Aßenmacher, Matthias:

The exposure-lag-response association between occupational radon exposure and lung cancer mortality

Master Thesis, Winter Semester 2016

Faculty of Mathematics, Computer Science and Statistics
UNSPECIFIED
UNSPECIFIED

Ludwig-Maximilians-Universität München

https://doi.org/10.5282/ubm/epub.39110

Master's Thesis

The exposure-lag-response association between occupational radon exposure and lung cancer mortality

Author

Matthias Aßenmacher

Supervisor

Prof. Dr. Helmut Küchenhoff
Department of Statistics

Department of Statistics
Ludwig-Maximilians-Universität München

Munich, 5th of December 2016

Abstract

Firstly, the aim of this master's thesis is to asses the exposure-lag-response relationship of occupational radon exposure and lung cancer mortality, based on data from the German uranium miners cohort (wismut cohort).
Secondly, it compares the obtained results to those which Gasparrini (2014) has obtained by investigating the data of the Colorado Plateau Uranium Miners (CPUM cohort).
In order to do so, the framework of the so-called *Distributed lag non-linear models* (DLNMs) is applied and seven different hypotheses, which are derived from the results of Gasparrini (2014), are tested or qualitatively evaluated. The whole DLNM framework is embedded in a Cox proportional-hazards model and B-Splines are used for the smooth estimation of the exposure-lag-response-relationship.
The first two, and most important, hypotheses are concerned with the overall form of the exposure-response function (linear versus non-linear) as well as with determining the latency period, until the exposure to a certain quantity of radon has a potential impact on the risk of dying from lung cancer. As a result, it can be stated that there seems to exist a latency period of at least two years for radon exposure. The form of the exposure-response relationship is found to be non-linear. These findings support the results of Gasparrini's work, who also used a latency period of two years in his models and reported a non-linearity in the exposure-response association.
The aim of the other five hypotheses is to check, whether the detailed characteristics of the estimated curves from Gasparrini (2014) are reproducible with this data. Although the exact results could not be verified in this case, the overall form and some of the main results can be confirmed with this data.
A third goal is to check whether the results obtained within the DLNM framework are comparable to the newly developed framework of the penalized piecewise exponential additive models by Bender et al. (2016). Some conformity of the results within these different frameworks is found, but some grave differences are reported as well.

Contents

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike information criterion
BfS	Federal Office for Radiation Protection
BIC	Bayesian Information Criterion
CPUM	Colorado Plateau Uranium Miners
df	degrees of freedom
DLM	Distributed lag model
DLNM	Distributed lag non-linear model
ERR	Excess relative risk
GAM	Generalized additive model
GAMM	Generalized additive mixed model
GDR	German Democratic Republic
GLM	Generalized linear model
Hyp	Hypothesis
JEM	Job-Exposure matrix
kBq	Kilobecquerel
mSv	Millisievert
NA	Not available (Missing value)
PAM	Penalized piecewise exponential additive model
ph-model	Proportional-hazards model
P-IRLS	Penalized iteratively re-weighted least squares
REML	Restricted Maximum Likelihood
RMSE	Root Mean Squared Error
SAG	Staatliche Aktiengesellschaft der Buntmetallindustrie
SDAG	Sowjetisch-Deutsche Aktiengesellschaft
TSCE	Two-stage clonal expansion model
WLM	working-level month

List of Symbols

All individuals under risk prior to t:	$R(t)$
All individuals with event at t_i:	D_i
Baseline hazard rate:	λ_0
Censoring Indicator:	δ_i
Censoring Time:	C_i
Cut-points for piecewise exponential models:	κ_j
Exposure-response function for radon:	$f(x_{t-\ell})$
Exposure-response function for silica dust:	$f(z_{t-\ell})$
Exposure variable for radon:	x
Exposure variable for silica dust:	z
Gaussian Random Effect:	b_{ℓ_i}
Hazard rate:	λ
Lag-response function for radon:	$w_x(\ell)$
Lag-response function for silica dust:	$w_z(\ell)$
Likelihood:	$\mathcal{L}$
Maximum Lag:	L
Minimum Lag:	ℓ_0
Partial Likelihood:	$\mathcal{PL}$
Number of events at t_i:	d_i
Number of uncensored events:	k
Observed survival Time:	t_i
Offset:	t_{ij}
Parameter estimates:	β
Survival Time:	T_i
Time-constant confounders in the PAM:	x_i^p
Vector of exposure history:	$\boldsymbol{q}_{x,t}$

Acknowledgements

This thesis was written under the supervision and with the support of Prof. Dr. Helmut Küchenhoff (Ludwig-Maximilians-University, Munich).

The author would also like to thank Dr. Michaela Kreuzer and Dr. Christina Sobotzki (Federal Office for Radiation Protection) for the excellent cooperation in providing the data.

Additionally, Dr. Antonio Gasparrini (London School of Hygiene and Tropical Medicine, London) as well as Dr. Christian Kaiser (Helmholtz Zentrum, Munich) contributed substantially to this work by supporting the author with their extensive knowledge in their respective fields.
This support was via E-Mails and Skype due to the spatial distance in the case of Dr. Gasparrini and in the form of regularly organized meetings at the Helmholtz Zentrum with Dr. Kaiser, where the author held short presentations about the progress of his work.

Further valuable support concerning the work and the coding within the PAM framework was provided by M.Sc. Andreas Bender (Ludwig-Maximilians-University, Munich).

1. Introduction

Lung cancer is a certain type of cancer where carcinoma are located either directly in the lungs or in the bronchia. In many cases it is diagnosed in an advanced stage and so it often takes a lethal course (Pharmazeutische Zeitung Online (2016)).
Besides smoking and other risk factors, (occupational) exposure to the radioactive noble gas radon is also considered to have an impact on the risk of dying from lung cancer. Since it has been classified as pulmonary carcinogen (IARC (1988)) by the International Agency for Research on Cancer, there is ongoing research on the association of radon exposure and different types of cancer and cancer mortality.
The motivation for this thesis is to characterize the exposure-lag-response relationship between occupational radon exposure and lung cancer mortality, i.e. to assess the effect of different levels of exposure at different points in time after the exposure.
It thereby contributes to current research by applying a relatively novel framework to a data set of extensive size which, up to now, has only been investigated using other approaches. This promises further insights in the above-mentioned association and allows comparisons to related analyses of different data sets as well as to different analyses of the same data set.

Overall, the thesis is structured as follows: In chapter 2 the data set and its origins are described, while chapter 3 gives an overview on the theoretical background for the modelling part, including a passage on the current state of research. Thereafter, chapter 4 and 5 contain the results of the statistical analysis. Eventually, the results are discussed in chapter 6 and a conclusion is drawn in chapter 7.

Concerning the technical side of this thesis, the whole code for the estimated models as well as all figures and tables was created with the statistical software `R` (R Core Team (2016)) using the surface R-Studio (RStudio Team (2015)). The individual packages which are used for the data analysis are cited at the respective passages in the text or just added to the list of references if there's no suitable passage.

2. Data

2.1. The Wismut Company

After World War II had ended in 1945, Germany was divided into four occupation zones controlled by the United States, the United Kingdom, France and the Soviet Union. Soon after the establishment of these zones, soviet experts started to explore parts of Eastern Germany in search for natural uranium deposits for their nuclear weapons program. As a result of the successful searching, mining started in the year 1946 to a minor degree.

In 1947, the Soviet stock company Wismut (SAG Wismut) was founded and soon became one of the world's largest producers of uranium, with a cumulative production of 231.000 tons from 1947 until 1990. A maximum of more than 100.000 miners were employed at the company to reach the high production output.

The first years until 1953 are also known as "The wild years" (Wismut GmbH (2016a)), which were characterised by "poor working conditions, complete disregard for the environmental concerns of the densely populated areas, and the destructive exploitation of resources" (Wismut GmbH (2016a)). These conditions prevailed until the mid-1950s when the policy of the company changed in several ways. In contrast to before, the aim wasn't to maximize the short-term profit, but to introduce efficient methods to ensure long-term operable mining.

In order to achieve this goal, the company intensified their scientific investigation of the mines, which also lead to some huge improvements concerning the underground working conditions.

But despite the above-mentioned improvements, the mean exposure of the workers to radon (which is known to be carcinogenic), as well as to silica dust (also potentially carcinogenic), did not change immediately after the change in mindset of the company, as the following graphics show[1]:

[1] graphics for the mean exposure to long lived radionuclides, gamma radiation, arsenic and fine dust are to be found in the appendix A.1

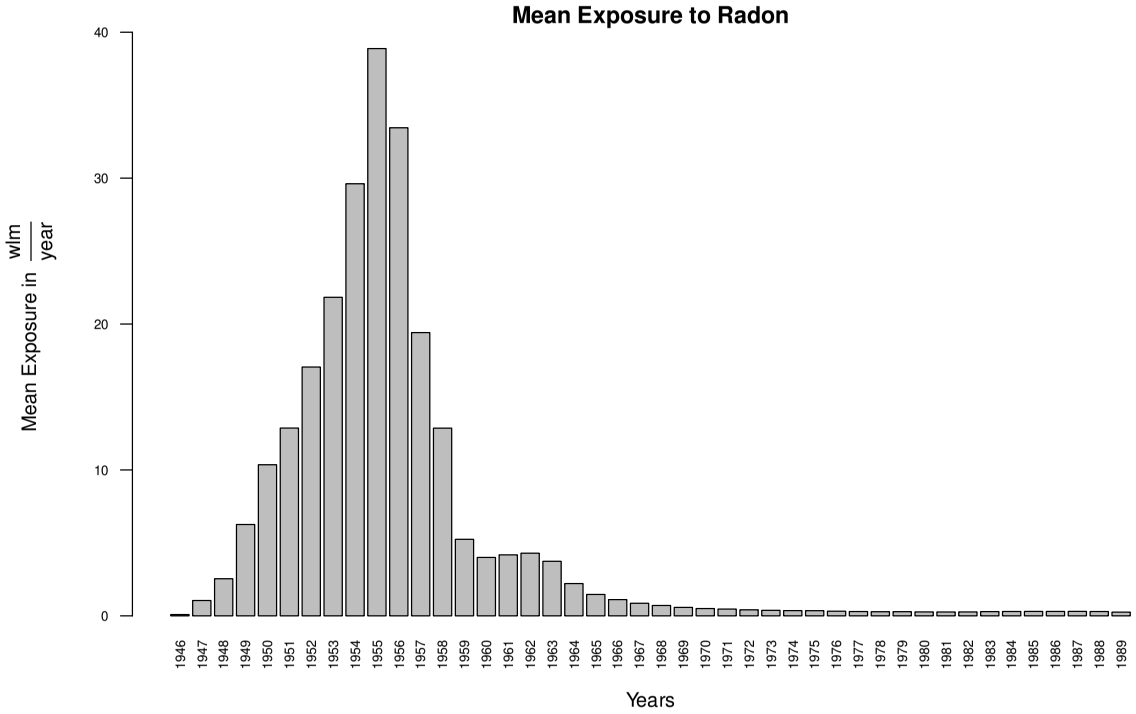

Figure 2.1.: *Mean Exposure to Radon in the Wismut Cohort (1946-1989)*

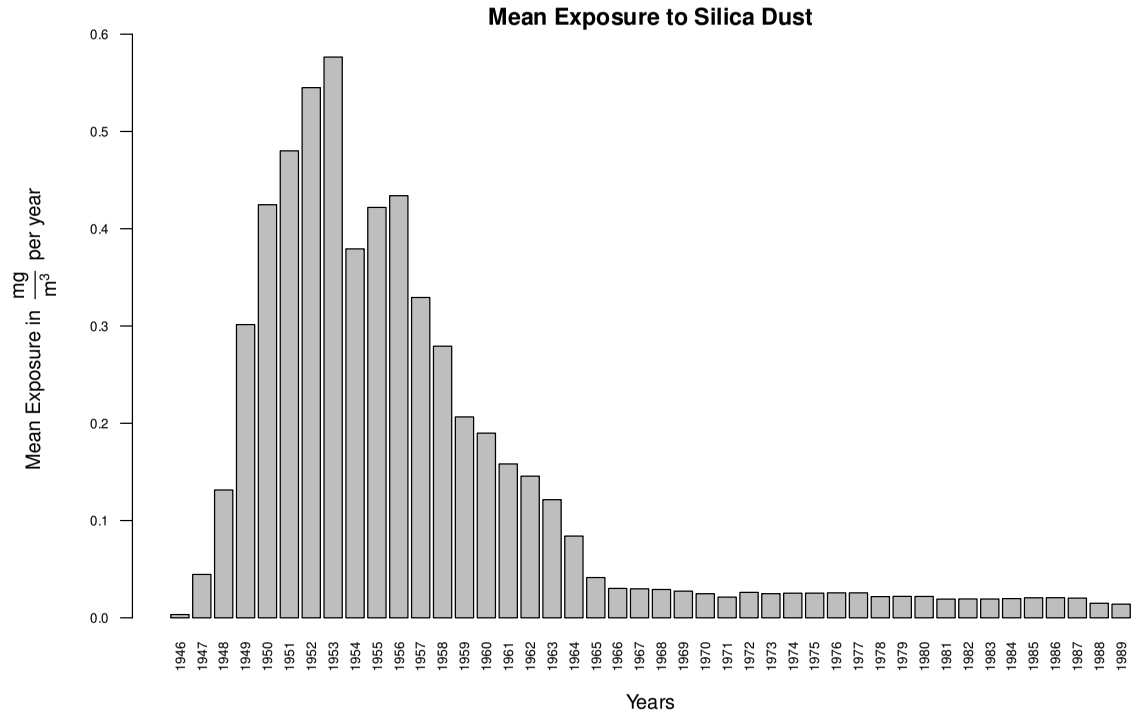

Figure 2.2.: *Mean Exposure to Silica Dust in the Wismut Cohort (1946-1989)*

First of all it has to be mentioned, that after the mean exposure of the miners to radon reached its peak in 1955, there was a decline in the following four years from an average of 38.87 wlm/year in 1955 to 5.24 wlm/year in 1959 (see figure 2.1). But it took until the mid- or the late 1960s to decrease the exposure permanently to a value that was below 1 wlm/year.
When looking at the mean exposure to silica dust (see figure 2.2), one observes a kind of related development, but also with some key differences. A peak in the exposure to silica dust was attained in 1953 with 0.57 mg/m^3 per year, and, unlike the radon exposure, there was no abrupt drop but rather a moderate decline until the mid-1960ies. During this period the average exposure still reached relatively high values, until it permanently dropped to a level of below 0.1 mg/m^3 per year on average.
Another facet that came along with the change of the mindset, was the foundation of a new company: The bi-national Soviet-German Company Wismut (SDAG Wismut). In the following years a permanent staff of about 45.000 miners, which was employed for more than three decades, emerged. This period from 1954 to 1991 is described as "State within the GDR state", as the company enjoyed many privileges by the GDR state (Wismut GmbH (2016b)). In contrast to the first years of the SAG Wismut, in this period lower exposures, coming along with the improved working conditions in the mines, prevailed.

After the German reunification in 1990, the mining was stopped permanently on December 31, 1990. In the following year, the USSR officially resigned on their shares of the SDAG Wismut in terms of a state treaty. So from that point in time on, the reunified Federal Republic of Germany obtained the sole ownership of the company and as a direct consequence, it was turned into a remediation company.
Another accompaniment of this step was the new formation and the renaming to the company's current name: Wismut GmbH (http://www.wismut.de/en/).
Its primary aims include "reclaiming former mining sites and restoring the environment for the benefit of man and nature" (Wismut GmbH (2016c)). Another aspect coming along with the rehabilitation process is the partnership between the Wismut GmbH and the regional economy, as many small and medium-sized businesses were assigned to the company's projects. So besides the environmental improvements that are being made, the Wismut GmbH also tries to strengthen the local companies.

2.2. The Wismut Cohort

The so called *wismut cohort* is the world's largest existing cohort data set of miners who have been occupationally exposed to several carcinogens. The data set contains information on 58.987 miners who were formerly employed at SAG Wismut or SDAG Wismut in the years from 1946 to 1989. It is provided by the Federal Office for Radiation Protection (BfS) in cooperation with the German Federal Ministry for the Environment, Nature Conservation, Building and Nuclear Safety (BMUB).
An update of the data is conducted every five years, when the vital status of the former miners is checked on a certain due date. If the vital status of a person has changed (i.e. he has died) within this period, it's being attempted to find out about the cause of death via residents' registration offices or health offices. Up to now, the previous dates of the follow-up have been December 31, 1998 (first follow-up), December 31, 2003 (second follow-up), December 31, 2008 (third follow-up) and December 31, 2013 (forth follow-up). This thesis is based on the data from the second follow-up from the year 2003[2]. A detailed overview as well as a further description of the data can be found in Kreuzer et al. (2002). Besides the research on the association between radon and lung cancer mortality, there are further topics which are investigated using this data set. These are amongst others the association between radon and leukemia or the joint impact of radon combined with the other substances.

2.3. Description of the data

General Information

The present data set is of dimension 58.987 x 271, as one row contains the information for one miner. In the first seven columns, the original data set contains an ID-variable (column 1), as well as information on the date of birth (column 2), the end of follow-up (column 3), the vital status at the end of follow-up (column 4), an indicator-variable for "death by lung cancer" (column 5) and the date of the begin (column 6) and the end (column 7) of employment of the miner at the Wismut Company. Besides these

[2]As the data is provided by the BfS within the limits of predefined project, the author has no influence on the which version of follow-up is made available

basic variables, there's also information of a person's exposure (in the form of job-exposure matrices[3]) to six different, potentially harmful, substances. Namely, these are radon (columns 8 to 51), long lived radionuclides (columns 52 to 95), gamma radiation (columns 96 to 139), arsenic (columns 140 to 183), fine dust (columns 184 to 227) and silica dust (columns 228 to 271).
Most important is the data on the exposure to radon, which, in 1988, has been officially classified as pulmonary carcinogen (IARC (1988)).
Furthermore silica dust is considered as an important confounder and will be included in the modelling part as well. Several forms of silica have also been investigated by the International Agency for Research on Cancer (IARC (1997)) and some forms of crystalline silica were considered as carcinogenic to humans. But despite this extensive and detailed amount of information, there are two important issues which aren't covered by this data set:

- *Smoking:* The data set spans over a period of time, in which a majority of the population, and especially of the miners, were considered to be heavy smokers. But, unlike some other data sets concerning occupational radon exposure (e.g. the CPUM cohort), the Wismut cohort doesn't contain any information on the smoking habits of the study population. This aspect, which is relevant insofar as smoking is also considered to be one of the major reasons for lung cancer, may limit the results obtained with this data to a certain extent.

- *Lung cancer incidence:* Another issue that can't be investigated using the data of the wismut cohort is the incidence of lung cancer, as it only contains information on lung cancer as a cause of death, but not on its occurrence. So the crucial information on the point in time, when a person actually got the diagnosis of lung cancer, is not given in the data.

Besides this general description on the data, the following part of this chapter contains more detailed insight in the form of descriptive statistics and quantitative information. At the end, there will also be an explanation on how the data was transformed to be suitable for the different types of models which were applied to it.

[3]for detailed information about the estimation of the job-exposure matrices see Lehmann et al. (1998) or Lehmann (2004)

Characteristics

To get a brief overview of the main characteristics of the data set, this section will start with a few tables containing information on the full cohort as well as different subgroups. The first two tables[4] incorporate data about the full cohort (2.1) and about all cases of death (2.2).

Full cohort	Cases:	58987		(100%)	
	Min.	1st Quartile	Median	3rd Quartile	Max.
Follow-up period ($years$)	15.33	47	60.33	70.17	103.2
Duration of Employment ($years$)	0.4167	4.25	10.17	21.5	45.5
Cum. Radon Exposure ($\frac{wlm}{year}$)	0	1.746	18.42	262.4	3224
Cum. Silica Dust Exposure ($\frac{mg}{m^3}/year$)	0	0.4036	1.761	8.524	55.98

Table 2.1.: Characteristics of the full cohort

All Deaths	Cases:	20920		(35.47%)	
	Min.	1st Quartile	Median	3rd Quartile	Max.
Follow-up period ($years$)	17.5	56.5	65.67	73.5	103.2
Duration of Employment ($years$)	0.4167	6.167	12.83	24.17	44.5
Cum. Radon Exposure ($\frac{wlm}{year}$)	0	6.629	182.1	749.6	3224
Cum. Silica Dust Exposure ($\frac{mg}{m^3}/year$)	0	2.014	7.406	16.53	55.98

Table 2.2.: Characteristics of the cases of death

While the full cohort contains 58.987 miners, slightly more than a third of them (20.920) have died during the follow-up period until December 31, 2003. When comparing the distributions of the follow-up times, the miners who died, tend to have had a longer follow-up than the full cohort. At first sight this seems a bit odd, but once one looks at the huge fraction of the observations which are censored (64.53%), this might be the reason for the occurrence of this unexpected difference.
Information on the duration of employment is to be found in the second row. When again comparing both tables, the cases of death have a somewhat higher 1st and 3rd

[4]All tables displayed in this thesis have been created via the `xtable`-package (Dahl (2016)).

quartile as well as a higher median. When looking at the distribution of the cumulative exposure to radon (measured in wlm/year) and the cumulative exposure to silica dust (measured in mg/m^3 per year), there are striking differences in the characteristics of the distributions. 1st and 3rd quartile and the median are much higher for the distribution of cumulative radon exposure as well as for the distribution of cumulative silica dust exposure. So, one can conclude that the miners who died had been exposed to a much higher amount of radon and silica dust than the full cohort.
The following two tables show summary statistics of all cases from table 2.2 separated by the cause of death. The cause is either classified as *lung cancer* or *other causes*, where *other causes* contains all cases with unknown or known, but not lung-cancer-related, causes of death.

Lung Cancer	Cases:	3016		(5.11%)	
	Min.	1st Quartile	Median	3rd Quartile	Max.
Follow-up period $(years)$	27.92	57.42	64.17	70.33	91.83
Duration of Employment $(years)$	0.5	8.417	17.08	27.08	43.58
Cum. Radon Exposure $(\frac{wlm}{year})$	0	75.65	564	1031	2990
Cum. Silica Dust Exposure $(\frac{mg}{m^3}/year)$	0	4.744	13.46	21.4	53.71

Table 2.3.: Characteristics of the lung cancer cases

Other causes	Cases:	17904		(30.35%)	
	Min.	1st Quartile	Median	3rd Quartile	Max.
Follow-up period $(years)$	17.5	56.25	65.92	74.08	103.2
Duration of Employment $(years)$	0.4167	5.833	12.17	23.5	44.5
Cum. Radon Exposure $(\frac{wlm}{year})$	0	4.629	138	678.3	3224
Cum. Silica Dust Exposure $(\frac{mg}{m^3}/year)$	0	1.777	6.68	15.48	55.98

Table 2.4.: Characteristics of the cases of death by other causes

When only looking at the cases of death, 14.42% of these occurred due to lung cancer, which equals about 5.11% of the full cohort. Compared to the cases of deaths by other cause (85.58% of the cases of death, 30.35% of the whole cohort), these cases had a much longer duration of employment (higher 1st and 3rd quartile, higher median) as well as a

much higher cumulative exposure to both radon and silica dust.
As far as the length of the follow-up period is concerned, there don't really seem to be big differences between both groups.

Data preparation

Before using the data to fit a model and even before transforming into the right form, some adjustments had to be made in advance.
First of all, the data was checked for missing values. While there were none in the basic variables as well as in the JEM for radon, 12848 missing values occurred in the JEM for silica dust. To get an idea of how much of the values are NA, one should take into consideration the dimension of a JEM of 58987 x 44, which yields 2595428 single values. So the fraction of missing values for the exposure to silica dust is just about 0.5%.
After taking a closer look at the structure of the NAs, it became clear that all 12848 of them were due to 292 persons in the data set who had missing values for silica dust for all 44 years. And while they are still included in the descriptive statistics above, these miners are from now on excluded from the further analysis of the data. A comparison of the values of the basic variables of these 292 miners to those from the whole cohort is to be found in the appendix A.2
A second point is that only those years, in which (a) a miner was either exposed to radon or silica dust in the current year or (b) he had been exposed to one of the substances in at least one of the previous 40 years, are included in the analysis. This decision was made, as the aim of this thesis is, to specify the association between occupational radon exposure and the risk of dying from lung cancer and so the years in which a person is not under risk don't matter for the estimation. As a consequence, all years before the actual begin of employment of a miner were discarded in the analysis. The years after employment had ended, are still kept in the data set, as the miners were still under risk in these years due to the lagged values of the exposure.
One further issue does concern only one worker (ID = 82524), whose begin of employment is given as "November 1945", which is insofar implausible, as the company wasn't founded before 1946. To be consistent with previous works, this is taken as given and the course of action in this case is as follows:
For the calculation of measures like e.g. the duration of employment this date is used. For the estimation in the models, this worker is included from 1946 onwards, as the JEM

starts in this year (i.e. no exposure values for the years before) and so this is technically the first point in time where he is under risk.

Transformation

In order to estimate the different proposed models (Distributed lag non-linear models and Penalized piece-wise exponential additive models) the original data set had to be transformed into two different versions of the data, as both of the models have different requirements concerning the form of the data.
The following explanations give a quick overview on how the data was transformed, starting with the DLNMs:

The estimation of the DLNMs is executed in `R` via the `coxph`-function from the `survival`-package (Therneau (2015), version 2.38, Therneau and Grambsch (2000)). But as the exposure history of a miner changes year by year and the aim of the model is to estimate the effects of certain exposure at a certain lag, the data set has to be expanded. So instead of containing one row per miner, the expanded data set contains one row per miner per year under risk. A second step which has to be taken is the separation of the date frame into two separate parts due to technical issues. The expanded data set just contains, besides the ID, information on the begin and the end of each interval (in this case: the begin and end of each year under risk), the status indicator for the event (death by lung cancer) and the other covariates (here: the age at first exposure and calendar time).
The second, separate, part includes the already above-mentioned exposure history created from the JEM. It includes for every miner for every year the current exposure as well as the lagged exposures from the previous year and is created in `R` via the `exphist`-function from the `dlnm`-package (Gasparrini (2011), version 2.2.7).

Besides some similarity concerning the expansion of the data frame to one row per miner per year under risk, the transformation for the estimation of the PAMs was slightly different: The PAMs are estimated in `R` via the `gam`- (Wood (2006)) or the `bam`-function (Wood et al. (2015)) from the `mgcv`-package (Wood (2014), version 1.8-12), which has different requirements to the form of the data. These functions require the information all to be in one data set. Of course, all the basic variables (ID, the interval, the status

indicator, age at first exposure, calendar time) are included, as well as a variable specifying the offset, which is practically always zero, except from (potentially) the last row of each person.
Additionally the following matrices are included in the data set:

- Exposure Matrices ("expMatRad" for radon and "expMatSil" for silica dust):
  This matrix is of dimension (years under risk) x 44 for every miner. It contains the complete information of the person's exposure in every row

- Interval Matrix ("intMat"):
  A matrix also of dimension (years under risk) x 44 which contains in every row 44 times the interval which a person is in

- Lag Matrix ("lagMat"):
  In this matrix, the a priori chosen lag structure is defined. It has the same dimension as the other matrices and contains zeros and ones that either exclude or include the respective lag. So in this case, the first two rows for every miner only consist of zeros, as no influence before lag 2 is assumed and from there on, every row contains forty ones, as an effect of up to forty years after the initial exposure is assumed to be possible. After that, it's a lower triangular matrix of zeros, as after forty years there's no further effect assumed

Further technical and theoretical aspects of the models will be explained in detail in chapter 3.
It will also give an overview on what has been discovered about the exposure-lag-response relationship between occupational radon exposure and lung cancer mortality up to now. The different findings of several authors will be divided according to which type of model they used in their research. A special focus will be on the results obtained by Gasparrini (2014) using DLNMs.
Additionally to this, the state of research on the exposure-lag-response-relationship of silica dust exposure and lung cancer mortality will also be reported if it was investigated in the respective papers.

3. Theoretical Background of the Modelling Framework

In this chapter, a literature review on the current state of research as well as the complete theoretical background for the performed analysis in this thesis will be provided. At first, the current state of research is summarized in section 3.1 while in the subsequent sections 3.2 up to 3.3.3 the theoretical knowledge needed for the modelling part is presented.

As a starting point for the theoretical explanations, the cox proportional-hazards model, as well as the the inference for estimation of the parameters of this model, are explained briefly in section 3.2. This type of models is used for survival data (also called time-to-event data), like present in this case. Survival data is the term for a special kind of data, where the variable of interest is not some binary or metric variable, but a period of time which is called failure time. So this model estimates the risk to have the event of interest at a certain point in time, given a set of covariates. These covariates usually include simple time-constant or time-varying variables.
But as in this case, we have more complex data than just simple survival data due to the additional complexity in the covariates. So the framework has to be extended in several ways. The additional complexity of the data comes from the detailed information on the occupational exposure to radon and silica dust, given in the form of a job-exposure-matrix.
This is incorporated by extending the the cox ph-model in a first step to the class of distributed lag models in section 3.3.1 which allows the estimated effect of the exposure on the hazard rate to vary smoothly depending on its temporal relation to the hazard rate. This effect is estimated by a, potentially non-linear, lag-response-function.
In a second step, taken in section 3.3.2, the DLM framework is extended to distributed lag non-linear models in which, additionally to before, the relationship between the response (i.e. the hazard rate) and the exposure is also potentially non-linear.

Finally, the theory needed to perform the analysis using the penalized piecewise exponential additive models will be explained in section 3.3.3. Since these models are based on the class of generalized additive mixed models (GAMMs), a different theoretical background will be needed.

3.1. Current State of Research

Up to now, there are already several studies which are concerned with the impact of the (occupational) exposure to radon and the mortality from lung cancer, other types of cancer or health outcomes in general. Silica dust and also smoking habits have been considered as important possible confounders in these studies. This chapter tries to give an overview on the current state of research and on some of the the well-established models for the analysis. A special focus will be on the work of Gasparrini (2014) as the modelling framework of this thesis orientates itself strongly by the framework applied in his work.

One thing that becomes obvious when studying the different papers, is that there's quite a variety of different frameworks in which the relationship between radon exposure and (lung cancer) mortality or other outcomes is investigated. The following list gives a short and non-exhaustive overview over some of these different types of models being used in current research, with the exemplarily summarized papers written in brackets behind:

- Distributed lag non-linear Models (Gasparrini (2014), Gasparrini et al. (2016))
- Poisson Regression Models (i.a. Grosche et al. (2006), Kreuzer et al. (2010), Walsh et al. (2010))
- Two-stage clonal expansion (TSCE) Models (Zaballa and Eidemüller (2016))

The summaries for analysis of radon exposure in the different frameworks are presented in form of a enumerated list for the case of Gasparrini (2014) and in form of short summaries for all other above-mentioned papers. Additional information on the potentially added confounders is given separately in each of the respective paragraphs.

3.1.1. Distributed lag non-linear Models

Gasparrini (2014): Modeling exposure-lag-response associations with distributed lag non-linear models

In his paper from the year 2014, Gasparrini shows three main results:

1. The exposure-response-function is not linear. In his selected model, he estimates it to have a more or less steady increase between the exposure values of 0 wlm/year and 50 wlm/year. At 50 wlm/year there seems to be a real break in the relationship, as the curve flattens out after this point. Misspecifying this relationship as linear has a severe impact on the results, as it massively influences the estimation of the lag-response-function. He shows this, by also estimating DLMs (with a linear exposure-response-function) and comparing them via the AIC and BIC. In case of the DLMs, this comparison indicates the lag-response-function to be a constant function along the lags, which is not realistic from a physiological point of view.

2. The lag-response-function is not a constant function. Gasparrini (2014) assumes the exposure to have a latency period of two years, i.e. there's no effect before lag two. From this point onwards, the estimated hazard ratio increases up to its peak at a lag of eleven years after the initial exposure. Increasing the temporal distance to the exposure beyond this point leads to a decreasing hazard ratio. The risk associated with occupational radon exposure eventually fades away completely approximately 35 years after a person was exposed.

3. In a simulation study which was also performed in the paper, the performance of the two information criteria (AIC and BIC) was evaluated with the following findings: In general, models that were selected using the AIC were stated to have a better performance than BIC-selected models. The term *performance* was measured in terms of relative bias, coverage and relative RMSE.
   The AIC-selected models were subject to moderate overfitting, which sometimes leads to the suggestion of overly complex models where simpler underlying scenarios are present. BIC-Selection on the other hand, showed a tendency in the other direction: These models showed severe underfitting, i.e. models with very simple

exposure-lag-response-associations were selected in cases where the underlying relationship was more complex. Particularly the assumption of linearity was affected by this lack of adequacy. These findings are completely in line with the general tendency of the BIC to select simpler models than the AIC.

Besides Radon, Gasparrini also considers Smoking habits as a confounder in his models. The form of how it is added to the model is chosen a priori to be fairly simple and isn't subject to further investigation in the process of finding the optimal model. Also the results of the estimated effects for Smoking aren't reported in this paper. The motivation for this proceeding is reported to be the limited information on the smoking histories in this particular data set.
Additionally, in order to control for a potential trend in lung cancer risk over time, he added another covariate to the model which contains the calendar time centered around the year 1970.

3.1.2. Poisson Regression Models

In contrast to the relatively new DLNM framework, Poisson regression models represent a very well established and frequently used framework in radiation epidemiology. So the following passage contains some more papers from a period of time over the years 2006 to 2010, in which this approach has been applied. All of these studies deal with different versions of the wismut data.

Grosche et al. (2006): Lung cancer risk among German male uranium miners: a cohort study, 1946-1998

This paper was the first one, to analyze a version of the wismut cohort data with respect to lung cancer risk. As it was already published about 10 years ago, the data of the first follow-up (until 31.12.1998) was used. The assumed latency period for the radon exposure was five years and a linear exposure-response-relationship was presupposed. Another difference is that, compared to the analysis in this thesis, in this paper the accumulated radon exposure was used. The estimation of the excess relative risk due to the exposure was stratified by 'attained age' ($<$ 55, 55-64, 65-74, and 75 and more

years) as well as by 'time since exposure' (5-14, 15-24, 25-34, and 35 and more years). The excess relative risk was found to be the highest in the second category of 'time since exposure' (15-24 years) and was reported to be significantly lower in the other three categories. For the variable 'attained age' a moderate decline in the ERR per wlm was reported. Overall, the ERR/wlm was estimated to be 0.21%.

Kreuzer et al. (2010): Radon and risk of death from cancer and cardiovascular diseases in the German uranium miners cohort study: follow-up 1946-2003

Kreuzer et al. investigated the same data set (wismut cohort, second follow-up until December 31, 2003) which is analyzed in this thesis. The focus of the paper was on exploring the relationship between the occupational exposure to radon an the risk of dying from (a) different types of cancer (lung cancer, extrapulmonary cancers and cancers of the extrathoracic airways and trachea) and (b) cardiovascular diseases.
As a latency period for the radon exposure, five years were assumed and again the cumulative exposure was added to the model linearly. Besides the exposure, 'attained age' and 'individual calender year' were considered as coavariates and the five potential confounders from the wismut cohort data set were added to the model separately.
While a significant increase in the risk in lung cancer (ERR/wlm = 0.19%) and cancer of the extrathoracic airways and trachea (ERR/wlm = 0.062%) were reported, no (or no significant) increase in risk for the other analyzed causes of death was found.

Walsh et al. (2010): Radon and the risk of cancer mortality-Internal Poisson models for the German uranium miners cohort

Like Kreuzer et al. (2010), the work of Walsh et al. is also based on the second follow-up of the wismut cohort and it also investigates the risks for more than one type of cancer. Further similarities are the use of a cumulative measure for the radon exposure and its linear modelling. The covariates 'age at median exposure', 'time since median exposure', centered around their respective means, and the so-called 'radon exposure rate' were added to the model. The exact definition of the 'exposure rate' can be found in the report from the National Research Council, Committee on the Biological Effects of Ionizing Radiation (1999).

In this paper, a also statistically significant effect of cumulative radon exposure, with an ERR/wlm of 0.20% was reported.

3.1.3. Two-stage Clonal Expansion Models

Zaballa and Eidemüller (2016): Mechanistic study on lung cancer mortality after radon exposure in the Wismut Cohort supports important role of clonal expansion n lung carcinogenesis

A very novel approach to this cohort is represented by the TSCE model by Zaballa and Eidemüller (2016) from the Institute of Radiation Protection in Munich. For the research in carcinogenesis it is a standard approach, which was already being applied to other cohort data sets (e.g. Luebeck et al. (1999), Kai et al. (1997)). Compared to the other presented models, it is a framework which is rather based on a biological point of view. Regarding the statistical side, the different stages of the model are modelled as Poisson processes.
Concerning the data, an exactly alike version of the second follow-up of the wismut cohort as in this thesis was used by excluding all workers with missing values for silica dust (as it is an important confounder in their model) exposure and considering a person to be at risk from their employment until the end of follow-up. The covariates 'attained age' and 'calendar year' are allowed to influence the baseline hazard and again, the 'exposure rate' is also added to the model. Another important part of the results-section is the comparison to an ERR-model (Poisson regression model).

3.2. The Cox proportional-hazards model

Theory

The cox ph-model was introduced by Sir David Roxbee Cox, a British statistician, in 1972. The remarks about this model in this thesis will mainly be based on the original paper from Cox (1975) and will orientate themselves notationally at the lecture notes from Kauermann (2014).
This whole modelling framework is mainly based on the concept of the hazard rate. The term hazard rate λ is defined as the probability that a subject who has survived up to a certain time point t will have an event in the next small time interval δt, divided by the length of this aforementioned small interval:

$$\lambda(t) = \lim_{\delta t \to 0} \frac{P(t < T < t + \delta t | T > t)}{\delta t} \tag{3.1}$$

In the cox ph-model the hazard rate is modelled dependent on the time point t and the included covariate values $\boldsymbol{x}_i$ of person i:

$$\lambda(t, \boldsymbol{x}_i) = \lambda_0(t) \cdot \exp(\boldsymbol{x}_i^T \beta) \tag{3.2}$$

The so-called baseline hazard rate $\lambda_0(t)$ is dependent on the time point t, while the time-independent covariate effects are incorporated by $\boldsymbol{x}_i^T \beta$. In the estimation process, the baseline hazard won't be specified as it is only viewed as a nuisance-parameter and the main interest lies in specifying the β-parameters. Another important thing to keep in mind is that the parameter vector does not contain an intercept.
The eponymous property of the model, the proportional hazard assumption, leads so some kind of parallel course of two subjects' hazard rates, where the relative risk of subject one compared to subject two is independent of the time t:

$$\frac{\lambda(t, \boldsymbol{x}_1)}{\lambda(t, \boldsymbol{x}_2)} = \frac{\lambda_0(t)}{\lambda_0(t)} \cdot \frac{\exp(\boldsymbol{x}_1^T \beta)}{\exp(\boldsymbol{x}_2^T \beta)} = \exp((\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \beta) \tag{3.3}$$

So in order to compute the relative risk of to individuals compared to each other one has

to calculate the difference of their covariates and insert the result in the model equation.

Inference

The following part describes briefly how the inference in the model is performed. It is based on both of the works from Cox (1972) and Cox (1975) and the notation is again inspired by the lecture notes from Kauermann (2014).
Cox rejects the use of an ordinary likelihood approach in favor of the so-called *partial likelihood* approach, since it is possible to skip the estimation of the nuisance parameter $\lambda_0(t)$ and thus reduce the dimensionality this way. The motivation and the theoretical derivation for this can be found in Cox (1975). In this thesis, only the direct consequences for the cox ph-model are presented.
One more thing that has to be taken into consideration is the possible right-censoring of the survival times, which also has an impact on the inference.
If one considers the formula given in equation 3.2, the full likelihood of the model is given as follows:

$$\mathcal{L}(\lambda_0(t), \beta) = \prod_{i=1}^{n} \lambda_0(t_i) \exp(x_i'\beta)^{\delta_i} \exp\left\{ - \int_0^{t_i} \lambda_0(s) \exp(x_i'\beta) \right\} \mathrm{d}s \tag{3.4}$$

with t_i being the observed survival time (i.e. $t_i = min(T_i, C_i)$) and δ_i being the censoring indicator (i.e. $\delta_i = I\{T_i \leq C_i\}$), taking a value of one if a persons survival time is observed and zero otherwise.
This likelihood from equation 3.4 can be reduced via some algebraic transformations to a partial likelihood of the following form:

$$\mathcal{PL}(\beta) = \prod_{i=1}^{k} \frac{\exp(x_{(i)}'\beta)}{\sum_{j \in R(t_{(i)})} \exp(x_j'\beta)} \tag{3.5}$$

with k being the number of uncensored events ($k \leq n$) and $R(t)$ being the number of individuals under risk just prior to t. The proceeding to obtain the estimates for the β-vector is the same as in the usual inference for the full likelihood, by maximizing it with respect to β. But more importantly, this equation only holds under the assumption of singularity of the individual events, i.e. there are no ties in the data.

As this is a rather unrealistic assumption, especially for big data sets, there are methods to adapt the partial likelihood function to this problem. Actually, the `coxph`-function in `R` offers three possible options:

- The Efron method:

$$\mathcal{PL}_{Efron}(\beta) = \prod_{i=1}^{k} \frac{\exp(\sum_{j \in D_i} x_j \beta)}{\prod_{j=1}^{d_i} \left[\sum_{k \in R(t_{(i)})} \exp(x_k \beta) - \frac{d_i - j}{d_i} \sum_{k \in D_i} \exp(x_k \beta) \right]} \tag{3.6}$$

- The Breslow method:

$$\mathcal{PL}_{Breslow}(\beta) = \prod_{i=1}^{k} \frac{\exp(\sum_{j \in D_i} x_j \beta)}{\left[\sum_{j \in R(t_{(i)})} \exp(x_j \beta) \right]^{d_i}} \tag{3.7}$$

- The Exact method:

$$\mathcal{PL}_{Exact}(\beta) = \prod_{i=1}^{k} \frac{\exp(\sum_{j \in D_i} x_j \beta)}{\sum_{Q \in Q_{(i)}} \exp(\sum_{j \in Q_{(i)}} x_j \beta)} \tag{3.8}$$

While the Breslow method works very well for a small d_i (i.e. a small number of events at the same time), the Efron method is considered as being more accurate for a large number of ties and as good as the Breslow method for a small number. The Exact method is computationally rather expensive and is appropriate for a small set of discrete time points.
Taking these properties of the different methods into consideration, in this thesis the Efron method will be used for tie handling. In Gasparrini (2014) the same decision about tie handling was made.

3.3. Models for exposure-response relationships

Flexible modelling of exposure-response relationships has been subject to research for quite a while now. The modelling part in this thesis will be based on an approach introduced by Gasparrini et al. (2010) and Gasparrini (2014).

But it was already much longer ago, when research on the topic of DLMs emerged (Almon (1965)) in the field of econometrics. In the following years, especially in more current research, it has been extensively used in the field of epidemiology.

While the work form Gasparrini et al. (2010) is mainly about the methodology in the context of time series analysis, the paper from Gasparrini (2014) is particularly about the application of the methodology in the context of modelling time to event data with the cox ph-model. In both of the mentioned publications, the first part is about describing simple DLMs which are used for modelling linear exposure-response relationships. Afterwards, this framework is extended to modelling non-linear exposure-response relationships. The structure used here will be identically, but will mainly focus on the modelling in the context of survival analysis. In Gasparrini et al. (2016) more elaborate methods for modelling the smooth functions in the exposure-lag-response-relationships in DLNMs, based on penalized splines, are presented.
But as these methods have not been officially published yet and due to the fact that the analysis of the B-Spline based models had already made much progress when author was able to have a look at these methods, they are not further pursued in this thesis. But nevertheless, they definitely have to be considered in future research.

Another approach for modelling these special relationships was introduced by Bender et al. (2016) who connect the concept of piecewise exponential models with the framework of Generalized additive mixed models. By doing this, they exploit the link between the likelihood of a piecewise exponential model and the likelihood of a generalized linear Poisson-Model with certain constraints.
This leads to another framework which is able to model the exposure-lag-response-association as a combination of smooth functions.

3.3.1. Distributed lag models

Distributed lag models are used in the attempt to describe the lag-response-relationship in the assumed presence of a linear effect. The term *lag-response-relationship* means the delayed effect of an influential variable on the dependent variable. These kinds of models are heavily used in econometrics (especially time series analysis), as well as in social sciences and epidemiology.
Gasparrini (2014) defines a function $s(x,t)$ which is generally applicable to several modelling frameworks and regression models. With this function he describes the dependency between the dependent variable and the influential variable by an exposure history of the influential variable x at a certain time point t.
The algebraic notion is given as follows:

$$s(x,t) = \int_{\ell_0}^{L} x_{t-\ell} \cdot w(\ell)\,\mathrm{d}\ell \quad \approx \quad \sum_{\ell=\ell_0}^{L} x_{t-\ell} \cdot w(\ell) \tag{3.9}$$

In the part of the equation on the left hand side of the approximately equal sign, the period $L-\ell_0$ defines the lag period over which the exposure has an effect on the outcome. The part on the right hand side is an approximation of the integral on the left hand side which is used due to computational purposes. This approximation is obtained through a discretization of the lag period into equally spaced time units.
In order to define a statistical model for (3.9), Gasparrini (2014) expresses the lag-response function $w(\ell)$ in matrix notation through the vector $\boldsymbol{q}_{x,t}$ of the exposure history:

$$\boldsymbol{q}_{x,t} = (x_{t-\ell_0}, ..., x_{t-\ell}, ..., x_{t-L})^\top \tag{3.10}$$

From this representation it is obvious that this vector is different for every time point t. The value ℓ_0 can be interpreted as the minimum lag at which an exposure affects the outcome, while analogously L is the maximum lag. Using (3.10) one can write the function $s(x,t)$ in a compact matrix notation:

$$s(x,t;\boldsymbol{\eta}) = \boldsymbol{q}_{x,t}^\top \boldsymbol{C}\boldsymbol{\eta} = \boldsymbol{w}_{x,t}^\top \boldsymbol{\eta} \tag{3.11}$$

A transformation of the lag vector ℓ of dimension $(L - \ell_0 + 1)$ with a vector defining the basis functions of dimension v_l yields $\boldsymbol{C}$. Hence, this matrix has the dimension

$(L - \ell_0 + 1) \times \upsilon_l$. So the result of (3.11), $\boldsymbol{w}_{x,t}^T \boldsymbol{\eta}$, is just the vectorial representation of the integral of $x \cdot w(\ell)$ over the interval $[\ell_0, L]$ with parameters $\boldsymbol{\eta}$.

It is also mentioned in Gasparrini (2014), that the equation in (3.11) is a more general representation than the one that was used by Gasparrini et al. (2010), because it isn't specifically tailored to the framework of time series analysis, but also applicable to other situations.
Information on the estimation and the prediction in the DLM framework will be given in section 3.3.2, as the DLMs can be seen as a special, simpler, case of the DLNMs.

3.3.2. Distributed lag non-linear models

Theory

As well as the previous section, this one will also start by giving an algebraic followed by a matrix representation. All remarks will get slightly more complicated due to the allowed non-linearity of the exposure-response relationship. Subsequently, this chapter will also contain information about the estimation process and prediction in DLNMs.
The model is extended by substituting the $x_{t-\ell}$ from (3.9) by an exposure-response function $f(x_{t-\ell})$, which yields the following expression:

$$s(x,t) = \int_{l_0}^{L} f(x_{t-\ell}) \cdot w(\ell)\, \mathrm{d}\ell \quad \approx \quad \sum_{\ell=\ell_0}^{L} f(x_{t-\ell}) \cdot w(\ell) \tag{3.12}$$

This equation is one step towards a more flexible representation since it allows non-linearity in the exposure-response function. But one clear and strong disadvantage is imposed by the assumption of independence of the exposure-response and the lag-response function. This would mean, that the exposure-response-function has to have the same form, no matter how long ago the exposure occurred. In many cases this assumption doesn't hold and thus limits the flexibility of the model. So a more general and more flexible representation is given by the following algebraic formula:

$$s(x,t) = \int_{\ell_0}^{L} f \cdot w(x_{t-\ell}, \ell)\, \mathrm{d}\ell \quad \approx \quad \sum_{\ell=\ell_0}^{L} f \cdot w(x_{t-\ell}, \ell) \tag{3.13}$$

Consequently, as $f \cdot w(x_{t-\ell}, \ell)$ is a bivariate function in x and in t, it is called the exposure-lag-response function. By using this representation, one is able to model the exposure-response association and the lag-response association simultaneously. In Gasparrini (2014) this results in the so-called *exposure-lag-response association* and graphically in a three dimensional *exposure-lag-response surface*.
For being able to model this kind of relationships, one needs a special kind of tensor products. This tensor product is parametrized through the so-called *cross-basis* (Armstrong (2006)). Gasparrini et al. (2010) give an algebraic representation of the cross basis, but the definition used in this thesis is the revised version from Gasparrini (2014). Additionally to the previously introduced dimension υ_l of the lag-basis, now the dimension υ_x for the basis representing $f(x)$ has to be taken into consideration when constructing the cross-basis. So in the same way as the matrix $\boldsymbol{C}$ was constructed for the lags, now the matrix $\boldsymbol{R}_{x,t}$ of dimension $(L - \ell_0 + 1) \times \upsilon_x$ is constructed for the vector $\boldsymbol{q}_{x,t}$ of the exposure history.
By defining

$$\boldsymbol{A}_{x,t} = \left(\mathbf{1}_{\upsilon_l}^{\top} \otimes \boldsymbol{R}_{x,t}\right) \odot \left(\boldsymbol{C} \otimes \mathbf{1}_{\upsilon_x}^{\top}\right), \tag{3.14}$$

the cross-basis function can be written in the following fashion:

$$s(x,t;\boldsymbol{\eta}) = \left(\mathbf{1}_{\upsilon_x \cdot \upsilon_l}^{\top} \boldsymbol{A}_{x,t}\right) \boldsymbol{\eta} = \boldsymbol{w}_{x,t}^{\top} \boldsymbol{\eta} \tag{3.15}$$

It is worth to note, that the simpler class of DLMs is also incorporated by this equations and can be created through the appropriate choice of the basis functions for $\boldsymbol{R}_{x,t}$.
Despite the quite unusual and complex form of the model it can be estimated by standard regression models, as is to be shown in the next paragraph.

Inference

In Gasparrini (2014) besides the theory, inferential methods for this framework are provided as well. As the association of the model is fully parametrized by the parameters in $\boldsymbol{\eta}$, the model can be estimated by the following steps:

1. The vector $\boldsymbol{q}_{x,t}$ of dimension $1 \times (L - \ell_0 + 1)$ is extended to a matrix $\boldsymbol{Q}$ of dimension $N \times (L - \ell_0 + 1)$ containing the exposure histories of all N observations of the data
2. Employing $\boldsymbol{Q}$ instead of $\boldsymbol{q}_{x,t}$ for constructing the matrix $\boldsymbol{R}_{x,t}$ in (3.14) leads to $\boldsymbol{A}_{x,t}$ and consequently yields a matrix $\boldsymbol{W}$ of transformed variables instead of the vector $\boldsymbol{w}_{x,t}$
3. This newly obtained matrix $\boldsymbol{W}$ is now included as design matrix in the regression model of choice (in this case: a cox ph-model) to estimate the parameters in $\boldsymbol{\eta}$
4. The number of degrees of freedom (df) which will be used for evaluating the model, is determined by the number of estimated coefficients, hence the dimension of $\boldsymbol{Q}$

Estimation

With the estimated model parameters two possible risk measures can be calculated: On the one hand side, exposure-specific risk contributions $\hat{\boldsymbol{\beta}}_{x_p}$ for a exposure x_p at lag l_p can be predicted. On the other hand, the cumulative risk measure $\hat{\beta}_c$, given a specific exposure history, can be calculated. If the exposure history is taken to be $\boldsymbol{q}_{x_p}$ with a constant exposure x_p over all lags and the matrices $\boldsymbol{R}_{x_p}$ and consequently $\boldsymbol{A}_{x_p}$ are computed as described in (3.14), the vector $\hat{\boldsymbol{\beta}}_{x_p}$ of the risk contributions and its covariance matrix $V\left(\hat{\boldsymbol{\beta}}_{x_p}\right)$ can be estimated as follows:

$$\hat{\boldsymbol{\beta}}_{x_p} = \boldsymbol{A}_{x_p} \hat{\boldsymbol{\eta}} \tag{3.16}$$

$$V\left(\hat{\boldsymbol{\beta}}_{x_p}\right) = \boldsymbol{A}_{x_p} V\left(\hat{\boldsymbol{\eta}}\right) \boldsymbol{A}_{x_p}^{\top} \tag{3.17}$$

This vector $\hat{\boldsymbol{\beta}}_{x_p}$ can be seen as a lag-response curve for a specific exposure level x_p and since it can be estimated for every exposure level from the range of x this yields, as described in Gasparrini (2014), a "bi-dimensional representation of the association". So consequently, the exposure-response curves $\hat{\boldsymbol{\beta}}_{l_p}$ along the exposures for a specific lag can be derived as well.

The cumulative risk can be computed in a very similar fashion:
By substituting the constant exposure history $\boldsymbol{q}_{x_p}$ in (3.16) and (3.17) for any arbitrarily defined exposure history $\boldsymbol{q}_h$, lag-specific risk contributions $\hat{\boldsymbol{\beta}}_h$ for every single exposure in this history are computed. These can be used to predict the cumulative risk as well as the associated covariance:

$$\hat{\beta}_c = \mathbf{1}_{v_x \cdot v_l}^\top \hat{\boldsymbol{\beta}}_h \tag{3.18}$$

$$V\left(\hat{\beta}_c\right) = \mathbf{1}_{v_x \cdot v_l}^\top V\left(\hat{\boldsymbol{\beta}}_h\right) \mathbf{1}_{v_x \cdot v_l} \tag{3.19}$$

Constraints

There are some points in which this model class has to be constrained due to identifiability issues. In this part, the practical consequences are explained briefly:

- If an intercept for $f(x)$ is included, the design matrix isn't of full rank and hence the parameters in $\boldsymbol{\eta}$ aren't identifiable. So there won't be an intercept included for the exposure-response functions in any model, which also makes sense from the biological point of view here, as an exposure of 0 wlm/year is not expected to raise the risk of dying from lung cancer
- Further constraints can be imposed by excluding an intercept from lag-response-curve, so that it is forced to start at a predicted risk of zero at the minimum of the lag period (Left constraint)
- The lag-response-curve can be forced to approach a predicted risk of zero at the maximum of the lag period (Right constraint). This can be achieved by a modification of the B-Spline bases which is introduced in the Appendix D1 in Gasparrini (2014)

A major advantage of all of the above-mentioned constraints is the possibility of specifying them without having to introduce any customized estimation methods for the estimation process Gasparrini (2014).

3.3.3. Penalized piecewise exponential additive models

The piecewise exponential approach to modelling survival data is not a novel approach itself, but an important extension is introduced by Bender et al. (2016). They have shown that this extension is specifically useful for data, where the subjects are exposed to different levels of a certain covariate at different points in time by analyzing data of critically ill patients in intensive care units with the information on their artificial nutrition. This model class is capable of incorporating the (potentially smooth) time-varying and cumulative effects of an exposure through advanced inference methods for generalized additive mixed models.
All of the formulae in this section are either directly taken from Bender et al. (2016) or are directly derived from their representations.

Theory

By partitioning the time axis into J intervals with $J + 1$ cut-points and assuming the baseline hazard $\lambda_0(t)$ to be constant in each of the J intervals, the cox ph-model from (3.2) can be transformed into a piecewise exponential model. The cut-points are chosen to be at $\kappa_0 < \kappa_1 < ... < \kappa_J$, with $\kappa_0 = 0$ and κ_J being the maximum of the follow-up period. The model equation takes the following (log-linear) form for the $j - th$ interval:

$$\log(\lambda_i(t|\boldsymbol{x}_i)) = \log(\lambda_j) + \boldsymbol{x}_i^T\beta \tag{3.20}$$

The likelihood of the model from (3.20) has been shown to be proportional to the likelihood of the following generalized linear model following a Poisson distribution (Holford (1980), Laird and Olivier (1981)):

$$\log(\mathbb{E}(y_{ij}|\boldsymbol{x}_i)) = \log(\lambda_{ij}t_{ij}) = \log(\lambda_j) + \boldsymbol{x}_i^T\beta + \log(t_{ij}) \tag{3.21}$$

with t_{ij} being the offsets, i.e. the time a subjects spends under risk in a certain interval ($t_{ij} = min(t_i - \kappa_{j-1}, \kappa_j - \kappa_{j-1})$). The proportionality of the likelihoods leads to a equivalence in the ML estimation of the model parameters which can be exploited by choosing the cut-points with respect to the temporal structure of the exposure history.
Bender et al. (2016) extend the above-mentioned GLM framework to the GAMM frame-

work in order to gain the ability to include effects which vary smoothly over time. They specify the hazard rate λ for individual i at time point t as follows:

$$\log(\lambda_i(t|\boldsymbol{x}_i, \boldsymbol{z}_i, \ell_i)) = f_0(t) + \sum_{p=1}^{P} f_p(x_i^p, t) + g(\mathcal{Z}_i(t), t) + b_{\ell_i} \tag{3.22}$$

The effects of all time-constant confounders (here: age at first exposure, calendar time) are captured by the term $\sum_{p=1}^{P} f_p(x_i^p, t)$ while the exposure-lag-response association is represented by $g(\mathcal{Z}_i(t), t)$. Additionally, b_{ℓ_i} is a Gaussian random effect for subject i.
As the two aforementioned time-constant confounders are included linearly, this part of the model equation simplifies to $\sum_{p=1}^{p} \beta_p x_i^p$.

Analogously to section 3.3.1 a time window in which the exposure history affects the hazard has so be defined. It is in this case denoted by $\mathcal{T}(j)$ and leads to the following exposure history to affect the hazard rate at time t

$$\mathcal{Z}_i(t) := \{z_i(t_e) : t_e \in \mathcal{T}_e(j)\} \tag{3.23}$$

where t_e denotes the time point at which the exposure actually occurred and consequently $z_i(t_e)$ the exposure history at t_e.
In a next step, Bender et al. (2016) specify the cumulative effects of the exposure histories $(g(\mathcal{Z}_i(t), t))$ as the integral over the partial effects $g(z_i(t_e), t)$

$$g(\mathcal{Z}_i(t_e), t) = \int_{t_e \in \mathcal{T}_e(j)} g(z_i(u), t) \mathrm{d}u \quad \approx \quad \sum_{k: t_{e,k} \in \mathcal{T}_e(j)} \Delta_k g(z_i(t_{e,k}), t) \tag{3.24}$$

or approximately as the sum over the partial effects multiplied by $\Delta_k = t_{e,k} - t_{e,k-1}$.
Furthermore, they specify partial effects to be a bivariate smooth function in t_e and t

$$g(z_i(t_e), t) = f(t_e, t) \cdot w_{ij} \tag{3.25}$$

with w_{ij} being an indicator function to show whether the exposure occurred inside of the predefined time window

$$w_{ij} = \begin{cases} z_i(t_e) & \text{if } t_e \in \mathcal{T}_e(j) \\ 0 & \text{else} \end{cases} \tag{3.26}$$

and $f(t_e, t)$ being modelled as a tensor product spline smooth:

$$f(t_e, t) = \sum_{m=1}^{M} \sum_{k=1}^{K} \gamma_{mk} B_m(t_e) B_k(t) = \sum_{m,k} \gamma_{mk} B_{mk}(t_e, t) \tag{3.27}$$

The shape of $f(t_e, t)$ is thereby controlled by the spline coefficients γ_{mk} and B_{mk} is a product of the marginal bases B_m and B_k.

Inference

Estimation and inferential procedures in the model class are based on stable likelihood-based methods introduced by Wood (2011) for penalized models.
To apply these methods, the model is represented by a sum of the model deviance $D(\boldsymbol{\gamma})$ and a penalty term controlling the smoothness:

$$D(\boldsymbol{\gamma}) + \sum_{p} \lambda_p \boldsymbol{\gamma}^\top \boldsymbol{K}_p \boldsymbol{\gamma} \tag{3.28}$$

The coefficient-vector $\boldsymbol{\gamma}$ contains all γ_{mk} and can, given the vector $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p)$, be estimated via the P-IRLS method. Convergence is ensured by basing the whole procedure on nested iterations.

4. Results for the DLMs and DLNMs

As described above, this thesis contains two different approaches to modelling the data of the wismut cohort. The focus, however, is on analyzing the data using distributed lag models and distributed lag *non-linear* models. In the process of the analysis, several different hypotheses are disposed and tested or evaluated. Furthermore, the behavior of the DLNMs under several restrictions (see section 3.3.2, Constraints) is investigated.
A model with a log-function as exposure-response function is also estimated and is compared to the optimal model in several ways. So in this chapter, in section 4.1 the different hypotheses are explained, as well as the strategy how they are to be checked. Section 4.2 contains information about the criteria that were used for selecting the best models.

4.1. Modelling strategy and Hypotheses

The algebraic notion of the estimated models looks as follows:

$$\lambda(t, \boldsymbol{\eta}_x, \boldsymbol{\eta}_z, \gamma, \delta) = \lambda_0(t) \cdot \exp\left[s_x(x, t; \boldsymbol{\eta}_x) + s_z(z, t; \boldsymbol{\eta}_z) + \gamma\, cal + \delta\, age\right] \tag{4.1}$$

In this model the hazard is modelled as a product of the baseline hazard $\lambda_0(t)$ and the exp() of the linear predictor. Of course it can also be converted to a form, where the log-hazard is modelled as a sum of the baseline hazard and the linear predictor:

$$\log\left[\lambda(t, \boldsymbol{\eta}_x, \boldsymbol{\eta}_z, \gamma, \delta)\right] = \log\left[\lambda_0(t)\right] + s_x(x, t; \boldsymbol{\eta}_x) + s_z(z, t; \boldsymbol{\eta}_z) + \gamma\, cal + \delta\, age \tag{4.2}$$

The cross-bases for radon and silica dust are represented by s_x for radon and s_z for silica dust. Apart from these two cross-bases, linear terms for the calendar time *cal* (centered

around the year 1970) and for the age at first exposure *age* are added.

DLMs

In a first step, simple distributed lag-models are applied to the data, where the exposure-response-relationship is assumed to be linear, while for the lag-response-relationship various approaches are applied. These approaches include a constant function for $w_x(\ell)$, i.e. an exposure always has the same effect no matter how long ago it occurred, a piecewise constant function, which allows differences in constant effects over several lags, and different kinds of B-Splines.
As for the B-Splines for the lag-response-relationship, different combinations of degrees of the splines (one to six) and numbers of equally spaced knots (one to five) on the quantiles of a weighted distribution[5] for the lags were analyzed. B-Splines of higher degrees (four to six) combined with a higher number of knots (three to five) were eventually considered unfeasible, because this led to highly volatile curves with several ups and downs over the lags as well as an upside at the end of the lag-response-curves in some cases. From a biological point of view, this is implausible and this is why these models weren't included in the further analysis of DLNMs.
It is obvious from the results of the study Gasparrini (2014) that this assumption of linearity is by far too restrictive, since the exposure-response-relationship has already been shown to be non-linear there. Whereas if one has a look at some of the other papers mentioned in section 3.1.2, one can see that in studies where the cumulative exposure is used as influential variable, the relationship is predominantly assumed to be linear. So the first hypothesis disposed to be falsified in this thesis is:

Hyp 1: "The exposure-response-relationship is of a linear form"

It is already partly answered by the literature, but will also be briefly addressed by an AIC-comparison of the DLMs and the DLNMs. Another reason, besides testing this hypothesis, why these simpler DLMs with a linear exposure-response relationship are fitted and analyzed anyway is to investigate, what imposing this restriction on the model does to the estimated lag-response-relationship.

[5]The weighted distribution for the lags is obtained by counting the incidence of non-zero exposure-values for each lag and weighting them with this factor. This leads to a higher weighting for the early lags, as every miner who (potentially) experienced exposures at high lags has to have experienced them at early lags as well. All of the mentioned quantiles of this distribution can be found in appendix A.3

Another crucial decision is, at which minimum lag ℓ_0 one allows the exposure to have an effect on the risk of dying from lung cancer in the model. As there will be shown in the further part of this chapter, this decision on the minimum lag has a huge impact on the results. Three different possibilities are considered here: On the one hand side, a lag of zero years is used at the very beginning. But during the process, this doesn't seem to be adequate at all, so a lag of two years like it was used in Gasparrini (2014), is also considered. Since this is not the only possible minimum lag mentioned in literature, a minimum lag of five years, like implemented by Kreuzer et al. (2010) or Walsh et al. (2010) was taken into account as well. Additionally, it was also proposed by project partner Dr. Christian Kaiser from the Institute of Radiation Protection in Munich.
So from this problem statement, one more hypothesis, with the aim to be falsified, is derived:

Hyp 2: "The minimum lag for the lag-response-relationship is zero"

Going along with this hypothesis, the aim was to determine a reasonable minimum lag. In order to do so, a lag of five years was considered as being the maximum, as no study using a higher minimum lag was found during the literature review.
For the maximum lag L at which the risk for lung cancer is possibly affected by the exposure, some restrictions are made. So in the following models it is theoretically just possible for the exposure to have an effect up to the maximum lag of 40 years, despite there is data of up to 57 years of follow-up after the beginning of employment. But one should keep in mind that at lags about 40 years or higher, there are certain problems concerning sparse data as not many miners had lived long enough to have experienced an exposure this long ago. All in all, this would lead to huge confidence intervals at high lags and also to possibly implausible courses of the estimated lag-response curves due to the data.
Another potential problem occurs due to the highly skewed distribution of the exposures, as there are many exposures very close to zero and only few very high exposures between 300 wlm/year and the maximum of 375 wlm/year. This may lead to some issues concerning the reliability of the estimated effects of these high exposures which also results in large confidence intervals at high exposure values or implausible courses of the estimated exposure-response curves.
Further hypotheses about the exact form of the exposure-lag-response relationship are tested withing the DLNM framework.

DLNMs

Concerning the modelling of the DLNMs a mostly similar proceeding (using the different variations of degrees of the B-Splines and number of knots) was applied, with the difference of the simultaneously differently modelled lag-response- and exposure-response-curves. From Gasparrini (2014), a hypothesis about the behavior of the lag-response curve at lags of 30 years or higher was derived:

Hyp 3: "The lag-response curve will approach one eventually"

To be precise, the lag-response-curve in Gasparrini (2014) approaches one at a lag of 35 years. So it is not only to be tested *whether* it approaches one, but also *when*. This leads to the forth hypothesis:

Hyp 4: "The lag-response curve will approach one at a lag of 35 years"

As the general form of the lag-response curve like it was estimated in Gasparrini (2014) with an increase up to a maximum effect at a certain lag and a steady decline with no more increases afterwards is seen to be plausible from a physiological point of view, this form is taken as given. The fifth hypothesis is concerned with the exact lag at which the maximum effect occurs:

Hyp 5: "The lag-response curve reaches its maximum at a lag of 11 years"

This hypothesis is also derived from Gasparrini (2014), as the lag-response curves of the final model in this paper reach their maximum approximately 11 years after the respective exposure occurred.
At last, an aspect concerning the exposure-response association is tested. The model from Gasparrini (2014) had a real breaking point in these curves, before which the increase was much more intense than afterwards. This breaking point was located at an exposure level of about 50 wlm/year. So the sixth and seventh hypotheses are formulated as follows:

Hyp 6: "There is a breaking point in the exposure-response association"

Hyp 7: "The break in the exposure-response curve is around 50 wlm/year"

Besides testing and evaluating these hypotheses, the aim is to get further, up to now undiscovered, insights in the exposure-lag-response association of occupational radon exposure and lung cancer mortality and check the overall compatibility of the results obtained by Gasparrini (2014).

4.2. Model selection and Diagnostics

The model selection is performed via the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), which are most commonly used for evaluating various kinds of models. These two criteria were adapted to survival analysis by Gasparrini (2014) and are used in this thesis, given by the following expression for the AIC:

$$AIC = -2 \cdot \mathcal{L}(\hat{\boldsymbol{\eta}}_x, \hat{\boldsymbol{\eta}}_z, \hat{\gamma}) + 2k \tag{4.3}$$

And respectively for the BIC:

$$BIC = -2 \cdot \mathcal{L}(\hat{\boldsymbol{\eta}}_x, \hat{\boldsymbol{\eta}}_z, \hat{\gamma}) + log(d) \cdot k \tag{4.4}$$

But there are some limitations of these criteria, which are already mentioned briefly in section 3.1.1, that have to be considered when evaluating the models. This chapter will explain the critical issues in more detail. The simulation study which was performed in Gasparrini (2014), was based on nine different scenarios for the exposure-lag-response relationship with 500 simulated data sets for each scenario. The number of subjects in this simulated data sets were either 200, 400 or 800 with a proportion of censored subjects of about 25%. Subsequently, the best-fitting model according to the AIC (or respectively BIC) was chosen out of a set of predefined models. The performance of the models with respect to the estimation of the overall cumulative effect was assessed using indices of the relative bias, the relative coverage and the relative RMSE.
This provided an insight on the limitations of the two criteria:

- AIC-selected models *generally* have a better performance
- Higher variability in AIC-selected models vs. higher bias in BIC-selected models
- The performance of AIC-based tests is not drastically affected by the sample size
- *Moderate* Overfitting for the AIC selection, leading to more flexible models
- *Severe* Underfitting for the BIC selection, leading to the selection of simpler (often also linear) models

Further details concerning the simulation study are to be found in the original paper and its appendix.

One more criterion which was applied during the analysis, pertains to the form of the exposure-lag-response surface. Like already briefly mentioned in chapter 4.1, there are some constraints from the physiological point of view. Unfortunately, the are no formal criteria or tests to check these, so that a graphical evaluation of the plotted exposure-response and lag-response curves was applied.
The exposure-response curve is subject to the expectation only to increase up to a certain level of exposure, where eventually a saturation is reached. This leads to the constraint that models with an exposure-response curve which increases everlastingly, are considered implausible and are hence discarded. Concerning the lag-response relationship, curves with wiggly courses aren't accepted. Wiggly is in this case defined as having more than one change in slope from positive to negative or vice versa. This is because from a physiological position the effect reaches a maximum at some point in time and has to decrease permanently thereafter. So if there's another increase after the estimated maximum in the lag-response curves of a model, this depicts a second change in slope and thus the model is discarded (*Criterion of wiggliness*).

4.3. Results for the DLMs

As a first approach, two different DLMs, both including a lag period from 0 to 40 years, are considered. Each of them consists of a linear function for $f(x_{t-\ell})$ for the radon exposure and does include the age at first exposure as well as the calendar time (centered around the value of 1970), but not silica dust yet, as a confounder. So these models spend 2 *df* on controlling for confounders.
The first model includes a constant function for $w_x(\ell)$, while it is chosen to be piecewise constant function with three cut-off points (at lag 10, lag 20 and lag 30) in the second model. In the following table, these two models will be referred to as "Model 1" and "Model 2". In a second step, silica dust is added to both of the two models, which will result in models named "Model 1a", and "Model 2a" respectively. The function $f(z_{t-\ell})$ for silica dust is specified as a linear threshold function, which estimates a linear effect on the log scale on the hazard ratio, if the exposure to silica dust takes values above an a priori chosen threshold and restrains it to an effect of zero for all values below this threshold. In this case, as well as for all the following models including the DLNMs, this a priori chosen threshold value is chosen to be 0.92 mg/m^3 per year (Zaballa and Eidemüller (2016)).

The lag-response function $w_z(\ell)$ for silica dust is defined as a piecewise constant function with two cut-off points at equally spaced quantiles of the distribution of the lags. The reason for this choice is the acceptable flexibility which is achieved under the condition of not spending to much *df* on a complicated modelling of the exposure-lag-response relationship of silica dust and lung cancer mortality. So the models which include silica dust, spend a total of 5 *df* on controlling for confounders.
The following graphics show the lag-response-curves of radon exposures to 50, 100, 150 and 200 wlm/year at a lag of zero years for each of the four above-mentioned models:

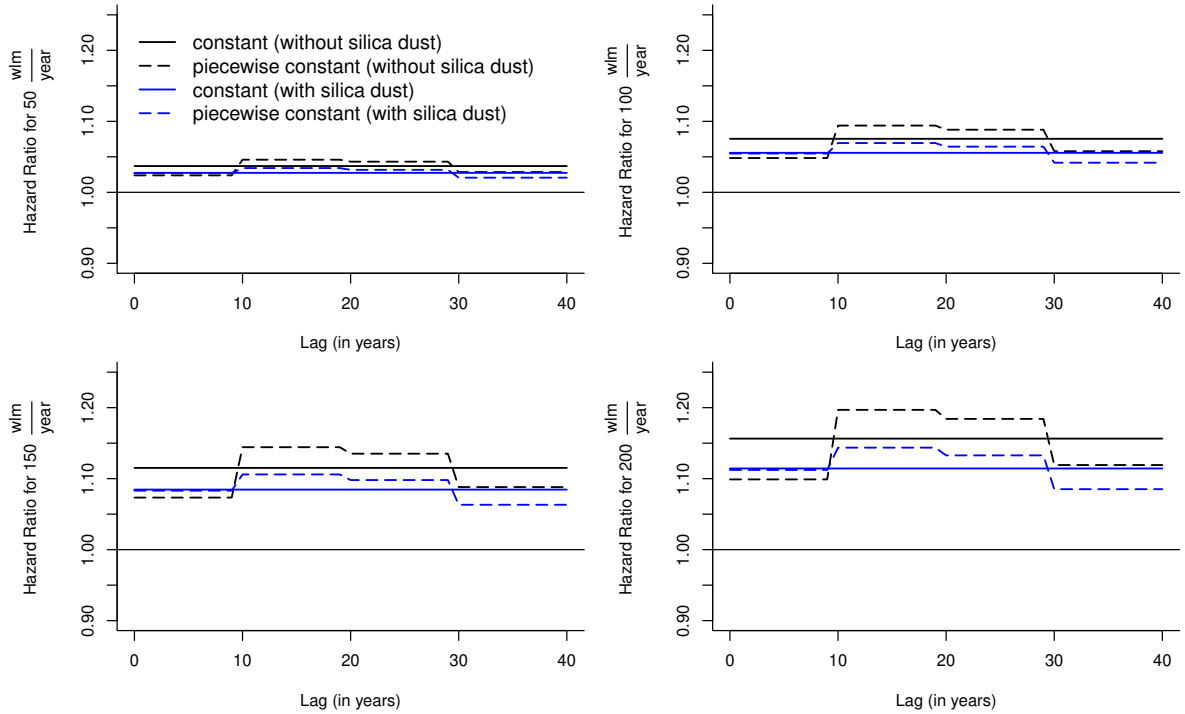

Figure 4.1.: *Comparison of the models 1, 2, 1a and 2a*

One thing that is observable in every one of the four plots in figure 4.1, is that the courses of the curves for the models 1 and 2 (black lines) are, besides the first interval from lag 0 to lag 10, consistently above those from the models 1a and 2a (blue lines). While this is the case, the general forms of the curves don't really change if the curves of model 1 and model 1a (solid lines) and respectively the curves of model 2 and model 2a (dotted lines) are compared to each other.

Another aspect that also supports the inclusion of silica dust to the model is a comparison of the AICs and BICs of the estimated models (see table 4.1).

Model	$f(x_{t-\ell})$	$w_x(\ell)$	AIC	BIC	*df*	Silica dust
Model 1	linear	constant	58427.86	58445.87	3	No
Model 1 a	linear	constant	58353.14	58389.17	6	Yes
Model 2	linear	piecewise constant	58404.23	58440.26	6	No
Model 2 a	linear	piecewise constant	58350.07	58404.11	9	Yes

Table 4.1.: Comparison of the models 1, 1a, 2 and 2a

First of all, if model 1 is compared to model 2, one observes an about 23 points lower AIC as well as an about 5 points lower BIC for model 2, which was being expected due to the more elaborate modelling of $w_x(\ell)$ in model 2. Both information criteria also indicate for both of the models that adding silica dust leads to an improvement of the fit. The AIC (BIC) of model 1a is about 75 (57) points lower than in model 1, for model 2a the AIC (BIC) reduces by about 54 (36) points compared to model 2.
A major decision that is derived from this insight is that all further models (DLMs as well as DLNMs) will include silica dust in the form of the already described threshold function. Nevertheless, one should always be aware of the fact, that there will be no further analysis on the form of the impact of silica dust, as it is just considered as confounder and the main interest is on the effect of radon exposure.
So eventually, there are two main things to take away from this:

1. When the effect of silica dust exposure on the hazard ratio for death from lung cancer is ignored, i.e. it is not added to the model through whatever kind of function, this leads to a severe overestimation of the effect of radon exposure. So it is crucial to consider silica dust as an important confounder in all of the models.

2. Despite the huge impact on the magnitude of the estimated effect of the radon exposure, adding silica dust does not seem to change the way *how* radon exposure affects the hazard ratio. This finding is justified by the observation that the forms of the curves for $w_x(\ell)$ do not change, when silica dust is added to the model.

In a second step, the aim was to determine a reasonable starting point for the lag-response-function going along with introducing a more complex form. This was achieved by combining B-Splines of degrees one to six with zero up to five knots on equally spaced quantiles of the lag-distribution. For the B-Splines with only one knot, three different knot positions were considered (33.3%-, 50%- and 66.6%-quantile). These B-Splines were all estimated with a possible intercept and including lags from zero up to 40, i.e. they

were allowed to have an effect at lag 0. This is, of course, an unrealistic assumption, but by doing this the resulting models can be examined with respect to the point, where $w_x(\ell)$ crosses the x-axis. As almost all the curves have an estimated intercept smaller than 1, the point where the curve crosses the axis can be interpreted as a hint to the time point, from where on the exposure influences the hazard ratio. Before this time point, the effect can be considered zero, because a protective effect of radon is not assumed at any time. A comparison of all the 48 models can be found in table A.4 in appendix A.4. In the last column, there is information in the minimum lag, where the estimated lag-response curves actually show a hazard ratio >1. In most of the models this is the case for a lag period of three years, but there are also some models for which it happens at a lag of two or at a lag of four. Almost none of the models show this behavior already at the lags of zero or one. So this observation strengthens the physiological arguments as well as the propositions by the project partners and the evidence from the literature, that there is no immediate increase in risk after the exposure occurs.
Due to this, **Hypothesis 2** can be pre-drawn and is rejected by this thesis.
In the next step, a similar procedure is applied, but now the minimum lag is chosen to be two years as most of the models from the prior step showed a hazard ratio >1 not until lag 3. The same model combinations as above were estimated. The results are displayed in table A.5 in appendix A.5.
When looking at the AICs of the different models, those with either a high degree of the B-Spline or a high number of knots (or both) exhibit the lowest AIC. But one thing that becomes obvious as soon as one has a look at the plotted lag-response curves, is that all models having a B-Spline of degree three or higher aren't acceptable. They all violate the *criterion of wiggliness* postulated in section 4.2, as they show several ups and downs in their curves. The BIC-selected models show a complete oppositional tendency: Among the models with the lowest BIC, solely simpler models appear. So the AIC selects the second most complex model (degree six with four knots) with 15 df, while the BIC proposes to choose the second simplest model (degree one with one knot) using just 7 df. The three figures[6] on the next page show the lag-response curves for the two above mentioned models selected by the information criteria as well as for one model which meets the *criterion of wiggliness* and simultaneously has the lowest AIC among those who else do. These models will in the further part be referred to as "Model 3", "Model 4" and "Model 5"[7].

[6]The range of the y-axis in figure 4.2 is chosen differently from all other figures on purpose, to show all ups and downs of these overly wiggly curves. In all other plots the range will be chosen equally.
[7]The exposure-response curves for model 5 are displayed in A.6

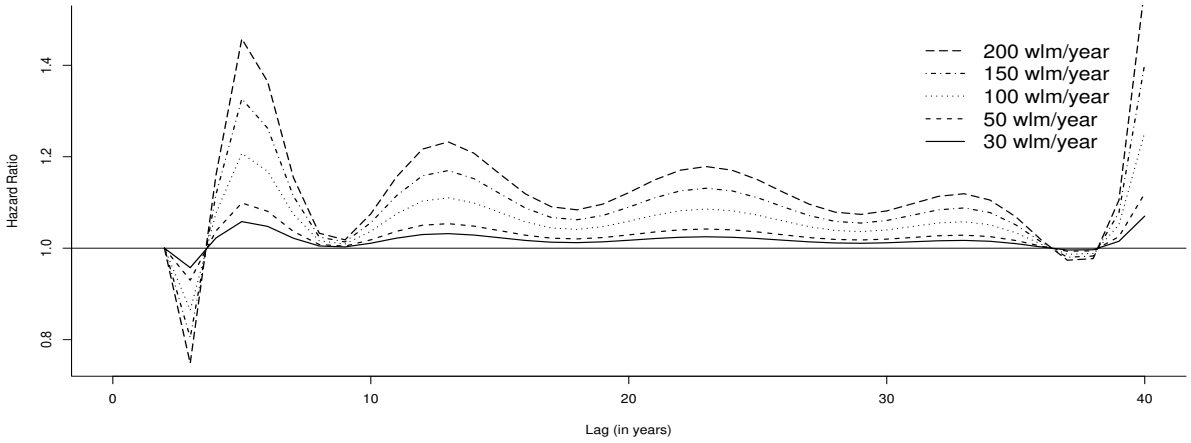

Figure 4.2.: *Lag-response curves of the AIC-selected DLM (Model 3)*

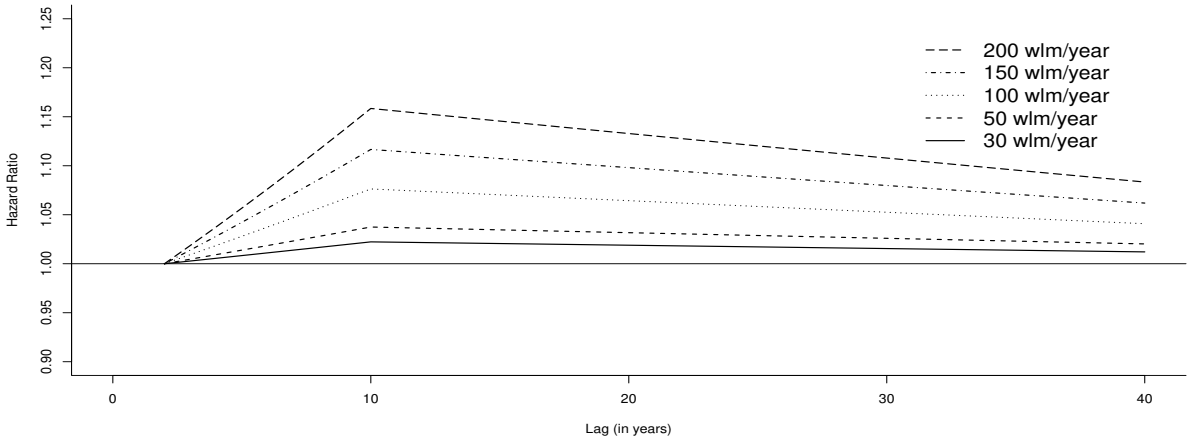

Figure 4.3.: *Lag-response curves of the BIC-selected DLM (Model 4)*

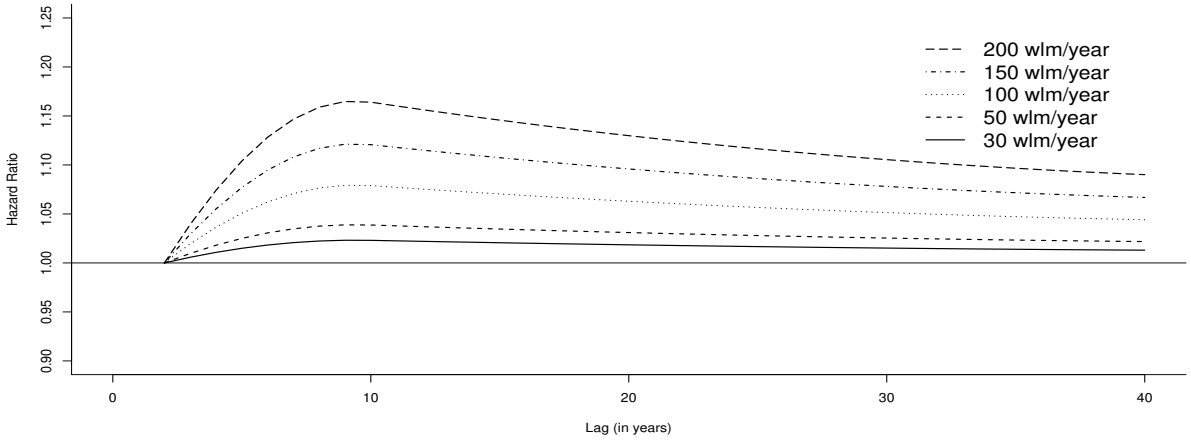

Figure 4.4.: *Lag-response curves of the AIC-best plausible DLM (Model 5)*

The following table, which is constructed similarly to table 4.1, gives an overview on the characteristics and the information criteria of the three models:

	$f(x_{t-\ell})$	$w_x(\ell)$	AIC	BIC	*df*	Silica dust
Model 3	linear	B-Spline (degree 6, 4 knots)	58333.21	58423.29	15	Yes
Model 4	linear	B-Spline (degree 1, 1 knots)	58345.03	58387.06	7	Yes
Model 5	linear	B-Spline (degree 2, 1 knots)	58345.58	58393.62	8	Yes

Table 4.2.: Comparison of the models 3, 4 and 5

The implausible AIC-selected model has an about 12 points lower AIC than the other two models, whose AICs are approximately the same. Concerning the BIC, it is much worse ranking about 36 points above the BIC-selected model and about 30 above the AIC-best plausible model.
It is also worth having a look at the estimated effects for the covariates *cal* and *age* which are included in each of the three models. Table 4.3 contains the estimates as well as the exp() of the estimates, their standard errors and their p-values, according to which both estimates are significant.

	Estimate	exp(Estimate)	Standard error	p-value
cal	-0.0184	0.9817	0.0042	0.0000
age	-0.0292	0.9712	0.0046	0.0000

Table 4.3.: Estimates of *cal* and *age* from model 5

The estimates for both of them have a negative sign and each of them is significant. The estimate for *cal* has a magnitude of -1.83%, which indicates a decreasing trend in lung cancer mortality risk over time. Additionally, with every more year of attained age at the time point of the first occupational radon exposure, the lung cancer mortality risk is estimated to decrease by about 2.88%. Theses two estimates will also be compared to the estimates from the DLNMs with respect to their magnitude and their significance. One more thing that will be compared is, whether the estimated exposure-lag-response-

relationship of lung cancer mortality and radon is of a plausible form.
For Model 5 it looks as follows:

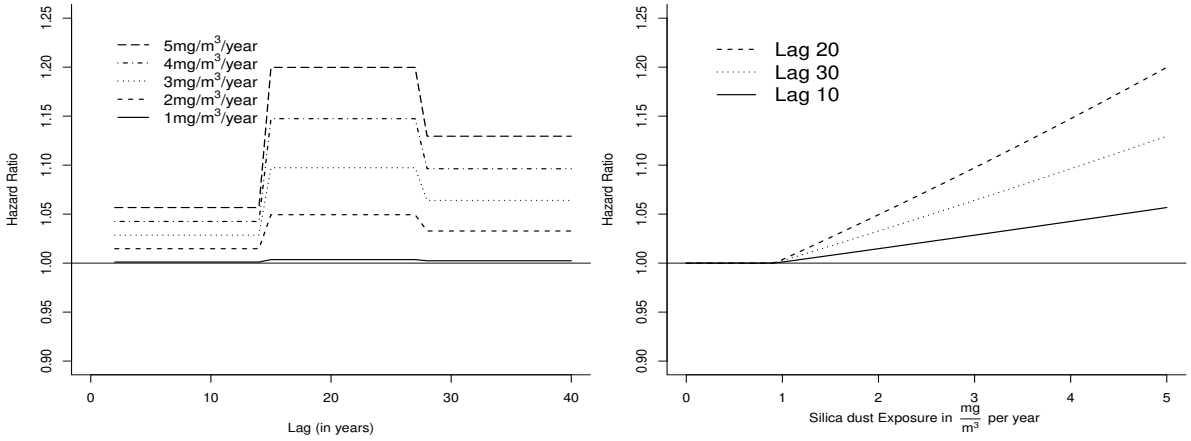

Figure 4.5.: *Lag-response and exposure-response curves for silica dust (Model 5)*

The estimated lag-response relationship of the most important confounder exhibits the highest impact between the two predefined breakpoints (lag 14.7 and lag 27.3) of the piecewise constant function. For the exposure-response relationship, the effect is zero for every silica dust exposure < 0.92 mg/m^3 per year, followed by a linear increase with different slopes for every one of the three intervals of the lag-response function. Exemplary for the three intervals (lag 2 - 14.7, lag 14.7 - 27.3, lag 27.3 - 40) the exposure-response curves were plotted at lags of 10, 20 and 30 years, as each of the three values lies within one of the intervals.

In order to validate a lag of two years as a minimum lag for the lag-response function and in order to show that a lag of five years would be inadequate, the following steps are taken:
A first model with same specifications as model 5 is estimated, with the sole difference of the inclusion of an intercept for the lag-response function. It is subsequently compared to model 5 and is also evaluated graphically. It will be referred to as "Model 5-1".
A second and a third model with the almost same specifications and the mere difference in the lag chosen as starting point for the lag-response function are estimated. One uses a lag of five years as a starting point and doesn't include an intercept ("Model 5-2"), whereas the other one does include an intercept ("Model 5-3"). These two models are compared to each other in appendix A.7 in the same fashion, as Model 5 and Model 5-1 are.

Figure 4.6 shows the lag-response curves of model 5-1 for different exposure levels:

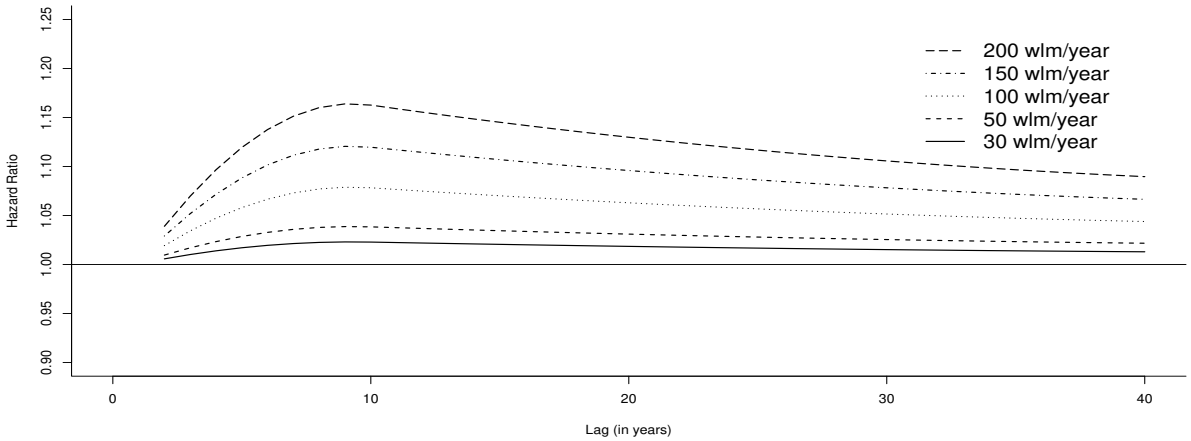

Figure 4.6.: *Lag-response curves for different radon exposures (Model 5-1)*

The overall course of these curves looks alike to that of the curves depicted in figure 4.4. But most importantly, when the estimated intercepts are looked at, they aren't estimated to be of great magnitude. This visual finding, as well as an comparison of the information criteria of the two models[8], indicates that a lag of two years is a reasonable starting point for a non-intercept model.

A comparison of the intercept model and the non-intercept model with a lag of five years as minimum lag (see appendix A.7) shows that a non-intercept would not be adequate in this case, as the estimated intercepts are of a greater magnitude. The surmise that radon exposure has a latency period of five years can hence be discarded.

The analysis of these models of the DLM class will not be pursued any further in this thesis, as the assumption of linearity does not seem to be adequate. This inadequacy is still to be shown in the further part of this chapter in section 4.4. The obtained results from this section will also be used for a detailed comparison to the models of the DLNM class in section 4.4.

[8]The AIC (BIC) of model 5-1 amounts to 58347.47 (58401.52), while the AIC (BIC) of model 5 is 58345.58 (58393.62)

4.4. Results for the DLNMs

For determining the optimal DLNM, considerably more model combinations had to be evaluated than in the simple DLM case. The huge increase in possible combinations is a result of the additional complexity due to the potentially non-linear modelling of the exposure-response function. The overall proceeding concerning the lag-response function is again similar to the DLM case, but with two differences: As it was already shown that using B-Splines of higher degrees is not really useful for modelling the lag-response association, B-Splines of degrees five and six aren't considered in this part of the analysis anymore. B-Splines of degree one are also discarded a priori as they would simplify the true underlying relationship to much. Concerning the knots and their placement, again zero to five knots on equally spaced quantiles of the distribution of the lags, with the exceptional placement for the one-knot-case, were considered (see: appendix A.3).
For the characterization of the exposure response-function from now on, B-Splines are considered as well. In this case, B-Splines of degree two, three and four in combination with zero to five knots at equally spaced quantiles of the exposure-distribution[9] are taken into consideration. In the one-knot-case, again three different positions of this knot were considered (33.3%-, 50%- and 66.6%-quantile).
Regarding the results, these are entirely different compared to the results of the DLM analysis despite some similarities concerning tendencies in AIC-selection versus BIC-selection. Starting with the differences, the AIC-selected models show clearly how the exposure-response-function is to be characterized: The 20 AIC-best models all include a B-Spline of degree two with two knots at 33.3% and the 66.6% quantiles of the exposure-distribution for this part of the association.
Within these models, there's more variation with regard to the specification of the lag-response relationship, but none of the AIC-best models contains a B-Spline with more than three knots. The two AIC-best models have to be discarded, as they violate the *criterion of wiggliness*, but the third best model (which exhibits an AIC only 2.8 points worse than the AIC-best model) satisfies the requirements concerning the wiggliness and has an AIC of about 101 points lower than the AIC-best DLM (Model 3). This model, which contains a cross-basis with B-Spline of degree two with two knots for the exposure-response function and a B-Spline of degree 2 with one knot at a lag of 20 years for the lag-response function, is selected to be the final model and will from now on be

[9] All relevant quantiles of the distribution of the radon exposure are to be found in table A.6 in appendix A.8. It is the distribution of all non-zero exposure values that appear in the exposure distribution during the period of lag 2 until lag 40.

referred to as "Model 6".
The BIC-selected models on the other hand, show the same lack of complexity as they already did in the analysis of the DLMs. Almost exclusively too simple models are selected, most of which are only estimated with a B-Spline with one knot or no knots at all. So to summarize the results of this, it can be stated that both of information criteria exhibit the expectable strengths and weaknesses.
The following three figures show the exposure-lag-response association for the selected model:

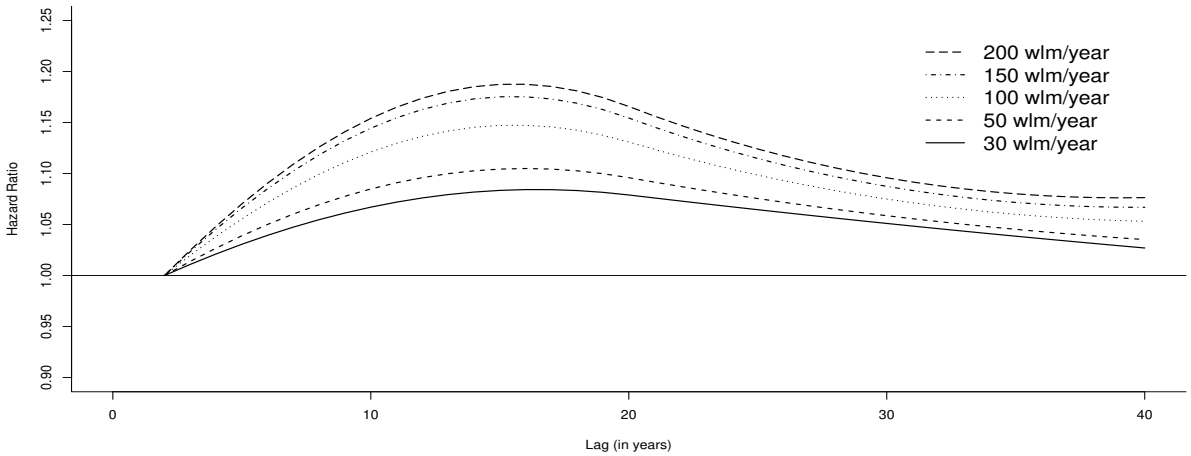

Figure 4.7.: *Lag-response curve of the final DLNM (Model 6)*

Compared to the lag-response curves from the AIC-best plausible DLM, mainly two things get obvious:
Firstly, the maximum of the hazard ratio is estimated to be much earlier for model 5, with maximum at a lag of about 9 years, than for model 6 with a maximum at a lag of about 16 years.
A second point is that the maximum as well as the decline afterwards is much more pronounced in the DLNM compared to the DLM. Model 6 exhibits a rather sharp decline after the peak and flattens out slowly at the end of the lag-period. Model 5 on the other hand, does not show this sharp decline but a rather constant process of flattening out over the rest of the lag-period.

Figure 4.8, which depicts the exposure-response curves from model 6, clearly shows the presence of non-linearity in the exposure-response association of occupational radon exposure and lung cancer mortality. This finding leads to the rejection of **Hypothesis 1**, as the DLNM also exhibits a better AIC than all of the DLMs.

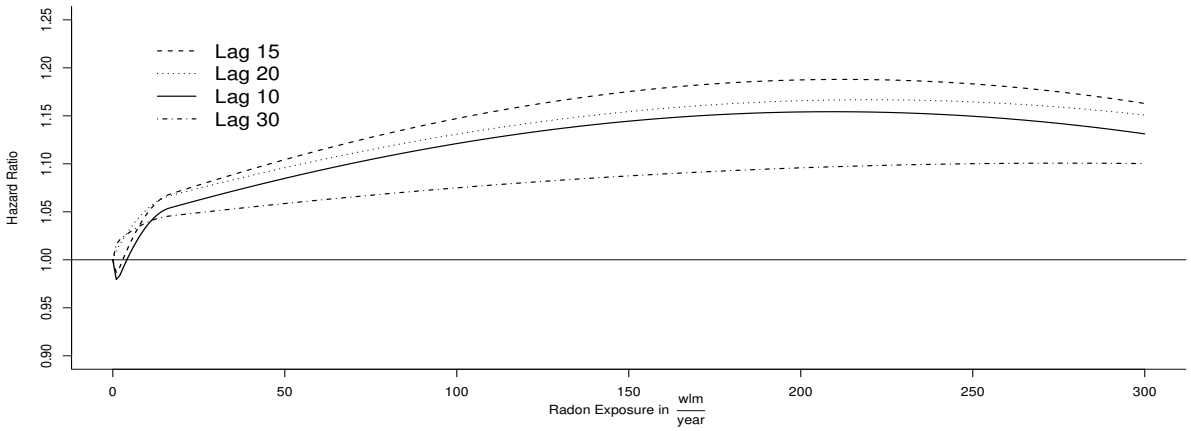

Figure 4.8.: *Exposure-response curve of the final DLNM (Model 6)*

One thing that catches the eye real quick is the break in the curves at an exposure of around 15 to 20 wlm/year. From this point onwards, the increase is a lot smaller until exposure-levels between 200 and 250 wlm/years are reached. Beyond this point there's no more increase in the hazard ratio due to a higher exposure, which can be interpreted as some kind of saturation. Another distinctive feature is the negatively estimated hazard ratio which some of the curves exhibit for very low exposures. This weird behaviour occurs for lags of up to 15 years, where the estimation of the hazard ratio for exposures < 3 wlm/year is *significantly* negative and for lags higher than 35 years where it is estimated *significantly* negative for exposures < 5 wlm/year.
The exposure-lag-response surface combines both curves to a three-dimensional surface:

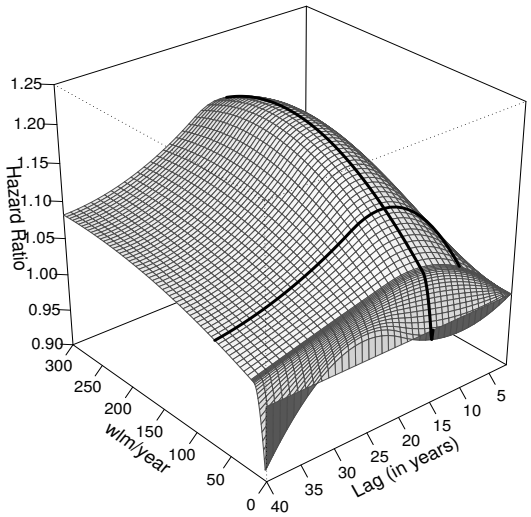

Figure 4.9.: *Exposure-lag-response surface of the final DLNM (Model 6)*

The two solid lines along the surface are a lag-response curve and an exposure-response curve, plotted exemplarily at a lag of 15 years and for an exposure of 70 wlm/year respectively. The choice of these two values, however, is arbitrary and is in this case just shown as an illustration.
Estimates for the two confounders *cal* and *age* are displayed in table 4.4 while the estimated exposure-lag-response association for silica dust is to be found in figure 4.10:

	Estimate	exp(Estimate)	Standard error	p-value
cal	-0.0101	0.9900	0.0044	0.0233
age	-0.0231	0.9772	0.0048	0.0000

Table 4.4.: Estimates of *cal* and *age* from model 6

The estimate for *cal* (-1.00%) indicates a decreasing trend over time, while the estimate for *age* of -2.28% indicates a decreasing lung cancer mortality risk for every more year of attained age at first exposure. Both of them are significant on the 5%-level. If these two estimates are compared to those from the AIC-best plausible DLM (*cal*: -1.83%; *age*: -2.88%), they are estimated 0.83 percentage points, and 0.60 percentage points respectively, smaller. This gives the impression that there might be a tendency towards overestimation of other covariates' effect if the exposure-response association is misspecified as linear.

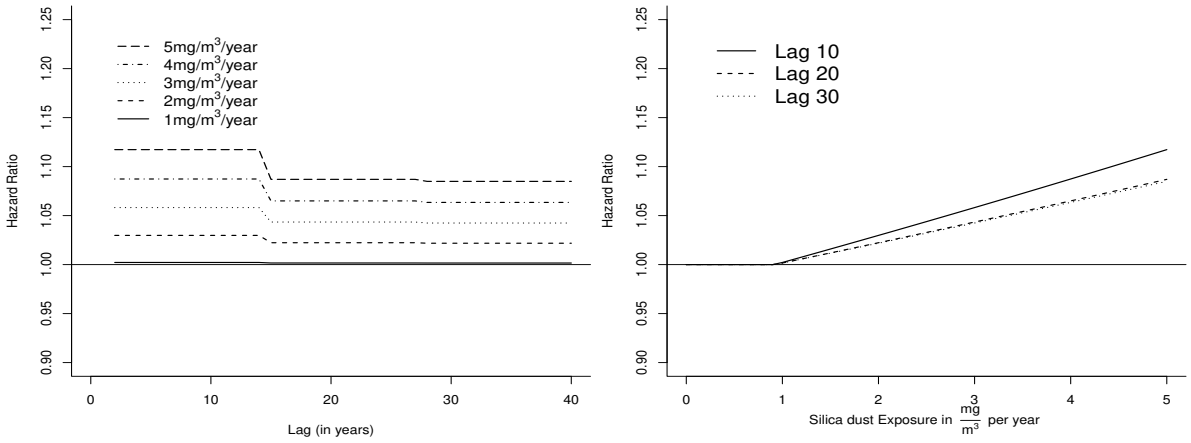

Figure 4.10.: *Lag-response and exposure-response curves for silica dust (Model 6)*

Silica dust exposure is estimated to have the highest effect in the first interval of the

piecewise constant function (lag 2 to 14.7) with a substantially lower estimated effect for the other two intervals. At the second breakpoint, there's merely a very small drop, so from lag 14.7 to lag 40 the lag-response curves remain almost constant. This can also be observed when looking at the exposure-response curves (which are again plotted exemplarily for the three intervals at lags of 10, 20 and 30 years), as those curves for lag 20 and lag 30 only differ marginally. The exposure-response curve plotted at a lag of 10 years however is taking its course clearly above the two aforementioned.
The impression that other covariates' effects are overestimated in the DLM is also supported by these figures, as the course of the lag-response curves is below the course of those from Model 5 and the slope of the exposure-response curves is not as steep (see again: figure 4.5). Further insights on the consequences of the misspecification are obtained by comparing model 3 (AIC-best DLM) and model 5 (AIC-best plausible DLM) to model 6 with respect to AIC, BIC and the *df*:

	$f(x_{t-\ell})$	$w_x(\ell)$	AIC	BIC	df
Model 3	linear	B-Spline (6, 4 knots)	58333.21	58423.29	15
Model 5	linear	B-Spline (2, 1 knot)	58345.58	58393.62	8
Model 6	B-Spline (2, 2 knots)	B-Spline (2, 1 knot)	58232.44	58334.52	17

Table 4.5.: Comparison of the AIC-selected DLM and the AIC-best plausible DLM to the AIC-selected DLNM

As already mentioned before, the AIC of model 6 is considerably lower compared to the AIC of model 3 (about 101 points) and obviously also compared to the AIC of model 5 (about 113 points). Even the BIC shows the same selection tendency as the AIC, as it also prefers model 6 over the two DLMs.
One more aspect worth mentioning is the similarity of model 3 and model 6 concerning the *df*. The complexity of model 3 with 15 *df* is much closer to the complexity of model 6 with 17 *df*, than if model 5 (7 *df*) is compared to model 6 with respect to model complexity. So in case of misspecification of $f(x_{t-\ell})$, the AIC seems to compensate this inadequacy by selecting models with an overly complex lag-response function, so that the overall complexity of the model is equal to the underlying, true complexity.
This is a very important point to take away from this section, as it gives a pretty good understanding of the severe consequences of misspecification of one part of the cross-

basis in the DLNM framework.
In a last step, in order to reassure ourselves that the minimum lag of 2 years (which was adopted from the results of the DLMs) is still adequate, "Model 6-1" is estimated. The only difference between model 6 and model 6-1 is the intercept included in model 6-1. A comparison of the AICs of these two models shows the superiority of model 6 to model 6-1, as model 6-1 exhibits an AIC of 58234.52 which is about 2 points higher than the AIC of model 6 (compare table 4.5). Besides the information criterion, the following graphical representation of the lag-response curves of model 6-1 also indicates that it might not be utterly necessary to include the intercept:

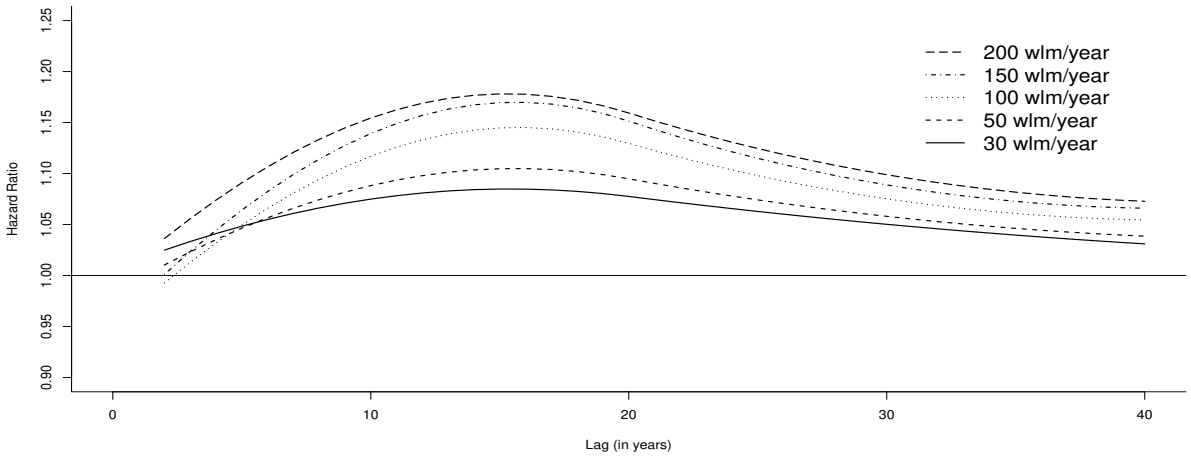

Figure 4.11.: *Lag-response curves for the DLNM with a minimum lag of two years including an intercept (Model 6-1)*

As the overall shape of the curves does not change compared to figure 4.7 and the intercepts are estimated to be rather small, this shows the expendability of an intercept.

If one now reclaims the hypotheses concerning the DLNMs postulated in 4.1, the results from this section lead to the following decisions about them:
Hypothesis 3 and consequently **Hypothesis 4**, which were both derived from the results published by Gasparrini (2014), are rejected.
In the lag-response curves shown in figure 4.7, the hazard ratio doesn't approach a value of one at any time inside the lag-period of two to forty years (except from the starting point at a lag of two years). As **Hypothesis 4** was just concerned with the exact lag where the lag-response curves would potentially have approached one, it is of no relevance anymore. In the further process of this chapter, an alternative model, which imposes a right-constraint on the model like it was explained in section 3.3.2, will be

presented, evaluated and compared to model 6.

Hypothesis 5 is concerned with the maximum of the lag-response curves, which Gasparrini (2014) found out to be at around 11 years after the initial exposure. This finding could not be verified either, as the maximum of the lag-response curves is estimated to appear at a lag of 16 years in model 6. So consequently, this hypothesis is rejected as well.

The last two hypotheses, **Hypothesis 6** and **Hypothesis 7** are aimed at the form of the exposure-response curves. The former of this two hypotheses cannot be rejected because the exposure-response curves of model 6 shown in figure 4.8 show indeed some kind of disruption at a certain exposure-level.

In a final step, **Hypothesis 7** however is rejected after a qualitative evaluation, as this aforementioned disruptive point in the exposure-response curves of model 6 is to be located clearly before the exposure value of 50 wlm/year postulated by the hypothesis.

Alternative Models

Additionally to the finally selected DLNM (Model 6), two other models were considered to be interesting and were thus investigated as well. To be precise, these additional models include:

(a) A model with a user-defined log-function for $f(x_{t-\ell})$ ("Model 7")

(b) A model where the right-constraint, mentioned in section 3.3.2, is applied to the B-Spline of $w_x(\ell)$ ("Model 8")

The following table shows a comparison of the two alternative models to model 6, with respect to the information criteria and the complexity of the models:

	$f(x_{t-\ell})$	$w_x(\ell)$	AIC	BIC	df
Model 7	Log-function	B-Spline (2, 1 knot)	58235.82	58283.86	8
Model 8	B-Spline (2, 2 knots)	RC-B-Spline (2, 1 knot)	58295.72	58349.77	9
Model 6	B-Spline (2, 2 knots)	B-Spline (2, 1 knot)	58232.44	58334.52	17

Table 4.6.: Comparison of the two alternative models to the AIC-selected DLNM

While model 7, which uses a log-function for the exposure-response function, exhibits an AIC only slightly higher than the final model, model 8 is by far worse according to the AIC. This supports the previous findings that the lag-response curves actually do not approach a hazard ratio of one at the and of the lag-period and that restraining them to do so leads to a worse fit.
What is interesting, however, is the good performance of the simplistic modelling of the exposure-response relationship as a log-function relative to model 6. Model 8 spends much less *df* than model 6 (8 vs. 17), but nevertheless comes close to its AIC value. One thing that is problematic about this model, is that no saturation of the exposure-response curve is reached due to the form of the log-function. But still, this is an important option that should be kept in mind when modelling these associations.
Graphical representations of the exposure-lag-response associations of these two alternatively estimated models are to be found in appendix A.9.

Predictions

It is not only possible to calculate the increase in hazard ratio due to a single exposure at a certain lag, but also to predict the *cumulative* hazard ratio induced by a certain exposure history. Such predictions will be calculated exemplarily for artificially created exposure histories in order to compare different models with respect to their predictive quality. For four of these six exposure histories, an exposure of 20 wlm/year each year over a period of ten years is assumed. This applies to either the last 10 years or 10 to 19 (20 to 29, 30 to 39) years ago. For the fifth exposure history, a yearly exposure to 100 wlm/years for the last 10 years is considered, while for the sixth exposure history a yearly exposure of again 20 wlm/year over a period of the last 20 years is examined.
Cumulative hazard ratios due to these exposure histories are calculated for model 5 and model 6 as well as for the two alternative models, model 7 and model 8. The results of these predictions are presented in table 4.7 below. All these estimated values are to be interpreted in comparison two an exposure history which only includes zero, i.e. the exposure history of a person who was not exposed at all.
In a first step, when concentrating on the comparison of the predictions of model 5 and model 6 (DLM versus DLNM), the following findings can be derived:
In general, the estimated cumulative hazard ratios are considerably higher in model 6. If only the first four exposure histories are considered, both of the regarded models exhibit

the highest estimated cumulative hazard ratio for an exposure of 20 wlm/year 10 to 19 years ago. For model 5, there's a relatively modest decline if the exposure is further in the past whereas this decline is more pronounced for model 6.
In a comparison based on the fifth history however, the estimate for model 5 is much more proximate to the one from model 6. For the last exposure history, there's again a big difference between the predictions of the two models, as model 6 predicts a much higher value for the cumulative hazard ration than model 5.

	Model 5	Model 6	Model 7	Model 8
20 $\frac{wlm}{year}$ (last 10 years)	1.08 (1.05-1.11)	1.25 (1.14-1.37)	1.21 (1.17-1.27)	1.33 (1.25-1.41)
20 $\frac{wlm}{year}$ (10-19 years ago)	1.15 (1.12-1.18)	1.94 (1.61-2.35)	1.74 (1.59-1.91)	2.09 (1.78-2.45)
20 $\frac{wlm}{year}$ (20-29 years ago)	1.12 (1.10-1.14)	1.79 (1.53-2.09)	1.55 (1.46-1.65)	1.46 (1.34-1.58)
20 $\frac{wlm}{year}$ (30-39 years ago)	1.10 (1.07-1.12)	1.43 (1.19-1.71)	1.39 (1.30-1.49)	1.06 (1.05-1.07)
100 $\frac{wlm}{year}$ (last 10 years)	1.47 (1.28-1.68)	1.59 (1.40-1.81)	1.53 (1.40-1.67)	1.67 (1.51-1.86)
20 $\frac{wlm}{year}$ (last 20 years)	1.24 (1.18-1.31)	2.43 (1.84-3.20)	2.11 (1.85-2.42)	2.77 (2.22-3.46)

Table 4.7.: Prediction of the cumulative hazard ratios (95%-CIs in brackets) for different predefined exposure histories in different models

If the other two models are taken into consideration and are compared to model 6, different insights are gained:
When at first only the exposure histories, which only exhibit exposures in the current past (i.e. the histories 1, 2, 5 and 6) model 7 predicts consistently *lower* cumulative hazard ratios than model 6 while the predictions obtained from model 8 are consistently *higher*. Concerning the exposure histories which possess exposures further in the past (i.e. the histories 3 and 4), other observations are made. For these histories, model 6 predicts the highest cumulative hazard ratios of the three models. The predictions of model 7 are in this case much more proximate to those from model 6 than the predictions from model 8, which might be due to the right-constraint imposed on the lag-response function in model 8.

5. Results for the PAMs

This chapter contains the results which were obtained by modelling the data within the framework of the piece-wise exponential additive models.
The main objective is to compare the predictive abilities of this framework to those of the DLNM framework and to spot potential differences. In order to avoid going beyond the scope of this whole thesis, the model will be kept relatively simplistic:
Firstly, the model equation is be specified (see equation (5.1)) and its single pieces are explained subsequently.
Secondly, the model is to be estimated which contains, besides the occupational radon exposure, only the age at first exposure as well as the calendar time (again centered around the year 1970) as covariates. Silica dust exposure is not included, as the first aim here is just to check whether the predictions of these two frameworks are comparable to each other for the simple case of only one exposure-lag-response association. So in order to compare it to the DLNM framework, another DLNM, containing the same covariates as the PAM, is estimated in this chapter ("Model 9").

The model formula of the PAM is specified as follows:

$$\begin{aligned} \log(\lambda_i(t|\boldsymbol{X}_i, \boldsymbol{x_{rad}}_i)) \quad &= \quad f_0(\tilde{t}) + \beta_{cal} x_i^{cal} + \beta_{age} x_i^{age} \\ &\quad + \sum_{t_e \in \mathcal{T}_e(j)} g(x_{rad}(t_e), \tilde{t}), \end{aligned} \tag{5.1}$$

where $f_0(\tilde{t})$ is the baseline hazard, which is estimated of the midpoints $\tilde{t}$ of the single time intervals. The linear effect of the two confounders calendar time and age at first exposure are incorporated by the terms $\beta_{cal} x_i^{cal}$ and $\beta_{age} x_i^{age}$, while $\sum_{t_e \in \mathcal{T}_e(j)} g(x_{rad}(t_e), \tilde{t})$ represents the cumulative effect of the radon exposure. This exposure-lag response association is estimated by a tensor product spline with the marginal bases being P-Splines (Eilers and Marx (1996)) of degree two with second order difference penalties. The degree of the splines was chosen to be two (and not the default of cubic splines) to be most

alike to the B-Splines of second degree used in the DLNMs[10].
Furthermore, the estimation of the model was performed by the `bam`-function (Wood et al. (2015)) from the `mgcv`-package (Wood (2014), version 1.8-12). This decision was made based on the expectedly high computational effort due to the large data set. During the estimation process, there were no problems with the convergence of the models.

The results are again displayed similar to the anterior chapters: Figure 5.1 shows the lag-response curves predicted from the results of the PAM in the top panel. In the bottom panel, the lag-response curves of model 9 are depicted for the reasons of comparison. Subsequently, in table 5.1, the estimates of the two covariates *cal* and *age* are presented in the same fashion as in table 4.4.

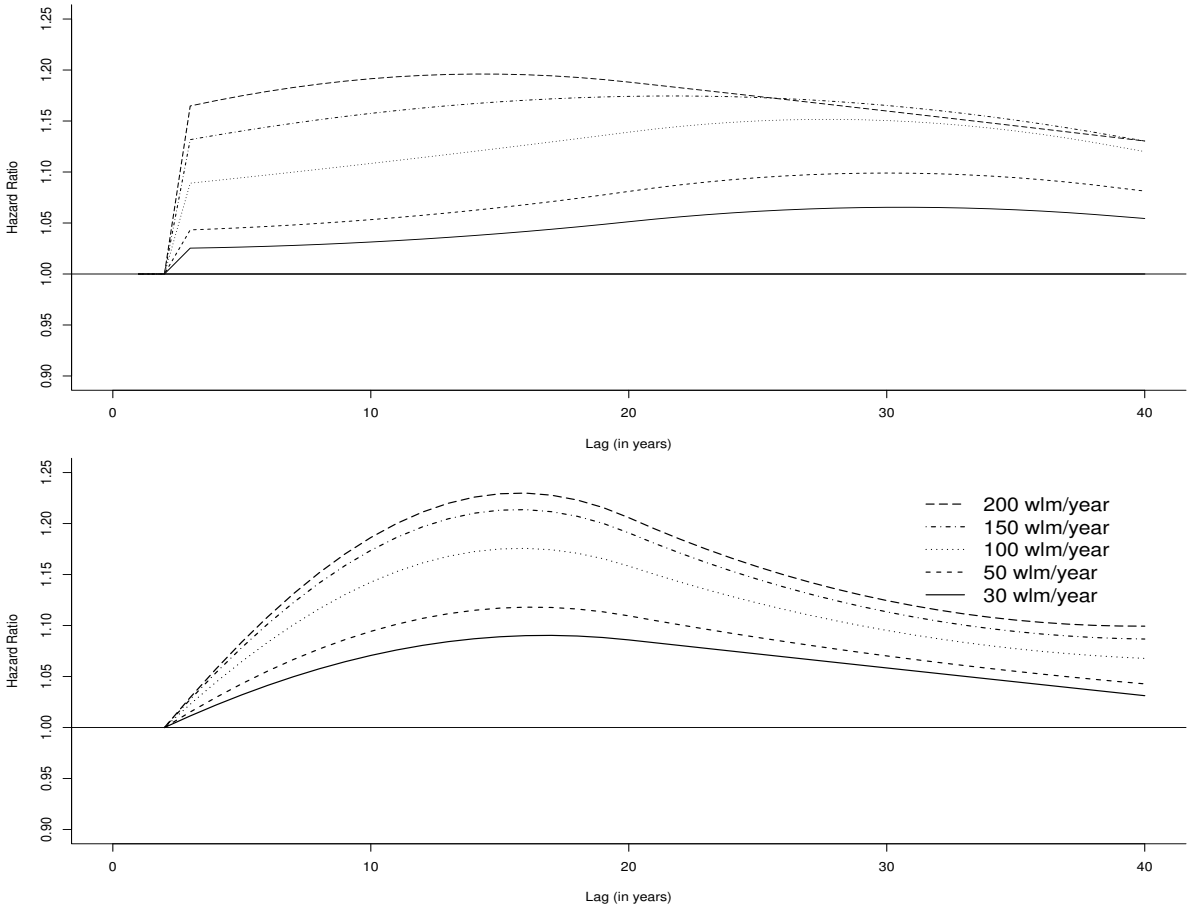

Figure 5.1.: *Lag-response curves of the PAM (top) and of model 9 (bottom)*

[10] Altogether, four different alternatives were considered: P-Splines of degree two and three combined with second order difference penalties and ridge penalties. A comparison of the AICs indicated the chosen model to be slightly better than the other options.

A comparison of the curves of the two models clearly shows some key differences.
Firstly, the PAM predicts the exposure to have an immediate, strong effect on the hazard ratio, as soon as it enters the predefined time window $\mathcal{T}_e(j)$ in which it is allowed to affect the hazard. The time window $\mathcal{T}_e(j)$ for the PAMs naturally chosen equally to the lag-period of the DLNMs to last from lag 2 to lag 40.
This immediate, strong effect is expressed by a steep increase in the curves for all different exposure-levels, followed by a very moderate further increase up to a maximum, which is different for the curves. The maximum of the effect on the hazard ratio tends to be earlier, the higher the initial exposure is and consequently later the lower it is. Thereafter, there's a smooth decline in all of them again.
So in a rough overall description, the curves sound kind of related to those from the DLNMs but with the key differences of the steep increase at the very beginning and the varying location of the maximum within the different curves.
One more thing to add when comparing the curves of model 9 to those of model 6 from figure 4.7: The curves of model 9 (where silica dust is not included as a confounder) are (a) for all exposure-levels consistently above them of model 6 and (b) of the same overall form like those of model 6. This confirms the findings from section 4.3, where the inclusion of silica dust as a confounder was justified by the same observations that were made in the simple DLMs.

	Estimate	exp(Estimate)	Standard error	p-value
cal	-0.0184	0.9818	0.0043	0.0000
age	0.0795	1.0827	0.0022	0.0000

Table 5.1.: Estimates of *cal* and *age* from the PAM

The inspection of table 5.1 reveals further differences between the two frameworks. While the estimate for the calendar time *cal* has at least the same sign as its estimate in model 9 has, they differ considerably concerning the magnitude. In the PAM, the effect is estimated to be -1.82%, whereas the estimate of model 9 is only 1.21%[11]. Somewhat more disturbing are the different signs of the estimates for the age at first exposure.
For the PAM a positive and very large effect of 8.27% is estimated, while for model 9 a relatively small negative effect of -2.31% is estimated.

[11]The estimates from model 9 are to be found in table A.7 in appendix A.10

6. Discussion

This chapter will mainly consist of three different topics: Firstly the differences in the results of the two different modelling frameworks have to be addressed. Secondly, single aspects of the DLNMs have to be discussed and thirdly advantages as well as limitations of the data are presented.

DLNMs versus PAMs

In this work two relatively novel modelling frameworks, which are still subject to research, are employed. So consequently, there are some aspects concerning the analysis, that have to be discussed.
The first point to be addressed here, are the different stages of technical development in which the two frameworks are currently embedded. As already mentioned before, the estimation of the PAMs relies on the stable routines of the `mgcv`-package and offers thus a variety of different splines which are implemented in the `gam`- and the `bam`-function. The estimation of the DLNMs, however, is performed via the `survival`- and the `dlnm`-package, which (in case of the `dlnm`-package) is a rather recently developed package compared to `mgcv`. So the `crossbasis`-function from `dlnm` does not offer a comparable variety of different splines (yet), as the `gam`-/`bam`-function does. But nevertheless, this is still a current topic of research to be worked on, as can be seen in the work of Gasparrini et al. (2016), who introduce P-Splines within the framework of DLNMs.
A second point is the difference in the estimation methods. DLNMs are estimated via the maximization of the partial likelihood in the cox ph-model (as explained in section 3.2), while the PAMs are estimated via a REML approach. This may also be a reason for potential differences.

Discussion of the DLNMs

One obviously criticizable point of the DLNMs is the arbitrary choice for specification of how to include silica dust exposure in the models. But note, that the aim of this work is not to characterize in exposure-lag-response association of silica dust exposure and lung cancer mortality but that of radon exposure and lung cancer mortality.
So the rationale for not digging deeper into this association and for specifying it a priori, is not to let the models become overly complex and to concentrate on the effect of radon exposure. This approach orientates itself by the course of action Gasparrini (2014) took, when he also specified the cross-basis for his most important confounder (Smoking) a priori and relatively simplistic.
Another aspect that might potentially become the target of criticism is the choice of the minimum lag. But since formal methods and tests for the determination of the minimum lag don't exist, the choice is kind of tricky and may always appear somewhat arbitrary. In order to debilitate potential criticism, different minimum lags used in literature (two and five years) as well as a lag of zero years were taken into account. After detailed comparisons, a minimum lag of two years seemed to be the adequate choice.

Data

Besides the technical details concerning the model, also some critical points on behalf of the wismut data are to be discussed.
First of all, as already mentioned in section 2.3, there is no information on the smoking status or the smoking habits of the miners in this data set. This point might depict a severe limitation of the data compared to other uranium miner cohort data sets and the studies based on them. But as it is not possible to get this information in some way, other variables have to be used as confounders in models fitted to this data set.
Secondly, the estimation of the job-exposure matrices of the different substances, but primarily that for radon exposure might have to be reviewed critically. The table below, which is taken from Kreuzer et al. (2011), page 22, shows how the exposure measurements actually were conducted:
For the whole time period from 1946-1955, which is included in this analysis and in which some of the highest exposures are to be found, no measurements were taken at all. All of the values of the JEM for this period were estimated retrospectively, based

"on the yearly production of ore, its uranium content, shaft geometry, techniques of uranium ore production and ventilation" (Kreuzer et al. (2011)). In the following years, partial measurements (1955-1965) and eventually regular measurements of radon and its progeny (from 1966 onwards) were conducted, but it was not until 1971 that individual monitoring was introduced.

Year	Measurement
Until 1954	No measurements
From 1955	Partial measurements of gamma exposure rates in a few objects
1955 – 1965	Partial measurements of radon in a few objects
1964 – 1965	Partial measurements of radon progeny (RnFP)
From 1966	Regular measurements of radon and its progeny
From 1967	Measurements of long-lived radionuclides (LRN)
From 1971	Individual monitoring of exposure

Figure 6.1.: *Exposure measurements in the wismut cohort (Kreuzer et al. (2011))*

The estimation of the JEM, as well as potential contained measurement or estimation errors, are topics that are still a subject of current research and that's why for this thesis, this data is taken as given.

But despite these uncertainties and weaknesses, the data has undoubtedly many strengths. Concerning the sample size, it's the world's largest existing cohort data set of uranium miners which gives the results a huge relevance. Furthermore, there's not only information on a person's cumulative exposure, but it's split up by year and is presented in a detailed JEM.
The third point to be mentioned here, is the vast amount of information on the exposure to other potentially harmful substances. This allows, as it is done with silica dust in this case, to take them into account as confounders as well.

So in order to summarize these points, one can state that a data set of this surpassing size gives researchers a unique possibility to gain further insight in the association between occupational (radon) exposure and lung cancer mortality.

7. Conclusion

After taking everything into consideration, the following points concerning the postulated hypothesis can be concluded:

1. Hypothesis 1 is rejected. The exposure-response relationship has been shown to be non-linear via a model comparison of the DLMs from section 4.3 and the DLNMs from section 4.4. The AIC of the best DLNMs is about 100 points lower than for the best DLMs and additionally more plausible models are selected via the information criteria in the DLNM framework.

2. Hypothesis 2 is rejected. Already in the DLM framework it was shown that the potential effect at lag zero would be estimated to be highly negative, which can't be the case. The literature review in the sections 3.1.1 and 3.1.2 also showed that none of the up to now published studies mentions such an effect. Additional to the pure rejection, an alternative minimum lag of 2 years has emerged through the analysis of this thesis.

3. Hypothesis 3 is rejected after a graphical evaluation of the lag-response curves. None of the curves of the selected DLNMs showed this behaviour. It could be artificially enforced by right-constraining the models as mentioned in section 3.3.2. This was also attempted by modifying the B-Spline basis of the lag-response function in a way described by Gasparrini (2014), but this didn't lead to an improvement regarding the AIC (see table 4.6 and figure A.9).

4. Hypothesis 4 is consequently rejected as well, as it was concerned with the exact lag where the lag-response-curve would potentially approach a hazard ratio of one. As this isn't the case at all for the selected model, this issue does not have to be discussed any further.

5. The decision about hypothesis 5 is a bit more tricky, as some of the models which were estimated during the process of finding the AIC-best DLNM peaked at earlier lags. There's the impression that the location of the maximum in the lag-response function depends strongly on the knot placement within the B-Splines. But it is ultimately rejected as the model, which was found to be the best model among the investigated models, has its peak at a lag of around 16 years after the initial exposure.

6. Hypothesis 6, which states that there exists some kind of breaking point in the exposure-response curves, can be confirmed by the results of this thesis.

7. Hypothesis 7 however is again rejected, as the graphical evaluation of the exposure-response curves shows that the location of the breaking point is at a considerably lower exposure level than postulated in the hypothesis.

One important thing to add at this point, is that a considerable part of the hypotheses could not be tested in a statistical sense of applying some kind of formal test, as no formal tests for these contexts exist. This remark applies primarily to the hypotheses five, six and seven. It partly applies to hypothesis number two, as an intercept model can be formally tested against a non-intercept model with the same minimum lag but it is difficult to assess the minimum lag via some kind of test.

Concerning the comparison of the two different modelling frameworks, the following can be stated:
The analysis of the two different models in chapter 5 reveals surprisingly big differences. These differences do not only concern the subject of primary interest, the effect of the occupational radon exposure, but also the estimates of the other covariates in the model. Thus, further investigations of this topic have to be done.

A. Appendix

A.1. Mean exposure to long lived radionuclides, gamma radiation, arsenic and fine dust

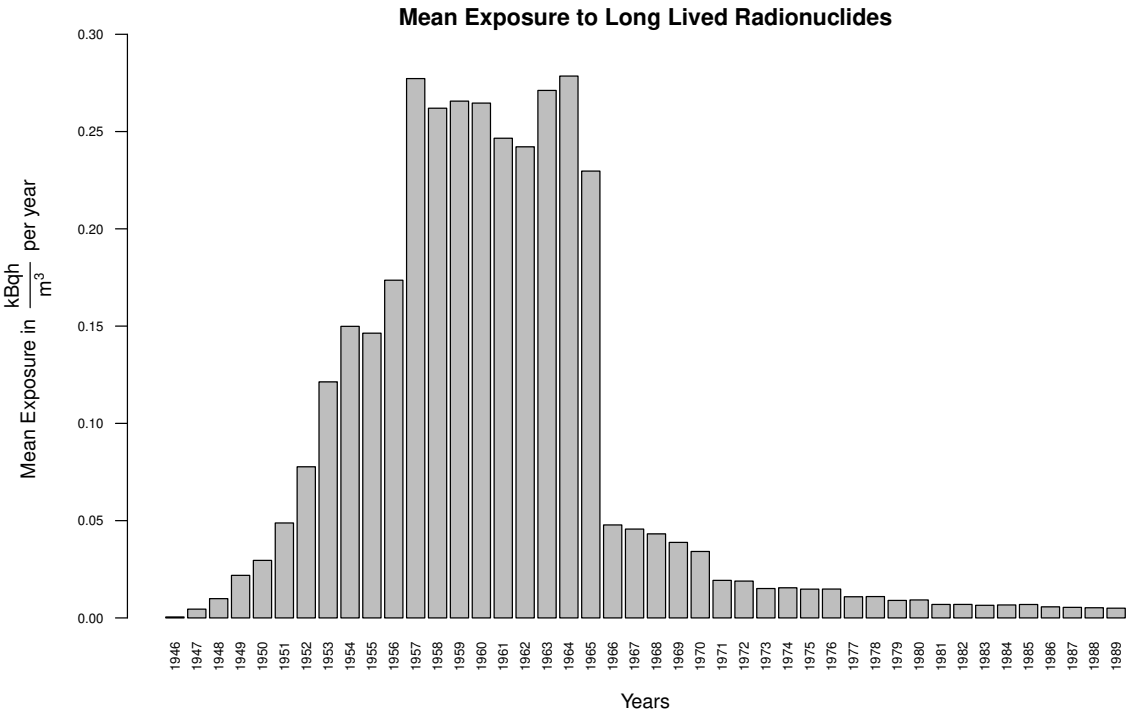

Figure A.1.: *Mean Exposure to Long lives radionuclides in the Wismut Cohort (1946-1989)*

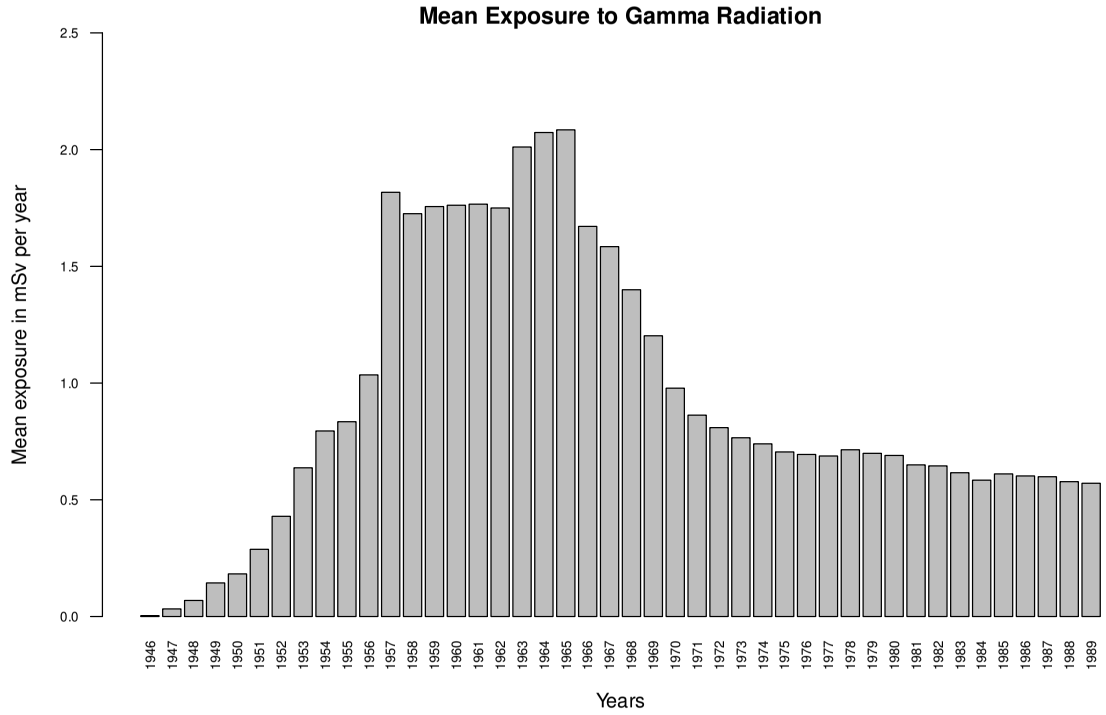

Figure A.2.: *Mean Exposure to Gamma radiation in the Wismut Cohort (1946-1989)*

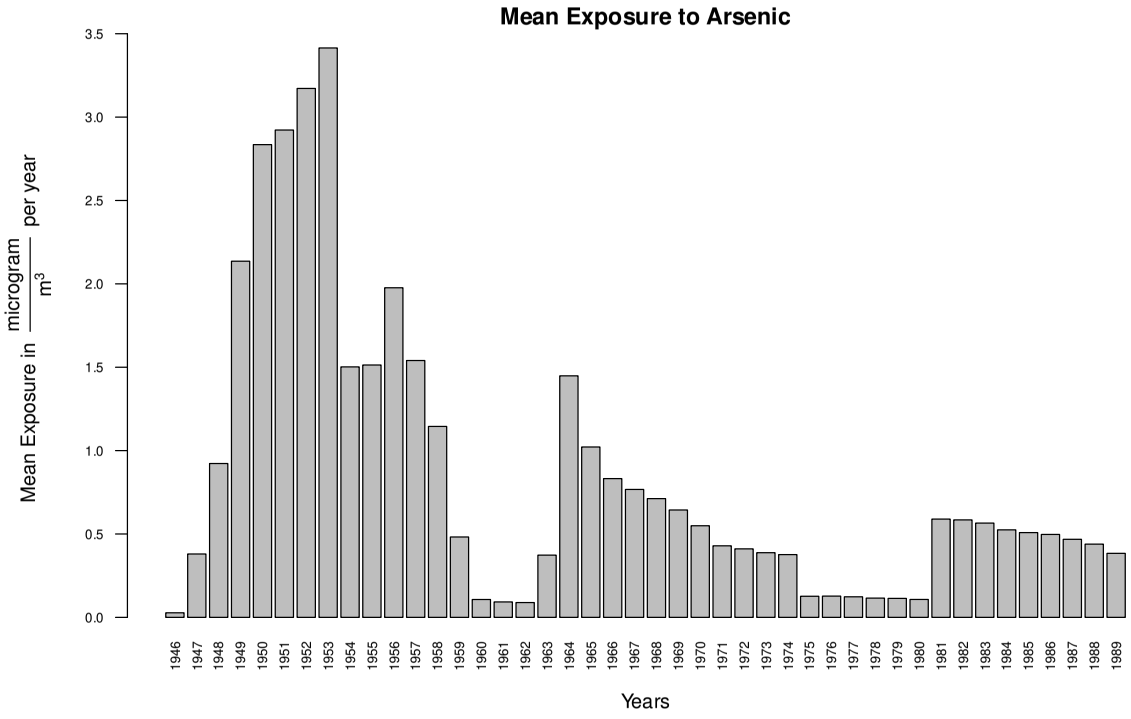

Figure A.3.: *Mean Exposure to Arsenic in the Wismut Cohort (1946-1989)*

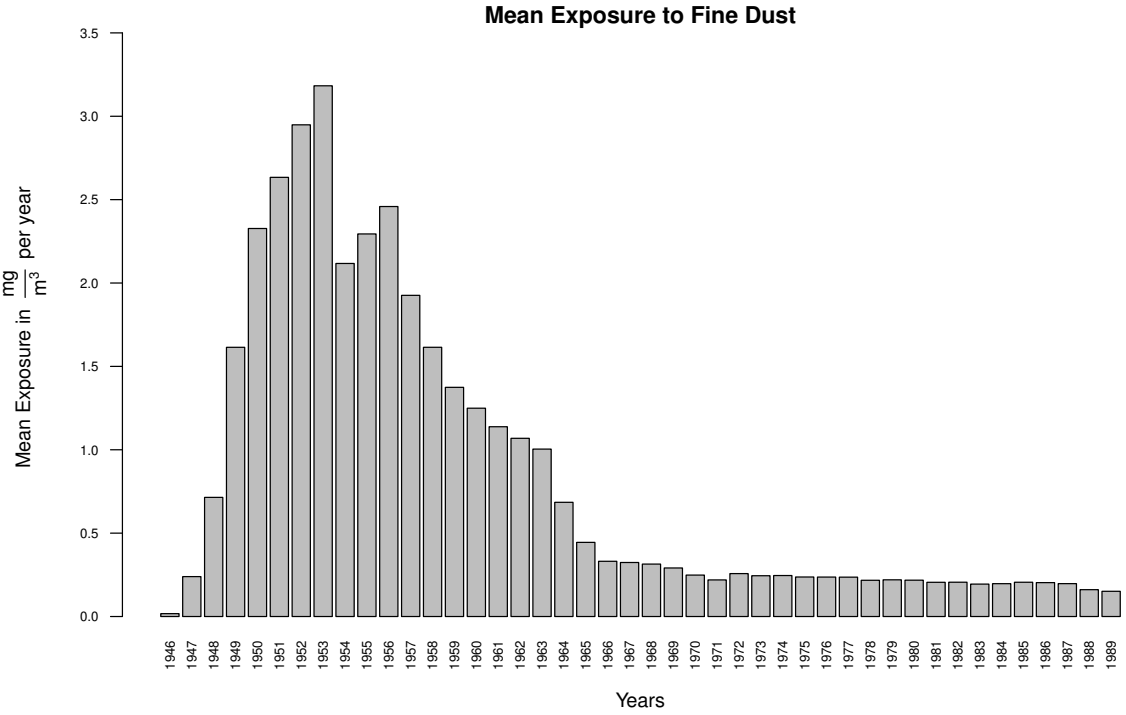

Figure A.4.: *Mean Exposure to Fine dust in the Wismut Cohort (1946-1989)*

A.2. Characteristics of the excluded miners

The following table contains the same information as the tables 2.1 to 2.4, but for the excluded workers due to missing values for the silica dust exposure:

Excluded cases	Cases:	292		(0.5%)	
	Min.	1st Quartile	Median	3rd Quartile	Max.
Follow-up period $(years)$	22.17	62.23	68.67	75.27	90.5
Duration of Employment $(years)$	0.6667	13.19	22.83	33.94	42.83
Cum. Radon Exposure $(\frac{wlm}{year})$	0	12.56	38.98	272.2	2645
Cum. Silica Dust Exposure $(\frac{mg}{m^3}/year)$	NA	NA	NA	NA	NA

Table A.1.: Characteristics of the Excluded cases due to missing values for silica dust

For reasons of comparison, the characteristics of the whole cohort from table 2.1 are displayed once again:

Full cohort	Cases:	58987		(100%)	
	Min.	1st Quartile	Median	3rd Quartile	Max.
Follow-up period (*years*)	15.33	47	60.33	70.17	103.2
Duration of Employment (*years*)	0.4167	4.25	10.17	21.5	45.5
Cum. Radon Exposure ($\frac{wlm}{year}$)	0	1.746	18.42	262.4	3224
Cum. Silica Dust Exposure ($\frac{mg}{m^3}/year$)	0	0.4036	1.761	8.524	55.98

There are striking differences in the length of the follow-up period, as the excluded miners have a higher minimum and maximum as well as consistently higher location parameters. The same observation is to be reported for the duration of employment. The excluded cases also exhibit a higher cumulative exposure to radon with a median which is more than twice as high as the median for the cumulative radon exposure in the full cohort.
This observation is consistent with the longer duration of employment among the excluded miners. The distributions of the cumulative exposures to silica dust can obviously not be compared due to the missing values in the group of the excluded miners.
A comparison of the amount of miners who died in the excluded cases and in the full cohort is to be found in the table below.

	Cases	All Deaths	Lung cancer	Other causes
Excluded Cases	292 (100%)	163 (55.82%)	20 (6.85%)	143 (48.97%)
Full Cohort	58987 (100%)	20920 (35.47%)	3016 (5.11%)	17904 (30.35%)

Table A.2.: Comparison of the excluded cases to the full cohort

A.3. Quantiles of the weighted Lag-distribution

Quantile	16.7%	20%	25%	33.3%	40%	50%	60%	66.7%	75%	80%	83.3%
Lag	6	7	8	10	12	15	18	20	24	26	28

Table A.3.: Quantiles of the weighted Lag-distribution

A.4. DLMs for different combinations of B-Spline degrees and numbers of knots (Minimum Lag 0)

	AIC	BIC	Degree Lags	No. Knots Lags	df	Min. Lag > 1
1	58350.62	58392.65	1	0	7	lag0
2	58348.24	58396.28	1	1	8	lag1
3	58350.27	58398.31	1	1	8	lag0
4	58351.06	58399.1	1	1	8	lag0
5	58350.23	58404.27	1	2	9	lag1
6	58350.04	58410.09	1	3	10	lag3
7	58350.45	58416.51	1	4	11	lag4
8	58347.54	58419.6	1	5	12	lag4
9	58351.75	58399.79	2	0	8	lag0
10	58346.97	58401.01	2	1	9	lag4
11	58349.16	58403.21	2	1	9	lag2
12	58349.55	58403.59	2	1	9	lag1
13	58348.94	58408.99	2	2	10	lag4
14	58341.05	58407.1	2	3	11	lag4
15	58337.89	58409.95	2	4	12	lag4
16	58326.89	58404.95	2	5	13	lag3
17	58346.04	58400.09	3	0	9	lag3
18	58345.57	58405.62	3	1	10	lag4
19	58347.94	58407.99	3	1	10	lag3
20	58347.49	58407.54	3	1	10	lag2
21	58334.33	58400.39	3	2	11	lag4
22	58331.33	58403.39	3	3	12	lag3
23	58321.78	58399.85	3	4	13	lag3
24	58318.12	58402.19	3	5	14	lag3
25	58347.06	58407.11	4	0	10	lag1
26	58327.7	58393.75	4	1	11	lag3
27	58332.46	58398.52	4	1	11	lag4
28	58334.11	58400.17	4	1	11	lag4
29	58329.7	58401.76	4	2	12	lag3
30	58318.94	58397.01	4	3	13	lag3
31	58314.41	58398.48	4	4	14	lag2
32	58311.7	58401.78	4	5	15	lag2
33	58332.56	58398.62	5	0	11	lag3
34	58326.15	58398.21	5	1	12	lag3
35	58331.79	58403.85	5	1	12	lag3
36	58333.84	58405.9	5	1	12	lag3
37	58318.6	58396.67	5	2	13	lag3
38	58313.11	58397.18	5	3	14	lag2
39	58306.19	58396.26	5	4	15	lag2
40	58302.72	58398.8	5	5	16	lag2
41	58334.06	58406.12	6	0	12	lag3
42	58314.21	58392.28	6	1	13	lag3
43	58321.55	58399.61	6	1	13	lag3
44	58324.83	58402.89	6	1	13	lag3
45	58314.8	58398.87	6	2	14	lag2
46	58301.24	58391.32	6	3	15	lag2
47	58299.4	58395.48	6	4	16	lag2
48	58288.53	58390.62	6	5	17	lag2

Table A.4.: Comparison of the DLMs with zero as minimum lag

A.5. DLMs for different combinations of B-Spline degrees and numbers of knots (Minimum Lag 2)

	AIC	BIC	Degree Lags	No. Knots Lags	df	Min. Lag > 1
1	58386.36	58422.39	1	0	6	lag3
2	58345.03	58387.06	1	1	7	lag3
3	58348.25	58390.28	1	1	7	lag3
4	58354.46	58396.49	1	1	7	lag3
5	58347.03	58395.07	1	2	8	lag3
6	58347.95	58401.99	1	3	9	lag3
7	58349.7	58409.75	1	4	10	lag3
8	58350.18	58416.23	1	5	11	lag3
9	58355.88	58397.91	2	0	7	lag3
10	58345.58	58393.62	2	1	8	lag3
11	58346.32	58394.36	2	1	8	lag3
12	58346.71	58394.75	2	1	8	lag3
13	58347.52	58401.56	2	2	9	lag3
14	58346.53	58406.58	2	3	10	lag3
15	58345.53	58411.58	2	4	11	lag3
16	58343.07	58415.13	2	5	12	lag3
17	58343.84	58391.88	3	0	8	lag3
18	58345.67	58399.72	3	1	9	lag3
19	58345.82	58399.87	3	1	9	lag3
20	58345.1	58399.14	3	1	9	lag3
21	58342.82	58402.87	3	2	10	lag3
22	58341.73	58407.79	3	3	11	lag3
23	58341.56	58413.62	3	4	12	lag3
24	58343.34	58421.4	3	5	13	lag3
25	58344.5	58398.55	4	0	9	lag3
26	58340.79	58400.84	4	1	10	lag3
27	58339.56	58399.61	4	1	10	lag3
28	58338	58398.05	4	1	10	lag3
29	58339.91	58405.96	4	2	11	lag3
30	58340.37	58412.43	4	3	12	lag3
31	58342.09	58420.15	4	4	13	lag3
32	58343.42	58427.49	4	5	14	lag3
33	58337.11	58397.16	5	0	10	lag3
34	58338.96	58405.02	5	1	11	lag3
35	58338.68	58404.73	5	1	11	lag3
36	58338.32	58404.37	5	1	11	lag3
37	58339.49	58411.55	5	2	12	lag3
38	58341.05	58419.12	5	3	13	lag3
39	58342.17	58426.24	5	4	14	lag3
40	58334.16	58424.23	5	5	15	lag4
41	58338.01	58404.07	6	0	11	lag3
42	58338.85	58410.91	6	1	12	lag3
43	58338.68	58410.74	6	1	12	lag3
44	58338.26	58410.32	6	1	12	lag3
45	58340.21	58418.27	6	2	13	lag3
46	58339.8	58423.87	6	3	14	lag3
47	58333.21	58423.29	6	4	15	lag4
48	58333.98	58430.06	6	5	16	lag4

Table A.5.: Comparison of the DLMs with two as minimum lag

A.6. Exposure-response curves of the AIC-best plausible DLM

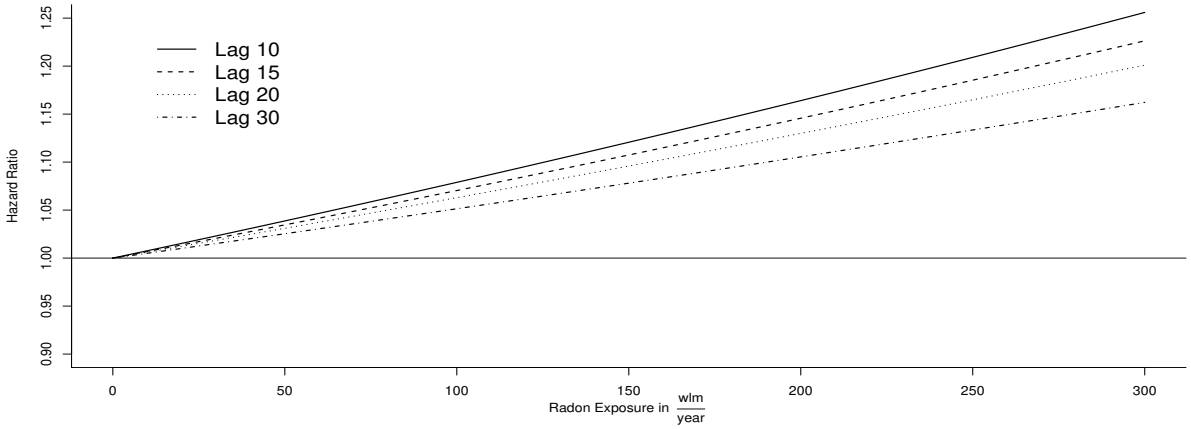

Figure A.5.: *The exposure-response curves for four different lags (Model 5)*

A.7. Comparison of an intercept model to a non-intercept model with a starting lag of five years

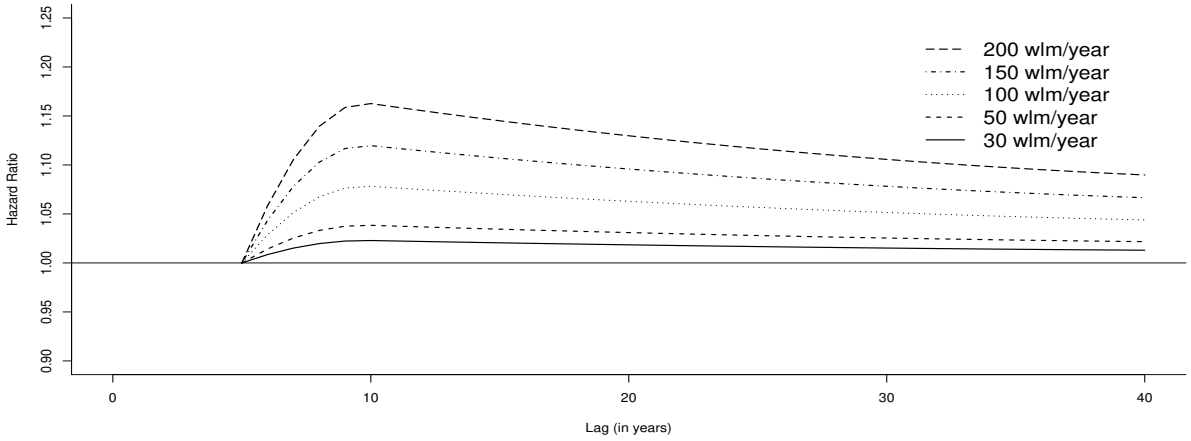

Figure A.6.: *Lag-response curves for different radon exposures (Model 5-2)*

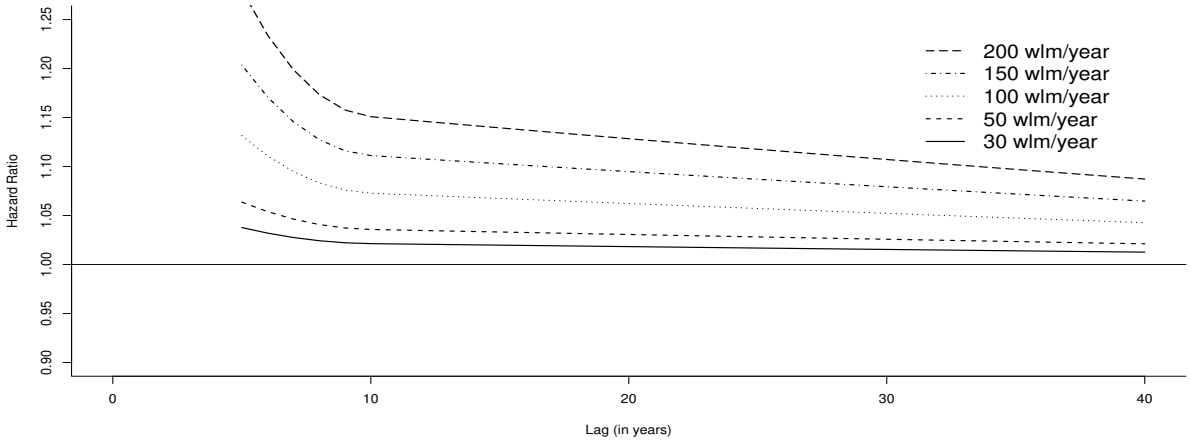

Figure A.7.: *Lag-response curves for different radon exposures (Model 5-3)*

The inappropriateness of the exclusion of an intercept in this case is clearly shown by the two figures above and by a comparison of the AICs (BICs) of the two models: While the AIC (BIC) of model 5-2 takes a value of 58348.42 (58396.46), the AIC (BIC) of model 5-3 amounts to 58346.64 (58400.69). The lower AIC for model 5-3 shows that a model with intercept is preferred, in case of 5 years being chosen as minimum lag. This indicates that a model with a minimum lag of five years without an intercept isn't adequate.

A.8. Quantiles of the Exposure-distribution

Quantile	16.7%	20%	25%	33.3%	40%	50%	60%	66.7%	75%	80%	83.3%
Exposure	0.6	0.75	1	1.6	2	3.6	7.75	16.52	31.66	41	55

Table A.6.: Quantiles of the Exposure-distribution

A.9. Alternative Models

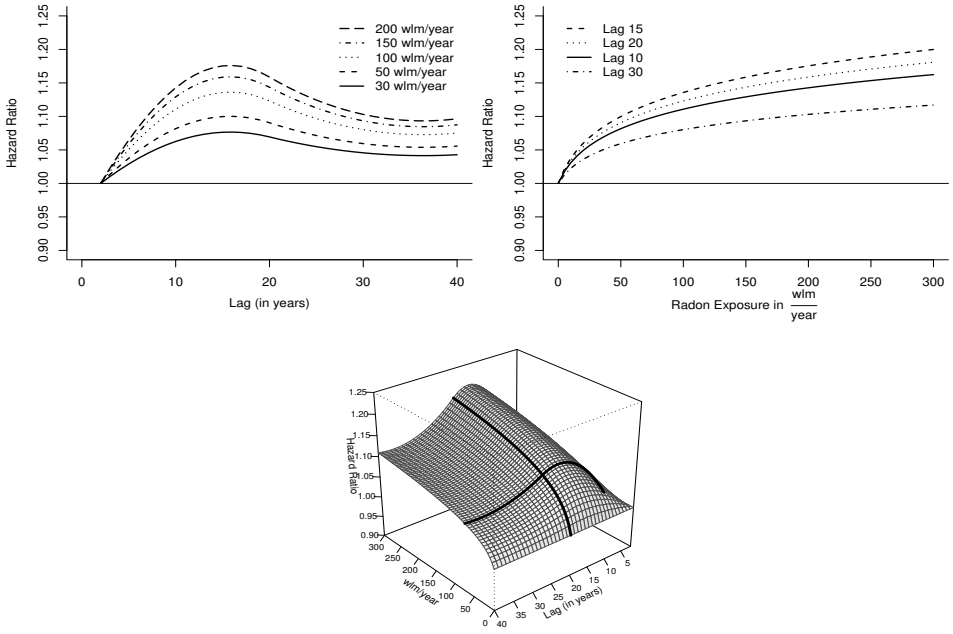

Figure A.8.: *Exposure-lag-response relationship for Model 7*

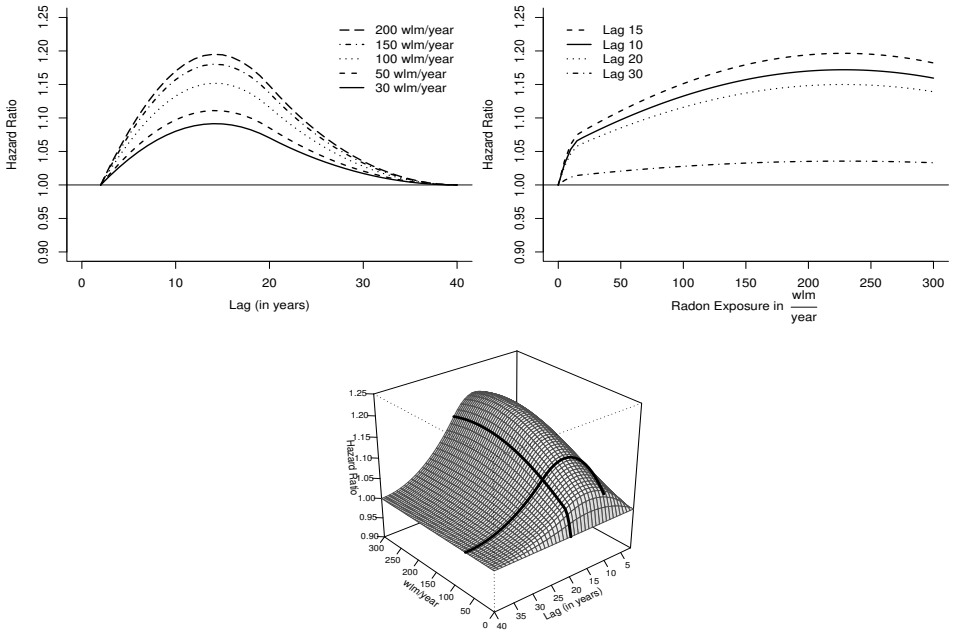

Figure A.9.: *Exposure-lag-response relationship for Model 8*

A.10. Estimates for the covariates from Model 9

	Estimate	exp(Estimate)	Standard error	p-value
cal	-0.0122	0.9879	0.0044	0.0057
age	-0.0233	0.9769	0.0048	0.0000

Table A.7.: Estimates of *cal* and *age* from model 9

A.11. Electronic appendix

You find a CD-Rom containing the electronic appendix attached to this study. The electronic appendix comprises six folders called "analysis", "data set", "documentation", "figures", "server" and "thesis".
The analysis folder contains the eight R files with the codes that were used to analyze the data. The first R file ("01-preparation.R") contains the code that was used to prepare the data and get it in the right form for the analysis. The code in the second R file ("02-exposure-history.R") was used to create the exposure histories for radon as well as for silica dust. In the third R file ("03-description.R), the code which was used to create the tables and figures in chapter 1 and in the appendices A.1 and A.2 can be found. The fourth and fifth R file ("04-dlm-modelling.R", "05-dlnm-modelling.R") contain the codes for the estimation and the analysis of the DLMs and the DLNMs. R file number six ("06-pam-modelling") contains the analysis of the penalized piecewise exponential additive models. Eventually the two R files "00-functions.R" and "99-packages.R" contain the used functions and the used packages.
In the folder called "server" those R codes are to be found, that were used to estimate the models for the model comparisons of the DLMs and the DLNMs ("dlm-server.R" and "dlnm-server.R"). Additionally, it contains the R code for the estimation of the penalized piecewise exponential additive models ("pam-server.R"). They all were run at the server of the Institute for Statistics due to the computational effort.
The data set folder just contains the data set, the figures folder contains all figures displayed in this thesis and in the thesis folder a PDF-version as well as the TEX-code of the thesis can be found. All presentations that were held, as well as important e-mails and status updates for the project partners are archived in the folder "documentation".

References

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica: Journal of the Econometric Society*, pages 178–196.

Armstrong, B. (2006). Models for the relationship between ambient temperature and daily mortality. *Epidemiology*, 17(6):624–631.

Bender, A., Scheipl, F., and Küchenhoff, H. (2016). Modeling exposure–lag–response associations with penalized piece-wise exponential models. *Department of Statistics (LMU): Technical Reports*, Nr. 192.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, 34(2):187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Dahl, D. B. (2016). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-2.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102.

Gasparrini, A. (2011). Distributed lag linear and non-linear models in R: the package dlnm. *Journal of Statistical Software*, 43(8):1–20.

Gasparrini, A. (2014). Modeling exposure-lag-response associations with distributed lag non-linear models. *Statistics in medicine*, 33(5):881–899.

Gasparrini, A., Armstrong, B., and Kenward, M. G. (2010). Distributed lag non-linear models. *Statistics in medicine*, 29(21):2224–2234.

Gasparrini, A., Scheipl, F., Armstrong, B., and Kenward, M. G. (2016). A penalized framework for distributed lag non-linear models. unpublished.

Grosche, B., Kreuzer, M., Kreisheimer, M., Schnelzer, M., and Tschense, A. (2006). Lung cancer risk among german male uranium miners: a cohort study, 1946-1998. *British journal of cancer*, 95(9):1280–1287.

Holford, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics*, pages 299–305.

IARC (1988). *IARC Monographs on the evaluation of the carcinogenic risk to humans. Man-made mineral fibres and radon*, volume 43. IARC, Lyon, France.

IARC (1997). *IARC Monographs on the evaluation of the carcinogenic risk to humans. Silica, Some Silicates, Coal Dust and para-Aramid Fibrils*, volume 68. IARC, Lyon, France.

Kai, M., Luebeck, E. G., and Moolgavkar, S. H. (1997). Analysis of the incidence of solid cancer among atomic bomb survivors using a two-stage model of carcinogenesis. *Radiation research*, 148(4):348–358.

Kauermann, G. (2014). Analyse von Überlebensdauern. Lecture Notes.

Kreuzer, M., Brachner, A., Lehmann, F., Martignoni, K., Wichmann, H.-E., Grosche, B., et al. (2002). Characteristics of the german uranium miners cohort study. *Health Physics*, 83(1):26–34.

Kreuzer, M., Grosche, B., Dufey, F., Schnelzer, M., Tschense, A., and Walsh, L. (2011). The german uranium miners cohort study (wismut cohort), 1946–2003. *Federal Office for Radiation Protection (BfS): Technical Reports*.

Kreuzer, M., Grosche, B., Schnelzer, M., Tschense, A., Dufey, F., and Walsh, L. (2010). Radon and risk of death from cancer and cardiovascular diseases in the german uranium miners cohort study: follow-up 1946-2003. *Radiation and environmental biophysics*, 49(2):177–185.

Laird, N. and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76(374):231–240.

Lehmann, F. (2004). Job-exposure-matrix ionisierende strahlung im uranerzbergbau der ehemaligen ddr (version 06/2003). *Gera: Bergbau BG*.

Lehmann, F., Hambeck, L., Linkert, K., Lutze, H., Meyer, H., Reiber, H., Renner, H., Reinisch, A., Seifert, T., and Wolf, F. (1998). Belastung durch ionisierende strahlung im uranerzbergbau der ehemaligen ddr. *St. Augustin: Hauptverband der gewerblichen Berufsgenossenschaften*.

Luebeck, E. G., Heidenreich, W. F., Hazelton, W. D., Paretzke, H. G., and Moolgavkar, S. H. (1999). Biologically based analysis of the data for the colorado uranium miners cohort: age, dose and dose-rate effects. *Radiation research*, 152(4):339–351.

National Research Council, Committee on the Biological Effects of Ionizing Radiation (1999). Health effects of exposure to radon - BEIR IV. *Washington, DC: National AcademyPress*.

Pharmazeutische Zeitung Online (2016). Lungenkrebs - Oft spät erkannt und kaum behandelbar. http://www.pharmazeutische-zeitung.de/index.php?id=1589. Accessed: 2016-11-30.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.

Walsh, L., Dufey, F., Tschense, A., Schnelzer, M., Grosche, B., and Kreuzer, M. (2010). Radon and the risk of cancer mortality-internal poisson models for the german uranium miners cohort. *Health physics*, 99(3):292–300.

Wickham, H. and Chang, W. (2016). *devtools: Tools to Make Developing R Packages Easier*. R package version 1.12.0.

Wismut GmbH (2016a). 1945 - 1953 SAG Wismut. http://www.wismut.de/www/webroot/en/sag_wismut.php. Accessed: 2016-06-13.

Wismut GmbH (2016b). State within the GDR state. http://www.wismut.de/www/webroot/en/sdag_wismut.php. Accessed: 2016-06-13.

Wismut GmbH (2016c). The Wismut rehabilitation project - revival of a region. http://www.wismut.de/en/wismut_gmbh.php. Accessed: 2016-09-13.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Wood, S. (2014). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*. R package version 1.8-12.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Wood, S. N., Goude, Y., and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):139–155.

Zaballa, I. and Eidemüller, M. (2016). Mechanistic study on lung cancer mortality after radon exposure in the wismut cohort supports important role of clonal expansion n lung carcinogenesis. unpublished.

Zeileis, A. and Grothendieck, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27.

Declaration of Authenticity

The work contained in this thesis is original and has not been previously submitted for examination which has led to the award of a degree.

To the best of my knowledge and belief, this thesis contains no material previously published or written by another person except where due reference is made.

Matthias Aßenmacher