


# BMJ Open Can smoking duration alone replace pack-years to predict the risk of smoking-related oncogenic mutations in non-small cell lung cancer? A cross-sectional study in Japan

Koichi Ogawa <sup>1</sup>, Yasuhiro Koh,<sup>2</sup> Hiroyasu Kaneda,<sup>3</sup> Motohiro Izumi,<sup>1</sup> Yoshiya Matsumoto,<sup>1</sup> Kenji Sawa,<sup>1</sup> Mitsuru Fukui,<sup>4</sup> Yoshihiko Taniguchi,<sup>5</sup> Naoki Yoshimoto,<sup>3</sup> Akihiro Tamiya,<sup>5</sup> Masahiko Ando,<sup>6</sup> Akihito Kubo,<sup>7</sup> Shun-ichi Isa,<sup>8</sup> Hideo Saka,<sup>9</sup> Akihide Matsumura,<sup>10</sup> Tomoya Kawaguchi<sup>1,3</sup>

**To cite:** Ogawa K, Koh Y, Kaneda H, *et al.* Can smoking duration alone replace pack-years to predict the risk of smoking-related oncogenic mutations in non-small cell lung cancer? A cross-sectional study in Japan. *BMJ Open* 2020;**10**:e035615. doi:10.1136/bmjopen-2019-035615

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2019-035615>).

Received 15 November 2019

Revised 24 June 2020

Accepted 04 August 2020



© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

## Correspondence to

Professor Tomoya Kawaguchi; [kawaguchi.tomoya@med.osaka-cu.ac.jp](mailto:kawaguchi.tomoya@med.osaka-cu.ac.jp)

## ABSTRACT

**Objective** To investigate whether smoking duration alone can replace pack-years to predict the risk of oncogenic mutations in non-small cell lung cancer (NSCLC).

**Design** A cross-sectional study using the baseline dataset from the Japan Molecular Epidemiology for Lung Cancer Study.

**Setting** Forty-three medical institutions nationwide in Japan.

**Participants** From July 2012 to December 2013, 957 patients with newly diagnosed stage I–IIIB NSCLC who underwent surgery were enrolled, and molecular analyses were performed on 876 samples (from 441 ever-smokers and 435 never-smokers).

**Main outcomes measured** We calculated the area under the receiver operating characteristic curve (AUC) values using logistic regression to compare between the predictive values of smoking duration and pack-years for mutational frequencies in the v-Ki-ras2 Kirsten rat sarcoma (*KRAS*), tumour suppressor p53 (*TP53*), and epidermal growth factor receptor (*EGFR*) genes and for cytosine-to-adenine base substitution (C>A).

**Results** For predicting *KRAS* mutations, the AUC values for smoking duration and pack-years were 0.746 (95% CI 0.682 to 0.800) and 0.759 (95% CI 0.700 to 0.810), respectively ( $p=0.058$ ). For predicting *KRAS* mutations in smokers, the AUC values for smoking duration and pack-years were 0.772 (95% CI 0.697 to 0.833) and 0.787 (95% CI 0.714 to 0.845), respectively ( $p=0.036$ ). There were no significant differences between the AUC values for smoking duration and pack-years in terms of predicting *TP53* and *EGFR* mutations and C>A. Pack-years was a significantly better predictor of *KRAS* mutations than smoking duration.

**Conclusion** Smoking duration was not significantly different from pack-years in predicting the likelihood of smoking-related gene mutations. Given the recall bias in obtaining smoking information, smoking duration alone should be considered for further investigation as a simpler alternative to pack-years.

## Strengths and limitations of this study

- This study is the first to show the comparison between two indices (smoking duration vs pack-years) in predicting the risk of smoking-related oncogenic mutations in non-small cell lung cancer (NSCLC).
- It focuses on the mutations in the v-Ki-ras2 Kirsten rat sarcoma (*KRAS*), tumour suppressor p53 (*TP53*) and epidermal growth factor receptor (*EGFR*) genes and cytosine-to-adenine base substitution (C>A), which were associated with pack-years based on the results of the Japan Molecular Epidemiology (JME) study, a prospective multicentre molecular epidemiology study.
- A limitation of this study was that the JME study data were obtained from targeted sequencing and not from whole genome or whole exome sequencing.

## INTRODUCTION

Lung cancer is the leading cause of cancer-related morbidity and mortality worldwide.<sup>1</sup> Low-dose CT has been shown to be effective as a screening test for lung cancer, but optimal eligibility for screening remains undetermined.<sup>2</sup> According to epidemiological studies, cancer development primarily occurs due to environmental factors.<sup>3</sup> Smoking is the most assessed cause of cancer and contributes to lung cancer development. There is convincing evidence that tobacco smoking strongly increases the risk of lung cancer, with a relative risk of approximately 4.4 in men and 2.8 in women for current smokers compared with never-smokers.<sup>4</sup>

V-Ki-ras2 Kirsten rat sarcoma (*KRAS*) and epidermal growth factor receptor (*EGFR*) mutations are well documented in the pathogenesis of lung adenocarcinoma according to

smoking status.<sup>5</sup> *KRAS* mutations show no sex predilection but are more frequent in Caucasians than in Asians, and most patients with these mutations are former or current cigarette smokers.<sup>6,7</sup> Unlike *KRAS* mutations, it has been reported that *EGFR* mutations are more frequently found in women, Asians, and never-smokers.<sup>8,9</sup> Information regarding pack-years has been widely used to assess the risks of lung cancer and chronic obstructive pulmonary disease (COPD).<sup>10,11</sup> Pack-years of smoking is calculated by multiplying the number of packs of cigarettes smoked per day by the number of years the person has smoked. The total number of base substitution mutations is positively correlated with pack-years smoked for all cancer types; based on these correlation rates, it is estimated that the approximate number of mutations accumulated in a normal cell of each tissue due to smoking a pack of cigarettes per day for a year is 150, particularly in the lungs.<sup>12</sup> A recent study showed that smoking duration alone reportedly provides stronger risk estimates of COPD than the composite index of pack-years.<sup>13</sup>

However, it remains unknown whether smoking duration can also replace pack-years for predicting smoking-related oncogenic mutations in non-small cell lung cancer (NSCLC). We previously reported a prospective, multicentric, molecular epidemiology study, the Japan Molecular Epidemiology (JME) study, which included comprehensive smoking information based on a detailed questionnaire and the mutational profiles of 72 genes using next-generation sequencing.<sup>14</sup> The prevalence of *KRAS*, tumour suppressor p53 (*TP53*) and *EGFR* mutations in this study were associated with smoking dose. In addition, cytosine-to-adenine base substitutions (C>A) were reported as the most significant smoking-related base substitution pattern.

Using the baseline dataset of this prospective cohort study, we investigated whether smoking duration alone could be an alternative index to pack-years in predicting the risk of oncogenic mutations in NSCLC.

## PATIENTS AND METHODS

### Study design and patient population

The eligibility criteria and questionnaire were previously described in the JME protocol (see online supplemental file 1).<sup>14</sup> Patients with newly diagnosed stage I–IIIB NSCLC who underwent surgery were considered eligible. Patients with a history of chemotherapy and/or radiotherapy and patients with a history of malignancies, other than adequately treated basal cell or squamous cell skin cancer or in situ cervical cancer, were excluded. The participants were required to complete a questionnaire before surgery that was modelled after the one designed for SWOG study S0424<sup>15</sup> to assess the following parameters in detail: smoking history, occupational exposures, reproductive and hormonal risk factors, weight loss, family history of cancer, medication history and current lifestyle (diet and exercise). The S0424 was originally designed to address the association between sex and lung

cancer carcinogenesis by using a detailed questionnaire and tissue specimens from smoker and never-smoker men and women with newly diagnosed stage I–III NSCLC.<sup>15</sup>

Formalin-fixed paraffin-embedded surgical tissues were sent to a central laboratory for genomic analysis and immunohistochemical staining. DNA was extracted from the samples, and quality-control assessments were performed as described previously.<sup>16</sup> Multiplexed, targeted deep sequencing was performed on MiSeq (Illumina, San Diego, California, USA) using a TruSeq Amplicon Cancer Panel and an additional custom panel (Illumina) to evaluate the tumours. Somatic mutations in 72 cancer-associated genes and copy numbers of five cancer-associated genes were selected based on previous reports<sup>17–19</sup> and were evaluated to cover the range of critical mutations.

### Statistical analyses

The JME study followed and extended the concept of S0424 by using the similar approach with the same questionnaire that would allow for direct comparison of the data. The sample size of the JME study was adjusted with reference to the sample size of S0424.

The correlations between smoking status (ever-smoker or never-smoker) and demographic factors, such as age, sex, histology and pathological stage, were examined using the  $\chi^2$  test. A logistic regression model was used for multivariate analysis. To evaluate the predictive values of pack-years and smoking duration for the above-mentioned mutations or C>A, we compared the area under the receiver operating characteristic curve (AUC) calculated using logistic regression, considering sex, age, stage and histology as covariates.

### Patient and public involvement

This research was performed without patient involvement. Patients were not invited to comment on the study design and were not consulted for determining patient-relevant outcomes or interpreting the results. Patients were not invited to contribute to the writing or editing of this document for readability or accuracy.

## RESULTS

### Clinical characteristics

From July 2012 to December 2013, 957 patients were recruited from 43 institutions of the National Hospital Organization, and information regarding environmental factors was obtained through questionnaires. For molecular analyses, 876 samples were successfully tested for gene mutations. Overall, 622 cases involved at least one mutation, and a total of 860 mutations were detected. Clinicopathological characteristics according to smoking status are shown in table 1.

There were 441 ever-smokers and 435 never-smokers in the JME study. The median smoking duration and pack-years were 41 (1–65) years and 43 (1–189) in ever-smokers, respectively. The frequencies of mutations in *KRAS*, *TP53*

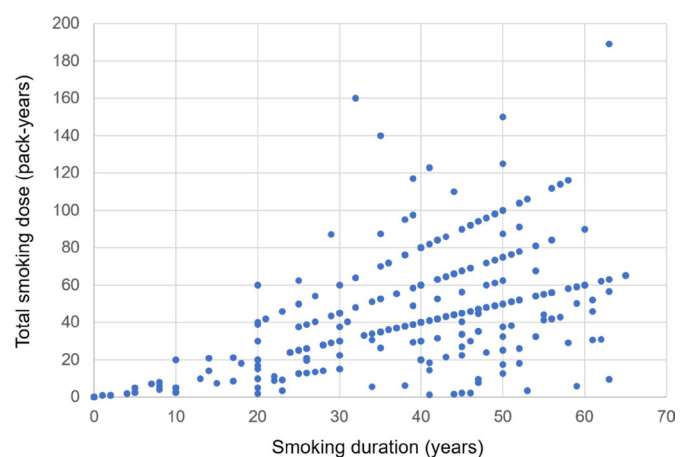
**Table 1** Clinicopathological characteristics according to smoking status

Characteristic	Ever smoker		Never smoker		P value
	Number of patients	%	Number of patients	%	
<b>Age at surgery, years</b>					
Median	69		71		
Range	41–89		23–92		
<b>Sex</b>					
Male	366	83	53	12.2	<0.001
Female	75	17	382	87.8	
<b>Smoking duration</b>					
Never smoker	0	0	435	100	<0.001
0<years<20	28	6.3	0	0	
20≤years<40	126	28.6	0	0	
40≤years	246	55.8	0	0	
<b>Pack-years</b>					
Never smoker	0	0	435	100	<0.001
0<pack-years<30	102	23.1	0	0	
30≤pack-years<60	184	41.7	0	0	
60≤pack-years	109	24.7	0	0	
<b>Histology</b>					
Adenocarcinoma	265	60.1	415	95.4	<0.001
Squamous cell carcinoma	135	30.6	7	1.6	
Others	41	9.3	13	3	
<b>Pathological stage</b>					
I	280	63.5	338	77.7	<0.001
II	81	18.4	50	11.5	
III	65	14.7	39	9	
IV	15	3.4	8	1.8	
Total	441	50.3	435	49.7	

and *EGFR* were 9.4%, 26.8% and 42.5%, respectively, and C>A was observed in 12.7% of cases. The distributions of smoking duration and pack-years are shown in [figure 1](#). In ever-smokers, the most frequent mutations were in *TP53* (38.3%), *EGFR* (20.2%) and *KRAS* (13.2%), and C>A was observed in 21.1% of cases, whereas in never-smokers, *EGFR* (65.1%), *TP53* (15.2%) and *KRAS* (5.5%) harboured the most frequent mutations, and C>A was observed in only 4.1% of cases.

#### Mutational frequencies associated with smoking duration or pack-years

We divided all cases into four groups according to smoking duration (never, light (0<duration<20 years), middle (20≤duration<40 years) and heavy (≥40 years)) and pack-years (never, light (0<packyears<30), middle (30≤packyears<60) and heavy (≥60 pack-years)). The frequencies of *KRAS* mutations in the never, light, middle and heavy smoking duration groups were 4.1%, 7.1%, 11.1% and 14.2%, respectively, while those in the never,



**Figure 1** Scatter diagram showing the distributions of smoking duration (longitudinal axis) and pack-years (horizontal axis). The frequency of smoking <20 cigarettes per day is higher than that of smoking ≥20 cigarettes per day.

light, middle and heavy pack-year groups were 4.1%, 8.8%, 13.0% and 14.7%, respectively (figure 2A). The frequencies of *TP53* mutations in the never, light, middle and heavy smoking duration groups were 15.1%, 25.0%, 38.9% and 41.5%, respectively; those in the never, light, middle and heavy pack-year groups were 15.1%, 32.4%, 39.1% and 46.8%, respectively (figure 2B). The frequencies of *EGFR* mutations in the never, light, middle and heavy smoking duration groups were 60.9%, 53.6%, 24.6% and 13.0%, respectively; those in the never, light, middle and heavy pack-year groups were 60.9%, 32.4%, 19.0% and 9.2%, respectively (figure 2C). The frequencies of C>A in the never, light, middle and heavy smoking duration groups were 4.1%, 10.7%, 23.0% and 22.3%, respectively; those in the never, light, middle and heavy pack-year groups were 4.1%, 15.7%, 19.6% and 17.4%, respectively (figure 2D).

We examined the associations between the frequency of mutations or C>A and the smoking duration or pack-year groups using logistic regression. The frequency of *KRAS* and *TP53* mutations or C>A in the smoking duration or pack-year groups increased significantly with an increase in smoking exposure (all groups:  $p < 0.001$ ). In contrast, the frequency of *EGFR* mutations in the smoking duration or pack-year groups decreased significantly with an increase in smoking exposure (all groups:  $p < 0.001$ ).

The ORs calculated using logistic regression are shown in table 2. Although the ORs for smoking duration were slightly higher than those for pack-years, no significant differences were observed.

### Comparison of mutational frequencies between smoking duration and pack-years

To compare between the predictive values of smoking duration and pack-years for mutational frequencies, we calculated the AUC values using logistic regression (table 3). For *KRAS* mutations in the overall population, the AUC values for smoking duration and pack-years were 0.746 and 0.759, respectively ( $p = 0.058$ ), whereas for *KRAS* mutations in cases involving smokers, the AUC values for smoking duration and pack-years were 0.772 and 0.787, respectively ( $p = 0.036$ ). There were no significant differences in the AUC values of smoking duration and pack-years for *TP53* and *EGFR* mutations and C>A. However, pack-years was a significantly better predictor of *KRAS* mutations than smoking duration in ever-smokers.

### DISCUSSION

This study showed that there were no significant differences in AUC values between smoking duration and pack-years for *TP53* and *EGFR* mutations and for C>A, but pack-years was a significantly better predictor of *KRAS* mutations than smoking duration in ever-smokers. Dogan *et al* reported that pack-years of smoking has a significant predictive value for *KRAS* and *EGFR* mutations in lung adenocarcinomas,<sup>20</sup> but they did not compare pack-years with smoking duration with respect to prediction of *KRAS*

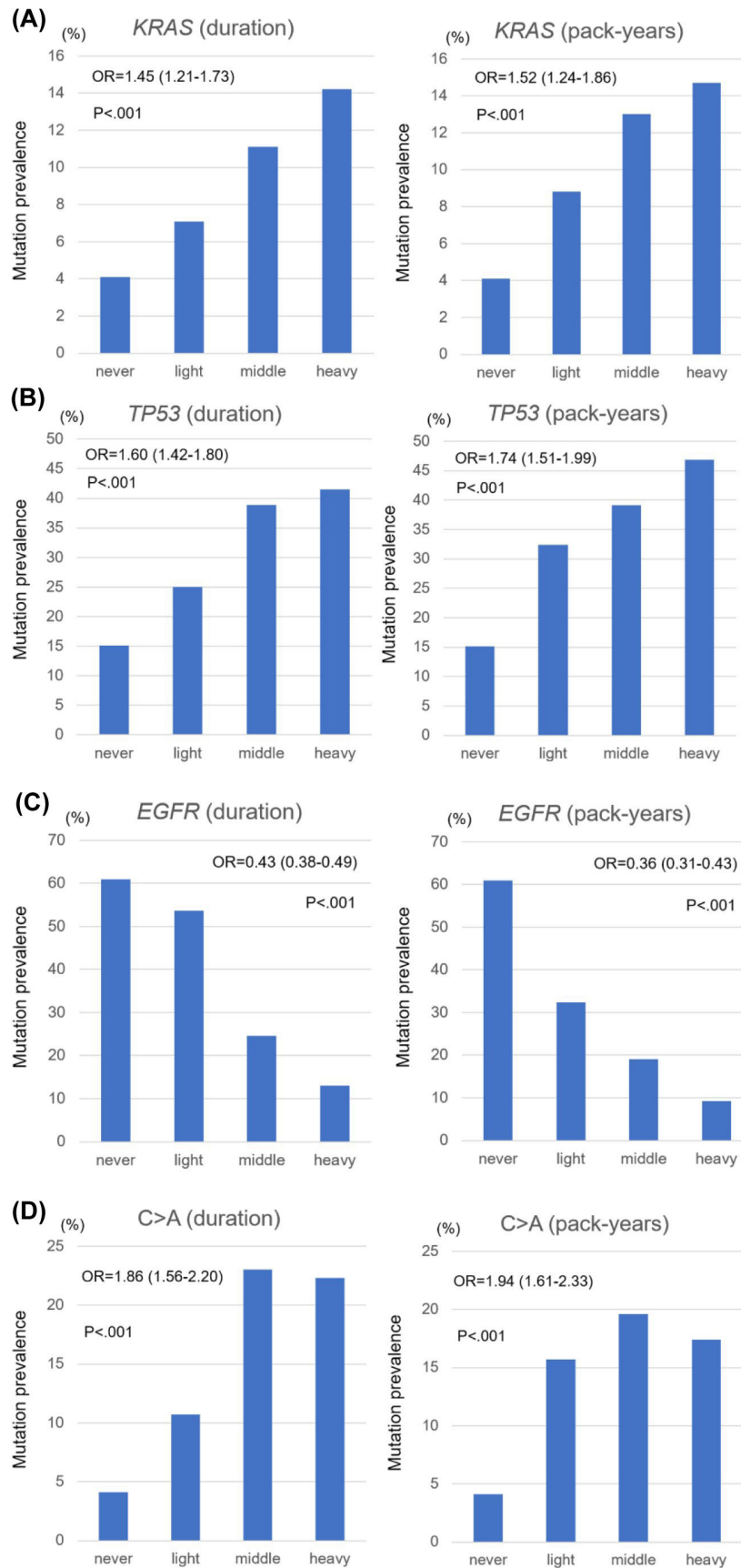
and *EGFR* mutations. To the best of our knowledge, our study is the first to compare these two indices (smoking duration vs pack-years) in predicting the risk of smoking-related oncogenic mutations in NSCLC. For COPD, the strength of the association between smoking duration and COPD was greater than that between pack-years and forced expiratory volume in 1s ( $FEV_1$ )/forced vital capacity, emphysema, gas trapping,  $FEV_1$ , 6min walking distance and St George's Respiratory Questionnaire.<sup>12</sup>

The relative contributions of smoking duration and cigarettes smoked per day to lung cancer incidence have been examined but not in terms of the incidence of driver gene alterations.<sup>21 22</sup> Smoking duration was more strongly associated with lung cancer development than cigarettes smoked per day,<sup>23</sup> but no comparisons were made in terms of pack-years. It is accepted that a longer duration of smoking is associated with increasing accumulation of genetic and epigenetic changes. A long smoking history is almost always self-reported, and it is conceivable that the lower predictive value of smoking intensity is due to the fact that smoking duration may be recalled and reported with greater accuracy than average daily intensity over a lifetime of smoking history.<sup>23</sup> We believe that the duration of smoking is more easily and accurately recalled than the average number of cigarettes smoked per day, which tends to fluctuate over time. It is also harder to accurately quantify the number of cigarettes smoked per day, and the measurements are correlated poorly with the biochemical assessments of smoking exposure.<sup>24</sup>

A retrospective analysis of individuals referred to centralised lung cancer screening programmes<sup>25–27</sup> serving a 5-hospital health services system in Seattle, Washington between October 2014 and January 2016 has been reported. The study assessed the eligibility of individuals referred for lung cancer screening and compared the information extracted from electronic medical records (EMRs) with the information derived from a shared decision-making conversation to determine the eligibility for lung cancer screening. They found a 96.2% discordance in pack-year smoking history between EMRs and shared decision-making conversations. The EMRs under-reported pack-years of smoking for 85.2% of participants. If the identification of eligible individuals relied solely on the accuracy of pack-year smoking history recorded in EMRs, 53.6% of participants would have failed to meet the 30-pack-year threshold for screening. Over-reliance on EMRs for the identification of individuals at risk may lead to missed opportunities for appropriate lung cancer screening.<sup>28</sup>

Although figure 2 shows similar graphs for smoking duration and pack-years for *KRAS*, *TP53* and *EGFR* mutations and C>A, pack-years was a significantly better predictor of *KRAS* mutations than smoking duration in ever-smokers; this was not the case for *TP53* and *EGFR* mutations and C>A. Figure 1 shows an unbalanced distribution of smoking duration and pack-years; therefore, it is possible that the exact smoking dose has not been reflected. It has been reported that a subtype of *KRAS*





**Figure 2** The frequency of (A) *KRAS*, (B) *TP53* and (C) *EGFR* mutations and (D) C>A according to smoking duration and pack-years. As the smoking dose increased, the frequencies of *KRAS* and *TP53* mutations increased in the smoking duration and pack-year groups, but the frequency of *EGFR* mutations decreased with the increase in smoking dose. As the smoking dose increased, the frequency of C>A tended to increase in the smoking duration and pack-year groups.

**Table 2** ORs of smoking duration and pack-years for predicting *KRAS*, *TP53* and *EGFR* mutations and C>A

Mutations or C>A	Smoking index	OR (95% CI)	P value
<i>KRAS</i>	Duration	1.03 (1.01 to 1.04)	$1.07 \times 10^{-3}$
	Pack-years	1.01 (1.01 to 1.02)	$1.14 \times 10^{-3}$
<i>TP53</i>	Duration	1.02 (1.01 to 1.03)	<0.001
	Pack-years	1.01 (1.00 to 1.02)	$7.31 \times 10^{-3}$
<i>EGFR</i>	Duration	0.968 (0.957 to 0.978)	<0.001
	Pack-years	0.978 (0.969 to 0.987)	<0.001
C>A	Duration	1.04 (1.02 to 1.05)	<0.001
	Pack-years	1.01 (1.00 to 1.02)	$6.25 \times 10^{-3}$

**Table 3** The AUC values of smoking duration and pack-years for *KRAS*, *TP53* and *EGFR* mutations and C>A in all cases and in smokers

Mutations or C>A	Cases	Smoking index	AUC (95% CI)	P value
<i>KRAS</i>	All	Duration	0.746 (0.682 to 0.800)	0.058
		Pack-years	0.759 (0.700 to 0.810)	
	Smokers	Duration	0.772 (0.697 to 0.833)	0.036
		Pack-years	0.787 (0.714 to 0.845)	
<i>TP53</i>	All	Duration	0.700 (0.658 to 0.739)	0.894
		Pack-years	0.700 (0.658 to 0.738)	
	Smokers	Duration	0.627 (0.571 to 0.681)	0.774
		Pack-years	0.629 (0.573 to 0.682)	
<i>EGFR</i>	All	Duration	0.801 (0.770 to 0.829)	0.911
		Pack-years	0.801 (0.770 to 0.828)	
	Smokers	Duration	0.850 (0.803 to 0.888)	0.454
		Pack-years	0.844 (0.795 to 0.882)	
C>A	All	Duration	0.746 (0.693 to 0.792)	0.472
		Pack-years	0.736 (0.687 to 0.780)	
	Smokers	Duration	0.660 (0.593 to 0.721)	0.129
		Pack-years	0.644 (0.576 to 0.707)	

AUC, area under the receiver operating characteristic curve.

mutations was associated with smoking dose.<sup>29</sup> In Dogan *et al*'s study, the observed *KRAS* mutation subtypes were G12C (39.4%), G12V (20.7%), G12D (17.0%) and G12A (10.7%).<sup>20</sup> Never-smokers were significantly more likely to harbour transition mutations (G>A), rather than the transversion mutations known to be smoking-related (G>T or G>C), than ever-smokers.<sup>30</sup> G12C, a transversion mutation, was the most frequent mutation among ever-smokers, and G12D, a transition mutation, was the most frequent mutation among never-smokers.<sup>20,30</sup> In our study, the observed *KRAS* mutations were G12C (26.0%), G12V (24.7%), G12D (19.2%) and G12A (15.1%). The frequency of *KRAS* G12C mutations in our study was lower than that reported by Dogan *et al*. This may reflect the difference in the proportion of ever-smokers (50.3% in our study, 72.6% in their study), since *KRAS* G12C was found to be more strongly associated with smoking-associated signature four in lung adenocarcinoma than other *KRAS* mutations.<sup>29</sup>

To explain why pack-years was superior to smoking duration in predicting the frequencies of *KRAS* mutations, we divided *KRAS* subtypes (G12A, G12C, G12D, G12V) into four groups according to smoking duration and pack-years. The frequencies of *KRAS* subtypes in each group are shown in online supplemental figure 1. We examined the association between the frequencies of *KRAS* subtypes and smoking duration or pack-year groups using logistic regression. The frequencies of *KRAS* G12C in both groups and that of G12V in the pack-year group increased significantly with an increase in smoking dose (G12C (duration):  $p < 0.001$ , G12C (pack-years):  $p = 1.03 \times 10^{-3}$ , G12V (pack-years):  $p = 0.017$ ). There were no significant increases in the frequencies of G12A (duration, pack-years), G12D (duration, pack-years) or G12V (duration) with an increase in smoking dose. Based on the results of this subset analysis, it can be reasonable to conclude that pack-years was superior to smoking duration in predicting the frequencies of *KRAS* mutations.

The main limitation of our study was that the JME study data were obtained by targeted sequencing and not from whole genome or whole exome sequencing. We focused on cancer development and chose 72 oncogenic driver genes as the targets for mutational analysis, and C>A was examined in limited lesions. Various targeted sequencing panels have recently been developed to efficiently determine tumour mutation burden, and a strong correlation has been observed between targeted sequencing and whole genome sequencing in some studies.<sup>31,32</sup> Another limitation of our study was recall bias, which cannot be ruled out when obtaining smoking information, even in a prospective cohort study. Additionally, data were extracted only from Japanese patients. The frequencies of gene mutations are known to differ according to ethnicities.<sup>33</sup> Therefore, to confirm our results, further studies using data from a large cohort involving people of different ethnicities with detailed smoking information are needed.

## CONCLUSION

Smoking duration was not significantly different from pack-years in predicting the likelihood of smoking-related gene mutations. Given the recall bias in obtaining smoking information, smoking duration alone should be considered for further investigation as a simpler alternative to pack-years.

### Author affiliations

<sup>1</sup>Respiratory Medicine, Osaka City University Graduate School of Medicine, Osaka, Japan

<sup>2</sup>Third Department of Internal Medicine, Wakayama Medical University, Wakayama, Japan

<sup>3</sup>Clinical Oncology, Osaka City University Graduate School of Medicine, Osaka, Japan

<sup>4</sup>Laboratory of Statistics, Osaka City University Faculty of Medicine, Osaka, Japan

<sup>5</sup>Internal Medicine, National Hospital Organization Kinki-chuo Chest Medical Center, Sakai, Japan

<sup>6</sup>Advanced Medicine and Clinical Research, Nagoya University Hospital, Nagoya, Japan

<sup>7</sup>Division of Respiratory Medicine and Allergology, Department of Internal Medicine, Aichi Medical University Graduate School of Medicine, Nagakute, Japan

<sup>8</sup>Clinical Research Center, National Hospital Organization Kinki-chuo Chest Medical Center, Sakai, Japan

<sup>9</sup>Respiratory Medicine, Nagoya Medical Center, Nagoya, Japan

<sup>10</sup>Surgery, National Hospital Organization Kinki-chuo Chest Medical Center, Sakai, Japan

**Acknowledgements** We would like to thank all the participants and their advisors who were involved in this study.

**Contributors** KO, TK and YK contributed to the idea and the design of this study. KO led the data analysis with statistical advice from MA and MF. KO produced the first draft of the paper. KO, YK, HK, MI, YM, KS, MF, YT, NY, AT, MA, AK, Sil, HS, AM and TK contributed to and approved the final manuscript.

**Funding** The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** YT has received personal fees from Chugai Pharmaceutical, Boehringer Ingelheim, MSD, AstraZeneca, Mundipharma and Taiho Pharmaceutical and has received grants and personal fees from Bristol-Myers Squibb and Ono Pharmaceutical, outside of the submitted work. AT has received grants from Ono Pharmaceutical, Bristol-Myers Squibb and AstraZeneca and has received personal fees from Eli Lilly, Ono Pharmaceutical, Chugai Pharmaceutical, Boehringer Ingelheim, AstraZeneca, Bristol-Myers Squibb, MSD, Taiho, Pfizer and Kisse outside of the submitted work. AK has received grants from the Ministry of Health, Labor and Welfare, Japan; has received grants and personal fees from Boehringer Ingelheim, Lilly, Chugai, and Ono; and has received personal fees from Novartis, Taiho and Pfizer outside of the submitted work. HS has received grants and personal fees from AstraZeneca, MSD, Ono Pharmaceutical, Eli Lilly Japan, Chugai Pharmaceutical, Olympus, Boehringer Ingelheim Japan, Novartis Pharma, Pfizer, Parexel International and Boston Scientific; has received grants from Bristol-Myers Squibb, Quintiles Transnational Japan, Beyer Pharmaceuticals, West Japan Oncology Group, AC Medical, Japan Blood Products Organization, CMIC HOLDINGS, Takeda, A2 Healthcare, Otsuka, Harada, Nobelpharma and Taisho Toyama Pham; and has received personal fees from Covidien Japan, Taiho Pharmaceutical, Becton, Dickinson and Company, AMCO, Kyowa Hakko Kirina and Kyorin Pharmaceutical outside of the submitted work. All other authors declare no potential conflicts of interest.

**Patient consent for publication** Not required.

**Ethics approval** The study was approved by the central institutional review boards and ethics committees, as well as by all the participating institutions. All patients provided written informed consent before enrolment.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available upon reasonable request. We carried out this study using data from the JME study; the protocols and statistical analysis plans can be found in the publication describing the JME study. The data are present in a repository managed by core members of the JME. A request for data use can be made to the corresponding author.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

### ORCID iD

Koichi Ogawa <http://orcid.org/0000-0002-0722-9073>

## REFERENCES

- Global Burden of Disease Cancer Collaboration, Fitzmaurice C, Akinyemiju TF, *et al*. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the global burden of disease study. *JAMA Oncol* 2018;4:1553–68.
- National Lung Screening Trial Research Team, Aberle DR, Adams AM, *et al*. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011;365:395–409.
- Lichtenstein P, Holm NV, Verkasalo PK, *et al*. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* 2000;343:78–85.
- Wakai K, Inoue M, Mizoue T, *et al*. Tobacco smoking and lung cancer risk: an evaluation based on a systematic review of epidemiological evidence among the Japanese population. *Jpn J Clin Oncol* 2006;36:309–24.
- Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers—a different disease. *Nat Rev Cancer* 2007;7:778–90.
- Buttitta F, Barassi F, Fresu G, *et al*. Mutational analysis of the HER2 gene in lung tumors from Caucasian patients: mutations are mainly present in adenocarcinomas with bronchioloalveolar features. *Int J Cancer* 2006;119:2586–91.
- Suzuki M, Shigematsu H, Iizasa T, *et al*. Exclusive mutation in epidermal growth factor receptor gene, HER-2, and KRAS, and synchronous methylation of nonsmall cell lung cancer. *Cancer* 2006;106:2200–7.
- Shigematsu H, Lin L, Takahashi T, *et al*. Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers. *J Natl Cancer Inst* 2005;97:339–46.
- Chapman AM, Sun KY, Ruestow P, *et al*. Lung cancer mutation profile of EGFR, ALK, and KRAS: meta-analysis and comparison of never and ever smokers. *Lung Cancer* 2016;102:122–34.
- Saldias Peñafiel F, Elola Aránguiz JM, Uribe Monasterio J, *et al*. [Risk factors for the development of lung cancer in a cohort of adult smokers]. *Rev Med Chil* 2016;144:1382–90.
- Cigarette smoking and health. American Thoracic Society. *Am J Respir Crit Care Med* 1996;153:861–5.
- Bhatt SP, Kim Y-I, Harrington KF, *et al*. Smoking duration alone provides stronger risk estimates of chronic obstructive pulmonary disease than pack-years. *Thorax* 2018;73:414–21.
- Alexandrov LB, Ju YS, Haase K, *et al*. Mutational signatures associated with tobacco smoking in human cancer. *Science* 2016;354:618–22.
- Kawaguchi T, Koh Y, Ando M, *et al*. Prospective analysis of oncogenic driver mutations and environmental factors: Japan molecular epidemiology for lung cancer study. *J Clin Oncol* 2016;34:2247–57.
- SWOG. View protocol abstract: S0424. Available: <http://www.swog.org/Visitors/ViewProtocolDetails.asp?ProtocolID=2000> [Accessed 30 Apr 2013].
- Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012;489:519–25.
- Ding L, Getz G, Wheeler DA, *et al*. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 2008;455:1069–75.
- Imielinski M, Berger AH, Hammerman PS, *et al*. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* 2012;150:1107–20.
- Govindan R, Ding L, Griffith M, *et al*. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* 2012;150:1121–34.
- Dogan S, Shen R, Ang DC, *et al*. Molecular epidemiology of EGFR and KRAS mutations in 3,026 lung adenocarcinomas: higher susceptibility of women to smoking-related KRAS-mutant cancers. *Clin Cancer Res* 2012;18:6169–77.



- 21 Flanders WD, Lally CA, Zhu B-P, *et al.* Lung cancer mortality in relation to age, duration of smoking, and daily cigarette consumption: results from cancer prevention study II. *Cancer Res* 2003;63:6556–62.
- 22 Lubin JH, Caporaso NE. Cigarette smoking and lung cancer: modeling total exposure and intensity. *Cancer Epidemiol Biomarkers Prev* 2006;15:517–23.
- 23 Remen T, Pintos J, Abrahamowicz M, *et al.* Risk of lung cancer in relation to various metrics of smoking history: a case-control study in Montreal. *BMC Cancer* 2018;18:1275.
- 24 Etter JF, Perneger TV. Measurement of self reported active exposure to cigarette smoke. *J Epidemiol Community Health* 2001;55:674–80.
- 25 Moyer VA, U.S. Preventive Services Task Force. Screening for lung cancer: U.S. preventive services task force recommendation statement. *Ann Intern Med* 2014;160:330–338–8.
- 26 Boiselle PM. Computed tomography screening for lung cancer. *JAMA* 2013;309:1163–70.
- 27 Gould MK. Clinical practice. lung-cancer screening with low-dose computed tomography. *N Engl J Med* 2014;371:1813–20.
- 28 Modin HE, Fathi JT, Gilbert CR, *et al.* Pack-year cigarette smoking history for determination of lung cancer screening eligibility. Comparison of the electronic medical record versus a shared decision-making conversation. *Ann Am Thorac Soc* 2017;14:1320–5.
- 29 Temko D, Tomlinson IPM, Severini S, *et al.* The effects of mutational processes and selection on driver mutations across cancer types. *Nat Commun* 2018;9:9.
- 30 Riely GJ, Kris MG, Rosenbaum D, *et al.* Frequency and distinctive spectrum of KRAS mutations in never smokers with lung adenocarcinoma. *Clin Cancer Res* 2008;14:5731–4.
- 31 Garofalo A, Sholl L, Reardon B, *et al.* The impact of tumor profiling approaches and genomic data strategies for cancer precision medicine. *Genome Med* 2016;8:79.
- 32 Campesato LF, Barroso-Sousa R, Jimenez L, *et al.* Comprehensive cancer-gene panels can be used to estimate mutational load and predict clinical benefit to PD-1 blockade in clinical practice. *Oncotarget* 2015;6:34221–7.
- 33 Tomoya K, Hirata K, Philip CM. *Molecular epidemiology (Asian vs Caucasian). IASLC 18th world conference on lung cancer (October 15-18), 2017.*