**Supplementary Materials and Methods**

Supplementary Materials and Methods Text S1, Text S2, and Text S3, provide further details on the Taiwan Biobank dataset, the Genetic Epidemiology study of lung adenocarcinoma in Taiwan, and the Lung Cancer Pharmacogenomics Study in Taiwan. Supplementary Materials and Methods Text S4 details the procedure forming the age-matched case-control study of ever-smokers in Taiwan. Supplementary Materials and Methods Text S5 details the model recalibration procedures. Supplementary Materials and Methods Text S6 provides more information regarding Taiwan Cancer Registry Long Form.

Text S1. Taiwan Biobank baseline and follow-up data

Taiwan Biobank, started in 2008, is an ongoing community-based cohort of Taiwanese participants aged 30—70 who are cancer-free at enrollment and have data 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 before November 2020, including a total of 122,071 participants; among them, 27,209 had follow-up data. For individuals having follow-up data but having no cancer diagnosis before the follow-up appointment, this study used their information collected at the planned follow-up time; otherwise, this study used their data collected at enrollment. Among the 27,209 participants with follow-up data, 26,792 had no cancer diagnosis by the follow-up time. Eliminating never-smokers resulted in 24,495 ever-smokers. Eliminating those whose follow-up data were not consistent with his/her baseline data regarding the risk factors in the PLCO 2012 and those having missing covariates resulted in 23,195 participants in Figure 1A.

Now, we describe the steps to eliminate the inconsistencies between the baseline and follow-up data. If the education codes between baseline and follow-up had a difference larger than 1 for a participant, we deleted this participant; if the difference was 1, we chose the education level at the follow-up time. If a participant had a family history of lung cancer or COPD or was an ever-smoker according to his/her baseline data or follow-up data, then this participant was considered to have a family history of lung cancer or COPD or to be an ever-smoker, even though inconsistencies existed.

Text S2　GELAC (Genetic Epidemiology study of lung adenocarcinoma in Taiwan)

Cases in GELAC were Taiwan Han Chinese aged 18 or above with incident lung cancer diagnosed in the period from 2000 to 2015 in one of six medical centers in Taiwan. No limitations on gender, smoking status, histology, or stage were imposed in the enrollment. The controls in the case-control study component of the GELAC were recruited from the health examination centers of these hospitals, following an age and gender frequency matching design. Although efforts had been made to conduct education and ethnicity matching in the beginning, we found it unsuccessful and gave up later on. Beginning in 2008, lung cancer patients treated by resection were recruited from another medical center, China Medical University Hospital, Taichung, Taiwan. A structured questionnaire was administered to each of the above participants by a trained nurse. After data cleaning, there were 1119 ever-smoking lung cancer patients and 379 ever-smoking healthy controls; see Table S5C.

Data from GELAC have been used to study epidemiologic and genetic risk factors for lung cancer in never-smoking females and in ever-smokers; more information about GELAC can be found there [1-7].

Text S3 LCPG (Lung Cancer Pharmacogenomics Study)

The Lung Cancer Pharmacogenomics Study (LCPG), supported by the National Research Program of Biopharmaceuticals, Ministry of Science and Technology, Taiwan, recruited late stage lung cancer patients whose first line treatments were chemotherapy or targeted therapy in the period 2015-2017 in National Taiwan University Hospital; National Taiwan University Hospital Hsinchu Branch; National Taiwan University Hospital Yunlin Branch; Chang-Gung Memorial Hospital, Lin-Kou; and National Cheng-Kung University Hospital. A total of 447 patients were recruited to the LCPG and among them 124 were ever-smokers. More information about the LCPG can be found in an earlier study [8].

Text S4 AMCCSE (Age-matched case-control study of ever-smokers)

Using the ever-smoking lung cancer patients included in the case-control component of the GELAC and the LCPG and the ever-smoking healthy controls included in the case-control component of the GELAC and the Taiwan Biobank, we formed the age-matched case-control study of ever-smokers. Lung cancer patients are referred to as cases. We first linearly ordered all the patients aged 50—74 with those from the GELAC preceding those from the LCPG and linearly ordered all the healthy controls aged 50—74 with those from the GELAC preceding those from the Taiwan Biobank. We carried out five rounds of matching process. In the first round of matching, given a case, exactly one control of the same age was selected and matched to the case if such a control existed; if there were two controls of the same age, the one preceded had the priority of being selected and matched. The above matching was carried out for each case in the order of precedence. All the controls matched to the cases in the first round were deleted in the second round of matching, which was carried out in the same manner as the first round. Continuing the matching process for a total of five rounds, we obtained 798 age-matched case-control groups.

Text S5. Complete recalibration procedures for Approach 2

The PLCOM2012 model for Asian ever-smokers without personal cancer history specifies that the probability of lung cancer diagnosis in the upcoming six years for an ever-smoker with age  and risk factor  is



Here,



with (ethnicity, education – 4, BMI – 27, COPD, family history of lung cancer, personal history of cancer, smoking status, (10/(average number of cigarettes smoked per day)) – 0.4021541613, years smoked – 27, quit time – 10), =10, = age- 62. The constants , , can be read out from the above expression. The exact definition of these risk factors appeared in Table 2, Tammemagi et al.[9] and in the Supplementary Information of Tammemagi et al.[10]. Education in PLCOM2012 model 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), postgraduate or professional degree (6).

To consider the PLCOM2012 model in Taiwan, we used the same definitions of these risk factors. That there are few people with education level (3) and (4) in Taiwan suggests room for future improvement in coding education.

Let . To adapt the PLCOM2012 model for Taiwanese use, we considered a logistic regression model with the linear predictor . Namely, the adapted model PLCOT was of the form



Noting that individuals in each age-matched group in the AMCCSE shared the same value on the covariate , we first estimated the slope  by applying conditional maximum likelihood[11] to the AMCCSE. Given the estimated , our second step was to estimate the “intercept”  by requiring the equality between the NESLP2011, which was 17,374, and the sum of PLCOT risks of individuals in the SPES2010.

The liner predictor for the PLCOT model is



The 95% confidence interval for the calibration slope  is (0.443068232283459, 0.58501941078681).

Note that this parsimonious approach adapts those in the literature [12, 13].

Text S6. Taiwan Cancer Registry Long Form (TCRLF)

Beginning in 2011, the TCRLF collected information on smoking experience for each cancer patient in the Taiwan Cancer Registry (TCR). For each of the patients in the TCR, there may be more than one record in the TCRLF for this patient. In this study, we chose the record that reports more serious smoking behavior (pack-years first and quitting time next) and resulted in 19688 patients having information on the number of pack-years smoked and the number of years since quitting. If we had ignored the patients having more than one record in the TCRLF, there were 19075 patients having information on the number of pack-years smoked and the number of years since quitting. The main results of this paper remain unchanged when the latter data were used.

Table S4A indicates that over 83% of the patients in the TCR had their smoking status (never-smoking or ever-smoking) known and Table S5A indicates that among them, 44.60% were ever-smokers, according to the TCRLF during 2011—2016. Table S4B reports the estimated age- and sex-specific number of ever-smoking lung cancer patients in the TCR, using TCRLF. It was done by first estimating the age- and sex-specific percentage of ever-smokers using those having known smoking status information in the TCRLF and then multiplying this percentage with the corresponding age- and sex-specific number of lung cancer patients in the TCR.

We modified the above-mentioned procedure to estimate the number of lung cancer patients diagnosed in 2011—2016 who satisfied conditions defined by age limits and smoking experience at the beginning of 2011. Because smoking information of a cancer patient in the TCRLF pertained to the time at cancer diagnosis, the number of pack-years and quit time may be reduced and former smokers may become current smokers, etc. if the reference calendar year 2011 was used. For example, the status of a 60-year-old current smoking lung cancer patient having smoked 42 years and 20 cigarettes per day and was diagnosed with lung cancer in 2016 according to the TCRLF would have smoked 42-5.5=36.5 years at the beginning of 2011.

**References**

1. Lo YL, Hsiao CF, Chang GC, Tsai YH, Huang MS, Su WC, et al. Risk factors for primary lung cancer among never smokers by gender in a matched case-control study. Cancer causes & control : CCC. 2013;24(3):567-76. doi: 10.1007/s10552-012-9994-x. PubMed PMID: 22729933.

2. 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 genetics. 2010;6(8). doi: 10.1371/journal.pgen.1001051. PubMed PMID: 20700438; PubMed Central PMCID: PMC2916850.

3. 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. Nature genetics. 2012;44(12):1330-5. doi: 10.1038/ng.2456. PubMed PMID: 23143601.

4. Wang Z, Seow WJ, Shiraishi K, Hsiung CA, Matsuo K, Liu J, et al. Meta-analysis of genome-wide association studies identifies multiple lung cancer susceptibility loci in never-smoking Asian women. Human molecular genetics. 2016;25(3):620-9. doi: 10.1093/hmg/ddv494. PubMed PMID: 26732429; PubMed Central PMCID: PMCPMC4731021.

5. Hosgood HD, 3rd, Wang WC, Hong YC, Wang JC, Chen K, Chang IS, et al. Genetic variant in TP63 on locus 3q28 is associated with risk of lung adenocarcinoma among never-smoking females in Asia. Hum Genet. 2012;131(7):1197-203. doi: 10.1007/s00439-012-1144-8. PubMed PMID: 22367405; PubMed Central PMCID: PMC3875137.

6. 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. American journal of epidemiology. 2009;170(6):695-702. doi: 10.1093/aje/kwp019. PubMed PMID: 19671832.

7. Chang CH, Hsiao CF, Yeh YM, Chang GC, Tsai YH, Chen YM, et al. Circulating interleukin-6 level is a prognostic marker for survival in advanced nonsmall cell lung cancer patients treated with chemotherapy. Int J Cancer. 2013;132(9):1977-85. doi: 10.1002/ijc.27892. PubMed PMID: 23034889.

8. 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(2):452-9. Epub 2019/12/19. doi: 10.1158/1055-9965.EPI-19-1221. PubMed PMID: 31848206.

9. Tammemagi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, et al. Selection criteria for lung-cancer screening. The New England journal of medicine. 2013;368(8):728-36. doi: 10.1056/NEJMoa1211776. PubMed PMID: 23425165; PubMed Central PMCID: PMC3929969.

10. 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(12):e1001764. doi: 10.1371/journal.pmed.1001764. PubMed PMID: 25460915; PubMed Central PMCID: PMC4251899.

11. Gail MH, Lubin, J.H. and Rubinstein. L.V. . Likelihood calculations for matched case-control studies and survival studies with tied death times. Biometrika. 1980;68:703-7.

12. Cox DR. Two further applications ofa model for binary regression. Biometrika. 1958;45:562–5.

13. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European heart journal. 2014;35(29):1925-31. Epub 2014/06/06. doi: 10.1093/eurheartj/ehu207. PubMed PMID: 24898551; PubMed Central PMCID: PMCPMC4155437.