Biological-Based Models of Carcinogenesis in the Lung from Radiation and Smoking

Noemi Castelletti

München 2018

Kerstin Mayer, *Tree of Life*, 2018

Biological-Based Models of Carcinogenesis in the Lung from Radiation and Smoking

Noemi Castelletti

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität
München

vorgelegt von
Noemi Castelletti
aus Bozen

München, den 25. Oktober 2018

Erstgutachter: Prof. Dr. Helmut Küchenhoff
Zweitgutachter: Prof. Dr. Johannes Müller
Tag der Disputation: 18. Januar 2019

ABSTRACT

Lung adenocarcinoma and squamous cell carcinoma are the deadliest cancers worldwide. Smoking and ionizing radiation are potent carcinogens affecting strongly both lung cancer subtypes. Several biological analyses have been performed to characterise the genetic mutations leading to adenocarcinoma and squamous cell carcinoma, and different genomic spectra have been observed. Biological markers of smoking related damage could be found, leading to a deep knowledge of cellular smoking effects. Less is known about the biological effects of radiation in human carcinogenesis. Risks have been quantified with epidemiological studies of these carcinogens. Based on the biologically substantiated assumption that the number of mutations is linearly related to the dose, in radiation epidemiology it is standard to model effects linearly. These models do however not have a biological interpretation and are disconnected from general statistical methods. Here we fill both gaps. First we apply statistical generalised additive models to examine the functional relation between risk and smoking and radiation effects. Secondly, with mechanistic multi-scale models we integrate molecular biology and epidemiology to describe the carcinogenesis of lung adenocarcinoma and squamous cell carcinoma.

To investigate the incidence of lung adenocarcinoma and lung squamous cell carcinoma we analysed two cohorts: first the Life Span Study cohort of atomic bomb survivors of Hiroshima and Nagasaki, and second the Eldorado cohort of Canadian Uranium miners. Exposures differed strongly between cohorts. Residents of Hiroshima and Nagasaki were exposed to a relative high dose of γ radiation for a short time, while the miners were exposed to a protracted and lower exposure to γ and α radiation. Information about smoking habits is available only for the former cohort.

Three types of models were applied to analyse the effects of radiation and smoking: state-of-the-art statistical risk models of radiation protection, statistical generalized additive models and mechanistic risk models. Although there were quantitative differences in effect size and significance, each result is presented below only for a single model.

For lung adenocarcinoma the best mechanistic model was a two pathway model. Smoking and radiation effects showed markedly different patterns: both acted on the apoptosis rate of precancerous cells but on different pathways without any interaction. A linear radiation effect was found in one pathway and a linear-exponential smoking effect in the other pathway. Independently of these results we analysed genomic data of American patients. It is known that the genetic damage of people with adenocarcinoma can be grouped into three pathways: the

receptor mutant (R^{MUT}) pathway, the transducer mutant pathway (T^{MUT}), and other signatures (O^{WT}). We could show that signatures of T^{MUT} and the O^{WT} pathways do differ much less from each other than both differed to the R^{MUT} pathway. Therefore, there is also genetic evidence that adenocarcinoma fall into two main classes. The two pathways of the mechanistic model could be associated to the R^{MUT} and R^{MUT}+O^{WT} pathways by their risk patterns in age and smoking.

On the other hand, for squamous cell carcinoma one pathway was sufficient to describe the incidence data. Although effects of radiation appeared to be highly significant, they could be traced back to arise only from the first five years of follow up (11 cases therein). When the first five years were excluded, no significant radiation effect could be found. Interestingly, for lung squamous cell carcinoma the mechanistic models could fit the effects of cigarette smoking in initiation and promotion. This was different for lung adenocarcinoma, where the main effect of smoking was a promotion of already existing pre-cancerous clones. For both, lung adenocarcinoma and squamous cell carcinoma, no interaction between radiation and smoking could be fitted for the Life Span Study cohort.

Results from analysis of the Eldorado cohort were in line with the results presented above. For lung adenocarcinoma both, the state-of-the-art statistical risk models and the generalised additive models, could find only a significant effect of γ radiation exposure. For lung squamous cell carcinoma, *vice versa*, both models could find only a significant effect of α radiation exposure.

Concluding, we showed that lung cancer cannot be investigated as a single endpoint but the different subtypes have to be analysed separately. Different radiation qualities act differently to the different subtypes, indicating different biological processes. Analogously, although smoking is an important risk factor for all subtypes, its effects were different and with different magnitudes.

ZUSAMMENFASSUNG

Adenokarzinome und Plattenepithelkarzinome der Lunge sind die tödlichsten Krebsarten weltweit. Rauchen und ionisierende Strahlung wirken auf beide Lungenkrebs-Subtypen stark karzinogen. Die genetischen Mutationen, die zu Adenokarzinom und Plattenepithelkarzinom führen, sind aus biologischen Untersuchungen bekannt. Dabei wurden unterschiedliche genomische Spektren beobachtet. Mehrere biologische Marker für Lungenkrebs in Rauchern konnten gefunden werden, was zu einem fundierten Wissen über die von Rauchen verursachten zellulären Effekte geführt hat. Weniger ist allerdings über die biologischen Effekte der Strahlung in der Karzinogenese der Lunge bekannt. Epidemiologische Studien erlauben die Quantifizierung der Risiken durch diese Karzinogene. Basierend auf der biologisch fundierten Annahme, dass die Anzahl der Mutationen linear mit der Strahlungsdosis ansteigt, ist es in der Strahlen-Epidemiologie üblich, das Risiko als lineare Funktion zu modellieren. Auch wenn die Linearität biologisch motiviert ist, haben diese Modelle ansonsten keine biologische Interpretation, unterscheiden sich dadurch aber von den allgemein verwendeten statistischen Methoden, so dass neuere statistische Entwicklungen nicht in die Strahlen-Epidemiologie übernommen wurden. Diese Arbeit soll zu beiden Richtungen beitragen. Einerseits wenden wir generalisierte additive Modelle an, um die funktionale Beziehung von Rauchen und Strahlungseffekten zu untersuchen. Andererseits verwenden wir molekulare Biologie und Epidemiologie, um mit mechanistischen Multi-Skalenmodellen die Krebsentstehung von Adeno- und Plattenepithelkarzinomen der Lunge zu beschreiben.
Um die Inzidenz von Adeno- und Plattenepithelkarzinom zu untersuchen, wurden zwei Kohorten analysiert: die Life Span Study der Atombombenüberlebenden in Hiroshima und Nagasaki sowie die Eldorado-Kohorte von kanadischen Uran-Minenarbeitern. Die Expositionen unterscheiden sich zwischen den Kohorten stark. Die Bewohner von Hiroshima und Nagasaki waren einer relativ hohen Dosis an γ-Strahlung in kurzer Zeit ausgesetzt, während die Minenarbeiter einer Langzeitexposition mit geringerer Dosis an α- und γ-Strahlung ausgesetzt waren. Informationen zum Rauchverhalten der Teilnehmer sind nur von der ersten Kohorte bekannt.
Der Effekt von Strahlung und Rauchen wurde mit drei Arten von Modellen analysiert: mit den im Strahlenschutz derzeit üblichen und mit generalisierten additiven statistischen Risikomodellen, sowie mit mechanistischen Risikomodellen. Auch wenn es quantitative Unterschiede in den Effektgrößen und Signifikanzen gab, werden die Ergebnisse im Folgenden jeweils nur für ein Modell präsentiert.
Das beste mechanistische Modell für Adenokarzinome der Lunge war ein Modell mit zwei

Pfaden. Rauchen und Strahlungseffekte zeigten deutlich unterschiedliche Muster: zwar wirkten beide auf die Apoptose-Rate von intermediären, präkanzerösen Zellen, jedoch im jeweils anderen Pfad ohne jegliche Interaktion. Ein linearer Strahlungseffekt wurde in dem einem Pfad und ein linear-exponentieller Effekt des Rauchens im anderen Pfad gefunden. Unabhängig von diesen Ergebnissen haben wir genomische Daten von amerikanischen Patienten untersucht. Es ist bekannt, dass Erbgutschäden von Menschen mit Adenokarzinom in drei Pfade gruppiert werden können: dem Rezeptor-Mutations-Pfad (R^{MUT}), dem Transducer-Mutations-Pfad (T^{MUT}) und anderen Signaturen (O^{WT}). Wir konnten zeigen, dass T^{MUT} und O^{WT} sich voneinander deutlich weniger unterscheiden als vom R^{MUT} Pfad. Es gibt also auch genetische Evidenz, dass Adenokarzinome in zwei Hauptgruppen aufgeteilt werden können. Die zwei Pfade im mechanistischen Modell konnten mit den den biologischen Pfaden an Hand der Abhängigkeiten des Risikos von Alter und Rauchen in Beziehung gesetzt werden.
Für Plattenepithelkarzinome war ein Pfad ausreichend, um die Inzidenzdaten zu beschreiben. Der Effekt der Strahlung erschien zwar hochsignifikant, aber der Effekt bezog sich nur auf die ersten fünf Jahre der Beobachtungszeit (und damit 11 Fälle). Wenn die ersten fünf Jahre ausgeschlossen wurden, konnte kein signifikanter Effekt der Strahlung gefunden werden. Bemerkenswerterweise ergab der beste Fit mit dem mechanistischen Modell Effekte des Rauchens in Initiation und in Promotion des Plattenepithelkarzinoms. Dies war anders als beim Adenokarzinom, wo der Haupteffekt des Rauchens in der Promotion bereits existierender präkanzeröser Klone lag. Für beide Suptypen konnte keine Interaktion zwischen Strahlung und Rauchen in der Life Span Study beobachtet werden.
Die Ergebnisse der Eldorado-Kohorte waren im Einklang mit den oben beschriebenen Ergebnissen. Für Adenokarzinome der Lunge konnten sowohl das im Strahlenschutz derzeit übliche, als auch das generalisierte additive statistische Risikomodell nur einen signifikanten Effekt der γ-Strahlung finden. Für Plattenepithelkarzinome konnten beide Modelle nur einen signifikanten Effekt durch α-Strahlung feststellen.
Wir konnten also zeigen, dass Lungenkrebs nicht als ein einzelner Endpunkt analysiert werden kann, sondern dass die Subtypen unterschieden werden müssen. Unterschiedliche Strahlungsarten wirken sich auf die unterschiedlichen Subtypen verschieden aus, was auf unterschiedliche biologische Prozesse schließen lässt. Ebenso, auch wenn Rauchen ein wichtiger Risikofaktor für alle Subtypen ist, unterscheiden sich auch hier die Effekte und haben unterschiedliche Stärke.

ACKNOWLEDGMENTS

This work would not exist without the help and the encouragement of, actually, a lot of people.

First of all I want sincerely thank Prof. Dr. Helmut Küchenhoff, my supervisor at the Ludwig-Maximilians-Universität München, Faculty of Mathematics, Informatics and Statistics. Thank you very much for your expertise and your time. I learned a lot from you and from your way of teaching, growing a lot as a person and as a scientist.

I want to thank Dr. Jan Christian Kaiser, my supervisor at the Helmholtz Center Munich, Institute of Radiation Protection, head of the "Integrative Modelling" working group. You supported me in any situation and taught me, how to be a scientist. It was a pleasure to have the possibility to work on your side. Thank you a lot for all the help and encouragement you gave me in this years.

Thank you, Prof. Dr. Johannes Müller, my external supervisor at the Technical University of Munich. When I started my studies in Mathematics, in 2009, I did not even know the existence of "Biomathematics". One day, walking through the corridors of the mathematics building in Garching Forschungszentrum, I read the name of the chair you work in: "Chair of Mathematical Modeling of Biological Systems". The name fascinated me and I attended many of your courses. The way you explained to me this topic is still pushing me in not so easy times.

During my stay at the Helmholtz Center Munich I had the possibility to visit two oversea institution: the Radiation Effect Research Foundation (RERF) in Hiroshima, Japan, and the University of California San Francisco (UCSF). At the RERF my reference person was Prof. Dr. Kyoji Furukawa. Thank you so much for the knowledge you passed me and for the grate work we did together. Thank you also a lot for your hospitality in Hiroshima. I think, it is especially because of the kindness of people that I like the Japanese culture so much. I will not forget the respectful way of being that I learned in your country. At UCSF my reference person was Prof. Dr. Lydia Zablotska. Thank you so much for all the knowledge you gave me and the possibility to learn from you. Thank you for the amazing data I could analyse, under you advice. Thank you for your hospitality in San Francisco. I really appreciated our sightseeing through the city (p.s. every time I buy dahlias, I have to think of you).

Thank you, Dr. Cristoforo Simonetto, postdoc in the "Integrative Modelling" group. In the last years you more and more became a kind of additional supervisor for me. Every time I asked you, you had time for me and, of course, the right words. Thank you for all I could learn from you and for your advices. A lot of the results presented in this Thesis were inspired by discussions with you, together with Christian. Thank you very much for everything. You are a really beautiful person with a much more beautiful family. You remembered me many many times, not to forget to go home. Thanks also for that.

Before Cristoforo, I had the pleasure to work together with Dr. Ignacio Zaballa (alias Igi). I really enjoyed our mathematical discussions and our lunches at the Asian place. It was a honor to sit in your office.

During the last part of my PhD Alessia Mafodda entered our group. Thanks to you and to your partner Emilio for deciding to come to Munich. We spent and are spending great time together! Thanks for helping me to keep the smile, although sometime it's not so easy to do. Sometime I have the feeling we know each other since ages, even though that's not the case. Thanks for everything!!

I would like also to thank all the people that still are or were in the "ISAR" group: Dr. Michael Dischер, Dr. Christian Staudt, Dr. Markus Eidemüller, Dr. Werner Friedland, Dr. Pavel Kundrát, Dr. Denise Güthlin, Dr. Reinhard Meckbach, Dr. Elke Schmitt, Dr. Helmut Schöllnberger, Dr. Elena Shemiakina and Dr. Alexander Ulanovsky, Albrecht Wieser, Dr. Clemens Woda, Dr. Sascha Zöllner. Thank you for all the lunches together and for the great advices that each one of you could give me. Different working fields did not limit fruitful discussion but enriched them.

Starting my PhD I also started my work in the StaBLab group. Thank you Dr. André Klima and Dr. Veronika Deffner for your patience in showing me things. From your experience I could get the feeling of the StaBLab group: everyone learns from each other and working together makes things easier. Dear Alexander Bauer, I really owe you a cake (actually more than one, and vegan, that's clear). Sevag Kevork, thanks for all jokes, but also for your expertise in modelling.
Linda Marchioro, we know each other since a long time. Thank you for the passion to data analysis you transmitted to me every time I was frustrated. You are a good friend!

Some parts of the descriptive data analysis of this Thesis were inspired by some project at the "Anfängerpraktikum" at the LMU. I want to thank all the students involved for the brilliant ideas.

Dear Prof. Dr. Christina Kuttler, thanks. I tried to make a list of all the things I wanted to thank you and only the bullet points were too many. You grow me up from the university point of view. First you showed me the way through my studies, now you are also showing me how to guide others. Thanks a lot for your good mood, every smile makes everything better. So, concluding, thanks!!

I want to thank two students from the TUM that wrote a Master's and a Bachelor's Thesis under C. Kuttler's and my supervision. Thanks to Katharina Strahler and Jonathan Haas for the collaboration and the writing down of the mathematical formulas in LaTex.

I want to thank all my friends that supported me in this time.
Especially I want to thank Kerstin Mayer and Matthias Schmitt, you are really good friends, not much more writing is needed. Thanks Kerstin for the beautiful painting you made as the cover of this work!
Thanks also to Julia, Bastian and Luisa Kollmannsberger for all the adventures we share(d).
Dear Anna M. Mathes, thanks for everything! We studied together and managed to remain really good friends also after the university. Thanks for keeping calling me! Thanks for the advices in difficult time and for the loud laughs in good ones. Thanks my friend!
Now I want to thank the most important part of my life: my family + boyfriend. Since my family doesn't speak English, I would like to with to Italian.
Cara famiglia, finalmente un capitolo della tesi comprensibile anche a voi. Vi ringrazio per tutto, e intendo proprio tutto. Mi avete accompagnata fin dall'inizio, con pianti per ogni esame andato male e con gioie immense per ogni passo in avanti. Ora sono qui, a scrivere gli ultimi ringraziamenti ed a pensare indietro. Grazie a voi genitori, voi avete reso possibile che io studiassi. Mi avete cresciuta, educata e seguita, e continuate a farlo. Grazie per la famiglia che avete costruito, siamo uniti, tutti e 4 (5 adesso). Grazie sorella di tutto. Siamo cosí diverse, che di piú non é possibile, ma proprio per questo impariamo una dall'altra e ci sosteniamo a vicenda. Abbiamo litigato tanto, perché cosí tanto ci vogliamo anche bene. Ciao piccolo Edo. Tu sei nato nel mentre di questo lavoro e mi hai ricordato una cosa tanto grande, quanto piccolo tu sei: la vita. Niente stress, niente brutti pensieri, la vita serena come quella dei bambini (quando non piangono). Ti voglio bene piccola creatura fantastica.
Un pensiero va anche ai miei nonni, conosciuti, mai conosciuti, vivi e ormai piú felici di noi. Cara nonna che non ci sei piú, ti penso tanto, soprattutto quando cucino e quando cucio con la tua macchina vestitini per il piccolo Edo. Lo so che mi guardi da lassú, ti sento. Grazie per tutto quello che mi hai insegnato e trasmesso. Quando mi si parla di rispetto, penso sempre a te. Un bacio fin lassú, mora. Cara nonna che mi accompagni ancora nella mia vita, grazie per tutto quello che continui ad insegnarmi. Umiltá, rispetto e amore sono tue caratteristiche. Penso spesso a te col nonno quando discutiamo io e Klemi, mi sa che sembriamo proprio voi due. Grazie di tutto, un abbraccio!
Famiglia, grazie di tutto, vi voglio bene!!

Grazie a te Klemi! Cosí tanto tempo passato insieme. Senza di te non sarei mai arrivata dove sono. Grazie del tuo perenne sostegno e della tua pazienza infinita. Grazie delle tue parole: "Mi dispiace ma non ti lascio mollare e tornare a casa" nella crisi all'inizio dello studio. Come spesso, avevi ragione. Grazie di tutto quello che mi insegni, delle risate e della gioia di vivere. Se penso indietro ci sei sempre, e sono contenta che sia cosí. Grazie della tua generositá e del tuo aiuto al prossimo. Mi ricordi sempre cosa sia la vera caritá. Rimani come sei, che ti amo per questo. Grazie di tutto, di avermi aiutata ad arrivare nella vita fino a qui!!

CONTENTS

List of Abbreviations

3SCE Three Stage Clonal Expansion Model

EAR Excess Additive Risk Models

AIC Akaike's Information Criterion

CI Confidence Intervals

CNA Copy Number Alteration

$Stat_{ELDO}^{LADC}$ State-of-the-art statistical risk model for lung adenocarcinoma in the Eldorado cohort

$Stat_{LSS}^{LADC}$ State-of-the-art statistical risk model for lung adenocarcinoma in the Life Span Study cohort

$Stat_{ELDO}^{SQUAM}$ State-of-the-art statistical risk model for squamous cell carcinoma in the Eldorado cohort

$Stat_{LSS}^{SQUAM}$ State-of-the-art statistical risk model for squamous cell carcinoma in the Life Span Study cohort

EAR Excess Absolute Rates

edf Estimated Degrees of Freedom

ERR Excess Relative Risk

GAERM Generalized Additive Excess Risk Models

GAM Generalized Additive Model

GAM_{ELDO}^{LADC} State-of-the-art statistical risk model for lung adenocarcinoma in the Eldorado cohort

GAM_{LSS}^{LADC} Generalized Additive Model for lung adenocarcinoma in the Life Span Study Cohort

GAM_{ELDO}^{SQUAM} State-of-the-art statistical risk model for squamous cell carcinoma in the Eldorado cohort

GAM_{LSS}^{SQUAM} Generalized Additive Model for squamous cell carcinoma in the Life Span Study Cohort

GMRRM Generalized Multiplicative Relative Risk Models

H3SCE Hybrid Three Stage Clonal Expansion Model

IC Initial Condition

indel Insertion/Deletion

IR Ionizing Radiation

LADC Lung Adenocarcinoma

LET Linear Energy Transfer

LM Linear Model

LSS Life Span Study Cohort

M_3 Molecular Mechanistic Model

MI Multiple imputation

M_3^{LADC} Molecular Mechanistic Model for lung adenocarcinoma

ERR Excess Relative Risk Models

M_2^{SQUAM} Mechanistic Model for lung squamous cell carcinoma

NSCLC Non-Small Cell Lung Cancer

ODE Ordinary Differential Equation

PDE Partial Differential Equation

PYRs Person years

R^{MUT} Receptor Mutant

SNV Single nucleotide variant

SQUAM Lung Squamous Cell Carcinoma

T^{MUT} Transducer Mutant

TSCE Two Stage Clonal Expansion Model

WLM Working level month

CHAPTER

1

INTRODUCTION

The aim of this Thesis is to analyze carcinogenesis and the related risk in the lung under the effects of smoking and two types of ionising radiation. This Chapter will hence start with a bio-epidemiological introduction of lung cancer and his histological types. It will be illustrated that the histological types are different in position, genetic profile and cell structure and can be considered as distinct diseases. This motivated our decision to analyse the different subtypes of lung cancer separately. Two cohorts will be considered: the first cohort is the Japanese Life Span Study of atomic bomb survivors of Hiroshima and Nagasaki exposed to a mixed field of neutrons, gamma waves and smoking. Secondly, the Eldorado cohort of Canadian uranium mine workers, exposed to both radon and gamma radiation. To better understand the modelling of these three different exposures brief introductions of their effects will be given. This Chapter will then conclude with a general overview of the Thesis.

1.1 Lung cancer

In 2012 around 1.8 million new lung cancer cases occurred, about 13% of the total cancers diagnosed, becoming worldwide the leading cause of death for males and the second one for females, after breast cancer [25, 59]. There are many known risks factors for lung cancer: asbestos, arsenic, radon, outdoor pollution and smoking [53]. Lung cancer rates differ a lot between countries and this can be related to the different smoking behaviour [7]. Since a very large part of lung cancers could be avoided by smoking cessation, lung cancer can be considered one of the most preventable cancers [59].

Lung cancer can be characterised by two main properties:

- histology: cells under a microscope, and
- molecular profile (also signature profile, genomic profile, or bio-marker profile): signatures found in cancer tissue [39].

Each patient is treated depending on these two factors, in order to minimize side effects and maximize the efficiency of treatments.

Histologically considered, about 15% of lung cancers are small cell lung cancer, while about 85% are non-small cell lung cancers (NSCLC) [14, 39]. NSCLC can again be differentiated between:

- Adenocarcinoma (LADC),
- Squamous cell lung cancer (also called epidermoid carcinoma) (SQUAM), and
- Large cell lung cancer.

In this Thesis we will look only on NSCLC and more precisely only on LADC and SQUAM.

The classification of lung cancer on the molecular level is performed by gene mutation in cancer tissue. The molecular spectrum of the histological types of lung cancer differs a lot [3, 9, 10, 44].

In the next sections a brief discussion of the biology of LADC and SQUAM will be presented.

1.1.1 Biological characterisation of lung adenocarcinoma

LADC is the deadliest cancer worldwide: 42% of all lung cases in women and 28% in men were classified as LADC [63]. The shape of the tumour is highly variable and can develop multiple centers. LADC tumours often appear in the periphery of the lung and develop in smaller airways, such as bronchioles (see Figure 1.1).
LADC is mostly driven by tobacco smoke, often causing mutations in the KRAS oncogene [10, 63]. While these mutations are found in different frequencies in Caucasian and Asian LADC patients and while KRAS mutations are more frequent in smokers, other risk factors (e.g. oncogene links) are unknown and molecular risk prediction models are missing [3, 9, 42, 57].
Most LADC cases occur in smokers, but compared to other histological types it is more frequently found in never smokers, particularly in women [56, 63]. Based on markedly different molecular profiles, LADC in never smokers develops completely different than in ever smokers. LADC in never smokers is the common lung cancer form in young patients and its causes are still not clear [56]. Since the female portion in LADC in never smokers is much higher than in other types, there is also the possibility that gender-dependent hormones could play a role in the development of LADC [56].

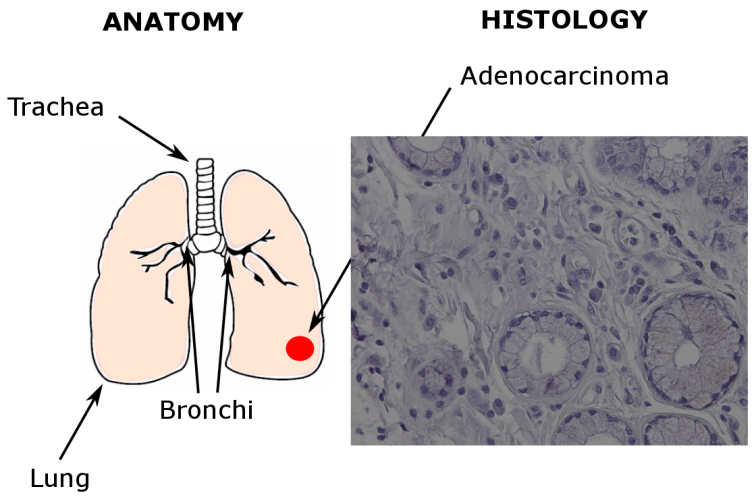


Figure 1.1: Histology and position in the lung of LADC. Histology provided by I. Gipanou [24].

The Cancer Genome Atlas Research [10] analysed LADCs from 230 patients and did a whole exome sequencing of the tissues. The molecular profile can be found in Figure 1.2. Columns represent patients, rows the analyzed variable (gender, smoking status and gene). The most

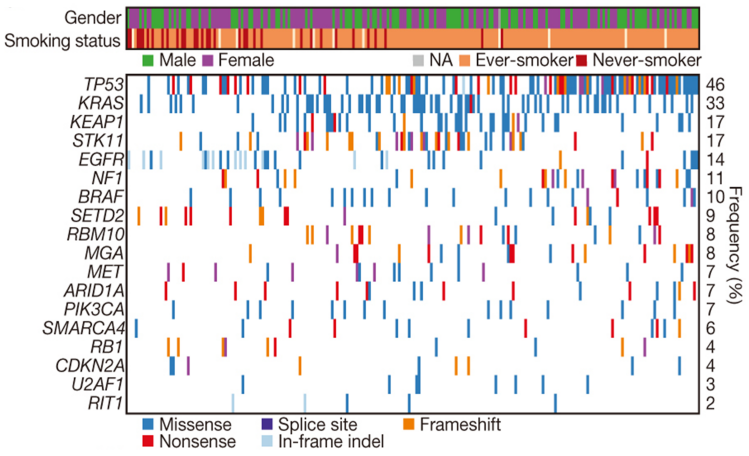


Figure 1.2: Mutational spectra from whole exome sequencing of 230 LADCs. Columns represent patients, rows the analyzed variable (gender, smoking status and gene). Figure taken from [10].

frequently mutated genes compared with normal tissue are: TP53, KRAS, KEAP1, STK11, EGFR, NF1 and BRAF. Looking at the columns of the table one can notice that, although all patients had LADC, there are patients with a lot of mutations and patients with just few mutations. Analysing the rows, it is clear that some genes are more frequently mutated than others. For the same histology, subtypes can be distinguished on a DNA level by different molecular profiles. Since the molecular form of these subtypes is different there must be also a different pathway leading to the endpoint.

1.1.2 Biological characterisation of lung squamous cell carcinoma

Forming 30% of all lung cancers, SQUAM is the second most common type and the gender proportion is reversed compared to LADC: 44% of lung cancers in men, and 25% in women [63]. Generally SQUAM tumours are of large size and may cavitate. A cavity is defined as a gas/fluid-filled space within a tumor mass or nodule. SQUAM is described as a malignant epithelial tumour that can show intercellular bridges coming from the bronchial epithelium [63]. SQUAM is usually positioned in the central part of the lung or in the main airways (bronchi) (see Figure 1.3). Due to location symptoms can be cough, trouble in breathing, blood in the sputum and chest pain [63].

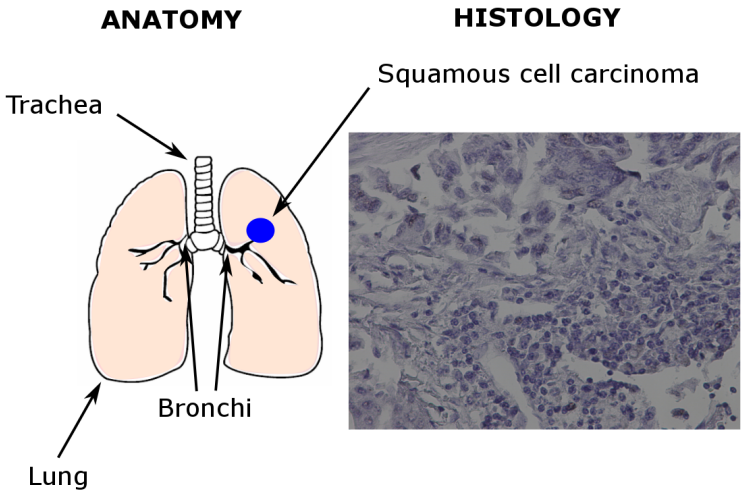


Figure 1.3: Histology and position in the lung of SQUAM. Histology figure provided by I. Gipanou [24].

SQUAM is the subtype of NSCLC that is more strongly associated with smoking: over 90% of SQUAMs occur in current or past smokers [63]. Other risk factors are age, genetic predisposition, exposures to passive smoke, mineral and metal dusts, asbestos and radon gas [63].

The TCGA consortium [44] analysed 178 SQUAMs by whole exome sequencing of cancer tissue. The mutational spectrum can be found in Figure 1.4. The most frequrntly mutated genes

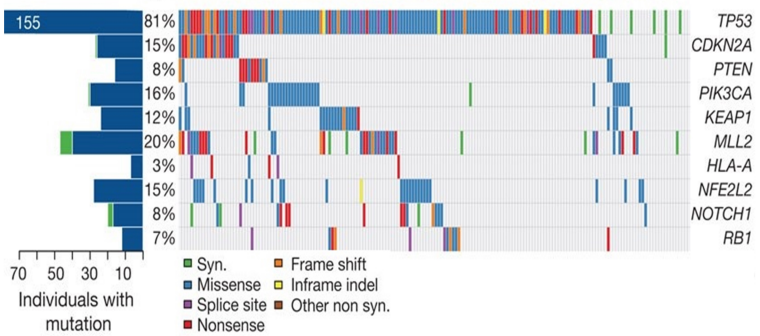


Figure 1.4: Significatly mutated genes of 178 SQUAMs. Columns represent patients, rows the analyzed genes. Figure taken from [44].

compared with normal tissue are: TP53, CDKN2A, PTEN and PIK3CA. Apart from TP53

all other genes differ from the mutated genes found for LADC (compare Figure 1.2). Looking at the columns of Figure 1.4 one can notice that the figure can be split in two groups: in the first group patients have a TP53 mutation and an additionally one (two mutational hits), in the second group patients have one or more mutations excluding TP53. For the first group this would mean that a TP53 mutation in necessary but not sufficient for SQUAM to arise, another mutation has to be present [43]. For the second group TP53 in not present, pointing to a different pathway to cancer.
This diversity in mutational spectra can be attributed to the fact that smoking damage is a complicated and extended process acting in different ways on the cells and on the DNA. The damage can hence be notably different from one patient to the other.

1.2 Exposure of biological tissue to ionising radiation

Radiation is defined as the energy emitted from a given source. Some examples are: sunlight, microwaves from an oven, medical X-rays and γ-rays from radioactive elements [1, 60]. In this Thesis we consider only ionising radiation (IR), which is radiation with sufficient energy to remove electrons from an atomic orbit, leaving the atom in a charged/ionized status. IR can be described either as particles or waves (particle wave dualism) [60].
In particulate radiation, consisting of (sub-)atomic (electrons, protons, etc.) the energy is carried in the form of kinetic energy or mass in motion. α-particles are ionising directly because they carry a charge and interact directly with the DNA when passing through a tissue [27].
The energy of electromagnetic radiation is carried by oscillating electrical/magnetic fields traveling through space with the speed of light. Forms of electromagnetic radiation are classified by wave length, especially X-rays and γ-rays of high energy are ionising. They are high energy waves with short wave length and high frequency. γ- and X-rays are considered as indirect IR because they are electrically neutral and do not interact with the DNA directly but mostly with the H_2O molecule producing free radicals, that then damage the DNA [27].
We consider only IR from γ-rays (atomic bombs of Hiroshima and Nagasaki, miner exposure) and α-particles radiation (miner exposure).
Independently from the type of IR, the damage caused to DNA are strand breaks and loss of base pairs [4]. The different types of DNA damage are represented in Figure 1.5, right panels.

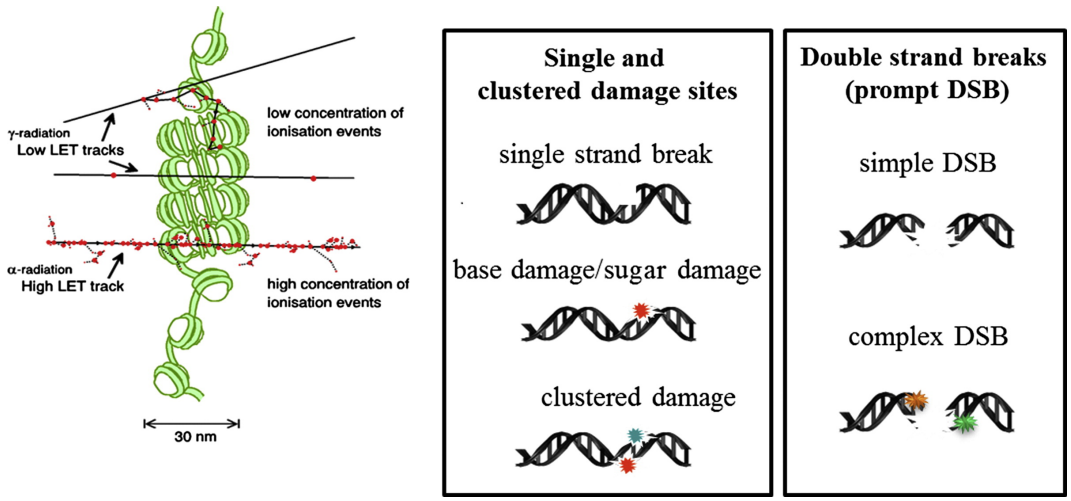


Figure 1.5: Left: Representation of DNA damage of low LET vs. high LET tracks. Right: Schematic representation of different types of strand breaks. Figure taken from [37].

Single strand breaks are breaks in only one strand of the DNA, the other one remains intact. The mechanism of DNA repair can normally reconstruct the original DNA in a very short time using the still intact part [4]. When double strand breaks occur, the repair of the DNA can be complicated and it is possible that wrong parts of DNA are glued together giving rise to the so called chromosomal aberration [4].

When IR comes into contact with tissue energy is transferred from particles/waves to the tissue itself. The release of energy along the covered path is called track. To quantify the energy transferred from IR per unit distance along a charged-particle track the concept of linear energy transfer (LET) was introduced [60]:

$$LET_{\Delta} = -\left(\frac{dE}{dx}\right)_{\Delta}, \tag{1.1}$$

where Δ is the maximal energy loss. LET_{Δ} is also called restricted stopping power.
γ-waves are characterized also for their low LET power, where α particles for their high LET one. This means that γ-waves produce spread damage but for a long track, where α particles cannot cover a long path but they release almost all energy in a restricted area creating huge local damage [32, 37]. A graphical representation of the DNA damage produce by high and low LET can be found in Figure 1.5, left panel.
Since the effect of γ rays and α particles to the tissue are physically and biologically different, this is a hint that carcinogenesis arising from this two types of radiation fields can be different for both LADC and SQUAM.

1.3 Exposure to cigarette smoke

Cigarette smoke is the main cause not only of lung cancer, but also of chronic lung and vascular diseases, and oral disease [35, 63]. Several toxins present in cigarette smoke have immunomodulatory effects. These components, together with other minor ones, induce chronic inflammation at the mucosal surfaces and modify the response of the host to external agents. One of the cigarette smoke components is also ^{210}Po, an α particle emitter [13, 28, 41]. Smokers can hence be considered exposed to ionizing radiation, although a dose quantification is difficult.
Reactive oxidative particles are produced in the burning cigarette, which inhalation can not be prevented by cigarette butt filters. These particles, also contained in the gaseous phase, are often short-lived, affecting primarily the epithelial cells lining in the upper airways inducing DNA damage [35]. Cigarette smoke components activate hence epithelial cell intracellular signaling cascades, leading to inflammatory processes (chronic immune cell recruitment and inflammation) [35]. Cigarette smoke can induced several toxic effects, particularly the induction of carcinogenesis as a result from direct genetic or epigenetic effects, denoted by altered gene functions (e.g. cell cycle, DNA repair, and tumor suppressor genes) [35].

1.4 Outline of the Thesis

The central question of this Thesis is the understanding of carcinogenesis and the related risks in lung adenocarcinomas and squamous cell carcinomas under the exposures of ionising radiation and cigarette smoke. To improve this understanding two different cohorts with information about these two lung carcinoma subtypes were analysed. The first one is the Japanese Life Span Study cohort of atomic bomb survivors of Hiroshima and Nagasaki. We only considered exposures to a mixed field of γ and neutron ionising radiation and to cigarette smoke. For about

40% of the cohort's members the smoking information were unknown and hence imputed. A complete descriptive analysis of this cohort can be found in Section 2.2. The second analysed cohort is the Eldorado cohort of mine workers. Most workers were uranium miners and mill workers employed at two mine sites (Port Radium and Beaverlodge, Canada) and workers employed at the radium and uranium refining and processing plant (Port Hope, Ontario). These people were exposed to radon gas, therefore to α particles and to γ waves. No information about cigarette smoke exposure is given for this cohort. A descriptive analysis of the Eldorado cohort can be found in Section 2.3.
The Life Span Study cohort was analysed using three different types of models

- state-of-the-art statistical risk models for radiation protection,
- generalised additive models, and
- biology-based mechanistic models.

A formal description of these three different methods can be found in Sections 3.3.2, 3.3.1 and 3.4, respectively. As we have seen in the paragraphs before, there are compelling reasons to consider lung adenocarcinoma and squamous cell carcinoma as different diseases. Therefore, for each subtype all three models were developed and compared. The effects of both cigarette smoke and ionizing radiation exposures were taken into account as influencing variables modifying the spontaneous cancer rate. Results and comments for the analysis of adenocarcinoma and squamous cell carcinoma can be found in Chapters 4 and 5, respectively.

From biology it is clear that smoking has a heavy impact in the development of lung cancers. Since for the Eldorado cohort no information about cigarette smoke exposure was provided, we decided not to apply biology-based mechanistic models to this cohort. Without this information no biological description of carcinogenesis can be provided. The models applied to this cohort were therefore

- state-of-the-art statistical risk models for radiation protection, and
- generalised additive models.

The effects on radiation risks estimates of α-particles and of γ waves were tested in different ways for both subtypes of lung cancer. Also for these analyses, results and comments for adenocarcinoma and squamous cell carcinoma can be found in Sections 6 and 7, respectively.

In Figure 1.6 the mind map of this Theses is represented. In the middle the central question: lung cancer risks in its subtypes. Left and right the two analysed cohorts with information about cigarette smoke exposure and/or ionising radiation exposure. This information were processed by the different models (yellow arrows), ending with new radiation and smoking risks for lung adenocarcinoma and small cell carcinoma (top and bottom central boxes).

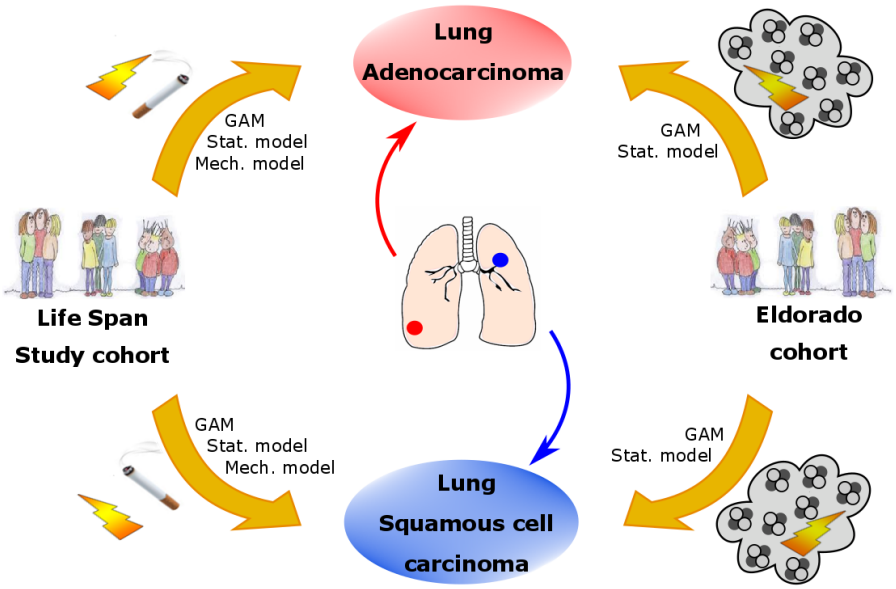


Figure 1.6: Outline of the Thesis. In the middle the central question: lung cancer risks in its subtypes. Left and right the two analysed cohorts with information about cigarette smoke exposure and/or ionising radiation exposure. This information were processed by the different models (yellow arrows), ending with new radiation and smoking risks for lung adenocarcinoma and small cell carcinoma (top and bottom central boxes).

Material and Methods

CHAPTER

2

THE COHORTS

This chapter in dedicated to the descriptive data analysis of the cohorts analysed in this Thesis. The first section describes the Japanese Life Span study cohort of atomic bomb survivors of Hiroshima and Nagasaki. Cohort members were both exposed mostly to a mixed field of neutron and γ waves and to cigarette smoke. Since for around 40% of the data the smoking information was missing, methods of imputation data were applied with both endpoints lung cancer (general) and squamous cell carcinoma. The importance of the endpoint by imputation processes will be demonstrated. A comparison of the original dataset with the imputed ones will be presented. Only results/analyses of imputed data are presented in this Thesis.
The second cohort taken into account is the Eldorado cohort, which is composed of only male mine workers of two mine sites (Port Radium and Beaverlodge, Canada) and workers employed at the radium and uranium refining and processing plant (Port Hope, Ontario). These workers were exposed to radon gas, therefore to α particles and to γ waves. No information about cigarette smoke exposure is given for this cohort. The definition of lagged dose will be introduced and explained. It will be shown how only from the raw data an association of LADC to γ waves and of SQUAM to α particles radiation exposure is visible.

2.1 Data shape

Epidemiologist investigate associations between exposure to possible detrimental risk factors and health effects in the general population. They use observational data of disease incidence characterised by covariables such as location, time and gender. This type of analyses can be described as a cohort design [23, p. 378]. In cohort studies people are recruited because their exposure status, independently of any knowledge about future disease risk, and followed up for disease occurrence after exposure. The cohort studies normally have a large number of participants in order to ensure enough cases for meaningful analysis [23, p. 203, p. 345].
For both cohorts observational data were available in grouped form as person-years tables, stratified buy different variables as, e.g., sex, city/facility, calendar year, age at exposure, age at end of followup, exposure intensity, smoking status. Person years (PYRs) quantify the time at risk which members have accumulated in a given stratum. A stratum does not correspond to an individual but is defined as a cell in the data space spanned by the covariables of the cohort [23, p. 701].
The following chapters give a detailed analysis of the two cohorts.

2.2 The Life Span Study of Japanese atomic bomb survivors

The Life Span Study cohort (LSS) of Japanese atomic bomb survivors has been the primary epidemiological basis for evaluating the long-term health effects of IR, dominated by γ rays of low LET in the range between 0-4 Gy. The LSS includes about 94,000 survivors who were in Hiroshima and Nagasaki at the time of bombing and about 27,000 people who were temporarily away from the cities at that time, and their mortality and cancer incidence have been followed up since 1950 and 1958, respectively [26, 45]. Cohort members experienced radiation exposure in all age groups and were not selected for pre-existing illness. The current analysis used the lung cancer incidence data [18, 21], for which lung cancer diagnoses and histological types were derived from a pathology review carried out during the follow-up period between 1958 and 1999 but no genotyping of cancer tissue was performed. The data used in this analysis consisted of a table of case counts and person years finely cross-classified by city, gender, radiation dose, follow-up period, attained age, age at exposure, and distance from the hypocenter.

2.2.1 Dosimetry

After the bombings of Hiroshima and Nagasaki studies have been performed, to collect data on survivor locations and shielding and to create systems to estimate individual doses from neutrons and γ rays [15].
Individual survivor dose estimation requires information on location and shielding at the time of the bombings. These data came mostly from detailed shielding histories obtained for proximal survivors and Master File cards created during the 1950s and 1960s for each person [15]. Like other methods of dosimetry for radiation epidemiology, atomic bomb survivor dosimetry assumes that a suitable measure of radiation for estimating quantitative relationships to health effects is the *absorbed dose in relevant tissues*, which is defined as the amount of energy deposited in the tissue from interactions of a specific type of ionizing radiation per unit mass of tissue. The unit is *grays* (1 Gy = 1 Jkg). To convert the rates at which neutrons or γ rays of a specified energy lose energy in, e.g., air or tissue, so called conversion factors are used [15].
Dose categories were defined in terms of weighted absorbed DS02 colon dose, which was calculated as the γ-ray dose plus 10 times the neutron dose in Gy, with additional adjustment to reduce bias in risk estimates due to the uncertainty involved in individual dose estimation.

A summary of the LSS can be found in Table 2.1.

Table 2.1: Summary of mean values for different covariables of the LSS cohort data broken down by gender.

	Women	**Men**	**Total**
Subjects	62515	42889	105404
Person years	1749254	1053713	2802967
Cases	731	1072	1803
LADC	321	315	636
SQUAM	76	254	330
other lung cancers	334	503	837
Case-weighted attained age (years)	71.78	70.61	71.09
Case-weighted age at radiation exposure (years)	34.63	35.64	35.23
PYRs-weighted age at radiation exposure (years)	24.11	20.73	22.84
Case-weighted radiation dose (Gy)	0.184	0.139	0.157
PYRs-weighted radiation dose (Gy)	0.096	0.101	0.098

2.2.2 Imputation of smoking information

While data on radiation dose were virtually complete in this cohort, data on smoking histories were available for about 60% of the members, which were mainly obtained from a series of mail surveys (at most three surveys for each individual) conducted between 1965 and 1991. The questionnaire at each data collection included a question about the smoking status (*never*, *current* or *past* categories). In addition, the age at starting smoking and the intensity (the average number of cigarettes smoked per day) were asked to ever-smokers and the age at quitting to past-smokers. For analyses, multiple survey sources of smoking data were combined and summarized as a relatively simple set of variables (start age, intensity, and quit age) to describe the smoking history for each individual having any information of smoking [18, 21]. In total, about 40% of the study subjects had no information on smoking habits due to non-response to all surveys, and the proportion of subjects with missing smoking data varied depending on sex, birth-year and radiation exposure [22]. In addition, the majority (about 60%) of the survey respondents had data from only one or two sources. In the earlier analyses [8, 18, 21, 26, 46], subjects with no smoking data were treated as having *unknown* smoking status throughout their time at risk. For those with smoking data, the smoking status during the period up to the first survey response was treated as 'unknown', and the status at the last response was carried forward to the end of follow-up. With a concern on potential impact of the incompleteness in smoking data, Furukawa et al. [22] applied a common missing data approach of multiple imputation (MI) (e.g., Sterne et al. [55]) to individual smoking histories in the LSS cohort and used the imputed data in analysis to evaluate the joint effects of radiation and smoking on LADC and SQUAM incidence. Two different sets of imputed data were used, each one composed of 50 imputed data sheets: one imputed with endpoint lung cancer and the other with endpoint SQUAM. A data set with a longer follow-up 1958-2009 could not be used since it lacked infor-

mation on histological types, and smoking imputation was not performed [8].
A comparison of the imputed data sheets vs. the original dataset containing missing values can be found in Table 2.2 (the row data used to calculated percentages can be found in Appendix A.1). In the following the original dataset containing missing information will be called original dataset.

Table 2.2: Comparison of cases after imputation (Imp) with the cases of the original dataset (OD) containing the category "unknown smoking information". Second and third columns summaries imputed data with endpoint lung for the lung cancer types lung in general and LADC, respectively. The last column is a summary of the imputed data with endpoint SQUAM for the subtype SQUAM.

	Lung total	**LADC**	**SQUAM**
Cases in OD	1803	636	330
% never smokers in OD	17	26	5
% ever smokers in OD	45	41	58
% cases without smoking info in OD	38	33	37
% imputed never smokers	44	35	50
% imputed ever smokers	35	31	35
$\frac{\textit{Never smokers OD}}{\textit{Ever smokers OD}}$	0.37	0.64	0.08
$\frac{\textit{Never smokers after Imp}}{\textit{Ever smokers after Imp}}$	0.43	0.69	0.11

The LSS has almost as double LADC cases than SQUAM (636 vs. 330 cases), where in the original dataset the amount of never smokers in much higher for the first subtype than for the second one (26% vs. 5%). This is a substantial difference between SQUAM and LADC that we have already seen in Chapter 1.1: a lot of LADC cases arise in never smokers, SQUAM is instead a smokers disease. For approximately 40% of the lung cancers no smoking information is available. A difference in percentage of imputed cases in never smokers can be observed between subtypes: lung and LADC have around 40% with no relevant difference between them, for SQUAM we have 50%. This discrepancy has to be attributed to the fact that the amount of never and ever smokers, and their smoking behaviours, differs a lot between subtypes. The last two rows of the table represent the shares of (n)ever smokers in the original dataset and, as average, in the imputed data sheets. Since the information are missing at random the shares before and after imputation should stay similar.
A summary of the LSS cohort data broken down by sex and smoking status for LADC (red) and SQUAM (blue) can be found in Table 2.3. The percentage of females never smokers in LADC case is approximately 5 time higher as in SQUAM cases, while for males there is almost no difference between subtypes. For current smokers a reversed behavoiur can be noticed: almost as double SQUAM cases than LADC cases in males, whereas almost no difference is noticeable for females. This is a substantial difference between SQUAM and LADC that we have already seen in Chapter 1.1: many LADC cases arise in never smokers, SQUAM is a smokers disease. For both subtypes the majority of never smokers were females. SQUAM cases are generally older than LADC cases with an increased smoking duration. Looking at the smoking variables *Age at begin smoking*, *Years since quitting* and *Smoking intensity* no relevant difference between LADC and SQAM can be noticed. SQUAM cases have a cumulative smoking amount higher than LADC cases, especially for female current smokers and male past smokers (26 vs 17 Packyears and 37 vs 32 Packyears, respectively). Since *Packyears* are calculated as the product *Smoking duration* and *Smoking intensity*, the diversity in smoking amount between SQUAM and LADC can be seen as the result of the combination of these two variables. The radiation

Table 2.3: Summary of mean values for age and exposure-related co-variables of the LSS cohort data broken down by smoking status and lung cancer subtype. For smoking related co-variables the means are taken over 50 data sheets with imputed smoking information, for LADC (in red) and SQUAM (in blue).

	Women			Men		
Smoking status	**Never**	**Past**	**Current**	**Never**	**Past**	**Current**
Cases (% of 636)	234 (37)	19 (3)	68 (11)	23 (4)	72 (11)	220 (35)
(% of 330)	24 (7)	14 (4)	38 (12)	9 (3)	48 (15)	197 (60)
Age at diagnosis (years)	68.1	72.0	70.4	71.1	68.9	67.7
	74.5	69.1	74.0	73.3	70.4	70.6
Age at begin	-	32.7	34.3	-	21.6	21.9
smoking (years)	-	29.0	32.0	-	21.7	22.1
Smoking	-	29.3	36.0	-	34.4	45.7
duration (years)	-	27.1	42.0	-	37.8	48.5
Years since	-	10.0	-	-	12.8	-
quitting (years)	-	13.0	-	-	10.5	-
Cumulative smoking	-	14.6	16.8	-	32.4	43.3
amount (pack-years)	-	16.6	26.2	-	37.5	44.9
Smoking intensity	-	9.1	9.3	-	19.2	19.4
(cigarettes/day)	-	10.9	12.3	-	20.1	18.9
Age at radiation	29.4	31.9	35.4	32.6	27.6	30.8
exposure (years)	38.7	32	38.6	42.7	33.1	37.7
Radiation dose (Gy)	0.194	0.085	0.213	0.041	0.113	0.143
	0.169	0.134	0.348	0.053	0.110	0.144

related variable *age at exposure* is much higher for SQUAM cases than for LADC cases with a relevant increase in dose, especially for past and current female smokers.

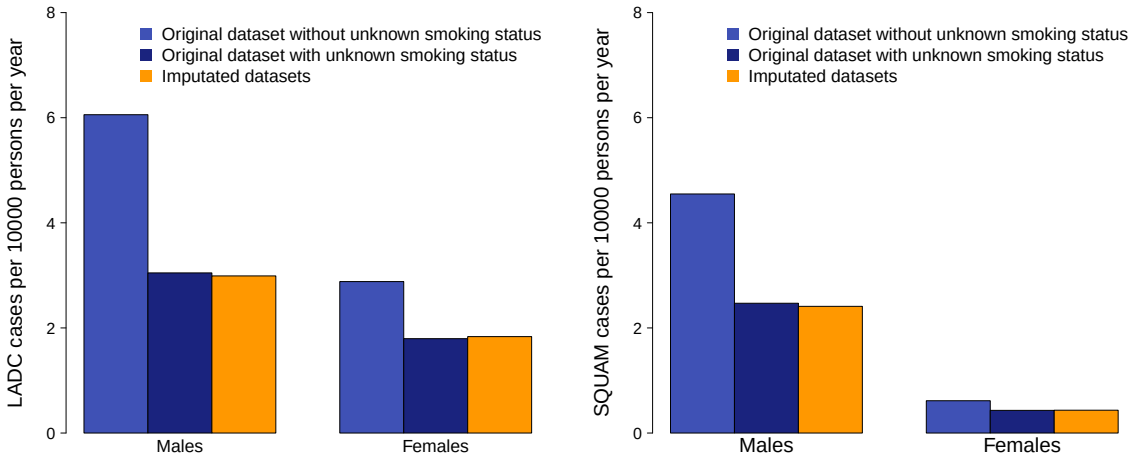
In the following the variation between data sheets and the difference between data sheet and original dataset will be analysed.

2.2.3 Original vs. imputed data sets

As we have seen in the section before, two different imputation data sets have been created for the analysis of LADC and SQUAM in the LSS. In this chapter we want first to analyse the differences between imputated and original data set (with and without unknown smoking information) and then we will proceed with the analysis of the variation between the 50 data sheets, variation in imputation outcome.

All previous studies that until now dealt with effects of smoking and radiation exposures to lung cancer in the LSS considered the unknown data as a special category, extrapolating in this way risk estimates for these two external effects [8, 18, 21]. Furukawa et al. [22] however showed that the best way to analyse this data is either to discharge the unknown data or to use imputed data sets. In the next figures we address this problem analysing the different data sets.
Form Figure 2.1 we can see that the deletion of unknown data would introduce a bias in the data: the crude rates of the reduced dataset (light blue) are in each category higher than that of the complete dataset and of the imputed data sheets. The crude rates of imputed and com-

plete data sets are almost identical, since the number of cases remains the same. The small difference has to be attributed to the small variation in the person years during the imputation process. The biggest bias can be found for both subtypes in the male category, indicating that the majority of the deleted data are healthy males. A higher crude rate denotes a deletion of person years but not of cases.



(a) LADC (b) SQUAM

Figure 2.1: Cases per 10^4 *persons per year for LADC (a) and SQUAM (b) by gender comparing the three data sets: the original data set without unknown smoking information (light blue), the original data set with unknown smoking information (dark blue) and the 50 imputed data sheets (in orange). The orange bar is the person years weighted mean over the 50 imputed datasets.*

Until now we just had a look to the difference between the original data set and the imputed data sheets, but we did not consider the variation between the imputed data sheets. In Figures 2.2 and 2.3 the huge bias between reduced data set and the imputed ones is detectable in each category. The variation of the related deviation from original value is relatively small and is maximal for SQUAM never smokers, both sexes. This behaviour has to be attributed to the fact that for SQUAM the amount of never smokers is small. The statistical power is therefore low causing less stability for imputation procedures.
Similar analysis were done for the imputed variables

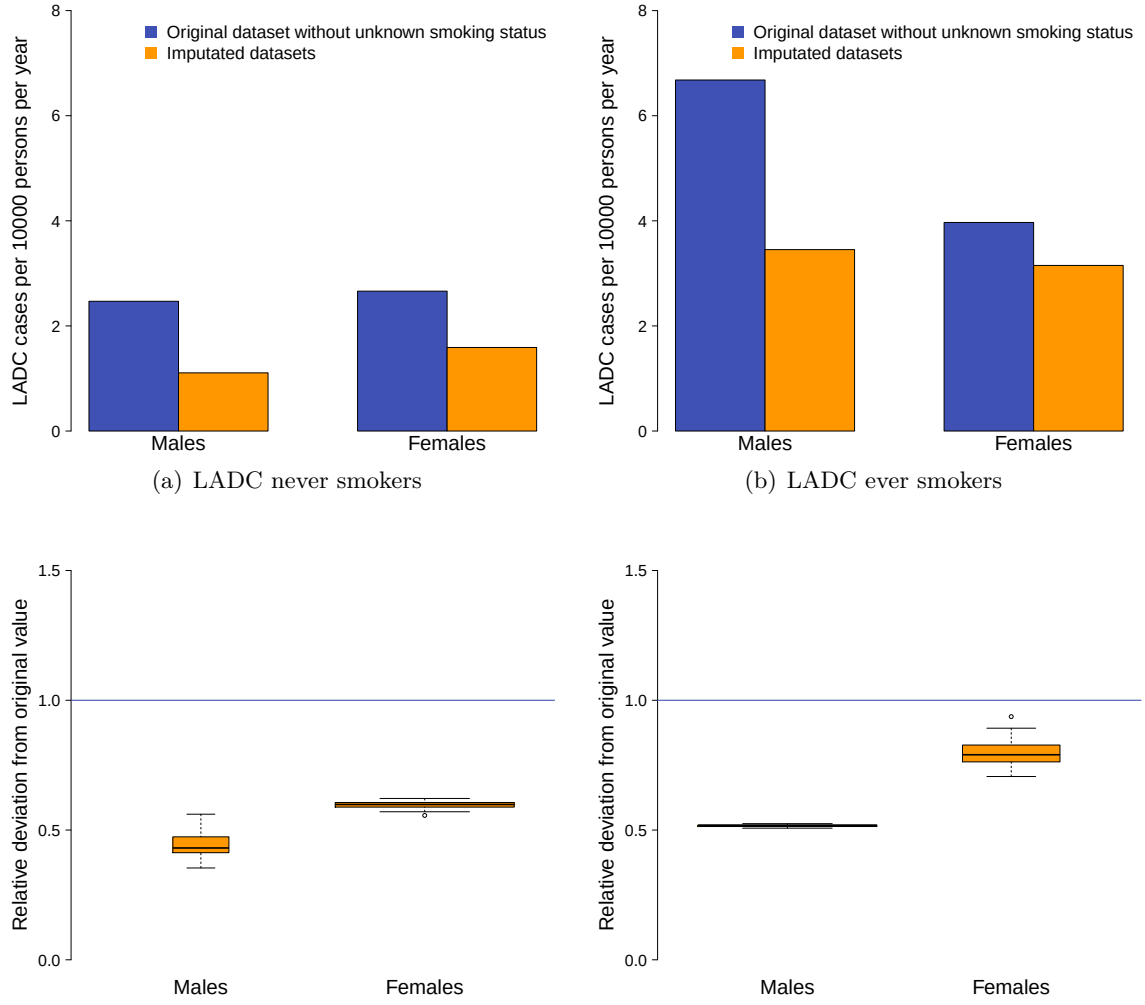
- smoking status (never, current and past smoker),
- smoking duration,
- smoking intensity (cigs/day), and
- years since quitting

and for the not imputed variables, split by imputed smoking status (never and ever smoker categories as the sum of past and current smokers)

- attained age, and
- lung dose (Gy).

Graphical representations can be found in Appendix A. Neither substantial differences between LADC and SQUAM nor major problems could be found in the data sets.

Furukawa et al. [22] addressed the problematic of handling the unknown smoking category. He could show that the best results can be obtained analysing imputed data. This is in line with the results from our above analysis. The analysis of LSS data will hence be done only with imputed data.



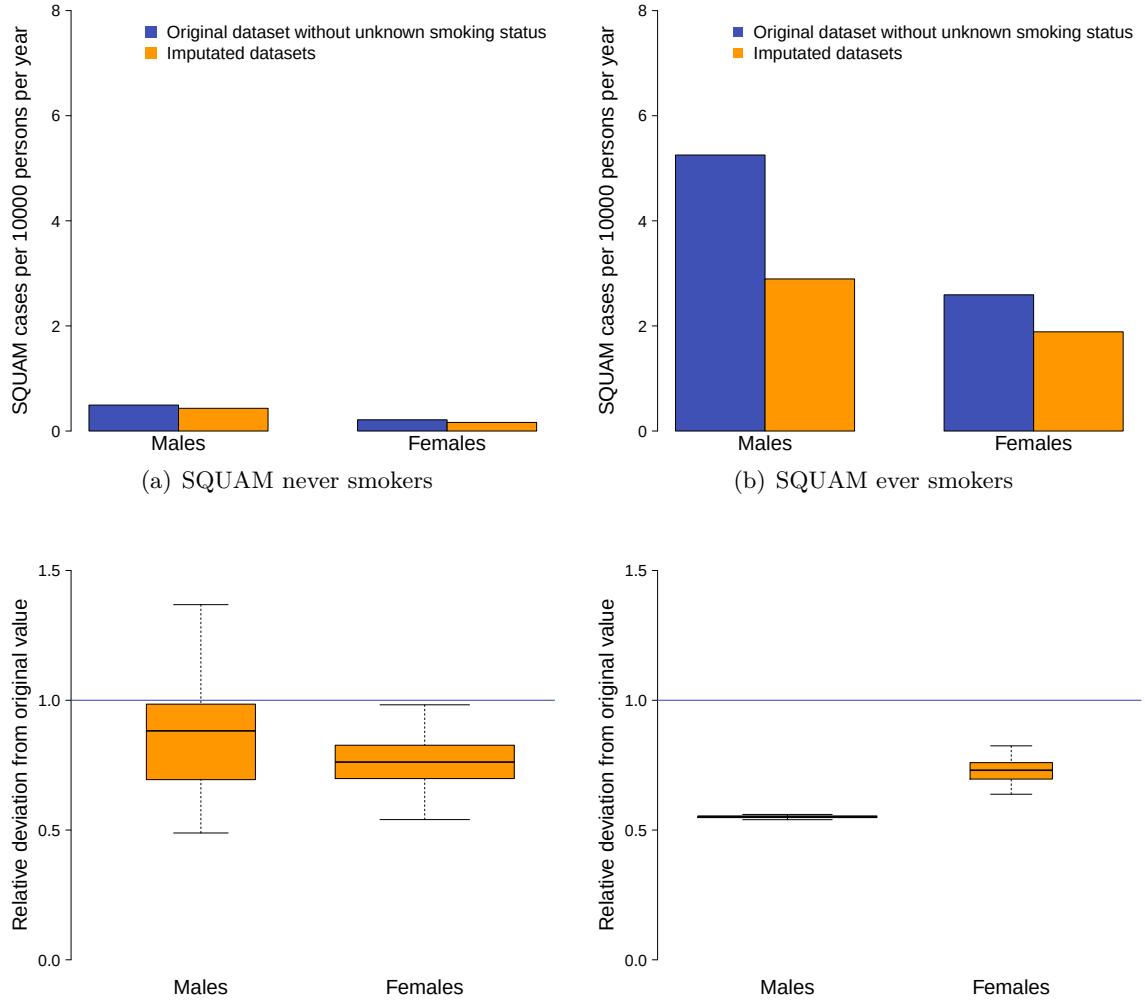
(a) LADC never smokers

(b) LADC ever smokers

(c) LADC never smokers

(d) LADC ever smokers

Figure 2.2: Cases per 10^4 *persons per year for LADC by gender and by smoking status (never smokers left panels and ever smokers right panels) comparing two data sets: the original dataset without unknown smoking information (light blue) and the 50 imputed data sheets (in orange). Panels (a) and (b), for never and ever smokers respectively, describe the difference of the incidence between the original dataset and the 50 imputed data sheets. The orange bars are the person years weighted mean over the 50 imputed data sheets. Panels (c) and (d) represent the relative deviation of the imputed data sheets from the original dataset. The relative deviation is calculated as incidence-ratio between original and imputed dataset. The boxplots represent the distribution of the relative deviations of the 50 imputed data sheets. The blue horizontal line represents the value one, it means a correspondence between original and imputed dataset.*



(a) SQUAM never smokers

(b) SQUAM ever smokers

(c) SQUAM never smokers

(d) SQUAM ever smokers

Figure 2.3: Cases per 10^4 *persons per year for SQUAM by gender and by smoking status (never smokers left panels and ever smokers right panels) comparing the two data sets: the original dataset without unknown smoking information (light blue) and the 50 imputed data sheets (in orange). Panels (a) and (b), for never and ever smokers respectively, describe the difference of the incidence between the original dataset and the 50 imputed data sets. The orange bars are the person years weighted mean over the 50 imputed data sheets. Panels (c) and (d) represent the relative deviation of the imputed dataset from the original dataset. The relative deviation is calculated as incidence-ratio between original and imputed dataset. The box-plots represent the distribution of the relative deviations of the 50 imputed data sheets. The blue horizontal line represents the value one, it means a correspondence between original and imputed dataset.*

2.3 The Eldorado cohort

The 17,660 male subjects of the Eldorado cohort were collected from the personnel records provided by the mines and processing sites operated by Eldorado Nuclear Ltd. The Eldorado Nuclear Ltd company was started in 1927 as Eldorado Gold Mines Limited to develop a gold mine in Manitoba. In 1930 radioactive deposits at Great Bear Lake were found. The Eldorado Mine at Port Radium was therefore developed. In 1933 a state-of-the-art refinery was built in Port Hope, Ontario. Between 1933 and 1940 radium was produced, together with silver, copper, and uranium salts. The mine at Port Radium was reopened in 1942 to supply the United States military with uranium products. The company was taken over by the Canadian Government in 1943, and in early 1944 the name was changed to Eldorado Mining and Refining Limited. With the discovery of the Port Radium deposits, the Beaverlodge Mine at Uranium City, Saskatchewan, was opened entering production in 1953. In the 1960s the Unated States military stopped purchasing of Canadian uranium ores for the purpose of atomic weapons, and from then on uranium was produced for power plants. The company was dismantled in 1988 and merged with assets of the Saskatchewan Mining Development Corporation (SMDC) to become Cameco Corporation [33].
The major part of the workers were uranium miners or mill workers employed at the two mine sites Port Radium (Northwest Territories, Canada) and Beaverlodge (northern Alberta, Canada) and workers employed at the radium and uranium refining/processing plant in Port Hope (Ontario) [33]. A small part of subjects were employed in "other sites" as for example head officers, aviators and researchers. The followup of the data analysed in this Thesis started in 1969 and ended in 1999 [33]. To be included in the study a subject had to fulfill the following characteristics:

- age at employment between 15 and 75 years,
- year at first employment between 1932 and 1980,
- end of followup after 1940, and
- being alive in year 1969.

People of all three facilities were exposed to both α and γ radiation. The α exposure derived from radon decay products and uranium. γ rays normally follow after α decay. Exposure to cigarette smoke is recorded for this cohort.

The individual annual exposures for this cohort is calculated in working level month (WLM) [68]. To calculate WLMs first a working level is calculated. A working level is the concentration of radon decay products per liter of air that would result in the ultimate release of $1.3 \cdot 105$ MeV of potential α-particle energy. One WLM is equivalent to one working month (170 h) in a concentration of 1 working level [68].

The exposure intensity differs between facilities: people in Port Hope were mostly subjected to γ radiation with very low α dose, all other facilities had and opposite major exposure to α with low γ radiation. The correlation of the different exposures for the three facilities are presented in Figure 2.4. A summary of the Eldorado cohort can be found in Table 2.4.
Please note that in previous analysis (e.g. [33]), where statistical models were applied, only lagged dose exposures were taken into account. A graphical description of lagged and non-lagged dose exposure is given in Figure 2.5. Let us consider the case in which a worker started working in a facility at age *Age at first employment* and stopped working at age *Age at last employment*.

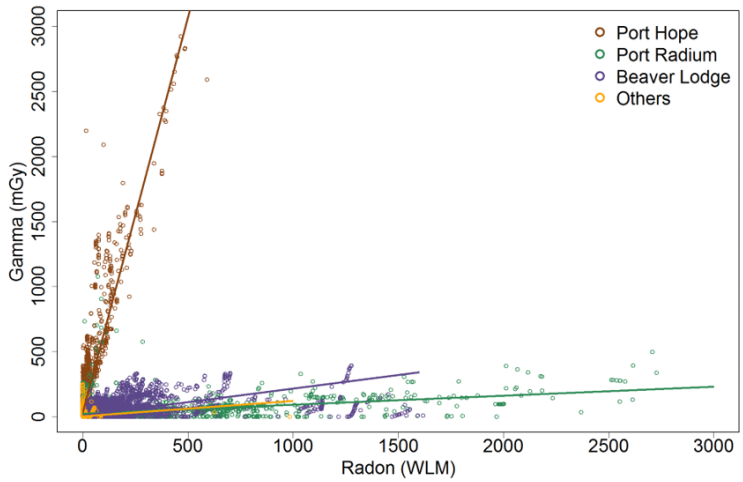


Figure 2.4: Gamma5 exposure as a function of Radon5 exposure in the Eldorado cohort differentiating between the different facilities: Port Hope (brown), Port Radium (green), Beaver Lodge (purple) and others (orange). The coloured lines represent linear regressions with γ intensity (mGy) as command variable and α intensity (WLM) as causing variable for the respective coloured facility.

Only if the end of followup happens in orange-marked age-interval we have a difference between lagged and non-lagged cumulative dose, that can be appreciated in the orange-marked region. The same plot can be done for γ radiation exposure. The idea behind a lagged dose is the assumption that any kind of exposure needs time in order to change the exposed environment and since descriptive models do no take into account any biological development, the latent period has to be introduce per hand [23, p. 673]. The impact between lagged and non-lagged dose in this data is really small (see Figure B.1).

As seen in Chapter 1.1, γ and α exposures damage tissues in different ways: γ waves go deep in the tissue while α particle stay on the surface. With this knowledge it is interesting to see if this association is visible in the raw data of this special cohort, where both exposures are known. The graphical representation of this association is presented in Figure 2.6. Cases (LADC in red and SQUAM in blue) are represented in dependency of both radiation exposures. A tendency of LADC to develop with γ and of SQUAM with α exposure is clearly identifiable. Since the majority of people were exposed to α particles (cf. Figure 2.4) it is clear that in this cohort the predominant lung cancer subtype is SQUAM (see Figure 2.7). In this case it is interesting to see how for both subtypes the raw rate decreases for higher age (cf. Figures A.2 and A.3 for LSS LADC and SQUAM, respectively). This fact can be attributed to a possible harvesting effect from an unknown smoking behaviour. Cases in smokers arise in younger ages, while older cases are not affected by smoking. The increment only for younger ages determines the decreasing behaviour. Since smoking information is not given for this cohort, only statistical models will be applied. A biological approach would not be reasonable without information about this important cause of lung cancer.

Different variables of the Eldorado cohort were analysed before model application. The results can be found in Appendix B.
Model results for to this cohort can be found in Chapters 6 and 7 for LADC and SQUAM, respectively.

Table 2.4: Summary of mean values for different covariables of the Eldorado cohort.

	Total	LADC	SQUAM
Subjects	15348	102	199
Person years	367079	1944	3469
Cases	593	102	199
Case-weighted attained age (years)	62.85	62.52	63.23
Case-weighted γ radiation dose (mGy)	56.23	76.37	48.99
PYRs-weighted γ radiation dose (mGy)	23.95	73.06	41.12
Case-weighted α radiation dose (WLM)	163.18	116.43	202.83
PYRs-weighted α radiation dose (WLM)	39.63	96.96	188.99

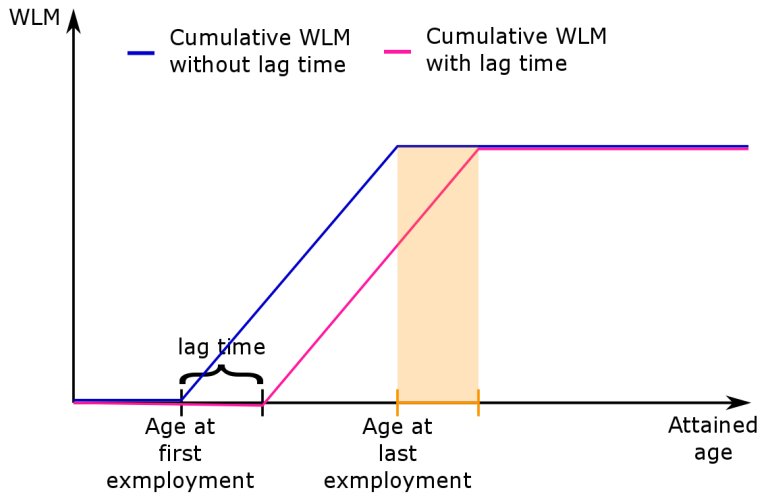


Figure 2.5: Description of the accumulation of WLM over age for a worker that started working in a facility at age Age at first employment and stops working at age Age at last employment. The blue(pink) line represents the cumulative WLM for this worker without(with) lag time. Only if the end of followup happens in orange-marked age-interval we have a difference between lagged and non-lagged cumulative dose, that can be appreciated in the orange-marked region.

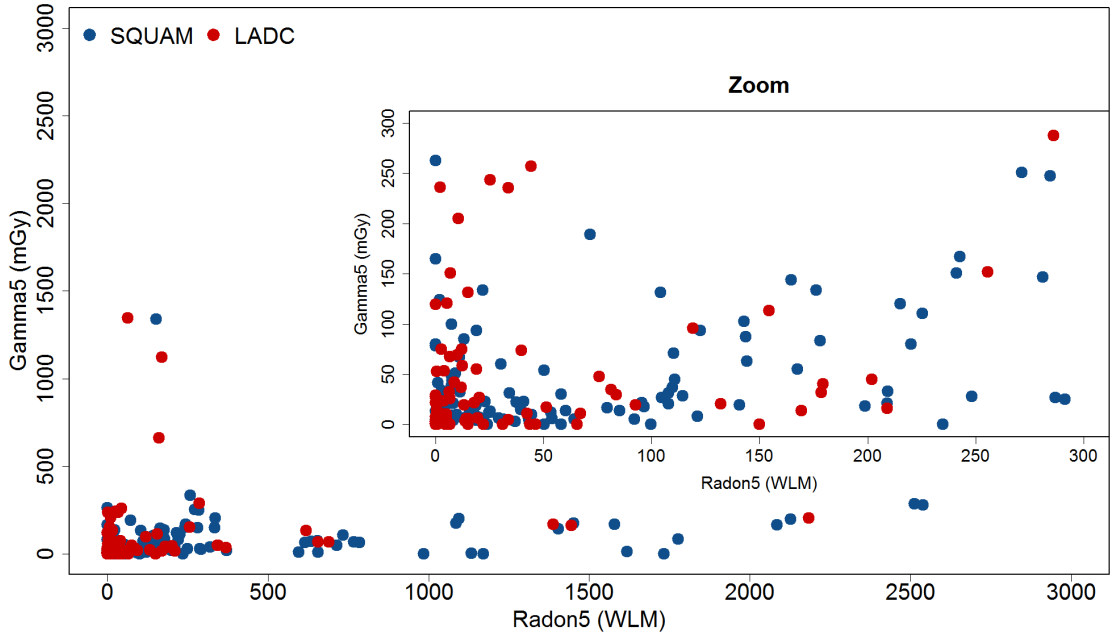


Figure 2.6: Incidence of the Eldorado cohort in relation to the variables Gamma5 and Radon5. LADCs are represented in red and SQUAMs in blue.

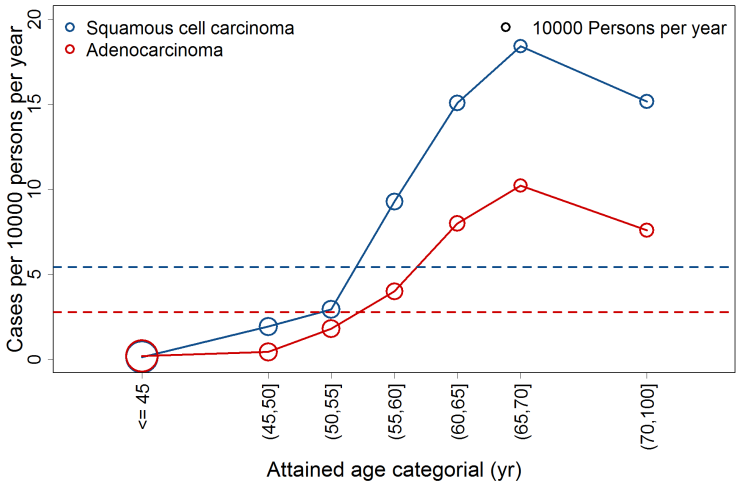


Figure 2.7: Age distribution of cases per 10^4 persons per year for LADC (red) and SQUAM (blue). The dashed lines represent the weighted mean values of the corresponding cancer types.

CHAPTER

3

METHODS OF DATA ANALYSIS

This chapter is dedicated to the models used for the analysis of Life Span Study and Eldorado cohort. The first part deals with epidemiology and statistical methods to compare models, while the second part is about the actual models used to fit the data.
The chapter starts with definitions of the survival and the hazard function. With these definitions Poisson regression is introduced, which allows us to test different models against the data. After a brief introduction of this regression method, the formula for the Poisson deviance will be derived. It is a measure for the discrepancy between the fitted and the real values of the response variable. Next, the definition of the Akaike's Information Criterion will be introduced, to allow us for comparison of non-nested models.
Having obtained the instruments to evaluate models, next the models themselves will be introduced. Two different classes of models are applied in this Thesis. Members of the first class we call "statistical models", as understanding and quantification of covariate effects to the outcome are established purely by statistical association. In contrast, the "biologically based molecular models" take into account biological knowledge for definition and covariate analysis. Within the class of statistical models, standard generalized additive models and explicit risk models that represent the state-of-the-art in radiation protection will be presented. Within the biologically based models the two stage clonal expansion model and the three stage clonal expansion model will be introduced. In the first model normal stem cells undergo two transformations to become malignant, in the second model three transformations.

3.1 Concepts of survival analysis

The motivation for this Thesis is to improve the understanding of carcinogenesis and the related risks in LADC and SQUAM after exposure to ionising radiation and cigarette smoke. One approach to this question involves the analysis of time periods from birth and exposures to a potential final event (LADC or SQUAM), and can be treated statistically within *Survival analysis* [23, p. 891].

Let us consider consider the non-negative random variables T_i as a description of the times at which given final events k_i occur. The distribution of T_i can be mathematically described by the survival function $S(t)$ [23, p. 893].

Definition 3.1.1: (Survival function)
The survival function S(t) describes the probability that at a given time t the investigated event k_i has not yet occurred

$$S_i(t) = P(T_i > t) \tag{3.1}$$

Of interest is also the probability that an event k_i occurs in a given time interval $t \leq T_i < t+\Delta t$. This probability can be described with the so called hazard function $h(t)$ [23, p. 893].

Definition 3.1.2: (Hazard function)
Let us consider the non-negative random variables T_i describing the time at which given events k_i occur. The hazard function or also failure rate $h(t)$ describes the probability that defined events k_i to occur in the determined time interval $[t, t + \Delta t]$

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \ \leq \ T \ < \ t + \Delta t \mid T \ \geq \ t)}{\Delta t}. \tag{3.2}$$

It is important to notice that the hazard function can be written as a function of the survival function.

Theorem 3.1.3:
Let us consider the non-negative random variables T_i describing the time at which given events k_i occur with corresponding survival function $S(t)$ and hazard function $h(t)$. The hazard function $h(t)$ can hence be written as a function of the survival function S(t)

$$h(t) = -\frac{d}{dt}\ln(S(t)). \tag{3.3}$$

The proof can be found in Appendix C.1.

Finally we are interested in numbers of people becoming sick, which is a count variable for which a Poisson distribution is assumed. In the next section a detailed introduction of Poisson regression is given.

3.2 Poisson regression

Let us consider observations $y_1, ..., y_n$ for the respective independently Poisson distributed response variables $Y_1, ..., Y_n$ with means $\mu_1, ..., \mu_n$ [23]. The probability of observing y_i observations in an interval is given by the equation

$$P(y_i|\mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y!}. \tag{3.4}$$

Of interest is the estimation of λ_i with different regression models, which can be generally described by the following form

$$\mu_i = \eta_i(\beta) = \eta(x_{1i}, ..., x_{ki}; \beta_1, ..., \beta_k), \tag{3.5}$$

where η is some regression function with regression parameters $\beta_1, ..., \beta_k$, with each component relating values $x_{i1}, ..., x_{ik}$ of explanatory variables to respective means [23].

Using equations (3.4) and (3.5) the log likelihood ℓ is thus given by the following equation [23]

$$\ell_y(\beta) = \sum_{i=1}^{n} y_i \ln[\eta_i(\beta)] - \eta_i(\beta) - \ln(y_i!) \tag{3.6}$$

The estimation of the model parameters $\beta_1, ..., \beta_k$ can hence be performed by the maximization of the (log)likelihood function. The statistic used in this Thesis to quantify the support of the data to a particular model is the so called deviance [23].

Definition 3.2.1: (Deviance in a model)
Let us consider model (3.5) with log likelihood (3.6). As a measure of distance between the model with the most likely parameters $\hat{\beta}$ and the model described by the "real" parameters β, the deviance is given as minus two times the logarithm of the normed likelihood L [23]

$$D_y(\beta) = -2 \ln \left(\frac{L_y(\beta)}{L_y(\hat{\beta})} \right) = -2(\ell_y(\beta) - \ell_y(\hat{\beta})) \tag{3.7}$$

$$= -2 \sum_{i=1}^{n} y_i \ln \left(\frac{\eta_i(\beta)}{\eta_i(\hat{\beta})} \right) - \eta_i(\beta) + \eta_i(\hat{\beta}) \tag{3.8}$$

To have a measure of goodness of fit comparable to the residual sum of squares in normal linear regression, the likelihood of the "real" model that perfectly fits the data can be compared to the likelihood of model under the other following definition of deviance

$$dev_y(\hat{\beta}) = -2 \sum_{i=1}^{n} y_i \ln \left(\frac{y_i}{\eta_i(\hat{\beta})} \right) - y_i + \eta_i(\hat{\beta}), \tag{3.9}$$

where y_i are the observations, here the number of people suffering from LADC or SQUAM [23], and η_i are the predicted cases. Since our data are given in a stratified form, the following relation between prediction and person years is given

$$\eta_i(\hat{\beta}) = h_i(\hat{\beta}) \cdot PYRs_i, \tag{3.10}$$

where h_i is the predicted hazard and $PYRs_i$ the person years of each stratum.

During data analysis a lot of models have to be compared to each other in order to find the "best" one, a parsimonious and well fitting model. The deviance is an appropriate statistic to compare nested hierarchical models. Let us consider two models. The first one may have the parameter set $\tilde{\beta} = [\beta_1, ..., \beta_n, r_1, ..., r_m]$. The other one may be "smaller", in the sense that it equals the first one but with m parameters set to zero $\zeta = [\beta_1, ..., \beta_n, 0, ..., 0]$. Let us assume that the small model is the correct one, then the difference in deviance is χ^2 distributed with m degrees of freedom. This is the so called likelihood ratio test with the null hypothesis for the smaller model against the larger model as the alternative [23]. We consider an improvement significant if the smaller model can be excluded on a 95% confidence interval.

To compare non-nested models the **A**kaike's **I**nformation **C**riterion (AIC) can be used. Models with a smaller AIC are preferred [47].

Definition 3.2.2: (AIC - Akaike's Information Criterion)
Let us consider a model η as defined in (3.5) with deviance dev (3.9) and k parameters. The Akaike's information criterion (AIC) is defined by

$$AIC = dev + 2k. \tag{3.11}$$

Having introduced the Poisson regression, the remaining part of this chapter is devoted to different models, instances of equation (3.5).

3.3 Statistical models

Two type of statistical models are described in this section. In this context the term "statistical model" is intended to underline the fact that understanding and quantification of covariate effects to the outcome are established purely by statistical association. In contrast, in the next sections biologically based molecular models will be presented. The difference is that for the latter models model definition and covariate analysis take into account biological knowledge.

3.3.1 The generalized additive models

This section is dedicated to the introduction of *Generalized Additive Models* (GAM). To have a better understanding of this type of models we will shortly discuss *additive models* and their characteristics. The peculiarity of GAMs is the usage of nonparametric functions (smoothing splines) that can model patterns of covariables that would remain hidden with simple parametrisation.

Let us consider the following Linear Model (LM)

$$h_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i \tag{3.12}$$

with the linear predictor

$$\zeta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} = x_i^T \beta, \tag{3.13}$$

where h_i denotes the counting response variable and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ are coefficients of the covariates $x_{1i}, ..., x_{pi}$ with i.i.d. $\epsilon_i \sim N(0, \sigma^2)$ as error term.
The class of *additive models* is an extension of LMs, equation (3.12), where the linear predictor (3.13) is generalized using smooth functions $f(\cdot)$

$$h_i = f_1(z_{1i}) + \ldots + f_q(z_{qi}) + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i, \tag{3.14}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ is a parameter vector of covariates assumed to have a linear effect [64]. Different smooth functions will be explained in the following.
A peculiarity of LM, and hence of *additive models*, is the additivity of the effects of the predictor. Due to assumptions of additivity and independence of the covariates, the effects of the predictor can be investigated separately. Holding all but one predictor fixed, the variation of the fitted response is independent of the other predictors. The functions of the covariates can hence be analysed individually [64].

GAMs are additive models described by equation (3.14) with the response following any distribution from the exponential family [64].

3.3.1.1 Smoothing splines

The assumption that the relation between response and covariate is not linear, can be tested with nonparametric functions. Since nonparametric functions per definition do not have a rigid form of dependence between response h_i and covariates $x_{1i}, \ldots, x_{pi}$, some smoothing techniques have to be introduced. This section is based on [64] and [29].

The first technique we introduce in this Thesis are **Polynomial Splines**.
Let us assume that a given data is given in the form of (h_i, x_i), $i = 1, \ldots, p$, where h_i are the observations of the response variable and x_i are the corresponding metric covariates. Let us also assume that the response variable can be described by a function $f(\cdot)$ and an error term ϵ_i

$$h_i = f(x_i) + \epsilon_i. \tag{3.15}$$

The first idea is to approximate $f(x_i)$ with a polynomial function

$$f(x_i) = \gamma_0 + \gamma_1 x_i + \ldots + \gamma_l x_i^l, \tag{3.16}$$

where $l \in \mathbb{N}$ and $\gamma_k \in \mathbb{R}$, $k \in \{0, \ldots, l\}$. The problem with a pure polynomial approach is that if the polynomials have low degrees, the "true" relation of the data might not be sufficiently explained. The other way around, if the degree is high, the model fit will be too wiggly.

Fahrmeir et al. [20] applied the idea to divide the codomain into m parts $k_0 < \cdots < k_m$ and approximate $f(x_{ij})$ in each interval $[k_j, k_{j+1})$, $j \in \{0, \ldots, m-1\}$ with l-th degree polynomials. However, the resulting piecewise estimated $f(x_{ij})$ is not necessary continuous.

B(asic)-splines can be introduced as a construction to guarantee that piecewise estimated functions on knots $k_1, ..., k_{m-1}$ are composed in a sufficient, $(l-1)$-times differentiable way. The smoothing function $f(.)$ can be estimated with B-splines so that $y = f(x) + \epsilon$ becomes a linear model. This is done by choosing specific *basic functions* $B_1(x), \ldots, B_d(x)$, $d = m + l - 1$, in formula

$$f(x) = \sum_{j=1}^{d} \gamma_j B_j(x), \tag{3.17}$$

where γ_j, $j = 1, ..., d$ are the coefficients [20].

B-splines are defined as non-zero functions only on a few intervals $[k_i, k_p]$, i,$p \in 0, \ldots, m$, with $i \neq p$. We can hence rewrite equation (3.15) as the following linear model

$$y = X\gamma + \epsilon, \tag{3.18}$$

with,

$$X = \begin{bmatrix} B_1(x_{12}) & \ldots & B_d(x_{12}) \\ \vdots & \ldots & \vdots \\ B_1(x_{(N_V-1)N_V}) & \ldots & B_d(x_{(N_V-1)N_V}) \end{bmatrix}, \gamma = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_d, \end{bmatrix}$$

where $y = (y_{12}, \ldots, y_{(N_V-1)N_V})^T$ and $\epsilon = (\epsilon_{12}, \ldots, \epsilon_{(N_V-1)N_V})^T$.

The parameter vector γ can hence be estimated by the ordinary least square method

$$\hat{\gamma} = (X'X)^{-1}X'y.$$

The parameter vector γ can not be interpreted in a reasonable way, but the form of the estimated function $\hat{f}(.)$,

$$\hat{f}(x) = B\hat{\gamma},$$

where $B = (B_1(x), \ldots, B_d(x))$, which is a result of $\hat{\gamma}$, can be interpreted.
With B-splines we now reached continuous and differentiable functions. The problem of $f(x_{ij})$ to be wiggly unfortunately persists [20].

The problem of having wiggly $f(x_{ij})$ functions can be bypassed with **P(enalized) splines**. The difference in applying P-splines is that instead of minimizing the deviance function $dev_y(\hat{\beta})$ (3.9), we minimize the following equation

$$dev_y(\hat{\beta}) + \lambda \int_C f''(x)^2 dx \tag{3.19}$$

with respect to $\hat{\beta}$. C represents the codomain of x and $f''(x)$ is the second derivative of function $f(x)$. Since the second derivative of a function yields information of a curvature of the function, the minimization of the second derivative penalizes models that are too wiggly. In order not to loose the balance between model's fit and smoothness the parameter λ can be adapted: $\lambda = 0$ correspond to a model without penalization, $\lambda \to \infty$ instead, leads to a linear regression of the data [20].

In this Thesis we want to use deviance and AIC to compare different models. These statistics are related only to likelihood and number of model parameters. P-splines additionally minimise the second derivative of the function. To compare models with and without penalization, for penalized models first a maximum likelihood estimate $\hat{\beta}$ from equation (3.19) is calculated. Then the evaluated deviance function $dev_y(\hat{\beta})$ (3.9) at the point $\hat{\beta}$ is given. This value is comparable with the deviance function $dev_y(\hat{\beta})$ (3.9) of non-penalized models.

3.3.2 State-of-the-art statistical risk models for radiation protection

State-of-the-art statistical risk models for radiation protection are explicit models commonly used in radiation epidemiology. They were developed from Preston [48] and are much more similar to models used in *Cox regression*, but with a flexible baseline rate [23]. Radiation effect is assumed to follow an **Excess Relative Risk** model (ERR)

$$h(t, z, d) = h_0(t, z)(1 + \rho(d)\epsilon(t, z)) \tag{3.20}$$

or an **Excess Additive Risk** model (EAR)

$$h(t, z, d) = h_0(t, z) + \rho(d)\epsilon(t, z), \tag{3.21}$$

where t denotes *attained age* and other functions of time, z is a vector of dose independent covariables and d is a vector of covariables describing the exposure. The function $h_0(\cdot)$ describes the background rates and is called baseline hazard, where $\rho(d)$ is the dose response function with effect modification $\epsilon(\cdot)$. In the models applied in Chapters 4 to 7 the general forms for $h_0(t, z)$, $\rho(d)$ and $\epsilon(t, z)$ will be

$$h_0(t, z) = e^{f_0(\Sigma_1, \ldots, \Sigma_n, z_{1i}, \ldots, z_{ni})} \tag{3.22}$$

$$\rho(d) = \gamma_0 \cdot d \tag{3.23}$$

$$\epsilon(t, z) = e^{f_1(\gamma_1, \ldots, \gamma_n, z_{1i}, \ldots, z_{ni})} \tag{3.24}$$

with parameter vector $\beta = [\Sigma_1, ..., \Sigma_n, \gamma_0, ..., \gamma_n]$ and covariates $z_{1i}, ..., z_{ni}$. Analogously to the model names, parameter γ_0 is called *excess relative risk* in ERRs and *excess absolute rate* per unit dose in EARs.

Since the LSS includes smoking-related information, risk models need to take into account also cigarette smoke exposure. Therefore Furukawa et al. [21] extended this kind of models and introduced also a term to describe synergistic effects of radiation and smoking. Equations (3.20) became the so called *Simple Multiplicative Model*

$$h(t, z, d) = h_0(t, z)(1 + \rho(d)\epsilon(t, z))(1 + \phi(smk)\kappa(t, z)) \tag{3.25}$$

and (3.21) the so called *Simple Additive model*

$$h(t, z, d) = h_0(t, z) + \rho(d)\epsilon(t, z) + \phi(smk)\kappa(t, z), \tag{3.26}$$

where $\phi(smk)$ is the smoking-response function with effect modification $\kappa(\cdot)$. In the models applied in Chapters 4 to 7 the general forms for $\phi(smk)$ and $\kappa(\cdot)$ will be

$$\phi(smk) = \phi_0 \cdot packyears \tag{3.27}$$

$$\kappa(t, z) = e^{f_2(\phi_1,...,\phi_n,z_{1i},...,z_{ni})} \tag{3.28}$$

with parameter vector $\beta = [\Sigma_1, ..., \Sigma_n, \gamma_0, ..., \gamma_n, \phi_0, ..., \phi_n]$ and covariates $z_{1i}, ..., z_{ni}$. Models with synergistic radiation-smoking effects he called **Generalized Multiplicative Relative Risk Models** (GMRRM) and **Generalized Additive Excess Risk Models** (GAERM) and are described by the following equations

$$h(t, z, d) = h_0(t, z)(1 + \rho(d)\epsilon(t, z) \cdot \omega(smk))(1 + \phi(smk)\kappa(t, z)) \tag{3.29}$$

and

$$h(t, z, d) = h_0(t, z) + \rho(d)\epsilon(t, z) \cdot \omega(smk) + \phi(smk)\kappa(t, z)), \tag{3.30}$$

respectively, where $\omega(\cdot)$ is a function of smoking variables with the following form

$$\omega(smk) = e^{\omega_1 \cdot \log(day\ packs\ +1) + \omega_2 \cdot \log^2(day\ packs\ +1)}. \tag{3.31}$$

The final parametric vector for GMRRMs and GAERMs is hence

$$\beta = [\Sigma_1, ..., \Sigma_n, \gamma_0, ..., \gamma_n, \phi_0, ..., \phi_n, \omega_1, \omega_2].$$

The parameters γ_0 and ϕ_0 are called *excess relative risks* for radiation and smoking in GMRRMs and *excess absolute rates* for radiation and smoking in GAERMs.

Excess Relative Risks (ERR) for people with only one kind of exposure can be easily derived solving model (3.20) for the parameter γ_0 or ϕ_0

$$ERR_d = \gamma_0 \cdot \epsilon(t, z) = \frac{h(t, z, d)}{h_0(t, z)} - 1 \tag{3.32}$$

$$ERR_{smk} = \phi_0 \cdot \kappa(t, z) = \frac{h(t, z, d)}{h_0(t, z)} - 1. \tag{3.33}$$

The ERR is best explained using an example. Let us consider an hypothetical ERR of 1.2. It means that per spontaneous case there were 1.2 exposure-induced cases. An ERR of 0 would

mean no difference in amount of cases between exposed and not exposed cohort.
Excess Absolute Rates (EAR) can be calculated with model (3.21)

$$EAR_d = \gamma_0 \cdot \epsilon(t,z) = h(t,z,d) - h_0(t,z) \tag{3.34}$$
$$EAR_{smk} = \phi_0 \cdot \kappa(t,z) = h(t,z,d) - h_0(t,z). \tag{3.35}$$

As the difference of total and baseline cases, the EAR describes the amount of exposure-induced cases.

3.4 Multistage mechanistic models

So far we presented two types of empiric models for hazard rates. Empiric models do not take into account any concrete biological process but are used to provide a parsimonious description of covariates' effects to ERRs and EARs. In this section we present mechanistic multistage models which aim is to approximate disease processes latent in the data. In this Thesis we consider two types of mechanistic models: the **Two Stage Clonal Expansion Model** (TSCE) and the **Three Stage Clonal Expansion Model** (3SCE), that were first analysed by Moolgavkar and Knudson [43]. Mechanistic multistage models describe carcinogenesis as a process with several distinct stages of mutation of stem cells: the TSCE has two stages while the 3SCE has three.

3.4.1 The two stage clonal expansion model

We first start with the description of a TSCE, which is schematically presented in Figure 3.1. The first transition may occur in any fraction of the large number of X healthy stem cells becoming initiated cells with yearly rate ν. Initiated cells can hence differentiate (symmetric division with rate α or inactivation with rate β) and grow into a clone (clonal expansion). Some of these intermediate cells can also transform again with rate μ becoming malignant cells, that after a time-lag Θt can create a detectable cancer lesion.

For mathematical implementation of the TSCE, mutation rates and rates of cell division or

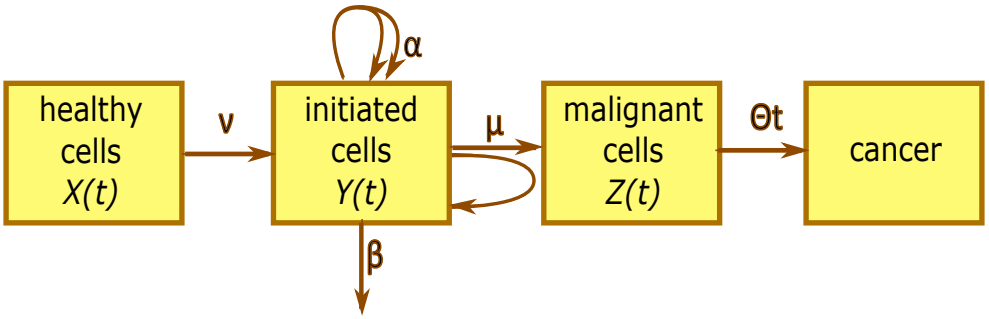


Figure 3.1: Schematic representation of the TSCE. Boxes represent cells in states with defined molecular properties. Arrows represent transitions between cell states. Rates of transition are denoted with Greek letters.

inactivation are treated as transient Poisson point processes of cell birth and death. Table 3.1 gives a summary of the possible transitions in a TSCE. Please note that the amount X of healthy stem cell is assumed to be so large not to be affected by the transition to initiated cells.

Every model has some different states it can adopt. The different possible states the system can take are described as the state space of the model.

Table 3.1: Transition rates of the TSCE. Referred to Figure 3.1

Summary of the possible transitions in the TSCE	
Rate	**Meaning**
ν	transition rate from healthy to initiated cells (increase Y)
α	symmetric division rate of initiated cells (increase Y)
β	inactivation/differentiation rate of initiated cells (decrease Y)
μ	asymmetric division rate of initiated cells, with a malignant daughter cell (increase Z)
Θt	time lag for malignant cells to become a tumor

Definition 3.4.1: (State space of the TSCE)
In the TSCE a cell may adopt three different states. The number of cell in each state is represented by:

- $X(t) = X$: constant number of healthy stem cells (ignoring lung growth during childhood for simplicity),
- $Y(t)$: number of intermediate cells at time t,
- $Z(t)$: number of malignant cells at time t.

It holds that X, $Y(t)$, $Z(t) \in \mathbb{N} \cup \{0\}$.

Since we want to adopt the stochastic version of the TSCE, the state of the model at time t is given by the probability that there are j intermediate and k malignant cells under the condition that at the start time t_0 we had j_0 intermediate and k_0 malignant cells, with $j, k, j_0, k_0 \in \mathbb{N} \cup \{0\}$.

Definition 3.4.2: (The state of the TSCE at time t)
Let j, k, j_0 and $k_0 \in \mathbb{N} \cup \{0\}$ and let us consider the start time t_0 were j_0 intermediate and k_0 malignant cells are present. Then the probability P to have j intermediate and k malignant cells at time t is defined by

$$P(j, k, t) = P(Y(t) = j, Z(t) = k | Y(t_0) = j_0, Z(t_0) = k_0). \tag{3.36}$$

If we want to predict the number of intermediate cells starting from time t after a specific time interval Δt we have to analyze the evolution of the stages over time. To do so we have to take a look at all possible transitions in the system that can happen until the time point $t + \Delta t$ leading to j intermediate and k malignant cells and the corresponding probabilities.

- initiation: $X \to Y$ with probability $X\nu P(j-1, k, t)\Delta t + \mathcal{O}(\Delta t)$
- symmetric cell division: $Y \to 2Y$ with probability $(j-1)\alpha P(j-1, k, t)\Delta t + \mathcal{O}(\Delta t)$
- inactivation: Y cell dies with probability $(j+1)\beta P(j+1, k, t)\Delta t + \mathcal{O}(\Delta t)$
- transition into malignant cell: $Y \to Z$ with probability $j\mu P(j, k-1, t)\Delta t + \mathcal{O}(\Delta t)$
- nothing happens: $(1 - X\nu\Delta t)(1 - j\alpha\Delta t)(1 - j\beta\Delta t)(1 - j\mu\Delta t)P(j, k, t) + \mathcal{O}(\Delta t)$
  Taking only terms of first order into account it remains $(1 - X\nu\Delta t - j\alpha\Delta t - j\beta\Delta t - j\mu\Delta t)P(j, k, t) + \mathcal{O}(\Delta t)$.

The sum over these five different possibilities leads to the probability to have j intermediate and k malignant cells at time $t + \Delta t$:

$$\begin{aligned} P(j,k,t+\Delta t) = & X\nu P(j-1,k,t)\Delta t + (j-1)\alpha P(j-1,k,t)\Delta t \\ & + (j+1)\beta P(j+1,k,t)\Delta t + j\mu P(j,k-1,t)\Delta t \\ & + P(j,k,t) - (X\nu\Delta t + j\alpha\Delta t + j\beta\Delta t + j\mu\Delta t)P(j,k,t) + \mathcal{O}(\Delta t). \end{aligned} \tag{3.37}$$

From equation (3.37) we can now derive the master equation of the system, which describes the evolution of the system over time.

Definition 3.4.3: (The Master Equation of the TSCE)
Let us consider equation (3.37) for the case of no intermediate cells $j_0 = 0$ and no malignant cells $k_0 = 0$ at time t_0. The time evolution of the TSCE is hence given by the following master equation

$$\begin{aligned} \frac{dP(j,k,t+\Delta t)}{dt} = & \lim_{\Delta t\to 0} \frac{P(j,k,t+\Delta t) - P(j,k,t)}{\Delta t} \\ = & X\nu P(j-1,k,t) + (j-1)\alpha P(j-1,k,t) \\ & + (j+1)\beta P(j+1,k,t) + j\mu P(j,k-1,t) \\ & - (X\nu + j\alpha + j\beta + j\mu)P(j,k,t), \end{aligned} \tag{3.38}$$

$$P(0,0,t_0) = 1. \tag{3.39}$$

Equation (3.38) defines the probability to have j intermediate and k malignant starting from time t after a time Δt. A variation in time of the number of intermediate and malignant cells is analysed. Equation (3.38) is a first order ordinary differential equation (ODE) with the initial condition (IC) (3.39) that describes the case in which no intermediate and no malignant cells have already developed at time t_0.

Introducing the definition of the probability generating function the problem (3.38) with IC (3.39) can be rewritten as a partial differential equation (PDE) of first order, which can be solved as presented in the following.

Definition 3.4.4: (The probability generating function of the TSCE)
The probability generating function of a discrete random variable is a power series representation of the probability mass function of the random variable. Let us consider the TSCE with constant number of healthy stem cells ($X(t) = X = const$). The corresponding probability generating function for the random variables $Y(t)$ and $Z(t)$ is hence given by

$$\Psi(y,z,t) = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} P(j,k,t)y^j z^k. \tag{3.40}$$

Please note that since $X(t) = X = const$ the probability generating function does not depend on the number of the healthy stem cells X.
Considering the case in which at the beginning no intermediate and no malignant cells are present, the initial condition of the probability generating function (3.40) has the following form

$$\Psi(y,z,t_0) = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} P(j,k,t_0)y^j z^k = P(j,k,t_0)y^0 z^0 = 1. \tag{3.41}$$

From the TSCE model we want to obtain a functional form describing the probability of having any number intermediate cell $Y(t)$, but no malignant ones. This can be connected to the definition of the survival function $S(t)$ (3.1). Written with probability generating functions it is the sum over all states where any number intermediate cell $Y(t)$ can develop, but no malignant cells:

$$S(t) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} P(j, k = 0, t) = \sum_{j_i=0}^{\infty} P(j, 0, t). \tag{3.42}$$

Proposition 3.4.5:
Let us consider master equation (3.38) with IC (3.39) and definition (3.4.4) of the probability generating function. The equation (3.38) can hence be rewritten as the following PDE

$$\begin{aligned} \frac{\partial}{\partial t}\Psi(y, z, t) &= F(y, t)\Psi(y, z, t) + G(y, z, t)\frac{\partial}{\partial y}\Psi(y, z, t) \\ \Psi(y, z, t_0) &= 1 \end{aligned} \tag{3.43}$$

with

$$\begin{aligned} F(y, t) &:= (y - 1)X\nu \\ G(y, z, t) &:= \mu yz + \alpha y^2 - [\alpha + \beta + \mu]y + \beta \end{aligned}$$

The proof can be found in Appendix C.2. Please note that since no transition departs from the box of the malignant cells (see Figure 3.1), also no partial derivative with respect to the variable z is present.

To solve PDE (3.43) the method of characteristics will be applied [36]. With this method the PDE will be transformed into a system of ODEs, whose equations are the characteristic curves of the PDE itself. For the definition of the characteristics a new variable is needed: s. s is just an auxiliary variable needed for the method of characteristic and does not have a direct interpretation.

Theorem 3.4.6:
Let us consider the PDE (3.43)

$$\begin{aligned} \frac{\partial}{\partial t}\Psi(y, z, t) &= F(y, t)\Psi(y, z, t) + G(y, z, t)\frac{\partial}{\partial y}\Psi(y, z, t) \\ \Psi(y, z, t_0) &= 1 \end{aligned}$$

with

$$\begin{aligned} F(y, t) &:= (y - 1)X\nu \\ G(y, z, t) &:= \mu yz + \alpha y^2 - [\alpha + \beta + \mu]y + \beta \end{aligned}$$

of the corresponding TSCE depicted in Figure (3.1) with the definitions

$$A := -\frac{1}{2}\left(\alpha - \beta - \mu + \sqrt{(\alpha - \beta - \mu)^2 + 4\alpha\mu}\right) \tag{3.44}$$

$$B := \frac{1}{2}\left(-(\alpha - \beta - \mu) + \sqrt{(\alpha - \beta - \mu)^2 + 4\alpha\mu}\right). \tag{3.45}$$

Then, PDE (3.43) with initial condition s_1 on the characteristic has the following solution

$$\Psi(s) = e^{\frac{X\nu}{\alpha}[\ln(f(s))-\ln(f(s_1))]}, \tag{3.46}$$

with characteristic

$$y(s) = \frac{w(s)}{\alpha} + 1, \tag{3.47}$$
$$t(s) = s + t_0, \tag{3.48}$$
$$z(s) = 0, \tag{3.49}$$

where s_1 is the starting point, with for generic initial conditions

$$w(s) = \frac{AB(e^{A(s-s_1)} - e^{B(s-s_1)}) - w(s_1)(Ae^{A(s-s_1)} - Be^{B(s-s_1)})}{(B - w(s_1))e^{A(s-s_1)} + (w(s_1) - A)e^{B(s-s_1)}} \tag{3.50}$$
$$f(s) = (B - w(s_1))e^{A(s-s_1)} + (w(s_1) - A)e^{B(s-s_1)}. \tag{3.51}$$

and for the specific initial condition $w(s_1) = 0$

$$w(s) = \frac{AB(e^{A(s-s_1)} - e^{B(s-s_1)})}{Be^{A(s-s_1)} - Ae^{B(s-s_1)}} \tag{3.52}$$
$$f(s) = Be^{A(s-s_1)} - Ae^{B(s-s_1)}. \tag{3.53}$$

Proof. To solve PDE (3.43) the method of characteristics will be applied. We look hence for the characteristic curves of the analysed equation (3.46). Therefore we introduce a new system of ODEs with the new variable s

- generating function:

$$\frac{d}{ds}\Psi(s) = (y(s) - 1)X\nu\Psi(s) \tag{3.54}$$
$$\Psi(s_0) = \Psi_0 = 1 \tag{3.55}$$
$$\Rightarrow \Psi(s) = \Psi_0 e^{\left(\int_{s_1}^{s}(y(s')-1)X\nu ds'\right)} = e^{\left(\int_{s_1}^{s}(y(s')-1)X\nu ds'\right)} \tag{3.56}$$

- t variable:

$$\frac{d}{ds}t(s) = 1 \rightarrow t(s) = s + t_0, t_0 = constant$$

- z variable:

$$\frac{d}{ds}z(s) = 0 \rightarrow z(s) = z_1 = 0$$

- y variable:

$$\frac{d}{ds}y(s) = -G(s) = -\alpha y(s)^2 + [\alpha + \beta + \mu]y(s) - \beta \tag{3.57}$$

Introducing the following transformation

$$y(s) = \frac{w(s)}{\alpha} + 1 \tag{3.58}$$

we can rewrite the equation (3.57) in a standard form of a so called Riccati equation

$$\frac{d}{ds}w(s) = -w(s)^2 - \gamma w(s) + \delta \tag{3.59}$$

with

$$\gamma = \alpha - \beta - \mu \tag{3.60}$$
$$\delta = \alpha\mu. \tag{3.61}$$

The Riccati equation (3.59) can be solved with standard methods [30]:
First we calculate a particular solution of it equating the right hand side of equation (3.59) to zero

$$\begin{aligned} & w(s)^2 + \gamma w(s) - \delta = 0 \\ \Rightarrow & A, B = \frac{-\gamma \mp \sqrt{\gamma^2 + 4\delta}}{2} \end{aligned} \tag{3.62}$$

A and B are the static solutions for constant w.
Then we introduce the following transformation

$$v(s) := \frac{1}{w(s) - B}. \tag{3.63}$$

Applying the transformation (3.63) to equation (3.59) we obtain the following first order linear ODE

$$\frac{d}{ds}v = [\gamma + 2B]v + 1, \tag{3.64}$$

that can be solved using the standard method of variation of constants. The solution of equation (3.64) for an arbitrary initial condition reads

$$v(s) = e^{(\gamma+2B)(s)}\tilde{c} - \frac{1}{\gamma + 2B}. \tag{3.65}$$

To get rid of the constant $\tilde{c}$ we rewrite it in terms of the general IC $v(s_1) = v_1$. Using equations (3.64) and (3.65) it follows

$$\begin{aligned} & v(s_1) = e^{s_1(\gamma+2B)}\tilde{c} - \frac{1}{\gamma + 2B} \stackrel{!}{=} \frac{1}{w(s_1) - B} \\ \hookrightarrow & \tilde{c} = \frac{\gamma + B + w(s_1)}{(w(s_1) - B)(\gamma + 2B)} e^{-s_1(\gamma+2B)} \end{aligned} .$$

A further simplification can be obtained using A already defined in (3.62):

$$\begin{aligned} -\gamma - B = -\gamma + \frac{\gamma}{2} - \frac{\sqrt{\gamma^2 + 4\delta}}{2} = -\left(\frac{\gamma + \sqrt{\gamma^2 + 4\delta}}{2}\right) = A \\ \hookrightarrow \tilde{c} = \frac{w(s_1) - A}{(w(s_1) - B)(\gamma + 2B)} e^{-s_1(\gamma+2B)}. \end{aligned}$$

The solution $v(s)$ therefore reads

$$v(s) = \frac{(w(s_1) - A)e^{(s-s_1)(\gamma+2B)} - (w(s_1) - B)}{(w(s_1) - B)(\gamma + 2B)}. \tag{3.66}$$

Now (3.66) can be transformed back to the original variable $w(s)$ using definition (3.63)

$$w(s) = \frac{AB(e^{A(s-s_1)} - e^{B(s-s_1)}) - w(s_1)(Ae^{A(s-s_1)} - Be^{B(s-s_1)})}{(B - w(s_1))e^{A(s-s_1)} + (w(s_1) - A)e^{B(s-s_1)}}. \tag{3.67}$$

In equation (3.67) we see that the numerator is the derivative in s of the denominator. With

$$f(s) := (B - w(s_1))e^{A(s-s_1)} + (w(s_1) - A)e^{B(s-s_1)}$$

it holds that

$$w(s) = \frac{\partial}{\partial s} f(s) \cdot \frac{1}{f(s)} = \frac{\partial}{\partial s} \ln(f(s)), \tag{3.68}$$

Finally equation (3.56) reads

$$\begin{aligned} \Psi(s) &= e^{X\nu \int_{s_1}^{s} (y(s')-1)ds'} = e^{\frac{X\nu}{\alpha} \int_{s_1}^{s} w(s')ds'} = e^{\frac{X\nu}{\alpha} \int_{s_1}^{s} \frac{d}{ds'} \ln(f(s'))ds'} \\ &= e^{\frac{X\nu}{\alpha}[\ln(f(s)) - \ln(f(s_1))]} \qquad (3.69) \\ \Psi(y, z, t_0) &= 1 \qquad (3.70) \end{aligned}$$

For the specific initial condition $w(s_1) = 0$ the functions $w(s)$ and $f(s)$ read

$$\begin{aligned} w(s) &= \frac{AB(e^{A(s-s_1)} - e^{B(s-s_1)})}{Be^{A(s-s_1)} - Ae^{B(s-s_1)}} \\ f(s) &= Be^{A(s-s_1)} - Ae^{B(s-s_1)} \end{aligned}$$

□

Remark 3.4.7:
Please note that $\Psi(s)$ is the survival function for the TSCE. We calculated it using probability generating functions as the sum over all states where any number of intermediate cells $Y(t)$ can develop, but no malignant cells. Equation (3.70) gives hence the probability that no intermediate and no malignant cells have already developed from normal healthy cells.

Using theorem 3.1.3 the hazard function can be derived out of the survival function

$$h(t) = -\frac{d}{dt} \ln(S(t)).$$

Taking the logarithm of equation (3.70)

$$\ln(\Psi(t)) = \ln(S(t)) = \frac{X\nu}{\alpha}[\ln(f(t)) - \ln(f(t_0))] \tag{3.71}$$

and inserting the relevant integration boundaries t (attained age) and $t_0 = 0$ (birth) one gets the final formula

$$\ln(S(t)) = \frac{X\nu}{\alpha}[\ln(B - A) \; + \; Bt \; - \; \ln(Be^{(B-A)t} - A)]. \tag{3.72}$$

The next and last step to the final hazard function is to take the derivative of equation (3.72) with respect to t

$$h(t) = -\frac{d}{dt} \ln(S(t)) = X\nu\mu \left[\frac{e^{(B-A)t} - 1}{Be^{(B-A)t} - A} \right]. \tag{3.73}$$

Table 3.2: Identifiable parameters of the TSCE. Referred to Figure 3.1

Summary of the identifiable parameters in the TSCE	
Parameter	**Meaning**
$C := X\nu\mu$	Initiation rate
$\gamma := \alpha - \beta - \mu$	Clonal expansion rate of the intermediate cells
$\delta := \alpha\mu$	Stochasticity parameter

The TSCE has only three identifiable parameters, shown in Table 3.2. Parameter C describes the so called initiation rate, which describes the turn over to malignant cell in the absence of clonal expansion. Parameter γ is the so called clonal expansion rate of the intermediate cells. It describes the net increase of the intermediate cells. δ is a so called stochasticity parameter and has no direct biological meaning.

So far we modeled the probability of a person to get a malignant cell without any external influence on it. If we want to build in the effects of external exposures (e.g. irradiation and smoking) on the hazard function, the parameters are assumed to depend on there exposures and thus vary with time.
Let us consider the case in which a person gets exposed to ionizing radiation at age *Age at bombing*, begins to smoke at age *Age begin smoking* and does not quit. Let us also assume that the irradiation exposure is a rather acute one with biological effects lasting for one week. Finally we assume that irradiation and smoking will affect the identifiable parameters. Figure 3.2 represents exemplary changes to the clonal expansion γ.

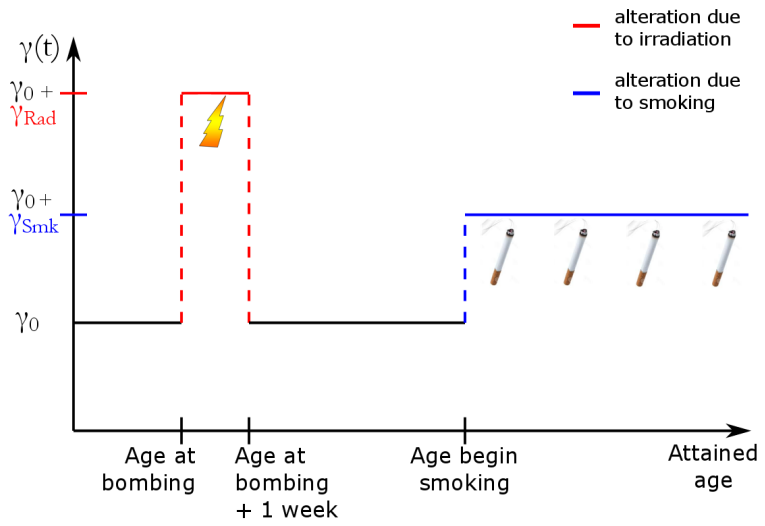


Figure 3.2: Schematic representation of the clonal expansion rate γ under the effects of ionising irradiation (red) and smoking (blue). The ionizing irradiation-exposure is acute and is assumed to have an effect of one week. Smoking is assumed to be continued at a fixed rate.

Within each of the age-intervals the parameters are assumed to be constant (piecewise-constant parameters). Therefore the hazard function can be derived as shown in the sections before for every single interval and then summed up.
The logarithm of the survival function with piecewise-constant parameters with k age-intervals

can be written similarly to equation (3.72)

$$\ln(S(t)) = \sum_{i=1}^{k} \frac{X\nu_i}{\alpha_i} \int_{s_{i-1}}^{s_i} w_i(s)ds = \sum_{i=1}^{k} \frac{C_i}{\alpha_i\mu_0} \int_{s_{i-1}}^{s_i} w_i(s)ds$$
$$= \sum_{i=1}^{k} \frac{C_i}{\delta_i} \int_{s_{i-1}}^{s_i} w_i(s)ds, \tag{3.74}$$

with

- $i = 1, ..., k$
- $C_i := X\nu_i\mu_0$,
- $\gamma_i := \alpha_i - \beta_i - \mu_i$,
- $\delta_i := \alpha_i\mu_0$,
- $A_i = -\frac{1}{2}(\gamma_i + \sqrt{\gamma_i^2 + 4\delta_i\theta_i})$,
- $B_i = \frac{1}{2}(-\gamma_i + \sqrt{\gamma_i^2 + 4\delta_i\theta_i})$,
- $\theta_i = \frac{\mu_i}{\mu_0}$.

A new identifiable parameter θ_i is needed since in the expressions A_i and B_i the piecewise-constant parameter δ_i do not contain the piecewise-constant parameter μ_i. All identifiable parameters for the piecewise constant TSCE are presented in Table 3.3.

Table 3.3: Identifiable piecewise-constant parameters of the TSCE. For definition of parameters, see Figure 3.1

Summary of the piecewise-constant parameters in the TSCE	
Parameter	**Meaning**
$C_i := X\nu_i\mu_0$	Piecewise-constant initiation rate
$\gamma_i := \alpha_i - \beta_i - \mu_i$	Piecewise-constant clonal expansion rate of the intermediate cells
$\delta_i := \alpha_i\mu_0$	Piecewise-constant stochasticity parameter
$\theta_i = \frac{\mu_i}{\mu_0}$	

In order for the survival function (3.74) to be a continuous function, boundary conditions for every interval of piecewise constant parameters are needed

$$w_{i-1}(s_{i-1}) = \frac{\alpha_{i-1}}{\alpha_i} w_i(s_{i-1}) = \frac{\delta_{i-1}}{\delta_i} w_i(s_{i-1}).$$

Remembering equations (3.51) and (3.53) of Theorem 3.4.6 it is known that the form of the function $f(s)$ differs for different initial conditions. In the first step of the recursion over all age-intervals it holds $w(s_i) = 0$. Therefore the formula for the function $f_i(s)$ is given by

$$f_i(s_i) = (B_i - w_i(s_i))e^{A_i(s_i - s_i)} + (w_i(s_i) - A_i)e^{B_i(s_i - s_i)} = B_i - A_i \qquad \text{for } i = k$$

For all other steps it holds that $w_i(s_i) = \frac{\delta_i}{\delta_{i-1}} w_{i-1}(s_{i-1})$, the formula for $f_i(s)$ is hence given by:

$$f_i(s_{i-1}) = (B_i - w_i(s_i))e^{A_i(s_{i-1} - s_i)} + (w_i(s_i) - A_i)e^{B_i(s_{i-1} - s_i)} \qquad \text{for } i \neq k$$

By inserting this result in equation (3.74) and applying equation (3.68) of $w(s)$

$$w(s) = \frac{\partial}{\partial s} \ln(f(s)),$$

the final formula of the survival function for the piecewise-constant reads

$$\ln(S(t)) = \sum_{i=1}^{k} \frac{C_i}{\delta_i} \int_{s_{i-1}}^{s_i} \frac{d}{ds} \ln f_i(s) ds = \sum_{i=1}^{k} \frac{C_i}{\delta_i} \ln \left(\frac{B_i - A_i}{f_i(s_{i-1})} \right)$$

For the explicit formula of the hazard function for piecewise-constant parameters one has to take the derivative with respect to t again. The hazard function for the TSCE with piecewise-constant parameters is hence given by

$$h(t) = -\frac{d}{dt} \ln(S(t)) = \sum_{i=1}^{k} \frac{C_i}{\delta_i} \frac{1}{f_i(s_{i-1})} \frac{d}{dt} f_i(s_{i-1}). \tag{3.75}$$

A backward recursion algorithm for the hazard function with piecewise-constant parameters can be found in Appendix C.3.

3.4.2 The three stage clonal expansion model

For certain cancer types two consecutive initiation steps might be necessary before cells gain a proriferative advantage. This model concept can be seen in Figure 3.3. As in the TSCE, despite

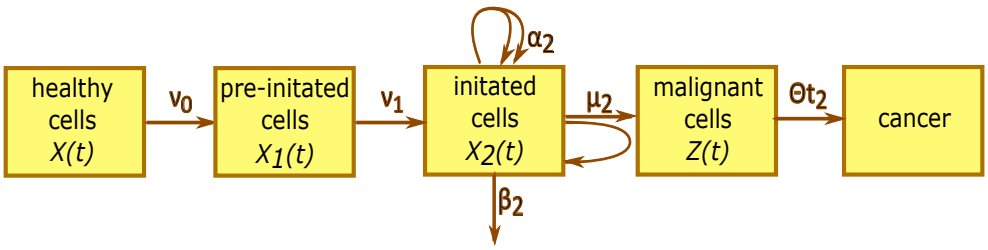


Figure 3.3: Schematic representation of the 3SCE. Boxes represent cells in states with defined molecular properties. Arrows represent transitions between cell states. Rates of transition are denoted with Greek letters.

the large number of healthy stem cells X, early molecular changes with yearly rate ν_0 leading to pre-initiated cells are rare. Pre-initiated cells may hence mutate again with rate ν_1 becoming initiated cells. Initiated cells can either divide symmetrically with rate α_2 or become inactivated with rate β_2. The final transformation stage summarizes a sequence of complex processes with effective rate μ_2. After a time-lag Θt_2 malignant cells can create a detectable cancer lesion.

With the same procedure as for the TSCE one can obtain ODEs for the variables $x_1(s)$ and $x_2(s)$ (cf. equation (3.57))

$$\frac{\partial}{\partial s} x_1(s) = -\nu_1 \cdot (x_2(s) - 1) x_1(s) \tag{3.76}$$

$$\frac{\partial}{\partial s} x_2(s) = -\alpha_2 x_2(s)^2 + (\alpha_2 + \beta_2 + \nu_2) \cdot x_2(s) - \beta_2 \tag{3.77}$$

with IC

$$x_1(s_1) = 1 \qquad \text{and} \qquad x_2(s_1) = 1.$$

As before, s runs over the domain of the characteristic curve. Analogously to equation (3.71) the logarithm of the survival function $S(t)$ is given by the following equation

$$\ln(S(s)) = \int_0^s X \cdot \nu_0 \cdot (x_1(s') - 1) ds' \tag{3.78}$$

Now we introduce two transformations

$$x_1(s) = w_1(s) + 1 \tag{3.79}$$

$$x_2(s) = \frac{1}{\alpha_2} w_2(s) + 1. \tag{3.80}$$

Inserting equation (3.79) and (3.80) into the formulas (3.76) and (3.77), respectively, we get two new ODEs in w_1 and w_2

$$\frac{d}{ds} w_1(s) = - \frac{\nu_1}{\alpha_2} \cdot w_2(s) \cdot (w_1(s) + 1) \tag{3.81}$$

$$\frac{d}{ds} w_2(s) = - w_2^2(s) - (\alpha_2 - \beta_2 - \nu_2) \cdot w_2(s) + \alpha_2 \nu_2. \tag{3.82}$$

with the IC

$$w_1(s_1) = 0 \tag{3.83}$$

$$w_2(s_1) = 0. \tag{3.84}$$

Equation (3.78) can hence be rewritten to:

$$\ln(S(s)) = X \cdot \nu_0 \int_0^s w_1(s') ds' \tag{3.85}$$

As for the TSCE there are identifiable parameters that can be introduced. They are presented in Table 3.4.

Table 3.4: Identifiable parameters of the 3SCE. Referred to Figure 3.3

Summary of the identifiable parameters in the 3SCE	
Parameter	**Meaning**
$C_2 := \mathrm{X} \cdot \nu_0 \cdot \nu_1 \cdot \mu_2$	Initiation rate
$\gamma_2 := \alpha_2 - \beta_2 - \mu_2$	Clonal expansion rate of the intermediate cells
$r_1 := \nu_1 \cdot \mu_2$	First stochasticity parameter
$\delta_2 := \mu_2 \alpha_2$	Second stochasticity parameter

With equations (3.81), (3.82) and (3.85) $w1(s)$, $w2(s)$ and $\ln(S(s))$ can be written as

$$\frac{d}{ds} w_1(s) = - \frac{r_1}{\delta_2} \cdot w_2(s) \cdot (w_1(s) + 1) \tag{3.86}$$

$$w_1(s_1) = 0$$

$$\frac{d}{ds} w_2(s) = - w_2^2(s) - \gamma_2 \cdot w_2(s) + \delta_2 \tag{3.87}$$

$$w_2(s_1) = 0$$

$$\ln(S(s)) = \frac{C_2}{r_1} \int_0^s w_1(s') ds'. \tag{3.88}$$

The ODE (3.87) is again a Ricatti equation and can be solved with the same procedure as for equation (3.59) in the proof of Theorem 3.4.6. The solution of the Riccati equation (3.87) has the following form

$$w_2(s) = \frac{\partial}{\partial s} \ln(l(s)) \tag{3.89}$$

with

$$l(s) = (E - w_2(s_1))e^{D(s-s_1)} + (w_2(s_1) - D)e^{E(s-s_1)}, \tag{3.90}$$

for generic initial conditions, and

$$l(s) = Ee^{D(s-s_1)} - De^{E(s-s_1)}, \tag{3.91}$$

for the specific initial condition $w_2(s_1) = 0$

$$D = \frac{-\gamma_2 - \sqrt{\gamma_2^2 + 4\delta_2}}{2} \tag{3.92}$$

$$E = \frac{-\gamma_2 + \sqrt{\gamma_2^2 + 4\delta_2}}{2}, \tag{3.93}$$

for specific initial conditions $w_2(s_1) = 0$.
The system composed of equations (3.86), (3.89) and (3.88) has to be solved and has the following form

$$\begin{aligned} \frac{d}{ds} w_1(s) =& -\frac{r_1}{\delta_2} \cdot w_2(s) \cdot (w_1(s) + 1) \\ w_1(s_1) =& 0 \\ w_2(s) =& \frac{\partial}{\partial s} \ln(l(s)) \\ \ln(S(s)) =& \frac{C_2}{r_1} \int_0^s w_1(s')ds'. \end{aligned}$$

Since the solution of $w_2(s)$ (3.89) depends on variable s, equation (3.86) is not easy to solve and there is no convenient solution for the logarithm of the survival function (3.88). Therefore a numerical solution is necessary.

As in the case of the TSCE, one can next introduce the effects of external agents and make the parameters variable in time, for example with piecewise-constant parameters. Since for the 3SCE a closed solution could not be found for constant parameters, this holds true for piecewise-constant ones. We prefer at this point to introduce a simplification of the 3SCE for which a closed solution could be found: the Hybrid Three Stage Clonal Expansion Model (H3SCE). This model will be introduced in the next section and used for data analysis in this Thesis instead of the 3SCE.

3.4.3 The hybrid three stage pre-initiation model

The H3SCE is intended to be an approximation of the 3SCE. It is assumed that the first transition can be approximated by the expected number of pre-initiated cells at time t: $E[X_1(t)]$. This model is called hybrid because the first step is modeled deterministic when all other transitions are modeled in a stochastic way. Figure 3.4 depicts a comparison of 3SCE and H3SCE.

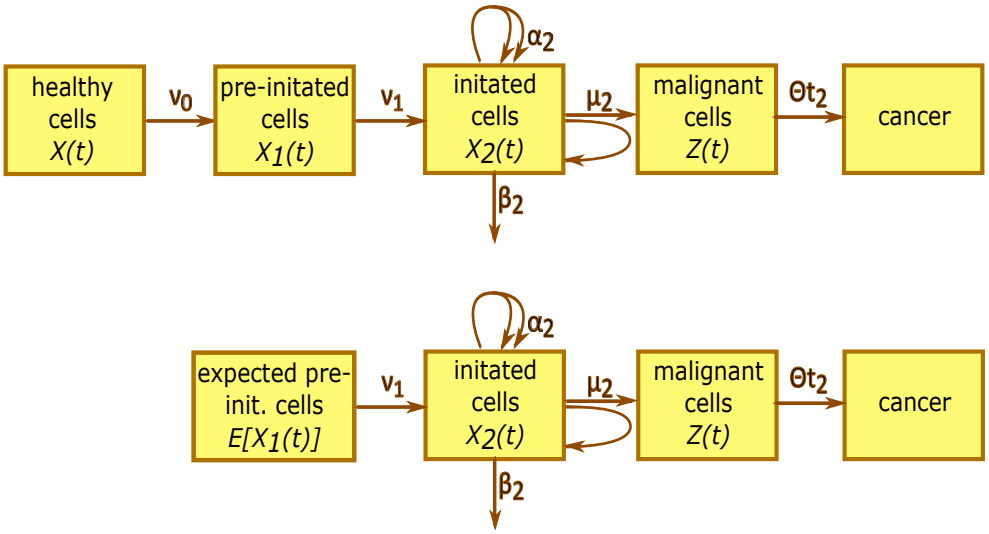


Figure 3.4: Schematic representation of the H3SCE in comparison with the 3SCE. Boxes represent cells in states with defined molecular properties. Arrows represent transitions between cell states. Rates of transition are denoted with Greek letters.

The ODEs for the variables $x_1(s,t)$, equation (3.76), and $x_2(s,t)$, equation (3.77), can be adopted from 3SCM

$$\begin{aligned}
\frac{\partial}{\partial s}x_1(s,t) &= -\nu_1 \cdot (x_2(s,t)-1)x_1(s,t) \\
x_1(t,t) &= 1 \\
\frac{\partial}{\partial s}x_2(s,t) &= -\alpha_2 x_2(s,t)^2 + (\alpha_2+\beta_2+\mu_2)\cdot x_2(s,t) - \beta_2 \\
x_2(t,t) &= 1.
\end{aligned}$$

Only the logarithm of the survival function, eqaution (3.78), changes a bit, from

$$\ln(S(t)) = X \cdot \nu_0 \int_0^t (x_1(s,t)-1)ds$$

to

$$\ln(S(t)) = \int_0^t E[X_1(s)](x_1(s,t)-1)ds \tag{3.94}$$

with

$$E[X_1(t)] = X\nu_0 \cdot t \tag{3.95}$$

Therefore equation (3.94) becomes

$$\ln(S(t)) = X\nu_0 \int_0^t s(x_1(s,t)-1)ds. \tag{3.96}$$

Equation (3.77) was already solved in the section before and its solution reads

$$\begin{aligned} w_2(s,t) &= \frac{\partial}{\partial s}\ln(l(s,t)) \\ l(s,t) &= (E - w_2(s_1,t))e^{D(s-s_1)} + (w_2(s_1,t) - D)e^{E(s-s_1)} \\ D &= \frac{-\gamma_2 - \sqrt{\gamma_2^2 + 4\delta_2}}{2} \\ E &= \frac{-\gamma_2 + \sqrt{\gamma_2^2 + 4\delta_2}}{2} \\ \delta_2 &= \mu_2\alpha_2 \\ \gamma_2 &= \alpha_2 - \beta_2 - \mu_2, \end{aligned}$$

remembering the application of transformation (3.80)

$$x_2(s,t) = \frac{1}{\alpha_2}w_2(s,t) + 1.$$

Now we analyze the ODE (3.76) for the $x_1(s,t)$ variable. It is a first order linear equation, the solution is therefore given by

$$x_1(s,t) = e^{\nu_1\int_s^t (x_2(s',t)-1)ds'}.$$

Assuming small values for the exponential function, we apply a Taylor expansion to it

$$x_1(s,t) = 1 + \nu_1\int_s^t (x_2(s',t)-1)ds' + \frac{\nu_1}{2}\left(\int_s^t (x_2(s',t)-1)ds'\right)^2 + ... \tag{3.97}$$

Inserting the terms of maximal first-order of equation (3.97) into equation (3.96) leads to

$$\ln(S(t)) = X\nu_0\nu_1\int_0^t s(x_2(s,t)-1)ds. \tag{3.98}$$

The integral of equation (3.98) can be solved by the method of integration by parts and the solution reads

$$\ln(S(t)) = X\nu_0\nu_1 t\int_0^t x_2(s',t) - 1ds'. \tag{3.99}$$

Recalling theorem 3.1.3, the hazard function can be easily calculated from the logarithm of the survival function

$$h(t) = -\frac{d}{dt}\ln(S(t)).$$

The hazard function of the H3SCE-model hence reads (for specific initial conditions)

$$h(t) = \frac{C_2}{\delta_2}t\frac{d}{ds}\ln(l(s,t)) \tag{3.100}$$

$$l(s,t) = Ee^{D(s-t)} - De^{E(s-t)} \tag{3.101}$$

$$D = \frac{-\gamma_2 - \sqrt{\gamma_2^2 + 4\delta_2}}{2} \tag{3.102}$$

$$E = \frac{-\gamma_2 + \sqrt{\gamma_2^2 + 4\delta_2}}{2} \tag{3.103}$$

Remark 3.4.8:
Please note that the final solution (3.100)- (3.103) of the hazard function for the H3SCE is equal to the final solution (3.73) of the hazard function for the TSCE just with different identifiable parameters and an additional factor linear in age.

As before, effects of external agents are introduced with the assumption that the parameters are constant in each interval. Accordingly, the linear dependence on s is approximated by its mean value on each interval.

$$\frac{s_i + s_{i+1}}{2}.$$

The hazard function for piecewise-constant parameters for the H3SCE hence reads

$$h(t) = \sum_{i=1}^{k} \frac{C_{2_i}}{\delta_{2_i}} \frac{(s_{i-1} + s_i)}{2} \frac{1}{l_i(s_{i-1}, t)} \frac{d}{dt} l_i(s_{i-1}, t). \tag{3.104}$$

A backward recursion of the hazard function in the case of piecewise constant parameters in the H3SCE is provided in Appendix C.4.

3.5 Combining parameter estimates from imputed data sheets

Applying the above presented methods to imputed data sheets, the definition of a MI overall point estimate has to be given. The gold standard for combining parameter estimates from imputed data sheets into one MI estimate is the so called Rubin's rule [40, 50]. Let us consider m imputed data sheets with single regression coefficients $\hat{Q_1}, ..., \hat{Q_m}$ and associated variances $U_1, ..., U_m$. The MI overall point estimate $\bar{Q}$ is the average of the m estimates of Q from the imputed data sheets

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i. \tag{3.105}$$

The corresponding total variance T for the overall MI estimate $\bar{Q}$ is

$$T = \bar{U} + \left(1 + \frac{1}{m}\right) \cdot B, \tag{3.106}$$

where

$$\bar{U} = \frac{1}{m} \sum_{i=1}^{m} U_i \tag{3.107}$$

is the estimated within imputation variance and

$$B = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{Q}_i - \bar{Q})^2 \tag{3.108}$$

is the between-imputation variance [40, 50].
The between-imputation variance is scaled by a factor $1/m$, reflecting the extra variability as a consequence of imputing the missing data using a finite number of imputations instead of an infinite number of imputations. If B dominates $\bar{U}$, m can be increased to improving the accuracy of estimates [40, 50].
These procedures can be extended in matrix form to combine k regression coefficients, where $\hat{Q}$ is a $k \times 1$ vector of these estimates and U is the associated $k \times k$ covariance matrix [40, 50].

The confidence interval (CI) of the parameters can hence be calculated with the standard formula

$$CI = \left[\bar{Q} - z_{1-\frac{\alpha}{2}}\frac{T}{\sqrt{m}}, \bar{Q} + z_{1-\frac{\alpha}{2}}\frac{T}{\sqrt{m}}\right], \tag{3.109}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the standardised normal distribution.
With this procedure, the CI is per definition symmetric.

The aim of this Thesis is the understanding of the risks for LADCs and SQUAMs under the exposures of ionising radiation and cigarette smoke. From the models presented in the above sections, risks estimates will be derived, whose CI may not be necessarily symmetric. An extension of the Rubin's rule, which includes the parameter correlation matrix, will hence be used in this Thesis. The extended procedure is depicted in Figure 3.5. As in the Rubin's rule, from

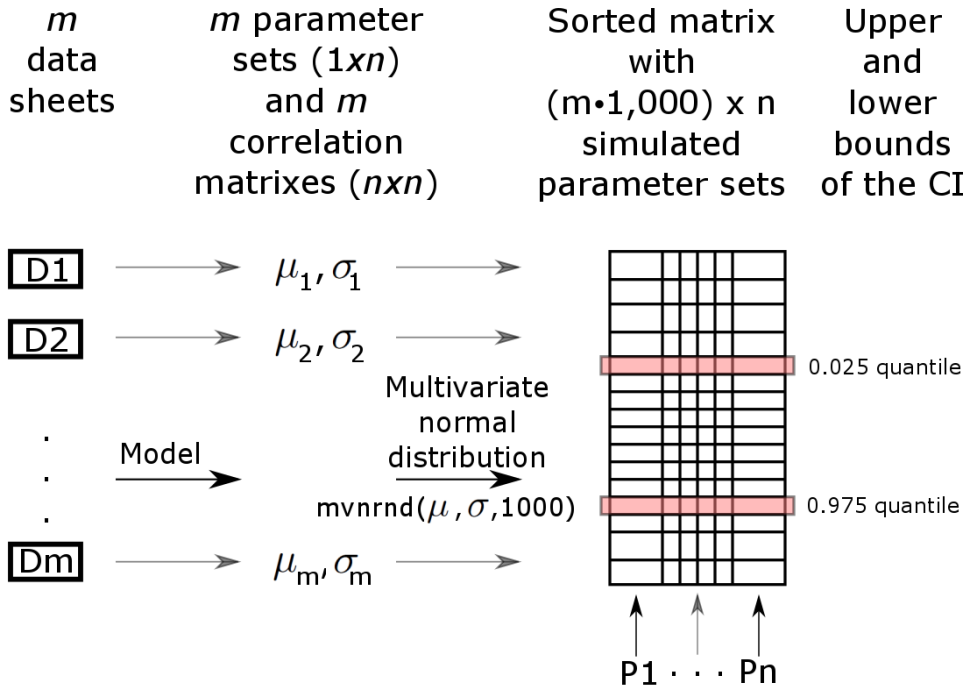


Figure 3.5: An extension of the Rubin's rule to calculate CIs of a MI overall point estimate.

m data sheets m maximum likelihood estimates will be fitted by Poisson regression, calculating also the respective correlation matrices. The MI overall point estimate $\bar{Q}$ is hence calculated as in equation (3.105). For the calculation of the CI, differently as in the the Rubin's rule, the correlation matrices play a role. Let us consider only one dataset $D1$ with maximum likelihood estimate μ_1 and correlation matrix σ_1. From a multivariate normal distribution with inputs μ_1 and σ_1 1000 simulated parameter sets are hence calculated. These procedure is hence applied to all m dataset, ending up with $m \cdot 1000$ simulated parameter sets. The parameter sets are hence put together in one single matrix having $m \cdot 1000$ rows and as many columns as the number of parameters contained in the parameter sets. From this matrix the 0.025 and the 0.975 quantiles are taken, giving a 95% CI for each parameter of the maximum likelihood estimate.
The same procedure can be applied to calculate CI for baseline hazards, hazards, EARs and ERRS. With the simulated parameters sets 1000 curves per data sheet will be calculated and hence the CI will be taken, as presented above. The asymmetry of the CI is not so much of interest for the parameters itself but more for the calculated risks.
In this extension of the Rubin's rule, all components presented in the Rubin's rule itself are

used. The within and the between imputation variances play both a role in the simulation of the parameter sets due to the correlation matrix. Since all parameters are hence merged into a single matrix and the quantiles are taken, number of imputed data sheets and variation in the imputation keep playing a role.

To compare models, cumulative deviance and AIC will be calculated as the sum of all deviances and AICs, respectively. A mean value and a descriptive representation via boxplots of the variation between deviances will be presented for each model in the respective appendix chapter. We preferred to present the total deviance and not to the mean value of the deviances in the main part of the Thesis since the deviance as a mean value would be difficult to interpret. The imputed data sheets differ in number of strata (number of rows in the tables). This is due to the fact that the imputed smoking information was also used to group the data sheets. Since the smoking information varies between data sheets, also the number of created strata varies. The difference in deviance between data sheets derives primarily from the number of strata/rows and not from a different goodness of fit of the model. The mean value contains all the information as the total deviance, but it would be difficult to interpret.

The best model to calculate the CI would be a Markov-Chain-Monte-Carlo simulation, which we decided however be a too complex and time costly method for our purposes.

Remark 3.5.1: (MI estimate of GAMs)
For GAMs the overall MI estimate is given as mean value over the 50 estimates coming from the imputed data sheets. Instead of the CI calculated from the correlation matrices, the standard deviation of the 50 parameters will be given. The variation between the calculated risks is given as boxplots over the 50 risks values for each point of the variable analysed (for attained age one boxplot for each age, for lung dose one boxplot for each dose).

3.6 Software

In this section we describe the software used to implement and fit the data.

- State-of-the-art statistical risk models were implemented and fitted with MATLAB (version R2017b) using the minimising algorithm *fminunc* already implemented in the program. The code is not presented in this Thesis since it is trivial: it just includes the equations given for each model in a Poisson regression.
- GAMs were implemented and fitted with R (version 3.3.0) using the *mgcv* package and the already implemented function *gam*. A code for this models is given for each analysis as output of the model.
- Mechanistic models were implemented and fitted with MATLAB (version R2017b) using the minimising algorithm *fminunc* already implemented in the program. The code for the recursion is presented in Appendix D.

For each model Poisson regression was performed. The different models give different estimations of the considered cases. To compare results from R and MALTAB it was ensured that both programs deliver the same estimates and deviance (AICs) for simple baseline models of state-of-the-art statistical risk models and GAMs (no usage of penalised splines in GAMs).

Results

CHAPTER

4

LUNG ADENOCARCINOMA IN THE LIFE SPAN STUDY COHORT

In this chapter we combine molecular and observational data to develop the first molecular mechanistic model for lung adenocarcinoma. Using two comprehensive genomic data sets from Eastern and Western patient populations, we determine that there are two broad molecular pathways to lung adenocarcinoma: one unique to EGFR-, EML_4/ALK-, and other transmembrane receptor-mutant (R^{MUT}) patients and one shared between KRAS-, BRAF-, NF_1-, PIK_3CA-, and other sub-membrane transducer-mutant (T^{MUT}) patients. Deploying information of smoke and irradiation exposure from the Life Span Study of Japanese atomic bomb survivors, we develop a mechanistic model to estimate the risk for lung adenocarcinoma by molecular pathway. The molecular mechanistic model accurately reproduces the observed incidence in the Life Span Study with moderately improved goodness-of-fit compared to standard epidemiological models. Amazingly, the molecular mechanistic model predicts for the first time the EGFR and KRAS mutation frequencies actually observed in different populations, a fact open to direct validation since for the Life Span Study genomic data are not jet available. Importantly, the molecular mechanistic model supports firm biological evidence for a close association between R^{MUT} cases with environmental radiation and T^{MUT} cases with smoking for the explanation of observational data.

4.1 State-of-the-art statistical risk models

In any mechanistic analysis, state-of-the-art statistical risk models are indispensable to put the results of mechanistic models into perspective. Here, the state-of-the-art statistical risk model model for LADC in the LSS ($Stat_{LSS}^{LADC}$) is inspired by the ERR model of Egawa et al. [18], applying additive interaction of smoking and radiation with the lowest AIC. Based on AIC, the additive action of smoking ($S = S(packyr, smkdyr, smkdqyr, smkint)$) and radiation ($R = R(D)$) is slightly favored, the EAR of Section 3.3.2. This action leads to a total hazard function

$$h = h_0 \cdot (1 + \rho(R) + \Psi(S)) \tag{4.1}$$

which applies the baseline hazard h_0 and the corresponding ERR $\rho(R)$ and $\Psi(S)$ according to

$$\begin{aligned} h_0 &= e^{\beta_1+\beta_2\cdot(city-1)+\beta_3\cdot agexp+\beta_4\cdot\ln(age)+\beta_5\cdot\ln(age)^2}, \\ \rho(R) &= \beta_6 \cdot D \cdot e^{\beta_7\cdot\ln(age)}, \\ \Psi(S)_{f,m} &= \beta_8 \cdot packyr \cdot e^{\beta_9\cdot smkdyr+\beta_{10}\cdot\ln(smkdqyr+1)+\beta_{11f,m}\cdot smkint}. \end{aligned} \tag{4.2}$$

The meaning of the variables introduce in the previous system of equations (4.2) is explained in Table 4.1.
The baseline hazard depends on city of residence (Hiroshima or Nagasaki), age at exposure and attained age. The radiation-related ERR $\rho(R)$ depends linearly on the lung dose and is modified by attained age.
The smoking-related ERR $\Psi(S)$ depends linearly on the cumulative smoking amount and is modified by years smoked, years since quit smoking, and smoking intensity. Only the last modifier was found to be sex-dependent.

Table 4.1: Explanatory variables for $Stat_{LSS}^{LADC}$. In the baseline hazard h_0 age at exposure is equivalent to birth year (birth year = 1945.7 - age at exposure). A pack contains 20 cigarettes. The only sex-dependent parameters $\beta_{11f,m}$ are related to smoking intensity. For the other parameters the sex-difference was found to be not statistically significant based on likelihood ratio tests on the 95% level.

Variable	Unit	Meaning
city	-	Hiroshima (1) or Nagasaki (2)
agexp	-	(age at exposure - 30 yr)/10 yr
age	-	attained age/70 yr
D	Gy	lung dose
packyr	packs/day × yr	cumulative amount of cigarette packs (packs smoked per day × years smoked)/50 yr
smkdyr	-	years smoked/50 yr
smkdqyr	-	years since quit smoking/50 yr
smkint	cigs/day	smoking intensity (cigs smoked per day)

MI overall point estimate with 95% Confidence Intervals (CI) for $Stat_{LSS}^{LADC}$ is presented in Table 4.2 (see section 3.5 for methodology). The deviances of all 50 data sheets can be found in Figure E.1.
A visualisation of risk estimates for $Stat_{LSS}^{LADC}$ is given in Section 4.2 for comparison with mechanistic models.

Table 4.2: MI overall point estimate for the state-of-the-art statistical risk model $Stat_{LSS}^{LADC}$ with 12 parameters. Central estimates are given as means from 50 imputed data sets with 95% CI simulated from multi-variate normal uncertainty distributions conditioned with the parameter correlation matrix (see section 3.5). Cumulative deviance/AIC are the sum over the 50 deviances/AICs.

Parameter estimates of the state-of-the-art statistical risk model $Stat_{LSS}^{LADC}$		
Name	Meaning	MI Mean (95% CI)
β_1	baseline	1.07 (0.95, 1.18)
β_2	baseline, city	0.23 (0.11, 0.35)
β_3	baseline, age at exposure	-0.26 (-0.32, -0.21)
β_4	baseline, attained age	4.19 (3.47, 4.90)
β_5	baseline, attained age (squared)	-4.58 (-6.08, -3.08)
β_6	radiation, linear resp. (Gy^{-1})	1.11 (0.62, 1.60)
β_7	radiation, attained age	-2.08 (-3.92, -0.23)
β_8	smoking, linear resp. (day × yr/packs)	5.82 (3.38, 8.57)
β_9	smoking, years smoked	0.91 (-0.18, 2.06)
β_{10}	smoking, years since quitting	-0.33 (-0.63, -0.08)
β_{11f}	smoking, smoking intensity females	-0.055 (-0.094, -0.018)
β_{11m}	smoking, smoking intensity males	-0.025 (-0.045, -0.005)
Cumulative deviance	252885	
Cumulative AIC	254090	

4.2 Molecular mechanistic models

In this section, we will first start with an analysis of biological data in order to give the motivation for the model design of the final molecular model. The biological data used here is already published in [9]. They compared the somatic profiles of LADCs and SQUAMs to identify novel genetic alterations. 660 LADC/normal paired exome sequences (including 274 previously unpublished cases, 227 previously described from [10] were analysed for LADC. We downloaded and analysed the same data of [9] concerning LADC.
Section 4.2.1 was developed in collaboration with G. T. Stathopoulos of the Comprehensive Pneumology Center (Ludwig-Maximilian University and Helmholtz Zentrum Muenchen.

4.2.1 Biological analysis

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 driver oncogene [9]. In addition to the available clinical information (smoking status, age, sex, etc.), 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 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 [9, 3]. In addition, we calculated the indel-SNV ratios, since such high ratios were found elsewhere to represent a direct molecular imprint of iatrogenic γ-IR [6].

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})).
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 mismatch repair-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 [6]. Copy number alteration indices were comparable across patients with different drivers, except from
ALK/RET/ROS1-fused patients that collectively displayed lower copy number alteration indices compared with all other patients (Figures 4.1, 4.2).

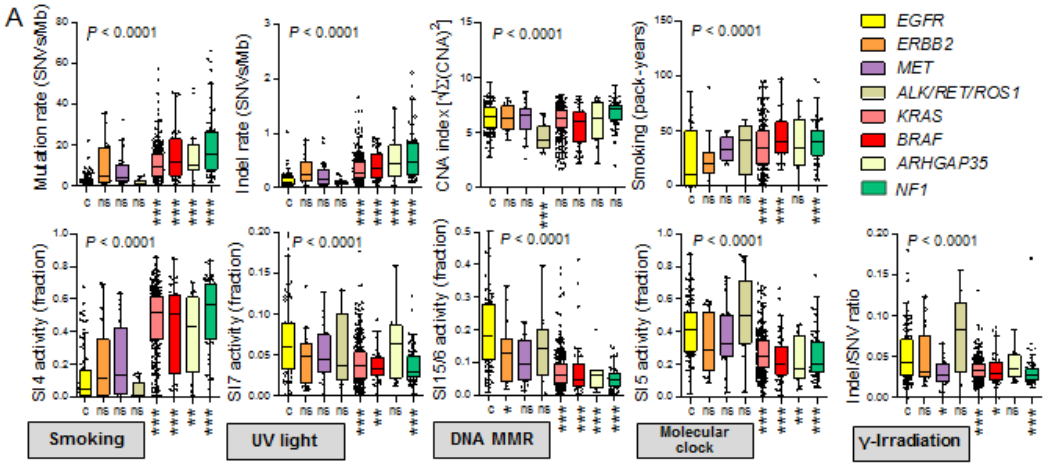


Figure 4.1: SNV rates, indel rates, CNA indices, smoking exposure, sex, genomic signatures of environmental carcinogen-induced base changes in the trinucleotide context (SI), indel/SNV ratios, and transversion status of 660 patients with LADC from the USA [9] grouped by the most frequent driver mutations. Significances $P \geq 0.05$, $P < 0.05$, $P < 0.01$, and $P < 0.001$ are coded as ns, $\star$, $\star\star$, and $\star\star\star$, respectively. Data are given as raw data points, median $\pm$ 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.

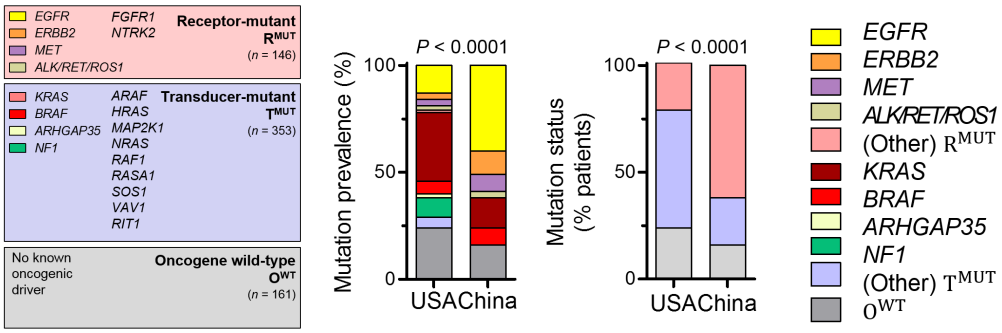
Based on this finding, we grouped US patients [9] and 101 LADC obtained from Asian patients [65] into R^{MUT}, T^{MUT}, and oncogene wild-type (O^{WT}; patient without R^{MUT}or T^{MUT}) groups, hypothesizing that these three groupings may represent distinct molecular pathways to LADC (Figure 4.3(a)). Individual mutation prevalence varied widely between East and West, translating into different frequencies of these pathways in Caucasian and Asian LADC (Figure 4.3(b)). A fact that has to be taken into account since the molecular analysis is 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

	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***
χ^2 ***P***	< 0.0001		< 0.0001			< 0.0001	

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

Figure 4.2: SNV rates, indel rates, CNA indices, smoking exposure, sex, genomic signatures of environmental carcinogen-induced base changes in the trinucleotide context, indel/SNV ratios, and transversion status of 660 patients with LADC from the USA [9] grouped by the most frequent driver mutations. Significances $P \geq 0.05$, $P < 0.05$, $P < 0.01$, and $P < 0.001$ are coded as ns, ⋆, ⋆⋆, and ⋆⋆⋆, respectively. 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).



(a) (b)

Figure 4.3: (a) Proposed grouping of US LADC patients [9] according to driver mutation into receptor-mutant (R^{MUT}), transducer-mutant (T^{MUT}), and oncogene-wild type (O^{WT}) molecular pathways. (b) Mutation rates and molecular pathway classification of 660 US LADC patients [9] and 101 LADC patients from China [65]. P, probability by χ^2 test.

distinct, while $T^{MUT}$and O^{WT}LADC were similar by all parameters examined except copy number alteration index (Figures 4.4, 4.5). 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 4.6). In the case of R^{MUT}LADC, 13 of the 18 analyzed covariables trended differently from the reference category $T^{MUT}$with high significance (Figure 4.6).

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 mismatch repair [52], and one shared between $T^{MUT}$and O^{WT}patients (hereafter referred to as T^{MUT}) with genomic marks of smoking exposure (Figure 4.7). 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 [17].

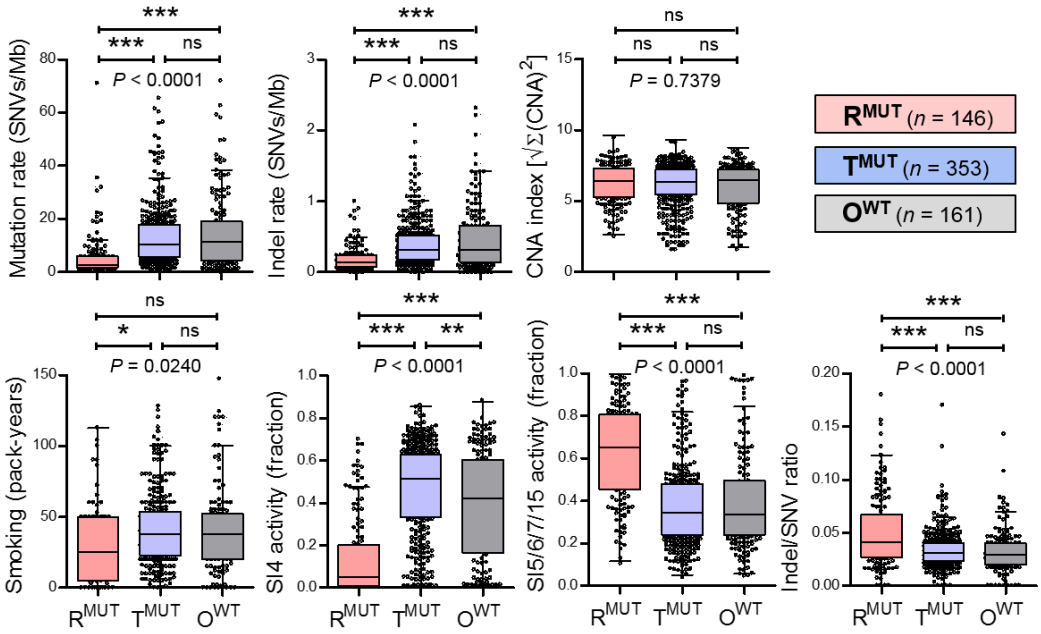


Figure 4.4: Single nucleotide variant (SNV) rates, insertion/deletion (indel) rates, copy number alteration (CNA) indices, smoking exposure, sex, genomic signatures of environmental carcinogen-induced base changes in the trinucleotide context, indel/SNV ratios, and transversion status of 660 patients with LADC from the USA [9] grouped by receptor-mutant (R^{MUT}), transducer-mutant (T^{MUT}), and oncogene-wild-type (O^{WT}) molecular pathways. Significances $P \geq 0.05$, $P < 0.05$, $P < 0.01$, and $P < 0.001$ are coded as ns, ⋆, ⋆⋆, and ⋆⋆⋆, respectively. 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

	Sex		Smoking			Transversion	
n	**Male**	**Female**	**Never**	**Former**	**Current**	**Low**	**High**
R^{MUT}	49	97^{c}	62	63	15^{c}	109	37^{c}
T^{MUT}	174	178**	14	229	83***	44	309***
O^{WT}	94	67***	17	85	46***	41	120***
χ^2 *P*	< 0.0001		< 0.0001			< 0.0001	

0-20 | 20-40 | 40-60 | 60-80 | 80-100

Frequency by pathway (%)

Figure 4.5: SNV rates, indel rates, CNA indices, smoking exposure, sex, genomic signatures of environmental carcinogen-induced base changes in the trinucleotide context, indel/SNV ratios, and transversion status of 660 patients with LADC from the USA [9] grouped by receptor-mutant (R^{MUT}), transducer-mutant (T^{MUT}), and oncogene-wild-type (O^{WT}) molecular pathways. Significances $P \geq 0.05$, $P < 0.05$, $P < 0.01$, and $P < 0.001$ are coded as ns, ⋆, ⋆⋆, and ⋆⋆⋆, respectively. 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).

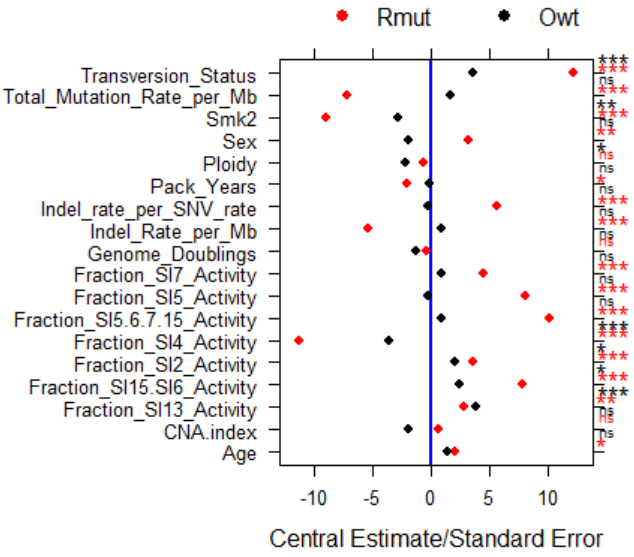


Figure 4.6: Points represent regression coefficients divided by their standard errors in univariate multinomial regression. 18 clinical and molecular variables of 660 US patients with LADC [9] stratified by molecular pathway were analyzed. Position on x-axis denotes deviation from the estimate in reference group T^{MUT}. Significance of deviation from the reference is color-coded (red: R^{MUT}; black: O^{WT}): ns, $\star$, $\star\star$, and $\star\star\star$: $P \geq 0.05$, $P < 0.05$, $P < 0.01$, and $P < 0.001$, respectively, for the indicated variables.

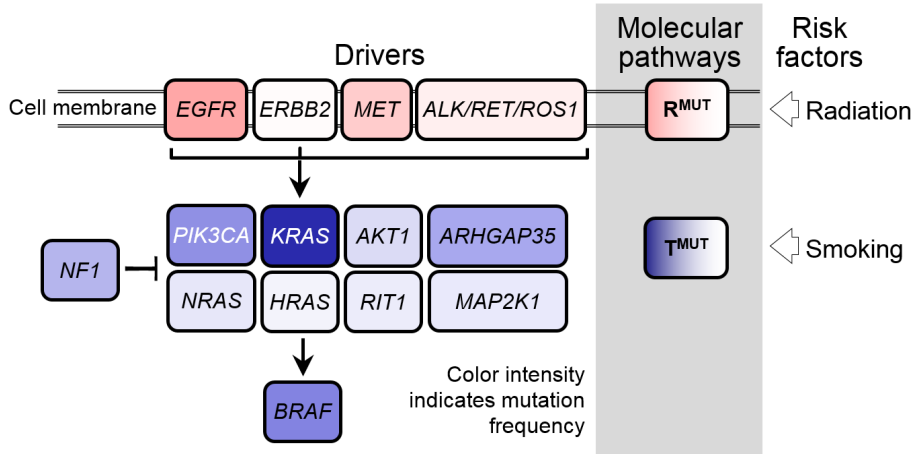


Figure 4.7: Schematic of the two proposed molecular pathways to LADC and the main risk factors for each pathway.

4.2.2 Development of the molecular mechanistic model

State-of-the-art epidemiological risk estimates from smoking and radiation exposure merely establish statistical associations without explicitly considering pathogenic processes and molecular data: molecular biology and epidemiology 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 [51].

The two molecular pathways (R^{MUT} vs. T^{MUT}) to LADC determine the conceptual model design as fundamental feature. With this constraint and due to the fact that Tomasetti et al. [58] argue that two/three driver mutations are involved in LADC, we considered only two- and three-stage clonal expansion models as candidates for both molecular pathways. The goodness of fit of the different models can be found in Table 4.3.

Table 4.3: Model pairs fitted for model selection of the R^{MUT} pathway (subscript $_R$) and of the T^{MUT} pathway (subscript $_T$).

Imputed data set no.	**AIC of last two candidate models** TSCE$_R$-3SCE$_T$	TSCE$_R$-TSCE$_T$
	AIC	
2	5076.9	**5075.6**
9	5037.4	**5036.9**
11	5082.3	**5082.2**
18	**5033.3**	5033.8
23	**4937.7**	4938.7
28	**5099.5**	5100.1
39	**4949.4**	4950.2
43	5085.5	**5084.7**
45	**5023.7**	5024.6
50	5039.1	**5038.2**
Cumulative AIC	**50364.8**	**50365.0**

To speed up the selection process, model pairs were fitted to 10 (out of 50) randomly chosen LSS data sets with imputed smoking information. Model selection was based on goodness-of-fit measured by the cumulative AIC for the 10 data sets. For the R^{MUT} pathway only a TSCE survived the test phase. For the T^{MUT} pathway a TSCE and 3SCE yielded the same AIC, when paired with the TSCE for the R^{MUT} pathway. Compared to the TSCE the 3SCE contains an additional mutational stage before clonal expansion but has the same number of parameters. The TSCE was chosen for the T^{MUT} pathway because it required substantially less computation time, nevertheless the 3SCE is also biologically plausible and cannot be excluded. The impact of this choice on the results is negligible. The conceptual design of the final preferred two-path molecular mechanistic model for LADC (M_3^{LADC}) is shown in Figure 4.8.

Smoking and radiation exposure are assumed to change biological parameters in mechanistic risk models. We tested actions on the rate ν of initiating mutations and the net clonal expansion rate γ using several functional forms: linear, linear-quadratic and linear-exponential responses. For smoking, model parameters were increased at smoking initiation and remained elevated for current smokers until end of follow-up. Baseline values were retained when past-smokers

quit. Judged by goodness-of-fit, the main biological effects of both smoking and radiation are represented by enhanced clonal expansion.
In the T^{MUT} pathway, smoking intensity (S) linearly enhances the clonal expansion rate

$$\gamma_T(S) = \alpha_T - \beta_T(S) - \mu_T = \gamma_{T0}\,[1 + g_{Sf,m}\,S\,\exp(-\kappa_{f,m}\,S)] \tag{4.3}$$

during a period of constant smoking intensity with an attenuated effect for high smoking intensity. The smoking parameters $g_{Sm,f}$ and $\kappa_{m,f}$ depend strongly on sex.
In the R^{MUT} pathway, a radiation dose D linearly enhances the clonal expansion rate

$$\gamma_R(D) = \alpha_R - \beta_R(D) - \mu_R = \gamma_{R0}\,[1 + g_R\,D] \tag{4.4}$$

after exposure for life. Since both pathways apply the same constant value of the stochasticity parameter $\delta = \alpha_R\mu_R = \alpha_T\mu_T$, increase of clonal expansion is solely caused by reduced cell inactivation β.

The MI overall point estimate of the M_3^{LADC} is presented in Table 4.4. The deviances of all 50 data sheets can be found in Figure E.1. Parameters $X_R = N\nu_R\mu_R$ and $X_T = N\nu_T\mu_T$ (where N is the number of healthy cells) are modified by city and age at exposure with the same functional form as in $Stat_{LSS}^{LADC}$ (4.1).

Table 4.4: Parameter estimates (95% CI) for the preferred mechanistic model M^3_{LADC} with 12 parameters which consists of two TSCE models pertaining to pathways R^{MUT} (subscript $_R$) and T^{MUT} (subscript $_T$). Parameter definitions correspond to Figure 4.9. Central estimates are given as mean values from 50 imputed data sets with 95% CI simulated from multi-variate normal uncertainty distributions conditioned on the parameter correlation matrix (see section 3.5). Cumulative deviance/AIC are the sum over the 50 deviances/AICs.

Parameter estimates of the preferred mechanistic model M^3_{LADC}			
		R^{MUT} pathway	T^{MUT} pathway
Name	Unit	MI Mean (95% CI)	
c_{city}		0.23 (0.11, 0.35)	
c_{agexp}		-0.24 (-0.29, -0.18)	
X_R, X_T	10^{-9} cells/yr^2	0.48 (0.11, 2.26),	4.64 (1.20, 18.9)
γ_{R0}, γ_{T0}	cells/yr	0.19 (0.16, 0.22),	0.092 (0.048, 0.128)
g_R	10^{-2} Gy^{-1}	0.58 (0.39, 0.77)	-
g_{Sf}	day/cigs	-	0.32 (0.057, 0.68)
g_{Sm}	day/cigs	-	0.086 (0.013, 0.180)
κ_f	day/cigs	-	0.14 (0.21, 0.078)
κ_m	day/cigs	-	0.031 (0.042, 0.021)
δ	10^{-7} cells/yr^2	2.73 (0.92, 8.06)	
Cumulative deviance		252520	
Cumulative AIC		253720	

The cumulative AIC from 50 imputed data sets is 370 points lower compared to $Stat_{LSS}^{LADC}$ (corresponding to 7.4 point per data set). M^3_{LADC} would hence be preferred against $Stat_{LSS}^{LADC}$.

M_3^{LADC} clearly revealed the two molecular pathways (R^{MUT} versus T^{MUT}) in observational incidence data of the LSS (Figure 4.9) although no genomic information of the LSS is available. Crude rate and predicted hazard (LADC cases in 10,000 persons per year) from M_3^{LADC} were plotted in 5 year-age groups from 40-45 up to 80-85 years. The hazard of R^{MUT}-related LADC cases peaks at age 70 yr, while the hazard in the T^{MUT} pathway becomes dominant at old ages. For cigarette smoke, clonal expansion in the T^{MUT}pathway was identified as the main biological target: smoking-related inactivation of initiated cells increased the net clonal growth rate γ_T for pre-neoplastic lesions. Sex-specific response curves exhibited markedly different shapes (Figure 4.10). For men (Figure 4.10 upper panel), the growth rate increased almost linearly up to a smoking intensity of 20 cigarettes/day and flattened thereafter. Clonal growth in women (Figure 4.10 lower panel) reacted much stronger to low-smoking intensity. The growth reduction after a peak at about 10 cigarettes/day is biologically not plausible but might be caused by a reporting bias. 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 pointing to lifelong radiation-induced inflammation caused by genetic damage.

Next we analysed 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 (Table 4.5). Refined resolution in exposure subgroups of low (5-100 mGy) and moderate (100+ mGy) radiation dose, and light (1-10 cigs/day), moderate (11-20 cigs/day) and heavy (20+ cigs/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. Interestingly, 60% of the total number of cases were estimated as spontaneous cases. A value that is comparable with the study presented by Takamochi et al. [57]. Only 6% of the total number of cases could be defined as radiation induced, compared to 34% for smoking. The baseline includes again the major amount of cases in the R^{MUT} pathway, where in the T^{MUT} the 80% of the cases were estimated as smoking induced. Please 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.

Summarizing, the main impact of smoking and radiation took effect in distinct molecular pathways without noticeable synergistic effects. Now we proceed with risk assessment.

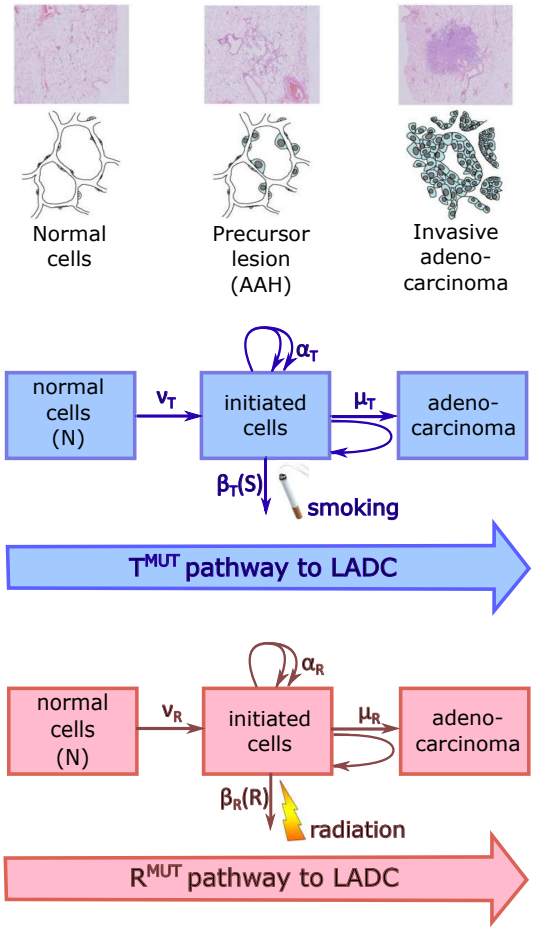


Figure 4.8: Top: Histological progression from normal cells over atypical adenomatous hyperplasia (AAH) as precursor lesions to invasive LADC [modified figure from Yatabe et al. [66]]. Bottom: Model implementation with two distinct molecular pathways pertaining to either T^{MUT} or R^{MUT} with two versions of the TSCE model. Boxes represent cells in states with defined molecular properties. Arrows represent rates of transition between cell states. Both agents of smoking and radiation cause the acceleration of clonal expansion by reduced cell inactivation. See model details and mathematical model derivation in Chapter 3.4. Parameter estimates are given in Table 4.4. The model algorithm and model implementation can be found in Appendix C.3 and Table D, respectively.

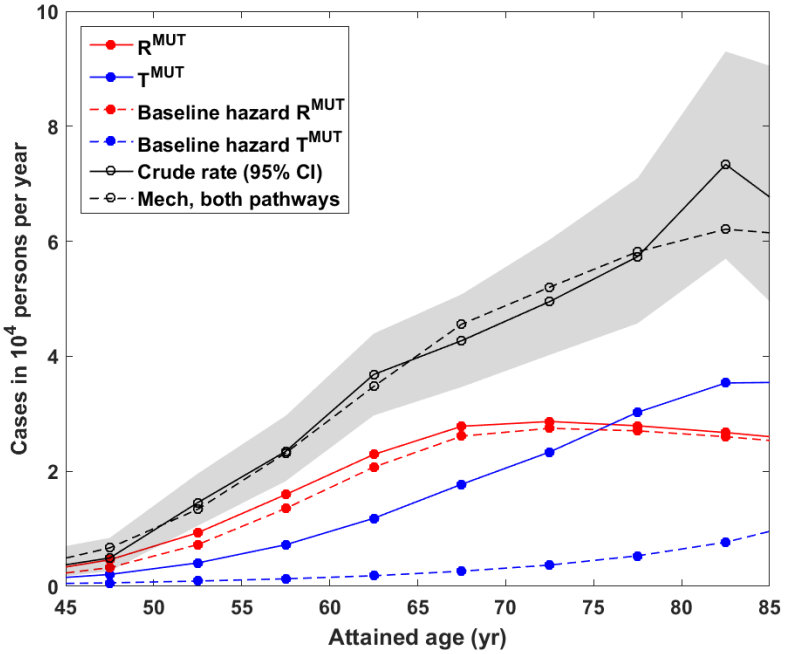
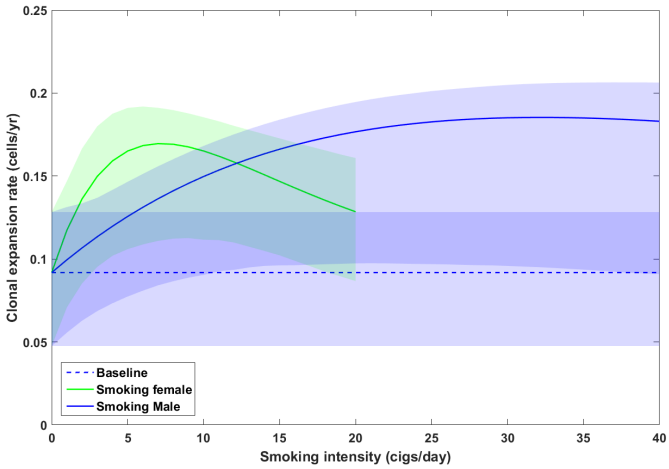
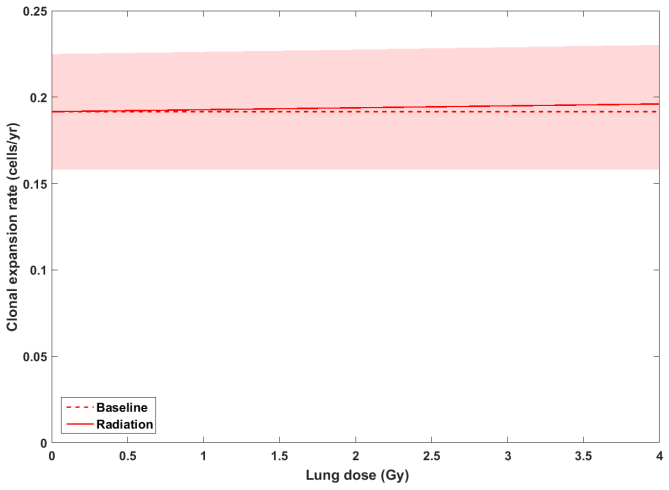


Figure 4.9: Crude rate and predicted hazard (LADC cases in 10,000 persons per year) from the preferred mechanistic model (Mech) for the LSS cohort in 5 year-age groups from 40-45 up to 80-85 years. The model clearly distinguishes pathway-specific hazards. The hazard of R^{MUT}-related LADC cases peaks at age 70 yr. The hazard in the T^{MUT} pathway becomes dominant at old ages. This is a model prediction of the LSS cohort without any genomic data.



(a) T^{MUT} pathway



(b) R^{MUT} pathway

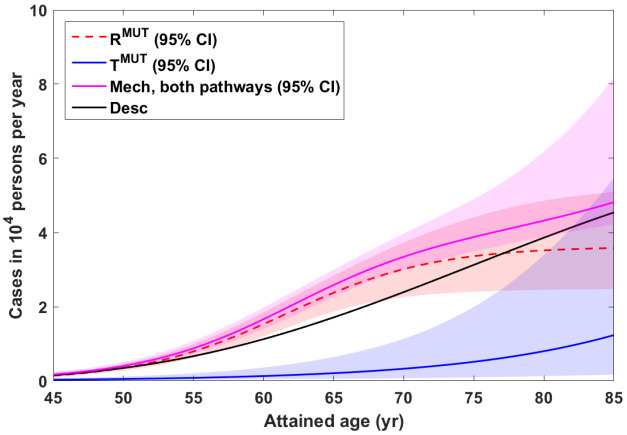
Figure 4.10: Clonal expansion rates for the two pathways T^{MUT} and R^{MUT} in M^3_{LADC}. (A) In the T^{MUT} pathway smoking intensity smkint linearly enhances the clonal expansion rate γ_T with an attenuated effect for high smoking intensity. The implausibly strong attenuation of the clonal expansion rate for females smoking more the 10 cigs/day is possibly caused by a reporting bias. (B) In the R^{MUT} pathway a radiation dose D linearly enhances the clonal expansion rate $\gamma_R(D)$, which remains permanently elevated after exposure for the whole life.

Table 4.5: M_3^{LADC} estimates for the breakdown of 636 LADC cases (% of 636 cases) 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 cigs/day), moderate (11-20 cigs/day) and heavy (20+ cigs/day) smoking intensity is made. 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.

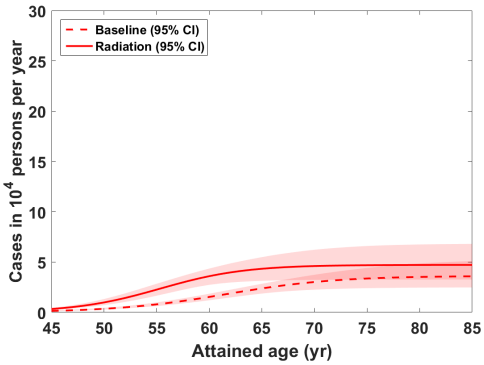
Radiation intensity (mGy)	Smoking intensity (cigs/day)	Observed cases (%)	Estimated cases (%)	R^{MUT} estimation (%)	Radiation induced estimated cases (%)	T^{MUT} estimation (%)	Smoking induced estimated cases (%)	Spontaneous estimated cases (%)
0-5	=0	**137 (21)**	**139 (22)**	**120 (19)**	**0 (0)**	**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)**

4.2.3 Risk assessment

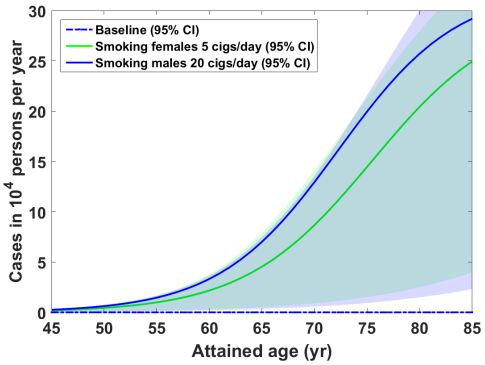
Before presenting EARs and ERRs for $M_3^{LADC}$and for $Stat_{LSS}^{LADC}$, let us start with a comparison of baseline hazard rates and hazard rates between the two molecular pathways.



(a) Baseline hazards



(b) Baseline hazard and hazard from radiation exposure in the R^{MUT} pathway



(c) Baseline hazard and hazard from smoking in the T^{MUT} pathway

Figure 4.11: (A) Baseline hazards in pathways R^{MUT} and T^{MUT} for radiation-induced LADC in the LSSCt. To eliminate the influence for city of residence person-year weighted city means are used. For comparison with the baseline hazard from $Stat_{LSS}^{LADC}$ (Desc) the total baseline hazard (as the sum of pathway-specific hazards) from the preferred mechanistic model M_{LADC}^3 is shown. (B) Baseline hazard and hazard from radiation exposure in the R^{MUT} pathway for a person exposed at 30 yr to a lung dose of 1 Gy (C) Baseline hazard and hazard from smoking in the T^{MUT} pathway for lifelong smokers starting at age 20 yr with smoking intensity 20 cigs/day (male) and 5 cigs/day (female).

From Figure 4.11(A) we can see that the total baseline hazard (as the sum of pathway-specific hazards) from the preferred mechanistic model M_{LADC}^3 and from the state-of-the-art statistical risk model 4.1 are of similar magnitude, where the blue line (the baseline of the T^{MUT} since no radiation-induce cases arise in this pathway) is really small. In Figure 4.11(B) the amount of case coming from radiation, compared to the baseline cases, is really small. Figure 4.11(C) the baseline is so small that it is almost not visible. The curve for female and males smoker are very similar, although the smoking intensity is four time higher for males.

We can also notice that the hazard rates for smokers are much higher than those for the R^{MUT}, indicating a strong relation between smoking and LADC. Another difference between the two pathways is the flattening of R^{MUT}, which is not present in the T^{MUT}.

Since for risk assessment the biological action presented in $M_3^{LADC}$of the previous chapter is better reflected in the EARs compared to ERRs, all respective ERRs will not be discussed in the main part of this Thesis but can be found in Appendix E.

4.2.3.1 Excess absolute rates for radiation

In the case of a two path model, where the pathways are mutually exclusive, the radiation-related EAR has the following form (to compare with equations 3.35)

$$EAR_{rad} = h_{tot}(n\,Gy, \cdot\frac{sigs}{day}) - h_{tot}(0\,Gy, \cdot\,\frac{sigs}{day}) \tag{4.5}$$

$$= h_{R^{MUT}}(n\,Gy) + h_{T^{MUT}}(\cdot\,\frac{sigs}{day}) - h_{R^{MUT}}(0\,Gy) - h_{T^{MUT}}(\cdot\,\frac{sigs}{day}) \tag{4.6}$$

$$= h_{R^{MUT}}(n\,Gy) - h_{R^{MUT}}(0\,Gy) \tag{4.7}$$

with $n > 0$, where h_{tot}, $h_{R^{MUT}}$ and $h_{T^{MUT}}$ represents the total hazard of the model, the hazard in the R^{MUT} pathway and the hazard in the T^{MUT} pathway, respectively. This formula applies only because radiation and smoking act separately on two different pathways independently.

Figure 4.12 depicts the EAR (as cases in 10,000 persons per year) for radiation-induced LADC for a person exposed at 30 yr. 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 (see Figure 4.10). To eliminate the influence for city of residence, person-year weighted city means are used. Figure 4.12(a) presents the bivariate EAR dependence on attained age and lung dose. The radiation risk maximizes at about 55 years for high lung dose. Figure 4.12(b) shows 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. Figure 4.12(c) instead shows 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. In each plot the EAR from $Stat_{LSS}^{LADC}$ (Desc) is shown for comparison.

The corresponding radiation-related ERR can be found in Appendix E.2, Figure (E.2).

4.2.3.2 Excess absolute rates for smoking

As for the radiation-related EAR, also the smoking-related EAR has a special structure in the case of a two path model, where the pathways are mutually exclusive

$$EAR_S = h_{tot}(\cdot\,Gy, m\,\frac{cigs}{day}) - h_{tot}(\cdot\,Gy, 0\,\frac{cigs}{day}) \tag{4.8}$$

$$= h_{R^{MUT}}(\cdot\,Gy) + h_{T^{MUT}}(m\,\frac{cigs}{day}) - h_{R^{MUT}}(\cdot\,Gy) - h_{T^{MUT}}(0\,\frac{cigs}{day}) \tag{4.9}$$

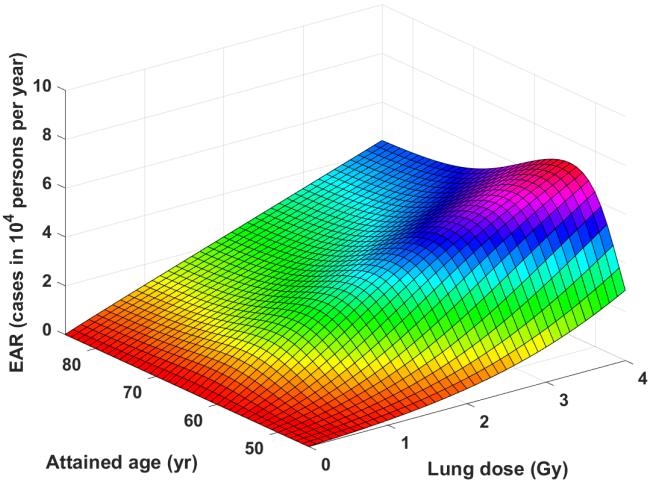
$$= h_{T^{MUT}}(m\,\frac{cigs}{day}) - h_{T^{MUT}}(0\,\frac{cigs}{day}) \tag{4.10}$$

with $m > 0$, where h_{tot}, $h_{R^{MUT}}$ and $h_{T^{MUT}}$ represents the total hazard of the model, the hazard in the R^{MUT} pathway and the hazard in the T^{MUT} pathway, respectively. This formula applies

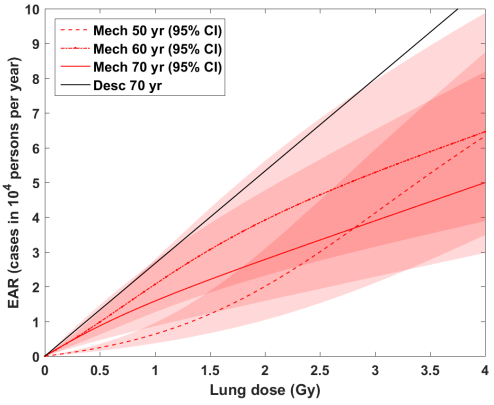
only because radiation and smoking act on two different pathways independently.

Figure 4.13 depicts the smoking-related EAR (as cases in 10,000 persons per year) for smoking-induced LADC for lifelong smokers starting at age 20 yr. 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 (see Figure 4.10). To eliminate the influence for city of residence person-year weighted city means are used. The bivariate EAR dependence on attained age and smoking intensity for female and male smokers is presented Figures 4.13(a) and 4.13(b), respectively. The implausibly strong attenuation of the clonal expansion rate for females smoking more the 10 cigs/day is possibly caused by a reporting bias. Panels (c) and (d) depict cross-sectional cuts to panels (a) and (b) for attained ages of 50, 60 and 70 yr. Panels (e) and (f) depict cross-sectional cuts to panels (a) and (b) for 5 cigs/day (males and females) and 20 cigs/day (males only). Female smokers of 5 cigs/day and male smokers of 20 cigs/day possess about the same risk. The EAR from $Stat^{LADC}_{LSS}$ (Desc) is shown for comparison.
The corresponding smoking-related ERR can be found in Appendix E.3, Figure (E.3).

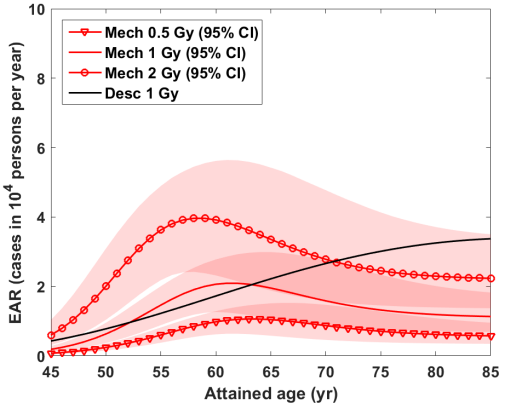
In Figures E.4 (females) and E.5 (males) the additive effect or radiation and smoking on the EAR is shown. The effect is additive since radiation and smoking act on differently pathways without any synergistic effect. Comparing Figures E.4 and E.5 with Figures 4.13(a) and (b) only the small increase in the EAR coming from radiation is detectable.



(a)



(b)



(c)

Figure 4.12: EARs (as cases in 10,000 persons per year) from M_3^{LADC} (Mech) for radiation-induced LADC for a person exposed at 30 yr. (a) Bivariate EAR dependence on attained age and lung dose. (b) Cross-sectional cuts to panel (a) for attained ages 50, 60 and 70 years. (c) Cross-sectional cuts to panel (a) for lung doses 0.5, 1 and 2 Gy. The EAR from $Stat_{LSS}^{LADC}$ (Desc) is shown for comparison.

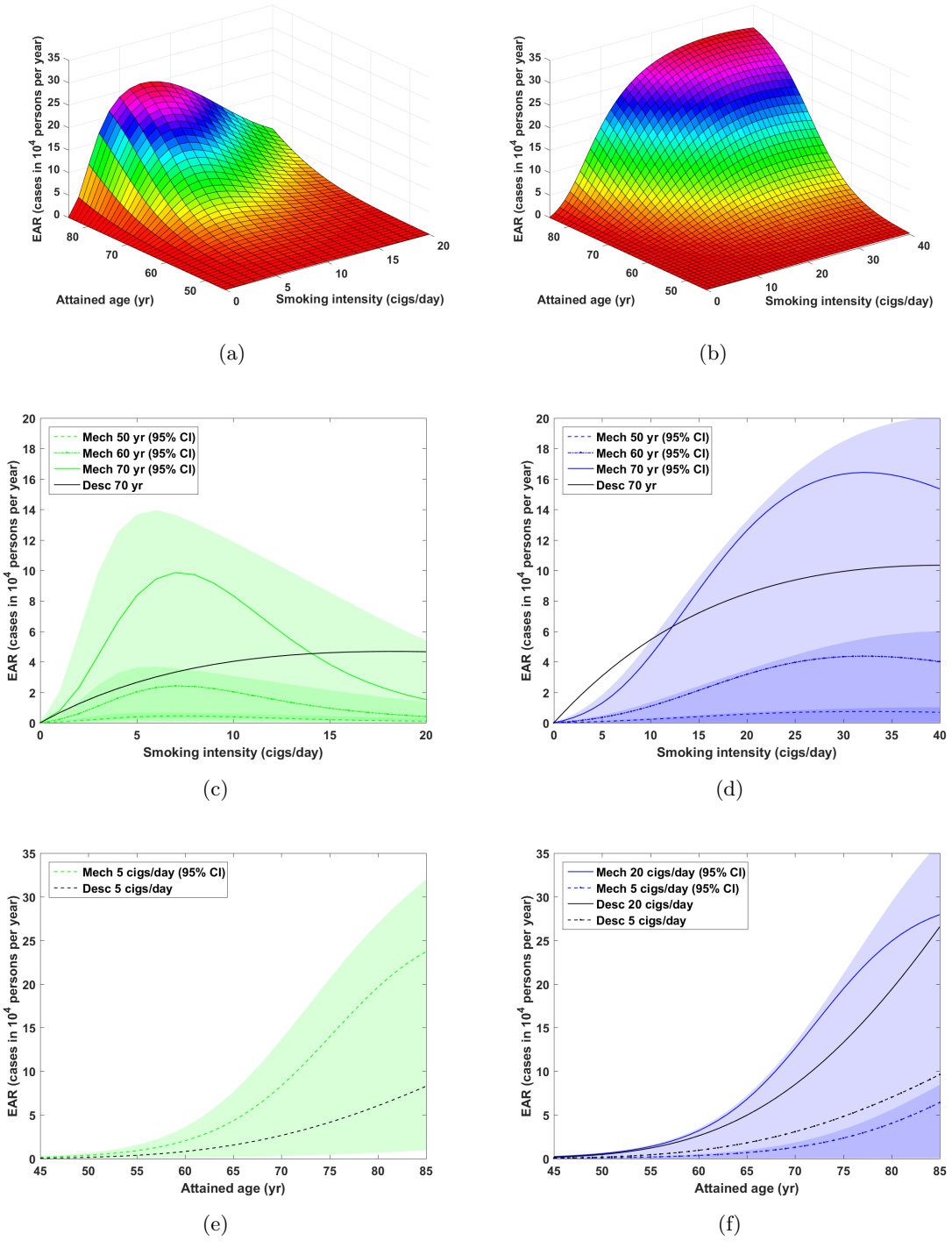


Figure 4.13: EARs (as cases in 10,000 persons per year) from M_3^{LADC} (Mech) for smoking-induced LADC for lifelong smokers starting at age 20 yr. Bivariate EAR dependence on attained age and smoking intensity for female smokers (a) and male smokers (b). Panels (c) and (d) depict cross-sectional cuts to panels (a) and (b) for attained ages of 50, 60 and 70 yr. Panels (e) and (f) depict cross-sectional cuts to panels (a) and (b) for 5 cigs/day (males and females) and 20 cigs/day (males only). The EAR from $Stat_{LSS}^{LADC}$ (Desc) is shown for comparison.

4.3 Generalized additive models

In model $Stat_{LSS}^{LADC}$ (4.2) the functional forms of attained age in the baseline and of radiation and smoking ERRs are postulated to be specific functions (linear quadratic and linear, respectively). For a more comprehensive analysis of those variables, a GAM for for LADC in the LSS (GAM_{LSS}^{LADC}) was developed on the basis of model $Stat_{LSS}^{LADC}$ presented in the previous Section 4.1. The preferred GAM_{LSS}^{LADC} has the following form

```
LADC_gam <- gam(adeno~1+I(city-1)+e30+s(I(age/70),d10gy)+s(I(age/70),
                I(packyrs/50))+I(smkyrs/50)+smkqyrs,
            offset = log(0.0001*PYR),
            data = imput_dat_39,
            family = poisson(link = "log"), method = "ML",
            optimizer = c("outer","newton"))
```

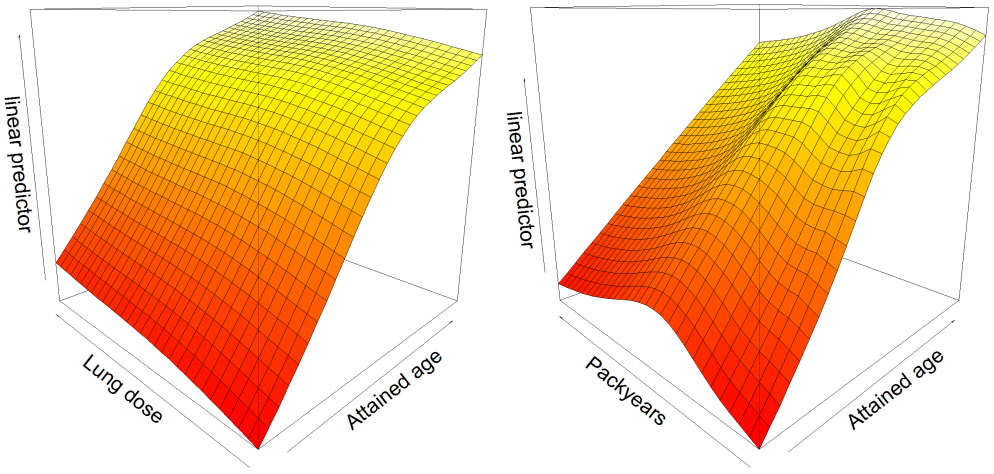
In model GAM_{LSS}^{LADC} parameters for baseline, city of origin and age at exposure ($e30$) were fitted linearly. Sex did not have any statistical significance. The variables for years smoked ($smkyrs$) and for years since quitting ($smkqyrs$) turned out to have a linear response. Two penalised splines were fitted to explore the functional form of radiation dose and attained age ($s(I(age/70), d10gy)$) and of cumulative smoking amount ($packyrs$) and attained age. In Table 4.6 the central estimates as means (Standard deviation) from 50 imputed data sets are given together with the total deviance/AIC (see chapter 3.5). The AIC was calculated following definition (3.11) for linear terms. The deviances of all 50 data sheets can be found in Figure E.1. For smoothing functions the contribution is given by $2 \cdot edf$, where is edf, the estimated degrees of freedom, describes the complexity degree of the fitted spline. Only the edf parameters show a large standard deviation. Boxplots of the 50 values are presented in Appendix F.1, Figure F.1. This fact has to be attributed to the variability of the multiple imputation process and to the flexibility of splines, that adapt to different data sets. The cumulative AIC from 50 imputed data sets is 999 points higher compared to M_3^{LADC} (corresponding to ca. 20 points per data set). M_3^{LADC} would also in this case be the preferred model.

Table 4.6: Parameter estimates for the GAM_{LSS}^{LADC}. Central estimates are given as means from 50 imputed data sets with the standard deviation between the 50 best estimates (see chapter 3.5). edf, the estimated degrees of freedom, describe the complexity degree of the fitted polynomial.

Parameter estimates of GAM_{LSS}^{LADC}	
Meaning	Mean (Standard deviation)
Intercept (10^{-3})	-3.14 (5.72)
City	0.23 (1.43 10^{-5})
Age at exposure	-0.26 (2.12 10^{-5})
Smoking duration (10^{-1})	9.24 (0.23)
Years since quitting (10^{-2})	-1.88 (1.92 10^{-3})
edf $s(I(age/70), d10gy)$	8.57 (25.23)
edf $s(I(age/70), I(packyrs/50))$	6.50 (29.73)
Cumulative deviance	252713
Cumulative AIC	254719

To underline the features of the GAM_{LSS}^{LADC} plots of linear predictions of age and lung dose and of age and cumulative smoking amount are presented in Figure 4.14. The imputed dataset 39 was chosen here as an example. In Appendix F, Figure F.4, the respective functions of age and lung dose $s(I(age/70), d10gy)$ and of age and smoking amount ($packyears$) $s(I(age/70), I(packyrs/50))$

are presented. Figure 4.14(a) shows that the linear predictor is not linear in age, nor in lung



(a) (b)

Figure 4.14: Results of GAM_{LSS}^{LADC} for the imputed data set 39. (a) The linear prediction (a) as a function of age and lung dose and (b) as a function of age and smoking amount (packyears).

dose but has a monotonic increasing shape. From Figure 4.14(b) we can notice again the monotonic increasing in age, but a peak in the cumulative smoking amount variable. The functional form of age is well modelled by the quadratic form of $Stat_{LSS}^{LADC}$, while the peak in the in the cumulative smoking amount variable would be missed.

To compare the model results of GAM^{LADC} predictions for different exposure scenarios were considered: unexposed never smokers (the baseline), exposed never smokers, unexposed smokers and exposed ever smokers. Radiation/smoking related hazards and baseline hazards are presented in Figure 4.15 for the scenarios. The boxplots represent the model predictions of the 50 imputed data sheets for LADC.

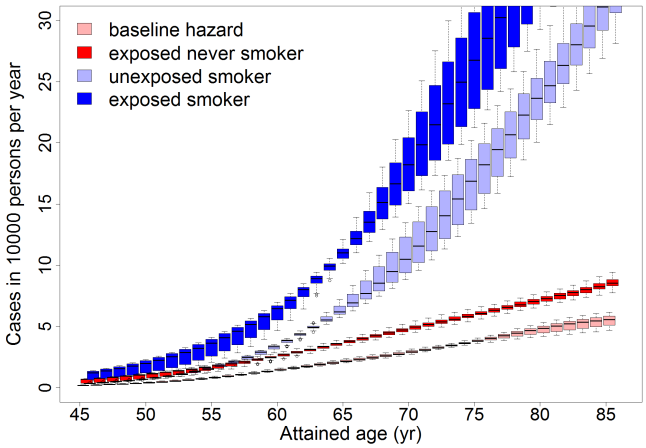
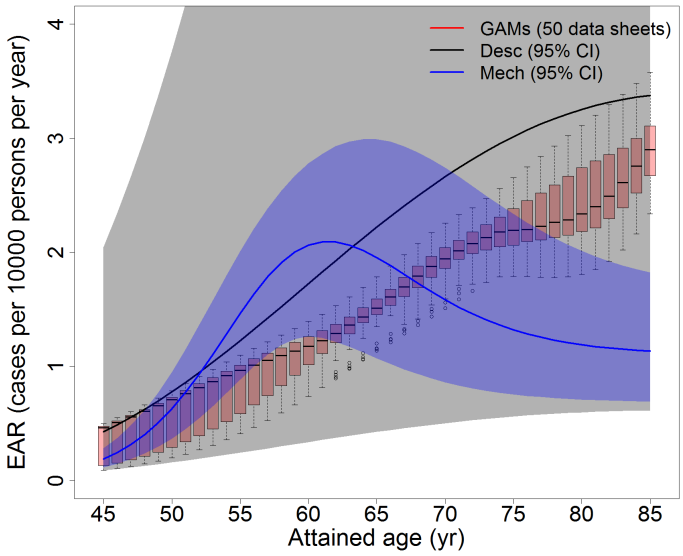


Figure 4.15: Baseline hazard and hazards from radiation and/or smoking for radiation-smoking induced LADC in the LSSC predicted from GAM_{LSS}^{LADC}. To compare with Figure 4.11. Age at exposure was fixed at age 30 yr and lung dose to 1 Gy. A smoking person began to smoke at age 20 yr and never stopped. The boxplots represent the variance of the 50 data sets.

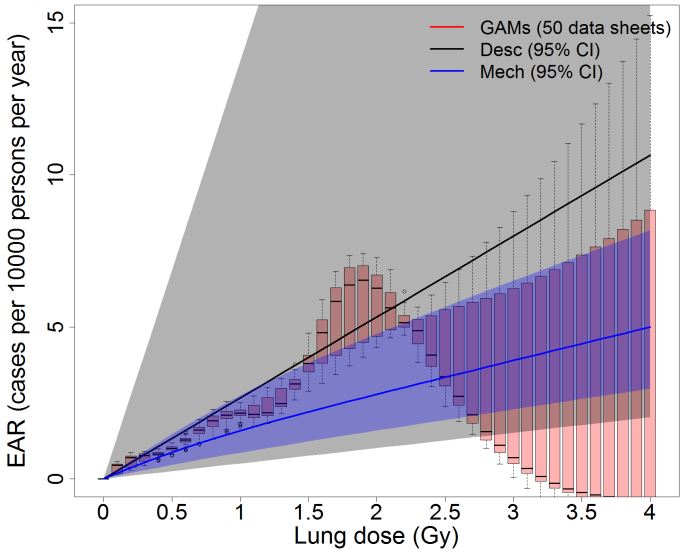
The curves of Figure 4.15 are similar to the results of M_3^{LADC} presented in Figure 4.11. The

increase of radiation-related hazards compared to the baseline hazard is indeed a bit higher in GAM_{LSS}^{LADC}. The smoking-related hazard is also predicted higher in GAM_{LSS}^{LADC}than in M_3^{LADC}.

Now we proceed with the analysis of excess risks. EARs for radiation and smoking are presented in Figures 4.16 and 4.17 (the respective ERRs can be found in Appendix F.2, Figures F.2 and F.3).
Both radiation and smoking EAR for the GAM_{LSS}^{LADC} are similar to the results of $Stat_{LSS}^{LADC}$ (see Figures 4.12 and 4.13 for more details). The peak modeled from M_3^{LADC} is however also not present in GAM_{LSS}^{LADC}.

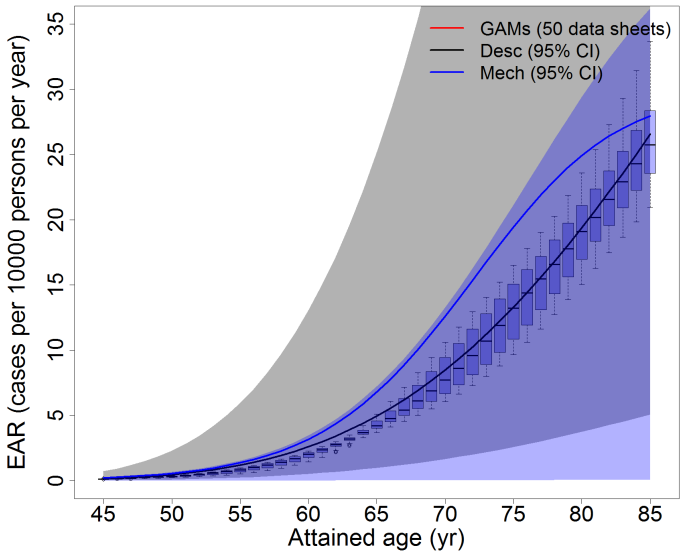


(a) EAR from GAM_{LSS}^{LADC} for radiation induced LADC as a function of attained age (yr)

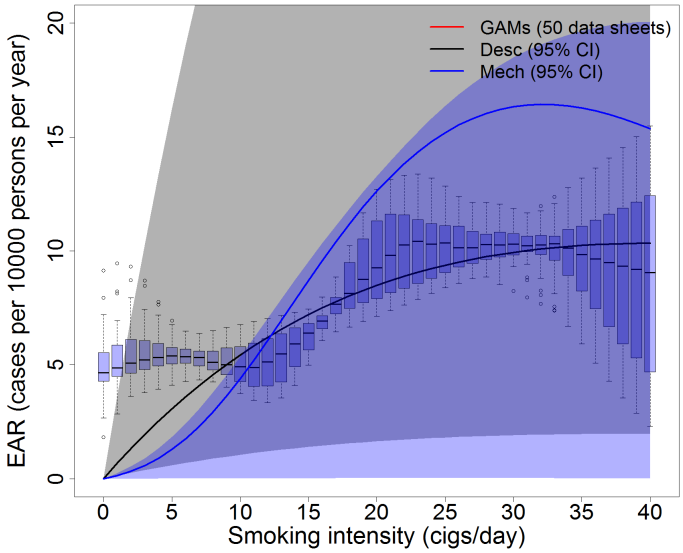


(b) EAR from GAM_{LSS}^{LADC} for smoking induced LADC as a function of lung dose (Gy)

Figure 4.16: Excess absolute rates (EARs as cases per 10.000 persons per year) from GAM_{LSS}^{LADC} for radiation-induced LADC for (a) a person exposed to 1 Gy lung dose at age 30 yr as a function of age, (b) a 70 years old person exposed at age 30 yr as a function of lung dose (Gy). The boxplots represent the variance of the 50 datasets by GAM_{LSS}^{LADC}. $Stat_{LSS}^{LADC}$ is presented in black, while M_3^{LADC} in blue for comparison.



(a) EAR from GAM_{LSS}^{LADC} for smoking induced LADC as a function of attained age (yr)



(b) EAR from GAM_{LSS}^{LADC} for smoking induced LADC as a function of smoking intensity (cigs/day)

Figure 4.17: Excess absolute rates (EARs as cases per 10.000 persons per year) from GAM_{LSS}^{LADC} for smoking-induced LADC for a current smoker that began at age 20 yr, never stopped (a) with smoking amount of 1 packyear as a function of attained age, (b) with varying smoking intensity. The boxplots represent the variance of the 50 datasets by GAM_{LSS}^{LADC}. $Stat_{LSS}^{LADC}$ is presented in black, while M_3^{LADC} in blue for comparison.

4.4 Summary of results

Result 1: LADC in the LSSC

In this Chapter we analysed the outcome of LADC in the LSSC and following results could be achieved

- Detection of two molecular pathways to LADC applying biological analysis to the Campbell data [9]: the R^{MUT} and T^{MUT} pathways
- Detection of R^{MUT} and T^{MUT} pathways using the molecular mechanistic model M_3^{LADC}
- Radiation and cigarette smoke exposure affect R^{MUT} and T^{MUT} pathways separately
- No synergistic effects of radiation and cigarette smoke exposures could be detected, neither in the state-of-the-art statistical risk model $Stat_{LSS}^{LADC}$ nor in model M_3^{LADC}
- Detailed functional description of effects using GAMs. An interaction of smoking and radiation ($s(smoking, radiation)$) was tested but resulted not significant.

CHAPTER

5

LUNG SQUAMOUS CELL CARCINOMA IN THE LIFE SPAN STUDY COHORT

This chapter is dedicated to the analysis of squamous cell carcinoma in the Life Span Study cohort. Of major interest is the understanding of the complex influence of smoking on this disease.
In the first section a state-of-the-art statistical risk model is developed and explained. Of fundamental importance is the finding that only the younger calendar year category was found radiation sensitive, with only 11 cases. Since it is not plausible that the radiation risk estimation of the whole cohort is based on one specific category with so less cases, this category was excluded from the analysis. A radiation related risk estimation for this type of lung cancer was possible but not significant.
The second part is about mechanistic risk models, that describe carcinogenesis on a cellular level. An effect of smoking could be found in initiation and promotion, with sex dependence for the latter process. For past smokers an extra gender-independent parameter for promotion could be found, that indicated a partial recovery of lung tissue.
Finally, generalised additive models were hence developed, to explore the functional form of the different effects.

5.1 State-of-the-art statistical risk models

State-of-the-art statistical risk models are the gold standard in radiation epidemiology and for each analysis this type of models are interesting to develop, since for each cohort different features of the models explained in Section 3.3.2 can be examined. For SQUAM in the LSS the state-of-the-art statistical risk models are based on [18, 21]. Looking at Table 4 in Egawa et al. [18] we noticed that for SQUAM the sex-averaged radiation-related ERR is relatively small, suggesting weak statistical support. We also notice an unexpected high value for the *age at exposure* parameter. At the same time the *attained age* parameter has a markedly negative value. This means that the radiation risk is maximal after exposure at young age. The combined dependence on *age at exposure* and *attained age* can be interpreted as a calendar year effect. To explore this hypothesis we analysed the original complete data set and first reproduced the results of Furukawa et al. [21] and Egawa et al. [18]. Than we did the same analysis as in [18] only for SQUAM but deleting all Poisson records containing the first calendar year category (from January, 1st 1958 to December 31st 1960, cal1). Only 11 cases were here excluded from the analysis, which all had unknown smoking information. Deviances and AICs can be found in Table 5.1. The model used for the analysis of Table 5.1 is the *Simple Additive Model* of

Table 5.1: Deviance and AIC from the state-of-the-art statistical risk model applied to the not imputed original dataset (second column) and to the not imputed original dataset without the first calendar year category, cal1 (third column). Five models were applied: only the baseline model, the baseline model with an extra parameter for the radiation-ERR, the baseline model with radiation-ERR and dose modifiers, the baseline model with only the full smoking function and finally the simple additive model. For a detailed description see model (G.1) *in Appendix G.1*

	Total dataset			**Dataset nor cal1**		
	Dev	**AIC**	**Δ Dev**	**Dev**	**AIC**	**Δ Dev**
Baseline	2757	2777		2667	2687	
Rad. ERR	2748	2770	9	2661	2683	6
Rad. fucntion	2734	2762	15	2655	2683	12
Smk. only	2586	2630	171	2499	2543	168
simple Add.	2566	2618	191	2492	2544	175

Egawa et al. [18]. We choose this model with dose effect modifiers, since we wanted to see if the effects of *attained age* and *age at exposure* changed without the first calendar year category. An explanation of the model can be found in Appendix G.1, model (G.1).
For the data set without cal1 the model fitting only the smoking function would be preferred under the AIC. Together with K. Furukawa we concluded that it is implausible that the total radiation response of a cohort can be supported by only 11 cases. The complete radiation-risk analysis would be determined by one category. We therefore proceed the analysis of SQUAM excluding people with calendar year cal1. A radiation-related risk analysis was performed for SQUAM in the LSS, but without significant results.

The derivation of the preferred statistical model for SQUAM in the LSS cohort was done using one imputed data sheet. The sequence of steps for this derivation is presented in a reduced form in Table G.1. The preferred state-of-the-art statistical risk model for SQUAM in the LSS

($Stat_{LSS}^{SQUAM}$) ended up to have the following form

$$h^{SQUAM_{LSS}} = h_0^{SQUAM_{LSS}} \cdot (1 + ERR_{smk}^{SQUAM_{LSS}}) \tag{5.1}$$

$$h_0^{SQUAM_{LSS}} = e^{\beta_0 + \beta_1 log(\frac{age}{70}) + \beta_{2F/3M} log^2(\frac{age}{70})} \tag{5.2}$$

$$ERR_{smk}^{SQUAM_{LSS}} = \beta_4 \frac{packyrs}{50} \cdot e^{\beta 5 log(years\ Quitting+1)}, \tag{5.3}$$

with the parameters reported in Table 5.2. The deviances of all 50 data sheets can be found in Figure G.1.

Table 5.2: Parameter estimates for $Stat_{LSS}^{SQUAM}$ 5.3 with 6 parameters. Central estimates are given as means from 50 imputed data sets with 95% CI simulated from multi-variate normal uncertainty distributions conditioned with the parameter correlation matrix (see section 3.5). Cumulative deviance/AIC are the sum over the 50 deviances/AICs.

Parameter estimates of $Stat_{LSS}^{SQUAM}$ 5.3		
Name	Meaning	MI Mean (95% CI)
β_0	baseline	-0.55 (-0.91, -0.21)
β_1	baseline, attained age	4.91 (4.26, 5.55)
β_{2F}	baseline fem., attained age (squared)	-23.50 (-32.52, -14.69)
β_{3M}	baseline mal., attained age (squared)	-10.33 (-13.96, -6.73)
β_4	smoking, linear resp. ($\frac{day \cdot yr}{packs}$)	22.52 (14.41, 32.91)
β_5	smoking, years since quitting	-0.29 (-0.43, -0.16)
Deviance	145396	
AIC	145996	

We could detect a markedly sex-dependent attained age response. Only the years since quitting modifier was found significant for the smoking ERR. Results will be visualised together with the mechanistic model M_2 for SQUAM (M_2^{SQUAM}) in Chapter 5.2.

5.2 Mechanistic model

For the development of M_2^{SQUAM} we did not consider multiple pathways for two reasons: firstly, 90% of DQUAM patients were smokers, suggesting a single pathway of cancer development; secondly, we found no hint for multiple pathways in the molecular data (see Chapter 1.1.2, Figure 1.4) although a through analysis is still lacking. We looked however for a more detailed description of the smoking damage. Compared to LADC this was only possible because in SQUAM the statistical power for smoking influence was much higher, since almost all people were smokers (see Table 2.3 in blue). A description of the preferred mechanistic model M_2^{SQUAM} can be found in Figure 5.1.
M_2^{SQUAM} is a TSCE with smoking responses in the clonal expansion γ_{SQUAM} and in the initiation X_{SQUAM}, with the following forms

$$\gamma_{SQUAM} = \gamma_{0f/m} \cdot (1 + \gamma_{Sf/m} \cdot smkint \cdot e^{\kappa_{f/m} \cdot smkint} + \gamma_{past} \cdot packyears), \tag{5.4}$$

$$X_{SQUAM} = X_0 \cdot (1 + X_S \cdot smkint). \tag{5.5}$$

Pack years are defined as *cigarette packs* (one pack contains 20 cigarettes) times the years smoked. Figure 5.2 depicts parameters γ_{SQUAM} and X_{SQUAM}. In γ_{SQUAM} smoking intensity (*smokint*) linearly enhances the clonal expansion rate during a period of constant smoking

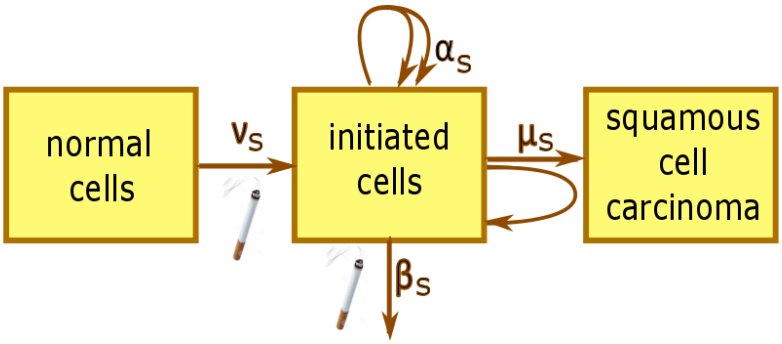


Figure 5.1: Schematic description of the preferred mechanistic model M_2^{SQUAM}. An effect of smoking on initiation and promotion could be found with the following forms: $\gamma = \gamma_{0f/m} \cdot (1 + \gamma_{Sf/m} \cdot smkint \cdot e^{\kappa_{f/m} \cdot smkint} + \gamma_{past} \cdot packyears)$ and $X = X_0 \cdot (1 + X_S \cdot smkint)$, respectively.

intensity with an attenuated effect for high smoking intensity. Note, that this is the same functional form found in the clonal expansion of M_3^{LADC}, T^{MUT} pathway. Baseline values were not retained when past-smokers quit, but an extra parameter (γ_{past}) linearly dependent the cumulative smoking amount (packs) could be fitted. The initiation parameter X_{SQUAM} depends linearly on smoking intensity. Only parameters of γ_{SQUAM} were sex dependent. An effect for radiation was tested but was not significant.

Parameter estimates of M_2^{SQUAM} can be found in Table 5.3. The deviances of all 50 data sheets can be found in Figure G.1. The cumulative deviance from the 50 imputed data sets is 92 points

Table 5.3: Parameter estimates for the preferred mechanistic model M_2^{SQUAM} with 10 parameters. Central estimates are given as means from 50 imputed data sets with 95% CI simulated from multivariate normal uncertainty distributions conditioned with the parameter correlation matrix (see section 3.5). Cumulative deviance/AIC are the sum over the 50 deviances/AICs.

Parameter estimates of the mechanistic model M_2^{SQUAM}		
Name	Unit	MI Mean (95% CI)
γ_{0f}	cells/yr	0.17 (0.12, 0.23)
γ_{0m}	cells/yr	0.22 (0.15, 0.29)
γ_{Sf}	day/cigs	0.14 (0.09, 0.20)
γ_{Sm}	day/cigs	0.035 (0.010, 0.066)
κ_f	day/cigs	-0.060 (-0.083, -0.037)
κ_m	day/cigs	-0.034 (-0.048, -0.019)
γ_{past}	$\frac{packs}{day \cdot yr}$	-0.49 (-0.73, -0.25)
δ	10^{-7}cells/yr^2	0.443 (0.025, 6.143)
X_0	10^{-10}cells/yr^2	0.393 (0.016, 7.424)
X_S	day/cigs	0.49 (0.23, 0.86)
Deviance		145478
AIC		146478

higher compared to $Stat_{LSS}^{SQUAM}$ (corresponding to 1.8 points per data set). The cumulative AIC is 482 points higher, corresponding to 10 points per data set. The cumulative deviances are very similar but the cumulative AIC is higher for M_2^{SQUAM}. The higher AIC is caused by the large number of parameters in M_2^{SQUAM}, which describe smoking-induced carcinogenesis. Apparently, the state-of-the-art descriptive model can describe this process effectively, with only

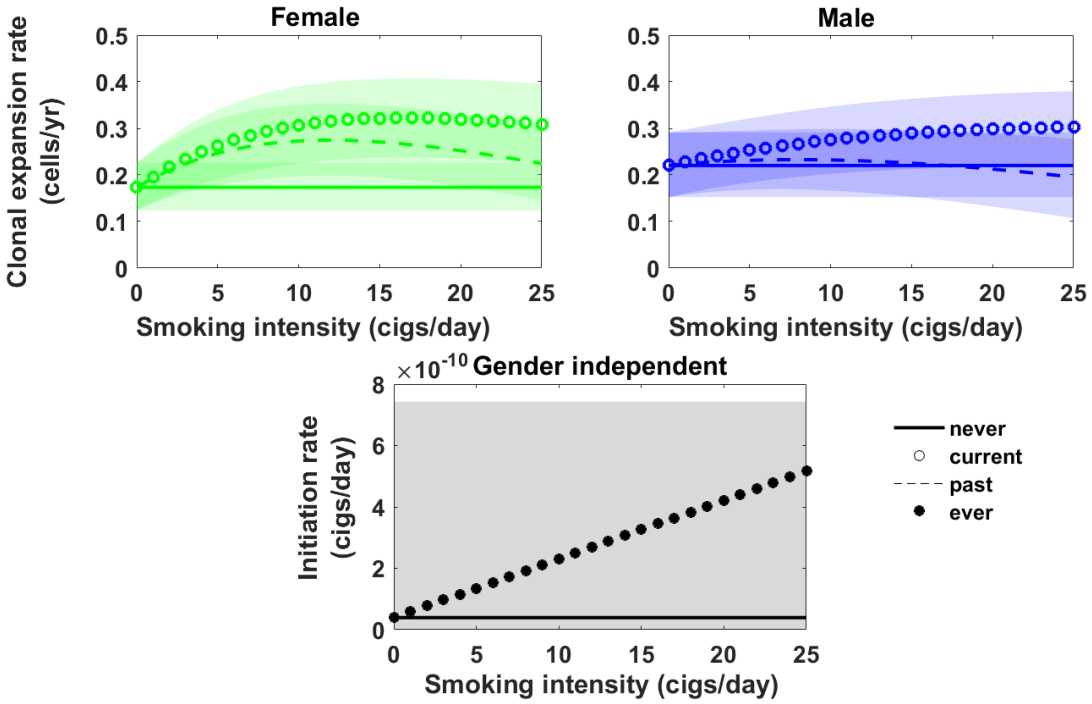


Figure 5.2: Upper panels: clonal expansion rates for female (left) and male (right) smoker person in M_2^{SQUAM}. Smoking intensity smkint linearly enhances the clonal expansion rate $\gamma_S(smkint) = \gamma_{0f,m}\,[1 + \gamma_{Sm,f}\,smkint\,\exp(-\kappa_{m,f}\,smkint) + \gamma_{past} \cdot packyears]$ with an attenuated effect for high smoking intensity and an extra parameter for the quitting smoking period. Pack years are defined as cigarettepacks (one pack contains 20 cigarettes) times the years smoked. Lower panel: sex-independent initiation rate linearly enhanced by smoking intensity $X_{tot} = X_0\,[1 + X_S\,smkint]$. Past smokers quit after 40 years of smoking.

two parameters. Only with M_2^{SQUAM} we can represent the complicated and of multiple facets action of smoking. An effects of smoking on the initiation parameter X can be biologically interpret as a mutational effects. The effect on γ as an inflammatory effect (see Chapter 1.3).

We analysed crude rate and predicted hazard (SQUAM cases in 10,000 persons per year) from M_2^{SQUAM}, plotted in 5 year-age groups from 40-45 up to 80-85 years (see Figure 5.3). In each group observed cases are estimated well by the model. Never and past smokers do not present a flattening for higher ages, that is however markedly present in current smokers.
Next we analysed M_2^{SQUAM} estimates for the breakdown of 319 SQUAM cases (% of 319 cases) in smoking-induces and spontaneous cases, cross-tabulated with exposure groups for smoking (Table 5.4). Refined resolution in exposure subgroups of light (1-10 cigs/day), moderate (11-20 cigs/day) and heavy (20+ cigs/day) smoking intensity is made. Also in this case 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. In contrast to LADC, for SQUAM almost 30% of the total number of cases were estimated as spontaneous cases, the half as found in LADC.

Summarizing, the effect of smoking could be quantified on initiation and promotion rates, with a sex dependence only on the second one. A specific rate for past smokers after quitting could

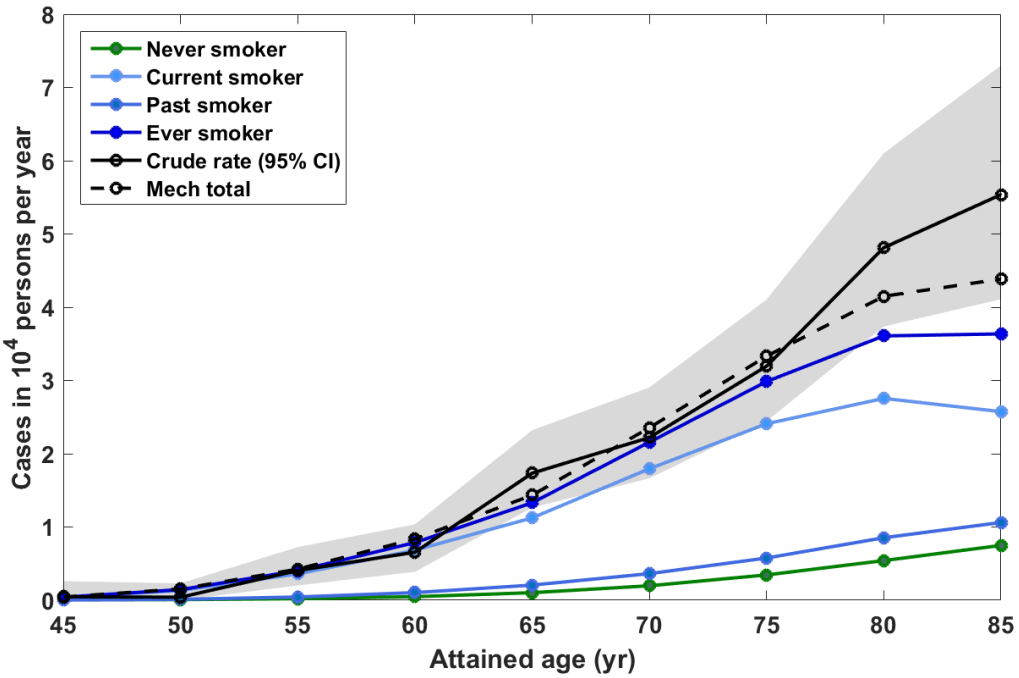


Figure 5.3: Crude rate and predicted hazard (SQUAM cases in 10,000 persons per year) from the M_2^{SQUAM} for the LSS cohort in 5 year-age groups from 40-45 up to 80-85 years.

be fitted as sex independent in the promotion. The most part of cases can be attributed to cigarette smoke exposure. Now we proceed with risk assessment.

Smoking status	Smoking intensity (cigy/day)	Observed cases (%)	Estimated cases (%)	Smoking induced estimated cases (%)	Spontaneous estimated cases (%)
never	=0	**32 (11)**	**37 (13)**	**0 (0)**	**37 (13)**
current	>0	**226 (70)**	**224 (69)**	**193 (60)**	**31 (9)**
	1-10	39 (12)	35 (11)	28 (9)	7 (2)
	10-20	135 (42)	132 (41)	114 (36)	18 (5)
	20+	52 (16)	57 (17)	51 (15)	6 (2)
past	>0	**61 (19)**	**58 (18)**	**40 (12)**	**18 (6)**
	1-10	13 (4)	14 (4)	8 (2)	6 (2)
	10-20	33 (10)	30 (10)	22 (7)	7 (3)
	20+	15 (5)	14 (4)	10 (3)	8 (1)
Total		**319 (100)**	**319 (100)**	**233 (72)**	**86 (28)**

Figure 5.4: M_2^{SQUAM} estimates for the breakdown of 319 SQUAM cases (% of 319 cases) from the LSS cohort cross-tabulated with exposure groups for smoking. Refined resolution in smoking status subgroups of never current and past smokers, and light (1-10 cigs/day), moderate (11-20 cigs/day) and heavy (20+ cigs/day) smoking intensity is made. 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.

5.2.1 Risk assessment

Next we sought to analyse baseline hazards and smoking-related hazards for $Stat_{LSS}^{SQUAM}$ (black lines) and mechanistic models (females in green and males in blue) M_2^{SQUAM} (see Figure 5.5). Age at begin of smoking was taken as 20 yr and age for stop smoking 60 yr. The smoking intensity was of 20 cigs/day.

The baseline hazard function is very low for both females and males in both mechanistic and state-of-the-art statistical models. The male baseline differs also strongly between the two type of models. Looking at the smoking-related hazards for past smokers we can see that compared to current smokers there is a decrease in the hazards, but we can also notice that the values never go down to the baseline hazard. Although in both models there is a sex dependency in the parameters (see Tables 5.2 and 5.3), the curves for female and males are quite similar. Only for current females smokers in state-of-the-art statistical risk model we can notice a decrease for higher ages, which is not present in the mechanistic models. All in all, both models agree with almost no sex-dependency detectable in the baseline hazard curves, although the model parameters suggest a strong sex-dependency.

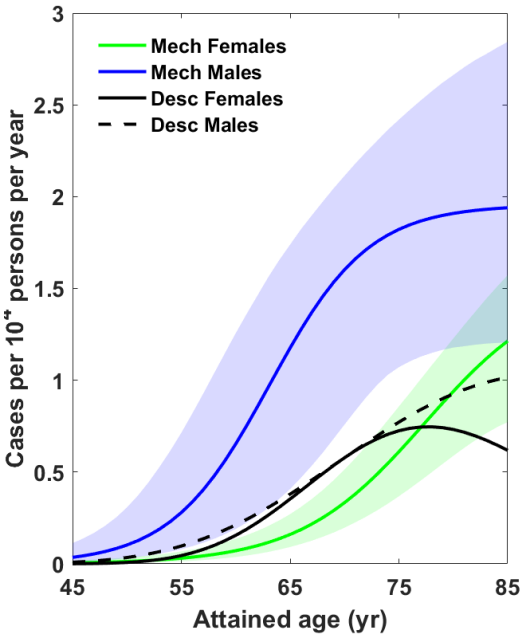
For risk assessment in the main part of this Thesis there will be only EARs presented because of two reasons: the first one is that the biological action presented in M_2^{SQUAM} of the previous chapter is better reflected in the EARs compared to ERRs, the second one is that since the baselines are really small and uncertainty affected, the division by these quantities for

the calculation of ERRs will be problematic. All respective ERRs can be found in Appendix G.4.

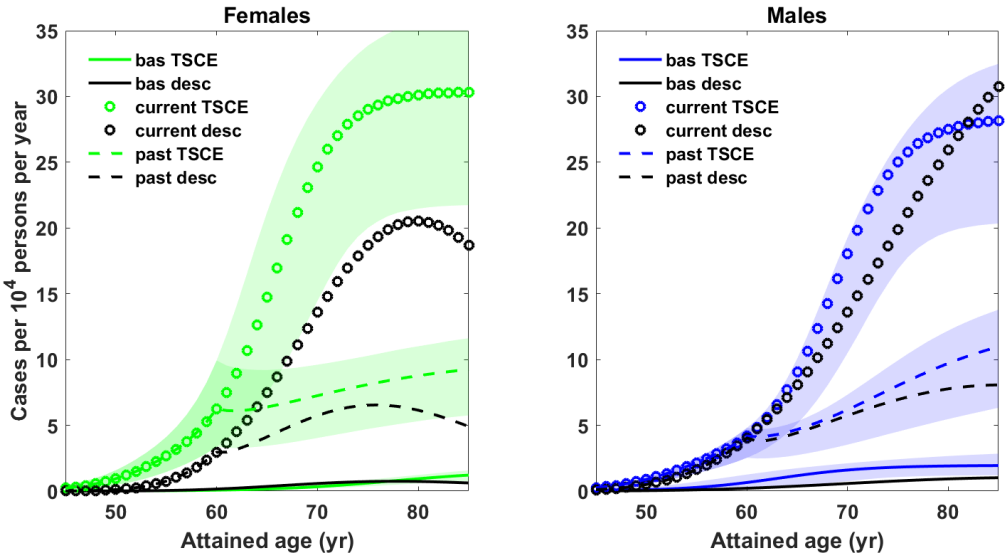
5.2.1.1 Excess absolute rates for smoking

Figures 5.6 and 5.7 present the EARs from M_2^{SQUAM}(Mech, females in green and males in blue) and $Stat_{LSS}^{SQUAM}$ (black lines) for smoking-induced SQUAM for lifelong and past (fe-)male smokers (left and right panels, respectively) starting at age 20 yr and for past smokers at quitting age 60 yr. In Figures 5.6 and 5.7(a) the EARs are presented as a function of age attained for different smoking intensities: 5, 10 and 20 cigs/day. For both lifelong and past smokers the EAR increases with smoking intensity. For past smokers a kink at age of quit smoking is visible, the EAR increases hence again. Past smokers have clearly lower risks as lifelong smokers. In Figures 5.6 and 5.7(b) the EARs are conversely presented as a function of smoking intensity (cigs/day) for different ages: 60, 70 and 80 years. For both lifelong and past smokers the EAR increases with age. Also in this cases past smokers have clearly lower risks as lifelong smokers. A sex difference in EARs is clearly visible only for past smokers, although also for current smokers the model parameters were significantly different and the past smoker parameter was gender-independent (see Table 5.3). The state-of-the-art statistical risk model (black lines) differs relevantly from the mechanistic one only for females, for which the statistical power is lower and hence the variability higher.

The corresponding smoking-related ERRs can be found in Appendix G.4, Figures G.2 and G.3.

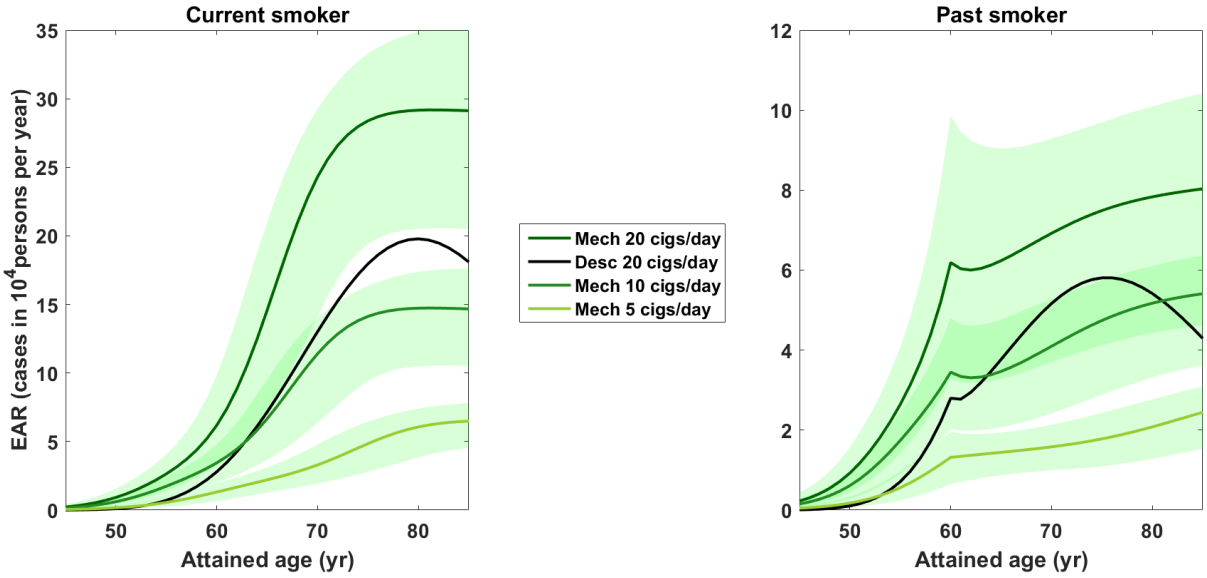


(a) Baseline hazards

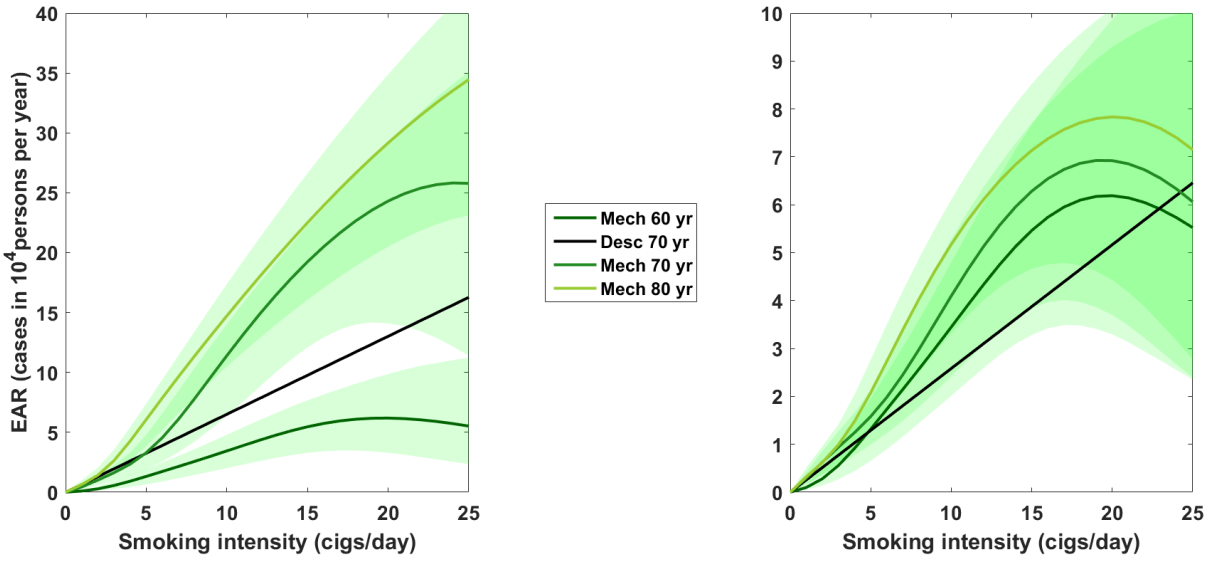


(b) Baseline hazards and smoking-related hazards

Figure 5.5: Baseline hazard (a) and smoking-related hazards (b) for state-of-the-art statistical risk model (Desc, black lines) and mechanistic model (females in green and males in blue) M_2^{SQUAM}. Age at begin of smoking was taken as 20 yr and age for stop smoking 60 yr. The smoking intensity was of 20 cigs/day. Solid lines denote never, dotted lines current and dashed lines past smokers.

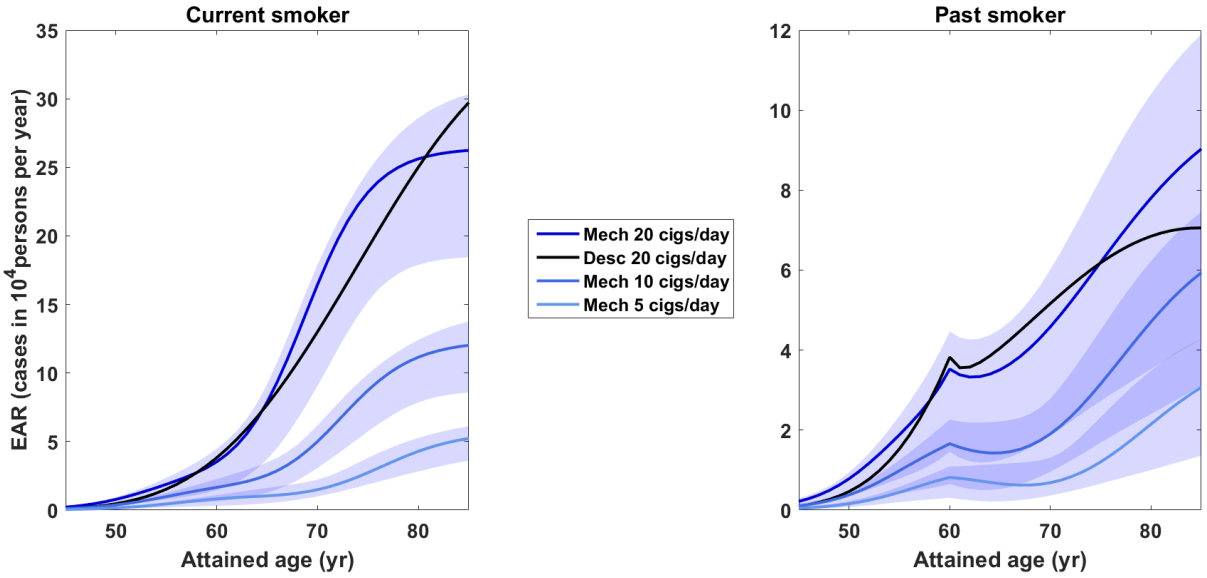


(a)

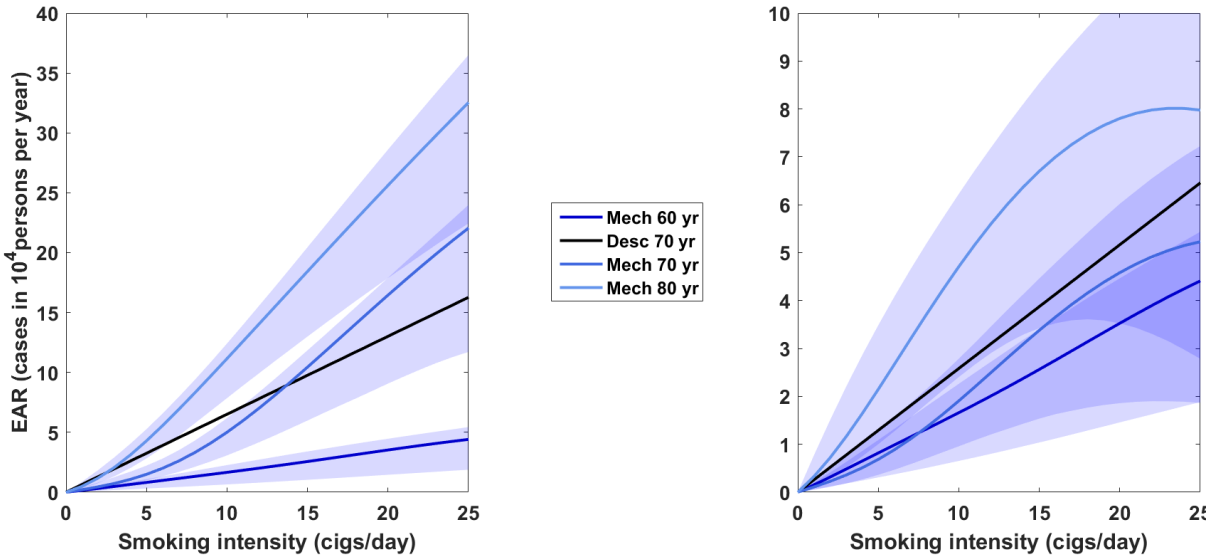


(b)

*Figure 5.6: EARs (as cases in 10,000 persons per year) from M_2^{SQUAM}(Mech) for smoking-induced SQUAM for lifelong and past **female** smokers (left and right panels, respectively) starting at age 20 yr and for past smokers at quitting age 60 yr. (a) EARs as a function of age attained for different smoking intensities: 5, 10 and 20 cigs/day. (b) EARs as a function of smoking intensity (cigs/day) for different ages: 60, 70 and 80 years.*



(a)



(b)

*Figure 5.7: EARs (as cases in 10,000 persons per year) from M_2^{SQUAM}(Mech) for smoking-induced SQUAM for lifelong and past **male** smokers (left and right panels, respectively) starting at age 20 yr and for past smokers at quitting age 60 yr. (a) EARs as a function of age attained for different smoking intensities: 5, 10 and 20 cigs/day. (b) EARs as a function of smoking intensity (cigs/day) for different ages: 60, 70 and 80 years.*

5.3 Generalized additive models

The development of GAMs for SQUAM in the LSS (GAM_{LSS}^{SQUAM}) was based on model $Stat_{LSS}^{SQUAM}$ and the preferred model has the following form

```
SQUAM_gam <- gam(squam~1 + s(I(ageMen/70)) +s(I(ageWomen/70))
                         +s(I(packyrs/50)) +s(smkqyrs),
              offset = log(PYR/10000),
              data = data_lss_squam,
              family = poisson(link = "log"),
              method = "ML",
              optimizer = c("outer","newton"))
```

with the parameters presented in Table 5.4. The deviances of all 50 data sheets can be found in Figure G.1. The cumulative deviance (see section 3.5) from 50 imputed data sets is 359(267)

Table 5.4: Parameter estimates of GAM_{LSS}^{SQUAM}. Central estimates are given as means from 50 imputed data sets with the standard deviation between the 50 best estimates (see section 3.5). edf, the estimated degrees of freedom, describe the complexity degree of the fitted penalised spline.

Parameter estimates of GAM_{LSS}^{SQUAM}	
Meaning	Mean (Standard deviation)
Intercept	-1.05 (0.02)
edf $s(I(ageMen/70))$	3.84 (0.05)
edf $s(I(ageWomen/70))$	3.14 (0.52)
edf $s(I(packyrs/50))$	5.08 (0.15)
edf $s(smkqyrs)$	1 (7.19 $10^{-}10$)
Cumulative deviance	145119
Cumulative AIC	146525

points lower compared to $M_2^{SUQAM}(Stat_{LSS}^{SQUAM})$ (corresponding to ca. 7(5) points per data set). The cumulative AIC from 50 imputed data sets is however 47(529) points higher compared to $M_2^{SUQAM}(Stat_{LSS}^{SQUAM})$ (corresponding to ca. 0.9(11) points per data set), indicating a high complexity of the fitted penalised spline.

All variables gave better fits with the application of spline functions instead of linear terms. Please note that for $s(smkqyrs)$ the fitted *edf* is equal one, it means a linear form. Fitting $s(smkqyrs)$ linearly showed a high correlation of this variable with the function $s(I(age_women/70))$, corrupting the whole fit. The usage of splines for the variable $smkqyrs$ is however not increasing the AIC of the model since for splines we calculate the AIC as $2 \cdot edf$. A graphical description of the polynomials fitted for one selected imputed dataset ($pydat - smk - imp - C23$) are given in Appendix G.5.
To better characterise the properties of model GAM_{LSS}^{SQUAM} we proceed with the analysis of the linear predictors for the imputed dataset $pydat - smk - imp - C23$ as an example.

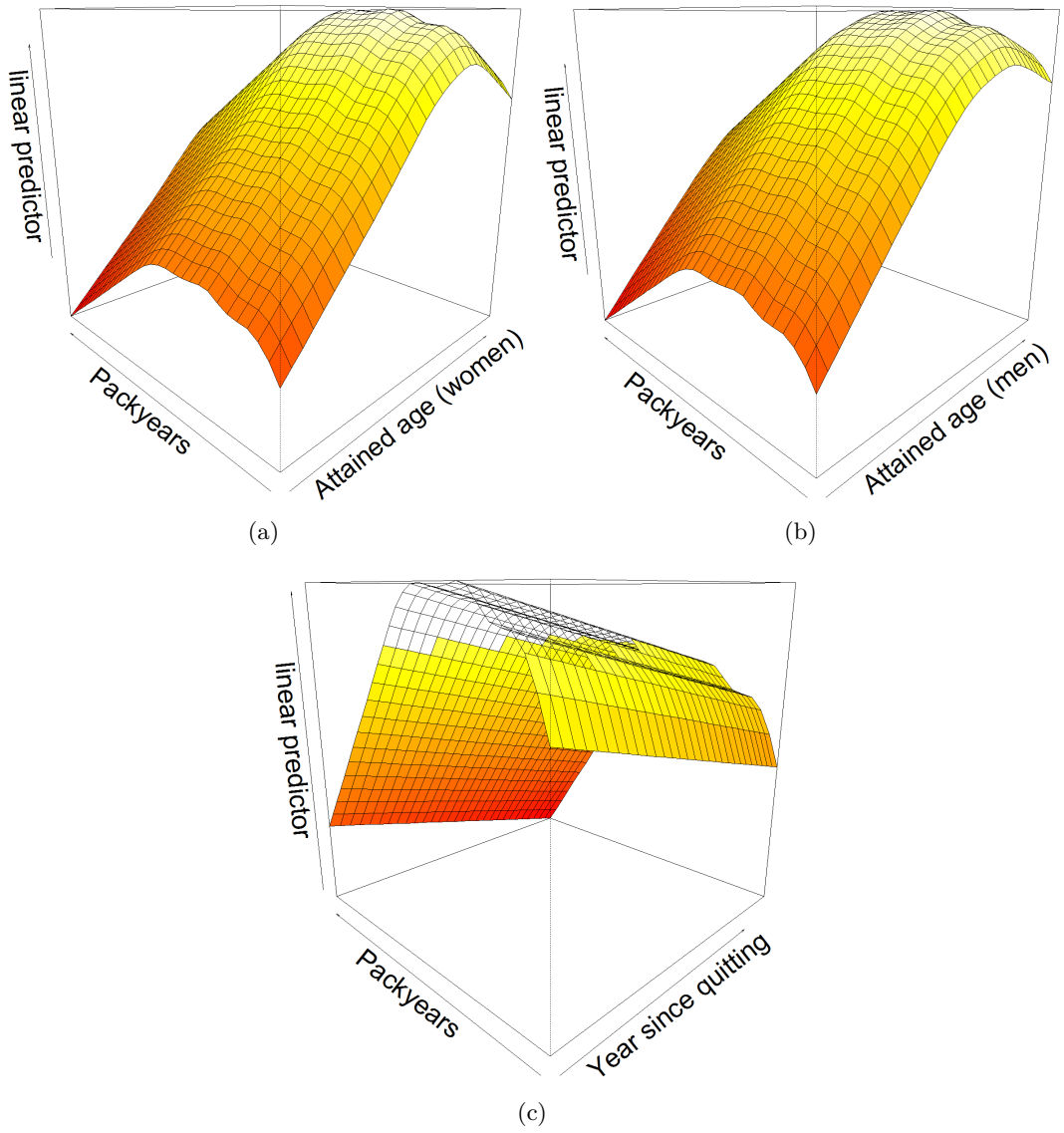


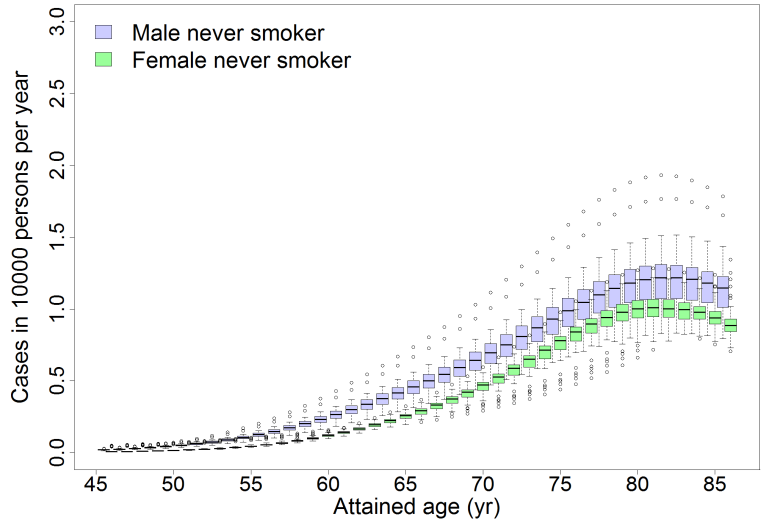
Figure 5.8: Linear predictors for one selected imputed dataset ($pydat - smk - imp - C23$) of model GAM_{LSS}^{SQUAM}. (a) and (b) Linear predictor as a function of cumulative smoking amount (packyears) and of attained age for men and women, respectively. (c) Linear predictor as a function of cumulative smoking amount (pack years) and of years since quitting.

Figures 5.8(a) and (b) present the linear predictor as a function of cumulative smoking amount (packyears) and of attained age for men and women, respectively. Figures 5.8(c), instead, presents the linear predictor as a function of cumulative smoking amount (pack years) and of years since quitting. Although (a) and (b) look very similar, the sex-dependence improved the fit of ca. 200 points. The difference between females and males can be seen in the flattening of the curves for higher ages, where for females the decrease is steeper.

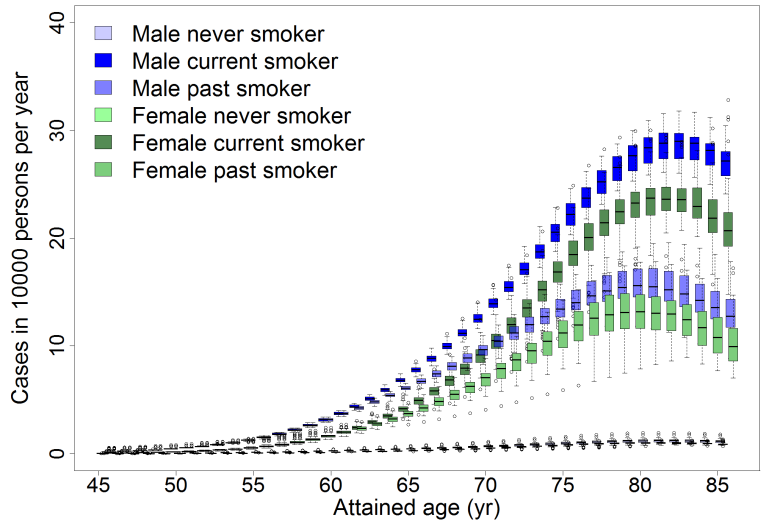
To compare model results of GAM_{LSS}^{SQUAM} with those of M_3^{SUQAM} and $Stat_{LSS}^{SQUAM}$ we made predictions for different scenarios: never smokers (females and males), current and past smoker (females and males), which are presented in Figure 5.9.
The baseline hazards are both clearly smaller than the smoking-related hazards. Past smokers show a reduction after quitting. The results from GAM_{LSS}^{SQUAM} are very similar of those of

model $Stat_{LSS}^{SQUAM}$ (see Figure 5.5).

From baseline hazards and smoking-related hazards we calculated the respective EARs, which are presented in Figures 5.10 and 5.11 for females and males. The ERRs can be found in Appendix G.6, Figures G.5 and G.6
Females and males have very similar risks. Past smokers have a clear attenuation of the EAR, which, however, never disappears. Also for risk estimation the results are very similar to those of model $Stat_{LSS}^{SQUAM}$ (see Figures 5.6 and 5.7 for females and males, respectively). Only with GAMs we can notice a decrease in the EARs for higher ages.

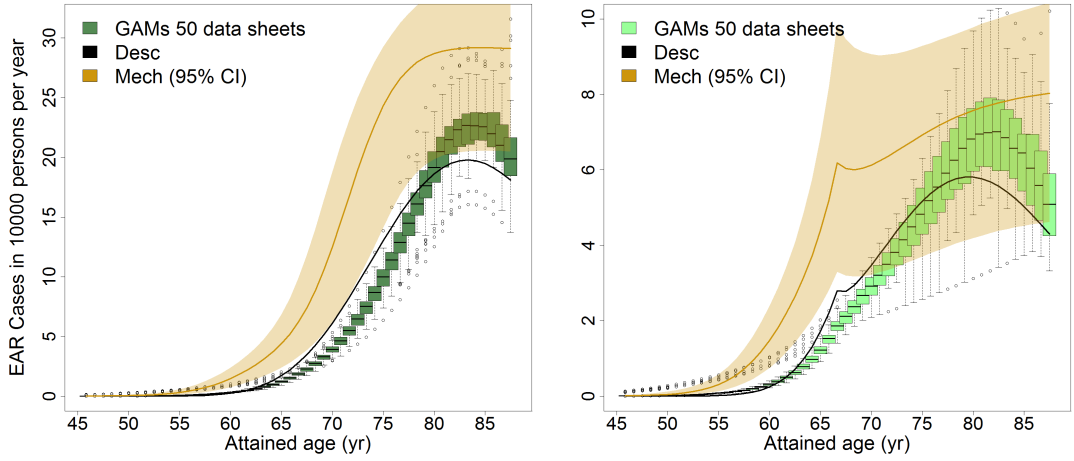


(a)



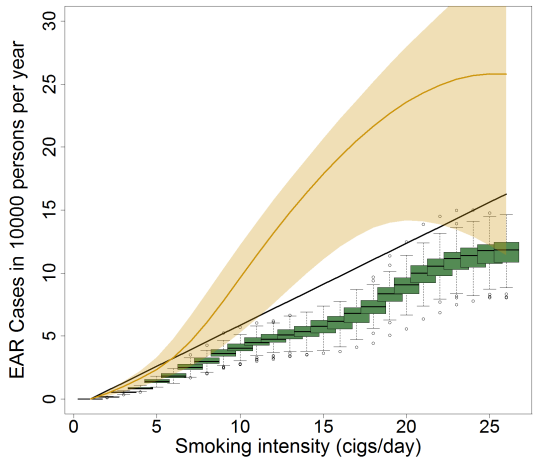
(b)

Figure 5.9: Baseline hazard (a) and smoking-related hazards (b) for GAM_{LSS}^{SQUAM} (females in green and males in blue) from model GAM_{LSS}^{SQUAM}. Age at begin of smoking was taken as 20 yr and age for stop smoking 60 yr. The smoking intensity was of 20 cigs/day. The boxplots represent the variance of the 50 data sets.

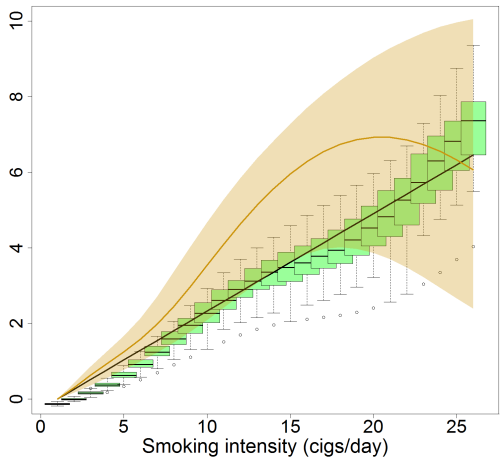


(a) Current female smoker

(b) Past female smoker

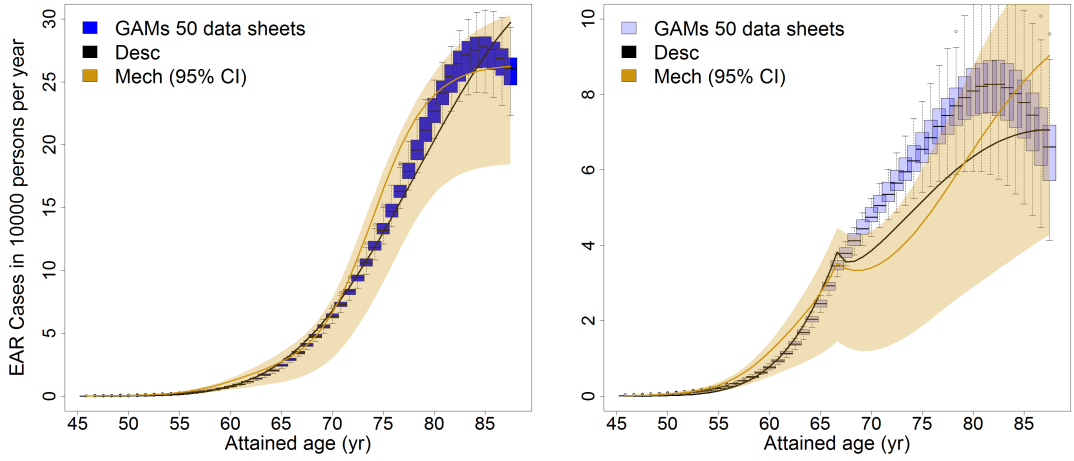


(c) Current female smoker



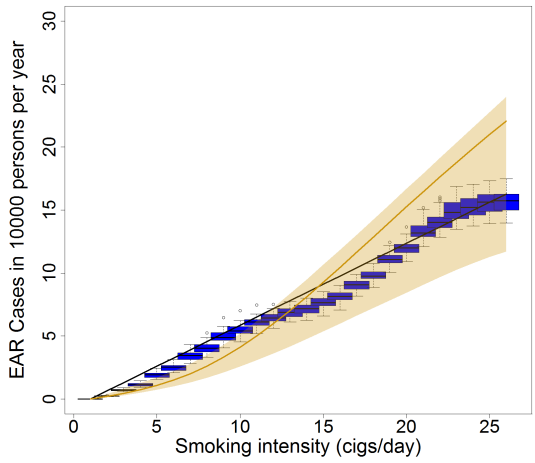
(d) Past female smoker

Figure 5.10: EARs (as cases per 10.000 persons per year) from GAM_{LSS}^{SQUAM} *(in green),* $Stat_{LSS}^{SQUAM}$ *(in black) and* M_2^{SUQAM} *(in orange) for smoking induced SQUAM for (a) current smoking females as a function of attained age, (b) past smoking females as a function of attained age, (c) current smoking females as a function of smoking intensity and (d) past smoking females as a function of smoking intensity. Age at begin of smoking was taken as 20 yr and age for stop smoking 60 yr. In figures as function of attained age the smoking intensity was of 20 cigs/day. In figures as function of smoking intensity the attained age was of 70 years. The boxplots represent the variance of the 50 data sheets.*

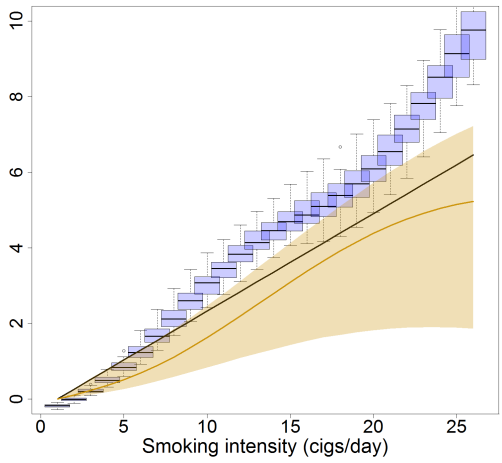


(a) Current male smoker

(b) Past male smoker



(c) Current male smoker



(d) Past male smoker

Figure 5.11: EARs (as cases per 10.000 persons per year) from GAM_{LSS}^{SQUAM} *(in green),* $Stat_{LSS}^{SQUAM}$ *(in black) and* M_2^{SUQAM} *(in orange) for smoking induced SQUAM for (a) current smoking male as a function of attained age, (b) past smoking male as a function of attained age, (c) current smoking male as a function of smoking intensity and (d) past smoking male as a function of smoking intensity. Age at begin of smoking was taken as 20 yr and age for stop smoking 60 yr. In figures as function of attained age the smoking intensity was of 20 cigs/-day. In figures as function of smoking intensity the attained age was of 70 years. The boxplots represent the variance of the 50 data sheets.*

5.4 Summary of results

Result 2: SQUAM in the LSS cohort

In this Chapter we analysed the outcome of SQUAM in the LSS and following results could be achieved

- No radiation sensitivity neither in the state-of-the-art statistical risk model $Stat_{LSS}^{SQUAM}$ nor in model M_2^{SQUAM}
- Detailed description of the damage of cigarette smoke in model M_2^{SQUAM} in initiation and promotion points to deleterious effects of smoking in the mutational spectrum and in inflammation
- Lower risks for past smokers after quitting, but never disappears
- Detailed functional description of effects using GAMs

CHAPTER

6

LUNG ADENOCARCINOMA IN THE ELDORADO COHORT

This chapter is dedicated to the analysis of lung adenocarcinoma in the Eldorado cohort. For this cohort there is information about exposures to α- and γ- radiation, but no information about cigarette smoke exposure is available. Because of this lack of knowledge, mechanistic models are not appropriate, since the major effect causing lung cancer cannot be modelled on a biological basis. Mechanistic models will hence not be developed for this cohort.
This chapter can be divided in two main sections: the first one is about development and findings of state-of-the-art statistical risk models, the second one about generalized additive models. The important finding coming from both models is that only the effects of γ radiation exposure was found significant. For both models an effect of α particles was fitted, but the results did not improve the goodness-of-fit.
Radiation risk estimation from both models will be finally presented and compared.

6.1 State-of-the-art statistical risk models

For the analysis of the Eldorado cohort only state-of-the-art statistical risks models and GAMs will be applied. In Appendix H.1 the model derivation of the preferred state-of-the-art statistical risks model for LADC in the Eldorado cohort ($Stat_{ELDO}^{LADC}$) is explained with model equations and model results (Table H.1). $Stat_{ELDO}^{LADC}$ was selected using the AIC criterion. The form of $Stat_{ELDO}^{LADC}$ is presented in model (6.1), with parameters presented in Table 6.1.

$$h^{LADC_{ELDO}} = h_0^{LADC_{ELDO}} \cdot (1 + ERR^{LADC_{ELDO}}) \tag{6.1}$$

with

$$h_0^{LADC_{ELDO}} = e^{\beta_0 + \beta_1 log(\frac{age}{60}) + \beta_2 log^2(\frac{age}{60}) + \beta_3 \cdot calendaryear} \tag{6.2}$$

$$ERR^{LADC_{ELDO}} = \beta_4 \cdot Gamma. \tag{6.3}$$

Table 6.1: Parameter estimates for model $Stat_{ELDO}^{LADC}$ with 5 parameters. Central estimates are given with 95% CI simulated from multi-variate normal uncertainty distributions conditioned with the parameter correlation matrix (see section 3.5).

Parameter estimates of model $Stat_{ELDO}^{LADC}$		
Name	Meaning	MI Mean (95% CI)
β_0	baseline	1.00 (0.55, 1.45)
β_1	baseline, attained age	3.32 (2.18, 4.45)
β_2	baseline, attained age (squared)	-17.88 (-23.48, -12.24)
β_3	baseline, calendar year	0.79 (0.36, 1.21)
β_4	radiation, ERR (10^{-2})	0.77 (0.27, 1.26)
Deviance	1477	
AIC	1487	

Remembering that the known exposures in the Eldorado cohort were two (α particles and γ rays), it is important to note that for LADC only a significant parameter for the exposure to γ rays could be fitted. An ERR for α particles was not significant when added to an ERR for γ (see Table H.1). Note that a relation between γ exposure and LADC was already suggested from the descriptive data analysis in Figure 2.6. In the baseline the first category of calendar years was takes as reference, all other categories were grouped together.

Since for the γ radiation ERR no effect modifier could be found significant (see Table H.1), we sought to proceed with a sensitivity analysis of the γ ERR. The ERR parameters were calculated using in each fit the preferred model 6.1 $Stat_{ELDO}^{LADC}$ but restricting the data to different maximal levels of exposure: 100 mGy, 250 mGy, 500 mGy, 750 mGy, 1000 mGy, 1500 mGy and complete dataset. The results can be found in Table 6.2.
Increasing the dataset there is a variability in the ERR-value, that remains however with the uncertainty near a value of one.

Table 6.2: Sensitivity analysis of the γ-ERR parameter of the model (H.1). *The ERR parameters were calculated restricting the data to different maximal levels of exposure (100 mGy, 250 mGy, 500 mGy, 750 mGy, 1000 mGy, 1500 mGy and complete dataset).*

Sensitivity analysis of model H.1			
Data (mGy)	γ-ERR (SE) (10^-2)	% cases	p-value
<= 100	1.66 (0.98)	81	0.09
<= 250	1.72 (0.69)	95	0.01
<= 500	1.26 (0.54)	97	0.02
<= 750	1.05 (0.46)	98	0.02
<= 1000	0.89 (0.42)	98	0.03
<= 1500	0.92 (0.40)	100	0.02
complete	0.77 (0.36)	100	0.03

6.2 Generalized additive models

To compare the goodness of fit of model $Stat_{ELDO}^{LADC}$ we also applied a GAM. The preferred GAM for LADC in the Eldorado cohort (GAM_{ELDO}^{LADC}) has the following form

```
gam_eldorado_adeno <- gam(adeno~1 +s(I(age/60)) +calendar_two_adeno
                          +s(I(age/60),gamma5),
                          offset = log(PYR/10000),
                          data = data_eldo_males,
                          family = poisson(link = "log"),
                          method = "ML",
                          optimizer = c("outer","newton"))
```

The corresponding parameters can be found in Table 6.3.

The functional form of GAM_{ELDO}^{LADC} is very similar to that of $Stat_{ELDO}^{LADC}$, but with a moderate

Table 6.3: Parameter estimates and estimated degrees of freedom for GAM_{ELDO}^{LADC}. Central estimates are given with SE, estimated degrees of freedom with reference degrees of freedom.

Parameter estimates of the GAM model	
Meaning	Mean (Standard deviation)
Intercept (10^-1)	0.58 (3.23)
calendar year (cats 2+3+4+5)	0.70 (0.31)
	Estimated degree of freedom (Reference df)
s(age/60)	3.63 (4.56)
s(I(age/60),gamma5)	3.69 (4.94)
Deviance	1462
AIC	1480

improvement of deviance and AIC (15 and 7 points, respectively). With the AIC one would prefer the GAM's description of the data. The improvement can be attributed to two facts: the baseline can be better fitted due to the higher complexity of the penalised spline, the variation of attained age with γ dose could be extrapolated from the data. The fitted functions of the model can be found in Appendix H.2.

To better appreciate GAM_{ELDO}^{LADC} we analysed the linear predictor of the model as a function of attained age (yr) and γ dose (Gy) (Figure 6.1).

The shape of γ dose has multiple peaks (for low and middle exposure) and can hence be better

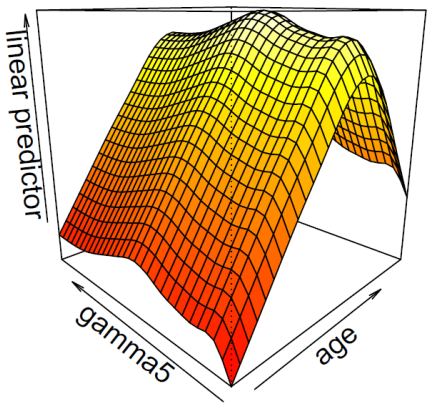
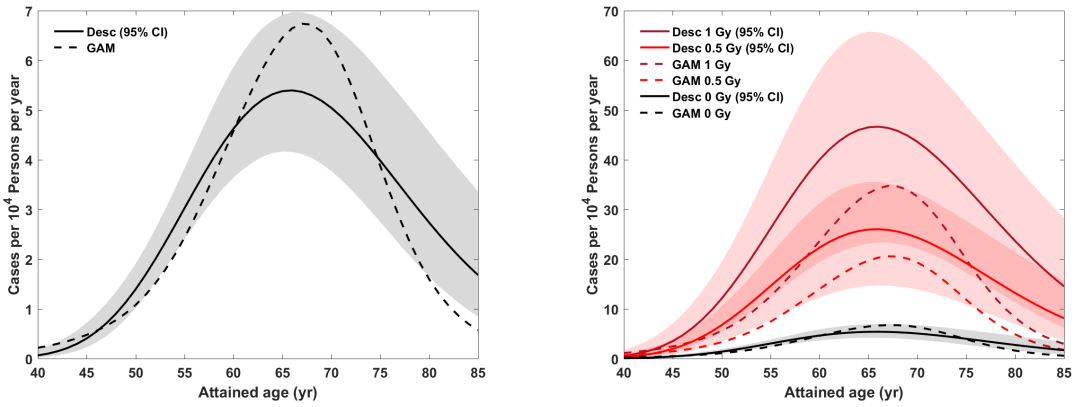


Figure 6.1: Linear predictor of the GAM model for LADC in the Eldorado cohort with respect to the variables attained age (yr) and γ dose (Gy).

represented by (penalized) splines, not used in state-of-the-art statistical risks models. The shape of attained age in more similar to the linear quadratic expression used in state-of-the-art statistical risks models, but can indeed be better fitted by complexer splines.

6.3 Risk assessment

The first analysis in risk assessment are baseline hazards and exposure-related hazards. In Figure 6.2 baseline hazards and γ exposure related hazards (cases per 10^4 persons per year) with different exposure intensities can be found. $Stat_{ELDO}^{LADC}$ (solid lines) and GAM_{ELDO}^{LADC} (dashed lines) are compared.



(a) Baseline hazards

(b) Baseline hazards and hazards

Figure 6.2: Cases per 10^4 persons per year for LADC baseline hazard (a) and baseline hazard with radiation-related hazard (b) as a function of attained age (yr) by different γ exposures comparing $Stat_{ELDO}^{LADC}$ (solid lines) and GAM_{ELDO}^{LADC} (dashed lines).

The baseline hazard of GAM_{ELDO}^{LADC} is slightly higher than that of $Stat_{ELDO}^{LADC}$ for ages 60 yr to 75 yr. In contrast, for radiation related hazard GAM_{ELDO}^{LADC} shows lower risks than $Stat_{ELDO}^{LADC}$. With higher baseline hazard but lower radiation related hazard we expect for GAM_{ELDO}^{LADC} the risks to be lower.

Since all models derived for LADC in the Eldorado cohort are ERR models (cf. equation (3.20)), ERR risk estimates present a better evaluation. In the next section only ERRs will be shown, while the corresponding EARs can be found in Appendix H.3.

6.3.1 Excess relative risks

Looking at the functional forms of $Stat_{ELDO}^{LADC}$ and GAM_{ELDO}^{LADC} one can notice that the γ ERRs are age-independent, but that they change with dose, with the functional form presented in Figure 6.3.

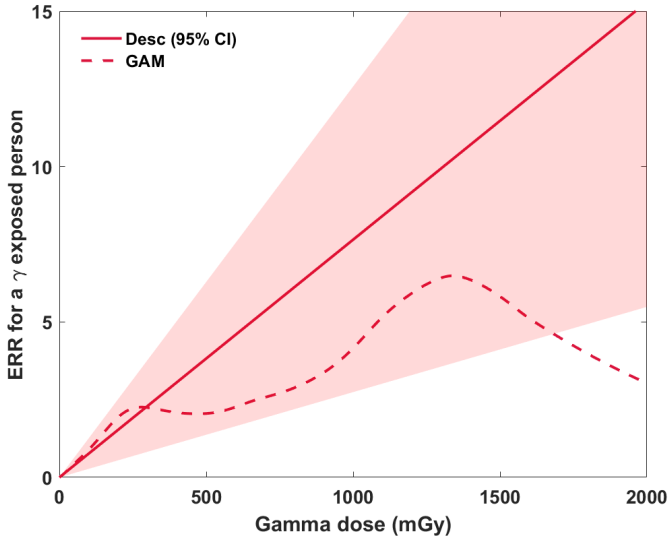


Figure 6.3: Age-independent ERRs for LADC in the Eldorado cohort from $Stat_{ELDO}^{LADC}$ (solid lines) and GAM_{ELDO}^{LADC} (dashed lines) as a function of γ-radiation exposure (Gy).

As expected, the values for GAM_{ELDO}^{LADC} are almost lower for all γ doses and decrease for higher dose. The ERR from $Stat_{ELDO}^{LADC}$ increases linearly with γ dose. Because of its wave-like form, the γ ERR description of GAM_{ELDO}^{LADC} lacks biological plausibility.

6.4 Summary of results

Result 3: LADC in the Eldorado cohort

In this Chapter we analysed the outcome of LADC in the Eldorado cohort and following results could be achieved

- Radiation sensitivity to α was not significant if combined with radiation sensitivity to γ for both $Stat_{ELDO}^{LADC}$ and GAM_{ELDO}^{LADC} models
- Detailed functional description of effects using GAMs

CHAPTER

7

LUNG SQUAMOUS CELL CARCINOMA IN THE ELDORADO COHORT

This chapter is dedicated to the analysis of lung squamous cell carcinoma in the Eldorado cohort. For this cohort there is information about exposures to α and γ radiation, but no information about cigarette smoke exposure is known. Because of this lack of knowledge, molecular mechanistic models are not appropriate, since one of the major effects to lung cancer would be missed. Molecular mechanistic models will hence not be developed for this cohort.
This chapter can be divided in two main sections: the first one is about development and findings of state-of-the-art statistical risk models, the second one about generalized additive models. The important finding coming from both models is that only the effect of α radiation exposure was found significant. For both models an effect of γ radiation was tested, but the results did not improve the goodness of fit.
Radiation risk estimation from both models will be finally presented and compared.

7.1 State-of-the-art statistical risk models

For the analysis of the Eldorado cohort only state-of-the-art statistical risks models and GAMs will be applied. In Appendix I.1 the model derivation of the preferred state-of-the-art statistical risks model for SQUAM in the Eldorado cohort ($Stat_{ELDO}^{SQUAM}$) is explained with model equations and model results (Table I.1). $Stat_{ELDO}^{SQUAM}$ was selected using the AIC criterion. The form of $Stat_{ELDO}^{SQUAM}$ is presented in equation (7.1), with parameters presented in Table 7.1.

$$h^{SQUAM_{ELDO}} = h_0^{SQUAM_{ELDO}} \cdot (1 + ERR^{SQUAM_{ELDO}}) \tag{7.1}$$

with

$$h_0^{SQUAM_{ELDO}} = e^{\beta_0+\beta_1 log(\frac{age}{60})+\beta_2 log^2(\frac{age}{60})+\beta_3 \cdot calendaryear_{group}} \tag{7.2}$$

$$+ e^{\beta_4 \cdot calendaryear_5+\beta_5 \cdot (timesinceexposure)+\beta_6 \cdot (timesinceexposure)^2} \tag{7.3}$$

$$ERR^{SQUAM_{ELDO}} = \beta_7 \cdot Radon \cdot e^{\beta_8 \cdot (timesinceexposure)}.$$

Table 7.1: Parameter estimates for model $Stat_{ELDO}^{SQUAM}$ with 9 parameters. Central estimates are given with 95% CI simulated from multi-variate normal uncertainty distributions conditioned with the parameter correlation matrix (see section 3.5).

Parameter estimates of model $Stat_{ELDO}^{SQUAM}$		
Name	Meaning	MI Mean (95% CI)
β_0	baseline	2.81 (2.30, 3.33)
β_1	baseline, attained age	4.50 (3.59, 5.42)
β_2	baseline, attained age (squared)	-17.07 (-21.10, -13.07)
β_3	baseline, calendar year (cats 2+3+4)	0.62 (0.34, 0.90)
β_4	baseline, calendar year (cat 5) (10^{-2})	4.22 (-35.85, 44.23)
β_5	baseline, time since last exposure (10^{-2})	-8.75 (-11.61, -5.90)
β_6	baseline, time since last exposure squared (10^{-3})	1.63 (1.20, 2.10)
β_7	radiation, ERR (10^{-2})	1.85 (0.47, 3.23)
β_8	radiation, time since last exposure (10^{-1})	-0.51 (-0.78, -0.24)
Deviance	2517	
AIC	2535	

Remembering that the known exposures in the Eldorado cohort were two (α particles and γ rays), it is important to note that for SQUAM only a significant parameter for the exposure to α particles could be fitted. An ERR for γ rays was not significant (see Table I.1). These are opposite results as for LADC in the Eldorado cohort. A relation between α exposure and SQUAM was already suggested from the descriptive data analysis, more specifically in Figure 2.6. In the baseline the first category of calendar years was takes as reference, than categories 2, 3 and 4 were grouped together. Category 5 showed a significant difference from the others.

7.2 Generalized additive models

The preferred GAM for SQUAM in the Eldorado cohort (GAM_{ELDO}^{SQUAM}) was developed based on model $Stat_{ELDO}^{SQUAM}$ and has the following form

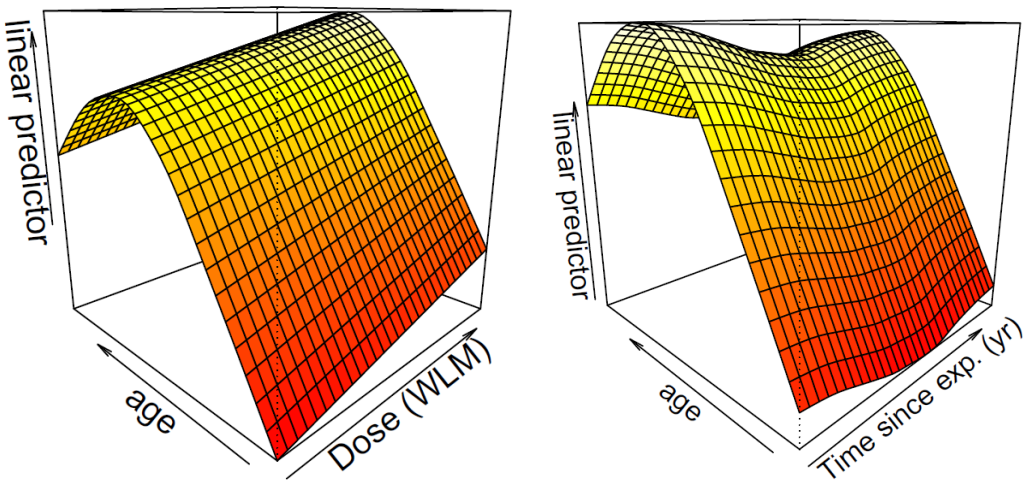
```
gam_eldorado_squam <- gam(squam~1 +s(I(age/60)) +calendar_two_squam
                          +calendar5 +radon5
                          +s(time_since_last_exp,I(age/60)),
                          offset = log(PYR/10000),
                          data = data_eldo_males,
                          family = poisson(link = "log"),
                          method = "ML",
                          optimizer = c("outer","newton"))
```

The corresponding parameters are presented in Table 7.2.

Table 7.2: Parameter estimates and estimated degrees of freedom of model GAM_{ELDO}^{SQUAM}. Central estimates are given with SE, estimated degrees of freedom with reference degrees of freedom.

Parameter estimates of model GAM_{ELDO}^{SQUAM}	
Name	Mean (Standard deviation)
Intercept	0.71 (0.25)
calendar year (cats 2+3+4)	0.59 (0.20)
calendar year (cat 5)	-0.02 (0.29)
radon5 (10^{-2})	0.13 (0.02)
Name	Estimated degree of freedom (Reference df)
s(age/60)	4.08 (5.05)
s(time since last exposure, age/60)	4.42 (5.97)
Deviance	2517
AIC	2542

The deviance of model GAM_{ELDO}^{SQUAM} is equal to the deviance of the respective state-of-the-art statistical risk model $Stat_{ELDO}^{SQUAM}$, while the AIC is 7 points higher. With AIC one would select model $Stat_{ELDO}^{SQUAM}$. The higher AIC of model GAM_{ELDO}^{SQUAM} has to be attributed to the higher complexity of the fitted penalised spline by the model (function presented in Figure I.1) that do not improve the deviance enough to decrease the AIC. To better appreciate the results of model GAM_{ELDO}^{SQUAM} we analysed the linear predictors of the model as a function of attained age and α dose (WLM) (Figure 7.1(a)) and of attained age and time since exposure (Figure 7.1(b)).
The attained age response has a shape similar to a linear quadratic one, normally used for state-of-the-art-statistical risk models. The *Time since exposure* response, instead, presents a wave-like behavior not present in state-of-the-art-statistical risk models.



(a) Linear predictor Radon dose (WLM) - Attained age (yr)

(b) Linear predictor Time since la exposure (yr) - Attaind age (yr)

Figure 7.1: Linear predictor of model GAM_{ELDO}^{SQUAM} with respect to the variables radon dose (WLM) and attained age (yr) (a) and Time since last exposure (yr) and attained age (yr) (b).

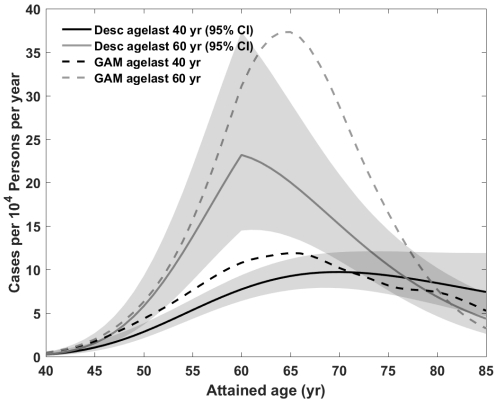
7.3 Risk assessment

To better understand the risks for SQUAM in the Eldorado cohort, we started the analysis of risk assessment with baseline hazards and α related hazards for different exposures (100 and 1000 WLM) and for different ages at exposure (40 and 60 years), which are graphically presented in Figure 7.2.
Baseline hazards and radiation related hazards both decrease for higher ages. For age at last exposure 40 yr both models present very similar results, while for age at last exposure 60 yr model GAM_{ELDO}^{SQUAM} presents higher baseline hazards but smaller radiation related hazards. This implicates hence smaller risks for this model.

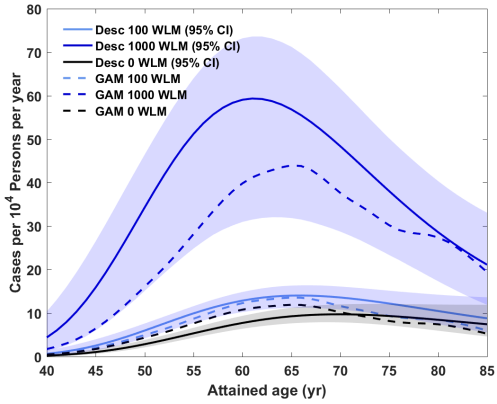
Since all models derived for SQUAM in the Eldorado cohort are ERR models (cf. equation (3.20)), ERR risk estimates present a better evaluation. In the next section only ERRs will be shown, while the corresponding EARs can be found in Appendix I.3.

7.3.1 Excess relative risks

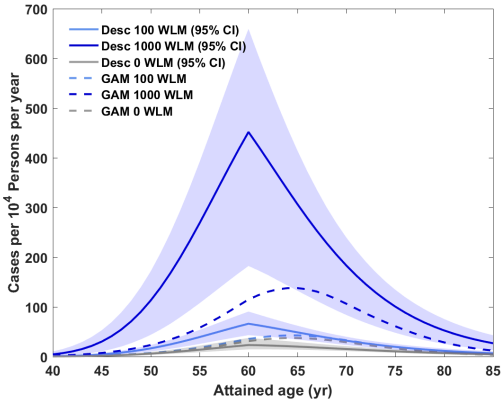
The ERRs from models $Stat_{ELDO}^{SQUAM}$ (solid lines) and GAM_{ELDO}^{SQUAM} (dashed lines) are presented in Figure 7.3.
In Figures 7.3(a) and (b) is it clear to see that, as expected, the risk extrapolation from GAM_{ELDO}^{SQUAM} is much lower than that of model $Stat_{ELDO}^{SQUAM}$. In Figures 7.3(c) and (d) we can notice that the ERR from model GAM_{ELDO}^{SQUAM} is independent on attained age. Since for model $Stat_{ELDO}^{SQUAM}$ the only risk modifier is the variable *age at last exposure*, we have different curves for different attained ages only if the examined attained age is bigger than the fixed age at last exposure. In Figures 7.3(c) and (d) the curves of model GAM_{ELDO}^{SQUAM} can be interpreted as a kind of interpolation of the results from the state-of-the-art statistical risk model.



(a) Baseline hazards

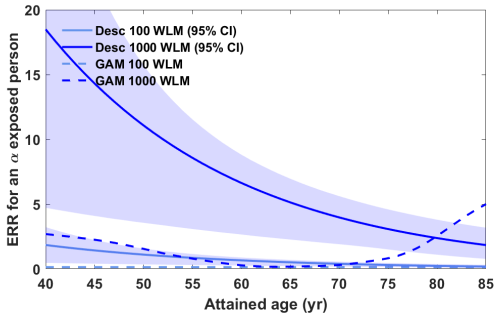


(b) Baseline hazards and hazards for age at last exposure 40 yr

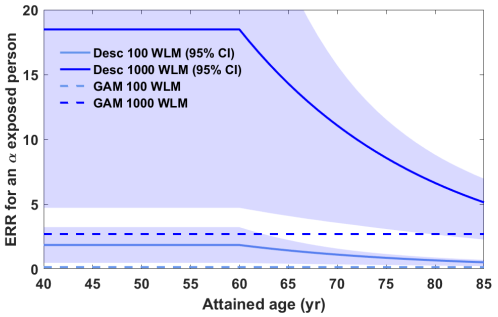


(c) Baseline hazards and hazards for age at last exposure 60 yr

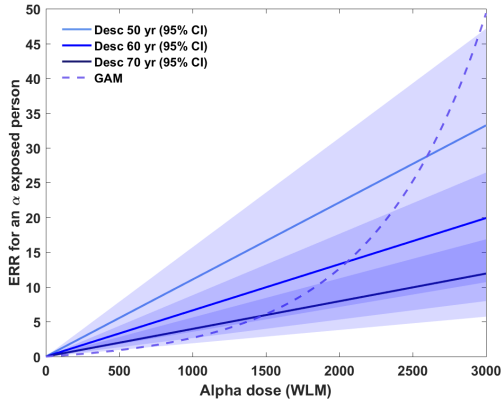
Figure 7.2: Cases per 10^4 *persons per year for SQUAM baseline hazard (a) and baseline hazard with hazard for age at last exposure 40 yr (b) and 60 yr (c) as a function of attained age (yr) by different radon exposures (WLM) comparing model* $Stat_{ELDO}^{SQUAM}$ *(solid lines) and model* GAM_{ELDO}^{SQUAM} *(dashed lines).*



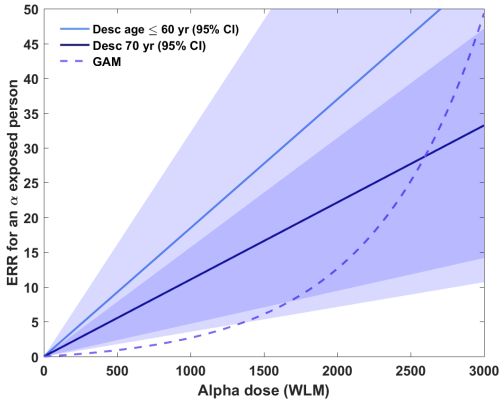
(a) ERR as a function of attained age for age at last exposure 40 yr



(b) ERR as a function of attained age for age at last exposure 60 yr



(c) ERR as a function of radon dose (WLM) for age at last exposure 40 yr



(d) ERR as a function of radon dose (WLM) for age at last exposure 60 yr

Figure 7.3: ERR for radon exposure (WLM) for SQUAM in the Eldorado cohort. The ERRs from models $Stat_{ELDO}^{SQUAM}$ are presented as solid lines wile those from model GAM_{ELDO}^{SQUAM} as dashed ones. (a)-(b) Radiation ERR as a function of attained age for age at last exposure 60 and 60 yr, respectively. (c)-(d) Radiation ERR as a function of α dose (WLM) for age at last exposure 60 and 60 yr, respectively.

7.4 Summary of results

Result 4: SQUAM in the Eldorado cohort

In this Chapter we analysed the outcome of SQUAM in the Eldorado cohort and following results could be achieved

- Radiation sensitivity only to α radiation exposure for both $Stat_{ELDO}^{SQUAM}$ and GAM_{ELDO}^{SQUAM} models
- Detailed functional description of effects using GAMs

CHAPTER

8

SUMMARY AND DISCUSSION

8.1 Data selection and model application

The aim of this Thesis is to analyse carcinogenesis and the related risks in the lung under the effects of smoking and two types of ionising radiation: γ waves and α particles. To explore this question two radio-epidemiological cohorts were examined by modelling incidence rates: the LSS and the Eldorado cohort [8, 18, 21, 33].

For the LSS Furukawa et al. [21] first and Cahoon et al. [8] next investigated the effects of radiation and smoking on lung cancer incidence. Both studies detected a strong interaction between radiation and cigarette smoking exposures. In [8] the models developed by [21] were just reapplied to the LSS with extended followup and obtained congruent results. Egawa et al. [18] extended the models developed by Furukawa et al. [21] to different subtypes of lung cancer: LADC, SQUAM and small cell carcinoma. Only the first two are investigated in this Thesis. For these subtypes the interaction of radiation and smoking was tested. Furukawa et al. [22] hence investigated the question how to cope with cohort strata with unknown smoking information. Three different approaches were compared:

1. create a separate category and fit an extra ERR for unknown smoking status,
2. ignore/delete strata with unknown smoking category, and
3. appeal imputation procedures to replace the missing information.

According to the analysis of [22] the best results can be obtained using imputed data, the second best deleting the strata with unknown smoking information and as last one considering the extra category. We confirmed this result with the descriptive data analysis of the LSS (see Chapter 2.2), although imputation has to be used carefully. In the analyses of [8, 18, 21] the considered procedure was the first one, the extra category. Based on the experience of Furukawa et al. [22] we decided to use only imputed data sets.

For the Eldorado cohort there are already risk studies for all histological subtypes of lung

cancer combined [33], but not for the different subtypes and for both γ and α exposures. For this cohort, no information about cigarette smoking exposure is available. Smoking is the leading causes to lung cancer, irrespective of subtype (see Chapter 1.1). Zaballa and Eidemueller [67] already applied mechanistic risk models to the Wismut cohort, another cohort of Uranium miner with protracted exposure to α radiation but without information about cigarette smoking. From this analysis it is clear that mechanistic models can be applied, but the interpretation of the effects is complicated, since information on a major risk factor is missing. The effects fitted for other covariables could be strongly affected by missing information. Because of this fact, no mechanistic risk modeling has been applied in this Thesis to the Eldorado cohort, since any biological interpretation would be speculative without cigarette smoking information.

8.2 Biology and modeling: lung adenocarcinoma in the Life Span Study cohort

Lung adenocarcinoma management and outcomes largely rely on tumor genotype [49]. However, current prediction 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: one unique to patients with mutations in transmembrane receptors (R^{MUT}) and another shared by patients with mutations in signal transduction genes and by patients with no known oncogene mutations (T^{MUT}). These molecular findings were combined with observational data of LADC incidence in Japanese atomic bomb survivors with known radiation/smoking exposure but unknown mutation status, to develop the first molecular mechanistic model (M_3^{LADC}) for LADC risk prediction stratified by two modelled molecular pathways. M_3^{LADC}provides risk prediction including molecular subtypes that can be tested for the LSS in the future, and explained for the first time the different mutation frequencies in Western and Eastern populations based on smoke and radiation exposures.

Just like standard epidemiological risk models, M_3^{LADC} accurately reproduced LADC incidence in the LSS, albeit with moderately improved goodness-of-fit. Lubin and Caporaso [38] 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 (Figure E.3). As a striking new feature, M_3^{LADC} clearly identified the two molecular pathways that emerged from the molecular analysis. Model predictions on the expected number of cases in both pathways were compatible with molecular measurements of KRAS and EGFR mutations in Japanese patients [57]. 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.

Previous molecular studies underpinned the biological plausibility of M_3^{LADC}. KRAS mutations are more common in smokers [10] and are suspected to confer resistance to radiotherapy [62], 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 this date, the risk factor that drives LADC development in never smokers is unknown, while these patients exhibit higher frequencies of EGFR mutations and EML4/ALK fusions [2, 56]. Here we show that radiation may drive disease development in these patients and provide a risk prediction model for this molecular class of LADC. This observation corresponds 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 TCGA analyses showed enhanced mutation rates in ever smokers of the T^{MUT} pathway [10]. In a mechanistic model, this observation should translate into an increase of the initiation rate in smokers. However, M_3^{LADC} works without such a plausible smoking effect because improvement in goodness-of-fit was inferior compared to a smoking action on clonal expansion. Hence, the model points to te main biological mechanism of smoking on LADC incidence to be associated with enhanced clonal growth. 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.

M_3^{LADC} also explains the higher susceptibility of women to smoke, evident by the current LADC pandemic in women [25]. A study of EGFR and KRAS mutations in 3000 LADC of Caucasian patients revealed a higher susceptibility of women to smoking exposure for KRAS-mutant cancers [16]. These findings are in line with a stronger increase of the smoking risk in the T^{MUT} pathway for female light smokers compared to male light smokers. Our results are concordant to the aforementioned study and can likely 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, as those observed in inbred strains of mice [54].

Risk prediction models, which are informed by adequate bioassays in addition to epidemiological variables, can predict lung cancer risk with high accuracy [19]. They do lack, however, a link between environmental agents and molecular risk stratification, which is provided by M_3^{LADC}. For example, this link suggests no elevated LADC risk even for heavy smokers in CT screening. It can be exploited in retrospective assessment to pin down the agent causing LADC based on the molecular profile of diseased tissue.

In conclusion, our invertigation of LADC in the LSS answers a longstanding question on the biological origins of age-risk patterns for LADC from concomitant exposure to smoking and radiation. Standard epidemiological models must inevitably rely on a vague description of synergistic effects, which are commonly couched in mathematical terms as either 'additive' or 'multiplicative' sometimes with further generalizations [8, 18, 21]. For risk assessment studies an effect is preferred based on statistical criteria of goodness-of-fit with scant biological justification. We have shown here that smoking and radiation drive the development of LADC along different molecular pathways with negligible interaction for doses below 4 Gy. Standard epidemiological models do not mirror these findings which have an impact on risk predictions especially in regions of the cohort dataspace with low case numbers. To conclude, 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.

8.3 No interaction between radiation and smoking exposures for lung cancer subtypes in the Life Span study cohort

The major result from Furukawa et al. [21], the first analysis of lung cancer in the LSS, is that radiation and smoking exposures act to this endpoint with clear significant interaction. Egawa et al. [18] extended the models developed by Furukawa et al. [21] to different histological subtypes of lung cancer: LADC, SQUAM and small cell carcinoma. Only the first two are of our interest. For these subtypes the interaction of radiation and smoking was tested.
Based on AIC for lung cancer as endpoint the generalized multiplicative model (the multiplicative model with interaction between smoking and radiation exposures) appeared to be the best model. But this was not the cases for LADC or for SQUAM alone [18, 21]. For LADC the preferred model was a simple additive model (radiation and smoking add their influence without any interaction), where for SQUAM the preferred model was a simply multiplicative (Table 4 in [18]). In [18] only the generalized multiplicative model was discussed, in order to compare the new findings with those of [21] for all lung cancer subtypes combined. This misleading representation is deceiving the community, which is convinced of the interaction of smoking and radiation exposures also for subtypes. However, from the model results one cannot exclude any interaction, which at the same time cannot be confirmed. The discrepancy between the results of Furukawa et al. [21] and of Egawa et al. [18] could be due to two factors: decrease in statistical power and superposition of effects. By splitting the total number of lung cancer in histological subtypes, the statistical power for each subtype is decreased. Based on molecular profiles we already know that the lung cancer subtypes can be considered as different diseases (see Chapter 1.1). For all lung cancer types combined, the model results can be considered as a superposition of effects pertaining to each subtype alone. Because radiation and smoking affect the subtypes with differential sensitivity, the interaction observed in [21] may possibly be considered as accidental.

An other important point for the analysis of [21] and [18] is that the strata in the data with unknown smoking information were classified separately and specific ERRs for smoking were fitted with a sex and age-dependent linear exponential shape. To test the radiation related ERR, however, no differentiation between strta with known and unknown smoking information was done. If the specific form used for the "unknown" smoking ERR is hence not good enough, the not fitted rest effect can be attributed to radiation, influencing also the interaction terms. We tested the influence of this oversight introducing extra ERR radiation related terms for the unknown smoking cells. The magnitude of the interaction still persisted but with a dramatic decrease.

Summarizing, with the recognition of the following facts

- possible superposition of effects in the analysis with endpoint lung ([21])
- actual absence of interaction already in [18]
- inadequate inclusion of the "unknown" smoking category for fitting of the radiation related ERR in both [21] and [18]

we are not surprised that also in our analysis no interaction between radiation and smoking could be found, in any lung cancer subtype.

In LADC the preferred molecular mechanistic models revealed the presence of two biological pathways in observational data: the R^{MUT} pathway comprised almost all spontaneous and

radiation induced cases, the T^{MUT} pathway contained very few spontaneous cases and is driven almost exclusively by smoking. We cannot exclude an effect of smoking in R^{MUT}, but we could not detect it based on biological and statistical analysis. From the biological analysis KRAS-mutant LADC and TCGA analysis showed enhanced mutation rates in ever smokers of the T^{MUT} pathway [10].
For SQUAM in the LSS no significant radiation response could be detected with any model. In Egawa et al. [18] (their Table 4) for SQUAM the gender averaged radiation-related ERR estimate is relatively small with border line significance. We also notice an unexpected high value for the *age at exposure variable* and a markedly negative value for *attained age*. The radiation risk is maximal after exposure at young age (a calendar year effect). We explored this hypothesis analysing the non imputed complete data set with and without strata of the first calendar year category. After this deletion the radiation-related ERR value was not significant anymore, for all the three types of models applied to this cohort (state-of-the-art statistical risk model of [21], molecular mechanistic models and GAMs). Together with K. Furukawa we concluded that it is implausible that the total radiation response of a cohort can be supported by only one category. We therefore proceed the analysis of SQUAM excluding strata with calendar year category 1. Radiation-related risks for SQUAM in the LSS were hence possible but not significant. Interestingly, for SQUAM the mechanistic models could fit the effects of cigarette smoking in initiation and promotion, differently as for LADC, indicating one more time the importance of the distinction between lung cancer subtypes.

8.4 Different radiation qualities act separately on different lung cancer subtypes

Although no information about cigarette smoke exposure is available, the Eldorado cohort is a peculiar data set for the analysis of radiation-related effects since those people were unfortunately exposed to both γ and α ionising radiation. In this Thesis, for the first time in radiation protection, different lung cancer subtypes were examined and compared after protracted exposure to different radiation qualities. Lane et al. [33] already analysed the Eldorado cohort, but neglecting the very important differentiation between subtypes. Kreuzer et al. [31], instead, payed attention to the different subtypes in a analysis of the Wismut cohort, a cohort of German Uranium miners, but not having an exposure to γ radiation and making only a case comparison. We developed standard state-of-the-art statistical risk models for radiation protection and statistical generalised additive models to explore the effects of both γ and α ionising radiation to LADC and SQUAM. Statistical generalized additive models have already succeed predicting lung cancer case, with a markedly increment of the goodness of fit [12]. Because of their easy way to cope with interactions, they are are precious instrument for our analysis.
For LADC both state-of-the-art statistical risk model and generalised additive models could find a significant radiation effects only for γ radiation exposure. The effects of exposure to α particles could be fitted but were not significant. For SQUAM, *vice versa*, both models could find only a significant response to α radiation exposure, while responses of exposure to γ particles again were not significant. These findings are in line with the results of [31] and with our results for the LSS, where for LADC radiation and smoking were acting on different molecular pathways and for SQUAM no significant response to γ radiation could be found. Moreover, these results may be explained biologically. It is known know that both α particles and cigarette smoke affect mostly the epithelium of the lung causing mostly SQUAM (see Chapter 1.3) [37]. Vahakangas et al. [61] and Choi et al. [11] analysed genetic radiation markers for radon. The following genes were significantly mutated: EGFR, TP53, NKX2.1 and PTEN. Since [11] analysed all lung cancer without subtypes, the presence of EGFR can be understood as stemming

from LADC lung cases. Moreover, genes TP53 and PTEN are also the genes that were found frequently mutated in [44] (see Chapter 1.1.2). Since almost all cases of both Choi et al. [11] and [44] were smokers, we can conclude that the radiation markers for α exposure do not differ from those of smoking. With the fact that α particles and cigarette smoke act additively/synergistically to the development of lung cancer [5, 34], it is not surprising that main effects of both risk factors relate to the same subtype, SQUAM. γ waves, instead, penetrate the tissue completely, damaging mostly all parts of the lung [37].
In syntheses, α radiation and smoking are the dominant risk factors for SQUAM, γ radiation is not visible but can still be important.
A summary of the ERRs calculated per method and lung cancer subtype is presented in Figure 8.1. For SQUAM of both LSS and Eldorado cohorts an ERR for α exposure was not

Table 8.1: Summary of ERR for γ radiation at 1 Gy at age 70 yrs. n.a. means not available. Values marked by a $\star$ are estimates from only one imputed data sheet.

ERR for γ radiation at 1 Gy at age 70 yrs				
Model	**LSS** age at exposure 30 yr		**Eldorado**	
	LADC	**SQUAM**	**LADC**	**SQUAM**
Mech	0.48 (0.14 1.38)	n.a.	n.a.	n.a.
Desc	1.23 (n.a.)	0.23 (0.91 SE)$\star$	7.7 (2.7 12.6)	-90.3610^{-5} (-21.3710^{-5} SE)$\star$
GAM	0.83 (n.a.)	n.a.	4.12 (n.a.)	n.a.

significant. The ERRs of the three models for γ exposure are concordant in each cohort, but differ in one order of magnitude between the cohorts. This difference may be due to the different type of exposure: short and large for the LSS, protracted for the Eldorado.

8.5 Lung cancer subtypes are different diseases

Finally, we want to stress once a gain a main result that was relevant throughout this Thesis: lung cancer cannot be analysed as a single endpoint, but the different histological subtypes have to be considered separately. Of course with this split the statistical power for each endpoint decreases. The results for the combined endpoint are, however, hard to interpret because of a likely superposition of effects. As we have seen, different radiation qualities act differently on the different subtypes, revealing different biological processes. Smoking was relevant to all analysed subtypes, but had different effects with different magnitudes. A deeper understanding of carcinogenesis in the lung with respect to radiation and smoking effects, can only be achieved with the consideration of lung cancer subtypes as different disease entities.

8.6 Limitations of this analysis

Limitations of this analysis relate to both the data and to the models applied.

In epidemiological data there is always some uncertainty in data collection. The match with cancer registries is involved and personal information in some instances may be connected wrongly. For some patients there is no clear cause of death.
As we have seen in the description of the cohorts, dosimetry is not an easy issue. New methods to reconstruct the internal dose for the LSS are still being developed. For the dosimetry in the Eldorado cohort, the protraction of the exposure plays a role. In the mine there are daily fluctuations in the concentration of radioactive substances, and therefore in the received dose.

No such detailed information is available. In the Eldorado cohort, some workers spent part of their employment duration outside the mine but this information was not available. Only the cumulative exposure during life was obtained. However, time-dependent information would be important for mechanistic analysis.
The importance of smoking as a risk factor for LADC and SQUAM was a leitmotiv in this Thesis. Uncertainty in this information, which was derived in the LSS primarily from surveys, is hence a major issue. In the past, smoking was restricted for women in Japan, who however smoked secretly and inhaled passive smoke. These facts definitely affect the results for female smokers. A comparison of smoking effects in different cohorts was not possible in this Thesis since no information was provided for the Eldorado cohort. This is still a point of major interest that should be investigated.

Each of model types applied has merits and defects. State-of-the-art statistical risk models for radiation protection generally can reproduce well the data after some model adaptations. Functional forms of the covariables' effects have to be assumed. The results do not have a biological meaning. Mechanistic models can be interpreted biologically but are simplistic descriptions of complex biological processes. Model derivation can be complicated and time consuming. Statistical generalized additive models are easy to derive but some control is necessary to avoid implausible results for risks estimates (wiggly functions). Therefore, the different models yield complementary information on the different cancer types.

As we have seen in the biological analysis of LADC in Chapter 4, results from a Japanese cohort may not be transferred easily to other ethnic groups. Comparing quantities can be misleading, though a qualitative analysis should remain consistent. The type of radiation the victims of the LSS were exposed to is also unique: a relatively uniform, high dose for a very short time. A transfer of risk estimates from this cohort to, e.g., medically exposed patients should be done with caution.

8.7 Outlook

From the findings of this Thesis several interesting questions arose.

For future data collection it is essential that attention is given to the different subtypes of lung cancer. If possible, also genetic information of the cancer tissue should be taken. It is important to understand that future epidemiological data analysis for lung cancer makes sense only when differentiating between subtypes. A global analysis on lung cancer would just be a superposition of the effects of the single subtypes, impeding the interpretation of the effects and the anyway uncertain transfer to other ethnic groups. To consolidate the findings about lung adenocarcinoma in the Life Span Study cohort, it would be helpful, if a portion of the tissue samples collected from the lung cancer patients would be genetically analysed. In this way, the radio-sensitivity of the R^{MUT} pathway could be tested directly.

For analysis of lung cancer, information on smoking is central for a correct quantification of other effects. Already a dichotomous information improves the analysis.

In this Thesis we exemplified how epidemiology and biology complement one another, to understand and prevent diseases. We warmly recommend a better cooperation between the different radiation fields: biology, epidemiology and treatment radiology. Moreover, because of the lack

of cooperation between fields, a lot of common statistical routines were never applied to radiation epidemiological data. Some steps in this direction were undertaken. In this Thesis for the first time generalized additive models were applied in radiation risk analysis although those models are a standard tool in pertinent statistical analysis. Present trends in statistics focus a lot on machine learning. This could help to reduce the potential impact of the researcher, which is appealing in a field under a strong societal controversy.

Risk assessment for radiation protection mostly relies on the results of state-of-the-art statistical risk models. For a more comprehensive investigation other modeling approaches, such as biologically-based models or more general classes of statistical models (i.e. generalized additive models) provide valuable insight.

Concluding, we want to underline that a very large part of lung cancers could be avoided if smoking was reduced in the population. Therefore, continuous efforts are necessary to prevent smoking initiation, especially in young people, and to promote cessation of smoking. Radiation showed to be a cause of adenocarcinoma, but secondary after smoking even in the victims of the atomic bombs. Although the much higher risk of smoking is well known to the scientific community, public perception disagrees completely. People should be better educated in respecting the natural phenomenon of radiation: it can be dangerous, indeed, it also saves lives daily, e.g. in cancer radiotherapy or medical diagnostic.

Appendix

APPENDIX

A

DESCRIPTIVE DATA ANALYSIS OF THE LIFE SPAN STUDY COHORT: SUPPLEMENTARY INFORMATION

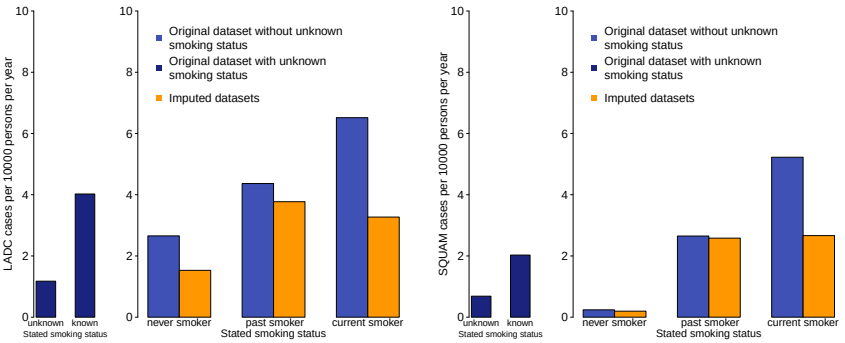
This Appendix Chapter contains supplementary information about the LSS cohort. Related to Chapter 2.2.

A.1 Comparison original vs. imputed data sets: raw data

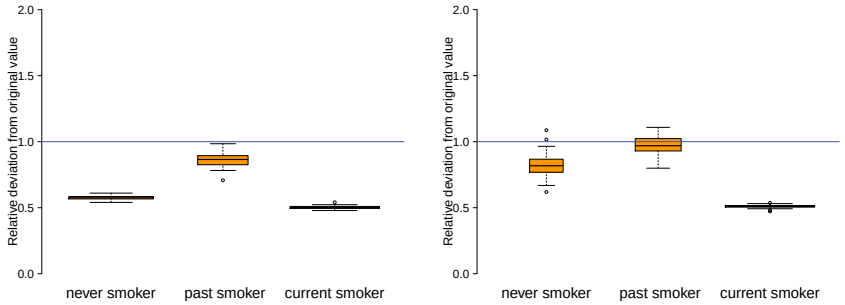
Table A.1: Comparison of cases after imputation (Imp) with the cases of the original dataset (OD) containing the category "unknown smoking information". Second, third and fourth columns summaries imputed data with endpoint lung for the lung cancer types lung in general, LADC and SQUAM, respectively. The fourth column is a summary of the imputed data with endpoint SQUAM for the subtype SQUAM.

	Lung total	**LADC**	**SQUAM**	**SQUAM**
Cases in OD	1803	636	330	330
Never smokers in OD	298	167	16	16
Ever smokers in OD	804	260	193	193
Cases without smoking info in OD	626	209	121	121
% cases without smoking info in OD	38	33	37	37
Imputed never smokers	257	98	30	**17**
Imputed ever smokers	480	126	97	**110**
Never smokers after Imp	555	265	46	**33**
Ever smokers after Imp	1284	386	290	**303**
% imputed never smokers	44	35	65	**50**
% imputed ever smokers	35	31	32	**35**
$\frac{\textit{Imputed never smokers}}{\textit{Never smokers OD}}$	0.86	0.59	1.88	**1.06**
$\frac{\textit{Imputed ever smokers}}{\textit{Ever smokers OD}}$	0.60	0.48	0.50	**0.57**
$\frac{\textit{Never smokers OD}}{\textit{Ever smokers OD}}$	0.37	0.64	0.08	0.08
$\frac{\textit{Never smokers after Imp}}{\textit{Ever smokers after Imp}}$	0.43	0.69	0.16	**0.11**

A.2 Smoking status of cases



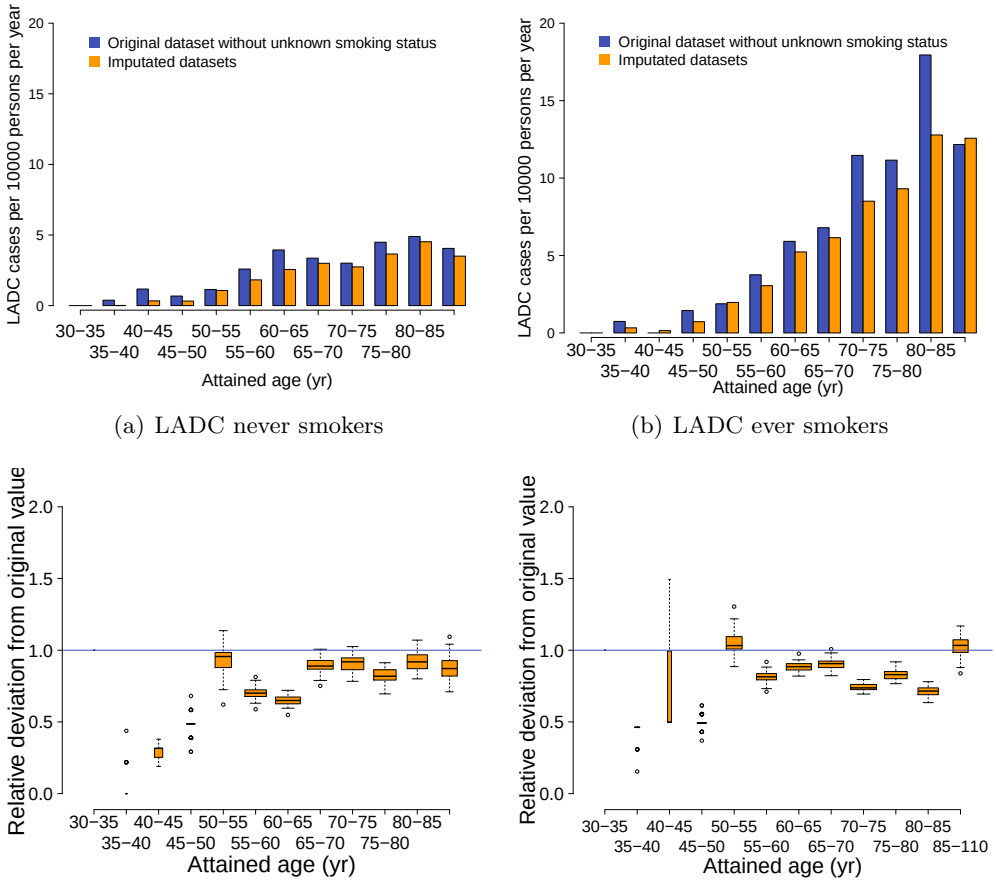
(a) LADC (b) SQUAM



(c) LADC (d) SQUAM

Figure A.1: Stated smoking status for LADC (left panels) and SQUAM (right panels) cases per 10^4 *persons per year comparing three data sets: the original data set without unknown smoking information (light blue), the original data set with unknown smoking information (dark blue) and the 50 imputed data sheets (in orange). Panels (a) right and (b) right describe the distribution of known and unknown smoking history in the original data set with unknown smoking information (dark blue). Panels (a) left and (b) left instead depict the difference of the incidence between the original data set without unknown smoking category and the 50 imputed sheets. The orange bars are the person years weighted mean over the 50 imputed sheets. Panels (c) and (d) represent the relative deviation of the imputed data sheets from the original data set without unknown smoking. The relative deviation is calculated as incidence-ratio between original and imputed dataset. The box-plots represent the distribution of the relative deviations of the 50 imputed sheets. The blue horizontal line represents the value one, it means a correspondence between original without unknown smoking category and imputed data set.*

A.3 Age distribution of cases by smoking status



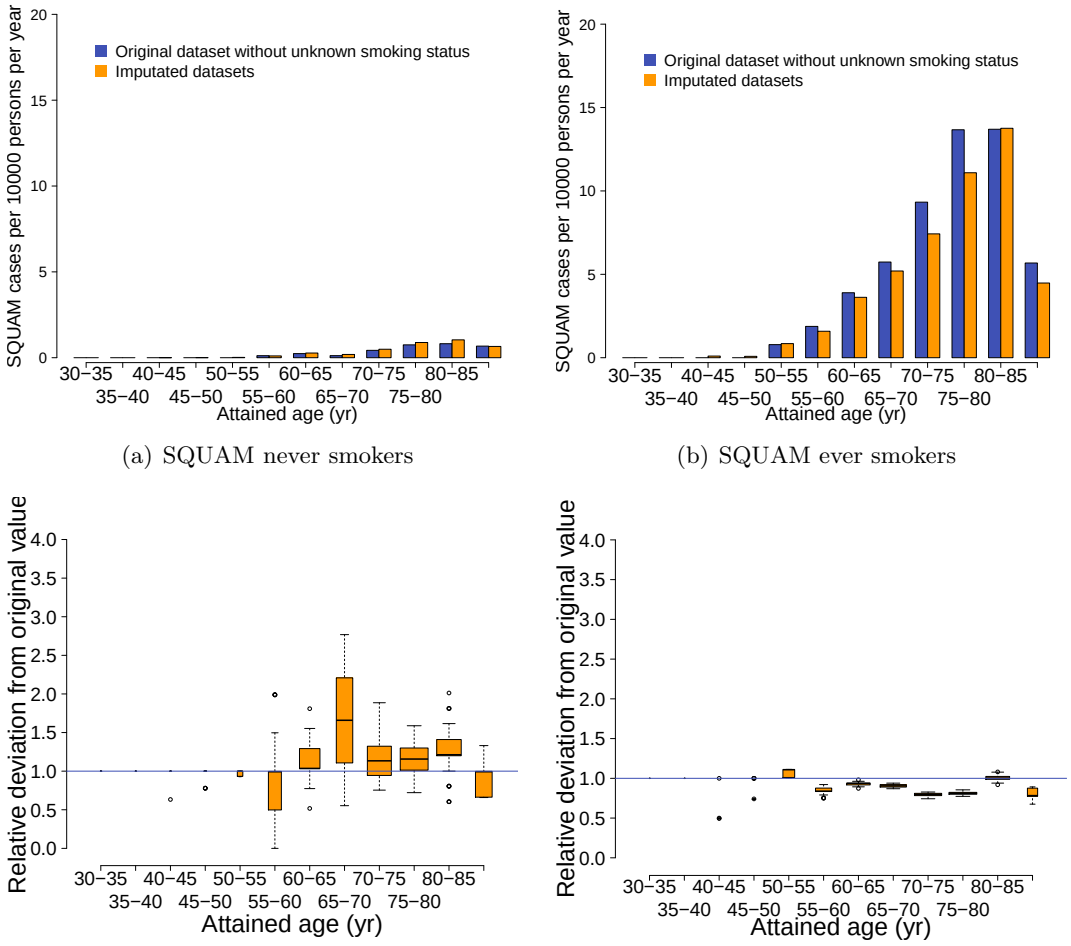
(a) LADC never smokers

(b) LADC ever smokers

(c) LADC never smokers

(d) LADC ever smokers

Figure A.2: Age distribution of LADC cases per 10^4 *persons per year by smoking status (never smokers left panels and ever smokers right panels) comparing two data sets: the original dataset without unknown smoking information (light blue) and the 50 imputed dataset (in orange). Panels (a) and (b), for never and ever smokers respectively, describe the difference of the incidence between the original data set the 50 imputed data sheets. The orange bars are the person years weighted mean over the 50 imputed data sheets. Panels (c) and (d) represent the relative deviation of the imputed sheets from the original dataset. The relative deviation is calculated as incidence-ratio between original and imputed dataset. The box-plots represent the distribution of the relative deviations of the 50 imputed data sheets. The blue horizontal line represents the value one, it means a correspondence between original and imputed dataset.*



(a) SQUAM never smokers

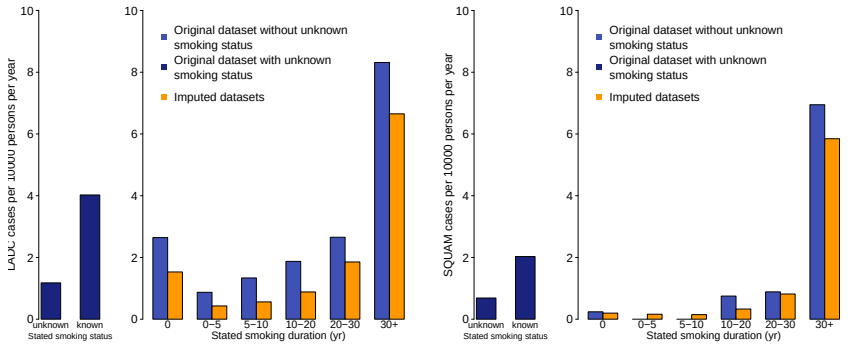
(b) SQUAM ever smokers

(c) SQUAM never smokers

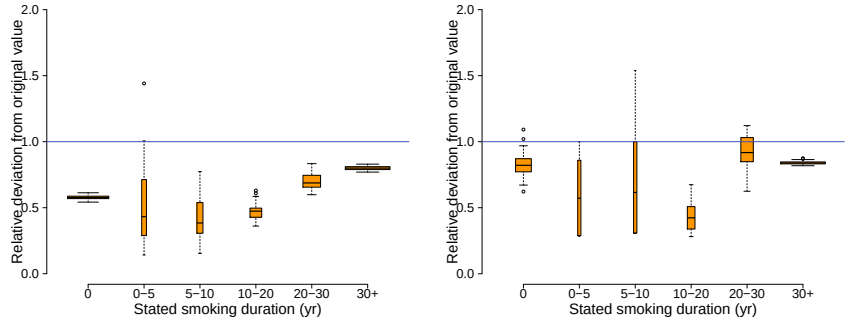
(d) SQUAM ever smokers

Figure A.3: Age distribution of SQUAM cases per 10^4 *persons per year by smoking status (never smokers left panels and ever smokers right panels) comparing two datasets: the original dataset without unknown smoking information (light blue) and the 50 imputed dataset (in orange). Panels (a) and (b), for never and ever smokers respectively, describe the difference of the incidence between the original data set the 50 imputed data sheets. The orange bars are the person years weighted mean over the 50 imputed data sheets. Panels (c) and (d) represent the relative deviation of the imputed sheets from the original dataset. The relative deviation is calculated as incidence-ratio between original and imputed dataset. The box-plots represent the distribution of the relative deviations of the 50 imputed data sheets. The blue horizontal line represents the value one, it means a correspondence between original and imputed dataset.*

A.4 Smoking duration distribution of cases



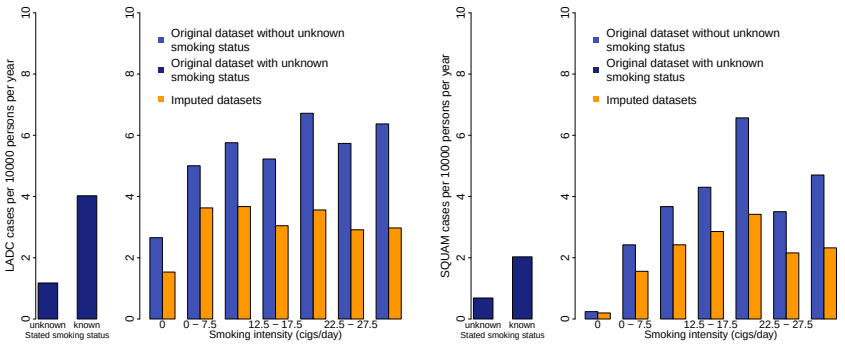
(a) LADC (b) SQUAM



(c) LADC (d) SQUAM

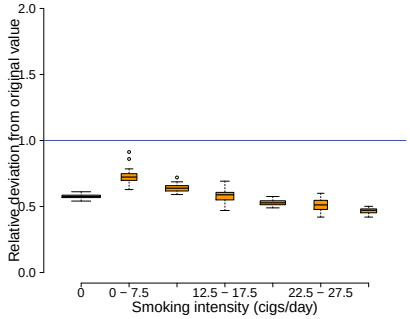
Figure A.4: Stated smoking duration for LADC (left panels) and SQUAM (right panels) cases per 10^4 *persons per year comparing three data sets: the original data set without unknown smoking information (light blue), the original data set with unknown smoking information (dark blue) and the 50 imputed data sheet (in orange). Panel (a) right and (b) right describe the distribution of known and unknown smoking history in the original data set with unknown smoking information (dark blue). Panel (a) left and (b) left instead depict the difference of the incidence between the original data set without unknown smoking category and the 50 imputed data sheets. The orange bars are the person years weighted mean over the 50 imputed data sheets. Panels (c) and (d) represent the relative deviation of the imputed data sheets from the original data set without unknown smoking. The relative deviation is calculated as incidence-ratio between original and imputed data set. The box-plots represent the distribution of the relative deviations of the 50 imputed data sheets. The blue horizontal line represents the value one, it means a correspondence between original without unknown smoking category and imputed data set.*

A.5 Smoking intensity distribution (cigs/day) of cases

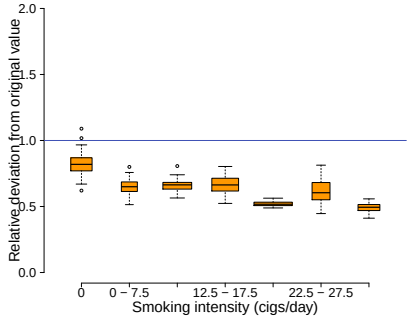


(a) LADC

(b) SQUAM



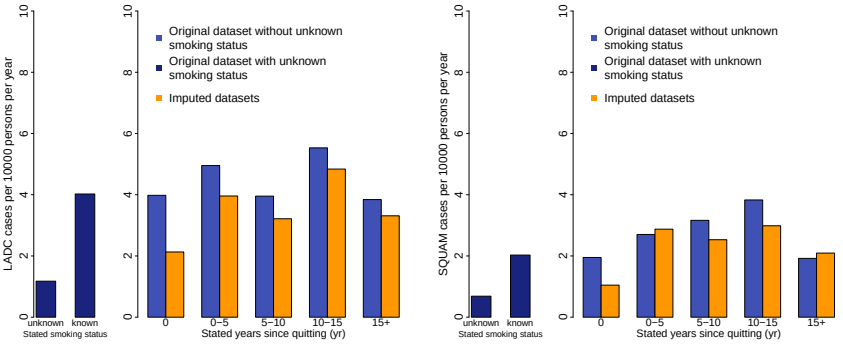
(c) LADC



(d) SQUAM

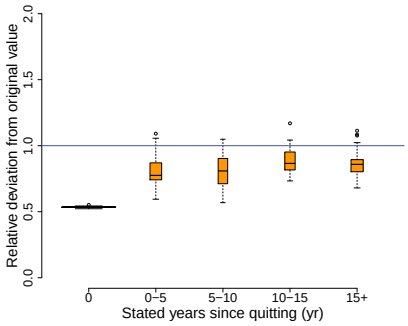
Figure A.5: Stated smoking intensity for LADC (left panels) and SQUAM (right panels) cases per 10^4 *persons per year comparing three data sets: the original data set without unknown smoking information (light blue), the original data set with unknown smoking information (dark blue) and the 50 imputed data sheet (in orange). Panel (a) right and (b) right describe the distribution of known and unknown smoking history in the original data set with unknown smoking information (dark blue). Panel (a) left and (b) left instead depict the difference of the incidence between the original data set without unknown smoking category and the 50 imputed data sheets. The orange bars are the person years weighted mean over the 50 imputed data sheets. Panels (c) and (d) represent the relative deviation of the imputed data sheets from the original data set without unknown smoking. The relative deviation is calculated as incidence-ratio between original and imputed data set. The box-plots represent the distribution of the relative deviations of the 50 imputed data sheets. The blue horizontal line represents the value one, it means a correspondence between original without unknown smoking category and imputed data set.*

A.6 Years since quitting distribution of cases

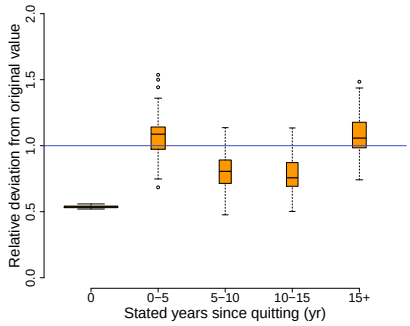


(a) LADC

(b) SQUAM



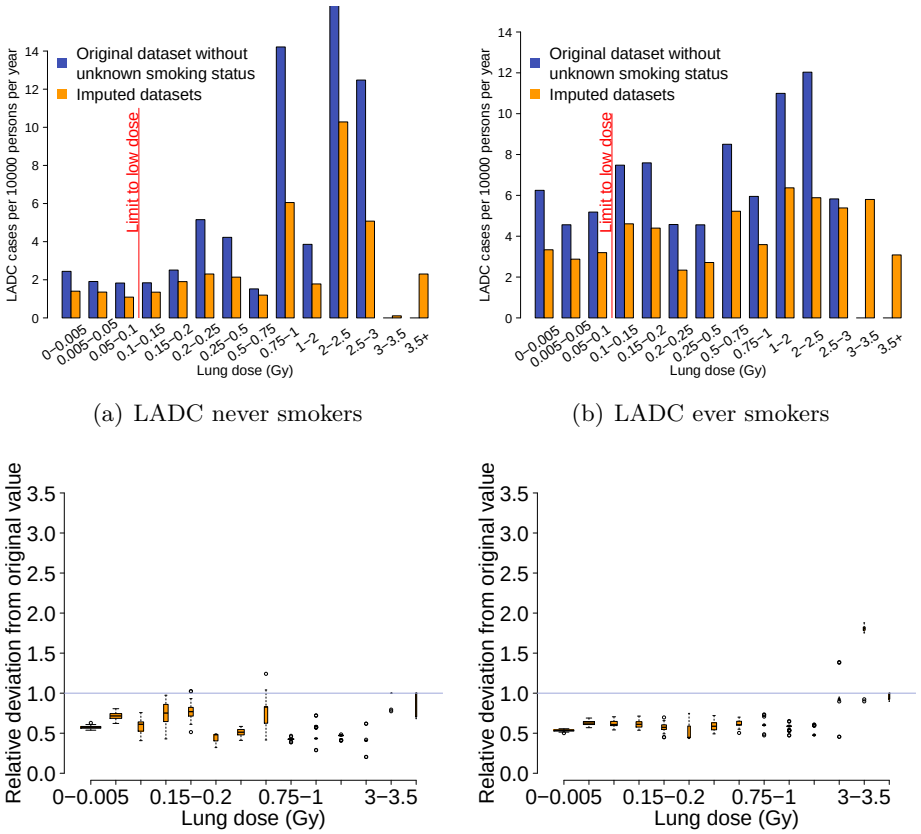
(c) LADC



(d) SQUAM

Figure A.6: Stated years since quitting smoking for LADC (left panels) and SQUAM (right panels) cases per 10^4 *persons per year comparing three data sets: the original data set without unknown smoking information (light blue), the original data set with unknown smoking information (dark blue) and the 50 imputed data sheet (in orange). Panel (a) right and (b) right describe the distribution of known and unknown smoking history in the original data set with unknown smoking information (dark blue). Panel (a) left and (b) left instead depict the difference of the incidence between the original data set without unknown smoking category and the 50 imputed data sheets. The orange bars are the person years weighted mean over the 50 imputed data sheets. Panels (c) and (d) represent the relative deviation of the imputed data sheets from the original data set without unknown smoking. The relative deviation is calculated as incidence-ratio between original and imputed data set. The box-plots represent the distribution of the relative deviations of the 50 imputed data sheets. The blue horizontal line represents the value one, it means a correspondence between original without unknown smoking category and imputed data set.*

A.7 Dose distribution of cases



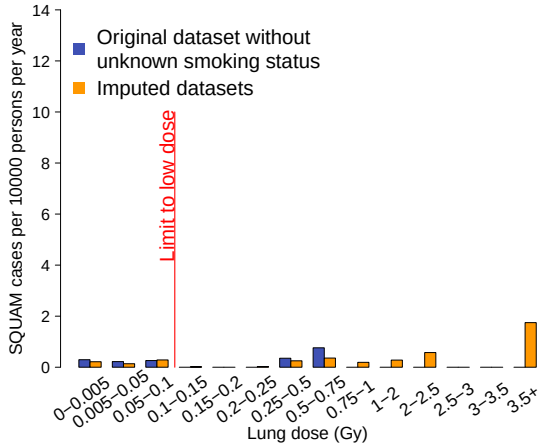
(a) LADC never smokers

(b) LADC ever smokers

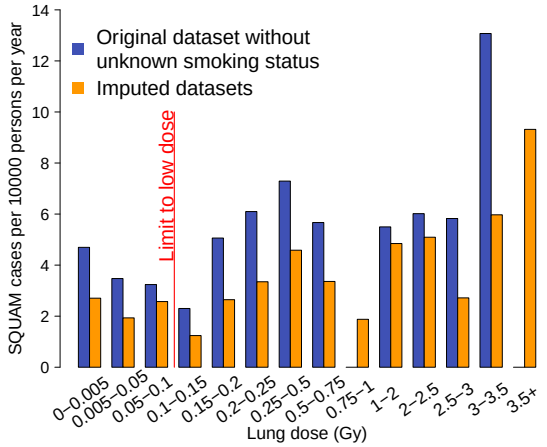
(c) LADC never smokers

(d) LADC ever smokers

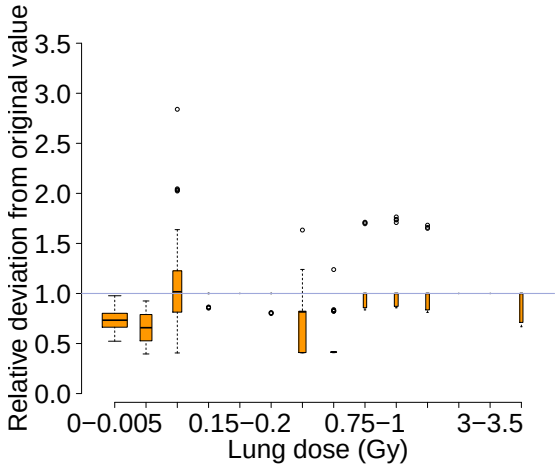
Figure A.7: Lung dose distribution (Gy) of LADC cases per 10^4 persons per year by smoking status (never smokers left panels and ever smokers right panels) comparing two data sets: the original data set without unknown smoking information (light blue) and the 50 imputed data sheets (in orange). Panels (a) and (b), for never and ever smokers respectively, describe the difference of the incidence between the original data set the 50 imputed data sheets. The orange bars are the person years weighted mean over the 50 imputed data sheets. The red line divides low from high dose. Panels (c) and (d) represent the relative deviation of the imputed data sheets from the original data set. The relative deviation is calculated as incidence-ratio between original and imputed data set. The box-plots represent the distribution of the relative deviations of the 50 imputed data sheets. The blue horizontal line represents the value one, it means a correspondence between original and imputed data set.



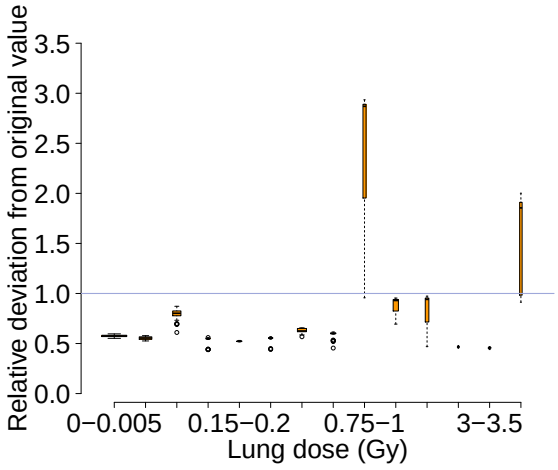
(a) SQUAM never smokers



(b) SQUAM ever smokers



(c) SQUAM never smokers



(d) SQUAM ever smokers

Figure A.8: Lung dose distribution (Gy) of SQUAM cases per 10^4 *persons per year by smoking status (never smokers left panels and ever smokers right panels) comparing two data sets: the original data set without unknown smoking information (light blue) and the 50 imputed data sheets (in orange). Panels (a) and (b), for never and ever smokers respectively, describe the difference of the incidence between the original data set the 50 imputed data sheets. The orange bars are the person years weighted mean over the 50 imputed data sheets. The red line divides low from high dose. Panels (c) and (d) represent the relative deviation of the imputed data sheets from the original data set. The relative deviation is calculated as incidence-ratio between original and imputed data set. The box-plots represent the distribution of the relative deviations of the 50 imputed data sheets. The blue horizontal line represents the value one, it means a correspondence between original and imputed data set.*

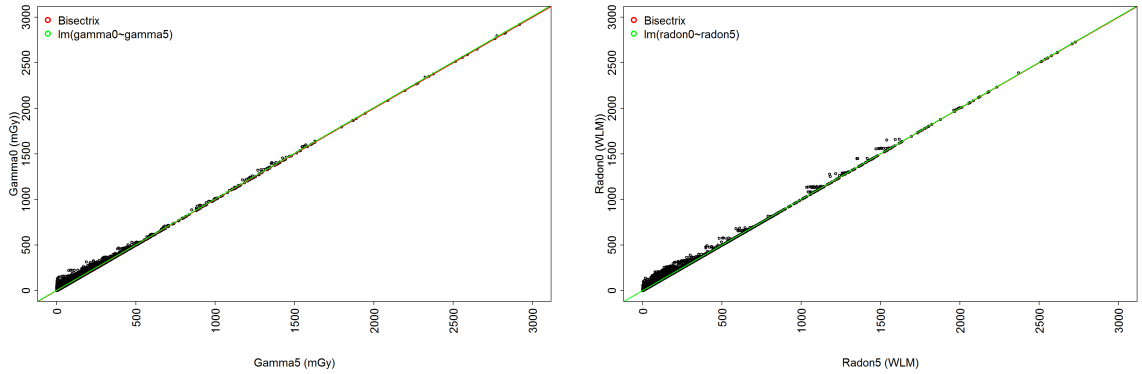
APPENDIX

B

DESCRIPTIVE DATA ANALYSIS OF THE ELDORADO COHORT: SUPPLEMENTARY INFOMATION

This Appendix Chapter contains supplementary information of the Eldorado cohort. Related to Chapter 2.3.

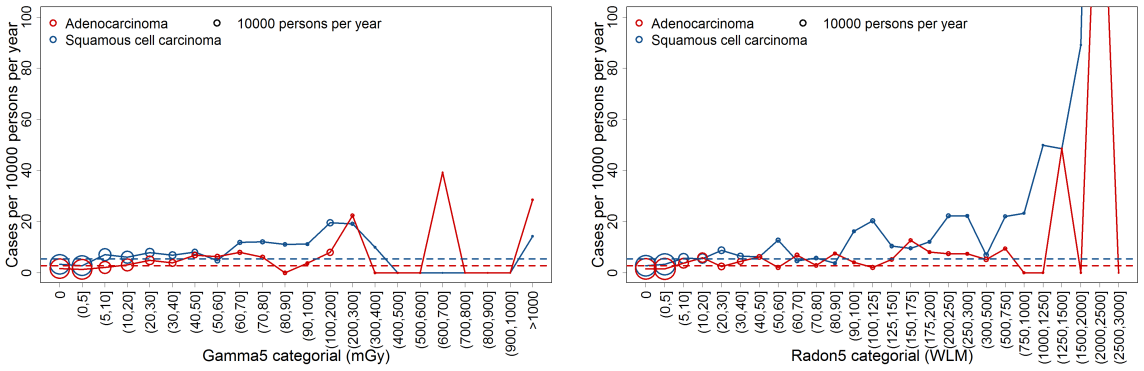
B.1 Lagged exposures vs. non-lagged exposures



(a) (b)

Figure B.1: Comparison of the variables Gamma5 (a) Radon5 (b) with the variables Gammao and Radono, respectively. For both variables the lagged variables almost do not differ from the not lagged ones.

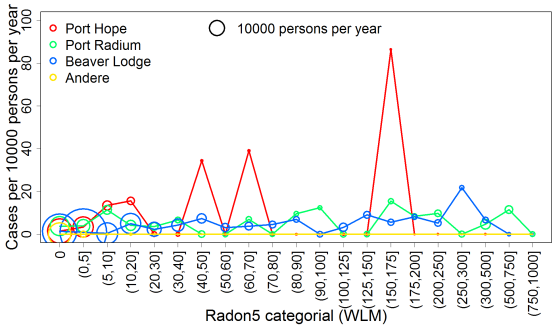
B.2 Dose distribution of cases



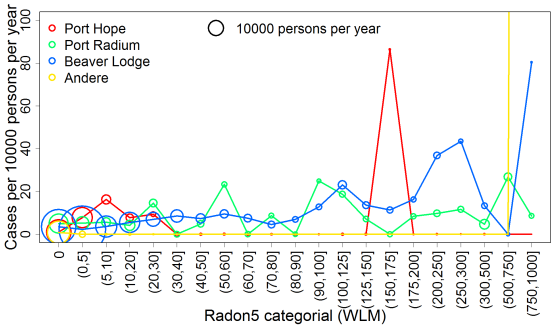
(a) *Gamma5* (b) *Radon5*

Figure B.2: LADC (in red) and SQUAM (in blue) per 10^4 persons per year of the Eldorado cohort as a function of the variable Gamma5 (a) and Radon5 (b).

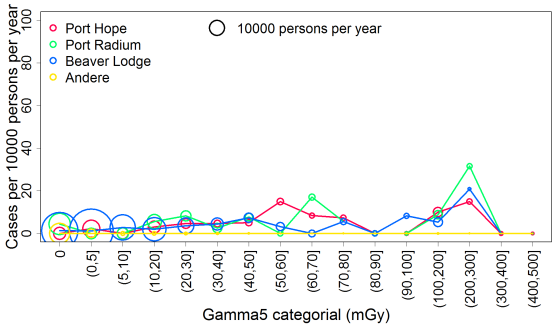
B.3 Dose distribution of cases by facility



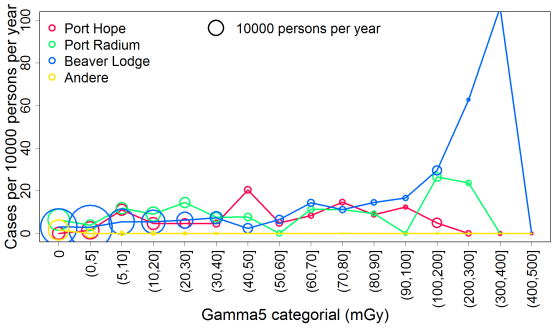
(a) LADC over *Radon5*



(b) SQUAM over *Radon5*



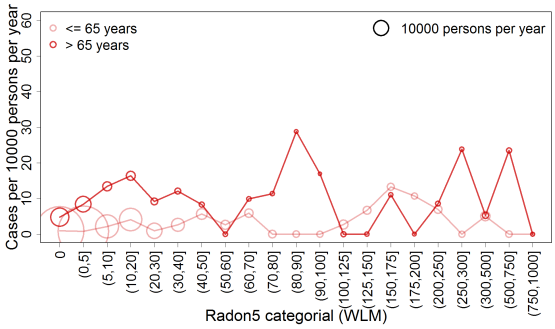
(c) LADC over *Gamma5*



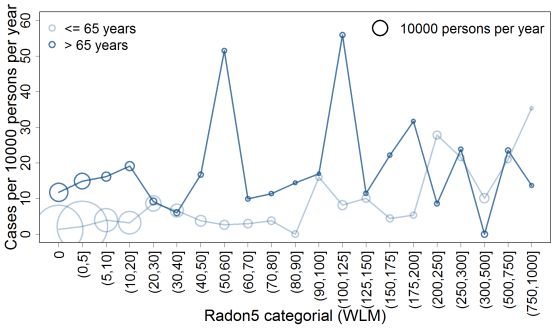
(d) SQUAM over *Gamma5*

Figure B.3: Cases per 10^4 *persons per year for LADC, left panels, and SQUAM, right panels, as a function of Radon5, panels (a) and (b), and of Gamma5, panels (c) and (d), differentiating between the different facilities: Port Hope (red), Port Radium (green), Beaver Lodge (blue) and others (yellow).*

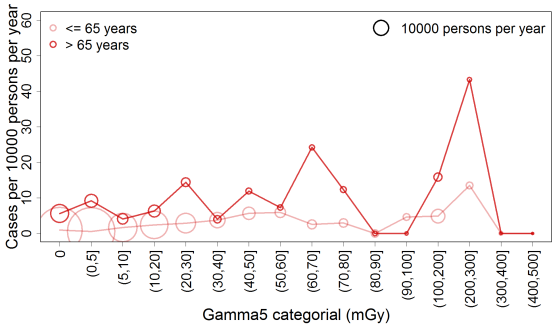
B.4 Dose distribution of cases by age categories



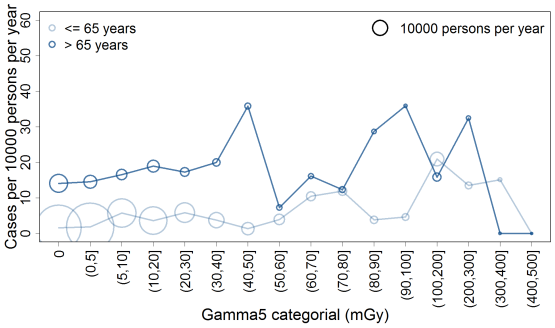
(a) LADC over *Radon5*



(b) SQUAM over *Radon5*



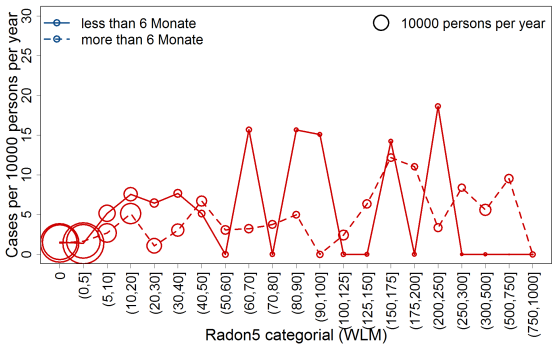
(c) LADC over *Gamma5*



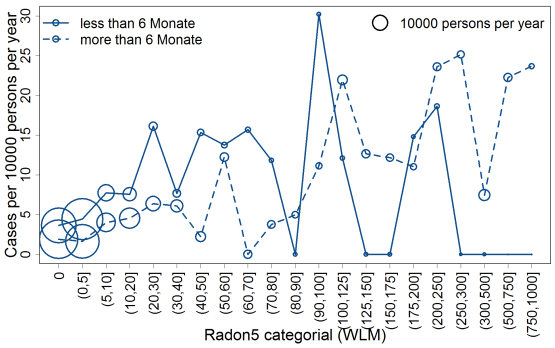
(d) SQUAM over *Gamma5*

Figure B.4: Cases per 10^4 *persons per year for LADC, left panels in red, and SQUAM, right panels in blue, as a function of Radon5, panels (a) and (b), and of Gamma5, panels (c) and (d), differentiating between different age categories of younger/equal than 65 years (soft colors) and older than 65 years (dark colors).*

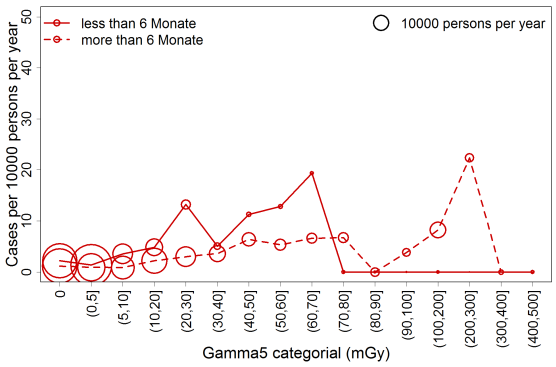
B.5 Dose distribution of cases by working duration categories



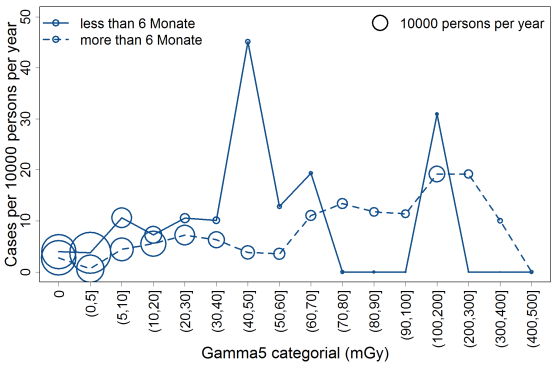
(a) LADC over *Radon5*



(b) SQUAM over *Radon5*



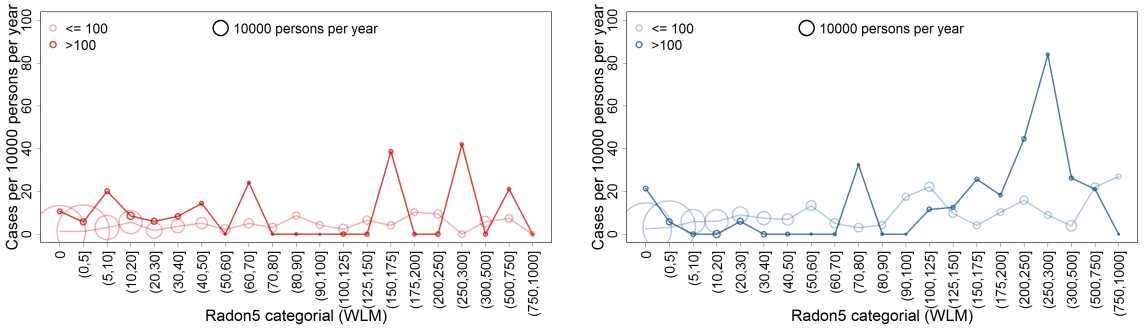
(c) LADC over *Gamma5*



(d) SQUAM over *Gamma5*

Figure B.5: Cases per 10^4 *persons per year for LADC, left panels in orange, and SQUAM, right panels in pink, as a function of Radon5, panels (a) and (b), and of Gamma5, panels (c) and (d), differentiating between different working durations of less than 6 months (solid line) and more than 6 months (dashed lines).*

B.6 Radon5 dose distribution of cases by Gamma5 categories



(a) LADC over *Radon5*

(b) SQUAM over *Radon5*

Figure B.6: LADC (a) and SQUAM (b) per 10^4 *persons per year of the Eldorado cohort as a function of the variable Radon5 differentiating between two categories of Gamma5: low dose in soft colors (≤ 100 mG) and high dose in dark colors (> 100 mG).*

APPENDIX

C

METHODS TO ANALYSE THE COHORTS: SUPPLEMENTARY INFORMATION

This Appendix Chapter contains supplementary information related to Chapter 3.

C.1 Survival function and hazard function

Proof. To prove Theorem 3.1.3 some more definitions are needed.
If the survival function $S(t)$ is defined as the probability that at a given time t no investigated event k_i have already occurred, the cumulative distribution of the complementary event $G(t)$ can be defined as following

$$G(t) = P(T \leq t). \tag{C.1}$$

The probability for the events k_i to occur in a given interval can be derived with the integral over the density function $g(t)$ and is given by

$$P([a, b]) = \int_a^b g(t)\,\mathrm{d}t. \tag{C.2}$$

With the definition of the distribution function $G(t)$ (C.1) the survival function can be rewritten as

$$S(s) = P(T > s) = 1 - G(s). \tag{C.3}$$

Now we have all definition to rewrite the hazard function.
With the formula for conditional probabilities the hazard function can be rewritten as following:

$$h(s) = \lim_{\Delta s \to 0} \frac{P(s \leq T < s + \Delta s,\ T > s)}{\Delta s P(T > s)} = \lim_{\Delta s \to 0} \frac{P(s \leq T < s + \Delta s)}{\Delta s P(T > s)}. \tag{C.4}$$

Remembering equations (C.2) for the numerator of formula (C.4) and (C.3) for the denominator, the formula of the hazard function can be approximated for small Δs as

$$h(s) = \lim_{\Delta s \to 0} \frac{P(s \leq T < s + \Delta s)}{\Delta s P(T > s)} = \frac{g(s)\Delta s}{\Delta s(1 - G(s))} = \frac{g(s)}{1 - G(s)}.$$

Noting that the density function is the derivative of the distribution function we get the final formula for the hazard function.

$$h(s) = \frac{g(s)}{1 - G(s)} = -\frac{d}{ds}\ln(1 - G(s)) = -\frac{d}{ds}\ln(S(s)).$$

□

C.2 Derivation of the two stage clonal expansion model

Proof of Proposition 3.4.5

Proof. The first step is to substitute equations (3.40) of definition (3.4.4) into equation (3.38):

$$\begin{aligned}
\frac{d}{dt}\Psi(y,z,t) &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\frac{d}{dt}P(j,k,t)y^j z^k \\
&= -X\upsilon\left[\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(P(j,k,t)-P(j-1,k,t))y^j z^k\right] && \text{(C.5)}\\
&\quad -\alpha\left[\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(jP(j,k,t)-(j-1)P(j-1,k,t))y^j z^k\right] && \text{(C.6)}\\
&\quad -\beta\left[\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(jP(j,k,t)-(j+1)P(j+1,k,t))y^j z^k\right] && \text{(C.7)}\\
&\quad -\mu\left[\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(jP(j,k,t)-jP(j+1,k-1,t))y^j z^k\right]. && \text{(C.8)}
\end{aligned}$$

Then we rewrite expression (C.5) to (C.8) in terms of the PGF.

- Expression (C.5)

$$\begin{aligned}
&-X\upsilon\left[\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(P(j,k,t)-P(j-1,k,t))y^j z^k\right]\\
=&-X\upsilon\left[\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(P(j,k,t)y^j z^k-\underbrace{\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}P(j-1,k,t))y^j z^k}_{j'=j-1}\right]\\
=&-X\upsilon\left[\underbrace{\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(P(j,k,t)y^j z^k}_{\Psi(y,z,t)}-\underbrace{y\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}P(j',k,t))y^{j'} z^k}_{y\Psi(y,z,t)}\right]\\
=&-X\upsilon\Psi(y,z,t)(1-y)
\end{aligned}$$

- Expression (C.6)

$$
\begin{aligned}
&-\alpha\left[\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(jP(j,k,t)-(j-1)P(j-1,k,t))y^j z^k\right]\\
=&-\alpha\left[\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(jP(j,k,t)y^j z^k-\underbrace{\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(j-1)P(j-1,k,t))y^j z^k}_{j'=j-1}\right]\\
=&-\alpha\left[y\underbrace{\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(jP(j,k,t)y^{j-1}z^k}_{\frac{\partial}{\partial y}\Psi(y,z,t)}-y\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}j'P(j',k,t))y^{j'}z^k\right]\\
=&-\alpha\left[y\frac{\partial}{\partial y}\Psi(y,z,t)-y^2\underbrace{\sum_{j'=0}^{\infty}\sum_{k=0}^{\infty}j'P(j',k,t))y^{j'-1}z^k}_{\frac{\partial}{\partial y}\Psi(y,z,t)}\right]\\
=&\alpha\frac{\partial}{\partial y}\Psi(y,z,t)(y^2-y)
\end{aligned}
$$

- Expression (C.7)

$$
\begin{aligned}
&-\beta\left[\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(jP(j,k,t)-(j+1)P(j+1,k,t))y^j z^k\right]\\
=&-\beta\left[\underbrace{\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(jP(j,k,t)y^j z^k}_{y\frac{\partial}{\partial y}\Psi(y,z,t)}-\underbrace{\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(j-1)P(j-1,k,t))y^j z^k}_{\frac{\partial}{\partial y}\Psi(y,z,t)}\right]\\
=&\beta\frac{\partial}{\partial y}\Psi(y,z,t)(1-y)
\end{aligned}
$$

- Expression (C.8)

$$
\begin{aligned}
& -\mu \left[\sum_{j=0}^{\infty} \sum_{k=0}^{\infty} (jP(j,k,t) - jP(j,k-1,t)) y^j z^k \right] \\
= & -\mu \left[\underbrace{\sum_{j=0}^{\infty} \sum_{k=0}^{\infty} (jP(j,k,t) y^j z^k}_{y \frac{\partial}{\partial y} \Psi(y,z,t)} - \underbrace{\sum_{j=0}^{\infty} \sum_{k=0}^{\infty} jP(j,k-1,t)) y^{(j-1)} z^k}_{k'=k-1} \right] \\
= & -\mu \left[y \frac{\partial}{\partial y} \Psi(y,z,t) - yz \underbrace{\sum_{j=0}^{\infty} \sum_{k'=0}^{\infty} jP(j,k',t)) y^{(j-1)} z^{k'}}_{\frac{\partial}{\partial y} \Psi(y,z,t)} \right] \\
= & \mu \frac{\partial}{\partial y} \Psi(y,z,t)(yz - y) \quad .
\end{aligned}
$$

These calculations lead to the final formula

$$
\begin{aligned}
\frac{\partial}{\partial t} \Psi(y,z,t) &= F(y,t) \Psi(y,z,t) + G(y,z,t) \frac{\partial}{\partial y} \Psi(y,z,t) \\
\Psi(y,z,t_0) &= 1
\end{aligned} \tag{C.9}
$$

with

$$
\begin{aligned}
F(y,t) &:= (y-1) X \nu \\
G(y,z,t) &:= \mu yz + \alpha y^2 - [\alpha + \beta + \mu] y + \beta
\end{aligned}
$$

□

C.3 Backward recursion of the two stage clonal expansion model

Table C.1: Recursion equations for the hazard $h(t)$ at age t of the TSCE-model with piecewise-constant parameters in k age-intervals.

Recursion algorithm for the hazard function with piecewise-constant parameters of the TSCE	
$h(t) = \sum_{i=1}^{k} \frac{C_i}{\delta_i} \frac{1}{f_i(s_{i-1})} \frac{\partial}{\partial t} f_i(s_{i-1})$	
$C_i = X\nu_i\mu_0$	
$\delta_i := \alpha_i\mu_0$	
$A_i = -\frac{1}{2}(\gamma_i + \sqrt{\gamma_i^2 + 4\delta_i\theta_i})$	
$B_i = \frac{1}{2}(-\gamma_i + \sqrt{\gamma_i^2 + 4\delta_i\theta_i})$	
$\theta_i = \frac{\mu_i}{\mu_0}$	
$\gamma_i := \alpha_i - \beta_i - \mu_i$	
$f_i(s_{i-1}) = B_i e^{A_i(s_{i-1}-t)} - A_i e^{B_i(s_{i-1}-t)}$	first step $i = k$
$f_i(s_{i-1}) = (B_i - w_i(s_i))e^{A_i(s_{i-1}-s_i)} + (w_i(s_i) - A_i)e^{B_i(s_{i-1}-s_i)}$	all steps $i \neq k$
$\frac{\partial}{\partial t} f_i(s_{i-1}) = A_i B_i (e^{B_i(s_{i-1}-t)} - e^{A_i(s_{i-1}-t)})$	first step $i = k$
$\frac{\partial}{\partial t} f_i(s_{i-1}) = -\frac{\partial}{\partial t} w_i(s_i)(e^{A_i(s_{i-1}-s_i)} - e^{B_i(s_{i-1}-s_i)})$	all steps $i \neq k$
$w_i(s_i) = 0$	first step $i = k$
$\frac{\partial}{\partial t} w_i(s_i) = \delta_i\theta_i$	first step $i = k$
$w_{i-1}(s_{i-1}) = \frac{\delta_{i-1}}{\delta_i} w_i(s_{i-1})$	all steps $i \neq k$
$\frac{\partial}{\partial t} w_{i-1}(s_{i-1}) = \frac{\delta_{i-1}}{\delta_i} \frac{\partial}{\partial t} w_i(s_{i-1})$	all steps $i \neq k$
$w_i(s_{i-1}) = \frac{A_i B_i (e^{A_i(s_{i-1}-t)} - e^{B_i(s_{i-1}-t)})}{B_i e^{A_i(s_{i-1}-t)} - A_i e^{B_i(s_{i-1}-t)}}$	first step $i = k$
$w_i(s_{i-1}) = \frac{A_i B_i (e^{A_i(s_{i-1}-s_i)} - e^{B_i(s_{i-1}-s_i)}) - w_i(s_i)(A_i e^{A_i(s_{i-1}-s_i)} - B_i e^{B_i(s_{i-1}-s_i)})}{(B_i - w_i(s_i))e^{A_i(s_{i-1}-s_i)} + (w_i(s_i) - A_i)e^{B_i(s_{i-1}-s_i)}}$	all steps $i \neq k$
$\frac{\partial}{\partial t} w_i(s_{i-1}) = \frac{A_i B_i (A_i - B_i)^2 e^{A_i(s_{i-1}-t)} e^{B_i(s_{i-1}-t)})}{(B_i e^{A_i(s_{i-1}-t)} - A_i e^{B_i(s_{i-1}-t)})^2}$	first step $i = k$
$\frac{\partial}{\partial t} w_i(s_{i-1}) = \frac{\partial}{\partial t} w_i(s_i) \frac{(A_i - B_i)^2 e^{A_i(s_{i-1}-s_i)} e^{B_i(s_{i-1}-s_i)}}{[(B_i - w_i(s_i))e^{A_i(s_{i-1}-s_i)} + (w_i(s_i) - A_i)e^{B_i(s_{i-1}-s_i)}]^2}$	all steps $i \neq k$

C.4 Backward recursion of the hybrid three stage clonal expansion model

Table C.2: Recursion equations for the hazard $h(t)$ at age t of the H3SCE-modelwith piecewise-constant parameters in k age-intervals.

Recursion algorithm for the hazard function with piecewise-constant parameters of the H3SCE-model	
$h(t) = \sum_{i=1}^{k} \frac{C_{2_i}}{\delta_{2_i}} \frac{(s_{i-1}+s_i)}{2} \frac{1}{l_i(s_{i-1},t)} \frac{\partial}{\partial t} l_i(s_{i-1},t)$	
$C_{2_i} = N\nu_{0_i}\nu_{1_i}\mu_{2_1}$	
$\delta_{2_i} := \alpha_{2_i}\mu_{2_i}$	
$D_i = -\frac{1}{2}(\gamma_{2_i} + \sqrt{\gamma_{2_i}^2 + 4\delta_{2_i}\theta_{2_i}})$	
$E_i = \frac{1}{2}(-\gamma_{2_i} + \sqrt{\gamma_{2_i}^2 + 4\delta_{2_i}\theta_{2_i}})$	
$\theta_{2_i} = \frac{\mu_{2_i}}{\mu_{2_0}}$	
$\gamma_{2_i} := \alpha_{2_i} - \beta_{2_i} - \mu_{2_i}$	
$l_i(s_{i-1}) = E_i e^{D_i(s_{i-1}-t)} - D_i e^{E_i(s_{i-1}-t)}$	first step $i = k$
$l_i(s_{i-1}) = (E_i - w_{2_i}(s_i))e^{D_i(s_{i-1}-s_i)} + (w_{2_i}(s_i) - D_i)e^{E_i(s_{i-1}-s_i)}$	all steps $i \neq k$
$\frac{\partial}{\partial t} l_i(s_{i-1}) = D_i E_i (e^{E_i(s_{i-1}-t)} - e^{D_i(s_{i-1}-t)})$	first step $i = k$
$\frac{\partial}{\partial t} l_i(s_{i-1}) = -\frac{\partial}{\partial t} w_{2_i}(s_i)(e^{D_i(s_{i-1}-s_i)} - e^{E_i(s_{i-1}-s_i)})$	all steps $i \neq k$
$w_{2_i}(s_i) = 0$	first step $i = k$
$\frac{\partial}{\partial t} w_{2_i}(s_i) = \delta_{2_i}\theta_{2_i}$	first step $i = k$
$w_{2_{i-1}}(s_{i-1}) = \frac{\delta_{2_{i-1}}}{\delta_{2_i}} w_{2_i}(s_{i-1})$	all steps $i \neq k$
$\frac{\partial}{\partial t} w_{2_{i-1}}(s_{i-1}) = \frac{\delta_{2_{i-1}}}{\delta_{2_i}} \frac{\partial}{\partial t} w_{2_i}(s_{i-1})$	all steps $i \neq k$
$w_{2_i}(s_{i-1}) = \frac{D_i E_i (e^{D_i(s_{i-1}-t)} - e^{E_i(s_{i-1}-t)})}{E_i e^{D_i(s_{i-1}-t)} - D_i e^{E_i(s_{i-1}-t)}}$	first step $i = k$
$w_{2_i}(s_{i-1}) = \frac{D_i E_i (e^{D_i(s_{i-1}-s_i)} - e^{E_i(s_{i-1}-s_i)}) - w_{2_i}(s_i)(D_i e^{D_i(s_{i-1}-s_i)} - E_i e^{E_i(s_{i-1}-s_i)})}{(E_i - w_{2_i}(s_i))e^{D_i(s_{i-1}-s_i)} + (w_{2_i}(s_i) - D_i)e^{E_i(s_{i-1}-s_i)}}$	all steps $i \neq k$
$\frac{\partial}{\partial t} w_{2_i}(s_{i-1}) = \frac{D_i E_i (D_i - E_i)^2 e^{D_i(s_{i-1}-t)} e^{E_i(s_{i-1}-t)})}{(E_i e^{D_i(s_{i-1}-t)} - D_i e^{E_i(s_{i-1}-t)})^2}$	first step $i = k$
$\frac{\partial}{\partial t} w_{2_i}(s_{i-1}) = \frac{\partial}{\partial t} w_{2_i}(s_i) \frac{(D_i - E_i)^2 e^{D_i(s_{i-1}-s_i)} e^{E_i(s_{i-1}-s_i)}}{[(E_i - w_{2_i}(s_i))e^{D_i(s_{i-1}-s_i)} + (w_{2_i}(s_i) - D_i)e^{E_i(s_{i-1}-s_i)}]^2}$	all steps $i \neq k$

APPENDIX

D

MATLAB CODE FOR TWO AND THREE STAGE CLONAL EXPANSION MODELS

```
function [sumh_tot] = recursion(u,par)

n                   = size(u,2);
delta               = exp(par(2));
theta               = 1;

%%
delta_beginning     = 1;
delta_beginning1    = delta;

%dw at the beginning is AB=\delta*theta,
%then it becomes the function written down
dwk_ui_1         = -delta.*theta;

%w at the beginning is =0, then it becomes the function written down
wk_ui_1             = 0.*ones(size(u,1),1);

%% hazard at the beginning, setting the first value for the sum
sumh                = 0;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% LOOP %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
for      i=n:-1:2

X                   = exp(par(3));
gamma               = par(1);
```

```
A                   = (-1./2).*...
                        (gamma + sqrt((gamma.^2)+4.*delta.*theta)).';
B                   = (1./2).*...
                        (-gamma + sqrt((gamma.^2)+4.*delta.*theta)).';

expA                = 1+ expm1(A.*(u(:,i-1) -u(:,i)));
expB                = 1+ expm1(B.*(u(:,i-1) -u(:,i)));

%% writing down the things for the dfi_ui_1
%boundary condition first step derivative: we have 1
dwi_ui              = (delta./delta_beginning1).*dwk_ui_1;
dfi_ui_1            = -dwi_ui.*(expA.'-expB.');

%% writing down all thing for function fi_ui_1
%boundary condition first step: we have 0
wi_ui               = (delta./delta_beginning).*wk_ui_1.';

fi_ui_1             = (B.' -wi_ui).*expA.' +(wi_ui -A.').*expB.';

%% writing down the hazard
%two stage clonal expansion model
sumh                = ((X.*(dfi_ui_1))./(delta.*(fi_ui_1))) +sumh;

%hybrid three stage clonal expansion model
sumh                = ((u(:,i-1) +u(:,i))./2)...
                       .*((X.*(dfi_ui_1))./(delta.*(fi_ui_1))) +sumh;

%% definig the new delta for the initial conditions for i not= k
delta_beginning     = delta;
%after this loop we have delta i, and the next loop will be with i-1
%-> delta(i-1)/delta(i)
delta_beginning1    = delta;

% definig the new wi(ui-1) for the initial conditions for i not= k
wk_ui_1             = (A.*B.*(expA-expB) ...
                      -wi_ui.'.*(A.*expA -B.*expB))./ fi_ui_1.';

% definig the new dwi(ui-1) for the initial conditions for i not= k
dwk_ui_1            = ((dwi_ui.'.*(A-B).^2.*expA.*expB)...
                          ./(fi_ui_1.').^2).';
end

sumh_tot                = sumh;
```

APPENDIX

E

RISK ASSESSMENT FOR LUNG ADENOCARCINOMA IN THE LIFE SPAN STUDY COHORT

This Appendix Chapter contains supplementary information related to Chapter 4.

E.1 Deviance comparison

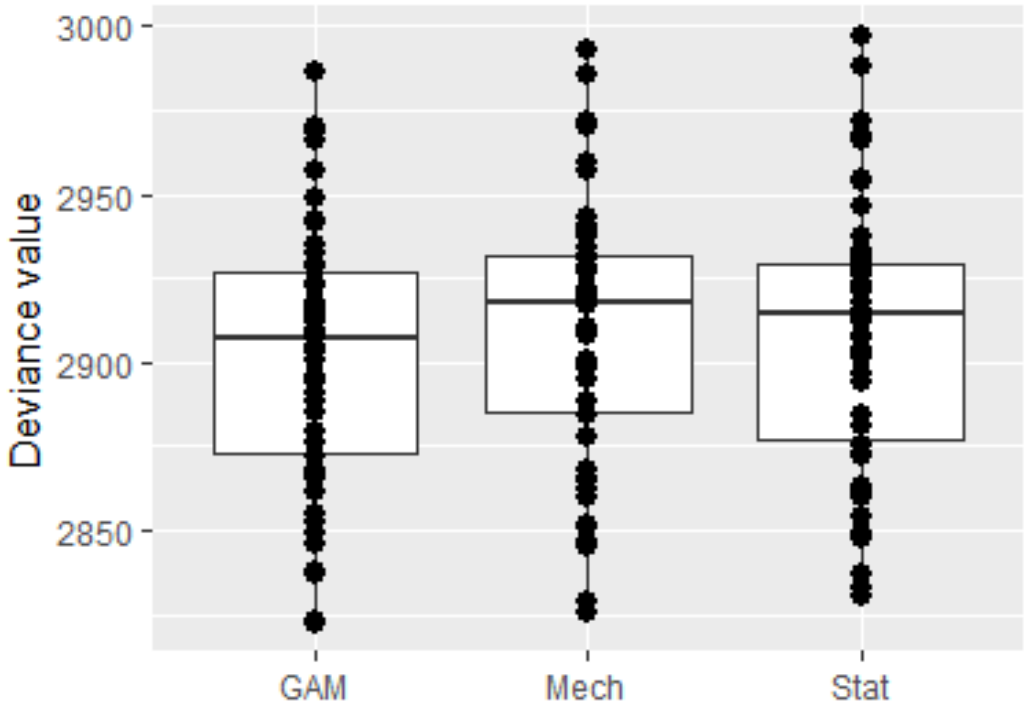


Figure E.1: Comparison of the deviance of models $Stat_{LSSC}^{LADC}$, M_3^{LADC} *and* GAM_{LSS}^{LADC}. *Boxplots represent the variation between the 50 imputed data sets. Mean values for GA, state-of-the-art descriptive and mechanistic models, respectively: 5054.257, 5057.7 and 5050.4.* GAM_{LSS}^{SQUAM} *gives a slightly better fit of the data compared to the other two modalities. No outliers are presented in the model. The difference in the deviance between the data sets in one model has to be attributed to the fact that with imputation each data set has a different number of strata.*

E.2 The radiation related excess relative risk for exposed never smokers

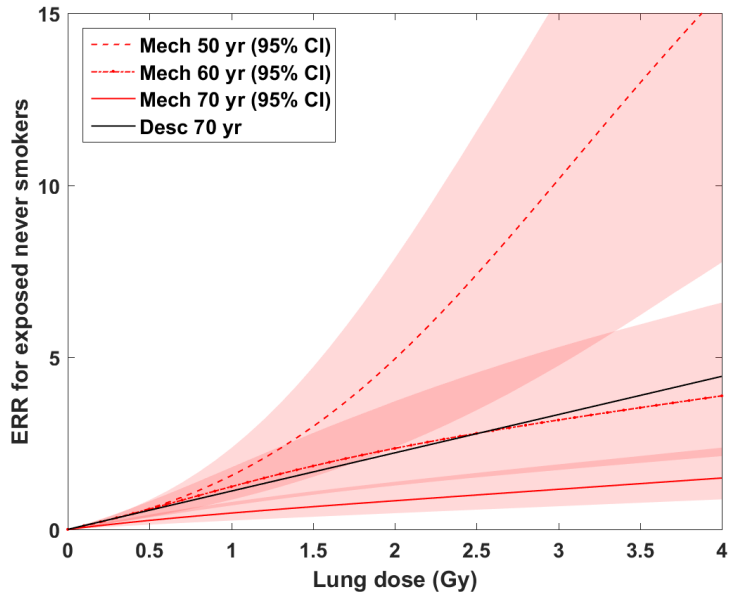
In the case of a two path model, where the pathways are mutually exclusive the rad-ERR reads

$$ERR_\gamma = \frac{h_{tot}(n\,Gy, 0\,\frac{sigs}{day})}{h_{tot}(0\,Gy, 0\,\frac{sigs}{day})} - 1 \tag{E.1}$$

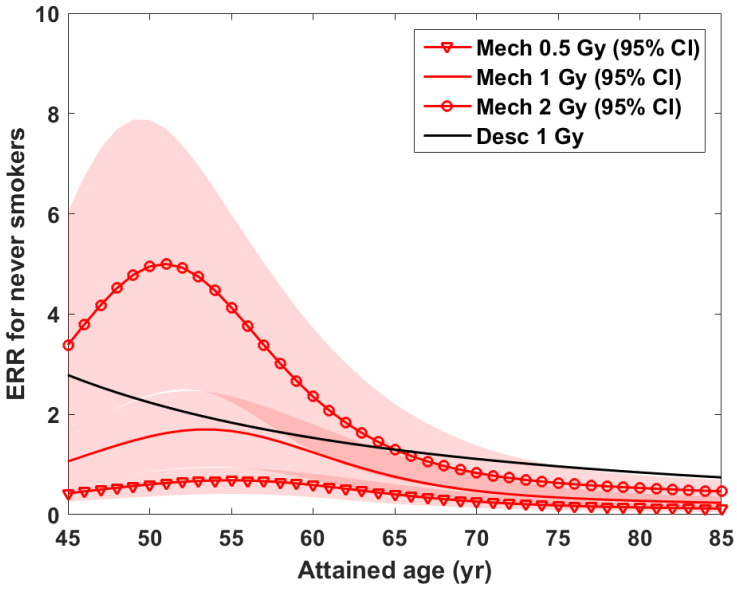
$$= \frac{h_{R^{MUT}}(n\,Gy) + h_{T^{MUT}}(0\,\frac{sigs}{day})}{h_{R^{MUT}}(0\,Gy) + h_{T^{MUT}}(0\,\frac{sigs}{day})} - 1 \tag{E.2}$$

$$= \frac{h_{R^{MUT}}(n\,Gy) - h_{R^{MUT}}(0\,Gy)}{h_{R^{MUT}}(0\,Gy) + h_{T^{MUT}}(0\,\frac{sigs}{day})} \tag{E.3}$$

with $n > 0$, where h_{tot}, $h_{R^{MUT}}$ and $h_{T^{MUT}}$ represents the total hazard of the model, the hazard in the R^{MUT} pathway and the hazard in the T^{MUT} pathway, respectively. This formula applies only because radiation and smoking act on two different pathways independently.



(a) Never smokers exposed at age 30 yr



(b) Never smokers exposed at age 30 yr

Figure E.2: Excess Relative Risk (ERR) from the preferred mechanistic model M_3^{LADC} (Mech) for LADC in the LSS cohort for never smokers exposed to radiation at 30 yr. Radiation only affects the R^{MUT} pathway independent of sex and smoking status. (a) For attained ages 50, 60 and 70 yr the ERR responds non-linearly to doses in the range 0 - 4 Gy. However, on the biological level radiation action is modeled by a linear increase of the clonal expansion rate in the R^{MUT} pathway which lasts for life. (b) For lung doses of 0.5, 1 and 2 Gy the ERR from the preferred mechanistic model peaks at decreasing age with increasing value. The ERR estimate at 1 Gy from $Stat_{LSS}^{LADC}$ 4.1 (Desc) is shown for comparison.

E.3 The smoking-related excess relative risk for unexposed smokers

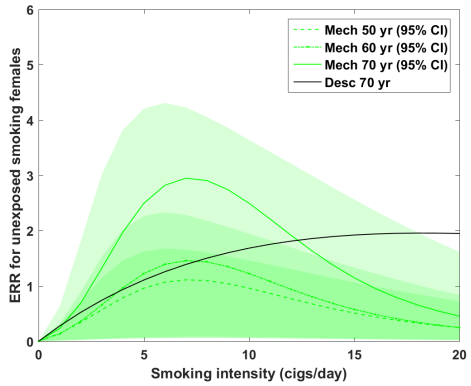
In the case of a two path model, where the pathways are mutually exclusive the smoking-ERR reads

$$ERR_{smk} = \frac{h_{tot}(0\,Gy, m\,\frac{sigs}{day})}{h_{tot}(0\,Gy, 0\,\frac{sigs}{day})} - 1 \tag{E.4}$$

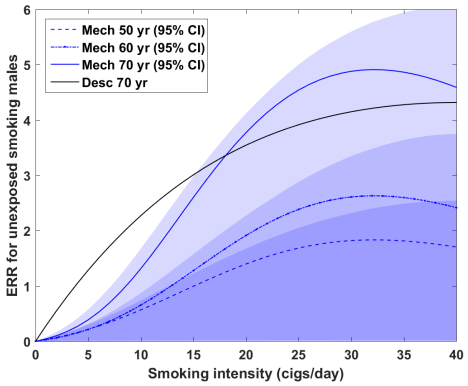
$$= \frac{h_{R^{MUT}}(0\,Gy) + h_{T^{MUT}}(m\,\frac{sigs}{day})}{h_{R^{MUT}}(0\,Gy) + h_{T^{MUT}}(0\,\frac{sigs}{day})} - 1 \tag{E.5}$$

$$= \frac{h_{T^{MUT}}(m\,\frac{sigs}{day}) - h_{T^{MUT}}(0\,\frac{sigs}{day})}{h_{R^{MUT}}(0\,Gy) + h_{T^{MUT}}(0\,\frac{sigs}{day})} \tag{E.6}$$

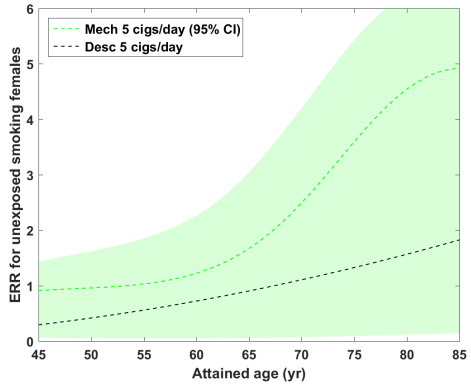
with $m > 0$, where h_{tot}, $h_{R^{MUT}}$ and $h_{T^{MUT}}$ represents the total hazard of the model, the hazard in the R^{MUT} pathway and the hazard in the T^{MUT} pathway, respectively. This formula applies only because radiation and smoking act on two different pathways independently.



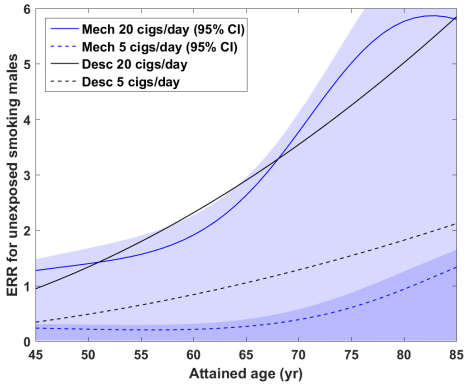
(a) Lifelong female smokers starting at age 20 yr



(b) Lifelong male smokers starting at age 20 yr



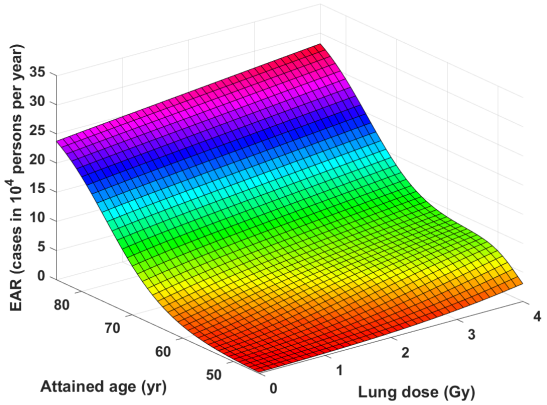
(c) Lifelong female smokers of 5 cigs/day starting at age 20 yr



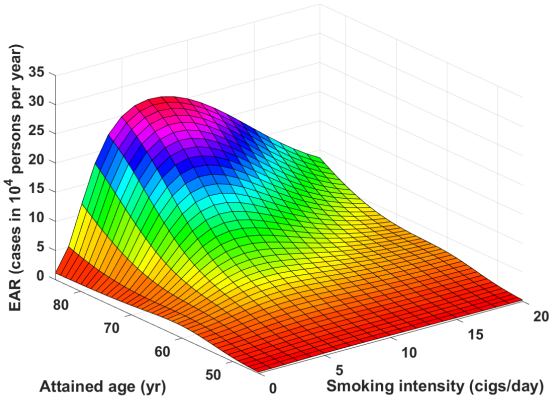
(d) Lifelong male smokers of 20 cigs/day starting at age 20 yr

Figure E.3: Excess Relative Risk (ERR) from the preferred mechanistic model M_3^{LADC} (Mech) for smoking-induced LADC in the LSS cohort for unexposed lifelong smokers starting at age 20 yr. Smoking only affects the T^{MUT} pathway with a markedly different response in both sexes but independent of radiation exposure. Panels (a) and (b) depict the ERR for attained ages of 50, 60 and 70 yr which is determined by the sex-dependent linear-exponential response of the clonal expansion rate in the T^{MUT} pathway. The implausibly strong attenuation of the EAR for females smoking more the 10 cigs/day is possibly caused by a reporting bias. Panels (c) and (d) depict the ERR over age for 5 cigs/day (males and females) and 20 cigs/day (males only). Female smokers of 5 cigs/day and males smokers of 20 cigs/day possess about the same risk. ERR estimates from $Stat_{LSS}^{LADC}$ (Desc) are shown for comparison.

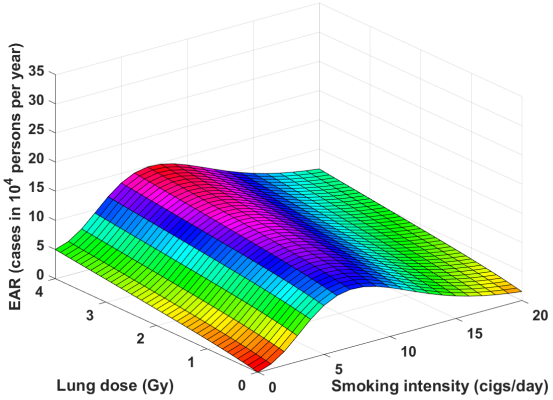
E.4 Excess absolute rates for a smoking irradiated person



(a) Females with smoking intensity 5 cigs/day

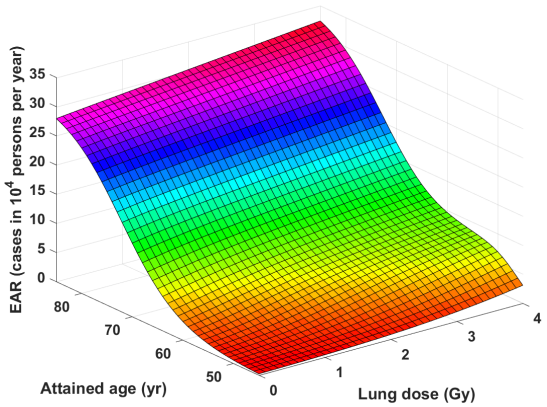


(b) Smoking females with lung dose 1 Gy

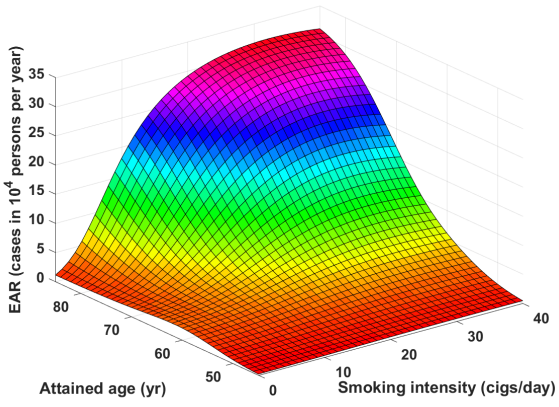


(c) Related to Figure 5. Smoking females at attained age 70 yr exposed to radiation at age 30 yr

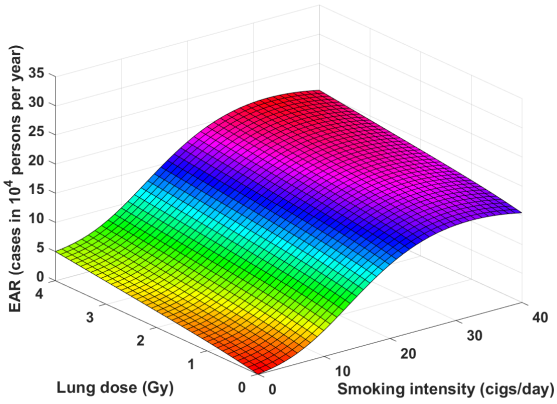
Figure E.4: Bivariate EARs (cases in 10,000 persons per year) for smoking-induced and radiation-induced LADC from the preferred mechanistic model M_3^{LADC} for lifelong female smokers starting at age 20 yr and exposed to radiation at age 30 yr. To eliminate the influence for city of residence person-year weighted city means are used. (a) Dependence on attained age and lung dose for comparison with Figure 4.13(a). (b) Dependence on attained age and smoking intensity. For a lung dose of 1 Gy comparison with Figure 4.12(a) reveals that the radiation effect on the EAR is negligible. (c) Additive effect of radiation and smoking at attained age 70 yr.



(a) Males with smoking intensity 20 cigs/day



(b) Smoking males with lung dose 1 Gy



(c) Smoking males at attained age 70 yr exposed to radiation at age 30 yr

Figure E.5: Bivariate EARs (cases in 10,000 persons per year) for smoking-induced and radiation-induced for LADC from the preferred mechanistic model M_3^{LADC} for lifelong male smokers starting at age 20 yr and exposed to radiation at age 30 yr. To eliminate the influence for city of residence person-year weighted city means are used. (a) Dependence on attained age and lung dose for comparison with Figure 4.13(b). (b) Dependence on attained age and smoking intensity. For a lung dose of 1 Gy comparison with Figure 4.12(b) reveals that the radiation effect on the EAR is negligible especially for heavy smokers. (c) Additive effect of radiation and smoking at attained age 70 yr.

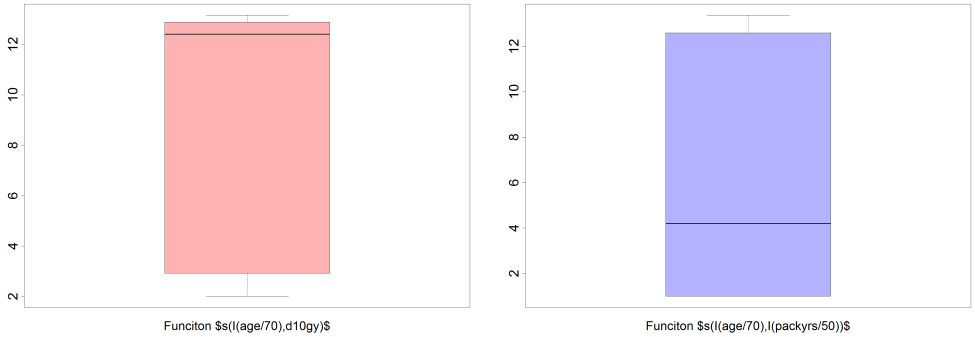
APPENDIX

F

GENERALIZED ADDITIVE MODELS FOR LUNG ADENOCARCINOMA IN THE LIFE SPAN STUDY COHORT: SUPPLEMENTARY RESULTS

This Appendix Chapter contains supplementary information related to Chapter 4.3.

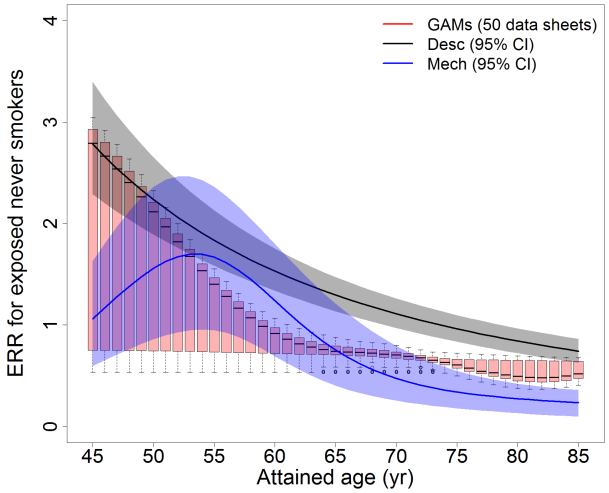
F.1 Variability of estimated degrees of freedom



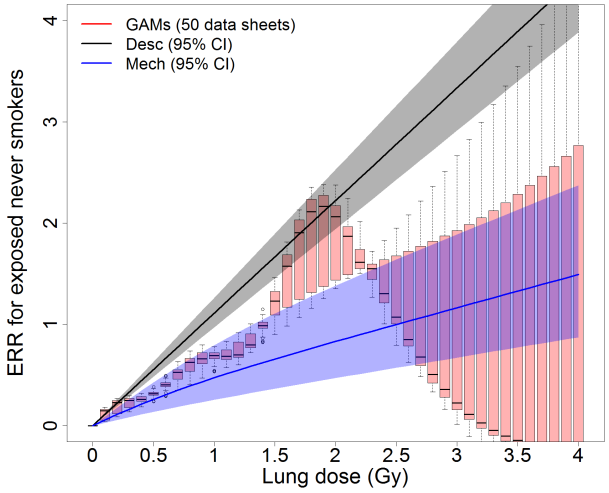
(a) (b)

Figure F.1: Results of GAM_{LSS}^{LADC}. *Boxplot of the values of estimated degree of freedom for (a) function* $s(I(age/70), d10gy)$ *and (b) function* $s(I(age/70), I(packyrs/50))$. *The edf parameter are the only parameters with a high variability. This fact has to be attributed to the variation of the multiple imputation process and to the flexibility of splines, that adapt to different data sets.*

F.2 ERRs of all three models

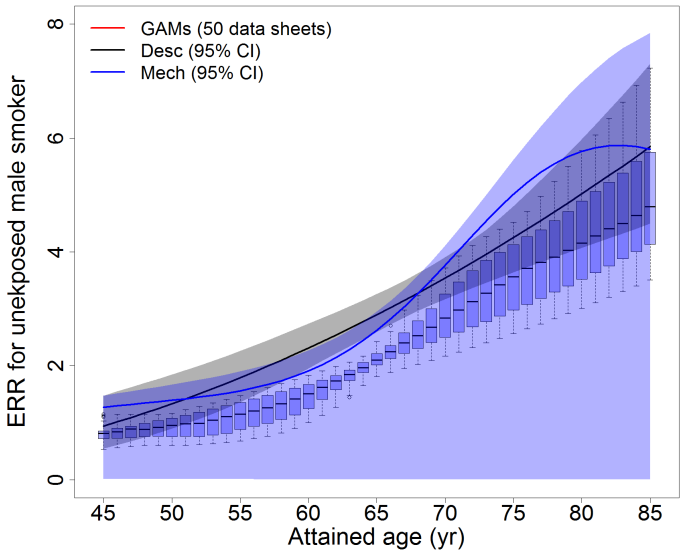


(a) ERR from GAM_{LSS}^{LADC} for radiation induced LADC as a function of attained age (yr)

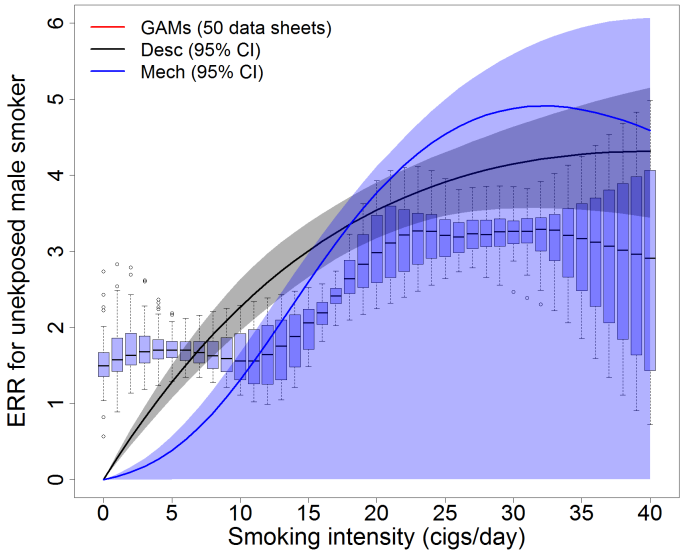


(b) ERR from GAM_{LSS}^{LADC} for smoking induced LADC as a function of lung dose (Gy)

Figure F.2: Excess relative risks (ERR) from GAM_{LSS}^{LADC} for radiation-induced LADC for (a) a person exposed to 1 Gy lung dose at age 30 yr as a function of age, (b) a 70 years old person exposed at age 30 yr as a function of lung dose (Gy). The boxplots represent the variance of the 50 datasets by GAM_{LSS}^{LADC}. $Stat_{LSS}^{LADC}$ is presented in black, while M_3^{LADC} in blue for comparison.



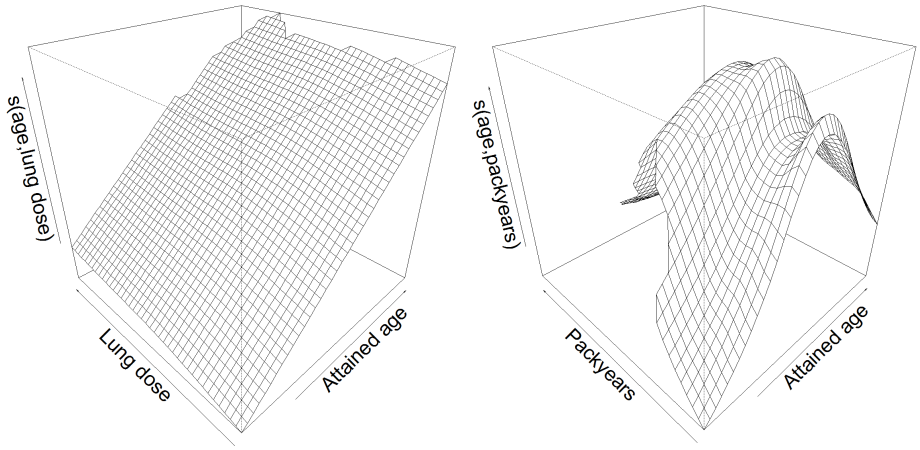
(a) ERR from GAM_{LSS}^{LADC} for smoking induced LADC as a function of attained age (yr)



(b) ERR from GAM_{LSS}^{LADC} for smoking induced LADC as a function of smoking intensity (cigs/day)

Figure F.3: Excess relative risks (ERR) from GAM_{LSS}^{LADC} for smoking-induced LADC for a current smoker that began at age 20 yr, never stopped (a) with smoking amount of 1 packyear as a function of attained age, (b) with varying smoking intensity. The boxplots represent the variance of the 50 datasets by GAM_{LSS}^{LADC}. $Stat_{LSS}^{LADC}$ is presented in black, while M_3^{LADC} in blue for comparison.

F.3 Spline functions



(a) (b)

Figure F.4: Graphical evaluation of some results of GAM_{LSS}^{LADC} *as 3D plots for the imputed dataset 39. (a) Function of age and lung dose* $s(I(age/70), d10gy)$*. (b) Function of age and smoking amount (packyears)* $s(I(age/70), I(packyrs/50))$*.*

APPENDIX

G

SQUAMOUS CELL CARCINOMA IN THE LIFE SPAN STUDY COHORT: SUPPLEMENTARY INFORMATION

This Appendix Chapter contains supplementary information related to Chapter 5.

G.1 Simple additive model

Simple additive model of Egawa et al. [18].

$$h^{SQUAM_{LSS}} = h_0^{SQUAM_{LSS}} \cdot (1 + ERR_{smk}^{SQUAM_{LSS}} + ERR_{rad}^{SQUAM_{LSS}}) \tag{G.1}$$

with

$$\begin{aligned} h_0^{SQUAM_{LSS}} &= e^{\beta_{0F/1M}+\beta_2(city-1)+\beta_3\frac{agexp-30}{10}+\beta_{4F/5M}log(\frac{age}{70})+\beta_{6F/7M}log^2(\frac{age}{70})} \\ ERR_{smk}^{SQUAM_{LSS}} &= \beta_8\frac{packyrs}{50} \cdot e^{\beta_9\frac{smokedYears}{50}+\beta_{10}(\frac{smokedyears}{50})^2+\beta 11log(yearsQuitting+1)} \\ ERR_{rad}^{SQUAM_{LSS}} &= \beta_{12} \cdot Gamma \cdot e^{\beta_{13}\frac{agexp-30}{10}+\beta_{14}log(\frac{age}{70})} \cdot (1 + msex) \end{aligned}$$

G.2 Development of the state-of-the-art statistical risk model

$$h^{SQUAM_{LSS}} = h_0^{SQUAM_{LSS}} \cdot (1 + ERR_{smk}^{SQUAM_{LSS}} + ERR_{rad}^{SQUAM_{LSS}}) \tag{G.2}$$

with

$$\begin{aligned} h_0^{SQUAM_{LSS}} &= e^{\beta_{0F/1M}+\beta_2(city-1)+\beta_3\frac{agexp-30}{10}+\beta_{4F/5M}log(\frac{age}{70})+\beta_{6F/7M}log^2(\frac{age}{70})} \\ ERR_{smk}^{SQUAM_{LSS}} &= \beta_8\frac{packyrs}{50} \cdot e^{\beta_9\frac{smokedYears}{50}+\beta_{10}(\frac{smokedyears}{50})^2+\beta 11log(yearsQuitting+1)} \\ ERR_{rad}^{SQUAM_{LSS}} &= \beta_{12} \cdot Gamma \end{aligned}$$

Table G.1: Interesting parameters of some models fitted to one imputed dataset for the derivation of the best state-of-the-art statistical risk model (of the form of a (general) additive model) for SQUAM in the LSS cohort. Only smoking modifiers were tested since already the radiation ERR was not significant. The imputed data set $pydat - smk - imp - C23$ was used.

Derivation of the state-of-the-art stat. risk model							
ERR γ	SE	ERR smk.	SE	Risk modif.	Value	SE	Dev.
							3120.9
		17.81	4.26				2913.5
0.51	0.26						3114.8
0.23	0.91	17.78	4.29				2913.4
0.13	0.85	20.19	4.92	*smk. dur. lin.*	1.42	0.51	2901.5
0.08	0.85	20.92	5.14	*smk. dur. lin.*	1.56	0.50	2900.1
				smk. dur. quad.	0.37	0.10	
0.15	0.89	20.44	5.00	*smk. dur. lin.*	0.63	0.55	2895.3
				smk. dur. quad.	0.26	0.11	
				years quitting	-0.22	0.10	

G.3 Deviance comparison

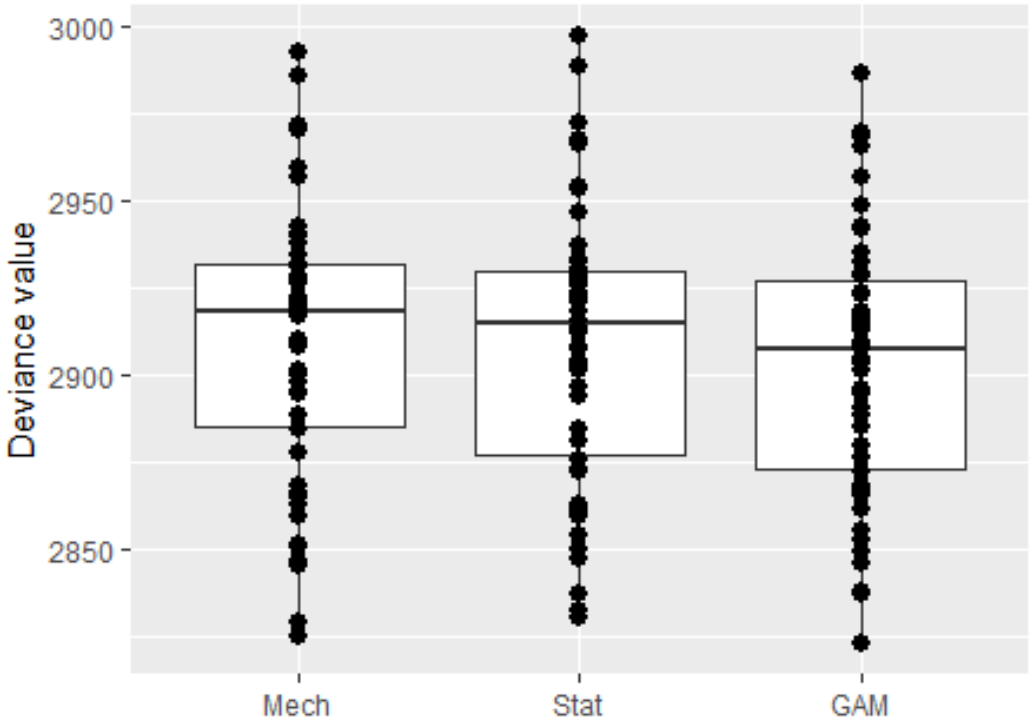
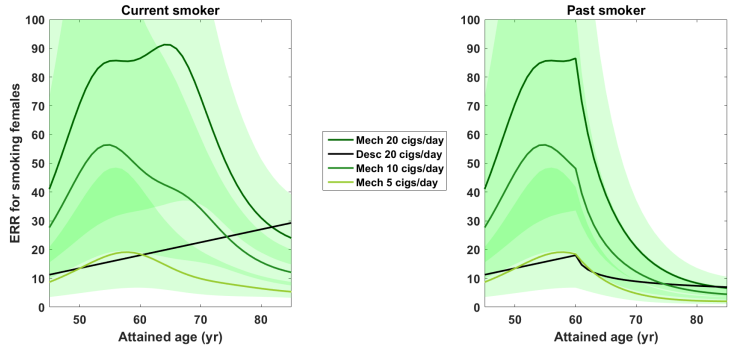
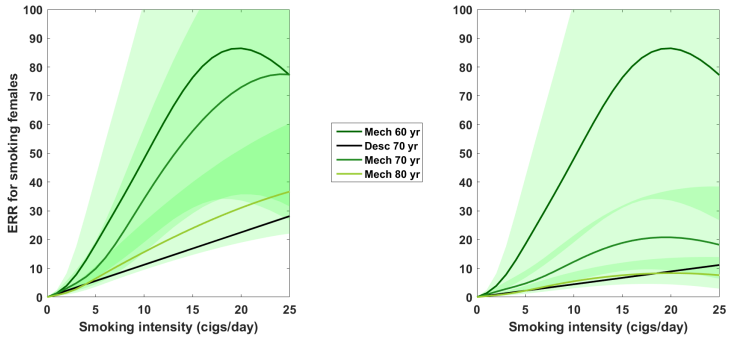


Figure G.1: Comparison of the deviance of models $Stat_{LSSC}^{SQUAM}$, M_2^{SQUAM} *and* GAM_{LSS}^{SQUAM}. *Boxplots represent the variation between the 50 imputed data sets. Mean values for mechanistic, state-of-the-art descriptive and GA models, respectively: 2909.563, 2907.926 and 2902.386.* GAM_{LSS}^{SQUAM} *gives a slightly better fit of the data compared to the other two modalities. No outliers are presented in the model. The difference in the deviance between the data sets in one model has to be attributed to the fact that with imputation each data set has a different amount of strata.*

G.4 Excess relative risks for smoking

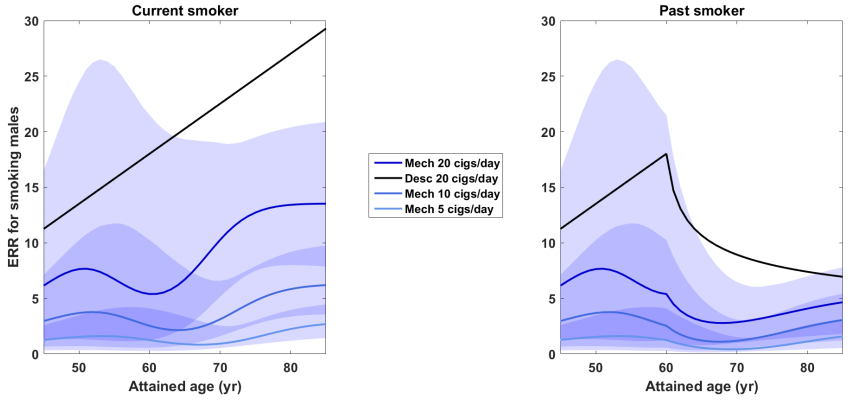


(a)

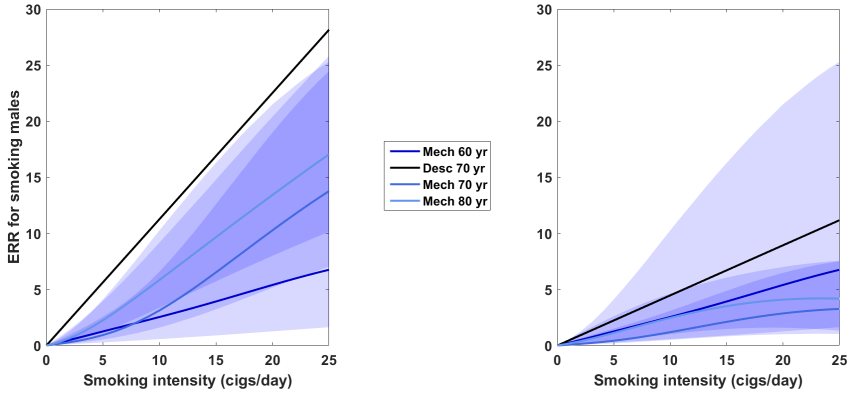


(b)

Figure G.2: ERRs from M_2^{SQUAM} (Mech) for smoking-induced SQUAM for lifelong and past female smokers (left and right panels, respectively) starting at age 20 yr and for past smokers at quitting age 60 yr. (a) ERRs as a function of age attained for different smoking intensities: 5, 10 and 20 cigs/day. For both lifelong and past smokers the ERR increases with smoking intensity. For past smokers a kink at age of quit smoking is visible, the ERR decreases hence exponentially. Past smokers have slightly lower risks as lifelong smokers. The undulated behavior is determined by the linear response to smoking intensity of the initiating parameter X. (b) ERRs as a function of smoking intensity (cigs/day) for different ages: 60, 70 and 80 years. For both lifelong and past smokers the ERR increases with age. In this cases past smokers have clearly lower risks as lifelong smokers.



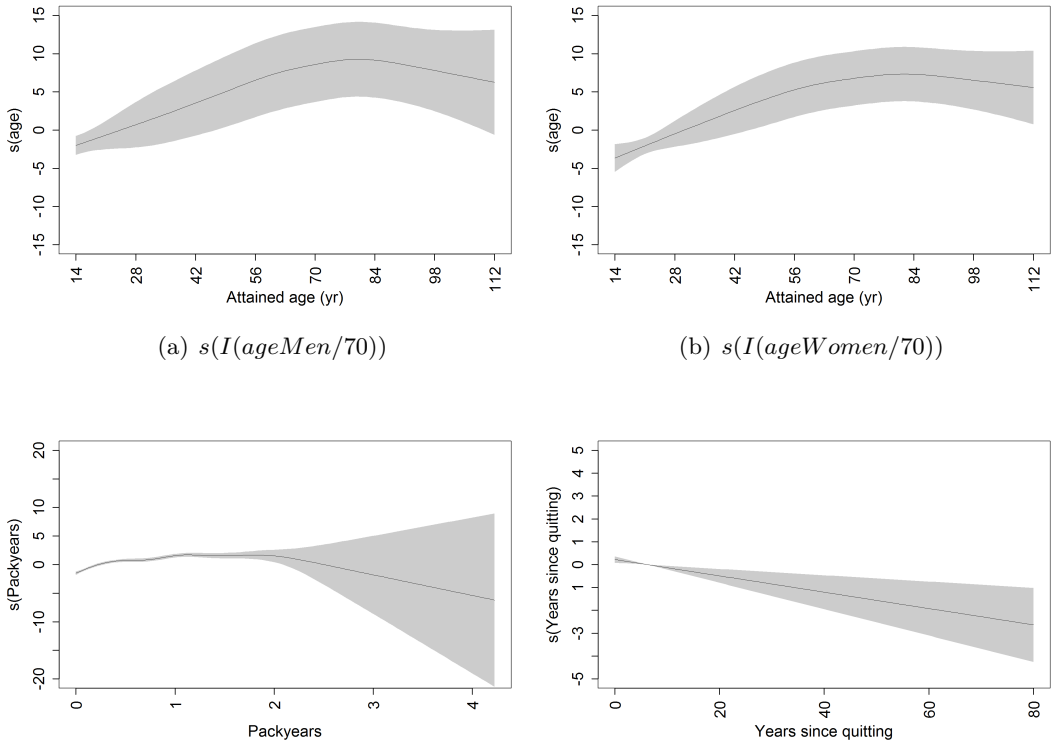
(a)



(b)

Figure G.3: ERRs from M_2^{SQUAM} (Mech) for smoking-induced SQUAM for lifelong and past male smokers (left and right panels, respectively) starting at age 20 yr and for past smokers at quitting age 60 yr. (a) ERRs as a function of age attained for different smoking intensities: 5, 10 and 20 cigs/day. For both lifelong and past smokers the ERR increases with smoking intensity. For past smokers a kink at age of quit smoking is visible, the ERR decreases hence exponentially. Past smokers have slightly lower risks as lifelong smokers. The undulated behavior is determined by the linear response to smoking intensity of the initiating parameter X. (b) ERRs as a function of smoking intensity (cigs/day) for different ages: 60, 70 and 80 years. For both lifelong and past smokers the ERR increases with age. In this cases past smokers have clearly lower risks as lifelong smokers.

G.5 Spline functions



(a) $s(I(ageMen/70))$

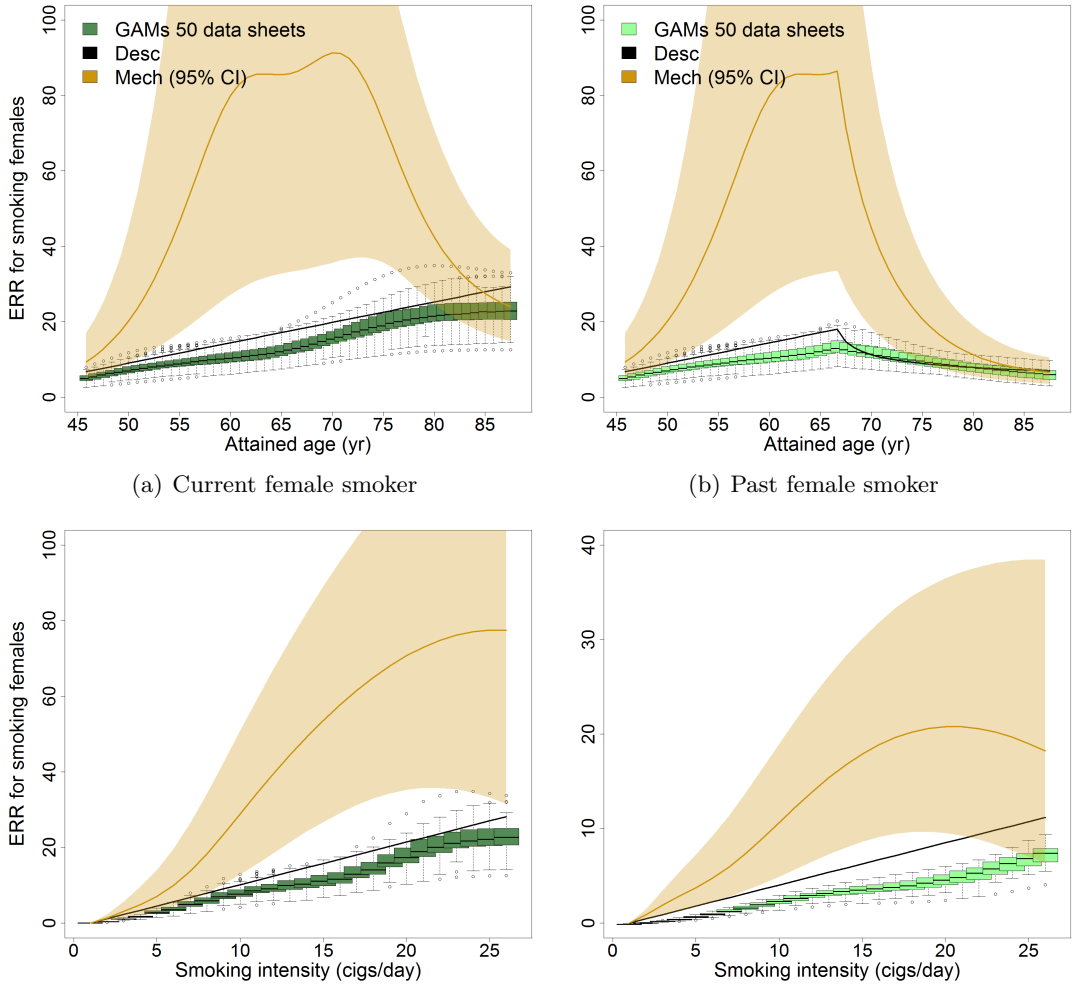
(b) $s(I(ageWomen/70))$

(c) $s(I(packyrs/50))$

(d) $s(smkqyrs)$

Figure G.4: Spline functions fitted for one selected imputed dataset ($pydat - smk - imp - C23$) in model GAM_{LSS}^{SQUAM}. (a) Spline for attained age for men, (b) spline for attained age for women, (c) spline for cumulative smoking amount (pack years) and (d) spline for the years since quitting. Although functions $s(I(ageMen/70))$ and $s(I(ageWomen/70))$ look very similar, the gender differentiation improved the fit of ca. 200 points.

G.6 ERRs of all three models



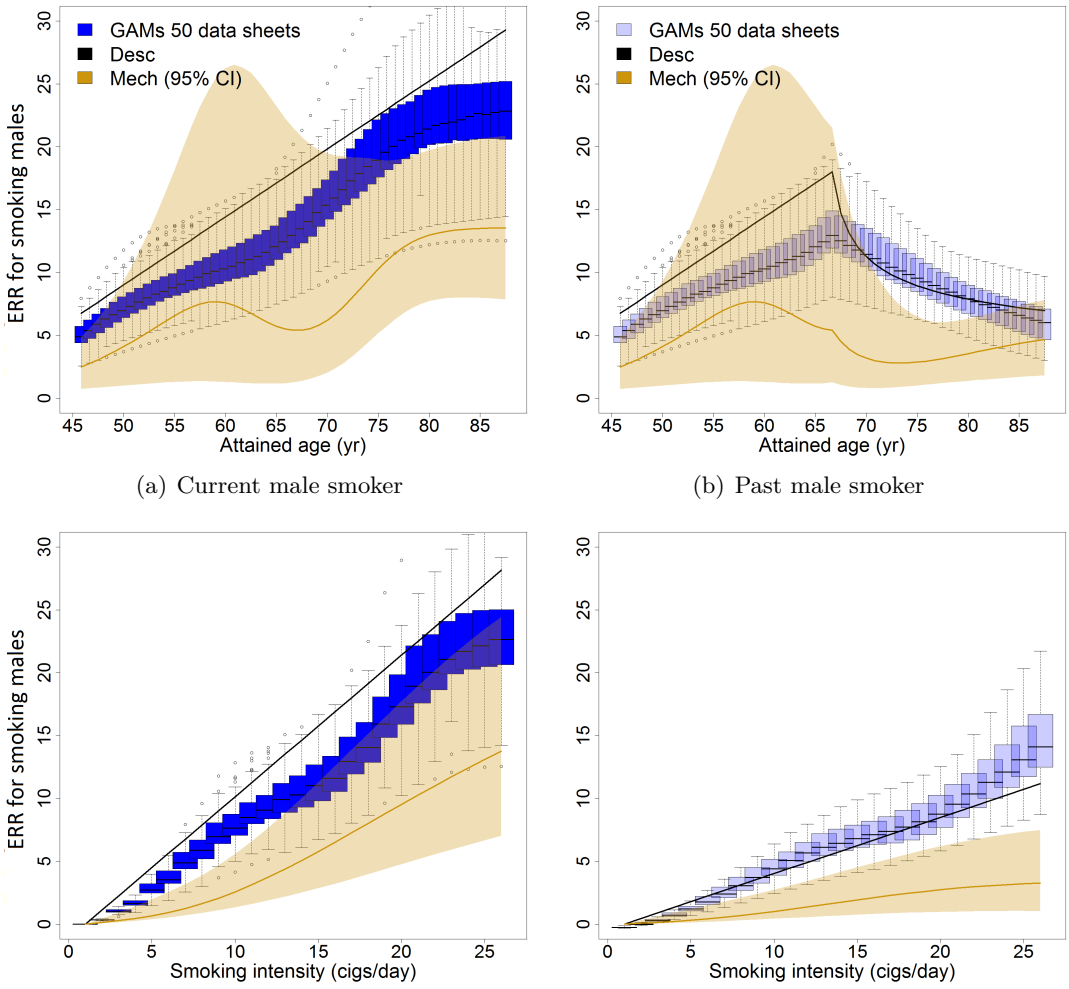
(a) Current female smoker

(b) Past female smoker

(c) Current female smoker

(d) Past female smoker

Figure G.5: ERRs from GAM_{LSS}^{SQUAM} (in green), $Stat_{LSS}^{SQUAM}$ (in black) and M_2^{SUQAM} (in orange) for smoking induced SQUAM for (a) current smoking females as a function of attained age, (b) past smoking females as a function of attained age, (c) current smoking females as a function of smoking intensity and (d) past smoking females as a function of smoking intensity. Age at begin of smoking was taken as 20 yr and age for stop smoking 60 yr. In figures as function of attained age the smoking intensity was of 20 cigs/day. In figures as function of smoking intensity the attained age was of 70 years. The boxplots represent the variance of the 50 data sheets.



(a) Current male smoker

(b) Past male smoker

(c) Current male smoker

(d) Past male smoker

Figure G.6: ERRs from GAM_{LSS}^{SQUAM} (in green), $Stat_{LSS}^{SQUAM}$ (in black) and M_2^{SUQAM} (in orange) for smoking induced SQUAM for (a) current smoking male as a function of attained age, (b) past smoking male as a function of attained age, (c) current smoking male as a function of smoking intensity and (d) past smoking male as a function of smoking intensity. Age at begin of smoking was taken as 20 yr and age for stop smoking 60 yr. In figures as function of attained age the smoking intensity was of 20 cigs/day. In figures as function of smoking intensity the attained age was of 70 years. The boxplots represent the variance of the 50 data sheets.

APPENDIX

H

ADENOCARCINOMA IN THE ELDORADO COHORT: SUPPLEMENTARY INFORMATION

This Appendix Chapter contains supplementary information related to Chapter 6.

H.1 Development of the state-of-the-art statistical risk model

$$h^{LADC_{ELDO}} = h_0^{LADC_{ELDO}} \cdot (1 + ERR_\gamma^{LADC_{ELDO}} + ERR_\alpha^{LADC_{ELDO}}) \tag{H.1}$$

with

$$h_0^{LADC_{ELDO}} = e^{\beta_0 + \beta_1 log(\frac{age}{60}) + \beta_2 log^2(\frac{age}{60}) + \beta_3 \cdot calendaryear} \tag{H.2}$$

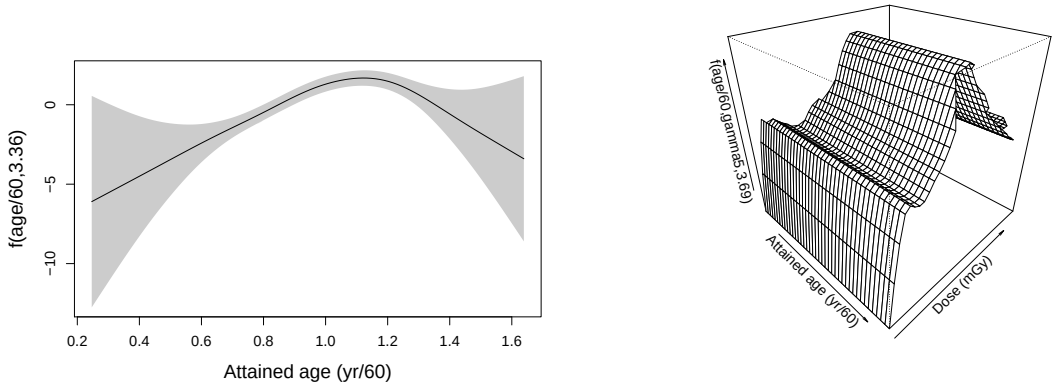
$$ERR_\gamma^{LADC_{ELDO}} = \beta_4 \cdot Gammae^{\beta_5 log(\frac{age}{60}) + \beta_6 \cdot dose\ rate + \beta_7 \cdot time\ since\ last\ exp} \tag{H.3}$$

$$ERR_\alpha^{LADC_{ELDO}} = \beta_8 \cdot Radon. \tag{H.4}$$

Table H.1: Interesting parameters of some models fitted for the derivation of the best state-of-the-art statistical risk model model for LADC in the Eldorado cohort.

Derivation of the state-of-the-art stat. risk model							
ERR γ (10^{-2})	SE (10^{-2})	ERR radon	SE	Risk modif.	Value	SE	Dev.
							1490.8
		0.20	0.12				1484.4
0.77	**0.36**						**1477.3**
0.66	0.34	0.17	0.13				1474.8
0.89	0.43			*age*	-2.0782	3.1645	1476.8
0.81	0.44			*dose rate*	-0.0007	0.0039	1477.2
0.15	0.13			*agelast*	0.06	0.02	1474.7

H.2 Spline functions

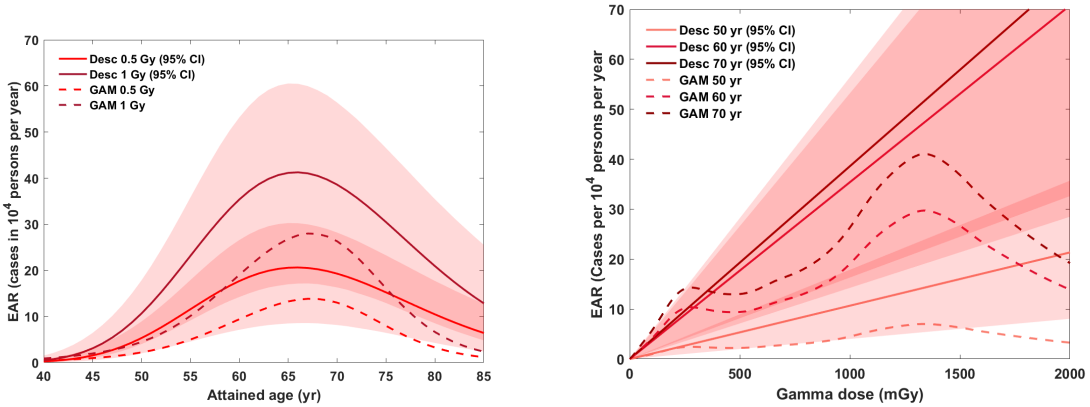


(a) Attained age spline

(b) Attained age-gamma spline

Figure H.1: Fitted spline of the GAM model for attained age (yr) (a) and for the interaction attained age-gamma exposure (Gy) (b).

H.3 Excess absolute rates



(a) EAR as a function of attained age

(b) EAR as a function of gamma dose (Gy)

Figure H.2: EAR for γ exposure (Gy) (cases in 10^4 persons per year) for LADC in the Eldorado cohort. Statistical models are presented as solid lines and GAM models as dashed lines. (a) The radiation EAR as a function of attained age is linear with age. (b) The EAR as a function of gamma dose (Gy) is independent of age.

APPENDIX

I

SQUAMOUS CELL CARCINOMA IN THE ELDORADO COHORT: SUPPLEMENTARY INFORMATION

This Appendix Chapter contains supplementary information related to Chapter 7.

I.1 Development of the state-of-the-art statistical risk model

$$h^{SQUAM_{ELDO}} = h_0^{SQUAM_{ELDO}} \cdot (1 + ERR_\gamma^{SQUAM_{ELDO}} + ERR_\alpha^{SQUAM_{ELDO}}) \quad \text{(I.1)}$$

with

$$h_0^{SQUAM_{ELDO}} = e^{\beta_0 + \beta_1 log(\frac{age}{60}) + \beta_2 log^2(\frac{age}{60}) + \beta_3 \cdot calendaryear_{group}} \quad \text{(I.2)}$$

$$+ e^{\beta_4 \cdot calendaryear_5 + \beta_5 \cdot (timesinceexposure) + \beta_6 \cdot (timesinceexposure)^2} \quad \text{(I.3)}$$

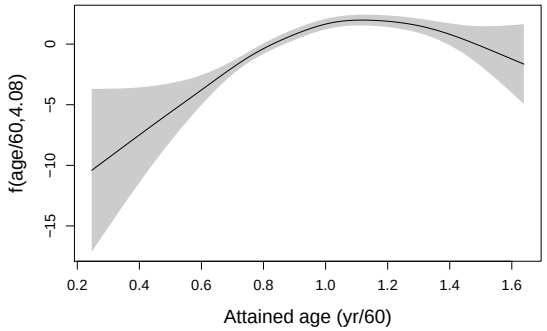
$$ERR_\alpha^{SQUAM_{ELDO}} = \beta_7 \cdot Radon e^{\beta_8 log(\frac{age}{60}) + \beta_9 \cdot dose\ rate + \beta_1 0 \cdot time\ since\ last\ exp}$$

$$ERR_\gamma^{SQUAM_{ELDO}} = \beta_1 1 \cdot Gamma. \quad \text{(I.4)}$$

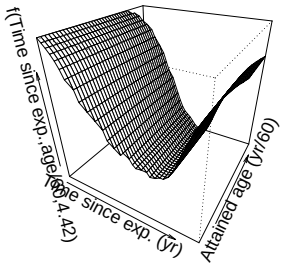
Table I.1: Interesting parameters of some models fitted for the derivation of the best state-of-the-art statistical risk model for SQUAM in the Eldorado cohort.

Derivation of the stat. model for SQUAM in the Eldorado cohort							
ERR γ (10^{-5})	SE (10^{-5})	ERR radon (10^{-2})	SE (10^{-2})	Risk modif.	Value (10^{-2})	SE	Dev.
							2570.4
-8.86	49.58						2570.4
		0.43	0.12				2524.4
-90.36	21.37	0.44	0.12				2521.6
		0.43	0.15	*age*	-2.80 10^{-09}	1.88	2524.4
		0.47	0.16	*dose rate*	-0.26 10^{-3}	0.71 10^{-3}	2523.9
		1.85	**1.00**	***agelast***	**-0.51** 10^{-1}	**0.19** 10^{-1}	**2517.1**

I.2 Spline functions



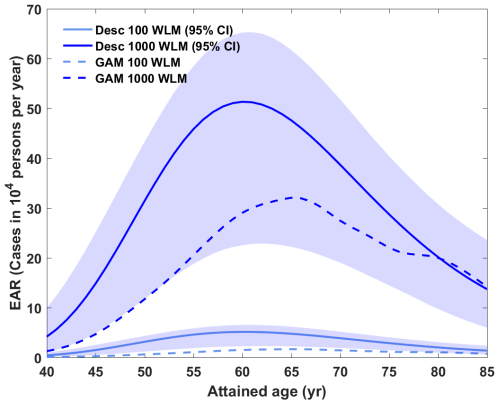
(a) Attained age spline



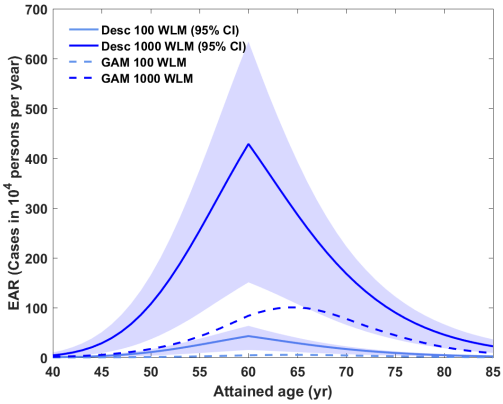
(b) Attained age-gamma spline

Figure I.1: Fitted spline of the GAM model for attained age (yr) (a) and for the interaction attained age-time since last exposure (b).

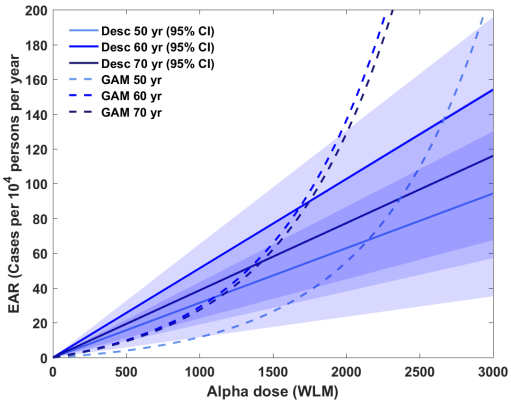
I.3 Excess absolute rates



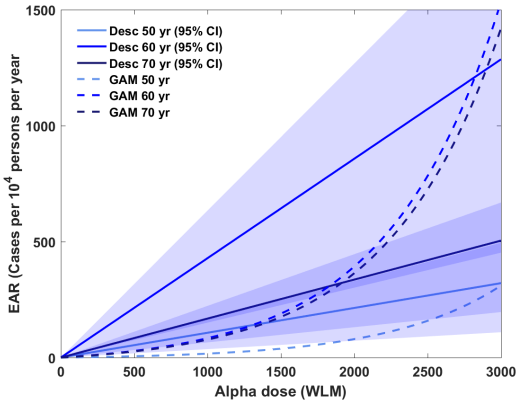
(a) EAR as a function of attained age for age quitting 40 yr



(b) EAR as a function of attained age for age quitting 60 yr



(c) EAR as a function of radon dose (WLM) for age quitting 40 yr



(d) EAR as a function of radon dose (WLM) for age quitting 60 yr

Figure I.2: EAR for radon exposure (WLM) (cases in 14^3 persons per year) for SQUAM in the Eldorado cohort. Statistical models are presented as solid lines and GAM models as dashed lines. (a) Radiation EAR as a function of attained age is linear with age. (b) The EAR as a function of gamma dose (Gy) is independent of age.

BIBLIOGRAPHY

[1] WHO 2018, July 2018. URL http://www.who.int/ionizing_radiation/about/what_is_ir/en/.

[2] A. J. Alberg, M. V. Brock, J. G. Ford, J. M. Samet, and S. D. Spivack. 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–29S, 2013. ISSN 1931-3543 (Electronic) 0012-3692 (Linking). doi: 10.1378/chest.12-2345. URL https://www.ncbi.nlm.nih.gov/pubmed/23649439.

[3] L.B. Alexandrov, S. Nik-Zainal, D.C. Wedge, S.A. Aparicio, S. Behjati, A.V. Biankin, G.R. Bignell, N. Bolli, A. Borg, A.L. Borresen-Dale, S. Boyault, B. Burkhardt, A.P. Butler, C. Caldas, H.R. Davies, C. Desmedt, R. Eils, J.E. Eyfjord, J.A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jager, D.T. Jones, D. Jones, S. Knappskog, M. Kool, S.R. Lakhani, C. Lopez-Otin, S. Martin, N.C. Munshi, H. Nakamura, P.A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J.V. Pearson, X.S. Puente, K. Raine, M. Ramakrishna, A.L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T.N. Schumacher, P.N. Span, J.W. Teague, Y. Totoki, A.N. Tutt, R. Valdes-Mas, M.M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L.R. Yates, Australian Pancreatic Cancer Genome I, Consortium IBC, Consortium IM-S, PedBrain I, J. Zucman-Rossi, P.A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S.M. Grimmond, R. Siebert, E. Campo, T. Shibata, S.M. Pfister, P.J. Campbell, and M.R. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500:415–421., 2013.

[4] E L. Alpen. *Radiation Biophysics*. Academic Press, 2014.

[5] J.M. Barros-Dios, A. Ruano-Ravina, M. PÃ©rez-RÃos, M. Castro-BernÃ¡rdez, J. Abal Arca, and M. Tojo-Castro. Residential radon exposure, histological types and lung cancer risk. a case-control study in galicia, spain. *Cancer Epidemiology, Biomarkers and Prevention*, 21(6):951–8, April 2012. doi: 10.1158/1055-9965.EPI-12-0146-T.

[6] S. Behjati, G. Gundem, D.C. Wedge, N.D. Roberts, P.S. Tarpey, S.L. Cooke, P. Van Loo, L.B. Alexandrov, M. Ramakrishna, H. Davies, S. Nik-Zainal, C. Hardy, C. Latimer, K.M.

Raine, L. Stebbings, A. Menzies, D. Jones, R. Shepherd, A.P. Butler, J.W. Teague, M. Jorgensen, B. Khatri, N. Pillay, A. Shlien, P.A. Futreal, C. Badie, Group IP, U. McDermott, Bova. G.S., A.L. Richardson, A.M. Flanagan, M.R. Stratton, and P.J. Campbell. Mutational signatures of ionizing radiation in second malignancies. *Nature Communications*, 7: 12605, 2016.

[7] F. I. Bray and E. Weiderpass. Lung cancer mortality trends in 36 european countries: secular trends and birth cohort patterns by sex and region 1970-2007. *International Journal of Cancer*, 126:1454–1466, 2010.

[8] E. K. Cahoon, D. L. Preston, D. A. Pierce, E. Grant, A. V. Brenner, K. Mabuchi, M. Utada, and K. Ozasa. Lung, laryngeal and other respiratory cancer incidence among japanese atomic bomb survivors: An updated analysis from 1958 through 2009. *Radiat Res*, 187 (5):538–548, 2017. ISSN 1938-5404 (Electronic) 0033-7587 (Linking). doi: 10.1667/RR14583.1. URL https://www.ncbi.nlm.nih.gov/pubmed/28323575.

[9] J. Campbell, A. Alexandrov, J. Kim, J. Wala, A.H. Berger, C.S. Pedamallu, S.A. Shukla, G. Guo, A.N. Brooks, B.A. Murray, M. Imielinski, X Hu, S. Ling, R. Akbani, M. Rosenberg, C. Cibulskis, A. Ramachandran, E.A. Collisson, D.J. Kwiatkowski, M.S. Lawrence, J.N. Weinstein, R.G. Verhaak, C.J. Wu, Hammerman P.S., A.D. Cherniack, G. Getz, Cancer Genome Atlas Research N, M.N. Artyomov, R. Schreiber, R. Govindan, and M. Meyerson. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat Genet*, 48:607–616, 2016.

[10] Network Cancer Genome Atlas Research. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–50, 2014. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature13385. URL https://www.ncbi.nlm.nih.gov/pubmed/25079552.

[11] J.R. Choi, S.-B. Koh, S.Y. Park, H.R. Kim, H. Lee, and D.R. Kang. Novel genetic associations between lung cancer and indoor radon exposure. *Journal of cancer prevention*, 22 (4):234–40, December 2017. doi: 10.15430/JCP.2017.22.4.234.

[12] M.S. Clements, B.K. Armstrong, and S.H. MOOLGAVKAR. Lung cancer rate predictions using generalized additive models. *Biostatistics*, 6(4):576–589, April 2005. doi: 10.1093/biostatistics/kxi028.

[13] B.S. Cohen, M. Eisenbud, and N.H. Harley. Alpha radioactivity in cigarette smoke. *Radiation Research*, 83(1):190–6, July 1980.

[14] L. CrinÃ^3, W. Weder, J. van Meerbeeck, E. Felip, and On behalf of the ESMO Guidelines Working Group. Early stage and locally advanced (non-metastatic) non-small-cell lung cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 21(5):103–115, May 2010. doi: doi:10.1093/annonc/mdq207. URL https://doi.org/10.1093/annonc/mdq207.

[15] H.M. Cullings, S. Fujita, S. Funamoto, E.J. Grant, G.D. Kerr, and D.L. Preston. Dose estimation for atomic bomb survivor studies: Its evolution and present status. *Radiation Research*, 166(1):219–254, 2006.

[16] S. Dogan, R. Shen, D. C. Ang, M. L. Johnson, S. P. D'Angelo, P. K. Paik, E. B. Brzostowski, G. J. Riely, M. G. Kris, M. F. Zakowski, and M. Ladanyi. 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(22):6169–77, 2012. ISSN 1078-0432 (Print) 1078-0432 (Linking). doi: 10.1158/1078-0432.CCR-11-3265. URL https://www.ncbi.nlm.nih.gov/pubmed/23014527.

[17] A.A. Efanov, A.V. Brenner, T.I. Bogdanova, L.M. Kelly, P. Liu, M.P. Little, A.I. Wald, M. Hatch, L.Y. Zurnadzy, M.N. Nikiforova, V. Drozdovitch, R. Leeman-Neill, K. Mabuchi, Tronko M.D., S.J. Chanock, and Y.E. Nikiforov. Investigation of the relationship between radiation dose and gene mutations and fusions in post-chernobyl thyroid cancer. *J Natl Cancer Inst*, 110:371–378, 2018.

[18] H. Egawa, K. Furukawa, D. Preston, S. Funamoto, S. Yonehara, T. Matsuo, S. Tokuoka, A. Suyama, K. Ozasa, K. Kodama, and K. Mabuchi. Radiation and smoking effects on lung cancer incidence by histological types among atomic bomb survivors. *Radiat Res*, 178 (3):191–201, 2012. ISSN 1938-5404 (Electronic). URL https://www.ncbi.nlm.nih.gov/pubmed/22862780.

[19] R. El-Zein, M. Schabath, C. Etzel, M. Lopez, J. Franklin, and M. Spitz. Cytokinesis-blocked micronucleus assay as a novel biomarker for lung cancer risk. *Cancer Research*, 66:6449–6456., 2006.

[20] L. Fahrmeir, T. Kneib, S. Lang, and Marx B. *Regression - Models, Methods, Applications*. Springer, 2013.

[21] K. Furukawa, D. L. Preston, S. Lonn, S. Funamoto, S. Yonehara, T. Matsuo, H. Egawa, S. Tokuoka, K. Ozasa, F. Kasagi, K. Kodama, and K. Mabuchi. Radiation and smoking effects on lung cancer incidence among atomic bomb survivors. *Radiat Res*, 174(1):72–82, 2010. ISSN 1938-5404 (Electronic) 0033-7587 (Linking). doi: 10.1667/RR2083.1. URL https://www.ncbi.nlm.nih.gov/pubmed/20681801.

[22] K. Furukawa, D. L. Preston, M. Misumi, and H. M. Cullings. Handling incomplete smoking history data in survival analysis. *Stat Methods Med Res*, 2014. ISSN 1477-0334 (Electronic) 0962-2802 (Linking). doi: 10.1177/0962280214556794. URL https://www.ncbi.nlm.nih.gov/pubmed/25348676.

[23] H. M. Gail and Benichou J. *Encyclopedia of Epidemiologic Methods*. Wiley, 2000.

[24] I. Gipanou. Personal correspondence.

[25] Collaboration Global Burden of Disease Cancer, C. Fitzmaurice, C. Allen, R. M. Barber, L. Barregard, Z. A. Bhutta, H. Brenner, D. J. Dicker, O. Chimed-Orchir, R. Dandona, L. Dandona, T. Fleming, M. H. Forouzanfar, J. Hancock, R. J. Hay, R. Hunter-Merrill, C. Huynh, H. D. Hosgood, C. O. Johnson, J. B. Jonas, J. Khubchandani, G. A. Kumar, M. Kutz, Q. Lan, H. J. Larson, X. Liang, S. S. Lim, A. D. Lopez, M. F. MacIntyre, L. Marczak, N. Marquez, A. H. Mokdad, C. Pinho, F. Pourmalek, J. A. Salomon, J. R. Sanabria, L. Sandar, B. Sartorius, S. M. Schwartz, K. A. Shackelford, K. Shibuya, J. Stanaway, C. Steiner, J. Sun, K. Takahashi, S. E. Vollset, T. Vos, J. A. Wagner, H. Wang, R. Westerman, H. Zeeb, L. Zoeckler, F. Abd-Allah, M. B. Ahmed, S. Alabed, N. K. Alam, S. F. Aldhahri, G. Alem, M. A. Alemayohu, R. Ali, R. Al-Raddadi, A. Amare, Y. Amoako, A. Artaman, H. Asayesh, N. Atnafu, A. Awasthi, H. B. Saleem, A. Barac, N. Bedi, I. Bensenor, A. Berhane, E. Bernabe, B. Betsu, A. Binagwaho, D. Boneya, I. Campos-Nonato, C. Castaneda-Orjuela, F. Catala-Lopez, P. Chiang, C. Chibueze, A. Chitheer, J. Y. Choi, B. Cowie, S. Damtew, J. das Neves, S. Dey, S. Dharmaratne, P. Dhillon,

E. Ding, T. Driscoll, D. Ekwueme, A. Y. Endries, M. Farvid, F. Farzadfar, J. Fernandes, F. Fischer, G. Hiwot TT, A. Gebru, S. Gopalani, et al. 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(4):524–548, 2017. ISSN 2374-2445 (Electronic) 2374-2437 (Linking). doi: 10.1001/jamaoncol.2016.5688. URL https://www.ncbi.nlm.nih.gov/pubmed/27918777.

[26] E. J. Grant, A. Brenner, H. Sugiyama, R. Sakata, A. Sadakane, M. Utada, E. K. Cahoon, C. M. Milder, M. Soda, H. M. Cullings, D. L. Preston, K. Mabuchi, and K. Ozasa. Solid cancer incidence among the life span study of atomic bomb survivors: 1958-2009. *Radiat Res*, 187(5):513–537, 2017. ISSN 1938-5404 (Electronic) 0033-7587 (Linking). doi: 10.1667/RR14492.1. URL https://www.ncbi.nlm.nih.gov/pubmed/28319463.

[27] E. Hall. *Radiobiology for the Radiologist*. Philadelphia„ 2001.

[28] H.S. Karagueuzian, C. White, J. Sayre, and A. Norman. Cigarette smoke radioactivity and lung cancer risk. *Nicotine and Tobacco Research*, 14(1):79–90, January 2012.

[29] S. Kevork. Separable temporal exponential random graph models. Master's thesis, Ludwig-Maximilians University Munich, Department of Statistics, 2017.

[30] M. A. Khamsi. Riccati equations, 2018. URL http://www.sosmath.com/diffeq/first/riccati/riccati.html.

[31] M. Kreuzer, K.M. Mueller, A. Brachner, M. Gerken, B. Grosche, T. Wiethege, and H.-E. Wichmann. Histopathologic findings of lung carcinoma in german uranium miners. *American Cancer Society*, 89(12):2613–21, December 2000.

[32] P. Kundrat. Personal correspondence.

[33] R. S. Lane, S. E. Frost, G. R. Howe, and L. B. Zablotska. Mortality (1950-1999) and cancer incidence (1969-1999) in the cohort of eldorado uranium workers. *Radiat Research*, 174(6):773–85, 2010. ISSN 1938-5404 (Electronic); 0033-7587 (Linking). doi: 10.1667/RR2237.1. URL https://www.ncbi.nlm.nih.gov/pubmed/21128801.

[34] P.M. Lantz, D. Mendez, and M.A. Philbert. Radon, smoking, and lung cancer: The need to refocus radon control policy. *American Journal of Public Health*, 103(3):443–447., March 2013. doi: 10.2105/AJPH.2012.300926.

[35] J. Lee, V. Taneja, and R. Vassallo. Cigarette smoking and inflammation. cellular and molecular mechanisms. *Journal of Dental Ressearch*, 91(2):142–149, February 2012. doi: 10.1177/0022034511421200.

[36] Julie Levandosky. Lecture Notes for Partial Differential Equations of Applied Mathematics. http://web.stanford.edu/class/math220a/handouts/firstorder.pdf, 2002. (last accessed: 30-08-2016).

[37] M. E. Lomax, Folkes L. K., and P. O'Neill. Biological consequences of radiation-induced dna damage: Relevance to radiotherapy. *Clinical Oncology*, 25(10):578–585, October 2013. URL https://doi.org/10.1016/j.clon.2013.06.007.

[38] J. H. Lubin and N. E. Caporaso. Cigarette smoking and lung cancer: modeling total exposure and intensity. *Cancer Epidemiol Biomarkers Prev*, 15(3):517–23, 2006. ISSN 1055-9965 (Print) 1055-9965 (Linking). doi: 10.1158/1055-9965.EPI-05-0863. URL https://www.ncbi.nlm.nih.gov/pubmed/16537710.

[39] Foundation LUNGevity. Accessed july 10, 2018, July 2018. URL https://lungevity.org/for-patients-caregivers/lung-cancer-101.

[40] A. Marshall, D. G. Altman, R. L. Holder, and P. Royston. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BioMediacal Central Medical Research Methodology*, 2009. doi: doi:10.1186/1471-2288-9-57.

[41] E.A. Martell. Radioactivity of tobacco trichomes and insoluble cigarette smoke particles. *Nature*, 249:215–7, May 1974.

[42] A. Midha, S. Dearden, and R2. McCormack. Egfr mutation incidence in non-small-cell lung cancer of adenocarcinoma histology: a systematic review and global map by ethnicity (mutmapii). *American Journal of Cancer Research*, 5(9):2892–2911, August 2015. ISSN 2156-6976/ajcr0013308. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633915/pdf/ajcr0005-2892.pdf.

[43] S. H. Moolgavkar and A. G. Knudson. Mutation and cancer: A model for human carcinogenesis. *Journal of the National Cancer Institute*, 66(6):1037–1052, 1981. doi: https://doi.org/10.1093/jnci/66.6.1037.

[44] Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489:519–525, September 2012.

[45] K. Ozasa, Y. Shimizu, A. Suyama, F. Kasagi, M. Soda, E. J. Grant, R. Sakata, H. Sugiyama, and K. Kodama. Studies of the mortality of atomic bomb survivors, report 14, 1950-2003: an overview of cancer and noncancer diseases. *Radiat Res*, 177 (3):229–43, 2012. ISSN 1938-5404 (Electronic) 0033-7587 (Linking). URL https://www.ncbi.nlm.nih.gov/pubmed/22171960.

[46] D. A. Pierce, G. B. Sharp, and K. Mabuchi. Joint effects of radiation and smoking on lung cancer risk among atomic bomb survivors. *Radiat Res*, 159(10):511–20, 2003. ISSN 0033-7587 (Print); 0033-7587 (Linking). URL https://www.ncbi.nlm.nih.gov/pubmed/12643796.

[47] D. Posada and Buckley T.R. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793–808, 2004. ISSN 1063-5157 print / 1076-836X online. doi: 10.1080/10635150490522304.

[48] D. Preston. Modeling radiaiton effects on disease incidence. *Radiation Research*, 124: 336–372, 1990.

[49] M. Reck, D. F. Heigener, T. Mok, J. C. Soria, and K. F. Rabe. Management of non-small-cell lung cancer: recent developments. *Lancet*, 382(513):709–19, 2013. ISSN 1474-547X (Electronic);0140-6736 (Linking). doi: 10.1016/S0140-6736(13)61502-0. URL https://www.ncbi.nlm.nih.gov/pubmed/23972814.

[50] D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys New York*. 2004.

[51] W. Ruhm, M. Eidemuller, and J. C. Kaiser. Biologically-based mechanistic models of radiation-related carcinogenesis applied to epidemiological data. *Int J Radiat Biol*, pages 1–25, 2017. ISSN 1362-3095 (Electronic) 0955-3002 (Linking). doi: 10.1080/09553002.2017.1310405. URL https://www.ncbi.nlm.nih.gov/pubmed/28346027.

[52] Y. Seki, T. Mizukami, and T. Kohno. Molecular process producing oncogene fusion in lung cancer cells by illegitimate repair of dna double-strand breaks. *Biomolecules*, 5(4): 2464–2476, September 2015. doi: 10.3390/biom5042464.

[53] M.R. Spitz, X. Wu, A Wilkinson, and Wei Q. *Cancer of the lung. In: Cancer Epidemiology and Prevention*, volume 638-658. Schottenfeld D., and Fraumeni, J. Jr., 3rd edition, 2006.

[54] G. T. Stathopoulos, T. P. Sherrill, D. S. Cheng, R. M. Scoggins, W. Han, V. V. Polosukhin, L. Connelly, F. E. Yull, B. Fingleton, and T. S. Blackwell. Epithelial nf-kappab activation promotes urethane-induced lung carcinogenesis. *Proc Natl Acad Sci U S A*, 104(519): 18514–9, 2007. ISSN 1091-6490 (Electronic); 0027-8424 (Linking). doi: 10.1073/pnas.0705316104. URL https://www.ncbi.nlm.nih.gov/pubmed/18000061.

[55] J. A. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338:b2393, 2009. ISSN 1756-1833 (Electronic) 0959-535X (Linking). doi: 10.1136/bmj.b2393. URL https://www.ncbi.nlm.nih.gov/pubmed/19564179.

[56] S. Sun, J. H. Schiller, and A. F. Gazdar. Lung cancer in never smokers–a different disease. *Nat Rev Cancer*, 7(10):778–90, 2007. ISSN 1474-1768 (Electronic) 1474-175X (Linking). doi: 10.1038/nrc2190. URL https://www.ncbi.nlm.nih.gov/pubmed/17882278.

[57] K. Takamochi, S. Oh, and K. Suzuki. Differences in egfr and kras mutation spectra in lung adenocarcinoma of never and heavy smokers. *Oncol Lett*, 6(5):1207–1212, 2013. ISSN 1792-1074 (Print) 1792-1074 (Linking). doi: 10.3892/ol.2013.1551. URL https://www.ncbi.nlm.nih.gov/pubmed/24179496.

[58] C. Tomasetti, L. Marchionni, M. A. Nowak, G. Parmigiani, and B. Vogelsteinf. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc Natl Acad Sci U S A*, 112(1):118–123, 2015. doi: 10.1073/pnas.1421839112.

[59] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal. Global cancer statistics, 2012. *CA Cancer J Clin*, 65(2):87–108, 2015. ISSN 1542-4863 (Electronic) 0007-9235 (Linking). doi: 10.3322/caac.21262. URL https://www.ncbi.nlm.nih.gov/pubmed/25651787.

[60] J.E. Turner. *Atoms, Radiation, and Radiation Protection*. Maxwell, R., 1986.

[61] K.H. Vahakangas, R.A. Metcalf, J.A. Welsh, W.P. Bennett, C.C. Harris, J.M. Samet, and D.P. Lane. Mutations of p53 and ras genes in radon-associated lung cancer from uranium miners. *The Lancet*, 339(8793):576–580, March 1992. doi: 10.1016/0140-6736(92)90866-2.

[62] M. Wang, A. M. Kern, M. Hulskotter, P. Greninger, A. Singh, Y. Pan, D. Chowdhury, M. Krause, M. Baumann, C. H. Benes, J. A. Efstathiou, J. Settleman, and H. Willers. Egfr-mediated chromatin condensation protects kras-mutant cancer cells against ionizing radiation. *Cancer Res*, 74(10):2825–34, 2014. ISSN 1538-7445 (Electronic) 0008-5472

(Linking). doi: 10.1158/0008-5472.CAN-13-3157. URL https://www.ncbi.nlm.nih.gov/pubmed/24648348.

[63] WHO. *Pathology & Genetics: Tumours of the Lung, Pleura, Thymus and Heart*. IARC-Press, 2004. ISBN 92 832 2418 3.

[64] N.S. Wood. *Generalized Additive Models*. Chapman & Hall/CRC and Taylor & Francis Group, 2006.

[65] K. Wu, X. Zhang, F. Li, D. Xiao, Y. Hou, S. Zhu, D. Liu, X. Ye, M. Ye, J. Yang, L. Shao, H. Pan, N. Lu, Y. Yu, L. Liu, J. Li, L. Huang, H. Tang, Q. Deng, Y. Zheng, L. Peng, G. Liu, X. Gu, P. He, Y. Gu, W. Lin, H. He, G. Xie, H. Liang, N. An, H. Wang, M. Teixeira, J. Vieira, W. Liang, X. Zhao, Z. Peng, F. Mu, X. Zhang, X. Xu, H. Yang, K. Kristiansen, J. Wang, N. Zhong, J. Wang, Q. Pan-Hammarstrom, and J. He. Frequent alterations in cytoskeleton remodelling genes in primary and metastatic lung adenocarcinomas. *Nat Commun*, 6:10131, 2015. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/ncomms10131. URL https://www.ncbi.nlm.nih.gov/pubmed/26647728.

[66] Y. Yatabe, A. C. Borczuk, and C. A. Powell. Do all lung adenocarcinomas follow a stepwise progression? *Lung Cancer*, 74(1):7–11, 2011. ISSN 1872-8332 (Electronic) 0169-5002 (Linking). doi: 10.1016/j.lungcan.2011.05.021. URL https://www.ncbi.nlm.nih.gov/pubmed/21705107.

[67] I. Zaballa and M. Eidemueller. Mechanistic study on lung cancer mortality after radon exposure in the wismut cohort supports important role of clonal expansion in lung carcinogenesis. *Radiation Environmental Biophisics*, 55(3):299–315, August 2016. doi: 10.1007/s00411-016-0659-0.

[68] L. B Zablotska, R. S. D. Lane, and S. E. Frost. Mortality (1950-1999) and cancer incidence (1969-1999) of workers in the port hope cohort study exposed to a unique combination of radium, uranium and gamma-ray doses. *BMJ Open*, 2013.

LIST OF FIGURES

LIST OF TABLES

EIDESSTATTLICHE VERSICHERUNG

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 25. Oktober 2018

Noemi Castelletti