Jr. Regression Modeling Strategies. Second ed: Springer; 2015.

6. Steyerberg EW. Clinical prediction models, a practical approach to development, validation, and updating. Second ed: Springer; 2019.
7. Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, et al. Tamoxifen for prevention of breast cancer: report of the national surgical adjuvant breast and bowel project P-1 study. J Natl Cancer Inst 1998;90: 1371–88.
8. Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst 1999;91:1541–8.
9. Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, et al. The LLP risk model: an individual risk prediction model for lung cancer. Br J Cancer 2008;98:270–6.
10. Chien LH, Chen CH, Chen TY, Chang GC, Tsai YH, Hsiao CF, et al. Predicting lung cancer occurrence in never-smoking females in Asia: TNSF-SQ, a prediction model. Cancer Epidemiol Biomarkers Prev 2020;29:452–9.
11. Pal Choudhury P, Maas P, Wilcox A, Wheeler W, Brook M, Check D, et al. iCARE: An R package to build, validate, and apply absolute risk models. PLoS One 2020;15:e0228198.
12. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35:1925–31.
13. Cox DR. Two further applications of a model for binary regression. Biometrika 1958;45:562–5.
14. Harrell FE Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.
15. Kullo IJ, Lewis CM, Inouye M, Martin AR, Ripatti S, Chatterjee N. Polygenic scores in biomedical research. Nat Rev Genet 2022;23:524–32.
16. Paulus JK, Wessler BS, Lundquist CM, Kent DM. Effects of race are rarely included in clinical prediction models for cardiovascular disease. J Gen Intern Med 2018;33:1429–30.
17. Tammemagi 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.
18. Tammemagi MC, Schmidt H, Martel S, McWilliams A, Goffin JR, Johnston MR, et al. Participant selection for lung cancer screening by risk modelling [the Pan-Canadian Early Detection of Lung Cancer (PanCan) study[: a single-arm, prospective study. Lancet Oncol 2017;18:1523–31.
19. Li K, Husing A, Sookthai D, Bergmann M, Boeing H, Becker N, et al. Selecting high-risk individuals for lung cancer screening: a prospective evaluation of existing risk models and eligibility criteria in the German EPIC cohort. Cancer Prev Res 2015;8:777–85.
20. Weber M, Yap S, Goldsbury D, Manners D, Tammemagi M, Marshall H, et al. Identifying high risk individuals for targeted lung cancer screening: independent validation of the PLCOm2012 risk prediction tool. Int J Cancer 2017;141:242–53.
21. Ten Haaf K, Jeon J, Tammemagi 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.
22. Katki HA, Kovalchik SA, Petito LC, Cheung LC, Jacobs E, Jemal A, et al. Implications of nine risk prediction models for selecting ever-smokers for computed tomography lung cancer screening. Ann Intern Med 2018;169:10–9.
23. Crosbie PA, Balata H, Evison M, Atack M, Bayliss-Brideaux V, Colligan D, et al. Second round results from the Manchester 'Lung Health Check' community-based targeted lung cancer screening pilot. Thorax 2019;74:700–4.
24. Crosbie PA, Balata H, Evison M, Atack M, Bayliss-Brideaux V, Colligan D, et al. Implementing lung cancer screening: baseline results from a community-based 'Lung Health Check' pilot in deprived areas of Manchester. Thorax 2019;74: 405–9.
25. Kavanagh J, Liu G, Menezes R, O'Kane GM, McGregor M, Tsao M, et al. Importance of long-term low-dose CT follow-up after negative findings at previous lung cancer screening. Radiology 2018;289:218–24.
26. Aggarwal R, Lam ACL, McGregor M, Menezes R, Hueniken K, Tateishi H, et al. Outcomes of long-term interval rescreening with low-dose computed tomography for lung cancer in different risk cohorts. J Thorac Oncol 2019;14:1003–11.
27. Teles G, Macedo ACS, Chate RC, Valente VAT, Funari MBG, Szarf G. LDCT lung cancer screening in populations at different risk for lung cancer. BMJ Open Respir Res 2020;7:e000455.
28. Ostrowski M, Binczyk F, Marjanski T, Dziedzic R, Pisiak S, Malgorzewicz S, et al. Performance of various risk prediction models in a large lung cancer screening cohort in Gdansk, Poland—a comparative study. Transl Lung Cancer Res 2021; 10:1083–90.
29. Hung RJ, Warkentin MT, Brhane Y, Chatterjee N, Christiani DC, Landi MT, et al. Assessing lung cancer absolute risk trajectory based on a polygenic risk model. Cancer Res 2021;81:1607–15.
30. Tammemagi MC. Selecting lung cancer screenees using risk prediction models—where do we go from here. Transl Lung Cancer Res 2018;7:243–53.
31. Chang IS, Jiang SS, Yang JC, Su WC, Chien LH, Hsiao CF, et al. Genetic modifiers of progression-free survival in never-smoking lung adenocarcinoma patients treated with first-line tyrosine kinase inhibitors. Am J Respir Crit Care Med 2017; 195:663–73.
32. Hsiung CA, Lan Q, Hong YC, Chen CJ, Hosgood HD, Chang IS, et al. The 5p15.33 locus is associated with risk of lung adenocarcinoma in never-smoking females in Asia. PLos Genet 2010;6.
33. Lan Q, Hsiung CA, Matsuo K, Hong YC, Seow A, Wang Z, et al. Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. Nat Genet 2012;44:1330–5.
34. Interior TMot. [cited 2020 May 7]. Available from: https://www.ris.gov.tw/app/portal/346.
35. Tammemagi MC, Church TR, Hocking WG, Silvestri GA, Kvale PA, Riley TL, et al. Evaluation of the lung cancer risks at which to screen ever- and never-smokers: screening rules applied to the PLCO and NLST cohorts. PLoS Med 2014;11:e1001764.
36. Gail MH, Lubin JH, Rubinstein. LV. Likelihood calculations for matched case–control studies and survival studies with tied death times. Biometrika 1980;68: 703–7.
37. Moyer VA. Screening for lung cancer: US Preventive Services Task Force recommendation statement. Ann Intern Med 2014;160:330–8.
38. USPSTF KAH, Davidson KW, Mangione CM, Barry MJ, Cabana M, et al. Screening for lung cancer: US Preventive Services Task Force recommendation statement. JAMA 2021;325:962–70.
39. Meza R, Jeon J, Toumazis I, Ten Haaf K, Cao P, Bastani M, et al. Evaluation of the benefits and harms of lung cancer screening with low-dose computed tomography: modeling study for the US Preventive Services Task Force. JAMA 2021; 325:988–97.
40. 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.
41. Chiang CJ, You SL, Chen CJ, Yang YW, Lo WC, Lai MS. Quality assessment and improvement of nationwide cancer registration system in Taiwan: a review. Jpn J Clin Oncol 2015;45:291–6.
42. Kao CW, Chiang CJ, Lin LJ, Huang CW, Lee WC, Lee MY, et al. Accuracy of long-form data in the Taiwan Cancer Registry. J Formos Med Assoc 2021;120: 2037–41.
43. Lu TH, Lee MC, Chou MC. Accuracy of cause-of-death coding in Taiwan: types of miscoding and effects on mortality statistics. Int J Epidemiol 2000;29:336–43.
44. Hsing AW, Ioannidis JP. Nationwide population science: lessons from the Taiwan National Health Insurance Research database. JAMA Intern. Med. 2015;175:1527–9.
45. Chang CH, Hsiao CF, Chang GC, Tsai YH, Chen YM, Huang MS, et al. Interactive effect of cigarette smoking with human 8-oxoguanine DNA N-glycosylase 1 (hOGG1) polymorphisms on the risk of lung cancer: a case–control study in Taiwan. Am J Epidemiol 2009;170:695–702.
46. Tseng CH, Tsuang BJ, Chiang CJ, Ku KC, Tseng JS, Yang TY, et al. The relationship between air pollution and lung cancer in nonsmokers in Taiwan. J Thorac Oncol 2019;14:784–92.
47. Chiang CY, Chang HY. A population study on the time trend of cigarette smoking, cessation, and exposure to secondhand smoking from 2001 to 2013 in Taiwan. Popul Health Metr 2016;14:38.
48. Matsuno RK, Costantino JP, Ziegler RG, Anderson GL, Li H, Pee D, et al. Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women. J Natl Cancer Inst 2011;103:951–61.
49. Kapoor PM, Mavaddat N, Choudhury PP, Wilcox AN, Lindstrom S, Behrens S, et al. Combined associations of a polygenic risk score and classical risk factors with breast cancer risk. J Natl Cancer Inst 2021;113:329–37.
50. National Lung Screening Trial Research T, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365:395–409.