



Development of a risk prediction model for lung cancer: The Japan Public Health Center-based Prospective Study

Hadrien Charvat  | Shizuka Sasazuki | Taichi Shimazu | Sanjeev Budhathoki |
Manami Inoue  | Motoki Iwasaki | Norie Sawada | Taiki Yamaji |
Shoichiro Tsugane | for the JPHC Study Group

Epidemiology and Prevention Group, Center for Public Health Sciences, National Cancer Center, Tokyo, Japan

Correspondence

Taichi Shimazu, Division of Prevention, Center for Public Health Sciences, National Cancer Center, Chuo-ku, Tokyo, Japan.
Email: tshimazu@ncc.go.jp

Funding information

Practical Research for Innovative Cancer Control, Grant/Award Number: 17CK0106270; National Cancer Center Research and Development Fund [23-A-31 (toku), 26-A-2, 29-A-4] and the Ministry of Health, Labour and Welfare of Japan

Although the impact of tobacco consumption on the occurrence of lung cancer is well-established, risk estimation could be improved by risk prediction models that consider various smoking habits, such as quantity, duration, and time since quitting. We constructed a risk prediction model using a population of 59 161 individuals from the Japan Public Health Center (JPHC) Study Cohort II. A parametric survival model was used to assess the impact of age, gender, and smoking-related factors (cumulative smoking intensity measured in pack-years, age at initiation, and time since cessation). Ten-year cumulative probability of lung cancer occurrence estimates were calculated with consideration of the competing risk of death from other causes. Finally, the model was externally validated using 47 501 individuals from JPHC Study Cohort I. A total of 1210 cases of lung cancer occurred during 986 408 person-years of follow-up. We found a dose-dependent effect of tobacco consumption with hazard ratios for current smokers ranging from 3.78 (2.00-7.16) for cumulative consumption ≤ 15 pack-years to 15.80 (9.67-25.79) for >75 pack-years. Risk decreased with time since cessation. Ten-year cumulative probability of lung cancer occurrence estimates ranged from 0.04% to 11.14% in men and 0.07% to 6.55% in women. The model showed good predictive performance regarding discrimination (cross-validated *c*-index = 0.793) and calibration (cross-validated $\chi^2 = 6.60$; *P*-value = .58). The model still showed good discrimination in the external validation population (*c*-index = 0.772). In conclusion, we developed a prediction model to estimate the probability of developing lung cancer based on age, gender, and tobacco consumption. This model appears useful in encouraging high-risk individuals to quit smoking and undergo increased surveillance.

KEYWORDS

Cohort study, competing risks, lung cancer, risk prediction model, tobacco smoking

1 | INTRODUCTION

Lung cancer is the leading cause of cancer-related mortality both worldwide¹ and in Japan: according to 2016 national statistics, it accounted for 19.8% of all cancer-related deaths.² By incidence, lung cancer represents the third most commonly diagnosed cancer in Japan, and accounted for 13.0% of all cancers diagnosed in 2013, with two-thirds of these occurring in men.²

Tobacco smoking has long been recognised as a major risk factor for lung cancer occurrence.^{3,4} A recent review reported median population attributable fractions of approximately 80% for men and 60% for women for smoking in relation to lung cancer⁵ (a Japanese study reported values of 67.5% in men and 23.9% in women⁶). This situation is especially concerning in Asia, where the prevalence of smoking is high, particularly among men.⁷ As lung cancer is a particularly deadly disease with an estimated all-stage 5-year net survival in the Japanese population of approximately 30%,⁸ primary prevention through information and education about the effect of smoking is important.

Moreover, lung cancer is often diagnosed at an advanced stage, when surgical resection cannot be considered or has a low probability of success. A study of stage-specific lung cancer survival in several countries showed that the 1-year survival of lung cancer patients was highly dependent on stage, with age-adjusted net survival values ranging from 64% to 88% for localized, 46% to 55% for regional, and 18% to 27% for distant non-small-cell lung cancer.⁹ These findings highlight the importance of identifying high-risk individuals who may benefit from increased surveillance and screening by low-dose computed tomography.¹⁰

In this study, we developed a prediction model based on a cohort of 59 161 individuals from Japan Public Health Center (JPHC) Study Cohort II to estimate the probability of developing lung cancer based on age and gender, as well as various smoking-related variables, and with adjustment for the competing risk of death from other causes. We also carried out an external validation study to assess the relevance of the developed model for risk stratification in the general Japanese population. This model might be useful in identifying high-risk individuals, and in encouraging them to undergo increased surveillance and adopt healthier lifestyles. In particular, the unequivocal decreasing trend in lung cancer occurrence with time since smoking cessation constitutes strong incentive to recommend quitting smoking.

2 | MATERIALS AND METHODS

2.1 | Study participants

Details of the study design have been described elsewhere.¹¹ Briefly, the participants were Japanese residents recruited into Cohorts I and II of the JPHC-based Prospective Study in 1990-1991 and 1993-1994, respectively. All participants answered a self-administered baseline questionnaire distributed at study entry. The starting

point was defined as the date of completion of the baseline questionnaire and individuals were followed up until the December 31, 2012. We identified 61 595 individuals in Cohort I and 78 825 individuals in Cohort II as the original population. After exclusion of individuals not fulfilling the inclusion criteria, the population used for analysis consisted of 47 501 individuals from Cohort I and 59 161 from Cohort II (Figures S1, S2). The study was approved by the institutional review board of the National Cancer Center, Tokyo, Japan.

2.2 | Follow-up and identification of lung cancer cases

Residency and death registration are required by the Basic Residential Register Law and Family Registry Law, respectively, and the registries are considered to be complete. We identified incident lung cancer cases (codes C34.0-C34.9 of the International Classification of Diseases for Oncology, 3rd edition)¹² by active patient notification from major local hospitals in each study area and by data linkage with population-based cancer registries.

2.3 | Model construction and crude probabilities estimation

Model development was based on the JPHC Study Cohort II population. For each individual, follow-up time was defined as the period from the date of acquisition of the baseline questionnaire to the date of lung cancer diagnosis, date of emigration from the study area, date of death, or end of follow-up, whichever came first.

Age was considered as a continuous variable in the model, and linear, quadratic, and quadratic with one knot (located at the mean age of the study population) effects were tested. Gender was introduced into the model and interaction between gender and age was considered. The following smoking-related variables were used: continuous age at smoking initiation, categories of cumulative consumption expressed in pack-years (PY) (≤ 15 PY, >15 PY and ≤ 30 PY, >30 PY and ≤ 45 PY, >45 PY and ≤ 60 PY, >60 PY and ≤ 75 PY, and >75 PY), and categories of time since smoking cessation for former smokers (>1 and ≤ 5 years, >5 and ≤ 10 years, and >10 years). In particular, individuals who had quit smoking for ≤ 1 year were considered as current smokers. Moreover, we tested the presence of an effect of passive smoking (defined as almost daily passive exposure to smoking in occupational or public settings) in non-smokers and family history of lung cancer (defined as at least one case of lung cancer in parents or siblings). All model comparisons were based on the Akaike Information Criterion. Model construction was based on a two-step approach. First, covariable selection based on the procedure described above was carried out using Cox proportional hazard regression and the proportionality assumption was tested using the Grambsch–Therneau test. Second, we constructed a flexible hazard regression model using the linear predictor determined by the above procedure and included, if necessary, the non-proportional effects identified.

In the second part of the analysis we estimated absolute risks of lung cancer occurrence. Because smoking is known to be related to other diseases (eg, other cancers and cardiovascular diseases) that affect the mortality of individuals, estimation of the cumulative probability of lung cancer occurrence was made in a competing risk setting, that is, by taking account of the fact that individuals might die before developing lung cancer. Consequently, we developed a second flexible hazard model in which the outcome was defined as death from any cause (except lung cancer, because these individuals were censored at the time of lung cancer diagnosis) and all other events censored. The same variables used in the lung cancer-specific model were used to construct the death-specific model using the same modelling procedure as described above. We then estimated the 10-year cumulative probabilities of lung cancer occurrence and death (Pr_{LC} and Pr_D , respectively) adjusted on the competing event for a vector of covariables \mathbf{X} using the following relationships:

$$Pr_{LC}(10, \mathbf{X}) = \int_0^{10} \lambda_{LC}(u, \mathbf{X}) S_{Tot}(u, \mathbf{X}) du$$

$$Pr_D(10, \mathbf{X}) = \int_0^{10} \lambda_D(u, \mathbf{X}) S_{Tot}(u, \mathbf{X}) du$$

where λ_{LC} and λ_D represent the lung cancer-specific and death-specific hazards, respectively, and:

$$S_{Tot}(t, \mathbf{X}) = \exp\left\{-\int_0^t (\lambda_{LC}(u, \mathbf{X}) + \lambda_D(u, \mathbf{X})) du\right\}.$$

Confidence intervals were obtained by the delta method.

2.4 | Predictive performance

The predictive performance of the final model was assessed in terms of discrimination and calibration. Discrimination was estimated on the lung cancer-specific model using Harrell's *c*-index. Calibration was assessed through the analogue of Hosmer–Lemeshow's chi-squared-test for survival analysis developed by Nam and d'Agostino¹³ using 10 years as the end-point and partitioning the study population into deciles of predicted risk, adjusted for the competing risk of death. Correction for optimism was obtained by 10-fold cross-validation¹⁴ using the model construction procedure described above for each fold.

2.5 | External validation

External validation of the developed model was based on JPHC Study Cohort I. Data preparation followed the same steps as for the previous analysis. All variables used in the previous model were available and identically defined in Cohort I. Using the models constructed in the previous step, relative and absolute risks were estimated for individuals in the external validation population and discrimination and calibration were calculated.

All analyses were undertaken with R statistical software (version 3.2.2; <https://www.r-project.org/>), particularly using the *mexhaz* package version 1.3¹⁵ to fit flexible survival models.

3 | RESULTS

A total of 1210 cases of lung cancer occurred in the derivation population (791 cases in the validation population) during 986 408 person-years of follow-up (944 685 person-years in the validation population). During the same period, 9381 individuals died without lung cancer (5646 individuals in the validation population). Baseline characteristics of individuals from both cohorts are summarized in Table 1.

Based on the model construction procedure described above, the final model for lung cancer occurrence included age, gender, and their interaction, smoking intensity, time since smoking cessation, and age at smoking initiation. Passive smoking in non-smokers and family history of lung cancer were found to have no statistically significant effect (hazard ratio = 1.17; 95% confidence interval, 0.89–1.53; and hazard ratio = 1.11; 95% confidence interval, 0.74–1.67, respectively; see Table S1) and were consequently not kept in the final model. There was no evidence of non-proportional effects of the variables included.

Table 2 summarizes the results in terms of effects of the covariables. Hazard ratios for the various combinations of smoking intensity and time since smoking cessation (taking non-smokers as the reference group) showed an increasing relationship between smoking intensity and the risk of lung cancer occurrence, whatever the status regarding smoking cessation (eg, 3.78 for individuals who smoked ≤ 15 PY and 15.80 for individuals who smoked > 75 PY in current smokers vs 1.16 and 4.85, respectively, in individuals who stopped smoking for > 10 years), and a decreasing relationship according to time since smoking cessation, whatever the smoking intensity.

Details about the final model for death occurrence are provided in Table S2. The same variables as the ones included in the model for lung cancer occurrence were used but, because the effect of smoking was found to be closely similar across categories of smoking intensity (ie, the effect of smoking on the risk of death was not dose-dependent), smoking status was kept as a dichotomous variable. The effect of age was modelled as gender-dependent, non-linear, and non-proportional.

Tables 3–6 summarize the 10-year cumulative probabilities of lung cancer occurrence adjusted for the competing risk of death from other causes for various combinations of gender, age, category of cumulative smoking intensity, and category of time since smoking cessation. In current smokers, values ranged from 0.14% to 11.14% in men and between 0.23% and 6.55% in women. Reflecting hazard ratio estimates, the gender-specific cumulative probabilities showed a gradient across the population: they increased with age and smoking intensity and decreased with time since smoking cessation. For example, 10-year cumulative probability estimates for men aged 70 years who had quit smoking for > 10 years showed an important

TABLE 1 Baseline characteristics of the derivation (Japan Public Health Center [JPHC] study Cohort II; 59 161 individuals) and validation (JPHC Cohort I; 47 501 individuals) populations

Baseline characteristic	Derivation population (Cohort II)		Validation population (Cohort I)	
	Men (n = 27 876)	Women (n = 31 285)	Men (n = 22 275)	Women (n = 25 226)
Age, y ^a	53.0 (46.0, 61.6)	53.6 (46.3, 62.1)	50.1 (44.1, 54.9)	50.2 (44.5, 54.7)
Smoking status, n (%)				
Never	4525 (16.2)	23 204 (74.2)	3639 (16.3)	18 305 (72.6)
Never with history of passive smoking	2195 (7.9)	5486 (17.5)	1661 (7.4)	4437 (17.6)
Current	14 751 (52.9)	2183 (7.0)	12 177 (54.7)	1966 (7.8)
Past, >1 and ≤5 y ^b	1201 (4.3)	115 (0.4)	1066 (4.8)	136 (0.5)
Past, >5 and ≤10 y	1546 (5.6)	99 (0.3)	1374 (6.2)	130 (0.5)
Past, >10 y	3658 (13.1)	198 (0.6)	2358 (10.6)	252 (1.0)
Age at initiation ^a	20 (19, 21)	25 (20, 33)	20 (19, 21)	25 (20, 34)
Family history (parents or siblings), n (%)				
No	27 330 (98.0)	30 694 (98.1)	21 793 (97.8)	24 609 (97.6)
Yes	546 (2.0)	591 (1.9)	482 (2.2)	617 (2.4)
Smoking intensity (in pack-years), ^c n (%)				
≤15 PY	3093 (14.6)	1399 (53.9)	3419 (20.1)	1626 (65.4)
>15 PY and ≤30 PY	6551 (31.0)	834 (32.1)	6340 (37.3)	650 (26.2)
>30 PY and ≤45 PY	6147 (29.0)	250 (9.6)	4525 (26.7)	166 (6.7)
>45 PY and ≤60 PY	3063 (14.5)	75 (2.9)	1704 (10.0)	30 (1.2)
>60 PY and ≤75 PY	1312 (6.2)	25 (1.0)	621 (3.7)	8 (0.3)
>75 PY	990 (4.7)	12 (0.5)	366 (2.2)	4 (0.2)

^aMedian (interquartile range).

^bTime since quitting.

^cProportions calculated among former and current smokers: Cohort II, n = 21 156 for men and n = 2595 for women; Cohort I, n = 16 975 for men and n = 2484 for women.

reduction compared with current smokers, ranging from approximately one-half (1.52% vs 2.80%) in individuals who smoked ≤15 PY to two-thirds (3.84% vs 11.14%) in individuals who smoked >75 PY. Corresponding values of the 10-year cumulative probabilities of death adjusted for the competing risk of lung cancer occurrence are provided in the Tables S3-S6.

The developed prediction model showed very good predictive performance in terms of discrimination (cross-validated c-index estimated at 0.793; see Figure S3) and calibration (cross-validated Nam-d'Agostino's χ^2 -test = 6.60; $P = .58$; see Figure S4). In terms of external validation, the model still showed high discriminative ability (c-index = 0.772; see Figure S3) but the Nam-d'Agostino test revealed significant differences between the observed and predicted number of events by category of predicted risks in the validation population (χ^2 -test = 24.99; $P = .002$; see Figure S4), with a tendency for the model to overestimate the number of cases in higher risk categories.

Finally, we provide a simple scoring system (Figure S5) based on the final lung cancer-specific model using the method described by Sullivan et al.¹⁶ Because this score cannot take account of the impact of competing risks, the score-specific probabilities are "net probabilities," that is, calculated under the assumption that individuals are not dying from other causes. Consequently, the probability of

lung cancer occurrence is overestimated. This reflects the fact that a scoring system is essentially a discrimination tool and might not be suited for estimation of the probability of disease occurrence when the impact of competing risks cannot be ignored.

4 | DISCUSSION

In this work, we developed a risk prediction model for lung cancer allowing the identification of high-risk individuals and estimation of the 10-year cumulative probability of lung cancer occurrence, adjusted for death from other causes, for different combinations of age, gender, and smoking-related variables. External validation confirmed the good discriminative ability of the model.

To our knowledge, this work represents the first attempt to build and validate a risk prediction model for the risk of lung cancer occurrence in the Japanese population based on detailed information about smoking history, and with consideration of the competing risk of death. Several risk prediction models for lung cancer have been published.¹⁷⁻²⁵ These differ by study design and the risk factors included and, in particular, by the way they handle smoking-related variables. Most of these models have been developed in European and North American populations, however, where the incidence of

TABLE 2 Summary of the hazard ratios associated with risk factors included in the survival model for lung cancer occurrence developed in the study population of 59 161 individuals from the Japan Public Health Center Study Cohort II

Risk factors		Hazard ratio (95% confidence interval)		
Effect of age, years ^a				
Women				
50		2.52 (1.94-3.28)		
60		5.01 (3.44-7.30)		
70		7.82 (5.10-11.99)		
Men				
40		0.63 (0.43-0.91)		
50		2.27 (1.54-3.35)		
60		6.49 (4.33-9.74)		
70		14.59 (10.26-20.75)		
Age at smoking initiation (for a 1-y increase)		0.97 (0.96-0.99)		
Smoking intensity, pack-years ^b	Current smokers ^c	Time since cessation >1 and ≤5 y	Time since cessation >5 and ≤10 y	Time since cessation >10 y
≤15 PY	3.78 (2.00-7.16)	1.89 (0.94-3.82)	1.27 (0.62-2.58)	1.16 (0.62-2.19)
>15 PY and ≤30 PY	6.17 (3.71-10.26)	3.09 (1.71-5.55)	2.06 (1.13-3.76)	1.89 (1.11-3.24)
>30 PY and ≤45 PY	9.03 (5.60-14.55)	4.52 (2.59-7.90)	3.02 (1.70-5.37)	2.77 (1.65-4.67)
>45 PY and ≤60 PY	10.60 (6.61-16.99)	5.30 (3.04-9.27)	3.55 (2.00-6.30)	3.25 (1.93-5.48)
>60 PY and ≤75 PY	14.84 (9.12-24.13)	7.43 (4.22-13.07)	4.97 (2.78-8.88)	4.55 (2.68-7.75)
>75 PY	15.80 (9.67-25.79)	7.91 (4.52-13.82)	5.29 (2.97-9.42)	4.85 (2.84-8.28)

^aBecause the effect of age was found to be non-linear and different between men and women, we summarize here a few hazard ratios for different combinations of age and sex, taking women aged 40 y as the reference group.

^bNon-smokers constitute the reference group.

^cCurrent smokers include past smokers who quit smoking for ≤1 year.

lung cancer and risk associated with tobacco consumption are reported to differ to those in Asian populations. As such, the need to develop similar tools more specific to Asian populations remains. Park et al²⁵ recently published the first such risk prediction model. Based on a large population of Korean men who participated in check-up visits and taking account of several risk factors (eg, body mass index), this model showed high discriminative ability (external validation *c*-index: 0.87). Although detailed information on various exposures were available, modelling of the relationship between smoking history and lung cancer was limited by the fact that no information on smoking consumption was available for past smokers. Moreover, the study population consisted of individuals who participated in check-up visits and thus might have been more health-conscious, which would in turn hamper the generalizability of their findings to the general population.

As a tool for estimating the cumulative probability of lung cancer occurrence by a certain time of follow-up, the developed model does not automatically classify individuals in groups of risk. However, in the context of primary prevention, our model might be interesting to inform individuals on their risk and entice them to modify their smoking habits. Moreover, it might be used together with other indicators (eg, availability, performance, and cost of screening procedures) to define groups of individuals that may benefit from screening procedures, similar to what was done by the US Preventive Services Task Force.²⁶

Tobacco consumption has long been established as a strong risk factor for lung cancer.^{3,4} In this work, we modelled in detail the relationship between smoking and lung cancer using several smoking-related variables, namely smoking intensity, age at smoking initiation, and time since cessation in former smokers. To reduce the collinearity between these variables and age, we chose to use cumulative consumption expressed in PY²⁷ and to express smoking intensity and time since cessation in categories. Consistent with previous studies, we found a clear dose-dependent increase in the risk of lung cancer with increasing cumulative smoking consumption. Moreover, we found a strong inverse relationship between time since cessation and risk of lung cancer occurrence,²⁸ and risk of death from other causes. This is important because it emphasizes the fact that offering advice and campaigns on smoking cessation might be an effective way to reduce the burden of tobacco smoking on lung cancer and death, even for individuals with a substantial history of tobacco consumption.

The impact of tobacco on health is not limited to lung cancer, and its relationship with various other diseases and mortality has been reported on several occasions.^{3,4,29} In the present study, smoking was found to have a strong and dose-independent effect on the risk of death before lung cancer occurrence, which translated into higher 10-year cumulative probabilities of death in smokers (eg, approximately 30% in smokers vs 19% in never-smokers for men aged 70 years). Given that smoking is a modifiable risk factor with a

TABLE 3 Ten-year cumulative probability of lung cancer occurrence (expressed in %) in non-smokers and current smokers according to gender, age, and intensity of smoking (expressed in pack-years) with consideration of the competing risk of death from other causes

Smoking intensity, pack-years ^a	Men; age, y			Women; age, y				
	40 ^b	50	60	70	40	50	60	70
Non-smokers	0.06 (0.04-0.09) ^c	0.23 (0.18-0.29)	0.64 (0.52-0.79)	1.35 (1.06-1.71)	0.10 (0.07-0.15)	0.26 (0.21-0.31)	0.50 (0.44-0.58)	0.75 (0.60-0.94)
≤15 PY	0.14 (0.09-0.22)	0.51 (0.36-0.74)	1.41 (0.99-2.03)	2.80 (1.92-4.09)	0.23 (0.14-0.37)	0.58 (0.39-0.86)	1.12 (0.75-1.66)	1.61 (1.04-2.49)
>15 PY and ≤30 PY	0.24 (0.17-0.33)	0.84 (0.69-1.01)	2.29 (1.90-2.78)	4.53 (3.63-5.64)	0.38 (0.25-0.56)	0.94 (0.71-1.25)	1.82 (1.39-2.38)	2.62 (1.90-3.60)
>30 PY and ≤45 PY	0.34 (0.25-0.48)	1.22 (1.05-1.43)	3.34 (2.91-3.84)	6.55 (5.44-7.88)	0.55 (0.37-0.82)	1.37 (1.06-1.79)	2.65 (2.07-3.39)	3.81 (2.80-5.17)
>45 PY and ≤60 PY	0.40 (0.29-0.57)	1.44 (1.19-1.73)	3.91 (3.32-4.61)	7.64 (6.35-9.18)	0.65 (0.43-0.97)	1.61 (1.22-2.13)	3.10 (2.39-4.03)	4.45 (3.26-6.07)
>60 PY and ≤75 PY	0.56 (0.39-0.81)	2.00 (1.61-2.50)	5.43 (4.45-6.62)	10.51 (8.37-13.15)	0.91 (0.59-1.39)	2.25 (1.66-3.04)	4.32 (3.25-5.73)	6.17 (4.40-8.61)
>75 PY	0.60 (0.41-0.88)	2.13 (1.67-2.71)	5.77 (4.63-7.17)	11.14 (8.74-14.15)	0.96 (0.62-1.49)	2.39 (1.74-3.28)	4.59 (3.41-6.17)	6.55 (4.62-9.25)

^aFor smokers, predictions are given for an age at smoking initiation of 20 y.
^bPredictions are given for the exact ages indicated.
^cConfidence intervals were estimated by the delta method.

TABLE 4 Ten-year cumulative probability of lung cancer occurrence (expressed in %) in past smokers (time since cessation >1 y and ≤5 y) according to gender, age, and intensity of smoking (expressed in pack-years) with consideration of the competing risk of death from other causes

Smoking intensity, pack-years ^a	Men; age, y			Women; age, y				
	40 ^b	50	60	70	40	50	60	70
≤15 PY	0.07 (0.04-0.12) ^c	0.26 (0.16-0.41)	0.72 (0.45-1.14)	1.46 (0.90-2.33)	0.12 (0.07-0.20)	0.29 (0.18-0.48)	0.57 (0.34-0.93)	0.82 (0.49-1.39)
>15 PY and ≤30 PY	0.12 (0.08-0.18)	0.42 (0.30-0.60)	1.17 (0.82-1.65)	2.35 (1.65-3.36)	0.19 (0.12-0.31)	0.47 (0.31-0.71)	0.92 (0.62-1.37)	1.34 (0.87-2.06)
>30 PY and ≤45 PY	0.17 (0.11-0.27)	0.62 (0.44-0.86)	1.70 (1.23-2.35)	3.43 (2.46-4.77)	0.28 (0.17-0.45)	0.69 (0.47-1.03)	1.34 (0.92-1.97)	1.96 (1.29-2.97)
>45 PY and ≤60 PY	0.20 (0.13-0.32)	0.72 (0.51-1.04)	2.00 (1.42-2.80)	4.01 (2.86-5.60)	0.33 (0.20-0.53)	0.81 (0.54-1.22)	1.58 (1.06-2.34)	2.29 (1.49-3.51)
>60 PY and ≤75 PY	0.28 (0.18-0.45)	1.01 (0.70-1.46)	2.78 (1.96-3.94)	5.56 (3.89-7.91)	0.45 (0.27-0.76)	1.13 (0.75-1.73)	2.20 (1.46-3.30)	3.19 (2.05-4.95)
>75 PY	0.30 (0.19-0.48)	1.08 (0.75-1.55)	2.96 (2.09-4.18)	5.91 (4.16-8.36)	0.48 (0.29-0.81)	1.21 (0.80-1.83)	2.34 (1.56-3.49)	3.40 (2.19-5.25)

^aPredictions are given for an age at smoking initiation of 20 y.
^bPredictions are given for the exact ages indicated.
^cConfidence intervals were estimated by the delta method.

TABLE 5 Ten-year cumulative probability of lung cancer occurrence (expressed in %) in past smokers (time since cessation >5 and ≤10 y) according to gender, age, and intensity of smoking (expressed in pack-years) with consideration of the competing risk of death from other causes

Smoking intensity, pack-years ^a	Men; age, y			Women; age, y				
	40 ^b	50	60	70	40	50	60	70
≤15 PY	0.05 (0.03-0.08) ^c	0.17 (0.11-0.28)	0.48 (0.30-0.78)	0.99 (0.60-1.62)	0.08 (0.04-0.14)	0.19 (0.12-0.33)	0.38 (0.23-0.63)	0.56 (0.32-0.96)
>15 PY and ≤30 PY	0.08 (0.05-0.13)	0.28 (0.19-0.42)	0.79 (0.54-1.15)	1.61 (1.09-2.36)	0.13 (0.08-0.21)	0.32 (0.21-0.49)	0.62 (0.40-0.95)	0.91 (0.58-1.43)
>30 PY and ≤45 PY	0.12 (0.07-0.19)	0.41 (0.29-0.60)	1.15 (0.81-1.64)	2.34 (1.63-3.37)	0.19 (0.11-0.31)	0.46 (0.30-0.71)	0.90 (0.60-1.36)	1.33 (0.85-2.08)
>45 PY and ≤60 PY	0.14 (0.08-0.22)	0.49 (0.33-0.72)	1.35 (0.93-1.95)	2.74 (1.89-3.97)	0.22 (0.13-0.37)	0.54 (0.35-0.84)	1.06 (0.69-1.63)	1.56 (0.98-2.45)
>60 PY and ≤75 PY	0.19 (0.12-0.31)	0.68 (0.46-1.01)	1.88 (1.28-2.75)	3.82 (2.59-5.61)	0.30 (0.18-0.52)	0.76 (0.49-1.19)	1.48 (0.96-2.29)	2.17 (1.36-3.47)
>75 PY	0.20 (0.12-0.33)	0.72 (0.48-1.08)	2.00 (1.37-2.93)	4.06 (2.75-5.96)	0.32 (0.19-0.56)	0.81 (0.52-1.27)	1.58 (1.02-2.43)	2.31 (1.44-3.69)

^aPredictions are given for an age at smoking initiation of 20 y.

^bPredictions are given for the exact ages indicated.

^cConfidence intervals were estimated by the delta method.

TABLE 6 Ten-year cumulative probability of lung cancer occurrence (expressed in %) in past smokers (time since cessation >10 y) according to gender, age, and intensity of smoking (expressed in pack-years) with consideration of the competing risk of death from other causes

Smoking intensity (pack-years) ^a	Men; age, y			Women; age, y				
	40 ^b	50	60	70	40	50	60	70
≤15 PY	0.04 (0.03-0.07) ^c	0.16 (0.11-0.24)	0.45 (0.30-0.66)	0.93 (0.63-1.39)	0.07 (0.04-0.12)	0.18 (0.12-0.28)	0.35 (0.23-0.54)	0.52 (0.33-0.83)
>15 PY and ≤30 PY	0.07 (0.05-0.11)	0.26 (0.19-0.36)	0.73 (0.54-0.98)	1.52 (1.11-2.06)	0.12 (0.07-0.19)	0.29 (0.20-0.42)	0.57 (0.40-0.82)	0.85 (0.57-1.27)
>30 PY and ≤45 PY	0.11 (0.07-0.16)	0.38 (0.28-0.52)	1.06 (0.79-1.43)	2.21 (1.63-3.01)	0.17 (0.10-0.28)	0.43 (0.29-0.62)	0.83 (0.58-1.20)	1.24 (0.82-1.86)
>45 PY and ≤60 PY	0.12 (0.08-0.20)	0.45 (0.32-0.63)	1.25 (0.91-1.71)	2.59 (1.90-3.54)	0.20 (0.12-0.33)	0.50 (0.34-0.74)	0.98 (0.67-1.43)	1.45 (0.96-2.20)
>60 PY and ≤75 PY	0.17 (0.11-0.28)	0.63 (0.44-0.89)	1.74 (1.25-2.42)	3.61 (2.57-5.05)	0.28 (0.17-0.47)	0.70 (0.46-1.06)	1.37 (0.92-2.03)	2.03 (1.31-3.13)
>75 PY	0.19 (0.11-0.30)	0.67 (0.46-0.96)	1.85 (1.31-2.61)	3.84 (2.70-5.43)	0.30 (0.18-0.50)	0.75 (0.49-1.14)	1.45 (0.97-2.18)	2.16 (1.38-3.36)

^aPredictions are given for an age at smoking initiation of 20 y.

^bPredictions are given for the exact ages indicated.

^cConfidence intervals were estimated by the delta method.

strong impact on health, this finding emphasizes again the need for better strategies for smoking control at both the individual level (information and prevention campaigns) and institutional level (eg, promotion of tobacco-free public places and control of cigarette prices).^{6,30}

In addition to the role of smoking history, age at start of follow-up was also found to have a substantial impact on lung cancer occurrence, with a differential effect between genders, the risk increasing more steeply in older men than in older women. This translates into a higher incidence of lung cancer in older men for a given level of exposure to other risk factors. A similar differential effect of age was found for gastric cancer³¹ and might be explained by unobserved factors that differentially affect men and women, especially in older birth cohorts, such as occupational exposures.

This work has several strengths. First, we used flexible parametric models to describe the relationship between age, sex, and several variables related to smoking history and both the occurrence of lung cancer and death from other causes. This allowed us to estimate cumulative probabilities adjusted on the competing event for various combinations of the risk factors. Second, the use of several variables related to smoking history allowed us to show the double gradient that existed between cumulative smoking consumption and increase in lung cancer occurrence on the one hand, and between time since cessation and the decrease in this risk on the other. Finally, the model's good discriminatory performance showed its potential for use as background information in implementing lung cancer screening policies in Japan.

This work also has some limitations. First, the JPHC questionnaires did not assess some of the factors that were previously shown to significantly affect the risk of lung cancer, such as a personal history of pulmonary diseases (particularly chronic obstructive pulmonary disease and emphysema), and exposure to environmental or occupational carcinogens such as asbestos. Consequently, these factors could not be included in our model. However, individuals suffering from chronic pulmonary diseases or with past exposure to environmental or occupational carcinogens are usually already identified as "at-risk," and consequently appropriately monitored and given recommendations on healthy lifestyle habits.

Second, although second-hand smoking has been shown to impact lung cancer occurrence, particularly among never-smokers, it was found to have no significant impact on the predictive ability of our model and was consequently removed from the final model. Because passive smoking was assessed by asking participants about tobacco exposure outside the home and workplace, it is possible that some non-smokers were wrongly classified as non-exposed to passive smoking at home, which would explain the absence of a significant relationship in this study. However, one should note that the absence of improvement in a risk prediction algorithm does not call into question the already established nocive impact of second-hand smoking,³² and preventing it remains an important measure of any policy aimed at reducing the occurrence of lung cancer, as well as cardiovascular and pulmonary diseases.

Finally, as a practical tool to inform on the absolute risk of lung cancer occurrence, our model does not distinguish between subtypes of lung cancer, although it is known that the strength of their association with tobacco smoking differs;³³ we verified that this association was higher for squamous cell and small-cell carcinomas than for adenocarcinomas (see Table S7).

In conclusion, we developed a risk prediction model for lung cancer occurrence in the Japanese population. Results emphasized the cumulative dose-dependent detrimental effect of tobacco consumption as well as the partial reversal of risk after smoking cessation. The model showed good discriminative ability in an external population. Consequently, it might be used in the Japanese population to identify high-risk individuals who may benefit from increased surveillance and screening.

ACKNOWLEDGMENTS

We would like to thank all members of the JPHC Study group (listed at <http://epi.ncc.go.jp/en/jphc/781/3838.html>) for their valuable contributions. This work was supported by Practical Research for Innovative Cancer Control from the Japan Agency for Medical Research and Development (grant no. 17CK0106270), the National Cancer Center Research and Development Fund [23-A-31 (toku), 26-A-2, 29-A-4] and the Ministry of Health, Labour and Welfare of Japan.

CONFLICT OF INTEREST

The authors have no conflict of interest.

ORCID

Hadrien Charvat  <http://orcid.org/0000-0003-3624-1394>

Manami Inoue  <http://orcid.org/0000-0003-1276-2398>

REFERENCES

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin*. 2015;65:87-108.
2. Cancer Statistics in Japan. 2016. http://ganjoho.jp/reg_stat/statistics/stat/summary.html. Accessed December 13, 2017.
3. Doll R, Peto R. Mortality in relation to smoking: 20 years' observations on male British doctors. *Br Med J*. 1976;2:1525-1536.
4. U.S. Department of Health and Human Services. *The Health Consequences of Smoking: 50 Years of Progress. A Report of the Surgeon General*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; 2014. Printed with corrections, January 2014.
5. Lee PN, Forey BA, Coombs KJ. Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. *BMC Cancer*. 2012;12:385.
6. Inoue M, Sawada N, Matsuda T, et al. Attributable causes of cancer in Japan in 2005 – systematic assessment to estimate current burden of cancer attributable to known preventable risk factors in Japan. *Ann Oncol*. 2012;23:1362-1369.

7. Katanoda K, Jiang Y, Park S, Lim MK, Qiao YL, Inoue M. Tobacco control challenges in East Asia: proposals for change in the world's largest epidemic region. *Tob Control*. 2014;23:359-368.
8. Allemani C, Weir HK, Carreira H, et al. Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet*. 2015;385:977-1010.
9. Walters S, Maringe C, Coleman MP, et al. Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004-2007. *Thorax*. 2013;68:551-564.
10. Sone S, Nakayama T, Honda T, et al. Long-term follow-up study of a population-based 1996-1998 mass screening programme for lung cancer using mobile low-dose spiral computed tomography. *Lung Cancer*. 2007;58:329-341.
11. Tsugane S, Sawada N. The JPHC study: design and some findings on the typical Japanese diet. *Jpn J Clin Oncol*. 2014;44:777-782.
12. World Health Organization. *International Classification of Diseases for Oncology*, 3rd edn. Geneva, Switzerland: World Health Organization; 2000.
13. D'Agostino RB, Nam BH. Evaluation of the performance of survival analysis models: discrimination and calibration measures. In: Balakrishnan N, Rao CR, eds. *Handbook of Statistics*, vol. 23. Amsterdam: Elsevier; 2004:1-25.
14. Pencina MJ, d'Agostino RB Sr, Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the Framingham Heart Study. *Circulation*. 2009;119:3078-3084.
15. Charvat H, Belot A. mexhaz: Mixed effect excess hazard models. R package version 1.3, 2017. <https://CRAN.R-project.org/package=mexhaz>. Accessed December 13, 2017.
16. Sullivan LM, Massaro JM, d'Agostino RB Sr. Presentation of multivariate data for clinical use: the Framingham Study risk score functions. *Stat Med*. 2004;23:1631-1660.
17. Hoggart C, Brennan P, Tjonneland A, et al. A risk model for lung cancer incidence. *Cancer Prev Res*. 2012;5:834-846.
18. Cassidy A, Myles JP, van Tongeren M, et al. The LLP risk model: an individual risk prediction model for lung cancer. *Br J Cancer*. 2008;98:270-276.
19. Cassidy A, Myles JP, Liloglou T, Duffy SW, Field JK. Defining high-risk individuals in a population-based molecular-epidemiological study of lung cancer. *Int J Oncol*. 2006;28:1295-1301.
20. Spitz MR, Amos CI, Land S, et al. Role of selected genetic variants in lung cancer risk in African Americans. *J Thorac Oncol*. 2013;8:391-397.
21. Tammemagi CM, Pinsky PF, Caporaso NE, et al. Lung cancer risk prediction: prostate, lung, colorectal and ovarian cancer screening trial models and validation. *J Natl Cancer Inst*. 2011;103:1058-1068.
22. Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. *J Natl Cancer Inst*. 2003;95:470-478.
23. Tammemagi MC, Katki HA, Hocking WG, et al. Selection criteria for lung-cancer screening. *N Engl J Med*. 2013;368:728-736.
24. Maisonneuve P, Bagnardi V, Bellomi M, et al. Lung cancer risk prediction to select smokers for screening CT – a model based on the Italian COSMOS trial. *Cancer Prev Res*. 2011;4:1778-1789.
25. Park S, Nam BH, Yang HR, et al. Individualized risk prediction model for lung cancer in Korean men. *PLoS ONE*. 2013;8:e54823.
26. Moyer VA, U.S. Preventive Services Task Force. Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med*. 2014;160:330-338.
27. Leffondre K, Abrahamowicz M, Siemiatycki J, Rachet B. Modeling smoking history: a comparison of different approaches. *Am J Epidemiol*. 2002;156:813-823.
28. Sobue T, Yamamoto S, Hara M, et al. Cigarette smoking and subsequent risk of lung cancer by the histologic type in middle-aged Japanese men and women: the JPHC Study. *Int J Cancer*. 2002;99:245-251.
29. Katanoda K, Marugame T, Saika K, et al. Population attributable fraction of mortality associated with tobacco smoking in Japan: a pooled analysis of three large-scale cohort studies. *J Epidemiol*. 2008;18:251-264.
30. Tabuchi T, Fujiwara T, Shinozaki T. Tobacco price increase and smoking behaviour changes in various subgroups: a nationwide longitudinal 7-year follow-up study among a middle-aged Japanese population. *Tob Control*. 2017;26:69-77.
31. Charvat H, Sasazuki S, Inoue M, et al. Prediction of the 10-year probability of gastric cancer occurrence in the Japanese population: the JPHC study cohort II. *Int J Cancer*. 2016;138:320-331.
32. Kurahashi N, Inoue M, Liu Y, et al. Passive smoking and lung cancer in Japanese non-smoking women: a prospective study. *Int J Cancer*. 2008;122:653-657.
33. Jedrychowski W, Becher H, Wahrendorf J, Basa-Cierpielek Z, Gomola K. Effect of tobacco smoking on various histological types of lung cancer. *J Cancer Res Clin Oncol*. 1992;118:276-282.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Charvat H, Sasazuki S, Shimazu T, et al. ; for the JPHC Study Group. Development of a risk prediction model for lung cancer: The Japan Public Health Center-based Prospective Study. *Cancer Sci*. 2018;109:854-862. <https://doi.org/10.1111/cas.13509>