Carcinogenesis, 2019, 1–11

doi:10.1093/carcin/bgz036
Advance Access publication February 19, 2019
Original Article

ORIGINAL ARTICLE

Risk of lung adenocarcinoma from smoking and radiation arises in distinct molecular pathways

Noemi Castelletti[1,*], Jan Christian Kaiser[1], Cristoforo Simonetto[1], Kyoji Furukawa[2], Helmut Küchenhoff[3] and Georgios T. Stathopoulos[4,5]

[1]Institute of Radiation Medicine (IRM), Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Neuherberg, Bavaria, Germany, [2]Biostatistics Center, Kurume University, 67 Asahi-machi, Kurume, Japan, [3]Department of Statistics, Ludwig-Maximilian University (LMU) Munich; Akademiestraße 1, 80799 Munich, Bavaria, Germany, [4]Laboratory for Molecular Respiratory Carcinogenesis, Department of Physiology, Faculty of Medicine; University of Patras; 1 Asklepiou Str., 26504 Rio, Achaia, Greece and [5]Comprehensive Pneumology Center (CPC) and Institute for Lung Biology and Disease (iLBD), University Hospital, Ludwig-Maximilian University (LMU) and Helmholtz Zentrum München, Member of the German Center for Lung Research (DZL), Max-Lebsche-Platz 31, 81377 Munich, Bavaria, Germany

*To whom correspondence should be addressed. Tel: +49 089 31872133; Fax: +49 089 31873363; Email: noemi.castelletti@helmholtz-muenchen.de
Correspondence may also be addressed to Jan Christian Kaiser, Tel: +49 089 31874028; Fax: +49 089 31873363; Email: christian.kaiser@helmholtz-muenchen.de

Abstract

KRAS mutations of lung adenocarcinoma (LADC) are associated with smoking but little is known on other exposure-oncogene associations. Hypothesizing that different inciting agents may cause different driver mutations, we aimed to identify distinct molecular pathways to LADC, applying two entirely different approaches. First, we examined clinicopathologic features and genomic signatures of environmental exposures in the large LADC Campbell data set. Second, we designed a molecular mechanistic risk model of LADC (M_3^{LADC}) that links environmental exposure to incidence risk by mathematically emulating the disease process. This model was applied to incidence data of Japanese atom-bomb survivors which contains information on radiation and smoking exposure. Grouping the clinical data by driver mutations revealed two main distinct molecular pathways to LADC: one unique to transmembrane receptor-mutant patients that displayed robust signatures of radiation exposure and one shared between submembrane transducer-mutant patients and patients with no evident driver mutation that carried the signature of smoking. Consistently, best fit of the incidence data was achieved with a M_3^{LADC} with two pathways: in one LADC risk increased with radiation exposure and in the other with cigarette consumption. We conclude there are two main molecular pathways to LADC associated with different environmental exposures. Future molecular measurements in lung cancer tissue of atom-bomb survivors may allow to further test quantitatively the M_3^{LADC}-predicted link of radiation to transmembrane receptor mutations. Moreover, the developed molecular mechanistic model showed that for low doses, as relevant e.g. for medical imaging, smokers have the same radiation risk compared with never smokers.

Introduction

Lung adenocarcinoma (LADC) is the number one cancer killer worldwide (1,2). LADC is mainly, but not exclusively, caused by tobacco smoke but also occurs in never smokers possibly due to both anthropogenic and environmental radiation exposures (3–5). The comprehensive genomic characterization of LADC from Caucasian and Asian patients has identified mutations in major driver oncogenes such as *KRAS*, epidermal growth factor receptor (*EGFR*) and others, with different frequencies observed in different populations (6,7). However, a biological concept explaining the relative contributions of cigarette smoking and radiation exposures to LADC incidence is still missing.

Received: December 21, 2018; **Revised:** January 30, 2019; **Accepted:** February 18, 2019

Abbreviations

3SCE	Three-Stage Clonal Expansion
AIC	Akaike's information criterion
CI	confidence interval
CNA	copy number alteration
EAR	excess absolute rate
EGFR	epidermal growth factor receptor
ERR	excess relative risk
γ-IR	ionizing γ-radiation
MMR	mismatch repair
LADC	lung adenocarcinoma
LSS	Life Span Study
SNV	single nucleotide variant
TSCE	Two-Stage Clonal Expansion

The Life Span Study (LSS) of Japanese atomic bomb survivors is arguably the most important cohort to investigate carcinogenic effects of radiation. Since exposure occurred to the general population, risk estimates from the LSS are applied to assess health effects from environmental and clinical radiation exposure in other populations around the world (8). Recently, a series of studies reported LSS risk estimates for lung cancer (9,10) and its histological types (11) from concomitant exposure to smoking and radiation. These epidomiological studies claim positive, more than multiplicative interaction between smoking and radiation. However, state-of-the-art epidemiological risk models merely establish statistical associations without explicitly considering pathogenic processes or molecular data. Importantly, different molecular pathways with specific age-risk patterns in observational data can lead to the same cancer classification in a given organ. This has been observed for colon cancer, which appears in two main molecular variants with differential growth dynamics (12). Radiation-induced papillary thyroid cancer exhibits a pertinent molecular footprint distinguished from the sporadic variant (13,14). However, when multiple endpoints are analyzed jointly in epidemiology with a state-of-the-art relative risk model, interactions or patterns of risk modifications may arise that are not substantiated by biology and can differ from cohort to cohort.

For lung carcinogenesis molecular biology and epidemiology still lack a common interface. Here, we bridge this gap by applying molecular mechanistic models (M_3) of carcinogenesis as tools to harness molecular data of LADC. M_3 treat carcinogenesis as a progression of cell-based key events on the pathway to malignancy and can detect in cancer incidence imprints from molecular events on recorded hazard or survival rates (15). When developing M_3 for the LSS, we faced the problem that molecular data from lung cancer tissue of LSS patients has not yet been generated. Therefore, we analyze molecular profiles from patients of the USA and China (6,7) to develop a biological concept which guides the design of M_3^{LADC}. The model design is checked in the LSS by means of goodness-of-fit and biological plausibility. It turned out that biological and model-identified pathways can be matched. We provide a comprehensive risk assessment for the LSS which relies on two non-interacting legs of M_3^{LADC} driven by either smoking or radiation. Should genomic data from the LSS become available, M_3^{LADC} is open to direct validation.

Materials and methods

Statistical analysis of molecular data

Mutation rates of 660 patients with LADC from the USA (6) and 101 from China (7) were extracted from the primary publications. For individual patients, clinical, exposure and mutation data from the US cohort were downloaded from the primary publication (6) and manually analyzed. Clinicopathologic and molecular data from (6) and (7) were examined for normality by Kolmogorov–Smirnov test, were found to be not normally distributed and are hence presented as median with Tukey's whiskers (boxes: interquartile range; bars: 50% extreme quartiles) and raw data points (dots). Differences in frequencies were examined by Fisher's exact or χ^2 tests and in medians of non-normally distributed variables by Kruskal–Wallis non-parametric analyses of variance with Dunn's post-tests. Survival was examined by Kaplan–Meier estimates with log-rank tests. Probability (P) is two-tailed and $P < 0.05$ was considered significant. Statistics and plots from clinico-pathologic and molecular data were done on Prism v5.0 (GraphPad, La Jolla, CA). Univariate multinomial regression analysis of clinico-pathologic and molecular data from (6) stratified by molecular pathways was done with R* (https://r-project.org/).

The LSS cohort of Japanese atomic bomb survivors

The LSS cohort has been the primary epidemiological basis for evaluating the long-term health effects of radiation, dominated by 0–4 Gy gamma rays of low linear energy transfer. It includes about 94 000 survivors who were in Hiroshima and Nagasaki at the time of bombing and about 27 000 who were temporarily away at that time and whose mortality and cancer incidence have been followed up since 1950 and 1958, respectively (16). As information on smoking is not available for all cohort members, missing data has been imputed (17) and analysis was repeated on 50 different imputed data sets (see Supplementary Material, available at *Carcinogenesis* Online). To put the results of mechanistic modeling into perspective, a descriptive risk model (11) has been applied (see Supplementary Material, available at *Carcinogenesis* Online). Supplementary Table S1, available at *Carcinogenesis* Online, summarizes the LSS cohort data broken down by sex and smoking status.

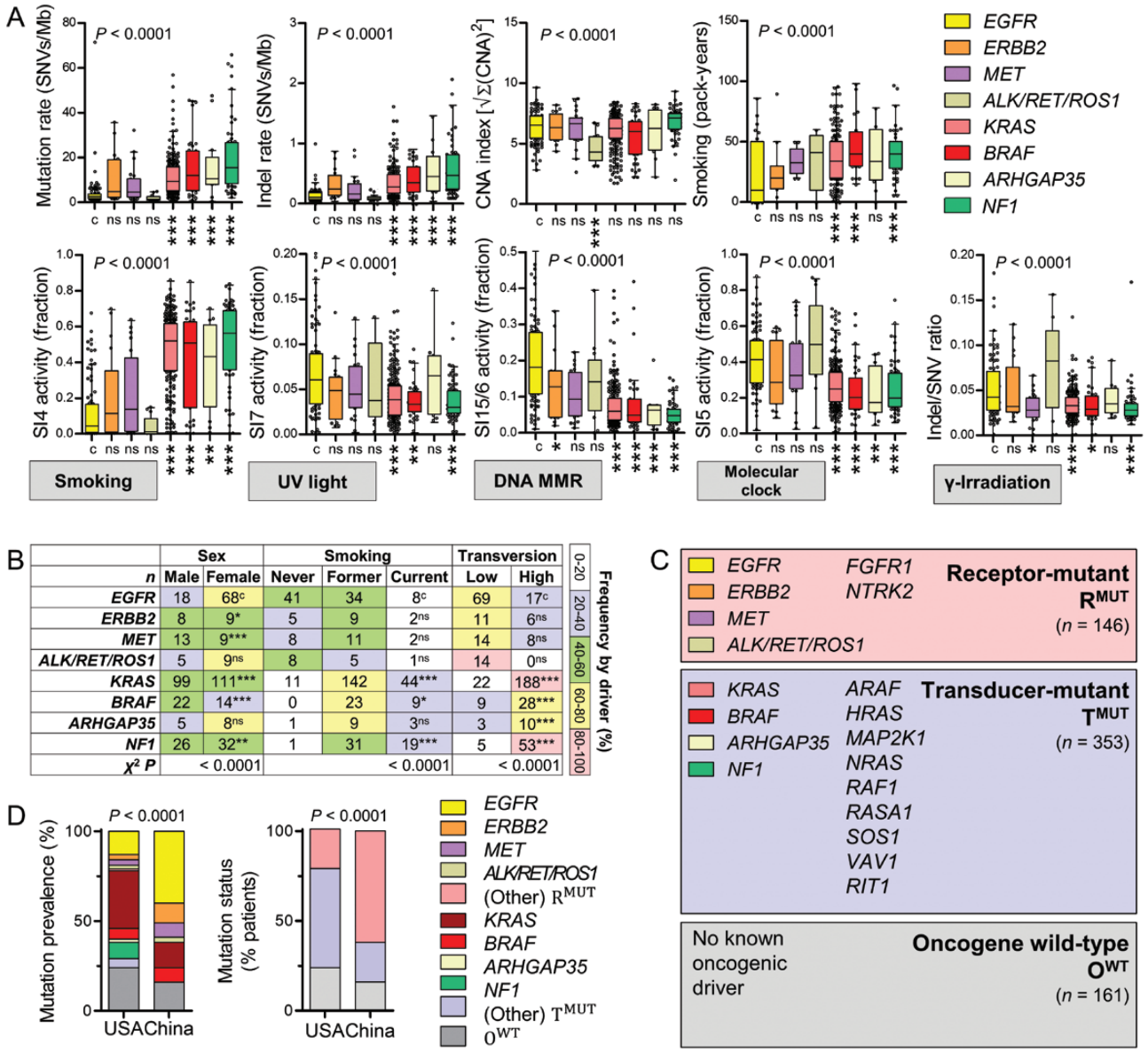
Mechanistic risk modeling

Mechanistic models have long been applied for the analysis of radio-epidemiological cohorts (15). For the present study, analysis of the LSS cohort was performed independently from the genomic results. But the model concept is motivated by the suggestion of two molecular pathways to LADC.

In a pathway-specific model, cancer develops in the lung epithelium from a large number of N healthy cells in homeostasis. Eventually, a very small fraction of healthy cells acquire initial mutations with rate ν. Initiated cells may either grow immediately (in the Two-Stage Clonal Expansion model, TSCE) into atypical adenomatous hyperplasia as precursor lesions in invasive LADC with net rate $\gamma = \alpha\text{-}\beta\text{-}\mu$ or only after acquiring a second mutation (Three-Stage Clonal Expansion model, 3SCE). The stem cell division rate α is reduced by a rate β of cell inactivation (i.e. apoptosis) and a transformation rate μ. In the transformation stage after clonal expansion, initiated cells suffer a 'final rare event' often as a mutation in a tumor suppressor gene, which turns them into cancer cells. During a lag time of several years cancer cells grow into a clinically relevant tumor (18). However, in the present LSS data set, inclusion of a lag time had only marginal impact on the results. Therefore, all complex processes after clonal expansion are summarized by a single late event with rate μ. An illustration of such a model but with two pathways can be found in Figure 3A.

For mathematical implementation of the TSCE model, mutation rates and rates of cell division or inactivation are treated as transient Poisson point processes of cell birth and death, which are couched in a set of master equations (19). This set has been transformed into a system of coupled ordinary differential equation of the Ricatti type, which is solved efficiently by a set of backward recursion relations given in Supplementary Table S5, available at *Carcinogenesis* Online. Meza *et al.* (20) present closed analytical forms for the hazard functions of the TSCE model and the 3SCE model. Their mapping of biological stages in cancer development to pertinent stages of age-dependent incidence is very instructive. For the TSCE, the recursion relations are presented in Supplementary Table S5, available at *Carcinogenesis* Online.

In Figure 3A, biological transition rates ν, μ, α and β are shown for each pathway. However, the recursion relations for the TSCE model are expressed in terms of the parameter combinations $X = N\nu\mu$, $\mu = \alpha\text{-}\beta\text{-}\mu$ and $\delta = \alpha\mu$, which possess the advantage of being identifiable from the



	Sex		Smoking			Transversion	
n	Male	Female	Never	Former	Current	Low	High
EGFR	18	68^{c}	41	34	8^{c}	69	17^{c}
ERBB2	8	9*	5	9	2^{ns}	11	6^{ns}
MET	13	9***	8	11	2^{ns}	14	8^{ns}
ALK/RET/ROS1	5	9^{ns}	8	5	1^{ns}	14	0^{ns}
KRAS	99	111***	11	142	44***	22	188***
BRAF	22	14***	0	23	9*	9	28***
ARHGAP35	5	8^{ns}	1	9	3^{ns}	3	10***
NF1	26	32**	1	31	19***	5	53***
χ² P		< 0.0001			< 0.0001		< 0.0001

Frequency by driver (%): 0-20, 20-40, 40-60, 60-80, 80-100

Figure 1. Identification of broad molecular pathways to lung adenocarcinoma. SNV rates, indel rates, CNA indices, smoking exposure, sex, genomic signatures of environmental carcinogen-induced base changes in the trinucleotide context (SI), indel/SNV ratios and transversion status of 660 patients with LADC from the USA (6) grouped by the most frequent driver mutations. **(A)** Data are given as raw data points, median ± Tukey's whiskers (lines: median; boxes: interquartile range; bars: 50% extreme quartiles). *P*, probabilities by Kruskal–Wallis test. Significances for comparison with *EGFR*-mutant control group (c) by Dunn's post-tests. **(B)** Data are given as number of patients (*n*). Color scale indicated frequency per row. *P*, probabilities by χ^2 test. Significances for comparison with *EGFR*-mutant control group (c) by χ^2 or Fischer's exact tests. Sample sizes were *EGFR* ($n = 86$), *ERBB2* ($n = 17$), *MET* ($n = 22$), *ALK/RET/ROS1* (pooled $n = 14$), *KRAS* ($n = 210$), *BRAF* ($n = 37$), *ARHGAP35* ($n = 13$) and *NF1* ($n = 58$). **(C)** Proposed grouping of US LADC patients (6) according to driver mutation into R^{MUT}, T^{MUT} and O^{WT} molecular pathways. **(D)** Mutation rates and molecular pathway classification of 660 US LADC patients (6) and 101 LADC patients from China (7). *P*, probability by χ^2 test. Significances $P \geq 0.05$, $P < 0.05$, $P < 0.01$ and $P < 0.001$ are coded as ns, *, ** and ***, respectively.

incidence data. The identifiability problem follows from the mathematical model structure and cannot be removed by increasing statistical power (12,21). The 3SCE model applies four mathematically identifiable parameters (20). The additional parameter accounts for fluctuations in the first mutational stage at young age but typically is not identifiable in practice due to low case numbers. As a consequence, it can be shown that applying $X = N\nu^2\mu \cdot age$ instead of $X = N\nu\mu$ in the recursion formulae for the TSCE model, the 3SCE model is approximated with very high accuracy. Exact recursion formulas for the 3SCE model can be used to derive the hazard functions but contain additional non-identifiable parameters (22).

One- and two-path models were tested with TSCE or 3SCE in either pathway (Supplementary Table S2, available at *Carcinogenesis* Online). Exposure to smoking and radiation are assumed to change the biological parameters in mechanistic risk models. We tested actions on the parameter X of initiating mutations and the net clonal expansion rate γ using several functional forms but did not test for effects on δ since δ affects the hazard incidence curve only at high ages. For radiation action, we chose linear, linear-quadratic and linear-exponential responses, which caused either acute or permanent parameter changes. For smoking action, we applied the same functional forms for the response to smoking intensity. The model parameters were increased at the beginning of smoking and remained elevated for current smokers until end of follow-up. Baseline values were retained when smokers ceased cigarette consumption because the implementation of residual effects in model parameters after smoking cessation did not significantly improve the fits. Here, we report only the main effects with optimal functional forms implemented

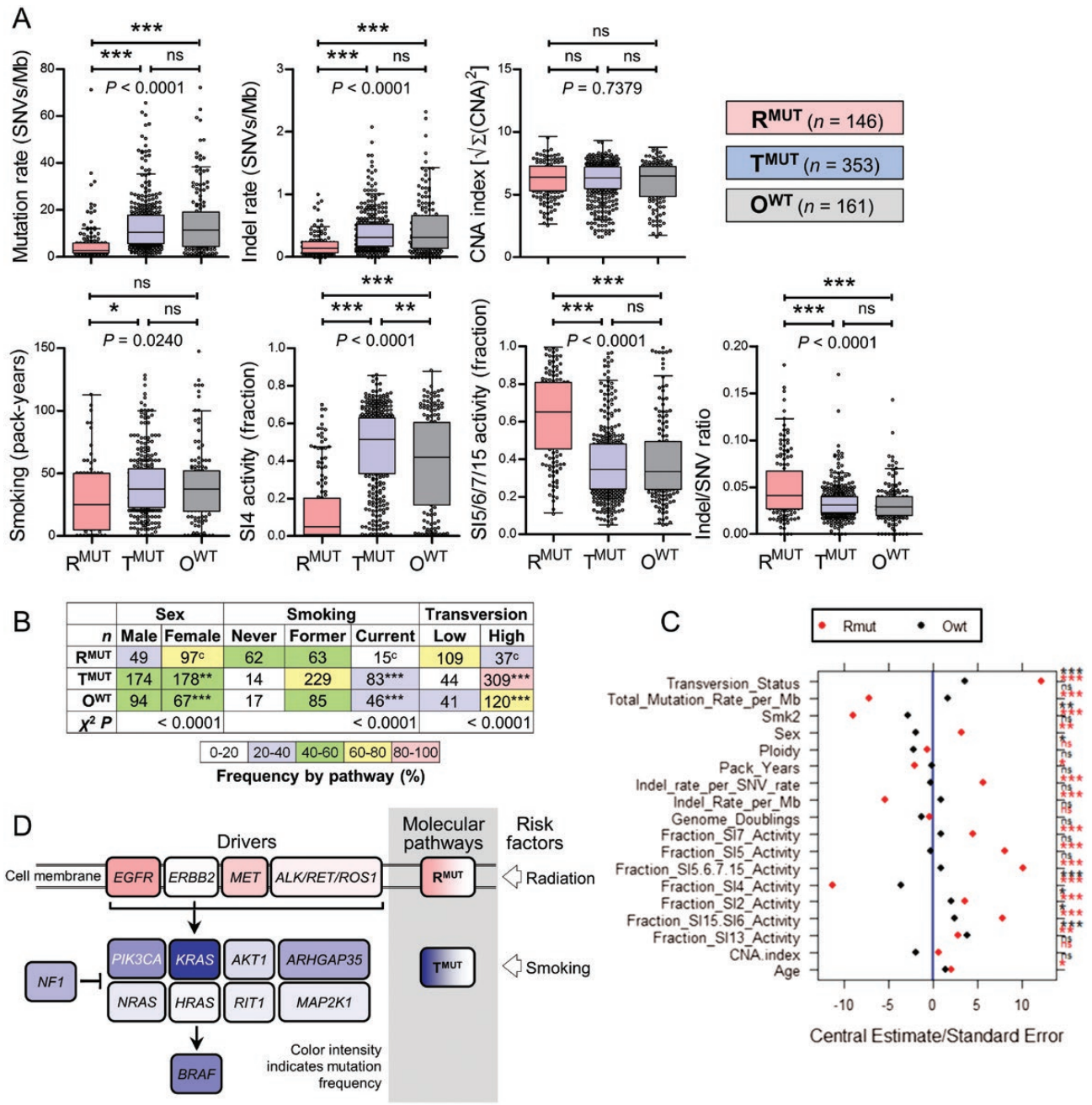


Figure 2. Clinical and molecular characteristics of 660 US LADC patients stratified by molecular pathway. SNV rates, indel rates, CNA indices, smoking exposure, sex, genomic signatures of environmental carcinogen-induced base changes in the trinucleotide context (SI), indel/SNV ratios and transversion status of 660 patients with LADC from the USA (6) grouped by R^{MUT}, T^{MUT} and O^{WT} molecular pathways. **(A)** Data are given as raw data points, median ± Tukey's whiskers (lines: median; boxes: interquartile range; bars: 50% extreme quartiles). *P*, probabilities by Kruskal–Wallis test. Significances are given for the indicated comparisons by Dunn's post-tests. **(B)** Data are given as number of patients (*n*). Color scale indicated frequency per row. *P*, probabilities by χ^2 test. Significances are given for the indicated comparisons by χ^2 or Fischer's exact tests. Sample sizes were *EGFR* (n = 86), *ERBB2* (n = 17), *MET* (n = 22), *ALK/RET/ROS1* (pooled n = 14), *KRAS* (n = 210), *BRAF* (n = 37), *ARHGAP35* (n = 13) and *NF1* (n = 58). **(C)** Points represent regression coefficients divided by their standard errors in univariate multinomial regression. Eighteen clinical and molecular variables of 660 US patients with LADC (6) stratified by molecular pathway were analyzed. Position on x-axis denotes deviation from the estimate in reference group T^{MUT}. Significance of deviation from the reference is color coded (red: R^{MUT}; black: O^{WT}). **(D)** Schematic of the two proposed molecular pathways to LADC and the main risk factors for each pathway. Significances $P \geq 0.05$, $P < 0.05$, $P < 0.01$ and $P < 0.001$ are coded as ns, *, ** and ***, respectively.

in the preferred model. In one pathway, smoking intensity *smkint* linearly enhances the clonal expansion rate

$$\gamma_T = \alpha_T - \beta_T(S) - \mu_T = \gamma_{T0}[1 + g_S \cdot smkint \cdot e^{-\kappa \cdot smkint}]$$

during a period of constant smoking intensity with an attenuated effect for high smoking intensity. In the other pathway, a radiation dose D linearly enhances the clonal expansion rate

$$\gamma_R = \alpha_R - \beta_R(D) - \mu_R = \gamma_{R0}[1 + g_R \cdot D]$$

after exposure for life. Adjustment for differences in the city of residence (Hiroshima or Nagasaki) and drifts in the birth cohort were performed for the mechanistic models with the same factors as in the descriptive model (see Supplementary Material, available at *Carcinogenesis* Online).

For each imputed LSS data set, model parameters were determined by Poisson regression. To limit the work load in the selection phase, candidate models were adjusted to only 10 out of 50 imputed LSS data sets. Model selection was based on goodness-of-fit measured by the cumulative Akaike's information criterion (AIC) = deviance +2 · no. of parameters for 10 data sets. Confidence intervals (CIs) were calculated by extending Rubin's rule (23). The workflow CI calculation is depicted in Supplementary Figure S1, available at *Carcinogenesis* Online.

Results

Identification of two causally and molecularly distinct pathways using molecular data

To identify possible clinical and/or molecular clusters of patients with LADC, we initially analyzed all data available from 660 Caucasian patients with LADC classified by oncogenes (6). In addition to the available clinical information, total single nucleotide variant (SNV) rates, insertion/deletion (indel) rates, copy number alteration (CNA) indices (calculated as the square root of the sum of all CNA squares of each tumor) as well as the contribution of established genomic signatures of environmental exposures were examined. These included a UV-related signature of C>T at TpCpC or CpCpC (COSMIC Signature 7, abbreviated SI7), a smoking-related signature of C>A transversions (SI4), a DNA mismatch repair (MMR) signature of C>T at GpCpG (SI15/SI6), two APOBEC-related signatures of C>G or C>T at TpCpT or TpCpA (SI13 and SI2) and a COSMIC signature 5 (SI5) with putative 'molecular clock' properties (6,24). In addition, we calculated the indel/SNV ratios, since high ratios were found elsewhere to represent a direct molecular imprint of γ-ionizing radiation (γ-IR) from molecular exposure (25).

Grouping of the 660 patients by the most frequent drivers (every driver with $n \geq 10$ patients available was examined) revealed that patients with *EGFR* ($n = 86$), *ERBB2* ($n = 17$), *MET* ($n = 22$) and *ALK/RET/ROS1* (pooled to achieve $n = 14$) mutations [hereafter collectively referred to as receptor-mutant (R^{MUT})] were different from patients with *KRAS* ($n = 210$), *BRAF* ($n = 37$), *ARHGAP35* ($n = 13$) and *NF1* ($n = 58$) mutations [hereafter collectively referred to as transducer-mutant (T^{MUT})]. To this end, R^{MUT} patients displayed lower SNV and indel rates and decreased smoking exposure evident by lower transversion rates and decreased activity of the smoking-related SI4 compared with T^{MUT} patients. At the same time, R^{MUT} patients were more frequently female and displayed increased activities of UV light-related SI7, of DNA MMR-related SI15/SI6 and of SI5 putatively reflecting molecular clock properties compared with T^{MUT} patients. Interestingly, R^{MUT} patients had higher indel/SNV ratios compared with T^{MUT} patients, indicating a molecular signature of γ-IR exposure (25). CNA indices were comparable across patients with different drivers, except from *ALK/RET/ROS1*-fused patients that collectively displayed lower CNA indices compared with all other patients (Figure 1A and B). Based on this finding, we grouped US patients (6) and 101 LADC obtained from Asian patients (7) into R^{MUT}, T^{MUT} and oncogene wild-type (O^{WT}; patient without R^{MUT} or T^{MUT}) groups, hypothesizing that these three groupings may represent distinct molecular pathways to LADC (Figure 1C).

Individual mutation prevalence varied widely between (6) and (7), translating into different frequencies of these pathways in Caucasian and Asian LADC (Figure 1D), a fact that has to be taken into account since the molecular analysis was done with American patients and the model analysis with a Japanese cohort. We next sought to compare the molecular profiles of the three candidate molecular pathways LADC to identify potential similarities and differences.

Interestingly, R^{MUT} LADC appeared distinct, whereas T^{MUT} and O^{WT} LADC were similar by all parameters examined, except SI4 activity (Figure 2A and B). This was also evident from univariate multinomial logistic regression analyses that showed a general pattern of O^{WT} LADC trending with T^{MUT} LADC (Figure 2C). In the case of R^{MUT} LADC, 13 of the 18 analyzed covariables trended different from the reference category T^{MUT} with high significance (Figure 2C). These findings indicated the existence of two distinct molecular pathways to LADC that bear different genomic marks of environmental exposures: one unique to R^{MUT} patients that features robust imprints of γ-IR and the associated DNA MMR (26) and one shared between T^{MUT} and O^{WT} patients (hereafter referred to as T^{MUT}) with genomic marks of smoking exposure (Figure 2D). Interestingly, the R^{MUT} pathway contained patients with ALK/RET/ROS1-fusions, which were recently shown to dose-dependently culminate from γ-IR in thyroid cancer (14).

Identification of two etiologically different pathways from incidence data

The preferred M_3^{LADC} was identified after multiple series of model testing which are summarized below. Models relying on only a single pathway to LADC did not provide a good description of the data with AIC values significantly higher than the preferred tw- path model M_3^{LADC} (Supplementary Table S2, available at *Carcinogenesis* Online). In single path models, interaction between smoking and radiation has been tested by multiplication of both covariables but was rejected based on goodness-of-fit. Smoking significantly increased both initiation and promotion, whereas radiation only enhanced promotion. Informed by these preliminary results, a large number of multi-stage models have been tested in pairs as candidates for preferred pathway-specific models (see Supplementary Table S2, available at *Carcinogenesis* Online). Tomasetti *et al.* (27) argue that two or three driver mutations are involved in LADC development. Thus, we considered only two- and three-stage clonal expansion models (TSCE and 3SCE, respectively) as candidates for both pathways. Whereas the TSCE model starts with one initial mutation, the 3SCE model applies two mutations in the early initiation phase.

After thinning out the set of candidate models we ended up with two almost identical two-path models in terms of goodness-of-fit (Supplementary Table S2, available at *Carcinogenesis* Online). Interestingly, for both models only one pathway turned out to be smoking dependent and the other to depend on radiation. Taking into account the results from the molecular analysis, we label the former T^{MUT} and the latter R^{MUT}. Whereas for the R^{MUT} pathway, a TSCE model yielded better fits than a 3SCE model, in the T^{MUT} pathway, an AIC-based selection of a TSCE model versus a 3SCE model was not possible for the T^{MUT} pathway. For practical reasons, the TSCE model was selected for the T^{MUT} pathway, although a 3SCE model cannot be excluded. The conceptual design of the final preferred two-path LADC model is shown in Figure 3A and the corresponding contributions of the different pathways to the hazard in Figure 3B. Supplementary Table S3, available at *Carcinogenesis* Online, lists the parameters as means over 50 data sets for both pathways. Central risk estimates were calculated with the parameters of Supplementary Table S4, available at *Carcinogenesis* Online. The preferred descriptive model formulated in Supplementary Equations (1)–(4), available at *Carcinogenesis* Online, was also fitted to all 50 imputed data sets and yielded an AIC higher by 7.4 points per data set compared with the preferred M_3^{LADC}.

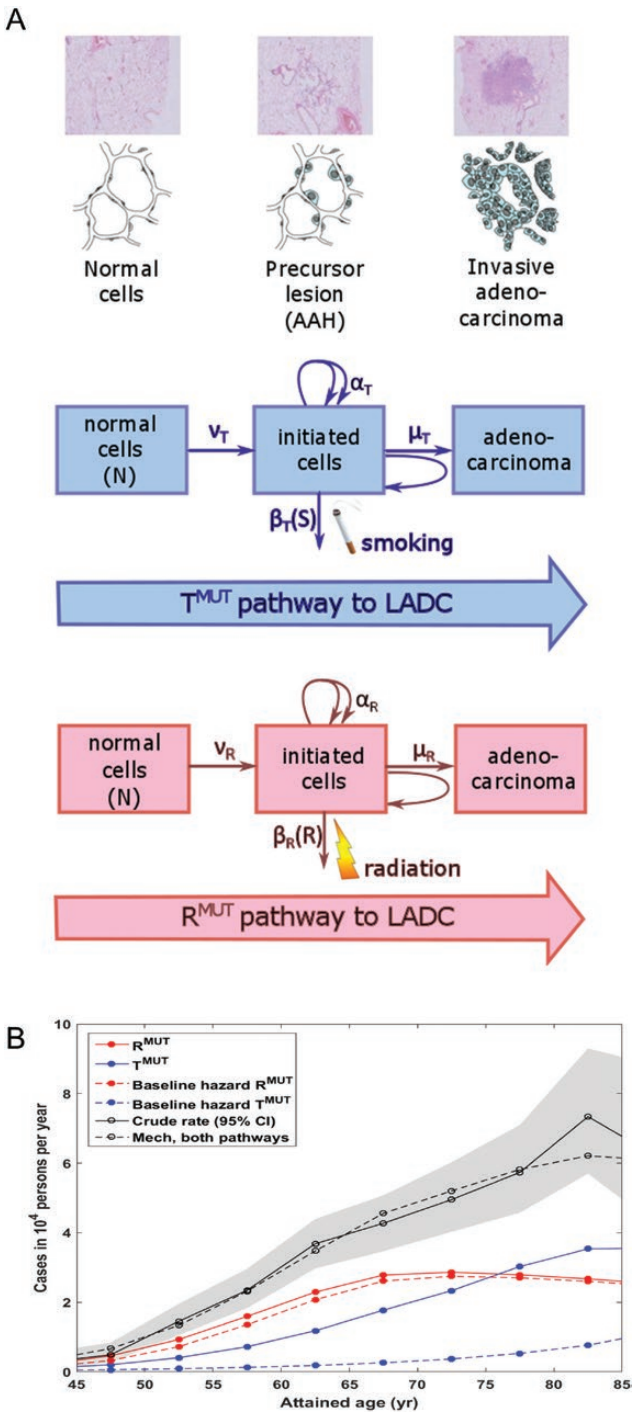


Figure 3. Molecular prediction of LADC risk stratified by molecular pathway. **(A)** Top: Histological progression from normal cells over atypical adenomatous hyperplasia as precursor lesions to invasive LADC [modified figure from Yatabe, Borczuk (40)]. Bottom: Model implementation with two distinct molecular pathways pertaining to either T^{MUT} or R^{MUT} with two versions of the TSCE model. Boxes represent cells in states with defined molecular properties. Arrows represent rates of transition between cell states. In both pathways, a tiny fraction of a large number of *N* healthy cells incurs early molecular changes with yearly rates ν. Initiated cells may either divide symmetrically with rates α or become inactivated with rates β. The final transformation stage summarizes a sequence of complex processes with effective rate μ. Both agents of smoking and radiation cause the acceleration of clonal expansion by reduced cell inactivation. See model details in Materials and Methods; mathematical model implementation and parameter estimates are given in Supplementary Tables S3–S5, available at *Carcinogenesis* Online. **(B)** Crude rate and predicted hazard (LADC cases in 10 000 persons per year) from the preferred mechanistic model M_3^{LADC} (Mech) for the LSS cohort in 5-year age groups from 40–45 up to 80–85 years. The model clearly distinguishes pathway-specific hazards. The hazard of R^{MUT}-related LADC cases peaks at age 70 years. The hazard in the T^{MUT} pathway becomes dominant at old ages. This is a M_3^{LADC} prediction for the LSS cohort without any genomic data.

For cigarette smoking, clonal expansion in the T^{MUT} pathway was identified as the main biological target. Sex-specific response curves exhibited markedly different shapes (Supplementary Figure S2, available at *Carcinogenesis* Online). For men, the growth rate increased almost linearly up to a smoking intensity of 20 cigarettes/day and flattened thereafter. Clonal growth in women reacted much stronger to low smoking intensity. The growth reduction after a peak at about 10 cigarettes/day is biologically not plausible (Supplementary Figure S2, available at *Carcinogenesis* Online). But growth models applying a continued increase or plateau were rejected due to markedly inferior goodness-of-fit.

The main radiation effect occurred in the R^{MUT} pathway. An acute radiation pulse yielded a linear permanent increase of the net clonal expansion γ_R consistent with lifelong radiation-induced inflammation caused by genetic damage. Summarizing, the main impact of smoking and radiation took effect in distinct molecular pathways without noticeable synergy. For risk assessment, this particular biological action is better reflected in the excess absolute rate (EAR) compared with the excess relative risk (ERR). Supplementary Figure S3, available at *Carcinogenesis* Online, presents a comparison of baseline hazard rates and hazard rates between the two molecular pathways. Figures 4 and 5 depict the EAR depending on smoking and radiation for pertinent exposure scenarios. In Supplementary Figures S4 and S5, available at *Carcinogenesis* Online, the additive effect from both agents on the EAR is shown. Supplementary Figures S6 and S7, available at *Carcinogenesis* Online, give the sex-specific ERR from smoking and radiation, respectively. Figure 6 presents a pathway-specific breakdown of expected LADC cases in different exposure groups for smoking and radiation.

Discussion

LADC management and outcomes largely rely on tumor genotype (28). However, current risk models of LADC do not provide molecularly stratified risks. We used molecular data from Caucasian and Asian patients with LADC to reveal two broad molecular fingerprints of the disease probably caused by different environmental exposures: one unique to patients with mutations in transmembrane receptors (R^{MUT}) featuring imprints of radiation and another shared by patients with mutations in signal transduction genes and by patients with no known oncogene mutations (T^{MUT}) displaying the molecular signature of tobacco smoking. In addition, we have developed M_3 and have fitted it to observational data of LADC incidence in Japanese atomic bomb survivors with known radiation/smoking exposure but unknown mutation status. This analysis provided independent evidence for the existence of two pathways to LADC, of which one depends on radiation and the other on smoking. Our combined genomic and epidemiologic analyses provide the first mechanistic link between radiation exposure and receptor mutations in LADC, including EGFR mutations and ALK/RET/ROS1 fusions. Importantly, the predictive power of M_3^{LADC} can be subject to rigorous validation by future measurements of the mutation status in LADC tissue of LSS patients.

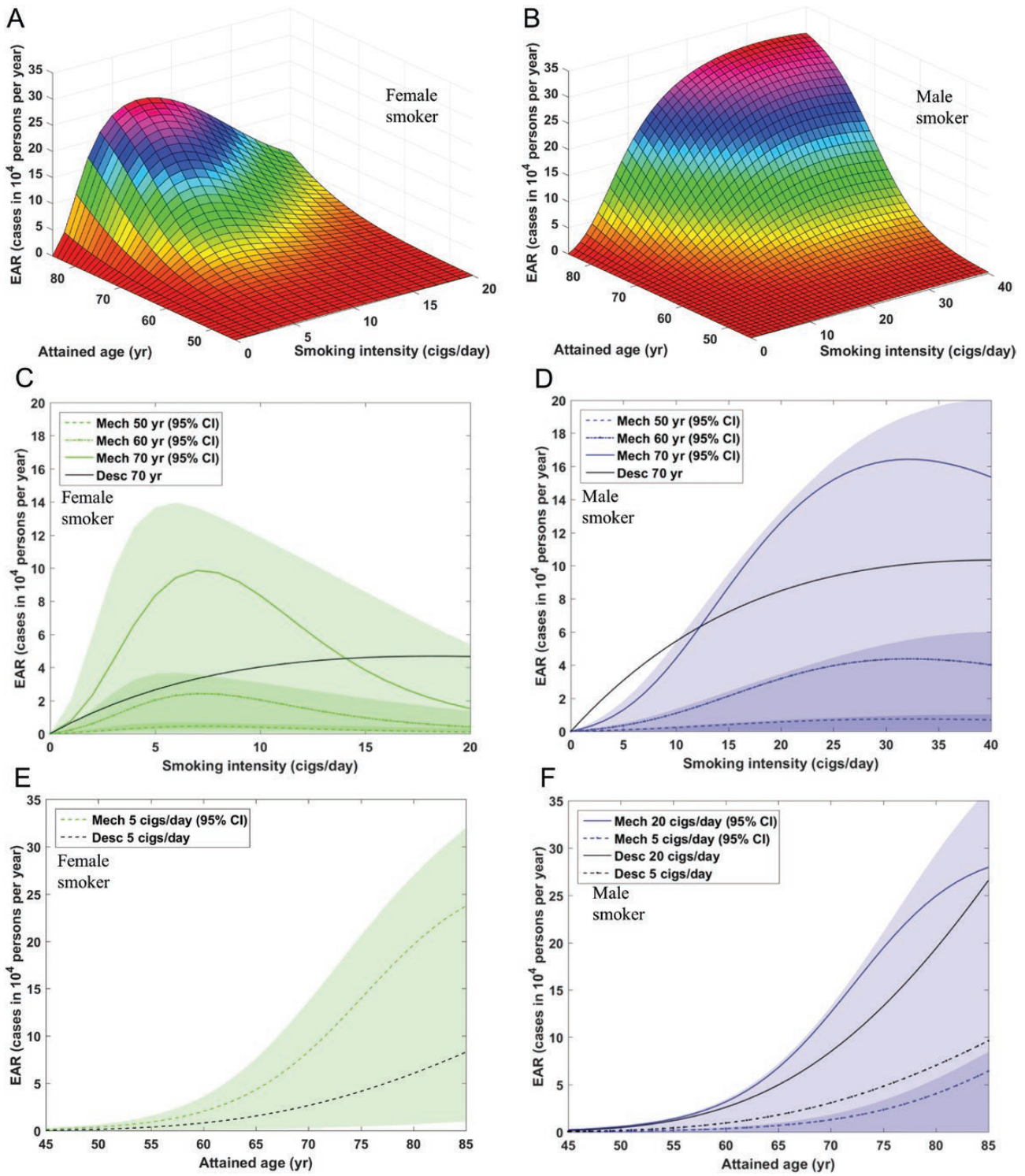


Figure 4. Excess absolute rates (EARs; as cases in 10 000 persons per year) from M_3^{LADC} (Mech) for smoking-induced lung adenocarcinoma in the Japanese LSS cohort for lifelong smokers starting at age 20 years. The EAR is determined by the sex-dependent linear-exponential response to the smoking intensity, which increases the clonal expansion rate in the T^{MUT} pathway independent of radiation (Supplementary Figure S2, top, available at *Carcinogenesis* Online). To eliminate the influence for city of residence, person-year weighted city means are used. Bivariate EAR dependence on attained age and smoking intensity for **(A)** female smokers and **(B)** male smokers. Panels **(C)** and **(D)** depict cross-sectional cuts to panels (A) and (B) for attained ages of 50, 60 and 70 years. Panels **(E)** and **(F)** depict cross-sectional cuts to panels (A) and (B) for 5 cigarettes/day (males and females) and 20 cigarettes/day (males only). Female smokers of 5 cigarettes/day and male smokers of 20 cigarettes/day possess about the same risk. The EAR from a descriptive risk model (Desc) is shown for comparison.

Midha *et al.* (29) provide a comprehensive summary of EGFR prevalence in LADC by ethnicity. In their Table 1, they report for China a prevalence of 48% (range 27–66% in 18 studies). In Japan, a prevalence of 45% (range 21–68% in 33 studies) was found. These shares are notably different from those in the USA of 23% (range 3–42% in 16 studies). The numbers for China comply with about 40% EGFR mutations found by Wu *et al.* (7), which we used to calculate R^{MUT} versus T^{MUT} shares for an Asian

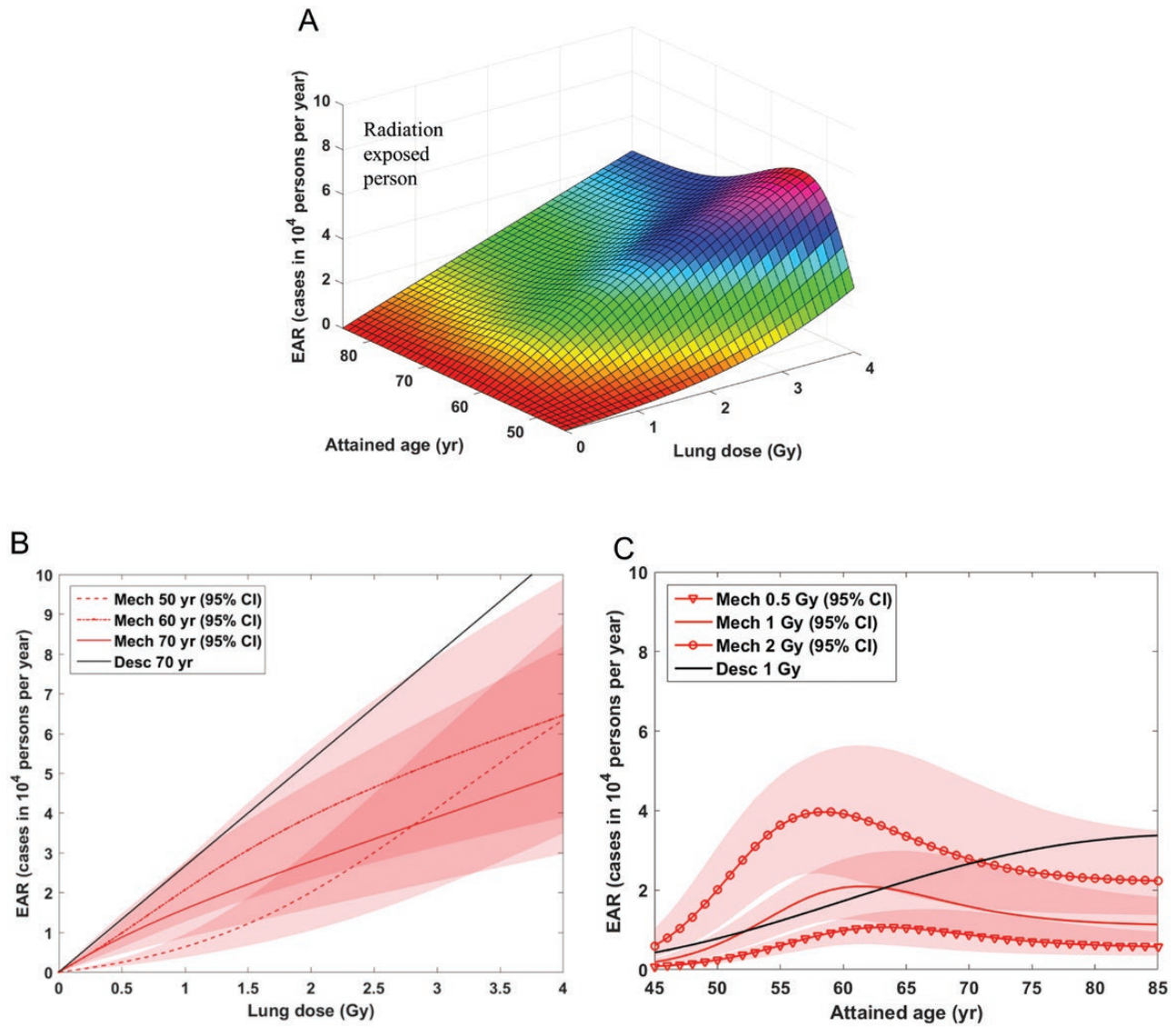


Figure 5. Excess absolute rates (EARs; as cases in 10 000 persons per year) from M_3^{LADC} (Mech) for radiation-induced lung adenocarcinoma in the Japanese LSS cohort for a person exposed at 30 years. The EAR is determined by the linear permanent response to an acute radiation pulse, which increases the clonal expansion rate in the R^{MUT} pathway independent of sex and smoking status (Supplementary Figure S2, bottom, available at *Carcinogenesis* Online). To eliminate the influence for city of residence, person-year weighted city means are used. **(A)** Bivariate EAR dependence on attained age and lung dose. The radiation risk maximizes at about 55 years for high lung dose. **(B)** Cross-sectional cuts to panel (A) for attained ages 50, 60, and 70 years. Over the dose range 0–4 Gy, the EAR responds non-linearly to a lifelong radiation-induced linear response of the clonal expansion rate in the R^{MUT} pathway. **(C)** Cross-sectional cuts to panel (A) for lung doses 0.5, 1 and 2 Gy. The radiation-induced EAR peaks at decreasing age with increasing value. The EAR from a descriptive risk model (Desc) is shown for comparison.

population. Since we performed our main molecular analysis with data from US patients, we considered it important to show the differential prevalence in an Asian (Chinese) data set, which is possibly close to Japanese prevalence in the LSS. Our LSS result for the R^{MUT} pathway predicts a share of 58% (from Figure 6). Since the majority of R^{MUT} cases would carry EGFR mutations, our prediction seems reasonable in view of the large variation in Japanese studies cited by Midha *et al.* (29).

Just like state-of-the-art epidemiological risk models, M_3^{LADC} accurately reproduced LADC incidence in the LSS, albeit with moderately improved goodness-of-fit. Our point estimate of the ERR for an unexposed male smoker starting at 20 years with one pack/day for 50 years of about 3.5 (Supplementary Figure S6, available at *Carcinogenesis* Online) is included in the 95% CI of the estimate 2.4 (1.4, 3,8) from Egawa *et al.* (11). The difference in point estimates might be related to the use of imputed data in our study compared with including data with unknown smoking status in the study of Egawa *et al.* (11). Lubin and Caporaso (30) analyzed a European lung cancer cohort with detailed smoking information using a generalized linear model in logistic regression. In their Figure 4, the sex-independent exposure response for LADC is measured in units of ERR/pack-year and shows remarkable agreement with our results for current male smokers (Supplementary Figure S6, available at *Carcinogenesis* Online). M_3^{LADC} suggests a higher susceptibility of women to smoke, evident by the current LADC pandemic in women (1). A study of EGFR and KRAS mutations in about 3000 LADC of Caucasian patients revealed a higher susceptibility of women to smoking exposure for KRAS-mutant cancers (31). These findings are in line with a stronger increase of the

Radiation intensity (mGy)	Smoking intensity (cigs/day)	Observed cases (%)	Estimated cases (%)	R^{MUT} estimation (%)	Radiation induced estimated cases (%)	T^{MUT} estimation (%)	Smoking induced estimated cases (%)	Spontaneous estimated cases (%)
0-5	=0	**137 (21)**	**139 (22)**	**120 (19)**	**0 (0)**	**19 (3)**	**0 (0)**	**139 (22)**
5+	=0	**121 (19)**	**116 (18)**	**103 (16)**	**19 (3)**	**13 (2)**	**0 (0)**	**97 (15)**
5-100		57 (9)	61 (9)	53 (8)	1 (0)	8 (1)	0 (0)	60 (9)
100+		64 (10)	55 (9)	50 (8)	18 (3)	5 (1)	0 (0)	37 (6)
0-5	>0	**209 (33)**	**209 (33)**	**74 (12)**	**0 (0)**	**135 (21)**	**124 (20)**	**86 (13)**
	1-10	41 (7)	43 (7)	19 (3)	0 (0)	24 (4)	21 (3)	22 (3)
	10-20	109 (17)	105 (16)	37 (6)	0 (0)	68 (11)	63 (10)	42 (7)
	20+	59 (9)	61 (9)	18 (3)	0 (0)	43 (6)	40 (6)	21 (3)
5+	>0	**169 (27)**	**172 (27)**	**71 (11)**	**16 (3)**	**101 (16)**	**93 (14)**	**63 (10)**
5-100		*87 (14)*	*92 (14)*	*33 (5)*	*1 (0)*	*59 (9)*	*54 (8)*	*37 (6)*
	1-10	24 (4)	20 (3)	9 (1)	0 (0)	11 (2)	9 (1)	11 (2)
	10-20	43 (7)	45 (7)	16 (3)	1 (0)	29 (4)	27 (4)	17 (3)
	20+	20 (3)	27 (4)	8 (1)	0 (0)	19 (3)	18 (3)	9 (1)
100+		*82 (13)*	*80 (13)*	*38 (6)*	*15 (3)*	*42 (7)*	*39 (6)*	*26 (4)*
	1-10	21 (3)	18 (3)	10 (2)	3 (1)	8 (1)	9 (1)	6 (1)
	10-20	34 (6)	34 (6)	15 (2)	6 (1)	19 (4)	16 (3)	12 (2)
	20+	27 (4)	28 (4)	13 (2)	6 (1)	15 (2)	14 (2)	8 (1)
Total		**636 (100)**	**636 (100)**	**368 (58)**	**35 (6)**	**268 (42)**	**217 (34)**	**384 (60)**

Figure 6. M_3^{LADC} estimates for the breakdown of 636 LADC cases (% of 636 cases) from the LSS cohort in modeled molecular pathways R^{MUT} and T^{MUT} cross tabulated with exposure groups for smoking and radiation. Refined resolution in exposure subgroups of low (5–100 mGy) and moderate (100+ mGy) radiation dose, and light (1–10 cigarettes/day), moderate (11–20 cigarettes/day) and heavy (20+ cigarettes/day) smoking intensity is made. Female smokers fall mostly in the light category. In each subgroup, observed cases are estimated well by the model. Exposure group numbers (bold faced) add up to total numbers (bold faced) in the bottom line. Exposure subgroup numbers add up to group numbers. Note that M_3^{LADC} estimates are derived from LADC incidence data in the LSS without genotyping. Model estimations for numbers and shares of cases in each molecular pathway would be directly accessible to measurements.

smoking risk in the T^{MUT} pathway for female light smokers compared with male light smokers. Our results are concordant to the aforementioned study and can probably be explained by genetic predisposition of women to persistent smoke-induced DNA damage, notwithstanding the possibility for sex-related differences in innate immune responses to tobacco smoke and its carcinogens, as those observed in inbred strains of mice (32). However, for smoking intensities above about 10 cigarettes per day, our model predicts a decrease in risk for females. This biologically implausible trend was already observed using a state-of-the-art epidemiological risk model (11). A possible explanation might be related to reporting bias, as smoking was not socially accepted for females in Japan in former years. As a striking new feature, M_3^{LADC} clearly identified by goodness-of-fit the two molecular pathways that emerged from the molecular analysis. Importantly, the predictive power of M_3^{LADC} can be subject to rigorous validation by future measurements of the mutation status in LADC tissue of LSS patients.

In the M_3^{LADC}, radiation actions are modeled as permanently increased proliferation of pre-neoplastic lesions after acute exposure. This increased clonal expansion rate for life after a single radiation hit best explains the epidemiologic LSS data, based on goodness-of-fit. In addition, this finding is biologically plausible based on multiple recent experimental lines of evidence that indicate that radiation-induced DNA damage exerts perpetual effects by altering intracellular sensing and pro-inflammatory signaling mechanisms (33–36). To this end, it is well known that ionizing radiation causes DNA damage directly and indirectly via reactive oxygen species production (33). This DNA damage was recently shown to be perpetuated by altered splicing and by aberrant regulation of tumor suppressor gene *TP53* (33). In experimental systems where DNA damage is not caused by radiation but by mere abnormal cellular turnover, caspase 8 was recently shown to mediate epigenetic changes that could very well lead to lifelong inflammation and liver cancer (34). Other groups have linked radiation-induced DNA damage with increased abnormal DNA pattern recognition within micronuclei, providing alternative explanations of how single radiation hits can lead to persistent inflammation (35) and have shown how targeting inflammation can prevent radiation-induced carcinogenesis in the skin (36). Hence, persistent radiation-induced inflammation can be the biological culprit for the lifelong tumor-promoting effects of a single radiation hit observed in the LSS.

Previous molecular studies underpin the biological plausibility of M_3^{LADC}. *KRAS* mutations are more common in smokers (6) and are suspected to confer resistance to radiotherapy (37), which is consistent with the lack of a radiation response in the T^{MUT} pathway in our study. Thus, the main contribution of radiation to LADC incidence is imparted via the R^{MUT} pathway and a possible contribution from the T^{MUT} pathway is too small for quantification. To date, the risk factor that drives LADC development in never smokers is unknown, whereas these patients exhibit higher frequencies of *EGFR* mutations and *EML4/ALK* fusions (3–6). Here, we show that radiation may drive disease development in these patients and provide a risk model for this molecular class of LADC. The genomic signatures of radiation and of MMR for radiation-induced DNA strand breaks were enriched in R^{MUT} tumors (24). R^{MUT} tumors also displayed elevated indel/SNV ratios, shown elsewhere to be a hallmark of secondary cancers induced by γ-IR (25). Moreover, gene fusions such as *EML4/ALK*, *KIF5B/RET* and *CD74/ROS1*, included here in the R^{MUT} pathway, have been linked with radiation in other cancers (14,26). These observations correspond to the radiation response of the R^{MUT} pathway as the most relevant

radiation effect proposed by M_3^{LADC}. Hence, we link for the first time radiation exposure to a molecular subset of LADC using molecular and epidemiologic evidence.

Smoking is linked with *KRAS*-mutant LADC and US patient analysis showed enhanced mutation rates in ever smokers of the T^{MUT} pathway (6). This observation should generate an increase of initiating mutations in smokers and was found significant in a one-path model (Supplementary Table S2, available at *Carcinogenesis* Online). However, M_3^{LADC} works without such a plausible smoking effect because improvement in goodness-of-fit was not increased compared with exclusive smoking action on clonal expansion. This result might be caused by insufficient statistical power. Hence, the main biological mechanism of smoking on LADC incidence is associated with enhanced clonal growth as already observed in (20) and (38). Initiated cells exhibit a growth advantage over healthy cells due to reduced cell death possibly caused by smoking-associated chronic inflammation. Hence our data build on the known linkage between smoking and *KRAS*-mutant LADC by expanding this link to T^{MUT} and O^{WT} LADC and by pinning the effects of smoke in time: at early time-points of smoking exposure. These results are relevant and important for the design of future chemoprevention strategies aimed to halt disease progression in smokers.

Prediction models, which are informed by adequate bioassays in addition to epidemiological variables, can forecast lung cancer occurrence with high accuracy (39). They do lack, however, a link between environmental agents and molecular risk stratification, which is provided by M_3^{LADC}. This link suggests no interactively increased LADC risk for heavy smokers in computerized tomography screening. It can be exploited in retrospective assessment to pin down the agent causing LADC based on the molecular profile of diseased tissue, which is highly relevant e.g. for compensation claims in the nuclear industry.

In conclusion, our study answers a longstanding question on the biological origins of age-risk patterns for LADC from concomitant exposure to smoking and radiation. To describe such patterns, standard epidemiological models must inevitably rely on a vague implementation of synergistic effects, which are commonly couched in mathematical terms as either 'additive' or 'multiplicative' sometimes with further generalizations (9–11). We have shown here by projecting signatures of environmental exposure into epidemiological cohorts that smoking and radiation drive the development of LADC along different molecular pathways with negligible interaction for doses below 4 Gy. The M_3^{LADC} approach provides a powerful tool for harnessing molecular data to improve studies of risk assessment and prediction in radiation protection and clinical applications. Our approach is of clinical relevance because we solidify cause–effect relationships in LADC development by integrating molecular and epidemiologic data. The cause of LADC can be inferred from their molecular alterations and the share of LADC with specific alterations can be predicted using the M_3^{LADC} model with possibly huge medical and socioeconomic implications'

Supplementary material

Supplementary data are available at *Carcinogenesis* online.

Funding

German Federal Ministry of Education and Research (BMBF) in the framework of Kompetenzverbund Strahlenforschung (KVSF) (FKZ 02NUK026 to N.C.); European Research Council Starting Independent Investigator (260524 to G.T.S.); Proof of Concept Grants (679345 to G.T.S.).

Acknowledgements

We thank Ignacio Zaballa for fruitful discussions on mechanistic model conception and development and Klemens M. Thaler for proof-reading the final version of the paper draft. This work could not be done without a strong cooperation with the Radiation Effect Research Foundation (RERF) that provided the data and helped in the analysis. The RERF of Hiroshima and Nagasaki, Japan is a public interest foundation founded by the Japanese Ministery of Health, Labor and Welfare (MHLW) and the US Department of Energy (DOE).
Conflict of Interest Statement: None declared.

References

1. Fitzmaurice, C. *et al.*; Global Burden of Disease Cancer C (2017) Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 Cancer Groups, 1990 to 2015: a systematic analysis for the global burden of disease study. JAMA Oncol., 3, 524–548.
2. Torre, L.A. *et al.* (2015) Global cancer statistics, 2012. CA Cancer J. Clin., 65, 87–108.
3. Hecht, S.S. (1999) Tobacco smoke carcinogens and lung cancer. J. Natl. Cancer Inst., 91, 1194–1210.
4. Alberg, A.J. *et al.* (2013) Epidemiology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest, 143(5 suppl.), e1S–e29S.
5. Sun, S. *et al.* (2007) Lung cancer in never smokers–a different disease. Nat. Rev. Cancer, 7, 778–790.
6. Campbell, J.D. *et al.*; Cancer Genome Atlas Research Network (2016) Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat. Genet., 48, 607–616.
7. Wu, K. *et al.* (2015) Frequent alterations in cytoskeleton remodelling genes in primary and metastatic lung adenocarcinomas. Nat. Commun., 6, 10131.
8. United Nations Scientific Committee on the Effects of Atomic Radiation (2012) *Sources and Effects of Ionizing Radiation*. UNSCEAR 2012 Report. UNSCEAR.
9. Furukawa, K. *et al.* (2010) Radiation and smoking effects on lung cancer incidence among atomic bomb survivors. *Radiat. Res.*, 174, 72–82.
10. Cahoon, E.K. *et al.* (2017) Lung, laryngeal and other respiratory cancer incidence among Japanese atomic bomb survivors: an updated analysis from 1958 through 2009. Radiat. Res., 187, 538–548.
11. Egawa, H. *et al.* (2012) Radiation and smoking effects on lung cancer incidence by histological types among atomic bomb survivors. Radiat. Res., 178, 191–201.
12. Kaiser, J.C. *et al.* (2014) Genomic instability and radiation risk in molecular pathways to colon cancer. PLoS One, 9, e111024.
13. Kaiser, J.C. *et al.* (2016) Integration of a radiation biomarker into modeling of thyroid carcinogenesis and post-Chernobyl risk assessment. Carcinogenesis, 37, 1152–1160.
14. Efanov, A.A. *et al.* (2018) Investigation of the relationship between radiation dose and gene mutations and fusions in post-Chernobyl thyroid cancer. J. Natl. Cancer Inst., 110, 371–378.
15. Ruhm, W. *et al.* (2017) Biologically-based mechanistic models of radiation-related carcinogenesis applied to epidemiological data. Int. J. Radiat. Biol., 93, 1–25.
16. Ozasa, K. *et al.* (2012) Studies of the mortality of atomic bomb survivors, Report 14, 1950–2003: an overview of cancer and noncancer diseases. Radiat. Res., 177, 229–243.
17. Furukawa, K. *et al.* (2017) Handling incomplete smoking history data in survival analysis. Stat. Methods Med. Res., 26, 707–723.
18. Luebeck, E.G. *et al.* (2013) Impact of tumor progression on cancer incidence curves. Cancer Res., 73, 1086–1096.
19. Feller, W. (1968) An Introduction to Probability Theory and Its Applications. Wiley, New York, NY.
20. Meza, R. *et al.* (2008) Age-specific incidence of cancer: phases, transitions, and biological implications. Proc. Natl. Acad. Sci. USA, 105, 16284–16289.

21. Little, M.P. *et al.* (2010) Parameter identifiability and redundancy: theoretical considerations. PLoS One, 5, e8915.
22. Luebeck, E.G. *et al.* (2008) Does folic acid supplementation prevent or promote colorectal cancer? Results from model-based predictions. Cancer Epidemiol. Biomarkers Prev., 17, 1360–1367.
23. Marshall, A. *et al.* (2009) Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med. Res. Methodol., 9, 57.
24. Alexandrov, L.B. *et al.*; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain (2013) Signatures of mutational processes in human cancer. Nature, 500, 415–421.
25. Behjati, S. *et al.*; ICGC Prostate Group (2016) Mutational signatures of ionizing radiation in second malignancies. Nat. Commun., 7, 12605.
26. Seki, Y. *et al.* (2015) Molecular process producing oncogene fusion in lung cancer cells by illegitimate repair of DNA double-strand breaks. Biomolecules, 5, 2464–2476.
27. Tomasetti, C. *et al.* (2015) Only three driver gene mutations are required for the development of lung and colorectal cancers. Proc. Natl. Acad. Sci. USA, 112, 118–123.
28. Reck, M. *et al.* (2013) Management of non-small-cell lung cancer: recent developments. Lancet, 382, 709–719.
29. Midha, A. *et al.* (2015) EGFR mutation incidence in non-small-cell lung cancer of adenocarcinoma histology: a systematic review and global map by ethnicity (mutMapII). Am. J. Cancer Res., 5, 2892–2911.
30. Lubin, J.H. *et al.* (2006) Cigarette smoking and lung cancer: modeling total exposure and intensity. Cancer Epidemiol. Biomarkers Prev., 15, 517–523.
31. Dogan, S. *et al.* (2012) 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., 18, 6169–6177.
32. Stathopoulos, G.T. *et al.* (2007) Epithelial NF-kappaB activation promotes urethane-induced lung carcinogenesis. Proc. Natl. Acad. Sci. USA, 104, 18514–18519.
33. Chen, J. *et al.* (2017) Identification of a DNA damage-induced alternative splicing pathway that regulates p53 and cellular senescence markers. Cancer Discov., 7, 766–781.
34. Boege, Y. *et al.* (2017) A dual role of caspase-8 in triggering and sensing proliferation-associated DNA damage, a key determinant of liver cancer development. Cancer Cell, 32, 342–359.e10.
35. Harding, S.M. *et al.* (2017) Mitotic progression following DNA damage enables pattern recognition within micronuclei. Nature, 548, 466–470.
36. Rigby, C.M. *et al.* (2017) Role of p53 in silibinin-mediated inhibition of ultraviolet B radiation-induced DNA damage, inflammation and skin carcinogenesis. Carcinogenesis, 38, 40–50.
37. Wang, M. *et al.* (2014) EGFR-mediated chromatin condensation protects KRAS-mutant cancer cells against ionizing radiation. Cancer Res., 74, 2825–2834.
38. Schöllnberger, H. *et al.* (2006) Analysis of epidemiological cohort data on smoking effects and lung cancer with a multi-stage cancer model. Carcinogenesis, 27, 1432–1444.
39. El-Zein, R.A. *et al.* (2006) Cytokinesis-blocked micronucleus assay as a novel biomarker for lung cancer risk. Cancer Res., 66, 6449–6456.
40. Yatabe, Y. *et al.* (2011) Do all lung adenocarcinomas follow a stepwise progression? Lung Cancer, 74, 7–11.