Peer Review File

Article Information: https://dx.doi.org/10.21037/tlcr-21-566

Reviewer A

Comment 1: As implementation of lung cancer screening proceeds in Asia, it is likely that risk prediction models tailored specifically to Asia will be needed to improve selection of individuals for screening. Thus, this is an important and timely topic and a worthwhile analysis. The manuscript is fairly easy to understand. The authors mention an Appendix but it was not provided to me.
I have some concerns about the new Korean risk model developed in this manuscript:
Reply 1: We apologize for the confusion. We have added the Appendix again through the submission website and we have provided explanations regarding your concerns. Please refer to the attached files.

Comment 2: Table 4 shows that the model substantially under-predicts risk for current smokers, while over-predicting risk for former smokers. This happens even in the training dataset, indicating that there is some important problem in the parameterization of the model. Adjustment of the model is needed, likely with additional terms to capture smoking status or interactions with smoking status. This is critical before proposing that the model be applied for screening selection.
Reply 2: As per your comment, in terms of calibration presented by E/O ratio, despite the good calibration of the model in ever-smokers, we observed underestimation in current smokers and overestimation in past smokers. However, the purpose of this study was to develop a lung cancer risk assessment model for ever-smokers and compare the model with the eligibility criteria of the Korean national lung cancer screening program, not to develop a prediction model for all subsets of the population. From this perspective, the overall calibration estimated by E/O in ever-smokers (not stratified by current smoking status) in both the training and validation sets was good.

Subject characteristics are systemically under- or over-estimated risks (1). Therefore, prediction models for both current and past smokers would have different calibrations, despite similar discrimination (1). These phenomena have also been observed in other studies. The prediction model developed for Korean men underestimated the risk in the high-risk group and overestimated the risk in the low-risk group (2). When nine lung cancer prediction models were applied to two population cohorts, more overestimations were observed when they were applied to the CPS-II cohort, in which the proportion of current smokers was about half of that in the other cohort (NHIC-AARP) (3).
If a model is over-fitted (by including too many predictors or excessively complex algorithms), subjects at low risk tend to have underestimated risk prediction and those at high risk tend to have overestimated prediction (1, 4). This model showed opposite results (risk underestimation in current smokers and overestimation in past smokers in both training and validation datasets) like previously developed models (3); thus, different calibrations between past and current

smokers were not caused by the over-fitting of the model. Therefore, we considered that the underestimated risk in current smokers and overestimated risk in past smokers were caused by the different characteristics of the two groups, not by the model itself.
Reflecting your comment, we added the following in the text for a more detailed explanation:
"The risk model showed good discrimination and calibration in the training and validation dataset overall. When divided by smoking history, the discrimination of the model was relatively lower in smokers with higher cumulative smoking exposure, which underestimated the risk. Otherwise, overestimation of risk was observed in past smokers or smokers with lower cumulative smoking. If this was caused by model fitting, there were underestimated risk prediction in low-risk populations and overestimated risk prediction in high-risk groups (1, 4). However, based on the direction of over- and underestimation of the model, it was caused by the characteristics of the sub-population (current smokers and past smokers). These results are consistent with those of a previous model validation study (3)."

Changes in text: page 17 line 338-346

Comment 3: The authors excluded participants who had any missing information for the variables considered. If 44% of the population is current/former smokers (from the discussion) then I calculate that they may have excluded around 2 million people from the dataset (to reach an analysis population of 969000) – is this correct? It is very important not to exclude a high proportion of participants due to missing data, as this can lead to bias. Multiple imputation methods should be used instead.
Reply 3: We apologize for the confusion. The initial considered population of 6,811,893 people included both smokers and non-smokers. After exclusion of 69,421 people with missing information on smoking history, 45,692 lung cancer patients before the date of health examination, and 54,091 people aged ≥80 years, 6,642,429 people remained. Among them, 1,422,858 were ever-smokers. After splitting the data into training and validation sets using random sampling, we included 996,004 subjects in the training set and 426,854 in the validation set.
In the training set, 317,170 (31.8%) had missing information on BMI, 32,814 (3.2%) had missing information on drinking, and 4,723 (0.5%) had missing information on physical activity (duplicate possible); thus, 678,407 were included in the initial model development. In the validation set, 135,725 (31.8%) had missing information on BMI, 14,067 (3.2%) had missing information on drinking, and 1,988 (0.5%) had missing information on physical activity (duplicates possible); thus, 290,994 were included in the initial model development. The missing proportions in the training and validation datasets were the same.
As per the reviewer's comment, we agree that exclusion of around 32% of participants due to missing data was a high proportion, leading to possible selection bias. The missing proportion of other variables was less than 4%, but the missing proportion of BMI was approximately 30%. Thus, we performed multiple imputation for BMI using the MICE package in R statistics. Other variables with missing data were physical activity and drinking, which were measured as categorical variables; thus, they were not imputed. After multiple imputation, 991,281 people without missing variables for the model were included in the analysis. A comparison of the

initial results, which excluded all people with missing values and revised results, excluding only 4,723 people with missing information on physical activity or drinking, is presented in the table below.

Multivariate lung cancer prediction model in Korean ever-smokers

	Including 678,407 people (Initial dataset)		Including 991,281 people (Imputed dataset)	
	Beta	Hazard ratio (95% Confidence interval)	Beta	Hazard ratio (95% Confidence interval)
Age				
Age-mean	0.14618	1.157 (1.150–1.164)	0.14453	1.156 (1.150-1.161)
(Age-mean)^2	-0.00242	0.998 (0.997–0.998)	-0.00239	0.998 (0.997-0.998)
Sex				
Male	0	1 (ref)	0	1 (ref)
Female	-0.38713	0.679 (0.611–0.754)	-0.40207	0.669 (0.610-0.733)
Pack-year				
Square root	0.17456	1.191 (1.174–1.208)	0.17577	1.192 (1.178-1.206)
Smoking status and years since cessation in past smokers				
Current smokers	1.03818	2.824 (2.543–3.137)	1.02502	2.787 (2.558-3.036)
<5 years	0.62246	1.864 (1.644–2.112)	0.64272	1.902 (1.718-2.105)
5–14.9 years	0.34404	1.411 (1.248–1.595)	0.32932	1.390 (1.259-1.536)
≥15 years	0	1 (ref)	0	1 (ref)
Physical activity				
<3/week	0	1 (ref)	0	1 (ref)
≥3/week	-0.06768	0.935 (0.889–0.983)	-0.07272	0.930 (0.892-0.969)
Number of days of alcohol consumption				
<5/week	0	1 (ref)	0	1 (ref)
≥5/week	0.05952	1.061 (1.001–1.125)	0.03217	1.033 (0.983-1.085)
Body mass index				
<18.5 kg/m^2	0.26841	1.308 (1.189–1.439)	0.43280	1.542 (1.428-1.664)
18.5–24.9 kg/m^2	0	1 (ref)	0	1 (ref)
≥25 kg/m^2	-0.24695	0.781 (0.741–0.824)	-0.25502	0.775 (0.741-0.810)
History of chronic pulmonary obstructive disease				
No	0	1 (ref)	0	1 (ref)
Yes	0.36037	1.434 (1.316–1.563)	0.39292	1.481 (1.378-1.592)
History of emphysema				
No	0	1 (ref)	0	1 (ref)
Yes	0.25687	1.293 (1.029–1.624)	0.31969	1,377 (1.146-1.654)
History of pneumoconiosis				
No	0	1 (ref)	0	1 (ref)
Yes	0.7179	2.05 (1.335–3.149)	0.54203	1.729 (1.186-2.494)
History of interstitial pulmonary disease				

No	0	1 (ref)	0	1 (ref)
Yes	1.48709	4.424 (3.475–5.633)	1.49383	4.454 (3.647-5.440)

After imputation, the results (Beta and hazard ratio) were slightly changed, but the direction of association and strength of the association were comparable. However, when Harrell's C-index and E/O ratio of the model from the initial dataset and revised dataset were compared, Harrell's C-index was slightly decreased in high pack-year smokers compared with the model from the initial dataset. In addition, the calibration (E/O) was lower in past-smokers or low pack-years smokers in both the training and validation datasets, suggesting over-fitted results (Please refer to Table).

Prediction performance of the lung cancer prediction model in Korean ever-smokers

Statistic	Including 678,407 people (Initial dataset)	Including 991,281 people (Imputed dataset)
	Value (95% CI)	Value (95% CI)
Harrell's C-index in the training dataset		
Ever-smokers	0.816 (0.810–0.822)	0.823 (0.819-0.825)
Current smokers	0.816 (0.808–0.824)	0.822 (0.818-0.824)
Past smokers	0.804 (0.790–0.818)	0.811 (0.801-0.816)
Smokers with <10 pack-years	0.787 (0.760–0.814)	0.810 (0.790-0.821)
Smokers with ≥10 pack-years	0.808 (0.802–0.814)	0.811 (0.807-0.813)
Smokers with ≥20 pack-years	0.792 (0.784–0.800)	0.787 (0.783-0.789)
Smokers with ≥30 pack-years	0.754 (0.746–0.762)	0.741 (0.735-0.744)
E/O ratio in the training dataset		
Ever-smokers	1.002 (0.979–1.024)	1.011 (1.010-1.011)
Current smokers	0.881 (0.858–0.904)	1.098 (1.095-1.101)
Past smokers	1.510 (1.444–1.580)	0.826 (0.823-0.828)
Smokers with <10 pack-years	1.143 (1.043–1.252)	0.896 (0.892-0.900)
Smokers with ≥10 pack-years	0.993 (0.970–1.016)	1.019 (1.017-1.021)
Smokers with ≥20 pack-years	0.974 (0.949–0.999)	1.043 (1.040-1.045)
Smokers with ≥30 pack-years	0.958 (0.930–0.988)	1.067 (1.055-1.080)
Harrell's C-index in the validation dataset		
Ever-smokers	0.816 (0.806–0.826)	0.823 (0.817-0.826)
Current smokers	0.816 (0.804–0.828)	0.821 (0.815-0.824)
Past smokers	0.803 (0.783–0.823)	0.813 (0.799-0.820)
Smokers with <10 pack-years	0.797 (0.758–0.836)	0.824 (0.797-0.839)
Smokers with ≥10 pack-years	0.808 (0.798–0.818)	0.810 (0.804-0.813)

Smokers with ≥20 pack-years	0.791 (0.779–0.803)	0.786 (0.778-0.790)
Smokers with ≥30 pack-years	0.753 (0.739–0.767)	0.737 (0.727-0.742)
E/O ratio in the validation dataset		
Ever-smokers	0.989 (0.956–1.023)	1.016 (1.013-1.02)
Current smokers	0.824 (0.793–0.857)	1.117 (1.113-1.122)
Past smokers	1.504 (1.404–1.611)	0.804 (0.800-0.808)
Smokers with <10 pack-years	1.072 (0.936–1.227)	0.923 (0.916-0.929)
Smokers with ≥10 pack-years	0.984 (0.905–1.019)	1.023 (1.020-1.027)
Smokers with ≥20 pack-years	0.962 (0.926–1.000)	1.051 (1.047-1.055)
Smokers with ≥30 pack-years	0.962 (0.919–1.007)	1.066 (1.060-1.072)

When multiple imputation was applied, data with missing values at random and data with missing values not at random could not be distinguished based on the data (5, 6). In addition, a recent study also showed that if the proportion of missing data is below 5% (missing data is negligible) or above 40% (missing data is substantial), it could be better to use observed data only, but discuss and report the extent of the missing data and the limitations (7) (Please refer to Figure).

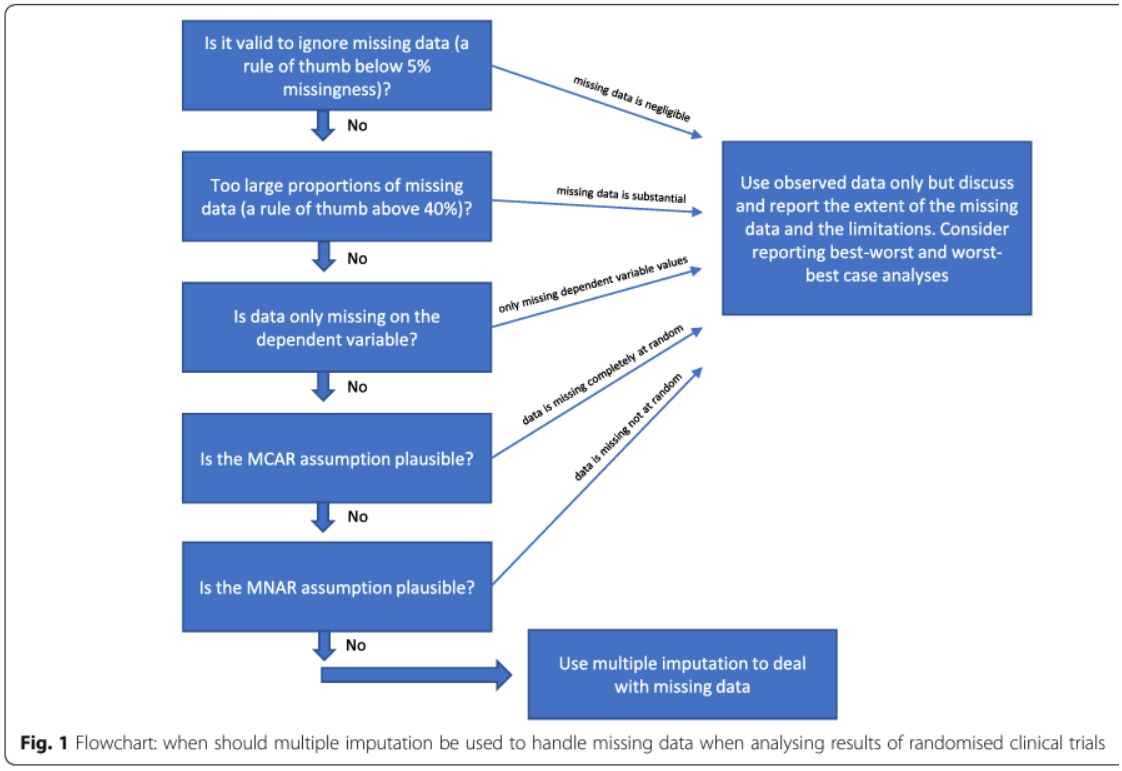

Fig. 1 Flowchart: when should multiple imputation be used to handle missing data when analysing results of randomised clinical trials

This figure was extracted from Jakobsen et al. BMC Medical Research Methodology (2017) 17:162.

Even when multiple imputation techniques are applied, it has been suggested to provide

analysis restricted to complete cases, for comparison with results based on imputed data (5). Therefore, based on the large proportion of missing data (around 32%) and possible over-fitting in consideration of calibration (underestimation of risk in the low-risk group and overestimation of risk in the high-risk group (1, 4)), we considered that it could be better to provide a model with complete data and show the extent of missing data and its limitation in the Discussion as follows:

"The proportion of missing information was highest for BMI (31.8% for all smokers), alcohol drinking (3.2%), and exercise (0.5%) (duplicates possible). The other variables did not contain missing information. Finally, 969,351 ever-smokers with all available predictors were included in the analysis (Figure 1)."

"In addition, we used the complete dataset to develop and validate the model. A large proportion of data (around 34%) mostly due to missing values in BMI (31.8%) were not included in the analysis. Although we tried multiple imputation using the MICE package of R statistics, the results suggested an over-fitted model (1, 4) and the assumption of multiple imputation could not be identified (7). Therefore, we only showed the training and validation results based on the complete dataset."

Changes in text: page 8 line 137-140, page 19 line 375-380

Comment 4: There is not enough detail about how the model was constructed, and what criteria were used for decision-making. Please add tables to show how it was decided which variables to include in the model. For the transformations, which transformations were considered for the continuous variables, and how was the final transformation chosen?

Reply 8: First, we selected variables using the univariate Cox proportional hazards model. Variables with significant results (P-value<0.05) were then applied to various cut-offs, and cut-offs with the lowest Akaike information criterion (AIC) were selected. For continuous variables, such as age, units of pack-years, and years since cessation, raw, log-transformed, squared, square root, and various categories were applied, and those with the lowest AIC were included in the model. We explained the detailed process in the Methods section as follows:

"First, we performed a univariate Cox proportional hazards model regression analysis and variables that showed a P-value of <0.05 were selected. The proportional hazards assumption for each variable was assessed using a log-log survival plot. At this stage, family history of cancer and medical history of cancer were excluded. For selected variables, various cut-offs were applied and cut-offs with the lowest Akaike information criterion (AIC) were selected. Continuous variables, including age, units of pack-years, and years since cessation, raw, log-transformed, squared, square root, and various categories were applied, and those with the lowest AIC were included in the model. Then, with selected variables, multiple Cox proportional hazards model regression analysis was performed."

Changes in text: page 10 line 183-191

Comment 5: Please confirm that the validation data was kept fully independent and never accessed during the model development process, and was only used to estimate the performance of the model after it was finalized. I was surprised that there was no evidence of optimism, i.e.

lower AUC in the validation dataset, which would normally be observed.
Reply 5: We agree with your concern. For internal validation, we divided the data into 70% and 30% initially, and constructed a model with 70% of the data. After dividing the dataset into two, we kept the validation dataset fully independent and did not assess the data during the model development process. In addition, when we performed imputation for BMI, the imputation was conducted separately.

However, we identified a similar discrimination and calibration of the model in the validation dataset. This was because the training and validation datasets were extracted from the same data sources. In combination with your comment 8, we described it as one of the limitations of the study as follows:
"Second, the dataset for model construction and validation were extracted from the same data source. Validation of the model in a different population, such as people who did not undergo NHIS health examination, would increase the external validity of the model. However, when the model was applied to select high-risk populations based on questionnaires and measurements during health examinations, training and validation among people who received health examinations was more appropriate for application in the real world."

Changes in text: page 19 line 381-386

Comment 6: Also in Table 4, please make the pack-year categories mutually exclusive. Further subgroup analyses also need to be provided, e.g. by age, sex, and other key variables.
Reply 6: We revised Table 4 for pack-year categories to make them mutually exclusive as follows:

Table 4. Prediction performance of the lung cancer prediction model in Korean ever-smokers

Statistic	Value (95% Confidence interval)
Harrell's C-index in the training dataset	
Ever-smokers	0.816 (0.810–0.822)
Current smokers	0.816 (0.808–0.824)
Past smokers	0.804 (0.790–0.818)
Smokers with <10 pack-years	0.787 (0.760–0.814)
Smokers with 10-19.9 pack-years	0.812 (0.800–0.818)
Smokers with 20-29.9 pack-years	0.826 (0.820–0.829)
Smokers with ≥30 pack-years	0.754 (0.746–0.762)
E/O ratio in the training dataset	
Ever-smokers	1.002 (0.979–1.024)
Current smokers	0.881 (0.858–0.904)
Past smokers	1.510 (1.444–1.580)
Smokers with <10 pack-years	1.143 (1.043–1.252)
Smokers with 10-19.9 pack-years	0.919 (0.915–0.924)
Smokers with 20-29.9 pack-years	0.984 (0.979–0.989)

Smokers with ≥30 pack-years	0.958 (0.930–0.988)
Harrell's C-index in the validation dataset	
Ever-smokers	0.816 (0.806–0.826)
Current smokers	0.816 (0.804–0.828)
Past smokers	0.803 (0.783–0.823)
Smokers with <10 pack-years	0.797 (0.758–0.836)
Smokers with 10-19.9 pack-years	0.819 (0.801–0.829)
Smokers with 20-29.9 pack-years	0.823 (0.809–0.830)
Smokers with ≥30 pack-years	0.753 (0.739–0.767)
E/O ratio in the validation dataset	
Ever-smokers	0.989 (0.956–1.023)
Current smokers	0.824 (0.793–0.857)
Past smokers	1.504 (1.404–1.611)
Smokers with <10 pack-years	1.072 (0.936–1.227)
Smokers with 10-19.9 pack-years	0.913 (0.906–0.919)
Smokers with 20-29.9 pack-years	1.036 (1.029–1.044)
Smokers with ≥30 pack-years	0.962 (0.919–1.007)

Changes in text: page 35 Table 4

Comment 7: I understand that 6.6 years of follow-up was the mean, but it would be more convenient to choose e.g. a 5-year timeframe for the model prediction. This is easier to interpret and align with existing risk models.

Reply 7: We selected 6.6 years of follow-up since it was the mean follow-up period of the study population. A 5-year timeframe for model prediction is often applied for breast cancer, for which prediction models have been most widely used. In the Gail model for breast cancer, women with a 5-year risk of >1.67% were considered to be in the high-risk group (8). However, for lung cancer prediction models, several time frames have been applied, such as 1-year for the Bach, TSCE, and Knoke models, 5-year for the LLP and LCRAT models, and 6-year for the $PLCO_{M2012}$ model (3, 9). In addition, unlike in the logistic regression-based model, in which the follow-up duration of each participant was assumed to be similar, we applied a Cox regression model, in which different follow-up times of participants were considered. Thus, we applied 6.6-year time frame.

Comment 8: A key limitation is the lack of external validation. Most models perform well in split-sample validation, but problems arise when transporting to new datasets. The paper would be much stronger with another validation in a new dataset that was not used to develop the model, though Reply:

Reply 8: We agree with your opinion that a lack of external validation was one of our key limitations. The dataset for the model construction and validation was obtained from national health examinees, considering possible data sources and future applications to select high-risk

populations based on questionnaires and measurements during health examinations. In combination with comment 5, we described it as one of the limitations of the study as follows: "Second, the dataset for model construction and validation were extracted from the same data source. Validation of the model in a different population, such as people who did not undergo NHIS health examination, would increase the external validity of the model. However, when the model was applied to select high-risk populations based on questionnaires and measurements during health examinations, training and validation among people who received health examinations was more appropriate for application in the real world."

Changes in text: page 19 line 381-386

Also, a few important points about the validation of Western risk models.

Comment 9: I did not understand why the authors didn't evaluate the models that were previously identified to be well-performing. Zeros can be imputed for the occupational variables as needed. Is the dataset lacking information on cause of death, to evaluate LCDRAT? The most important models to validate would be Bach, LCRAT, LCDRAT, PLCOm2012, and LLPv2 or LLPv3 – because these have previously been the best performing and most transportable. See Katki et al Annals of Internal Medicine 2018; ten Haaf et al PLoS Medicine 2017; Robbins et al Br J Cancer 2021; and others.

Reply 9: We tried to evaluate as many models as possible. The LCDRAT model predicts the 5-year cumulative risk of lung cancer death; however, in our data, we had information on death but did not have information on the cause of death. Thus, we did not evaluate the LCDRAT model. However, we agree with your opinion that more lung cancer incidence models developed in Western countries need to be evaluated in the Korean population. We additionally evaluated the Bach and LCRAT models; thus, five models, including Bach, LCRAT, $PLCO_{M2012}$, Pittsburgh, and LLPi, were evaluated. For the Bach model, all individuals were assumed to have no asbestos exposure, and for the LCRAT model, all individuals were assumed to have some college level of education and no family history of lung cancer. The results are presented in Table 2.

Table 2. Predictive performance of previously developed models

	Area under the curve	Expected/observed ratio
Bach	0.661 (0.598–0.665)	2.23 (2.07–2.40)
LCRAT	0.811 (0.807–0.814)	4.73 (4.50–4.96)
$PLCO_{M2012}$ 2013†	0.772 (0.768–0.777)	1.24 (1.22–1.26)
Simplified $PLCO_{M2012}$ 2013‡	0.781 (0.776–0.785)	1.10 (1.08–1.16)
Pittsburgh 2015	0.781 (0.778–0.784)	1.21 (1.19–1.23)
LLPi 2015§	0.803 (0.800–0.806)	3.25 (3.20–3.31)
Simplified LLPi 2015¶	0.803 (0.800–0.806)	3.21 (3.16–3.26)
Korean model	0.816 (0.810–0.822)	0.995 (0.973–1.017)

LCRAT, Lung Cancer Risk Models for Screening; $PLCO_{M2012}$, the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial Model 2012; LLPi, Liverpool Lung Project models
† The race was assumed to be Asian. The educational level was assumed to be some college. A family history of lung cancer was imputed as a family history of any cancer.
‡ The race was assumed to be Asian. The educational level was assumed to be some college. A family history of lung cancer was imputed as none.
§Family history of lung cancer was imputed as a family history of cancer. All participants with a family history of any cancer were assumed to have a late onset.
¶Family history of lung cancer was imputed as none.

Changes in text: page 33 Table 2

We have added the Abstract, Method, Results, and Discussion sections as follows:
"Performance of Bach, Lung Cancer Risk Models for Screening, $PLCO_{M2012}$, Pittsburgh, and Liverpool Lung Project models were evaluated."

Changes in text: page 3 line 57-58

"The Korean lung cancer risk model showed better discrimination and calibration than previously developed models in Western population."

Changes in text: page 3 line 71-72

"Among the nine models previously validated in the US population(11), models for lung cancer death(8, 21) were not considered because in the data, information on the cause of death was not available. Subsequently, five models, including Bach (22), lung cancer risk models for screening (LCRAT), (8) the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial Model 2012 ($PLCO_{M2012}$), Pittsburgh, and Liverpool Lung Project models (LLPi)(10, 23, 24), were applied to the participants."

Changes in text: page 8 line 149-154

"When these five models were applied to Korean population, the range of AUCs was 0.661–0.811, and Bach showed lowest discrimination and LCRAT showed better discrimination than other models."

Changes in text: page 12 line 227-229

"Regarding calibration, all models overestimated the risk in Koreans, with an E/O ratio of 1.10–4.73. Specifically, the LCRAT and LLPi models overestimated the lung cancer risk."

Changes in text: page 12 line 235-237

"When models developed for ever-smokers in the Western population were applied to the

Korean population, they moderately discriminated people who would develop and those who would not develop lung cancer (AUC, 0.66–0.81).”

Changes in text: page 16 line 306-308

Comment 10: In the text, the authors say they imputed the ‘highest level’ of education, but the Table 2 footnote implies that they imputed ‘some college’. This could have a strong effect on the E/O’s. This needs to be clarified, and an important sensitivity analysis could be to calculate E/O’s after imputation of different levels of education.

Reply 10: In the $PLCO_{M2012}$ model, the odds ratio of a one-level increase in education level was 0.922 (beta coefficient -0.081274). The category of education in the $PLCO_{M2012}$ model had six ordinal levels: lower than high-school graduates (level 1), high-school graduates (level 2), some training after high school (level 3), some college (level 4), college graduate (level 5), and postgraduate or professional degree (level 6), and the model was centered on level 4. Therefore, imputation with a lower level of education would increase the E/O ratio. Based on Education at a Glance published by OECD (Available at: https://www.oecd.org/education/education-at-a-glance/), the proportion of Korean people with some college or more was around 50%. Thus, we considered imputation value with ‘some college” as appropriate for the Korean population. We have corrected the Method section as follows:

“Therefore, we assumed that none of the participants had family history of lung cancer and all the participants had ‘some college’ level which was a reference for $PLCO_{M2012}$(10) despite the biased estimates (25), similar to a previous study (12). For asbestos exposure, all participants were assumed to be non-exposed.”

Changes in text: page 9 line 157-160

In addition, we added it as one of the limitations of this study as follows:

“In addition, for the unavailable variables in our data, we considered them as a single value. It could cause biased estimates (28) and overall over- or under-estimation of the risk.”

Changes in text: page 20 line 393-395

Comment 11: It would also be interesting to evaluate previously developed risk models for Asia. The authors mention that a few have been developed but do not explain why they weren’t evaluated. Please note, even if a model were developed among both never and ever-smokers, it can certainly be evaluated in a population of ever-smokers.

Reply 11: We tried to evaluate two previously developed lung cancer risk models for Asians: one was developed for Korean men and the other was developed for Japanese people. These two models include age at smoking initiation as one of the factors (age < 16, $16 \leq$ age < 19, $19 \leq$ age 30, $30 \leq$ age ≤ 40, and age ≥ 40 years for Korean men, and for 1-year increments for the Japanese population). However, in our study population, information on the age at smoking initiation was not available. Thus, we could not evaluate the previously developed risk models for Asians. In addition, several studies have demonstrated that never smokers are unlikely to reach levels of risk that allow them to benefit from screening; thus, we did not consider models

for both smokers and non-smokers. We included this as one of the limitations of the study as follows:
"Third, we evaluated the previously developed models, but due to the unavailable information included in the previously developed models, only five models could be evaluated and models developed in Asian countries could not be evaluated."

Changes in text: page 19-20 line 391-393

Minor
Comment 12: I would suggest adding 'in Korea' or something similar to the title of the manuscript, to differentiate it from other risk models.
Reply 12: We revised the title of the manuscript as follows;
"Risk-based prediction model for selecting eligible populations for lung cancer screening among ever-smokers in Korea"

Changes in text: page 1 line 3-4

Comment 13: The main criticism of risk models for lung screening eligibility is that they lead to older people being screened. It would be useful to report the median age of people screened under NLST vs risk model criteria in this population.
Reply 13: As per your comment, we described the median age of people screened under NLST vs. risk model criteria in this population in the Results section as follows:
"The median ages of selected people through the NLST criteria or 6.6-year lung cancer risk cut-off of >2.1% were 63.1 and 69.6 in the training and validation dataset, respectively."

Changes in text: page 15 line 287-289

Comment 14: The correct terminology for split-sample design can be 'training and validation' or 'training and testing' but not 'test and validation'.
Reply 14: We changed the term "test dataset" to "training dataset". Please refer to the revised manuscript for clarity.

Comment 15: In the results, please provide the IQR of follow-up time.
Reply 15: We provided the IQR of the follow-up time as follows;
"The mean and interquartile range of the follow-up times were 6.6 years and 6.2-7.1 years, respectively."

Changes in text: page 12 line 222-223

Comment 16: Please check the description of ref 17 in the discussion – it didn't apply PLCOm2012 in never smokers (it is not valid for never smokers) but developed a different model for never smokers.
Reply 16: We included reference 17 because the study was against lung cancer screening among never smokers, stating that never smokers should not be screened. If the reviewer thinks that

the reference would not be appropriate, we will exclude it.

Comment 17: The explanation in the discussion regarding who changes their eligibility status based on the risk model is quite useful. I would suggest moving this to the results section and presenting a table to fully present the data. This could include the median ages mentioned above.
Reply 17: We moved the explanation in the discussion regarding who changed their eligibility status based on the risk model to the Results section. The following sections were moved from the Discussion to Results:
"When we applied the model-based cut-off risk of >2.1% instead of the NLST criteria, 73.5% of the participants remained ineligible, 8.4% remained eligible, and 18.1% changed eligibility statuses. For individuals who changed from ineligible using the NLST criteria to eligible using the model-based cut-off, 3.7% developed lung cancer within 6.6 years. For individuals who changed from eligible to ineligible, only 1.6% developed lung cancer. Individuals who changed from ineligible to eligible were older, predominantly women, and had more underlying pulmonary diseases (Table 5)."

Changes in text: page 14-15 line 281-287

Comment 18: Last paragraph of the discussion – there are clear risk thresholds that have been proposed for the PLCOm2012, LCRAT/LCDRAT, and LLPv2 models at minimum. See the draft protocol for NHS England lung screening, and papers by Katki, Landy, Tammemagi, etc. One could also define thresholds for these models using the same approach (equal number of people screened as NLST criteria) and then use those thresholds to compare directly with the new Korean risk model. This approach would allow the most direct comparison.
Replay 18: As per your comment, the other models also applied a cut-off with an equal number of people screened as the NLST criteria. Thus, we excluded that part from the Discussion.

Comment 19: Please clarify what data were used to generate Table 2 (training, validation, or both combined).
Reply 19: In Table 2, the whole dataset, including both the training and validation sets, was applied. We clarified this as follows in the Methods section.
"The performance of each model in the whole dataset was presented as discrimination (receiver operating characteristic curve and area under the curve [AUC]) and calibration (expected/observed [E/O] ratio)."

Changes in text: page 9 line 167-169

Comment 20: I would suggest some additional editing of the manuscript. There are a few sentences throughout that don't make sense either because of language/sentence structure or logic that isn't clear. There is also some repetition in the manuscript and some inconsistencies.
Reply 21: We submitter the initially submitted version of our manuscript for English editing. Based on your comment, we received submitted the revised manuscript for English editing again. Please refer to the revised manuscript for clarity.

Reviewer B

This study did contribute knowledge to the field of lung cancer prevention, To better improve the manuscript, please consider incorporating the following comments in the subsequent revision.

Comment 1. Background information on lung cancer issues (e.g., incidence, prevalence, and mortality rate) among Korean should be addressed in the Introduction section.
Reply 1: We have included a brief description of lung cancer incidence and mortality in the Introduction section as follows:
"In Korea, lung cancer is the third most common cancer with 28,628 new cases and is the most common cause of cancer-related deaths in 2018. The crude incidence and mortality rate of lung cancer was 55.8 and 34.8 per 100,000 person-years, respectively (14)."

Changes in text: page 5 line 94-96

2. Line#111-117: It will be clear to provide a flow chart to show subject inclusion and exclusion.
Reply 2: We have added a flow chart of the selection process of the study population in Figure 1. Please refer to Figure 1.

3. Line#162: more information on how the 6.6-year was obtained is needed.
Reply 3: We have included the calculation of follow-up time as follows:
"The follow-up time (person-years) was calculated from the date of the health examination to December 31, 2014, date of death, or date of lung cancer diagnosis, whichever came first."

Changes in text: page 9 line 175-177

4. Line#179-181: this sentence might be confusing to the audience. Please use a plain language to describe how the threshold was determined.
Reply 4: We have revised the sentence for better understanding as follows:
"We selected a model-based lung cancer risk threshold at which the equal number of people screened using the NLST criteria was selected."

Changes in text: page 11 line 210-212

5. Line#195: check accuracy for this sentence: "In all the models, the AUC was 0.772–0.803"
Reply 4: For clarity, we have revised the sentence as follows:
"When these five models were applied to Korean population, the range of AUCs was 0.661–0.811, and Bach showed lowest discrimination and LCRAT showed better discrimination than other models"

Changes in text: page 12 line 227-229

6. Table 3 results should be briefly described in the text.
Reply 6: As per your comment, the results of Table 3 are described as follows:
“History of interstitial pulmonary disease (HR 4.424, 95% CI=3.475–5.633) was the most significant predictor of lung cancer, followed by smoking status (HR of current smokers 2.824, 95% CI=2.543–3.137) and history of pneumoconiosis (HR 2.05, 95% CI=1.335–3.149). The HR of the square root of pack-years was 1.191 (95% CI=1.174–1.208).”

Changes in text: page 13 line 250-253

7. Line#222: It's "Table 4".
Reply 7: This was a typographical error. We have changed “Table 3” to “Table 4.”

8: Line#225-227: these were results from the validation dataset. Also, past smokers had a lower C-Index value, did it mean past smokers have lower discrimination than current smokers (0.803<0.816)?
Reply 8: Based on the Harrell’s C-index, it could be suggested that past smokers have lower discrimination than current smokers by 1.3%. However, the authors deemed that the 1.3% of the lower C-index was minimal and not significant.

9. Please have some discussions on the application of the new prediction model in the Korean American population
Reply 9: To apply this model to the Korean American population, another validation study would be needed. Therefore, it would be too early to discuss the application of the new prediction model to the Korean American population without a validation study. Thus, we did not include this point in the Discussion section.

10. What are the study impluications for policy, practice and future research?
Reply 10: We have included the implications of the study for policy, practice, and future research as follows:
“Korea started the national lung cancer screening program based on the eligibility criteria of NLST in 2019. It could be expected that a combination of lung cancer prediction models tailored for the Korean population and a national lung cancer screening program would provide a more efficient nationwide screening program. Further research on the cost-effectiveness of the model-based and current criteria of the national lung cancer screening program in the Korean population is needed.”

Changes in text: page 20 line 402-407

References

1. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. BMC Medicine. 2019;17(1):230.
2. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. The New England journal of medicine. 2011;365(5):395-409.
3. 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. Annals of internal medicine. 2018;169(1):10-9.
4. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Medical decision making : an international journal of the Society for Medical Decision Making. 2015;35(2):162-9.
5. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ (Clinical research ed). 2009;338:b2393.
6. Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P. Strategies for dealing with missing data in clinical trials: from design to analysis. The Yale journal of biology and medicine. 2013;86(3):343-58.
7. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Medical Research Methodology. 2017;17(1):162.
8. Wang X, Huang Y, Li L, Dai H, Song F, Chen K. Assessment of performance of the Gail model for predicting breast cancer risk: a systematic review and meta-analysis with trial sequential analysis. Breast Cancer Research. 2018;20(1):18.
9. 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 medicine. 2017;14(4):e1002277.