Statistical analysis of association between long-term exposure to air pollution and repeated hospitalizations for pneumonia

Kristina Ranc

Kongens Lyngby, 2011
Master Thesis

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

Summary

This thesis deals with statistical methods and their application on the association between long-term exposure to traffic-related air pollution (for up to 39 years) in Copenhagen and hospital admissions for pneumonia, in a prospective cohort study. The purpose of this study is to investigate whether the exposure to air pollution is a risk factor for pneumonia hospitalizations, as well as it is associated with recurrent admissions.

The Danish Cancer Society provided data on 57053 participants of Danish Cancer, Diet and Health cohort, aged 50-65 years at baseline (1993-1997), which were followed in Danish hospital discharge register for all hospital admissions for pneumonia up to 2010. Traffic pollutants considered are nitrogen dioxide (NO_2) and nitrogen oxides (NO_x), available as mean annual levels estimated at residential addresses since 1971. We modelled the association between mean NO_2 and NO_x levels and hospitalizations for pneumonia using the Cox regression, in the full cohort and separately for people with and without previous hospital admissions for pneumonia and with and without co-morbidities defined by Charlson index.

In order to explore the association between the exposure to air pollution and the first or recurrent pneumonia hospitalizations this thesis contains a variety of statistical survival methods both standard and extended. The applied models are the ordinary Cox model, Andersen-Gill model, Conditional Andersen-Gill model, Frailty model, and Conditional Frailty model. The model are first introduced and then applied.

The investigation showed that during 12.7 years' mean follow-up, 3024 (5.7%) out of 53239 eligible people were admitted to hospital for pneumonia, and among those individuals 626 (1.2%) had more than one pneumonia admission. Mean NO_2 levels were significantly positively associated with risk for first pneumonia hospitalization in the full cohort (hazard ratio and 95% confidence interval per double mean exposure: 1.25; 1.14-1.36); in 46462 people without earlier hospitalizations for pneumonia or co-morbid conditions defined by Charlson (1.23; 1.11-1.36), and in 6292 people with history of co-morbid conditions defined by Charlson (1.22; 1.02-1.46).

The highest risk was observed in 485 people with a history of pneumonia hospitalizations (1.68; 1.01-2.81) which led to the idea of investigating the effect of exposure to air pollution on recurrent pneumonia hospitalizations. Conditional Frailty model revealed that mean NO_2 levels were also significantly positively associated with risk for recurrent pneumonia hospitalization in full cohort, up to 3 admissions per subject (1.30; 1.19-1.41).

From these findings we concluded that living in areas with high traffic-related air pollution increases the risk of hospitalization for pneumonia. The effect was highest in people with prior hospitalizations for pneumonia.

Preface

This thesis was prepared at the Department of Informatics Mathematical Modelling, the Technical University of Denmark (DTU) in partial fulfillment of the requirements for acquiring the Master of Science degree (M.Sc.) in engineering.

The thesis deals with statistics with the main focus on statistical models of survival data and the application of those on air pollution and pneumonia data provided by the Danish Cancer Society. The ordinary and extended Cox regression model has been applied on time to first event data as well as recurrent data.

I would like to thank my supervisors on this project, Zorana Jovanovic Andersen for the opportunity to do this project in collaboration with the Danish Cancer Society, for her guidance, patience and lots of motivation in many aspects of life; and Per Bruun Brockhoff, for the guidance and support throughout the project. I would also like to thank a number of people who made it possible to carry out this project, the Institute of Cancer Epidemiology at the Danish Cancer Society, for providing data for this project; people in the department of Environment and Cancer for help and pleasant working environment that made me feel like the part of the group; and MD Reimar W. Thomsen for help with the understanding the risk factors for pneumonia. Last but not least, a special thank to my family for all the support, encouragement and love I received even though they were so far away.

Kongens Lyngby, July 2011

Kristina Ranc

Contents

Summary i

Preface iv

Chapter 1 5

Introduction 5

1.1 Epidemiology and the Burden of Disease 5

1.2 Pneumonia 6

1.3 Air Pollution Epidemiology 6

1.4 Air Pollution and Pneumonia 7

1.5 Purpose of this Study 8

Chapter 2 9

Cohort and health outcome 9

2.1 Cohort Studies 9

2.2 The Danish Diet, Cancer and Health (DCH) Cohort Design 9

2.3 Health Outcome - Pneumonia 10

2.3.1 Danish Health Registries 11

2.4 Potential Confounders 11

2.5 Co-morbidity - Major Chronic Diseases 14

Chapter 3 16

Air Pollution 16

3.1 Classification of Air Pollutants 16

3.1.1 Gasses 16

3.1.2 Particulate matter 17

3.2 AirGIS Model 18

3.3 Exposure assessment 20

Chapter 4 22
Methodology 22
4.1 Introduction to Survival Analysis 22
4.1.1 Censoring and truncation 22
4.2 Survival function and hazard rate 24
4.3 Counting process formulation 25
4.4 Estimation 26
4.5 Cox proportional hazard model 27
4.5.1 Estimation 28
4.5.2 Test statistics 29
4.5.3 Functional Form 30
4.5.4 Testing proportional hazards assumption 31
4.6 Extending the Cox model 31
4.6.1 Robust variance for recurrent events 32
4.6.2 Models for recurrent events 33
4.6.2. a Intensity – Based model 33
4.6.2. b Andersen - Gill model 34
4.6.2. c Frailty model 35
4.6.2. d Conditional models 36
Chapter 5 38
Results 38
5.1 Study population and event incidence 38
5.2 Descriptive Data Analysis 39
5.2.1 Testing the potential confounders 39
5.2.2 Air pollution exposure 43
5.2.3 Cumulative hazard rates and Survival curves 47
5.3 Time to first event analysis using ordinary Cox model 51
5.3.1 Association between NO_2 and NO_x exposure and first 51
pneumonia occurrence in DCH cohort 51
5.3.2 Association between traffic proxies exposure and pneumonia incidence in DCH cohort .. 55
5.4 Recurrent events analysis using extended Cox model 56
5.4.1 Association between NO_2 and NO_x exposure and recurrent 58

pneumonia occurrence in DCH cohort 58
5.5 Model validation 62
Chapter 6 65
Conclusions and Discussion 65
6.1 Conclusion 65
6.2 Discussion 67
Chapter 7 68
Considerations and Further Work 68
7.1 Considerations 68
7.1.1 Limitations 69
7.1.2 Strengths 69
7.2 Further work 69
Appendix A 71
Definitions 71
Appendix B 75
Acronym table 75
Appendix C 76
Supplementary figures 76
C.1 Survival Curves and Cumulative Hazards 76
C.2 Checking PH assumption – Schoenfeld residuals 79
Appendix D 81
R programming 81
D.1 Data preparation 81
D.2 Testing potential confounders – Univariate Cox regression 86
D.3 Modeling the exposure to air pollution 89
D.4 Ordinary Cox regression models - time to first pneumonia 91
D.4 Extended Cox regression models – recurrent pneumonias 97
Bibliography 106

Chapter 1

Introduction

1.1 Epidemiology and the Burden of Disease

Epidemiology is the study of how disease is distributed in populations and the factors that influence or determine this distribution. The premise underlying epidemiology is that any heath condition is not at random; rather certain characteristics of individual predispose a person to, or protect against, a variety of different diseases. The characteristics may be primary genetic in origin, or may be the result of exposure to certain environmental factor. However, the most often the interaction of genetics and environment determine the development of disease. Epidemiology informs evidence-based medicine for identifying risk factors for disease and determining optimal treatment approaches to clinical practice and for preventive medicine [1].

Investigating the cause and risk factors for disease, gives valuable information that can be used in prevention and reduction of a risk from a disease. Chronic diseases, characterized by long duration and slow progression, such as cardiovascular diseases (CVD), cancer, chronic respiratory diseases, and diabetes, are by far the leading cause of mortality in the world, representing 60% of all deaths [2]. One of the biggest and still unsolved concerns is cancer. Just couple of years ago the world's leading cause of death was CVD disease. However, treatment improvements, successful risk factor management, and prevention have reduced CVD incidence and cancer has become the number one cause of death with steady rates over recent years [2]. During the second half of nineteenth century the cancer registries have been implemented and facilitated epidemiological studies which have shown that there is also strong relationship between lifestyle, in particular smoking and diet, and cancer [3,4]

Other diseases also impose large public health burden and present challenges. Chronic respiratory diseases are in top ten leading causes of morbidity and mortality in the World. Chronic obstructive respiratory disease (COPD), mainly caused by smoking, but also occupational and environmental exposures to particles and dust, is projected to be the third leading cause of death and the fifth leading cause of disability by 2020 [5]. Asthma and allergic diseases are also on rise, both in children and adults [6]. Despite dramatic reduction in mortality from infectious disease in this century, respiratory infections still present a big problem in developing (low-income) countries [2], but also considerable problem in the

developed world. Namely, lower respiratory infection is in top four leading causes of death with increasing rates over years [7].

1.2 Pneumonia

The most common infections that can affect the lower respiratory tract are pneumonia and bronchitis, whereas influenza affects both the upper and lower respiratory tracts. Pneumonia is a form of acute respiratory infection that affects the lungs. The lungs are made up of airways and small air sacs with thin walls called alveoli, which fill with air when a healthy person breathes, and where oxygen exchange with blood stream takes place. When an individual has pneumonia, the alveoli are filled with pus and fluid, which makes breathing painful and limits oxygen intake. Pneumonia is caused by a number of infectious agents, including viruses, bacteria, and fungi. The symptoms of pneumonia are rapid or difficult breathing, cough, fever, chills, loss of appetite, wheezing (more common in viral infections). Pneumonia is age – related with the vast majority among those over 65 years [2].

During the past decade, hospitalizations with pneumonia have increased by 20–50% in Western population. In the USA, pneumonia combined with influenza is the eight leading cause of death and the most frequent due to infectious disease [7-9]. Also the European Union recent statistics shows very high death rates for pneumonia, which are the highest in the United Kingdom, Belgium, Ireland, Portugal and Denmark [10]. With treatment and prevention improvements the average life length is increasing, therefore also the number of elderly, as well as the number of hospitalizations among older people [2]. The economic burden associated with hospital care, medications, and years of work lost due to morbidity and mortality is projected to escalate with increasing number of older people with chronic diseases in next few decades [8,11,12]. In Denmark, over 14000 people are admitted to hospital for pneumonia annually, and over 1600 dies from pneumonia, mainly women. Furthermore, the number of people hospitalized for pneumonia over last decade is increasing in Denmark, whereas admissions for bronchitis remain stable [13].

1.3 Air Pollution Epidemiology

Technological improvements and economical development lead to more comfortable life styles, better health care, and constant improvements in life expectancy. However, some drawbacks of economical prosperity have introduced new public health challenges; obesity and physical inactivity associated with modern lifestyle have contributed to a large CVD burden and recent diabetes epidemic [14]. Environment around us has also suffered from technological revolution and affected the human health. Side-products of economic development, increasing industrial activity, massive growth in transport sector (motorized vehicle and air), and accompanying need for more energy have lead to soil, water, and air contaminations which affect human health. Environmental epidemiology, defined as the epidemiologic study of the health

consequences of exposure that are involuntary and that occur in the general environment (air, water, diet, soil, etc.), attempts to explain how environment around us can cause a disease. A common feature in environmental epidemiology is that data are observed, and usually involve low-level exposure to the general public, which are difficult to measure and difficult to link to disease [13]. This is also true for air pollution, which was only recently (in last 60 years) recognized as a risk factor for a number of diseases. Air pollution epidemiology is a part of environmental epidemiology, discerning the complex link between air pollution and disease [15].

The past air pollution problems (several decades ago) in the western world cities were mainly caused by emissions from combustion fossil fuels such as wood and coal burning used for domestic (heating and cooking) and industrial purposes. These sources of air pollution have been successfully controlled by policies limiting their use and providing alternatives, such as introduction of central heating in the major cities in the developed world, which contributed to major reductions in pollution for sulfur dioxide (SO_2). Along with the reduction on emissions from fossil fuels, new threat to clean air both in developed and rapidly industrializing countries is now posed by traffic emissions. Petrol and diesel-powered motor vehicles emit a wide variety of pollutants, principally particulate matter (PM), carbon monoxide (C0), nitrogen oxide ($N0_x$), and volatile organic compounds (VOC_s), a mix of affect urban air quality. Traffic pollution problems are worsening worldwide, leading accordingly to recent increasing number of epidemiological studies focusing on this source of air pollution [16].

It is well established that exposures to elevated levels of air pollution over several days can exacerbate respiratory and cardiovascular disease triggering hospitalizations and death [17-19]. Accumulated effects of air pollution due to chronic, long-varying exposure to air pollution over many years have also been shown to cause the development of chronic respiratory and cardiovascular disease [17]. Also in Denmark, air pollution was linked to the risk for stroke [20], the respiratory diseases, such as asthma with children and adults [2,11], as well as COPD [12].

Furthermore, the increase in chronic conditions such as heart disease, diabetes, chronic obstructive pulmonary disease and cancer have been suggested as important factors underlying this increasing trend of pneumonia hospitalizations.

1.4 Air Pollution and Pneumonia

The idea that air pollution can cause infectious disease such as pneumonia is rather new. Exposure to pollutants in air affects lungs by causing oxidative stress and inflammation in lung tissues, which is a biological mechanism behind COPD and asthma association with air pollution [11,12,21]. With respect to infectious disease, it is believed that long-varying exposure to air pollution and accumulated damage from this exposure in lung tissue predisposes individuals to

pneumonia. Specifically, combined with other risk factors, such as age, nutrition, smoking habits, alcohol intake, occupational exposure etc., exposure to air pollution reduces the ability of organism to defend against viruses and bacteria, especially in elderly, thus increasing the risk for pneumonia [22]. Data from animal experiments have illustrated that exposure to nitrogen dioxide (NO_2) can impair the function of alveolar macrophages and epithelial cells, thus increasing the risk of lung infections, such as influenza and pneumonia [23].

Epidemiological evidence regarding the link between air pollution and pneumonia is very limited. Only single study to date has examined a link between long-term exposure to air pollution and risk of pneumonia [24]. This case-control study from Ontario, Canada, has recently found a link between long-term exposure to air pollution at home and pneumonia hospitalizations among elderly. This study lacked information on long residential address history, and thus long-term exposure was defined only as 2 to 9 years mean exposure prior to pneumonia diagnoses. Furthermore, inherent limitation of case-control studies is the recall and information bias when collecting confounder information retrospectively, after defining cases and controls. Finally, Neupane et al. did not have information on co-morbid conditions, which are well known to be important determinants for the risk of pneumonia, and possibly modifiers of air pollution effect.

1.5 Purpose of this Study

Here we studied the association between air pollution at residence for up to 40 years and the risk for first ever, as well as recurrent hospital admission for pneumonia in an elderly Danish cohort. We present several novel aspects in respect to literature [24]. First, we have a well defined large elderly cohort (57000 individuals) with a prospective assessment of risk factors for pneumonia. Secondly, pneumonia was assessed objectively from a nationwide hospital register. Finally, we tested for the first time whether the effect of air pollution was modified by a number of lifestyle factors as well as co-morbidities; and whether people with co-morbidities were more susceptible to the effect of air pollution than healthy people (without any disease at baseline), using Charlson co-morbidity index.

Chapter 2

Cohort and health outcome

This chapter consists of introduction to the cohort used in this study and the definition of the health outcome of interest – pneumonia. The cohort includes many variables, some of which information lie beyond the aim of this study. All relevant variables are described and corresponding characteristics have been further investigated.

2.1 Cohort Studies

In a cohort study a group of people is identified and followed over a period of time to see how their exposures affect their health outcomes. For ethical reasons, randomized people cannot be exposed to potentially harmful substance; therefore this is not a randomized study design. This type of study, called observational study, is normally used to look at the effect of suspected risk factors that cannot be controlled experimentally. For example, in order to study the association between some of the personal habits, lifestyle characteristics, uncontrolled exposures and occurrence of disease.

There are two types of cohort studies. A *prospective cohort study* is where the investigator identifies the original population at the beginning of the study and accompanies the subjects concurrently through calendar time until the certain point where disease develops or doesn't develop. The problem with this design is a need for long follow-up calendar time. The other type of cohort design is *retrospective* where the exposure is ascertained from past records and outcome is ascertained at the time the study has begun. It is also possible to conduct a study that is a combination of previous two types.

2.2 The Danish Diet, Cancer and Health (DCH) Cohort Design

The Danish Diet, Cancer and Health cohort used in this analysis consists of 57053 people (27178 males and 29875 females) aged 50-65 years from Denmark, who lived in Copenhagen and Aarhus between December 1993 and May 1997. This cohort was conducted to investigate relations between lifestyle: dietary components, food and nutrition (by single item or combinations) and the incidence of cancer and chronic diseases. First, at baseline (1993-1997) all participants filled in a questionnaire concerning lifestyle factors. The questionnaire includes basic daily habits and more specific known or suspected risk factors for cancer development, such as smoking habits, alcohol intake, diet, occupational history etc. The information from

questionnaires is combined with biological specimens in order to investigate genetic susceptibility and gene-environment interactions with regard to diet, dietary components, and the risk of disease development [25].
DCH prospective cohort study enables us to analyze diseases other than cancer, by linking people under the study to health registries, such as hospital registry.

2.3 Health Outcome - Pneumonia

Pneumonia is one of the leading causes of death from infectious disease with increasing rates all over the world [9]. Therefore, we are interested it investigating the association between lifestyle and air pollution exposure, and the risk for pneumonia hospitalizations in Denmark.

DCH cohort study was primarily conducted for studding the risk of cancer and chronic diseases but since pneumonia can occur as co-morbidity in relation to many other chronic diseases, it is relevant outcome which explains some of the burden of chronic disease. In favor of this study is also the fact that the DHC cohort is constructed and planed to be used for cancer related investigations. The participants were aware of that when received the questionnaires, which might lead to having the biased answers. Therefore, use of this cohort in studding non-cancer related outcome, like pneumonia, reduces possible information and recall bias that could come from the awareness of investigated people about DCH cohort's main use when answering questions about confounders.

The unique civil registration number (CPR) allows for linkage of DCH cohort participant to the Danish National Hospital Discharge Register for extraction of their hospitalizations and corresponding diagnoses, defined by International Classification of Diseases (ICD) codes. To obtain date of death or emigration and detailed residential address history from 1971 to 2010 we have used the Central Population Registry and for geographical coordinates the Danish Address Database. ICD is the international standard diagnostic classification of disease given by the Worlds Health Organization (WHO) for all general epidemiological, health management purposes and clinical use [26]. Relevant diagnosis are pneumonia (ICD-10 codes J12.x-J18.x), ornithosis (ICD-10 code A709.x), or legionelosis (ICD-10 code A481.x) occurring between the baseline and the end of follow-up, 31st of December 2009. (Corresponding ICD-8 codes are: 480.xx-486.xx, 0.73.xx, and 471.xx respectively).

2.3.1 Danish Health Registries

All Danish residents have a unique personal identification number called CPR, encoding sex and date of birth, which is administrated by the Danish Civil Registration System. Most public administrative records use this number for identification and linkage of citizens.

The Central Population Registry together with the Danish Address Database contains information about emigration, death and change of address.

The Danish health system provides free health care and the National Health Insurance Service Registry (NHISR) contains information about all services provided by general and specialist practitioners in Denmark. Furthermore, the National Patient Register (NPR), established in 1977, is the base of all patient – discharges from the hospitals together with given diagnosis, dating back to 1976. Diagnoses are coded corresponding to ICD which has couple of versions involving by time. Current classification follows ICD – 10, whereas before 1999 it was ICD – 8. The Register of Medical Product Statistics (RMPS), established in 1993 contains information of all prescriptions from Danish pharmacies including prescriptions by date, type, and amount.

2.4 Potential Confounders

When the relationship between exposure and the outcome of interest has to be examined one has to take into account that other factors could influence this relation. These factors are called confounders. Confounding occurs when a variable is associated with both the exposure and the disease under study. Therefore, in epidemiology the effect of the exposure under study on the disease (outcome) can be mixed with that of a third factor that is associated with the exposure and an independent risk factor for the disease. The consequence of confounding is that the estimated association between exposure and the outcome is not the same as true effect, which leads to wrong conclusions, since the effect attributed to the exposure of interest is actually caused by something else. The confounders in some cases can completely remove the effect of exposure, but they can also just change the strength of the relationship [1].

For studying the effect of air pollution on pneumonia in DCH cohort, we first needed to examine which of available personal information could influence the risk of pneumonia hospitalization. Thus, before testing the relationship between air pollution exposure and pneumonia hospital admissions we need to examine potential confounding of other factors (*Figure 1*).

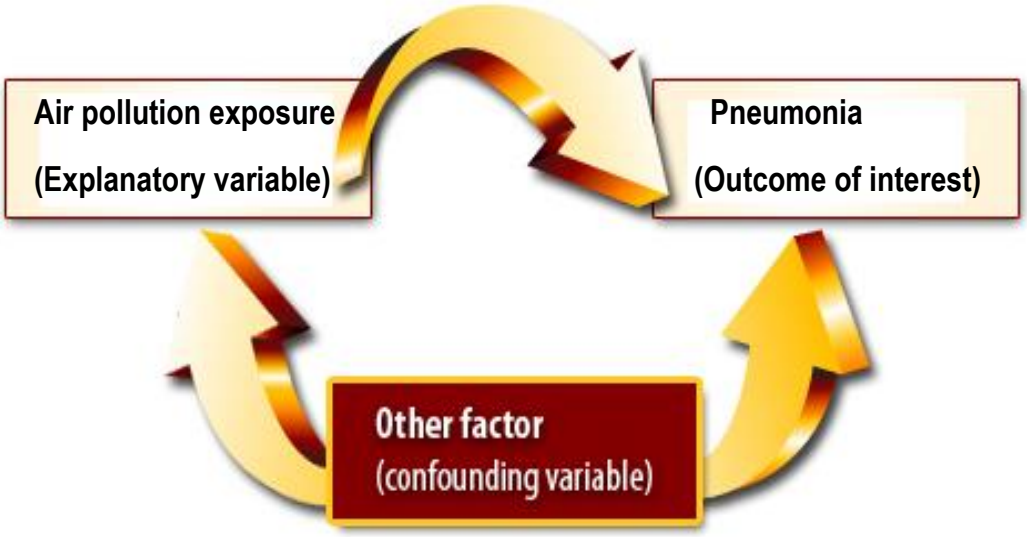

Figure 1: Confounding

First, we considered well established risk factor for any disease age and gender. The risk for most diseases, including pneumonia, increases with age. Age at the baseline (from 1993 until 1997) in modeled as the continuous variable (age as underlying time scale) or categorized in two levels around the mean. Gender is known to be a common determinant of disease risk, reflecting many factors that differ between genders, including biological differences, but also life-style, occupation, utilization of health care, prevention, etc.

Secondly, lifestyle factors which have been found to be linked to risk of pneumonia in existing literature are considered, and these include: body mass index (BMI); smoking habits as smoking status, intensity, duration and exposure to environmental tobacco smoke; alcohol consumption as status for consuming some or no alcohol as well as intensity; nutrition habits as fruit and fat intake given in grams per day; physical activity in hours per week; and occupational exposure.

Smoking status is defined as never, previously or currently smoker. Smoking intensity was calculated by equating a cigarette to 1g, a cheroot or a pipe to 3g, and a cigar to 5g of tobacco. Smoking related characteristic is also environmental tobacco smoke (ETS) which is the indicator of exposure to second-hand smoke at home or work for minimum 4 hours per day. Intensity of alcohol intake is defined as the number of drinks per week. Occupational exposure is defined as a minimum of 1 year employment in: mining; electroplating; shoe or leather manufacture; welding; painting; steel mill; shipyard; construction (roof, asphalt, or demolition); truck, bus, or taxi driver; asbestos or cement manufacture; asbestos insulation; glass, china, or pottery

manufacture; butcher; auto mechanic; waiter; or cook; and reflects occupation earlier related to chronic lung disease, with focus on lung cancer, as this cohort was designed primarily to study cancer.

Additional potential predictor is socio-economic-status (SES) defined as yearly income on municipality levels in Copenhagen.

All potential confounders are defined as shown in *Table 1*.

Risk factor	Categories
Age	< 56 vs. ≥ 56
Gender	Female vs. male
Education	< 8 years 8-10 years ≥ 10 years
BMI	Underweight (< 20 kg/m^2) Normal (20-30 kg/m^2) Obese (> 30 kg/m^2)
Nutrition fruit intake **fat intake**	Mean in 100g/day Mean in 100g/day
Sports	Not physically active < 3.5 hours/day ≥ 3.5 hours/day
Smoking	Never Previously Current < 15 g/day Current 15-25 g/day Current ≥ 25 g/day
ETS	Yes / No
Alcohol	No alcohol use 1-20 drinks/week ≥ 20 drinks/week
Occupational exposure	Yes / No
SES	Yearly income/municipality

Table 1: Definition of the potential confounders

2.5 Co-morbidity - Major Chronic Diseases

The Charlson index is a co-morbidity scoring system that includes weighting factors on the basis of disease severity. The system was developed originally as a prognostic indicator on the basis of patients with a variety of conditions admitted to a general medical service. It is commonly used in outcome studies to account for the impact of co-morbid conditions of patients and has been adapted and validated or use with hospital discharge data in ICD databases for the prediction of short – and long – term mortality [27].

The Charlson index includes 19 major disease categories, such as congestive heart failure, peripheral vascular disease, COPD, diabetes, tumor, leukemia, AIDS etc., all of which are known to increase risk of pneumonia [28]. Additionally three more disease categories relevant for cases of pneumonia are included in co-morbidity scoring. Those are diagnosis of Hypertension, HIV (in addition to AIDS) and Gastro - oesophageal reflux. All the co-morbid diagnoses are presented in *Table 2***Error! Reference source not found.**.

Since diabetes is quite important risk factor for pneumonia, it needs to be treated more carefully [29]. Therefore, diabetes diagnoses are extracted from the Danish National Diabetes Register (NDR), which gives more details then using only LPR data. NDR contains information from 3 different sources, such as the National Patient Register (NPR), the health insurance databases (NHISR) and pharmacies records (RPMS) [30].

The Danish National Registry of Patients is used to obtain previous diagnosis for each disease included in the Charlson index. We extracted diagnosis for each study member using hospital discharges, which are coded according to ICD – 8 and ICD – 10.

	Disease	ICD 8	ICD 10	Score
1	Myocardial infarction	410	I21;I22;I23	1
2	Congestive heart failure	427.09; 427.10; 427.11; 427.19; 428.99; 782.49	I50; I11.0; I13.0; I13.2	1
3	Peripheral vascular disease	440; 441; 442; 443; 444; 445	I70; I71; I72; I73; I74; I77	1
4	Cerebrovascular disease	430-438	I60-I69; G45; G46	1
5	Dementia	290.09-290.19; 293.09	F00-F03; F05.1; G30	1
6	Chronic pulmonary disease	490-493; 515-518	J40-J47; J60-J67; J68.4; J70.1; J70.3; J84.1; J92.0; J96.1; J98.2; J98.3	1
7	Connective tissue disease	712; 716; 734; 446; 135.99	M05; M06; M08; M09; M30; M31; M32; M33; M34; M35; M36; D86	1
8	Ulcer disease	530.91; 530.98; 531-534	K22.1; K25-K28	1
9	Mild liver disease	571; 573.01; 573.04	B18; K70.0-K70.3; K70.9; K71; K73; K74; K76.0	1
10	Diabetes type1 Diabetes type2	249.00; 249.06; 249.07; 249.09 250.00; 250.06; 250.07; 250.09	E10.0, E10.1; E10. E11.0; E11.1; E11.9	1
11	Hemiplegia	344	G81; G82	2
12	Moderate to severe renal disease	403; 404; 580-583; 584; 590.09; 593.19; 753.10-753.19; 792	I12; I13; N00-N05; N07; N11; N14; N17-N19; Q61	2
13	Diabetes with end organ damage - type1 - type2	 249.01-249.05; 249.08 250.01-250.05; 250.08	 E10.2-E10.8 E11.2-E11.8	2
14	Any tumor	140-194	C00-C75	2
15	Leukemia	204-207	C91-C95	2
16	Lymphoma	200-203; 275.59	C81-C85; C88; C90; C96	2
17	Moderate to severe liver disease	070.00; 070.02; 070.04; 070.06; 070.08; 573.00; 456.00-456.09	B15.0; B16.0; B16.2; B19.0; K70.4; K72; K76.6; I85	3
18	Metastatic solid tumor	195-198; 199	C76-C80	6
19	AIDS	079.83	B21-B24	6
(20)	Hypertension	400-404	I10-I15	1
(21)	HIV (in addition to AIDS)		B20	1
(22)	Esophageal reflux	530.99	K21	1

Table 2: Discharge diagnoses translation of the co-morbidity diseases defined by Charlson and additional 3

Chapter 3

Air Pollution

Air pollution is ubiquitous exposure that affects most people, especially the majority of population living in urban areas. Our main interest is to investigate the effect of traffic-related air pollution to risk of pneumonia. Therefore, the aim of this chapter is to introduce air pollution exposure used in the analysis as a short introduction by its classification, followed by data available and used in this analysis.

Traffic-related pollution is nowadays the major threat to clean air in urban areas. In epidemiological studies traffic - related air pollution is defined typically by measure (central) exposure or modeled estimated (at residence) exposure to NO_2 , SO_2 , $PM_{2.5}$ or UFPs, and/or more simple proxy such as residential proximity to busy roads, calculated by GIS (Geographic Information System) [16].

3.1 Classification of Air Pollutants

Common ambient air pollution can be grouped into two large classes: gasses, which are, measured by their chemical composition, and include sulfur dioxide (SO_2), nitrogen oxide (NO_x), carbon monoxide (CO), and ozone (O_3), and particles (PM), which have mixed and complex chemical structure and are thus measured by their physical properties, such as mass and number.

3.1.1 Gasses

Sulfur dioxide (SO_2) is prevalent in all raw materials, including crude oil, coal, and ore that contains common metals like aluminum, copper, zinc, lead, and iron. In the atmosphere SO_2 originates mainly from combustion of fossil fuels from stationary sources (heating, power generation) and in motor vehicles.

Nitrogen oxide (NO_x) is the generic term for a group of highly reactive gasses containing nitrogen and oxygen in varying amounts and it is form when fuel is burned at high temperatures, as in a combustion process. The primary sources are motor vehicles, and all the sources that burn fuels.

NO_2 is generated from reaction of NO and O_3 in the ambient air and it is a respiratory tract irritant that causes a spectrum of adverse health effects, depending on the dose of exposure. It

may also contribute to susceptibility to respiratory infections, especially in young and elderly, while in confined spaces, severe injury and even death may occur [31].

3.1.2 Particulate matter

Particulates, or particulate matter (PM), are tiny particles of solid or liquid suspended in the air. It is container or mix of many different components (chemical elements) from various sources, with local and regional variation affecting its toxicity. PM is the pollutant that has been most studied and most consistently associated with health effects. Particulate matter is commonly presented in size cuts, which are given in µm.

$PM_{2.5}$ (particles with aerodynamic diameter of 2.5 µm or less) is known as fine particles (FPs). It is measured by its mass or mass concentration, typically in unit $\mu m/m^3$.

The smallest particles, those with particles aerodynamic diameter of 0.1 µm or less, are known as ultrafine particles (UFPs). They are different from the large PM fractions because they contribute very little to the mass, but occur in magnitude higher numbers. Thus, UFPs are instead of mass, measured by numbers of number concentrations (number of particles/m^3) [32].

Deposition of PM is the airways depend on the particle size, anatomy of the airways and breathing. Coarse particles are deposited mainly in the upper airways. Particles less than 10 µm can be deposited further down in the bronchi, whereas particles with smaller diameters (FPs and UFPs) can travel all the way into alveoli, affecting lungs [33,34].

3.2 AirGIS Model

The Danish air pollution and human exposure modelling system (AirGIS model [35]) is based on a geographical information system (GIS), and used for estimating traffic-related air pollution with high temporal (an hour) and spatial (individual address) resolution. AirGIS calculates air pollution at a location as the sum of three contributors:

1) Regional background, estimated from trends at rural monitoring stations and from national vehicle emissions [36].
2) Urban background, calculated from a simplified urban background (SUB) procedure that takes into account urban vehicle emission density, city dimensions (transport distance), and average building height (initial dispersion height) [37].
3) Local air pollution from street traffic, calculated with the Operational Street Pollution Model (OSPM) from data on traffic (intensity and type), emission factors for each vehicle type and EURO class, street and building geometry, and meteorology [38].

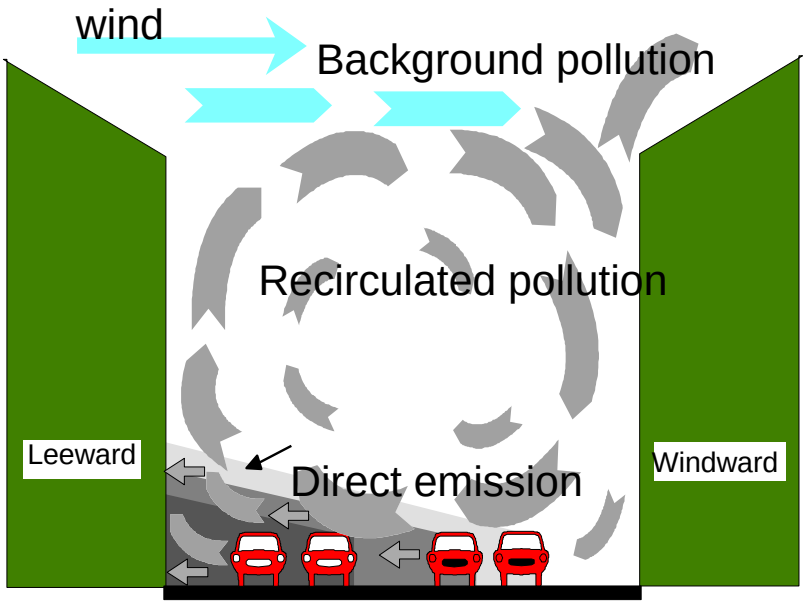

Figure 2: Schematic illustration of the flow and dispersion inside a street canyon (Berkowitz, 2000)

Input data for the AirGIS system come from various sources: a GIS-based national street and traffic database, including construction year and traffic data for the period 1960–2005 [39], and a database on emission factors for the Danish car fleet [40], with data on light - and heavy - duty vehicles dating back to 1960, built and entered into the emission module of the OSPM. A national GIS database with building footprints supplemented with construction year and building height from the national building and dwelling register, national survey and cadastre

data-bases, and a national terrain-evaluation model, provided the correct street geometry for a given year at a given address. The geocode of an address refers to the location of the front door with a precision within 5 m for most addresses. With a geocoded address and a year, the starting point is specified in place and time, and the AirGIS system automatically generates street configuration data for the OSPM, including street orientation, street width, building heights in wind sectors, traffic intensity and type, and the other data required for the model. Air pollution is calculated in 2 m height at the façade of the address building.

The dispersion models used to assess NO_2 levels have been successfully validated against measured values. It has also applied in several studies, for instance in the studies of asthma, lung cancer and COPD in this cohort [11,12]. The AirGIS mode has been validated in two major ways. One way was to look at the correlation between modeled and measured half - year mean of NO_2 concentrations at 204 positions in the greater Copenhagen area, which gave us a correlation coefficient (r) of 0.90 with measured concentrations being on average 11% lower than the modeled [37]. We also compared modeled and measured one - month mean concentrations of NO_x and NO_2 over a 12 - year period (1995 - 2006) in a busy street in Copenhagen (Jagtvej, 25 000 vehicles per day, street canyon), which showed correlation coefficients (r) of 0.88 for NO_x and 0.67 for NO_2. The modeled mean NO_x concentration over the whole 12-year period was 6% lower than the measured [41]. Thus, the model predicted both geographical and temporal variation well.

However, there are always some limitations that we have to be aware of. The exposure assessment method considers only outdoor concentrations at the residential addresses but not the indoor neither the work address, which might have some effect on the overall exposure. As we have no data on work address, outdoor concentrations of NO_2 at residence will be used as a proxy of personal exposure, which results in some exposure misclassification. The use of outdoor levels of air pollution is a gold-standard in air pollution epidemiology [12,17,18,42], since personal measurements are expensive and not feasible in cohort studies. Furthermore, it has been documented that outdoor concentrations are reasonable proxies of personal exposure, since indoor penetration of traffic-related air pollution is high, and correlation between personal and outdoor concentrations for particles is high where for gases it should be even higher.

3.3 Exposure assessment

The Danish GIS – based air pollution and human exposure modeling system (AirGIS) was used to model outdoor concentrations of traffic pollution at the residential addresses since 1971. The air pollution concentration values are taken for all cohort members with 80% or better residential history. Missing values due to missing address or missing geographical coordinates were substituted by the levels calculated for the proceeding address or, when the first address was missing, for the subsequent address.

For each cohort member the exposure was assessed from the residential address history since 1971, which was used to model outdoor levels of nitrogen dioxide (NO_2) and nitrogen oxides (NO_X) with the Danish AirGIS dispersion modeling system.

Input for AirGIS model, as already explained in previous section, is:

- Street / building geometry (street width, distances, building height, open sector)
- Street network and traffic data (emission factor, density, speed, types, variation patterns over time)
- Meteorology (temperature, wind speed, wind direction, solar influx)

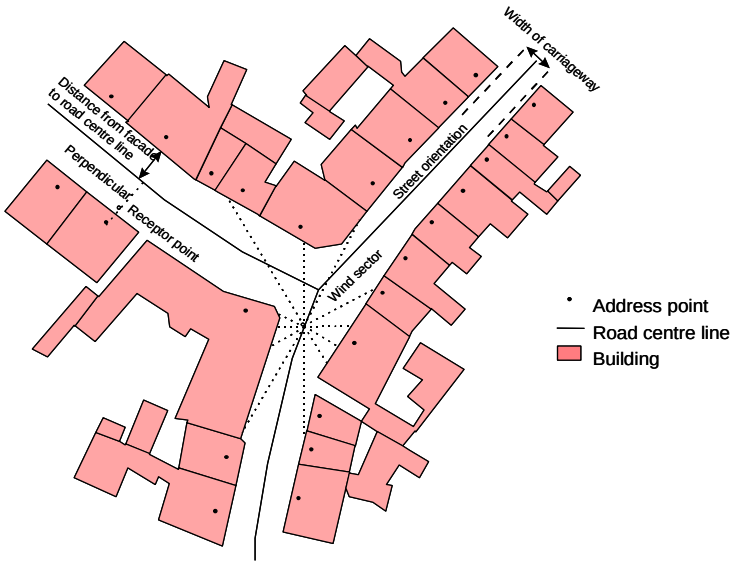

Figure 3: The 2½ dimensional Urban Landscape Model of the AirGIS system that automatically generates required street configuration and traffic input data for the Operational Street Pollution Model (OSPM)

The output is air pollution exposure in terms of yearly mean NO_2 and NO_X concentrations at the residential addresses for all cohort members since 1971.

We also defined six air pollution proxies based on traffic data at the residential address at recruitment (1993 – 1997):

- The presence of major road (density ≥ 5 000 vehicles/day) within a 50m radius
- The presence of major road (density ≥ 5 000 vehicles/day) within a 100m radius

- The presence of major road (density ≥ 10 000 vehicles/day) within a 50m radius
- The presence of major road (density ≥ 10 000 vehicles/day) within a 100m radius

- Traffic load, as the total number of kilometers driven by vehicles within a 100m radius
- Traffic load, as the total number of kilometers driven by vehicles within a 200m radius

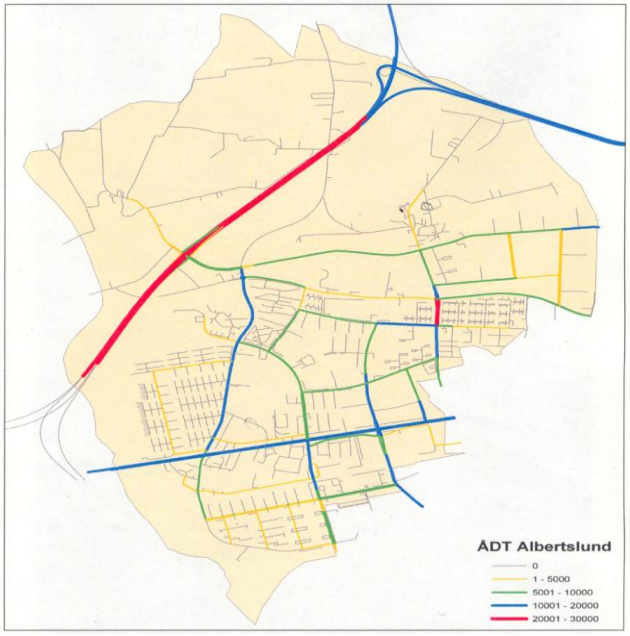

Figure 4: Schematic representation of traffic loads in Albertslund, Denmark

Chapter 4

Methodology

This section gives some general theoretical background of analyzing the survival data. The statistical approaches used in this study are presented. First, an introduction to survival analysis is given, by basic definitions with notation, followed by its most used estimations. Then the main concept of the Cox proportional hazard model is presented from the theoretical aspect, with interpretation and validation, and also possible extensions as improvements of the basic model.

4.1 Introduction to Survival Analysis

The techniques for studding the outcome variable of interest as *the time until an event* were primarily developed in the medical and biological sciences. The event of interest in this case is most often the occurrence of the disease or death, giving the name Survival Analysis. The procedures used for analyzing the survival data are widely used in other areas too. For example in economics and sociology, so called duration analysis, or in engineering when one might wish to study time in use of a machine, which is called failure time analysis. Nevertheless, our focus is on biomedical data analysis [43].

In a survival analysis, we usually refer to the time variable as survival time. This name comes from the concept that an individual had "survived" over some follow-up time, which can be measured as the calendar time in years, months, weeks, days, etc. or alternatively age of individual, from the beginning of follow – up period until the event occurs. It doesn't have to mean that event is a negative individual experience; it can also be the time until person recovers, or goes back to work. The person's survival time is denoted by T, and any specific value of interest for the random variable T is denoted by t.

4.1.1 Censoring and truncation

The duration of the study is most often limited in time. Therefore, in survival analysis one has to consider the subjects key analytical problem called *censoring*. In essence, censoring occurs when we have some information about the individual survival time, but don't know it exactly. Hence, the data consists of complete and incomplete observations so ordinary linear regression

or other standard statistical methods can't be applied and that is why survival data require specific statistical theory.

The incomplete observations are termed censored survival times. The reasons for censoring might be when a person does not experience the event before the study end, a person is lost to follow – up during the study (e.g. moved) or when a person withdraws from the study because of some other event occurs that affects outcome of interest (e.g. death in case of studding the certain disease occurrence) or some other reason. We generally refer to this kind of data as *right – censored.* This is simply denoted by indicator variable with value 1 for event occurrence, or 0 for censorship.

Furthermore, in a clinical study the initial event could be time of entry the study, time of admission to hospital, time of diagnosis etc, which corresponds to time 0 in the study time scale. The set of individuals for whom the event has not occur before the given time *t*, and who has not been censored before *t*, is termed the risk set at time *t*. Quite often there is a case of having different starting times for subjects under observation [43,44]. Although modeling survival data with age as time scale has similar expression in the models with time-on-study or calendar time as time scale, implicit mechanisms are many ways different. For example, at a given age, some subjects are not yet under observation whereas others may not be anymore. Therefore, the number of subjects at risk does not vary monotonically with age and risk sets are not nested. This structure defines and an *open cohort*, under which a subject's observation is conditional to some characteristics at the recruitment, like pre-existing health condition, place of birth etc. Thus, using age as the time scale implies delayed entry with left-truncation occurring at the age at inclusion. Alternative time scale is calendar time with models adjusted for age, however age as underlying time scale is documented as the most unbiased and therefore mostly recommended time scale [45]. (*Figure 5*)

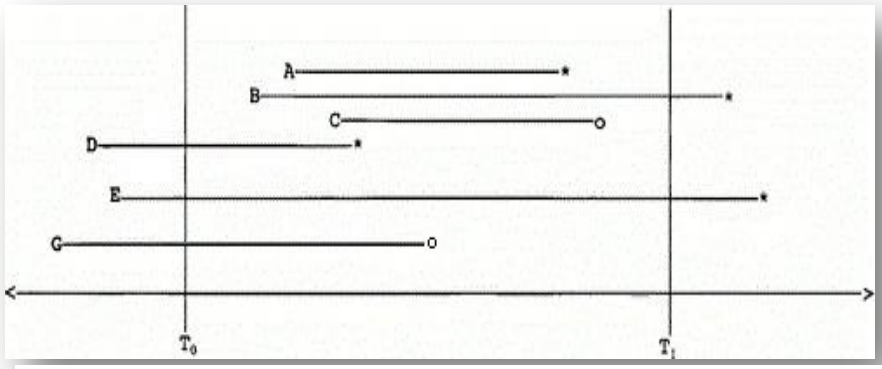

Figure 5: Graphical presentation of left-truncated data

4.2 Survival function and hazard rate

The survival data can't be analyzed by ordinary statistical methods because of censoring and truncation. However, the concept for these analyses is not complicated. Two important terms needed are *survival function* and *hazard rate*.

Basic terms needed to easier explain the concept are the probability density function (*pdf*) $f(t)$ of a continuous random variable:

$$\boldsymbol{P(t \leq T < t + \Delta t) = f(t)\Delta t} \qquad (4.1)$$

which describes the relative likelihood for an individual to have an event of interest in the time interval $[t, t + \Delta t)$. And cumulative distribution function (*cdf*) is:

$$F(t) = P(T \leq t) \qquad (4.2)$$

The survival function, $S(t)$, gives the expected proportion of individuals for whom the event has not yet happened by time t, for the predefined set of followed individuals. So, the survival function specifies the unconditional probability that the event of interest has not happened by time t.

$$S(t) = P(T > t) = \int_t^{\infty} f(s)ds = 1 - F(t) \qquad (4.3)$$

The visualization of this can be done by plotting the survival curves of the survival functions. Theoretically, time is a continuous random variable ranged from zero to infinity, so that gives the smooth curve starting at study time 0 where all the individuals are under the risk, and decreasing over time tending to 0 when the time goes to infinity (*Figure 6 – left*).

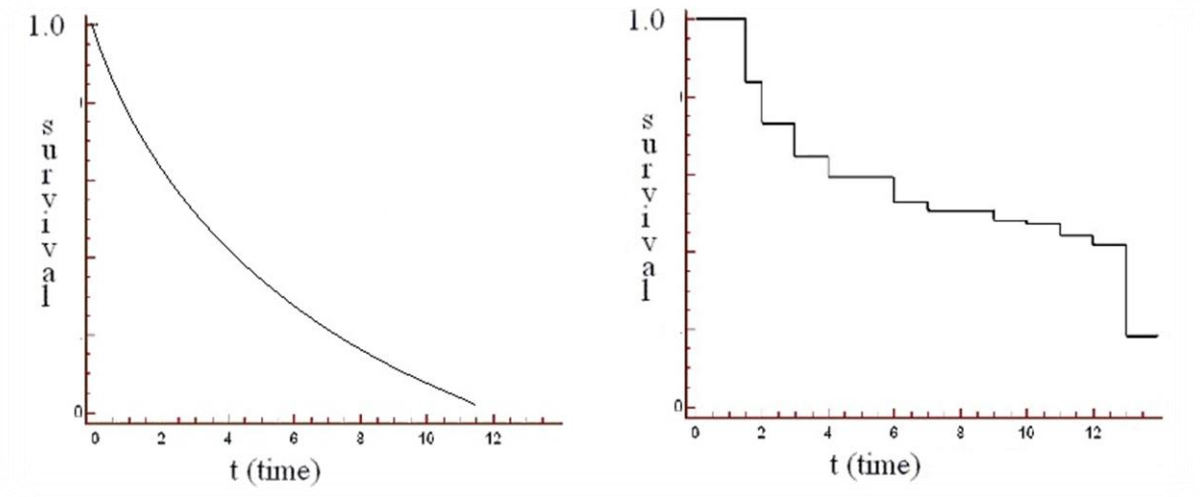

Figure 6: Graphical presentation of Survival curves – example
Left: Smooth curve - in theory; right: Step function – jumps at the end of intervals – real case scenario

In practice the situation is a bit different. The survival curves are step function rather than smooth curves with jumps at the end of time intervals. It is also quite usual that the survival function decreases towards a positive value at the study end (*Figure 6 – right*) [43].

The hazard rate, $\lambda(t)$, gives the instantaneous potential per unit time for the event to occur, given that the individuals have been under the risk up to time t. In contrast to the survival function, the hazard rate is defined by means of a conditional probability. Assuming that T is continuous, that it has probability density, one looks at the individuals who have not yet experienced the event of interest by time t and considers the probability of having the event in the small time interval stating at t, $[t, t + \Delta t)$.

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t \mid T \geq t) \qquad (4.4)$$

Note that, the hazard rate and survival function are giving opposite information. The survival function focuses on not experiencing the event, i.e. surviving, and the hazard rate focuses on occurrence of event, i.e. failing; and while the survival curve is a function that starts at 1 and declines over time, the hazard rate can essentially be any nonnegative function. The relation between hazard and survival function is given as:

$$\lambda(t) = \frac{f(t)}{S(t)} \qquad (4.5)$$

This relation makes it fairly easy to obtain both functions by knowing only one [43,46].

4.3 Counting process formulation

Comparing to the basic description of survival data, where we only account for time to the event of interest (T_i) and censoring status (δ_i), the concept of counting processes broads the scope of survival analyses to more elaborate processes. Counting process replaces the pair of variables (T_i, δ_i) with the pair of functions $(N_i(t), Y_i(t))$, where $N_i(t)$ represents the number of observed events within the interval $[0, t]$ for subject i and $Y_i(t)$ the status variable at time t defined as:

$$Y_i(t) = \begin{cases} 1, & \text{subject } i \text{ is under oservation and at risk at time } t \\ 0, & \text{otherwise} \end{cases}$$

Here, $Y(t)$ is left-continuous deterministic function based on past – predictable process, whose value at any time t is known infinitesimally before t. And $N(t)$ is right-continuous step function - counting process.

$N(t)$ represents the total number of events precisely at time t, and $N_i(t)$ the number of events at time t for each subject i under observation. Whereas $Y(t)$ presents the number of subjects under observation and at risk at time t [46,47].

$$N(t) = \sum_{i=1}^{n} N_i(t) \qquad Y(t) = \sum_{i=1}^{n} Y_i(t)$$

This formulation generalizes analysis to multiple events and multiple at-risk intervals. However, the later is out of the scope of this study.

4.4 Estimation

The most common estimator of the survival function is the *Kaplan – Meier estimator,* which is the product limit method and estimates the survival function directly from the continuous survival time. It is expressed as:

$$\hat{S}_{KM} = \prod_{t_i \le t} \left(1 - \frac{N(t_i)}{Y(t_i)}\right) \tag{4.6}$$

Where the time interval $[0, t]$ is partitioned into smaller time intervals $0 = t_0 < t_1 \ldots < t_k = t$, and $N(t_i)$ events in the time interval up to time t_i , and $Y(t_i)$ individuals at risk prior to t_i [43].

Another estimator for the survival function was suggested by Therneau and Grambsch, and that is *Breslow estimator*:

$$\hat{S}_B = \prod_{t_i \le t} \exp\left\{-\frac{N(t_i)}{Y(t_i)}\right\} \tag{4.7}$$

It is quite similar to Kaplan – Meier estimator when there are many subjects at risk. For the finite samples, the relation $\hat{S}_B(t) \geq \hat{S}_{KM}(t)$ holds, since $e^{-x} \geq 1 - x$.

The estimation of hazard rate is in literature proven to be much easier on the cumulative hazard

$$\Lambda(t) = \int_0^t \lambda(s) ds \tag{4.8}$$

instead of hazard function itself, which follows from the fact that it easier to estimate cumulative distribution function than probability density function [44].

The *Nelson-Aalen* is the most common non-parametric estimator of the cumulative hazard function based on a right censored data:

$$\widehat{\Lambda}(t) = \sum_{t_i \leq t} \frac{N(t_i)}{Y(t_i)} \qquad (4.9.1)$$

Intuitively, this expression is estimating the hazard at each distinct time of event t_i as the ratio of the number of events to the number at risk. The cumulative hazard up to time t is simply the sum of the hazards at all event times up to t, and has a nice interpretation as the expected number of events in $(0, t]$ per unit at risk. This estimator has a strong justification in terms of the theory of counting processes [46].

The relation between cumulative hazard and survival function is $\Lambda(t) = -log\big(S(t)\big)$, where the survival function can be based on Kaplan – Meier or Breslow estimate.

The Nelson - Aalen estimator is essentially a method of moments estimator and thereby the variance can be estimated consistently by:

$$\widehat{\Lambda}(t) = \sum_{t_i \leq t} \frac{N(t_i)}{[Y(t_i)]^2} \qquad (4.9.2)$$

However, Therneau and Grambsch suggest the alternative as the approximation for the log-transformation because it improves the accuracy of the confidence intervals.

4.5 Cox proportional hazard model

The Cox model is a well - recognized statistical technique for analyzing survival data. The purpose of the model is to simultaneously explore if there is an effect of one or several variables on the survival. The Cox model is semi-parametric that specifies the hazard of ith subject as:

$$\lambda_i(t) = \lambda_0(t) e^{(X_{i1}\beta_1 + X_{i2}\beta_2 + \ldots + X_{ip}\beta_p)} \Leftrightarrow \lambda_i(t) = \lambda_0(t) e^{X_i \beta} \qquad (5.1)$$

Where first part is non-parametric, unspecified nonnegative function of time λ_0, which can take any form, is called the baseline hazard. X_i is a covariate for ith subject under the observation; and β is a p - dimensional column vector of coefficients representing the effect of the covariates. The exponential form ensures that the estimates are physically possible, since the event rates can't be negative because once we have the event it can't "unhappen".

The advantages of the Cox model are the simplicity of direct influence of covariates through their linear or log-linear combination and flexibility that baseline hazard gives to the model since no specific distribution is assumed for the baseline group.

Another advantage is very easy interpretation of the regression parameters as relative or log-relative risks. The value of parameter may be interpreted as the change in relative risk when the covariate is increased by one unit and the model is corrected for the other covariates.

The name *proportional hazard* model comes from the fact that the hazard ratio is constant over time. For two subjects i and j with fixed covariates X_i and X_j we have:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)e^{X_i\beta}}{\lambda_0(t)e^{X_j\beta}} = \frac{e^{X_i\beta}}{e^{X_j\beta}} \tag{5.2}$$

Proportionality of the hazards is the key assumption of the Cox regression model.

4.5.1 Estimation

Because of the semi-parametric nature of the model, one can't use ordinary likelihood methods to obtain estimates. Therefore, for estimating covariates parameters β, Cox developed a nonparametric method he called partial likelihood. Estimation of parameter values is then obtained by use of maximum partial likelihood estimation [46].

For uncensored subjects $i = 1, \dots, d$ and censored $= d+1, \dots, n$, the *partial likelihood* is presented by:

$$L(\beta) = \prod_{i=1}^{d} \frac{exp\ (X_i\beta)}{\sum_{j\in R(t_i)} exp\ (X_j\beta)} \tag{5.3}$$

where $R(t_i)$ in denominator is summing over all individuals in the risk set $Y_j(t)$.

By log-transforming partial likelihood we get:

$$l(\beta) = logL(\beta) = \sum_{i=1}^{d}[X_i\beta - log\ (\sum_{j\in R(t_i)} exp\ (X_j\beta))] \tag{5.4}$$

naturally called *log partial likelihood*.

In general, the partial likelihood is not ordinary likelihood in sense of being proportional to the probability of an observed dataset, however it can still be treated as a likelihood for purposes of asymptotic inference [46].

The differentiated log partial likelihood $l(\beta)$ with respect to β, is the $p \times 1$ gradient vector called *score vector* of the form:

$$U(\beta) = \frac{\partial}{\partial \beta} l(\beta) = \sum_{i=1}^{d} [X_i - E(\beta)] \tag{5.5}$$

where the expectation is

$$E(\beta) = \frac{\sum_{j \in R(t_j)} \exp(X_j \beta) X_j}{\sum_{j \in R(t_j)} \exp(X_j \beta)} \tag{5.6}$$

And the maximum partial likelihood estimator $\hat{\beta}$ is found by solving the partial likelihood equation:

$$U(\hat{\beta}) = 0$$

For the real data with big dimensions this is very demanding and the computer algorithms are designed to deal with it. Functions which are used to fit a Cox proportional hazard regression model most often use the Newton-Raphson algorithm for solving the partial likelihood equation [46].

In large sample cases the maximum partial likelihood estimators have properties similar to ordinary maximum likelihood. In particular, $\hat{\beta}$ is in large samples approximately multivariate normally distributed around the true parameter value β with a covariate matrix that may be estimated by the inverse of the expected information matrix $I(\hat{\beta})^{-1}$ [44].

4.5.2 Test statistics

In order to test the null hypothesis $H_0: \beta = \beta_0$, one may apply the usual likelihood-based tests. Three most common test statistics will be presented here. Those are likelihood ratio, score and Wald test statistics.

- The Likelihood ratio test st.: $\chi^2_{LR} = 2(logL(\hat{\beta}) - logL(\beta_0) = 2(l(\hat{\beta}) - logl(\beta_0))$
- The score test statistics: $\chi^2_S = U(\beta_0)^T I(\beta_0)^{-1} U(\beta_0)$
- The Wald test statistics: $\chi^2_W = (\hat{\beta} - \beta_0)^T I(\hat{\beta})(\hat{\beta} - \beta_0)$

These three test statistics are asymptotically equivalent and all have chi-square distribution with *p* degrees of freedom under the null hypothesis giving the consistent parameters estimator $\hat{\beta}$ [46,47].

4.5.3 Functional Form

The Cox model assumes the proportional hazard structure with a log-linear model for the covariates, that is, with fixed covariates X_i :

$$\lambda_i(t) = \lambda_0(t)e^{X_i\beta}$$

For continuous variables this implies that the ratio is the same for all subintervals on the variable scale. However, the data may show the threshold effects, usually for upper and lower range. For that reason, one should explore the correct functional form for the covariates.

One of the simplest suggested approaches to examine if the proportional hazard assumption is violated is to plot the residuals from a null model against each covariate separately and superimposing a scatter-plot smoother. Intuitively, this is similar to the ordinary plots of response variable against each predictor used for uncensored data in linear models. However, as in uncensored data cases, this method may fail when correlations are present and also one should address appropriate weighting of the observations to account for different follow-up time.

Another approach, which addresses linear and nonlinear relationship of covariates, is by using Poisson regression approach. In this case one can use any modeling tools available in programs for Poisson regression analysis and tease out the appropriate functional form. This is quite good method if the tools for Poisson data exist, but still involves very complex manipulations.

However, an alternative is to model the functional form directly in the Cox model functions by fitting some available functions of the covariates. Particularly useful classes of functions for this purpose are *regression* and *smoothing splines* as flexible fitting functions.

Splines are a very good way for exploring nonlinear relationship of covariates. Regression splines are a general tool in many statistical softwares, and therefore easy to implement. However, even though smoothing splines are easy to understand but they have high computational requirements. Therefore, we restrict ourselves here on regression splines.

The regression splines have several important properties. One useful property is locality of the influence, which, for example, doesn't hold for polynomials. Then, those curves can be constrained to be linear beyond the last control point and that form of spline fit is often called *natural splines* or *restricted cubic splines*. The spline curves are controlled by the number of degrees of freedom and after the choice has been made it can directly be implemented in the proportional hazard model. Therefore, this is the most recommended way for analyzing the functional form of the covariates, especially in non-linear case. They are easy to fit, and they are computed within the Cox model standard tests of hypothesis. Confidence intervals are easily added as well [46,47].

4.5.4 Testing proportional hazards assumption

The key assumption of the Cox regression model is the proportional hazard structure. This might fail in many ways. Therefore, the assumptions need to be validated in order to verify the use of Cox model, i.e. the relationship between subjects for any variable in the model needs to be independent of time.

Plotting the residuals against time is one way of evaluation. This is visualization method where a line can be fit to the plot followed by a test for zero slope. If the test shows a slope significantly different from zero, one has the evidence of validation of the hazard proportionality assumption.

Furthermore, many statistical softwares have an implemented function for checking the proportional hazard assumption, so it is a trivial check when performing the data analysis. It gives the test statistic and $p - value$ for significance. In cases with high number of observations this method is more appropriate since the plots require sometimes reduced number of observations to be readable [43,46,47].

4.6 Extending the Cox model

In the ordinary Cox model only one kind of event and just one (first) event occurrence per subject is considered. The procedures for this kind of analysis are well-developed, simple to implement and interpret, which have very useful properties. However, the concern is still a waste of available information and therefore more details could be included. This improves the statistical power of the analysis and is used to more detailed investigations, but on the other hand the models to be performed are more complex.

There is number of possible extensions of the general survival models and they are quite intuitive. First possible extension is when each subject can have the same type of event multiple times; that is ***recurrent events***. E.g. occurrence of the same type of disease for a number exposed subjects might be of interest more than only once during the follow-up, since the effect of exposure might be stronger after already experiencing the disease.

Second is called ***competing risks*** analysis, which is performed when only one event per subject is of interest, but that event might be of different types. E.g. the event of interest might be death but from different causes.

And third kind of extension of the general model is when we have both several events and several types of events per subject, which belong to a group of ***multi - state models***.

In ordinary Cox analysis of survival data each observation is considered as independent. In case of multiple events this becomes a problem when estimating the parameters' variance, since a single subject with multiple events has number of rows which are not independent, and that needs to be taken into account. Also, heterogeneity across individuals should be addressed because some individuals might be more prone to disease than the others, in relation to some known or unknown factors. These are some of the important issues to be addressed when extending the ordinary Cox analysis.

Furthermore, it is important to understand if the data set is unordered or ordered. In case of unordered data, there might be correlated groups within the data set, but still the outcome within the group is unordered, meaning that the event of interest for all the individuals in the group might occur without any within-group-order. On the other hand, ordered data's expected outcome are sequential multiple events of the same type per subject [46].

This study has one specific event of interested – pneumonia, which can occur multiple times for each subject, so our focus is on the recurrent data. This section will be just a part of all possible Cox model extensions, and aims to describe different models for recurrent events of ordered data.

4.6.1 Robust variance for recurrent events

An important issue to be addressed, when working with multiple events, is the estimation of the variance component. The variance estimation for $\hat{\beta}$ parameters of the data with recurrent events can't be obtained as it was for a single event per subject, because all the observations are no longer independent. Now the observations within each subject with multiple events must be considered as a cluster. This can be done by using *grouped jackknife* as the variance estimate for correlated data [46].

In general, *jackknife* is an alternative method used to derive a robust estimate of the variance for the Cox model. Jackknifing is a method that calculates the difference between the estimated parameters ($\hat{\beta}$) and the estimates of the parameters leaving one observation out at the time ($\hat{\beta}_{(i)}$). The variance estimation is therefore systematically recomputed for each observation.

Grouped jackknife does the same but for correlated data. This means, if the observations are not independent within a subject, grouped jackknife takes that into account and in every step leaves out one group of correlated data at the time, instead of one observation at the time. This way it provides a sober and robust estimate of the variance for the parameters in the model.

However, a disadvantage of jackknife method is from the computational aspect. It is very intensive because recomputations are done in each step of estimation. Nevertheless, in some statistical softwares (e.g. R, SAS, Stata) there is straight forward implementation, just by

introducing *cluster* term for subjects with multiple events in the model. So the robust variance estimate for correlated data has high computational requirement but is possible to be obtained [46].

4.6.2 Models for recurrent events

The survival data with recurrent ordered outcome, i.e. sequential multiple events of the same type over time per subject, is lately of increasing interest. Modeling occurrence of repeated events is very important from medical point of view, since many diseases are expected to be recurrent [48]. This type of the analysis can be done by several approaches. However, the biggest challenge in this case is to set up the data and, in contrast to ordinary Cox model which is quite general and simple to use, these models have very high computational requirements.

The main issues to address when considering recurrent event data is potential correlation among events, which violates the Cox model's assumption that events occur independently, and dissimilarity of studied population. This leads to two important consequences: Cox model is both biased and inefficient. Therefore variations of the Cox model have been proposed for estimation with recurrent events to account for the events correlation and individual heterogeneity, those are:

- Variance-corrected models,
- Frailty models, and
- Conditional models

It is usually suggested to start the analysis with very simple model and proceed using models with increasing complexity, by correcting the variance, accounting for events correlation and individual heterogeneity. Thus, we present the simple intensity-based model and then more complex, extended Cox models of recurrent event analysis: Andersen-Gill, Conditional Andersen-Gill; Frailty and Conditional frailty model [46,49].

4.6.2. a Intensity – Based model

This is the simplest approach based on ordinary Cox model. It has strong assumptions but it very easy to implement and visualize. This class is suitable to model the full dynamics of the recurrent events process and the likelihood is available like in the general Cox model. It is recommended for calendar time data [50].
The model is defined as

$$\lambda_i(t) = \lambda_0(t)e^{X_i\beta}$$

which is formally identical to the ordinary Cox model for survival data. Data is presented the same as in the proportional hazard mode, $N_i(t)$ as number of event at time t for $i\ th$ subject,

$Y_i(t)$ presents at-risk process for $i\ th$ subject prior to time t, and $i\ th$ subject's covariates X_i, where the subjects i=1,2, … , n are independent.

Here we allow recurring events per subject which results in several observations per each subject. Thus, data set will contain one row for each observation and time intervals correspond to time since entry to first event or from the last occurrence of event until a new one, and so on until the censoring or end of follow-up. Individuals without events have only a single observation and thus one row of data.
Difference from the ordinary Cox model is in at-risk process $Y_i(t)$. In time to first event case, it would go from one to zero at the time t when an event occurs, and in the multiple events case it remains one as events occur. Visual form of this model can be seen on *Figure 7*, where arrow represents and event, and box the at-risk set, which is in this case the same after each event.

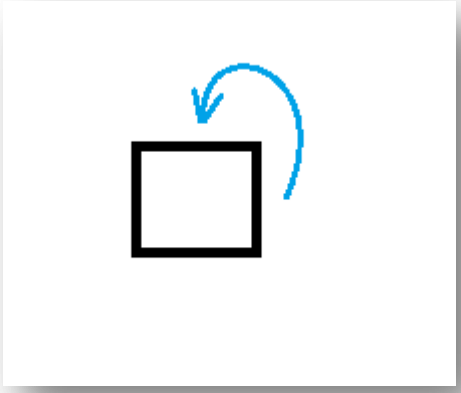

Figure 7: The intensity-based model for recurrent data – schematic
Arrow = event, Box = risk set

Even though intensity-based model takes multiple events into account, its disadvantage is that it assumes independency between all the events and it doesn't distinguish when events belong to a certain subject.

4.6.2. b Andersen - Gill model

Very similar to intensity-based model is Andersen-Gill model. It is a counting process model, very easy to implement and interpret, but also with strong assumptions. Their schematic form is the same (*Figure 7*).

Anderson-Gill model also assumes that all events are independent but the improvement compared to intensity-based model is that every subject is considered as one individual cluster. This way each subject contributes to the risk set for an occurrence of event as long at the subject is under observation at the time it experiences the event. Therefore, Andersen-Gill model considers correlation due to multiple events per subject by adjusting the standard error

estimates using robust variance described in Section 4.6.1. Still, disadvantage of this model is that it can't make the difference between the occurrence of first and second, second and third event and so forth. So, the only improvement that this model gives is in robust variance estimation [46].

This variance-corrected model presents one way of dealing with the efficiency problem produced by heterogeneity across individuals. Even though Andersen-Gill corrects the variance, it still does not incorporate the heterogeneity into the estimates themselves and therefore remain biased [49].

4.6.2. c Frailty model

In contrast to variance-corrected models, such as Andersen-Gill, frailty or random effect models deal with subject's heterogeneity by making assumptions about frailty distribution and incorporating it into the model estimates. The idea of frailty model is to consider that individuals are dissimilar, i.e. some subjects are more susceptible to experience the event of interest than are others, and that some are more likely to have second, third and so forth event, than are others. To consider the possible heterogeneity of a subject, the distribution can be at least approximated [49].

To visualize this approach we can again using arrows as events and boxes as individuals, and the schematic form is as presented in *Figure 8*.

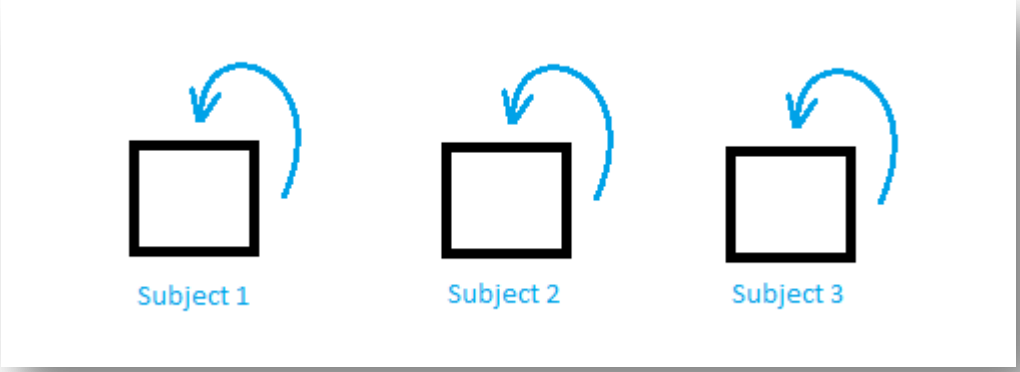

Figure 8: The frailty model for recurrent data – schematic
Arrow = event, Box = risk set

The basic concept of frailty model is to add an unmeasured random effect in the hazard function to account for heterogeneity in subjects. Therefore, the structure of frailty model corresponds to proportional hazard framework with additional random effect as a continuous variable that describes excess risk or frailty for individuals.

It has the form:

$$\lambda_i(t) = \lambda_0(t)e^{X_i\beta + W_i u} \tag{5.7}$$

where W_i is a frailty term from a probability distribution with mean 0 and variance 1, u vector of coefficients, and again, $\lambda_0(t)$ is the baseline hazard function, β vector of coefficients and X_i matrix of observed covariates. For zero parameter u we get standard proportional hazard model, and if it's not zero then we allow for unmeasured factors, which affect the hazard rate.

When considering the same model in another form:

$$\lambda_i(t) = Z_i\lambda_0(t)e^{X_i\beta} \tag{5.8}$$

where $Z_i = e^{W_i u}$, we can see that frailty term acts multiplicatively on the hazard rate. The hazard rate is now conditional on both the covariates and the frailty. The assumption is that the distribution of Z_i needs to be specified with mean 1 and unknown variance equal to some parameter θ. This parameter is a measure of heterogeneity of the subjects. The event times are assumed to be independent conditional on the chosen parametric distribution, so inference may be made in standard fashion. The distribution of frailty term can in general be any positive distribution, such as Gaussian, t distribution or the most often used gamma distribution [43,49].

4.6.2. d Conditional models

When analyzing medical data, it is frequently the case of having repeated events but correlated. This violates the ordinary Cox model's assumption of all observations being independent. Therefore, the conditional models were proposed.

The main assumption of conditional model is that events are correlated, meaning that subject cannot be at risk for second event before it experience the first and so on; in general, a subject is not at risk for kth event before the $k-1th$ event occurs. To accomplish this each event is assigned to a separate stratum which allows the underlying intensity function to vary from event to event. Schematic form of conditional model for a subject with recurrent events is presented in *Figure 9*.

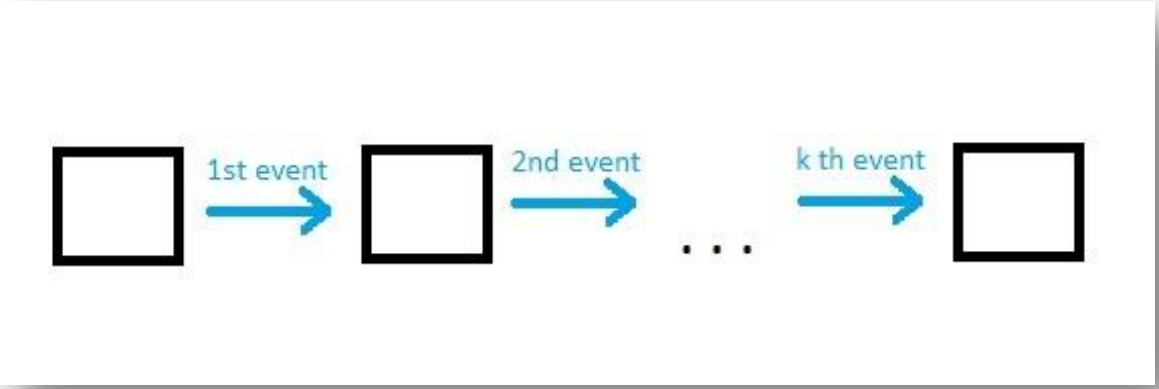

Figure 9: The conditional model for recurrent data – schematic
Arrow = event, Box = risk set

By accounting for events dependency we can improve variance-corrected models as well as random-effect models. Therefore, presented Andersen-Gill and frailty model can be stratified on event number and which leads to more appropriate models for fitting the recurrent data. Thus, these models are called Conditional Andersen-Gill and Conditional frailty model [46,49].

Chapter 5

Results

This chapter presents main findings of this thesis. We start with description of the study population and pneumonia incidence rates. Then, the explorative analysis of cohort members' characteristics and air pollution exposure is given. And finally, we present results of the modeling association between air pollution exposure at home over many years and risk for first and multiple hospital admissions for pneumonia in DCH cohort.

5.1 Study population and event incidence

Study population for this study consisted of 57053 DCH cohort members aged 50 – 65 years and who lived in Copenhagen or Aarhus between December 1993 and May 1997, who were followed in the Danish Hospital Discharge Register until the event (pneumonia hospitalizations), emigration or death registered in CPR (censoring date), or 31st of December 2009. Several exclusion criteria were applied prior to definition of final study population. Initially, 571 people have excluded due to history of cancer before baseline. We have also excluded 962 cohort members for missing residential address at recruitment, 948 for whom less than 80% of residential address history was available from 1971 until the end of follow-up, and 1333 with missing information on one or more covariates, giving the of 53239 people eligible for the study.

Of 53239 people we have found 3024 (5.7%) cases of admitted DCH cohort members for the first pneumonia between baseline and 31st of December 2009, with an average follow-up of 12.7 years. The overall incidence rate was 4.5 cases per 1000 person-years. Among 3024 individuals 626 (1.2%) had more than one pneumonia admission, that is repeated events. The repeated event (new hospital admission for pneumonia) was defined as at least 30 days after previous pneumonia hospitalization, which was suggested by medical experts as reasonable time window. In this cohort, there are up to 10 pneumonia cases per person during follow-up.

To be able to study whether association between air pollution and hospitalizations for pneumonia is modified by some other pre-existing health condition; we have performed the analysis on different subsets of DCH population (*Table 2*). First, we consider dividing study population into healthy sub-population (no disease prior to baseline) and those with prior disease, and defined the co-morbid conditions by Charlson index of diseases before and after

baseline. See *Section 2.5* and *Table 2* for precise definition of co-morbidity in this cohort by Charlson index. We have found that of 53239 total people on study, 46947 (88.2%) individuals had no co-morbid diseases before baseline and were considered healthy population, whereas 6292 (11.8%) had co-morbid conditions before baseline. Secondly, we considered the cases of pneumonia hospitalizations before the baseline, since we want to examine if air pollution effect differs for healthy individuals and ones with history of this specific disease. Of total of 53239 people, 46462(87.3%) individuals did not have history of hospital admissions for pneumonia, whereas 485 (0.9%) individuals were admitted to hospital for pneumonia before baseline.

5.2 Descriptive Data Analysis

Before performing the main analysis we performed descriptive-analysis in order to describe the cohort data, air pollution exposure, and event incidence in DCH. Therefore, an exploratory data analysis is conducted to determine confounders, their effect on air pollution, and the relationship between pollution levels and the risk of experiencing pneumonia.

5.2.1 Testing the potential confounders

We needed to distinguish between effect of exposure to air pollution as the main risk factor for pneumonia, and potential effect of some characteristics of subjects under observation, which might affect air pollution - pneumonia association. To be able to do that we first need to test potential confounders. The covariates which we tested are recognized risk factors for pneumonia [51], and are presented by their definition and categorization used in this analysis, in *Table 1*. Furthermore, *Table 3* shows distribution of these covariates in DCH cohort for total population and population subgroups considering history of pneumonia and Charlson index before baseline.

Risk factor	Categories	Total population $n = 53239$	No pneumonia no CI diseases $n = 46462$	Pneumonia no CI diseases $n = 485$
Age	< 56 years	23923 (48.69%)	23343 (50.24%)	204 (42.06%)
	≥ 56 years	27316 (51.31%)	23119 (49.76%)	278 (57.94%)
Gender	Female	27857 (52.32%)	24803 (53.38%)	258 (53.20%)
	Male	25382 (47.68%)	21659 (46.62%)	227 (46.80%)
Education	< 8 years	17546 (32.96%)	14700 (31.64%)	156 (32.16%)
	8-10 years	24586 (46.18%)	21741 (46.79%)	230 (47.42%)
	≥ 10 years	11107 (20.86%)	10021 (21.58%)	99 (20.42%)
BMI	Underweight	1906 (3.58%)	1612 (3.47%)	32 (6.60%)
	Normal	43625 (81.94%)	38519 (82.90%)	368 (75.88%)
	Obese	7708 (14.48%)	6331 (13.63%)	85 (17.52%)

Nutrition fruit	Mean intake	181.63 g/day	182.39 g/day	179.32 g/day
fat	Mean intake	85.29 g/day	85.21 g/day	87.17 g/day
Sports	Not ph. active	24387 (45.81%)	20560 (44.25%)	243 (50.10%)
	< 3.5 h/day	23598 (44.32%)	21225 (45.68%)	188 (38.76%)
	≥ 3.5 h/day	5254 (9.87%)	467 (10.07%)	54 (11.13%)
Smoking	Never	18876 (35.46%)	17210 (37.05%)	131 (27.01%)
	Previously	15265 (28.68%)	13087 (28.17%)	152 (31.34%)
	Current:			93 (19.18%)
	< 15 g/day	9813 (18.43%)	8396 (18.07%)	
	15-25 g/day	6497 (12.20%)	5429 (11.68%)	65 (13.40%)
	≥ 25 g/day	2782 (5.23%)	2334 (5.03%)	44 (9.07%)
ETS	Yes	34148 (63.14%)	29283 (63.03%)	331 (68.25%)
	No	19091 (35.86%)	17179 (36.97%)	154 (31.75%)
Alcohol	No alcohol	4435 (8.53%)	3719 (8.16%)	322 (68.80%)
	1-20 dr./week	36706 (70.55%)	32381 (71.04%)	39 (8.34%)
	≥ 20 dr./week	10885 (20.92%)	9480 (20.80%)	107 (22.86%)
Occupational exposure	Yes	14904 (27.99%)	12346 (26.77%)	134 (27.63%)
	No	38335 (72.01%)	34026 (73.23%)	351 (72.37%)
SES	Mean income (10000dkk/year)			

Table 3: Characteristics of Diet, Cancer and Health cohort for incidence of pneumonia and Charlson index diseases at follow-up

The analysis of potential confounding effects is conducted using univariate Cox proportional hazard model

$$\lambda_i(t) = \lambda_0(t)e^{X_i\beta}$$

for covariate X_i. The effects of all potential confounders by hazards rates, together with corresponding 95% - confidence intervals, and their significance level are given in *Table 4*. The univariate Cox analysis shows significant effect of most of tested covariates.

All covariates except socio-economic status (SES) at neighborhood level showed significant association with hospitalization for pneumonia. High education, fruit intake and physical activity had significant preventive effect, whereas BMI, fat intake, alcohol intake, occupational exposure, and especially smoking significantly increased the risk for pneumonia hospitalizations. According to these results, the final models for testing the effect of exposure to air pollution on hospital admission for pneumonia should be corrected for all confounders.

Socio-economic status (SES) was not associated with hospital admissions for pneumonia, which may be explained by the fact that SES was defined as an average income at municipality level.

However, pneumonia is known as age-related disease and this is elderly cohort, so the independence can't be assumed. Also age in only two levels may not be enough, so we will still correct the final models for age by using it as underlying time scale, which is a typical unbiased time scale for cohort studies [45,52].

Risk factor	Categories	HR (95% CI)	$p - value$
Age	< 56 years	1	
	≥ 56 years	0.94 (0.85 – 1.04)	0.214
Gender	Female	1	
	Male	1.23 (1.16 – 1.33)	2.66×10^{-9}
Education	< 8 years	1	
	8-10 years	0.84 (0.78 – 0.91)	1.94×10^{-5}
	≥ 10 years	0.80 (0.72 – 0.88)	1.24×10^{-5}
BMI	Normal (20-30 kg/m^2)	1	
	Underweight (<20 kg/m^2)	2.03 (1.75 – 2.35)	2×10^{-16}
	Obese (>30 kg/m^2)	1.18 (1.07 – 1.30)	7.89×10^{-4}
Nutrition fruit	Mean intake (100 g/day)	0.93 (0.90 – 0.95)	1.74×10^{-8}
fat	Mean intake (100 g/day)	1.34 (1.20 – 1.51)	4.48×10^{-7}
Sports	Not ph. active	1	
	< 3.5 h/day	0.68 (0.67 - 0.86)	2×10^{-16}
	≥ 3.5 h/day	0.76 (0.65 - 0.87)	2.83×10^{-5}
Smoking	Never	1	
	Previously	1.33 (1.20 - 1.48)	3.77×10^{-8}
	Current < 15 g/day	1.96 (1.76 - 2.17)	2×10^{-16}
	Current 15-25 g/day	2.68 (2.40 - 2.99)	2×10^{-16}
	Current ≥ 25 g/day	3.10 (2.69 - 3.56)	2×10^{-16}
ETS	No		
	Yes	1.665 (1.53 - 1.81)	2×10^{-16}
Alcohol	1-20 drinks/week	1	
	No alcohol use	1.27 (1.13 - 1.44)	1.1×10^{-4}
	≥ 20 drinks/week	1.36 (1.25 - 1.48)	2.9×10^{-12}
Occupational exposure	No	1	
	Yes	1.34 (1.25 – 1.45)	2.08×10^{-14}
SES	Mean income (10000dkk/year)	1.02 (0.91 – 1.13)	0.782

Table 4: Univariate Cox analysis for potential confounders

Since the Cox proportional hazard model assumes log-linear structure for covariates we also need to consider their functional form. For that reason continuous variables are interesting to be analyzed in more detail to make sure we have the right form used in the models. As discussed in *4.3.3*, good way to check the functional form is by fitting regression splines directly into the Cox model.

From all defined confounders we decided to adjust for, first we addressed the form of age, as an important continuous variable. As already mentioned, this is the elderly cohort, so the risk for pneumonia is not expected to be different for different age groups. Therefore, there is no need for changing the form of age and models are adjusted for age as a linear function.

Second interesting confounder is body mass index (BMI) of the study participants, as known pneumonia predictor [51]. It is continuous variable categorized in three levels, underweight, normal and obese. The assumption of linearity might be wrong here, since underweighted and obese people are proven to have higher chances to suffer from diseases. Therefore the functional form is presented by fitting restrictive cubic spline with 3 degrees of freedom (df) (*Figure 10*). It can be seen that the risk for pneumonia is not increased linearly with increasing BMI, but is low for normal BMI, whereas both low and high BMI are associated with increased risk for pneumonia, Due to this violation of linearity, in this study we modeled BMI with spline, allowing for U shape.

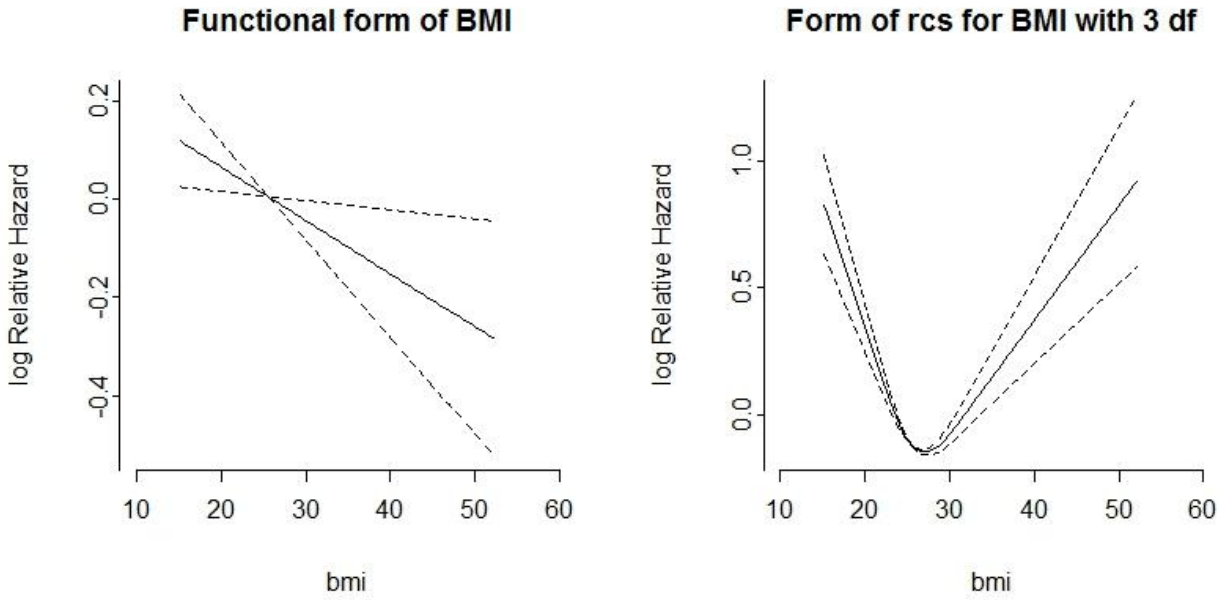

Figure 10: BMI functional form for DCH members – Original BMI values (left) and fitted restricted cubic spline with 3 df (right)

5.2.2 Air pollution exposure

In DCH cohort, traffic-related air pollution is given at residential addresses for participants in two ways. First, the modeled exposure (airGis dispersion model) to nitrogen dioxide (NO_2) and nitrogen oxides (NO_x) available as yearly average values since 1971 for all the addresses in Copenhagen are the main exposure of interest in this study. Secondly, a more naive proxies pf exposure to traffic-related air pollution were defined as indicators of presence of major roads within 50 and 100m radius around the residence at baseline address, as well as, the intensity of traffic around the participants' residential addresses at baseline.

Personal exposure to NO_2 and NO_x for DCH cohort members was defined for analyses as cumulative mean of annual mean values since 1971 until the event, censoring date or end of follow-up (31th of December 2009). These mean levels varied widely among cohort members (*Figure 11*).

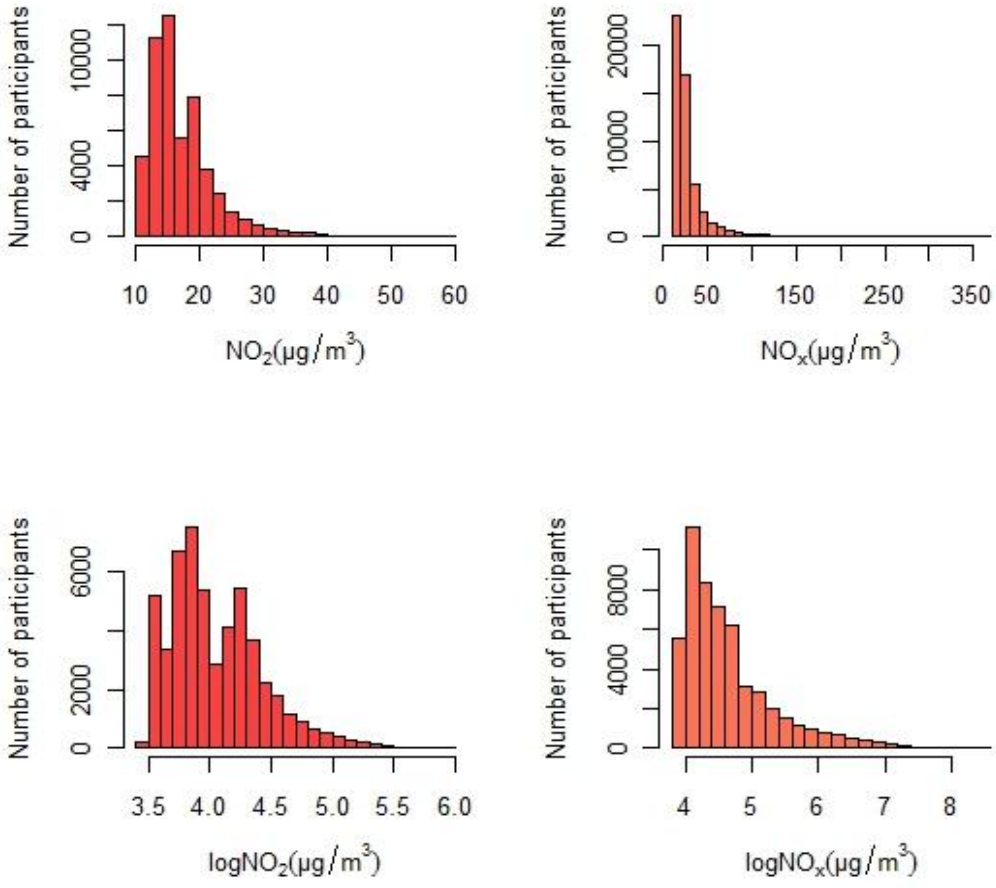

Figure 11: Frequencies of 39-year accumulated mean levels of original and log-transformed nitrogen dioxide (NO2) and nitrogen oxides (NOx) at residences of 53239 DCH cohort members

From frequency plots (*Figure 11* – upper panel) we can see very skewed distributions, i.e. there are not many observations with very high exposure values for both NO_2 and NO_x. Therefore, log-transformation of NO_2 and NO_x levels was performed and used in the analyses.

Since the main aim is to investigate what is the effect of exposure to air pollution on the risk for pneumonia in DCH cohort, clear and easy interpretation of resulting rates is of course very important. Log-transformation is then a good choice, because it does not change the results, it deals with distribution skewness and is very easy to interpret. We used logarithm with base 2 transformation, which gives estimated effect rates when doubling the exposure values (*Figure 11* – bottom panel).

Moreover, the functional form of the relationship between NO_2 and NO_x and pneumonia is expected to be linear (from existing literature), and was estimated and presented in *Figure 12*, with the original mean values of the left panel and log-transformed version on the right. Transformation clearly improves the functional form, especially for NO_x values. NO_2 becomes very closer to linear, and NO_x also until certain value where it becomes constraint. The behavior on the right end of NO_x is explained by very few observations with quite high exposure values.

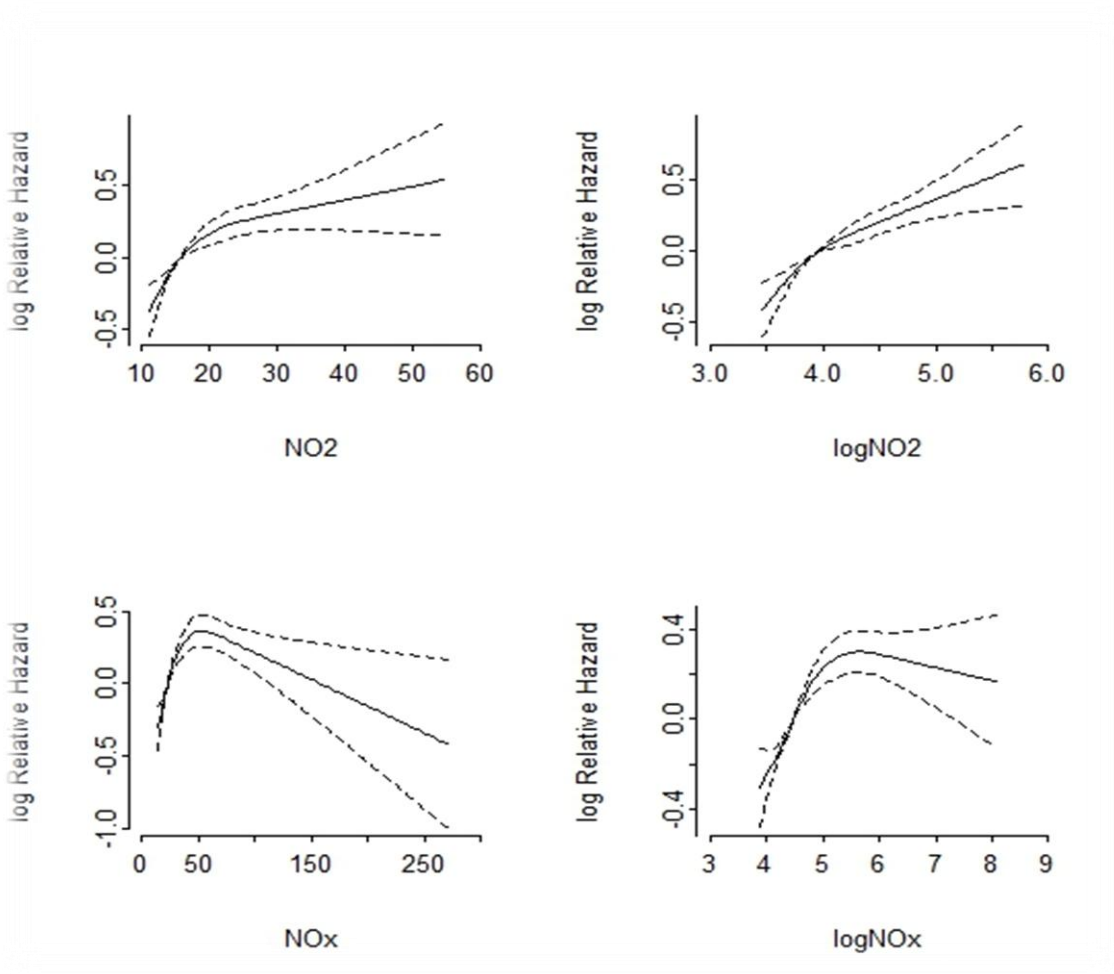

Figure 12: Functional form of mean original and log-transformed NO_2 and NO_x exposure values in DCH cohort

The other proxy of exposure to air pollution data is based on traffic intensity data around the residence at baseline (time of recruitment into the cohort). We have defined indicator variables for the presence of major roads (those with traffic density of 10.000 and 5.000 vehicles per day) within 100m and 50m radius around the residential addresses, as well as the traffic load vraibles, which is the total number of kilometers driven by vehicles within 200m and 100m radius around residential addresses of cohort members at requirement.

Distribution of traffic-related air pollution exposure for DCH cohort members by NO_2 and NO_x levels and traffic proxies is presented in *Table 5*.

Lengths of exposure	Air pollution	Total population $n = 53239$		No pneumonia no CI diseases $n = 46462$		Pneumonia no CI diseases $n = 485$	
Modeled NO_2 and NO_x exposure		***Mean (SD)***	***Median***	***Mean (SD)***	***Median***	***Mean (SD)***	***Median***
39 years (1971 -)	NO_2 ($\mu g/m^3$)	17.52 (5.42)	15.80	17.47 (5.40)	15.76	17.85 (5.72)	15.83
	NO_x ($\mu g/m^3$)	29.65 (22.23)	22.51	29.50 (22.12)	22.51	30.87 (22.68)	23.03
29 years (1981 -)	NO_2 ($\mu g/m^3$)	18.80 (6.35)	16.32	18.74 (6.32)	16.25	19.15 (6.71)	16.50
	NO_x ($\mu g/m^3$)	31.70 (26.59)	23.04	31.51 (26.44)	22.88	32.99 (28.49)	24.10
19 years (1991 -)	NO_2 ($\mu g/m^3$)	18.74 (6.90)	15.79	18.69 (6.88)	15.79	19.04 (7.29)	15.79
	NO_x ($\mu g/m^3$)	31.52 (29.42)	21.71	31.35 (29.32)	21.46	32.86 (31.68)	20.60
1 year mean at baseline	NO_2 ($\mu g/m^3$)	17.19 (5.19)	15.42	17.15 (5.17)	15.38	17.51 (5.46)	15.68
	NO_x ($\mu g/m^3$)	28.76 (21.44)	22.17	28.61 (21.32)	22.06	29.91 (22.77)	22.51
1 year mean at end-date	NO_2 ($\mu g/m^3$)	16.92 (5.21)	15.18	16.88 (5.19)	15.12	17.28 (5.55)	15.33
	NO_x ($\mu g/m^3$)	27.68 (20.17)	21.45	27.52 (20.04)	21.31	28.94 (21.87)	21.82
1 year mean at baseline		***Mean (SD)***	***Median***	***Mean (SD)***	***Median***	***Mean (SD)***	***Median***
Traffic load (10^5 vehicles km/day) within 100m radius		10.88 (17.31)	3.48	10.72 (17.24)	3.37	11.15 (17.33)	3.89
Traffic load (10^5 vehicles km/day) within 200m radius		46.30 (54.00)	25.58	45.75 (53.78)	24.95	46.24 (55.40)	22.66
1 year mean at baseline		***No. of cases (%)***		***No. of cases (%)***		***No. of cases (%)***	
Major road (5000 v/day) within 50m radius		8754 (16.44)		7536 (16.22)		86 (17.73)	
Major road (10000 v/day) within 50m radius		4317 (8.11)		3699 (7.96)		47 (9.69)	
Major road (5000 v/day) within 100m radius		16849 (31.65)		14471 (31.15)		156 (32.78)	
Major road (10000 v/day) within 100m radius		9023 (16.95)		7671 (16.51)		92 (18.97)	

Table 5: Description of air pollution exposure in Diet, Cancer and Health cohort for incidence of pneumonia and Charlson index diseases at follow-up

5.2.3 Cumulative hazard rates and Survival curves

In exploring the data, it is recommended to start with some simple univariate analysis and have a look at the shape of survival curves and hazard rates of covariates (predictors). Kaplan-Meier survival estimates give the insight into the shape of survival functions and Nelson-Aalen estimates of cumulative hazards for each predictor.

Figure 13 presents Kaplan-Meier survival and Nelson-Aalen cumulative hazard, based on the null Cox model of DCH cohort data. The survival function shows the expected survival over the time, i.e. time in the study (age as time scale) without experiencing pneumonia, and cumulative hazard gives us information about event intensity.

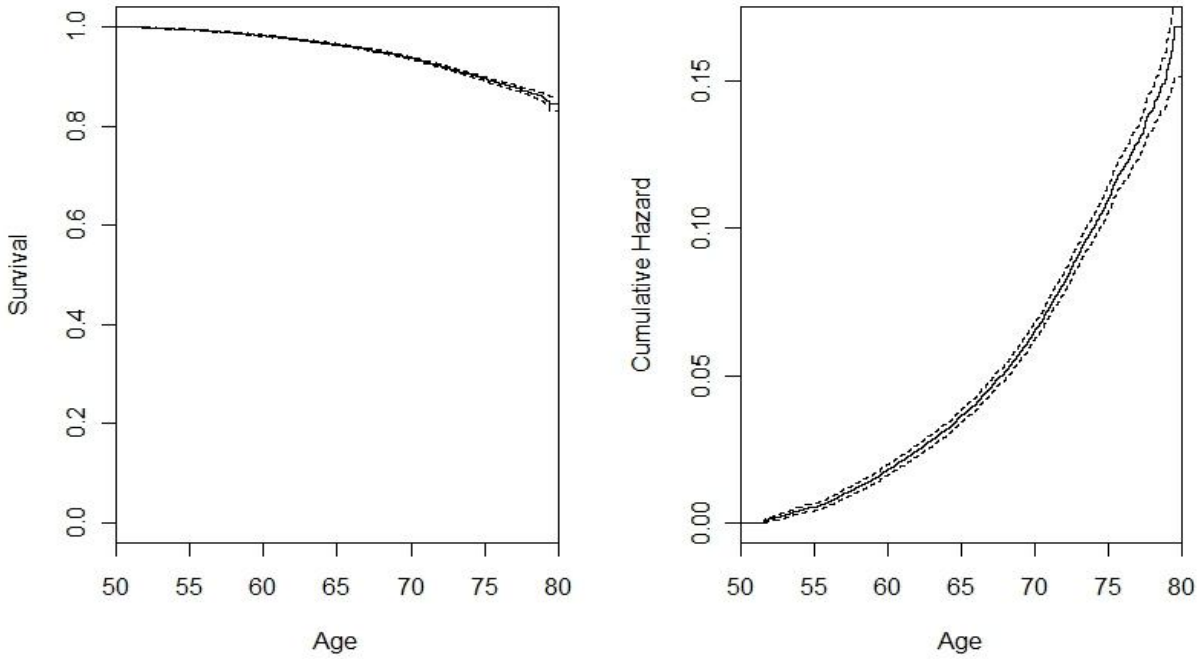

Figure 13: Survival curve (left) and Cumulative hazard (right) based on null Cox Model

Note that survival curves and hazard rates are giving information from the opposite perspectives, which have predefined relationship. Thus, knowing one we can easily calculate the other because cumulative hazard is simply negative logarithm of the survival ($\Lambda(t) = -\log S(t)$). Therefore, we will for simplicity analyze only cumulative hazards further on.

It is important to explore the difference in cumulative hazards among factors of air pollution exposures which is our main predictor of interest, and also for all confounding variables relevant for this study.

The exposure assessment of log-transformed NO_2 and NO_x in DCH cohort is grouped in quartiles and corresponding Nelson-Aalen estimates of cumulative hazards are presented in *Figure 14*. We can see that for first 5 to 10 years (from age of 50 until age of 55 – 60) all exposure groups are almost identical and intensity of pneumonia hospitalizations among DCH cohort members is low. From the age of 60, the intensity functions are increasing and higher air pollution exposure values have higher event rates. We have observed more pneumonia hospitalizations among people with higher air pollution exposure, and this result is consistent with our hypothesis.

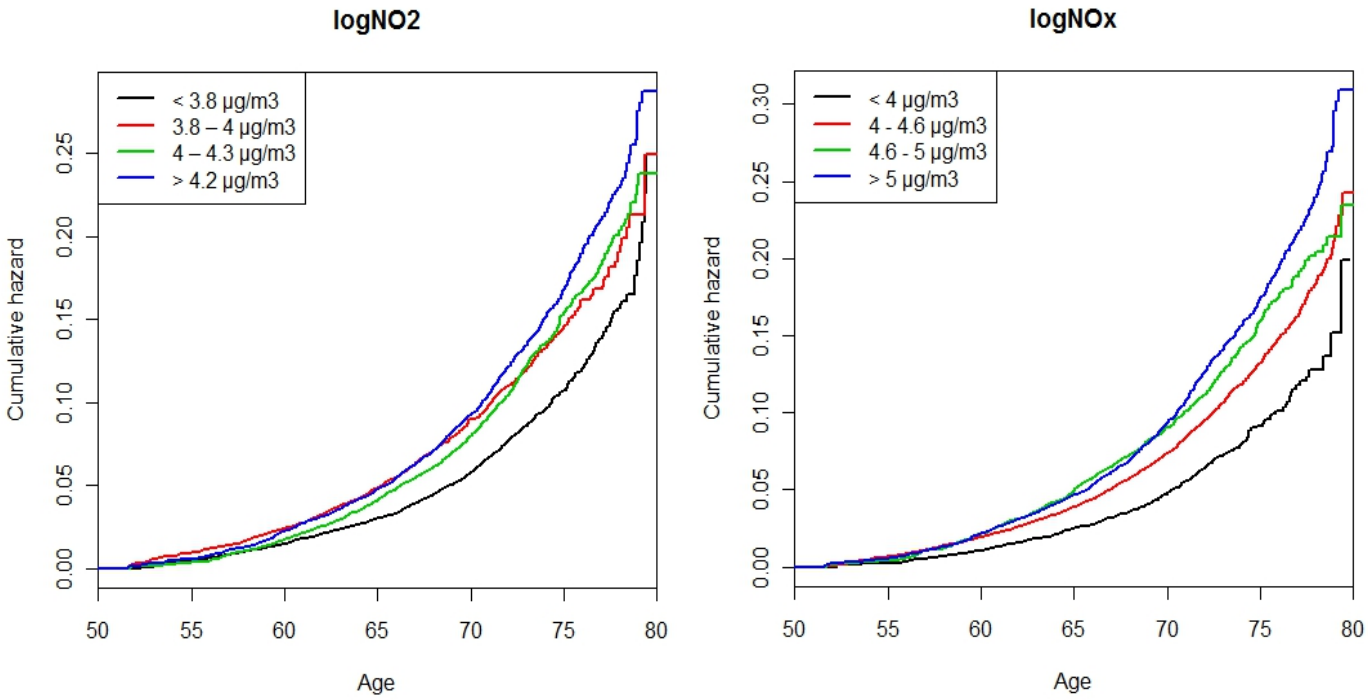

Figure 14: Cumulative hazard functions of exposure grouped in quartiles of DCH cohort

The confounders with most significant effect on risk for pneumonia are discovered to be BMI, smoking and alcohol, and therefore cumulative hazard functions and their 95% confidence intervals of each of these are presented here, *Figure 15*, *Figure 16* and *Figure 17*. The plots for the rest of covariates with confounding effect can be found in the *Appendix C*.

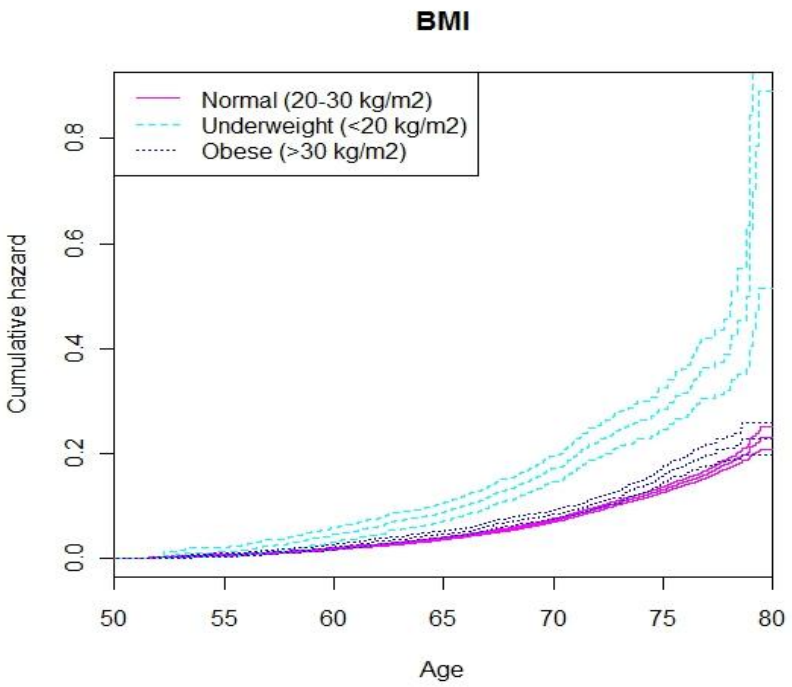

Figure 15: Cumulative hazard functions for BMI of DCH cohort members categorized in three groups

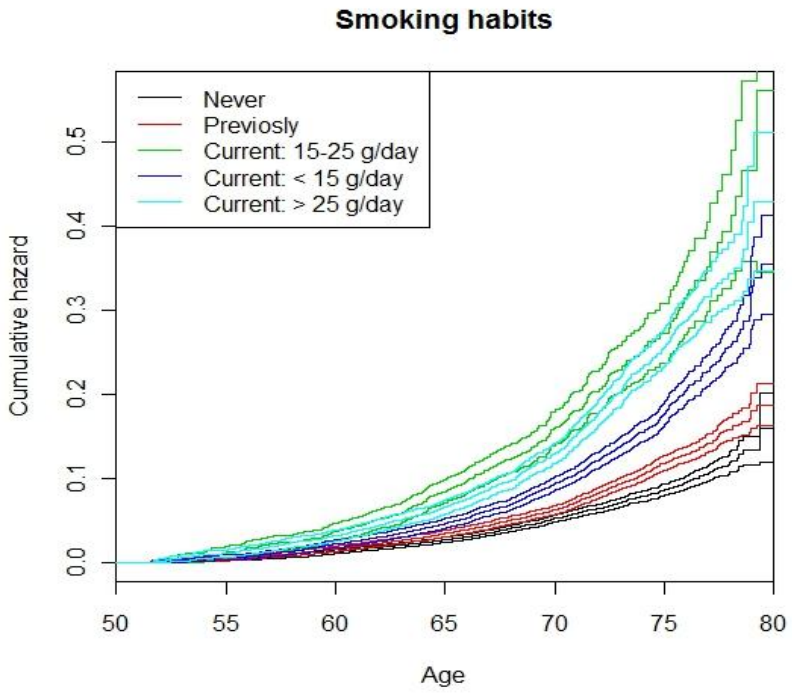

Figure 16: Cumulative hazard functions for smoking habits of DCH cohort members categorized in five groups

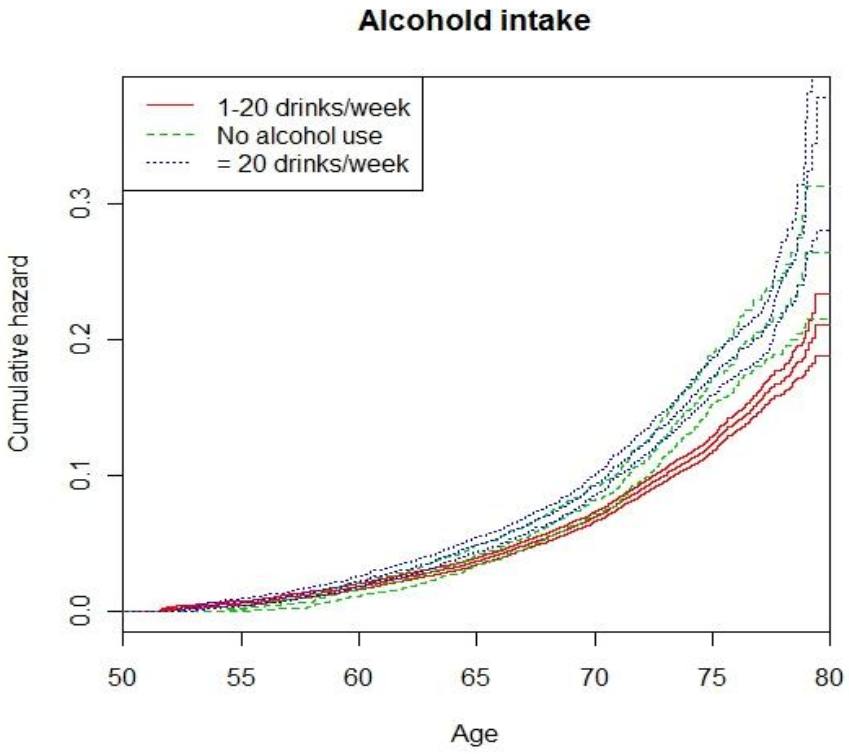

Figure 17: Cumulative hazard functions for alcohol intake of DCH cohort members categorized in three groups

Estimated cumulative hazards for BMI of DCH cohort members show noticeable difference for underweight group comparing to the other two, except at the beginning of the interval, where all are very similar. This implies that underweighted people are more likely to have higher intensity rate for pneumonia hospitalizations than the others.

Next, we observed that smoking status of study participants is very important pneumonia predictor. Cumulative hazard rates reveal that smoking has very bad influence on pneumonia hospitalizations, meaning that pneumonia occurrence is more frequent for intensive smokers. Smoking more than 15 grams of tobacco per day has the highest pneumonia rates. This result is consistent with the fact, that smoking can make a serious damage to the lungs, and lead to respiratory diseases.

The alcohol consumption is defined as number of drink per week for an individual. From estimated cumulative hazards we can see that middle group of 1-20 drinks per week has noticeable lower event intensity compared to other two. Non-drinkers and people with high alcohol consumption, more than 20 drinks per week, have very similar rates and are more likely to experience pneumonia. At the first look, it might seem strange to have group of people that

do not consume alcohol with so high event intensity. However this may be due to the fact that some of them can be already sick, even “too sick to drink”, since pneumonia can occur as co-morbidity in relation to many other chronic diseases, or had former alcohol abuse [51].

5.3 Time to first event analysis using ordinary Cox model

The design of this analysis of DCH cohort data includes follow-up since time of recruitments until the first pneumonia hospitalization, or censoring caused by loss to follow-up by death or emigration, or the end of follow-up (31st of December 2009). The Cox analysis has been performed in order to investigate if there is an association between exposure to air pollution caused by traffic, and hospital admissions for pneumonia in Copenhagen area.

Considering the confounders of cohort characteristics, there are three versions of fitted Cox model. First simple model was only adjusted for age. Second model was adjusted for age and for confounders which are known risk factors for pneumonia and are related to air pollution in terms of inhalation of harmful particles lungs, and include smoking status, duration and intensity, environment tobacco smoke (ETS) and occupational exposure. And the third model was adjusted for all relevant confounders.

The risk for experiencing pneumonia associated to traffic-related air pollution exposure of DCH cohort members is investigated in two ways. First, the risk associated with the modeled exposure to NO_2 and NO_x available as yearly mean since 1971, and defined as cumulated mean in several exposure lengths. And secondly, more naïve proxies of exposure to air pollution were defined from data on traffic density around residence at baseline address (time to recruitment), which are describing the presence of major roads and traffic intensities around residential addresses of study participants.

5.3.1 Association between NO_2 and NO_x exposure and first pneumonia occurrence in DCH cohort

Looking at traffic proxy of NO_2 and NO_x exposure first, we have modeled the exposure length of total cohort population (53239) in several ways. Using full available data of measured air pollution levels in Copenhagen since 1971 up to the end of follow-up, we have calculated cumulative mean exposure of up to 39 years. Secondly model includes exposure since 1981, which is at most 29 years when the study was ended, where we leave 10 years of so-called “burn-in” period. And third, up to 19 years long exposure time, from 1991 which is just couple of year before the recruitment. There are also two simple yearly exposure models, pollution level for the year at recruitment and for the year of hospitalization for pneumonia, censoring or end of follow-up.

Table 6 shows estimated hazard ratios (HR) and 95% confidence intervals for log-transformed NO_2 and NO_x exposure of 53239 DCH cohort participants, with 3024 cases of pneumonia hospitalization observed during follow-up. We discovered significant positive association between air pollution levels and admissions for pneumonia. All exposure models are consistent with this result and no significant difference in estimates using different exposure lengths is found. Therefore, we decided to use only one exposure proxy for the further analysis and that will be model with entire available exposure data, since 1971.

Cox Model / *Exposure for Total population*	*log*	Adjusted for age **HR (95%CI)**	Adjusted for age, smoking, ets and occupational exposure **HR (95%CI)**	Fully adjusted* **HR (95%CI)**
Cumulative mean exposure since **1971** (to event, censoring or 31. December 2009)	**NO_2**	1.42 (1.30 – 1.54)	1.29 (1.18 – 1.41)	1.25 (1.14 – 1.36)
	NO_x	1.19 (1.14 – 1.25)	1.13 (1.08 – 1.18)	1.11 (1.06 – 1.16)
Cumulative mean exposure since **1981** (to event, censoring or 31. December 2009)	**NO_2**	1.39 (1.28 – 1.50)	1.28 (1.18 – 1.38)	1.24 (1.14 – 1.34)
	NO_x	1.17 (1.12 – 1.22)	1.11 (1.07 – 1.17)	1.10 (1.05 – 1.15)
Cumulative mean exposure since **1991** (to event, censoring or 31. December 2009)	**NO_2**	1.35 (1.25 – 1.45)	1.26 (1.16 – 1.35)	1.22 (1.13 – 1.32)
	NO_x	1.16 (1.11 – 1.21)	1.15 (1.05 – 1.26)	1.09 (1.05 – 1.14)
Mean exposure at cohort baseline (1993–1997)	**NO_2**	1.46 (1.35 - 1.58)	1.36 (1.26 - 1.48)	1.33 (1.23 - 1.44)
	NO_x	1.20 (1.15 – 1.25)	1.15 (1.10 – 1.20)	1.14 (1.09 – 1.18)
Mean exposure at the end of follow-up (event date, censoring date or 31.Dec2009)	**NO_2**	1.20 (1.11 - 1.29)	1.13 (1.04 - 1.22)	1.11 (1.02 - 1.20)
	NO_x	1.16 (1.11 – 1.21)	1.12 (1.07 – 1.17)	1.11 (1.06 – 1.16)

Table 6: Association between NO_2 and NO_x exposure of different length and pneumonia incidence (n=3027) among 53239 DCH cohort participant

The Spearman rank correlation coefficient between 39-year mean NO_2 and NO_x reveals that they are highly correlated (0.96). Thus, it is enough to consider one of these two and we will take NO_2 for the further analysis.

When considering different predefined subpopulation of total population, according to co-morbidities prior to cohort baseline, we have found the same result of significant strong positive association between NO_2 exposure since 1971 and hospital admissions for pneumonia. It is presented in *Table 7* bellow, for categorical and linear exposure. Categorization is made in four groups as quartiles of log-transformed mean value.

Model / *Population*	$\log_2 NO_2$	Adjusted for age HR (95%CI)	Adjusted for age, smoking, ets and occupational exposure HR (95%CI)	Fully adjusted* HR (95%CI)
Total population $n = 53239$ ***(3024 pneum.cases)***	Linear trend	1.42 (1.30 – 1.54)	1.29 (1.18 – 1.41)	1.25 (1.14 – 1.36)
	< 3.8 μg/m^3	1	1	1
	3.8 – 4 μg/m^3	1.34 (1.21 – 1.49)	1.30 (1.17 – 1.45)	1.28 (1.16 – 1.43)
	4 – 4.3 μg/m^3	1.36 (1.23 – 1.50)	1.30 (1.18 – 1.44)	1.28 (1.16 – 1.42)
	≥ 4.3 μg/m^3	1.55 (1.41 – 1.71)	1.40 (1.27 – 1.55)	1.35 (1.23 – 1.49)
No Charlson index hospitalizations before baseline $n = 46947$ ***(2305 pneum. Cases)***	Linear trend	1.40 (1.27 – 1.55)	1.28 (1.16 – 1.41)	1.24 (1.13 – 1.38)
	< 3.8 μg/m^3	1	1	1
	3.8 – 4 μg/m^3	1.32 (1.17 – 1.49)	1.29 (1.14 – 1.45)	1.28 (1.13 – 1.44)
	4 – 4.3 μg/m^3	1.37 (1.23 – 1.54)	1.31 (1.17 – 1.47)	1.30 (1.16 – 1.45)
	≥ 4.3 μg/m^3	1.52 (1.36 – 1.70)	1.38 (1.23 – 1.54)	1.34 (1.20 – 1.50)
History of co-morbid conditions defined by Charlson index $n = 6292$ ***(719 pneum.cases)***	Linear trend	1.37 (1.15 – 1.63)	1.28 (1.07 – 1.53)	1.22 (1.02 – 1.46)
	< 3.8 μg/m^3	1	1	1
	3.8 – 4 μg/m^3	1.28 (1.03 – 1.58)	1.24 (1.00 – 1.54)	1.23 (0.99 – 1.53)
	4 – 4.3 μg/m^3	1.26 (1.02 – 1.56)	1.24 (1.01 – 1.53)	1.22 (0.99 – 1.51)
	≥ 4.3 μg/m^3	1.55 (1.27 – 1.89)	1.42 (1.16 – 1.73)	1.35 (1.11 – 1.66)
No Charlson index and no pneumonia before baseline $n = 46462$ ***(2231 pneum.cases)***	Linear trend	1.39 (1.26 – 1.53)	1.26 (1.14 – 1.40)	1.23 (1.11 – 1.36)
	< 3.8 μg/m^3	1	1	1
	3.8 – 4 μg/m^3	1.33 (1.18 – 1.50)	1.30 (1.15 – 1.47)	1.29 (1.14 – 1.45)
	4 – 4.3 μg/m^3	1.37 (1.22 – 1.54)	1.31 (1.17 – 1.47)	1.30 (1.15 – 1.45)
	≥ 4.3 μg/m^3	1.51 (1.35 – 1.69)	1.36 (1.22 – 1.53)	1.33 (1.18 – 1.49)
History of pneum. hosp. without Charlson index before baseline $n = 485$ ***(74 pneum.cases)***	Linear trend	1.74 (1.04 – 2.90)	1.71 (1.03 – 2.85)	1.68 (1.01 – 2.81)
	< 3.8 μg/m^3	1	1	1
	3.8 – 4 μg/m^3	0.94 (0.44 – 1.97)ᵗ	1.04 (0.49 – 2.23)ᵗ	1.07 (0.49 – 2.32)ᵗ
	4 – 4.3 μg/m^3	1.33 (0.71 – 2.50)ᵗ	1.45 (0.76 – 2.74)ᵗ	1.49 (0.77 – 2.88)ᵗ
	≥ 4.3 μg/m^3	1.91 (1.04 – 3.48)	1.90 (1.03 – 3.49)	1.92 (1.03 – 3.58)

**= adjusted for age, gender, bmi, smoking, ets, alcohol, education, occupational exposure, sport, fruit, fat*

ᵗ = statistically non-significant $(p - value > 0.05)$

Table 7: Association between NO_2 exposure and pneumonia incidence on different DCH population groups

5.3.1. a Discussion

Overall for the total DCH population of 53239 individuals, the risk for hospital admission for pneumonia is 25% higher when doubling the exposure to NO_2 in a fully adjusted model (*Table 7*). When looking at the quartiles of log-transformed NO_2 exposure, the group exposed to more than 4.3 $\mu g/m^3$ (corresponding to around 20 $\mu g/m^3$ of original exposure) has the highest hazard rate of 1.35 (1.23 – 1.49), which means that individuals from the group with highest NO_2 exposure, compared to the low exposure group of less than 3.8 $\mu g/m^3$ (app.14 $\mu g/m^3$ originally), have approximately 35% higher risk to be hospitalized for pneumonia.

Considering co-morbidities defined by Charlson index specification of diseases, there is no significant difference in association between exposure and pneumonia incidence among people with or without co-morbid diseases. Hazard rates are consistent with the ones for the total population, so for the simplicity we will conduct further analysis only for the total DCH population.

The strongest association can be seen among the group of DCH cohort members which have been hospitalized for pneumonia before baseline. However, the sample is quite small; of 485 people only 74 had readmissions for pneumonia after baseline. Therefore, only marginally significant effect is obtained for the group with highest NO_2 exposure (more than 4.3 $\mu g/m^3$ of log-transformed exposure) and the other groups have no significant association. However, even though the effect is marginally significant, association is very strong. The hazard rate for doubling the exposure is 1.68 (1.01 – 2.81) and for the group exposed to more than 4.3 $\mu g/m^3$ even 1.92 (1.03 – 3.58). In other words, the individuals with history of pneumonia before baseline exposure to high level of air pollution are up to 92% more likely to experience a new event due to exposure to high levels of NO_2.

5.3.2 Association between traffic proxies exposure and pneumonia incidence in DCH cohort

In analyzing the effect of exposure to NO_2 on pneumonia hospitalizations we also considered traffic proxy variables. Those are variables describing the exposure to air pollution coming from the presents of major roads and high traffic intensity around residential addresses at recruitment.

Using Cox proportional hazard model we have also discovered significant positive association between traffic-related air pollution exposure and risk for pneumonia hospitalizations. In the *Table 8* , we can see hazard rates and 95% confidence intervals for these 6 variables. All the traffic proxy variables are highly significant and imply that, for example, having heavy traffic roads within 50 or 100m meter radius from the residential address increases the risk for pneumonia from 10 to 15%.

Cox Model / Traffic-related air pollution exposure	Adjusted for age **HR (95%CI)**	Adjusted for age, smoking, ets and occupational exposure **HR (95%CI)**	Fully adjusted* **HR (95%CI)**
Major road (5000 v/day) within 50m radius	1.19 (1.08 – 1.31)	1.11 (1.02 – 1.22)	1.09 (0.99 – 1.19)
Major road (10000 v/day) within 50m radius	1.27 (1.12 – 1.43)	1.18 (1.04 – 1.33)	1.15 (1.01 – 1.29)
Major road (5000 v/day) within 100m radius	1.22 (1.14 – 1.32)	1.15 (1.07 – 1.24)	1.12 (1.03 – 1.20)
Major road (10000 v/day) within 100m radius	1.26 (1.15 – 1.38)	1.18 (1.07 – 1.29)	1.14 (1.04 – 1.24)
High traffic load (10^6 vehicles km/day) within 100m radius	1.057 (1.038 - 1.075)	1.041 (1.022 - 1.060)	1.035 (1.016 - 1.054)
High traffic load (10^6 vehicles km/day) within 200m radius	1.023 (1.017 - 1.029)	1.018 (1.012 - 1.024)	1.016 (1.010 - 1.022)

**= adjusted for age, gender, bmi, smoking, ets, alcohol, education, occupational exposure, sport, fruit, fat*

Table 8: Association between presence of major roads (5000 or 10000 v/day) and high traffic loads around residential addresses and pneumonia incidence for total DCH cohort population ($n = 53239$)

5.4 Recurrent events analysis using extended Cox model

After finding quite strong positive association between air pollution exposure and first pneumonia hospitalization, we went a step further to utilize available data on all hospital admission for DCH cohort members, and examined if having multiple pneumonia admissions can also be related to air pollution. In addition to DCH cohort data used in previous (time to first event) analysis, we have added all the pneumonia cases after the first amission until the end of follow-up for all cohort members. In order to distinguish between hospital admissions for two different cases of pneumonia on a single subject, considered time gap between two distinct events is 30 days.

In DCH cohort there are up to 10 pneumonia hospitalizations per subject, but with very low frequency (*Figure 18*). In order to get significant result we will reduce the number of repeated event to maximum 3 per subject.

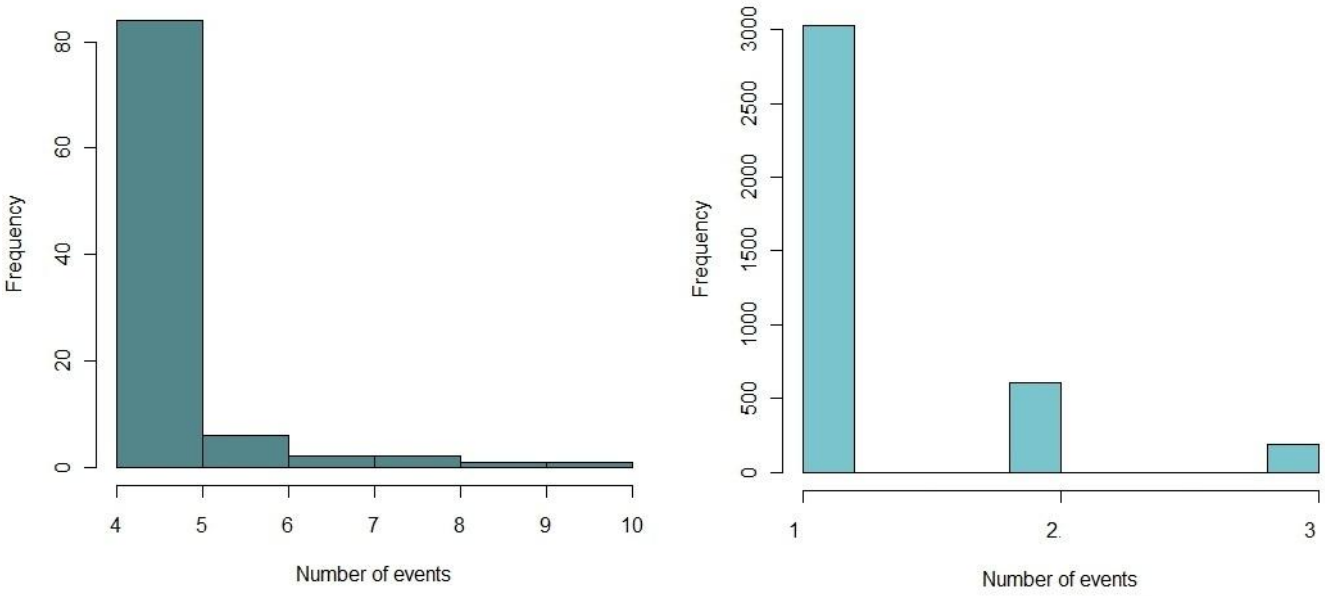

Figure 18: Distribution of number of events per subject in DCH cohort
4 or more events (left) ; From simple to 3 events (right)

In the analysis of survival data with multiple events difference occurs for individuals with recurrent events because events might be correlated and survival time is not the same. For example, subjects with 3 events have time defined as time at entry to the study (baseline: 1993 – 1997) until first pneumonia hospitalization, from first until second and from second until third. This gives us multiple lines (rows) in the data ser for all the individuals with more than

one event occurrence. The end of time intervals could be censoring date if censoring happened earlier than subsequent admission.

Before proceeding to with recurrent events models, it is interesting to see how survival curves and cumulative hazard rates look, for different number of events. The Kaplan – Meier estimate of survival and Nelson – Aalen estimate of cumulative hazard with 95% confidence intervals, for events number in DCH cohort, are presented in *Figure 19*.

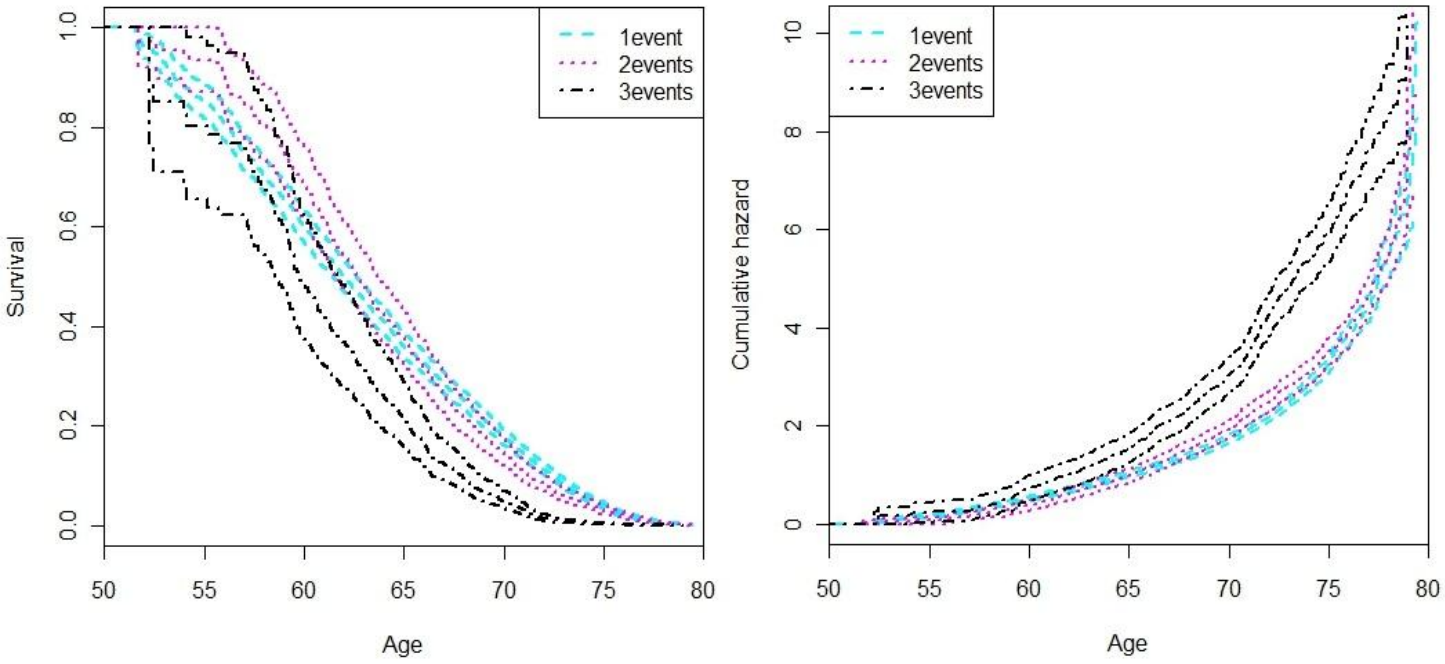

Figure 19: Cumulative hazard functions for events number of DCH cohort

From the plots it can be seen that having three pneumonia hospitalizations is different than having only one or two. The survival rates at the beginning of the time scale are a bit lower for three events group but further, after the age of 55 or 60, the difference is larger. The cumulative hazards show that the group with three pneumonia hospitalizations per subject has higher event frequency than the other two. Having one or two events seems to have very similar properties except from around age of 70 to age of 75 where two events cases cumulative hazard and it's 95% confidence interval are above the ones with only single event. This results leads to the conclusion that having pneumonia multiple times, more than twice, implies higher risk for new pneumonia occurrences. Therefore, we would expect that individuals with more than two pneumonia hospitalizations are more likely to experience it again.

5.4.1 Association between NO_2 and NO_x exposure and recurrent pneumonia occurrence in DCH cohort

There are couple of different models for recurrent data, as discussed in Section 4.6. Those are various extensions of ordinary Cox model and some are implemented in our study. We start with simple intensity-based model which is just ordinary Cox model of multiple events data. Next is variance-corrected Andersen Gill model, random effect Frailty model and, accounting for events dependency, Conditional models.

Fitted extended Cox models are still given in three versions considering adjustments for confounders. Those are age adjusted, then model adjusted for age, environment tobacco smoke and occupational exposure, and last, fully adjusted model.

Following tables, *Table 9* and *Table 10*, present hazard ratios (HRs) and 95% confidence intervals of different Cox model extensions for recurrent DCH cohort data. Among total DCH population of 53239 individuals, there are 3823 cases of pneumonia hospitalizations. The analysis reveals significant strong positive association between exposure to NO_2 and NO_x and risk for repeated pneumonia hospitalizations. Furthermore, among 485 DCH cohort members with history of pneumonia before baseline (before 1993 – 1997), there are 108 cases of pneumonia hospitalizations, which only marginally significant positive association. The significance is not strong due to a small sample size, i.e. after the baseline we have 74 out of 485 individuals that have experienced pneumonia, and only 20 with more than one event after the baseline.

For modeling recurrent data we first used the simple Intensity - based model and the Andersen-Gill model, which corrects for robust variance. Those two models are very similar and both have big disadvantage of make very strong assumptions. Log-hazard rates of those two models are almost identical, as expected, and one of the assumptions is that all events are independent which is very unlikely to be the case in our data. For that reason we have fitted the conditional Andersen-Gill model, which corrects for the robust variance, but also includes events as dependent, which means second event can't occur before the first nor third can before the second event. The hazard rates are still almost the same as in the previous models, so there still seems to be no improvement.

Since the individuals under the study are dissimilar, meaning that we can't expect them all to have same chances for being hospitalized for pneumonia, first or repeated, we need to account for subject's heterogeneity. The random individual effect is included by fitting the Frailty model. Frailty model's estimates are very significant and obviously higher than previous. Moreover, the model adjusted for age, smoking, ETS, and occupational exposure and fully adjusted model also returns very significant frailty estimate, allowing us to conclude that frailty term is needed in the model of recurrent DCH cohort data.

Again we should also include events correlation structure by stratifying on the event number. This is done by fitting the Conditional Frailty model.

Extended Cox Model / *Recurrent models for total population (3823 pneum. cases)*	*log (linear trend)*	Adjusted for age **HR (95%CI)**	Adjusted for age, smoking, ETS and occupational exposure **HR (95%CI)**	Fully adjusted* **HR (95%CI)**
Intensity-based Model	**NO_2**	1.43 (1.32 – 1.54)	1.30 (1.20 – 1.40)	1.25 (1.16 – 1.35)
	NO_x	1.21 (1.16 – 1.26)	1.14 (1.10 – 1.19)	1.12 (1.08 – 1.17)
Andersen – Gill model	**NO_2**	1.43 (1.31 – 1.56)	1.30 (1.19 – 1.42)	1.25 (1.15 – 1.37)
	NO_x	1.21 (1.15 – 1.26)	1.14 (1.09 – 1.20)	1.12 (1.07 – 1.14)
Conditional Andersen – Gill model	**NO_2**	1.42 (1.32 - 1.53)	1.33 (1.23 - 1.44)	1.30 (1.20 - 1.44)
	NO_x	1.20 (1.15 – 1.25)	1.16 (1.11 – 1.21)	1.14 (1.09 – 1.19)
Frailty model	**NO_2**	1.85 (1.51 – 2.27)	1.54 (1.27 – 1.86)	1.43 (1.19 – 1.73)
	NO_x	1.44 (1.28 – 1.62)	1.29 (1.16 – 1.44)	1.24 (1.11 – 1.37)
Conditional Frailty model	**NO_2**	1.45 (1.34 - 1.58)	1.33 (1.23 – 1.45)	1.30 (1.19 – 1.41)
	NO_x	1.22 (1.17 – 1.28)	1.16 (1.11 – 1.22)	1.14 (1.09 – 1.20)

*= *adjusted for age,gender, BMI, smoking, ETS, alcohol, education, occupational exposure, sport, fruit, fat*

Table 9: Association between NO_2 and NO_x exposure and pneumonia incidence ($n = 3823$) among 53239 DCH cohort members using different extended Cox model for recurrent

Extended Cox Model / *Recurrent models for population with pneumonia history (108 pneum. cases)*	*log*	Adjusted for age **HR (95%CI)**	Adjusted for age, smoking, ETS and occupational exposure **HR (95%CI)**	Fully adjusted* **HR (95%CI)**
Intensity-based Model	**NO_2**	1.78 (1.17 – 2.71)	1.69 (1.11 – 2.58)	1.75 (1.14 – 2.69)
	NO_x	1.30 (1.04 – 1.63)	1.26 (1.01 – 1.58)	1.29 (1.02 – 1.63)
Andersen – Gil model	**NO_2**	1.78 (1.06 – 2.98)	1.69 (1.02 – 2.81)	1.75 (1.07 – 2.87)
	NO_x	1.30 (1.00 – 1.70)	1.26 (0.97 – 1.64)	1.29 (1.00 – 1.67)
Conditional Andersen – Gill model	**NO_2**	1.35 (0.88 – 2.10)†	1.24 (0.78 - 1.98)†	1.25 (0.80 - 1.95)†
	NO_x	1.17 (0.93 – 1.48)†	1.11 (0.86 – 1.43)†	1.11 (0.88 – 1.42)†

† = *statistically non-significant* ($p - value > 0.05$)

*= *adjusted for age, gender, BMI, smoking, ETS, alcohol, education, occupational exposure, sport, fruit, fat*

Table 10: Association between NO2 and NOx exposure and pneumonia incidence (n=108) among 485 DCH cohort members with history of pneumonia before baseline using different extended Cox models for recurrent events

5.4.1. a Discussion

The analysis of recurrent DCH cohort data is conducted using simple intensity-based model, which is ordinary Cox model on the recurrent data, and then the Andersen – Gill model of recurrent data, which only improvement compared to former is robust variance. These two models result in same hazard rates, approximately 1.25 for log-transformed NO_2 values using fully adjusted model, which is also the same as in the analysis of the single pneumonia occurrence, presented in *Section 5.3*.

Further, the analysis is conducted using more complicated models: frailty model, which includes the random individual effect, conditional Andersen – Gill and conditional Frailty model, which are no longer assuming independence between repeated events.

Among the total DCH population of 53239 individuals, the conditional Andersen – Gill model still didn't show any significant improvement compared to previous models. However, fitted the Frailty and Conditional Frailty model are showing higher association between air pollution and multiple pneumonia hospitalizations. Frailty model adjusted for smoking, ETS and occupational exposure, fitted in the total DCH population for log-transformed NO_2 exposure, results in 1.54 (1.27 – 1.86) hazard rate (95% confidence interval). This means that doubling the NO_2 exposure value we would expect approximately 50% higher risk for recurrent pneumonia occurrence. The hazard rate corresponding to the conditional frailty model is 1.33 (1.23 – 1.45), which is a bit less than not accounting for the events dependency. However, it is very important to stratify on the event number, to make sure proper correlation structure for events is considered, so the conditional frailty model is suggested and reveals 33% higher risk for experiencing pneumonia for doubling NO_2 exposure value.

The strongest association can be seen among the group of DCH cohort members which have already had pneumonia before baseline, which is consistent with the result obtained in analysis of the single pneumonia occurrence. However, the sample is quite small, and that leads to only marginal significance of intensity-based and Andersen – Gill model, whereas for more complicated models association is no longer significant. Stronger association between air pollution exposure and recurrent data using these two models, compared to single event data, is probably only due to more pneumonia cases we get, since the fitted models are not different.

On the first look, the fact that more complicated models of recurrent data, like frailty and conditional models, are not reviling significant association between air pollution exposure and pneumonia incidence among people with previous pneumonia hospitalizations might be a bit strange. However, if we consider that this group of people has experienced pneumonia before baseline at least once; having repeated pneumonia occurrences after the baseline doesn't necessarily give any new information. Hence, if a person had previous pneumonia hospitalizations, then having multiple pneumonias during follow up will just add more cases to that person and won't give more valuable information.

5.5 Model validation

When fitting the Cox proportional hazard model, it is necessary to check model's key assumption - proportionality. This assumption implies that the hazard ratio is constant over time for any two subjects with any combination of covariates.

Tests and graphical diagnostics of the proportional hazard assumptions are based on scaled Schoenfeld residuals. Proportionality test of all relevant predictors (covariates) is done by correlating scaled Schoenfeld residuals with the log-transform age (underlying time scale). Each of the predictor's proportional hazard test from the fully adjusted Cox model of total DCH population (given in *Table 7*) in presented in *Table 11*.

	rho	χ^2 *test stat.*	$p-value$
log NO_2	0.05	7.03	0.01
Gender	0.04	4.24	**0.04**
BMI	0.04	5.97	0.02
Smoking: previously	-0.01	0.33	0.57
Smoking: low	0.03	2.80	0.10
Smoking: medium	0.02	0.94	0.33
Smoking: high	-0.003	0.04	0.85
ETS	0.01	0.64	0.42
Alcohol	0.02	1.08	0.30
Education: Medium	-0.01	0.22	0.63
Education: High	0.005	0.08	0.77
Occup. Exposure	-0.04	4.17	**0.04**
Physical activity: low	0.004	0.07	0.79
Physical activity: high	-0.007	0.13	0.71
Fruit intake	0.03	2.04	0.15
Fat intake	-0.004	0.07	0.79

Table 11: Checking proportional hazard assumption

Test statistics mostly have low values, which corresponds to high $p-value$ meaning that proportional hazard assumption in not violated. Only couple of predictors have low $p-value$ (less than 5%), where the proportionality is under the question. Those are gender with $p=0.04$ and occupational exposure with $p=0.03$. The log-transformed NO_2 and NO_x exposure variables have respectively $p=0.06$ and $p=0.19$ significance levels of proportionality which validates the assumption.

Graphical presentation helps to interpreting the results, thus plots of the scaled Schoenfeld residuals for each of the predictors are obtained. Residual plots for log-transformed NO_2 and NO_x exposure variables are presented in *Figure 20* , for which test doesn't shown any violation of the assumption; and in *Figure 21* for gender and occupational exposure, where test shows that proportionality doesn't hold. Plots of other predictors are presented in *Appendix C*.

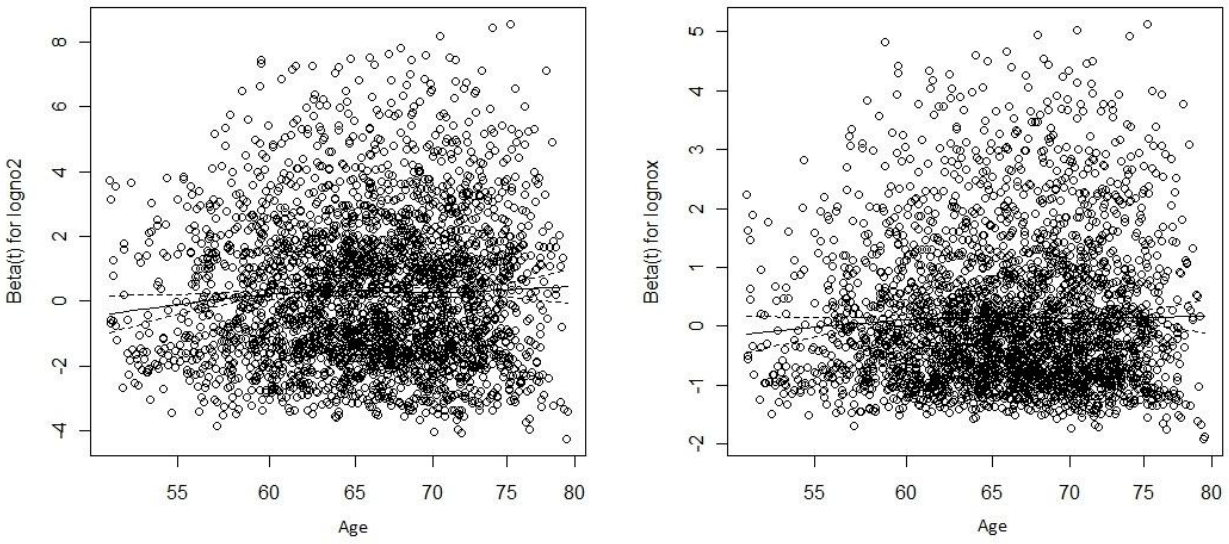

Figure 20: Plots of scaled Schoenfeld residuals against log-transformed time scale (age) for $logNO_2$ (left) and $logNO_x$ (right)

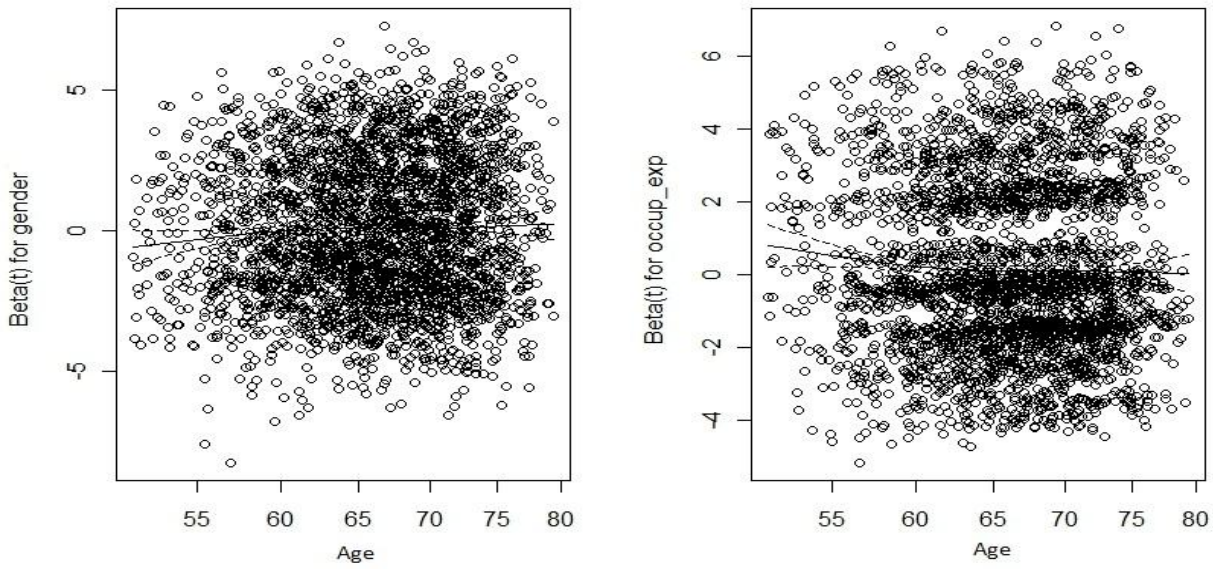

Figure 21: Plots of scaled Schoenfeld residuals against log-transformed time scale (age) for gender (left) and occupational exposure (right)

The solid line in the plot is smoothing - spline fit and dashed lines represent ± 2 standard errors around the fit. The proportional hazard assumes coefficients to be constant over time, i.e. $\beta(t) = \beta$, and thus the fit against age should be horizontal line with zero slope.

In our case proportionality test shows significant violation for gender and occupational exposure but from the residuals plot, we can see that the smoothing-spline fit differ from horizontal line only at the beginning and the end of time scale. There we can also notice fewer numbers of observations which may change the fit shape. Except at the ends of time scale, fit seems to be very good.

Since, the residuals plots where the most of the observations are is expected horizontal line, we can disregards the ends of time scale and conclude that the proportional hazard assumption does hold and discovered positive association between air pollution exposure and risk for pneumonia hospitalizations using Cox model is valid.

Chapter 6

Conclusions and Discussion

6.1 Conclusion

This thesis was aimed to investigate epidemiological data available from the Danish Cancer Society. Applying statistical methods of the survival analysis we have been able to successfully show evidence of association between traffic-related air pollution and pneumonia hospitalizations in the city of Copenhagen.

From the analysis among DCH cohort participants, both for first pneumonia hospitalization and recurrent, we have found that the exposure to NO_2 and NO_x as well as presence of high traffic loads around the residential addresses, is positively significantly associated with the risk of hospital admission for pneumonia.

Considering the exposure length we have found no significant difference in the effect of up to 39-, 29- or 19- year cumulative exposure. Therefore, all the maximum available data has been used as the main exposure proxy, which is since 1971 until the end of follow-up, and applied in all the analysis. Using 39-year exposure and fully adjusted model on the total DCH population showed strong positive association with pneumonia hospitalization, for NO_2 exposure (HR: 1.25; 95%CI: 1.14 – 1.36) and a bit lower for NO_x (HR: 1.11; 95%CI: 1.06 – 1.16). Since the NO_2 and NO_x exposure are highly correlated (corr. coeff. 0.96), investigating effect of one is enough.

A history of co-morbid diseases defined by Charlson index had no influence on the association between air pollution and pneumonia, as the identical associations were observed in cohort subset of people who were healthy (no co-morbidity conditions) and those with prior co-morbid conditions. The conducted analysis on the total DCH population ($n = 53239$) of the time to first pneumonia hospitalization revealed 35% higher risk for being hospitalized for pneumonia among individuals exposed to more than 20 $\mu g/m^3$ of NO_2, as compared those living in clean areas The strongest association is found among the individuals with pneumonia hospitalizations before baseline, where the risk is 92% higher for people exposure to high level of NO_2 (20 $\mu g/m^3$) than the ones with the lowest possible pollution levels (around 14 $\mu g/m^3$).

Furthermore, repeated pneumonia hospitalizations have never been studied before with respect to risk associated with exposure to air pollution, which makes this analyses novel.

Pneumonia is not a chronic disease, but it is a co-morbid condition of many other diseases and can occur multiple times, especially in elderly. For those reasons we wanted to see if recurrent occurrence is also related to air pollution exposure.

The analysis of recurrent DCH data (up to three events) showed that the strong association between air pollution exposure and repeated pneumonia occurrences is still present. It was conducted using the same approach as in first pneumonia case, but also improved by considering individuals with repeated events as a cluster and thus correcting the variance, adding random individual effect since participants are dissimilar, and even further, accounting for events dependency.

The intensity-based model, as well as Andersen – Gill model give similar results as the analysis of the first pneumonia data, but can't be considered adequate for recurrent data. For that reason frailty and conditional models are preferable.

Considering the individuals with multiple events as clusters and accounting for the events dependency using conditional Andersen – Gill model revealed strong association (for NO_2: 1.30; 1.20-1.40) and adding the random effect to account for the individual heterogeneity even stronger (for NO_2: 1.43; 1.19-1.73).

However, the most informative model to be used but also computationally most intensive is conditional frailty model. This model includes random individual effect, which doesn't assume individuals to be the same, as well as accounts for events dependency, allowing for example second event to occur only if the first already did etc., thus it is the most recommended suitable model for our data. Fitting fully adjusted conditional frailty model to the total DCH population revealed 30% higher risk of recurrent events for doubling NO_2 values and 14% for NO_x.

The analysis of recurrent data showed strong positive association between traffic-related air pollution exposure and hospital admissions for pneumoni, but still the effect on the first pneumonia occurrence is a bit stronger. In the first pneumonia occurrence during follow-up, the strongest association was found among individuals with history of pneumonia before baseline which leaded to potentially significant analysis of recurrent pneumonias. Among the same group of individuals recurrent analysis didn't show significant results, since the sample is quite small, but it can also be irrelevant since the first pneumonia after baseline for the individuals with history of pneumonia is already repeated event. For that reason, we can conclude that having a history of pneumonia admissions and high air pollution exposure significantly and considerably increases the risk for new pneumonia occurrences. However, the risk is not limited to this group. Exposure to air pollution is associated with risk for

hospitalization for pneumonia in the whole cohort, also in healthy individuals and those with other co-morbid conditions.

6.2 Discussion

The idea of investigating the effect of air pollution on the occurrence of pneumonia among elderly is quite new. It came from the existing literature, which has shown association between air pollution and risk for experiencing other disease, such as asthma, COPD and stroke; but also lack of investigations about long-term exposure to air pollution and pneumonia hospitalizations (only one study to date). Since pneumonia is known to be age-related disease and exposure to air pollution is known to be positively related to all major chronic lung diseases, the purpose of this study was to investigate how traffic-related air pollution in the area of Copenhagen influences the risk of experiencing pneumonia once or multiple times among elderly.

The main founding of this study is that there is a strong positive association between long – term exposure to traffic-related air pollution and pneumonia, as well as repeated pneumonia admissions. This association is the strongest for readmission for pneumonia in individuals with prior pneumonia hospitalizations.

This thesis has contributed with an example of the methods used in survival analysis on the “real world” data. The Danish Diet, Cancer and Health cohort and all relevant data from the Danish registries systems ware obtained from the Danish Cancer Society and analyzed using ordinary Cox regression as well as its several extensions required for recurrent data analysis.

Chapter 7

Considerations and Further Work

7.1 Considerations

We found that the traffic-related air pollutants NO_2 and NO_x are relevant for pneumonia hospitalization, NO_2 having the strongest effect. NO_2 is an airway irritant, which even at the low concentrations found in everyday life can cause respiratory tract infections by interacting with the immune system and may play a role in lung inflammation [53].

We cannot conclude from our study that NO_2 is of pathological significance in pneumonia or just an indicator of other harmful pollutants originating from traffic, especially particles. In any case, the results of this study provide evidence that traffic-related urban air pollution contributes to the development of pneumonia and that reductions in traffic emissions would be beneficial for public health.

A consideration of our study regarding health outcome is that only pneumonia hospitalizations were considered, since we didn't have self-reported data about pneumonia. This means that pneumonia occurrence is underestimated here since not all pneumonia cases are hospitalized, and that with pneumonia hospitalizations we studied the most serious pneumonia cases.

Furthermore, we have couple of consideration from the statistical modeling point of view. In analyzing the survival data our goal was to use the Cox model, ordinary and extended. However, other possibilities could also be parametric models, such as exponential model, the Weibull model or log – logistic model [54]; or non-parametric models, such as additive hazard model [55]. Both advantages and disadvantages could arise using these models compared to the applied Cox models but it would be relevant to apply some of these models and see how well they fit the data.

However, available literature says that fitting parametric model in case of testing couple of groups with different exposures or treatments, and looking at their comparisons, may need to be adjusted for other subject's characteristics, which makes a lot of model assumptions. In this case it's recommended to apply ordinary Cox regression, like in this study [54]. The situations of non-proportional hazard, which is not the case in this study, frequently occur when the covariates are time-varying. According to the literature, a possible way to handle that is to use

nonparametric Aalen additive hazard model which is very flexible and provides a good fit [55]. However, it is very important to test if all the covariates are time-varying or some of them are constant over time. In case of mixture of these two, a semi- parametric model would be more correct to apply since it gives more precise information about the estimated effects [55].

7.1.1 Limitations

Every "real – life" study has some limitations, so does this one. The dispersion models we used to assess NO_2 and NO_x concentrations at the addresses of study participants are only surrogates of real exposures and are inevitably associated with some exposure misclassification. However, the model have been validated [37] and applied in Denmark [11,12,56,57] and the United States [55] and possible misclassification is likely to be non-differential with respect to development of pneumonia. A previous comparison of measured NO_2 concentrations with those calculated from a Danish dispersion model showed that misclassification was primarily of the Berkson type, typically associated with exposures predicted from the model. This type of error is not expected to bias the estimates, although it may decrease their precision [56].

A further limitation of the exposure assessment method is that we assessed only outdoor concentrations and lacked information on work address, commuting habits, and personal activities.

7.1.2 Strengths

The main strength of this study is the objective definition of incidence of pneumonia as the first admission, as well as repeated admissions, for this condition in the Danish Hospital Discharge Register, a nationwide register of routinely collected data with no loss to follow-up. Another strength is the large prospective cohort with available residential address history and information on potential confounders collected before admissions for pneumonia. As the Danish Diet, Cancer, and Health cohort was not originally designed to study pneumonia or air pollution; and pneumonia hospitalizations, vital status, and the information on addresses used in modeling air pollution were obtained from the reliable population-based registries; the possibility of information or recall bias with respect to health outcome or differential bias with respect to exposure is minimal.

7.2 Further work

The analyses of an effect of long-term exposure to air pollution could be extended with a study of several other relevant issues. First, it would be interested to study whether the effect of exposure to air pollution on pneumonia is modified by any other factors, for example by smoking. The effect may be same in smokers and non-smokers (implying no interaction), or perhaps stronger among smokers than non-smokers, suggesting additive harmful effects of

simultaneous exposure to tobacco smoke and air pollution. It could also be so that smokers have marginal additional harmful effect from exposure to air pollution, and that effect of air pollution is limited only to non-smokers. No studies have yet provided data on modification of the effect on air pollution by smoking. Similarly, it would be interested to test for other potential effect modifiers, such as gender, age, or BMI.

When dealing with recurrent data, an important question arises in the area of frailty models. The frailty distribution must be positive, thus mostly recommended is gamma distribution, also used in this study. However, other frailty distributions, such as Gaussian distribution, are also an option. For that reason, it might be relevant to investigate if another frailty distribution could capture more information from the data than the other, and thus leads to the frailty model with better performance.

Other possibly very relevant analyses would be to incorporate the effect of short-term exposure to air pollution, that is levels of air pollution on the same or day before hospitalization. The short-term exposure to air pollution has documented effect on pneumonia and other lung diseases. However, these analyses would require data on daily mean levels of air pollution, which are routinely measured at background air pollution monitors in Copenhagen and Aarhus located centrally in these cities, and which cannot be assumed representative for all DCH cohort members. Also, different statistical considerations would be necessary for this analysis, and that is why no studies exist yet that have looked into effect of short and long-term exposure to air pollution simultaneously on any lung disease.

Appendix A

Definitions

A

Alveoli: Small air sacs or cavities in the lung that give the tissue a honeycomb appearance and expand its surface area for the exchange of oxygen and carbon dioxide. Branches off the bronchioles. [26]

Andersen-Gill model: is a model used for recurrent event data. The model considers every patient as one individual cluster instead of every event as independent and thereby a robust variance can be estimated for the model. [46]

B

BMI: The Body Mass Index is a statistical measure which compares a person's weight and height, and it is a widely used diagnostic tool to identify weight problems within a population. The formula for BMI is $weight(kg)/(height(m))^2$. [2]

Berkson misclassification (error model): is description of random error (or misclassification) in measurement. Unlike classical error, Berkson error causes little or no bias in the measurement. [26]

C

Cardiac diseases: are a group of disorders of the heart and blood vessels. [2]

Censoring: is in association with survival analysis when an individual steps out of a trial early (of irrelevant reasons relative to the trial). If the patient do not have any events during the trial period, and is therefore still under risk, he will also be defined as censored in the end of the trial. [1]

Confounding: In statistics, a confounding variable (also confounding factor, lurking variable, or confounder) is an extraneous variable in a statistical model that correlates (positively or negatively) with both the dependent variable and the independent variable. [26]

Cohort: In statistics, a group of subjects with a common defining characteristic – typically age group. But despite from an open population, which can include or exclude people, if they do not meet a certain defined criteria, the cohort is a specific group of people to be followed from one point in time to another. The start time could be at birth, or when diagnosed with a disease. [1]

Counting process: is a process that records the (uncensored) events in e.g. survival data as time proceeds. [50]

Cox regression: is a class of survival models in statistics. It is a statistical technique that is used to determine the relationship between survival and several independent exploratory variables. Cox regression relates the time that passes before some event occurs to one or more covariates that may be associated with that quantity. [26]

D

Diabetes: is a metabolic disease affecting the way the bodies use digested food for growth and energy. Diabetics have a high blood sugar level, either because the body does not produce enough insulin, or because cells do not respond to the insulin that is produced. [2]

E

Epidemiology: is the study of how disease is distributed in population and the factors that influence or determine this distribution. [1]

Environmental epidemiology: is the epidemiologic study of the health consequences of exposure that are involuntary and that occur in the general environment (air, water, diet, soil, etc.). [2]

F

Frailty model: is a model that incorporates an unmeasured random effect in the hazard function to account for heterogeneity in subjects, usually used on recurrent data. [46]

H

Hazard function: is a function used in survival analysis and is a product of two functions. Partly the baseline hazard function that characterizes how the hazard function changes as a function

of survival time. And partly of a function that characterizes how the hazard function changes as a function of subject covariates. [46]

Hypertension: is a chronic medical condition in which the blood pressure is elevated. It is also referred to as high blood pressure or shortened to HT, HTN or HPN. [26]

J

Jackknife method: provides an alternative and reasonably robust method for determining the error from the data to the parameters. The jackknife derives estimates of the parameter of interest from each of several subsamples of the parent sample and then estimates the variance of the parent sample estimator from the variability between the subsample estimates. [46]

K

Kaplan-Meier: is an estimator of the survival function and is also called the product limit estimator. This estimator is widely used in survival analysis. [43]

L

Likelihood ratio test: is a statistical test used to compare the fit of two models. The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other. [26]

N

Non-parametric model: is a family of distributions that cannot be described using a finite number of parameters and is contrasted with the parametric model. [50]

Null Cox model: is a Cox model without any explanatory variables; hence β_0 is the overall population risk. [46]

P

Parametric (survival) models: is a family of distributions that can be described using a finite number of parameters. More specific, parametric survival model is one in which survival time is assumed to follow a known distribution. [43]

Particulate matter or fine particles: are tiny subdivisions of solid or liquid matter suspended in a gas or liquid. Particles often have irregular shapes with actual geometric diameters that are difficult to measure. [26]

Pneumonia: is a form of acute respiratory infection that affects the lungs. The lungs are made up of small sacs called alveoli, which fill with air when a healthy person breathes. When an individual has pneumonia, the alveoli are filled with pus and fluid, which makes breathing painful and limits oxygen intake. [2]

R

Recurrent event data: is data where a single event can occur multiple times in a subject in a certain period. [46]

Relative risk: is a ration of the probability of the event occurring in the exposed group versus a non-exposed group. It is also referred to as hazard ratio and is in Cox regression also called the odd ratio. [27]

Robust variance: in ordinary Cox models with recurrent events data, where estimate of variance for the covariate effects treats each of the observations as independent, a robust variance can be introduced. If a subject do have multiple events it is therefore reasonable to use another estimation of variance which could be a grouped jackknife estimate. The estimate of variance would thereby be more robust. [46]

S

Semi-parametric model: In statistics is a model that has parametric and nonparametric components. [50]

Survival analysis: is just another name for time to a single event analysis. The term survival analysis is used predominately in biomedical sciences where the interest is in observing time to death occurrence of disease. [43]

Survival function: gives the probability of observing a survival time greater than or equal to some stated value. In most applied settings the interest lies in describing how long the subjects live, which is fundamental to a survival analysis. [43]

Appendix B

Acronym table

Acronym	Term
BMI	Body Mass Index
CI	Confidence interval
CVD	Cardiovascular Diseases
df	Degree of freedom
ETS	Environmental tobacco smoke
HR	Hazard ratio
PM	Particulate Matter
SES	Socio-economic status

Appendix C

Supplementary figures

This appendix contains supplementary figures of survival and cumulative hazard plots for each of the relevant confounders for this analysis. Then the Schoenfeld residuals of covariates from the fully adjusted Cox model where the proportional hazard assumption is valid are displayed.

C.1 Survival Curves and Cumulative Hazards

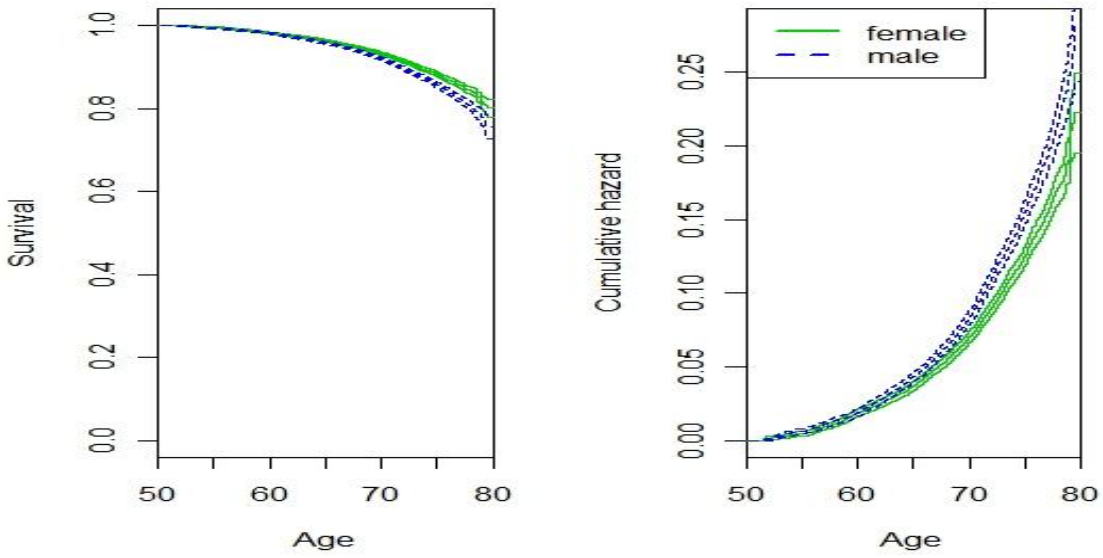

Figure 22: Survival func. and cumulative hazard func. for Gender of DCH cohort members

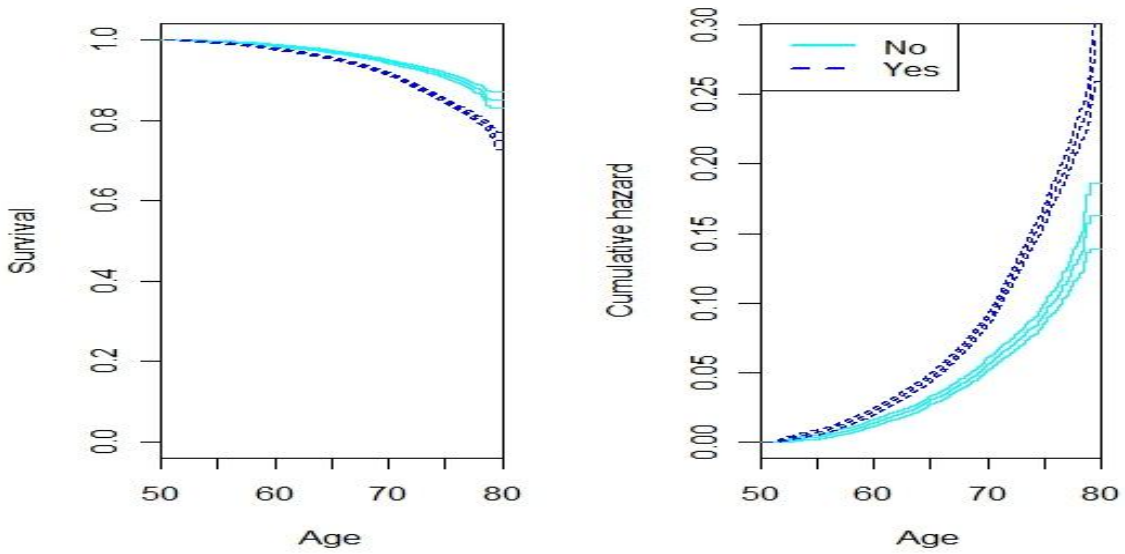

Figure 23: Survival func. and cumulative hazard func. for ETS status of DCH cohort members

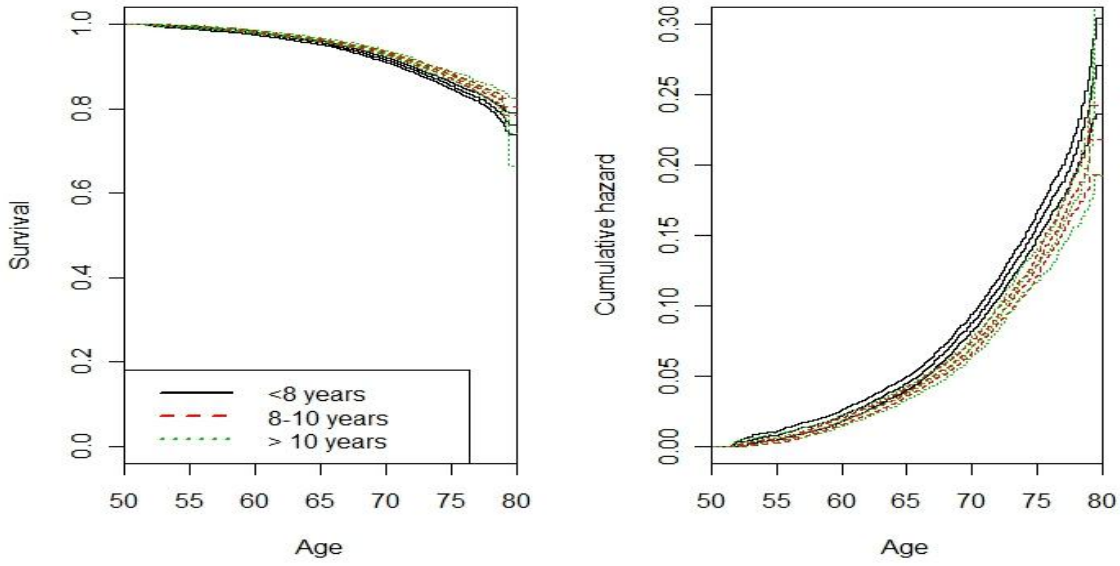

Figure 24: Survival func. and cumulative hazard func. for Educational status of DCH cohort members

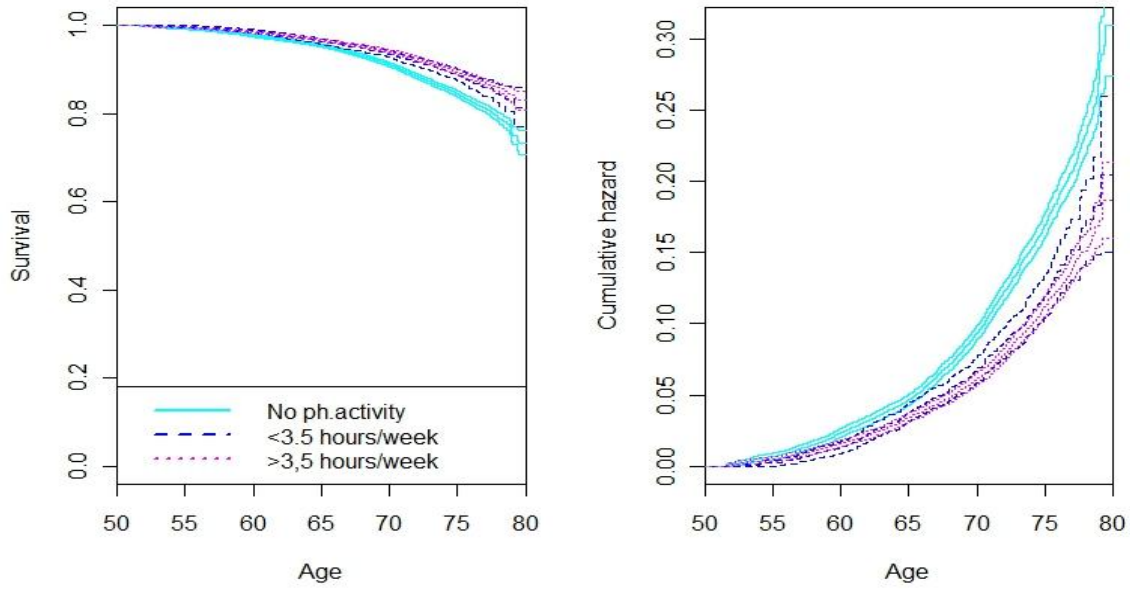

Figure 25: Survival func. and cumulative hazard func. for physical activity of DCH cohort members

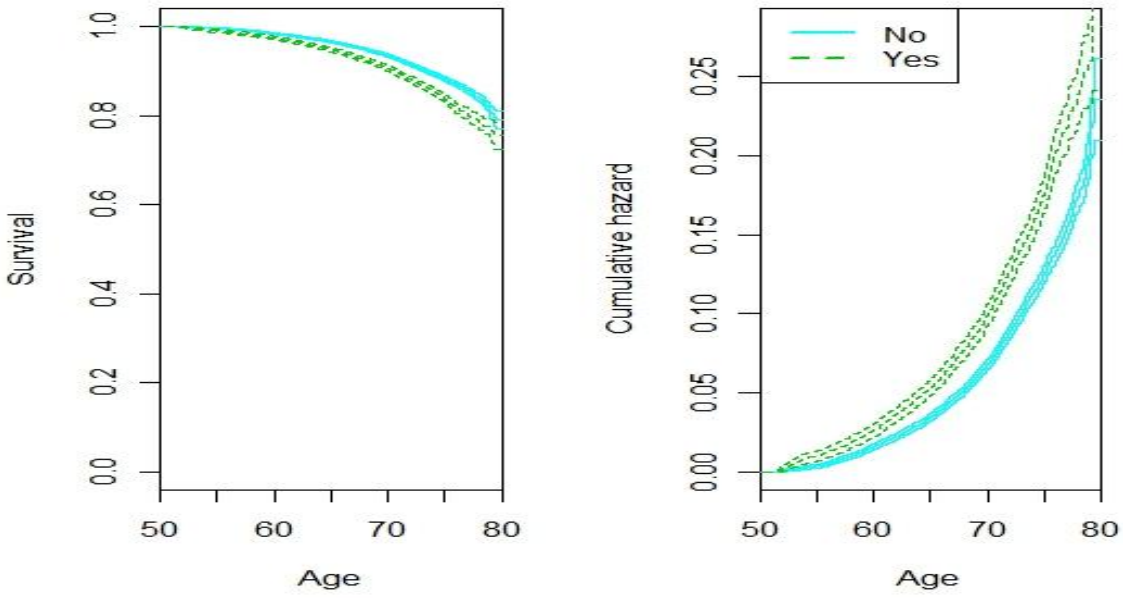

Figure 26: Survival func. and cumulative hazard func. for Occupational exposure of DCH cohort members

C.2 Checking PH assumption – Schoenfeld residuals

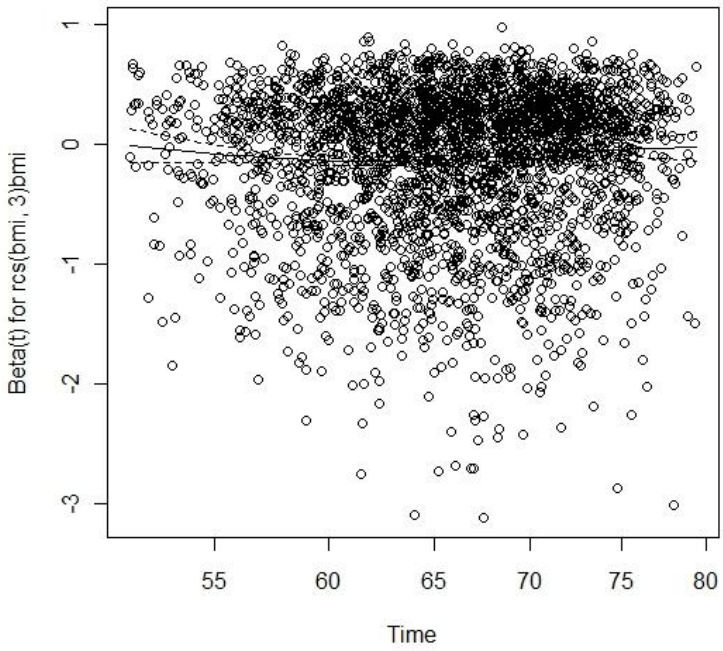

Figure 27: Plots of scaled Schoenfeld residuals against log-transformed time scale (age) for BMI

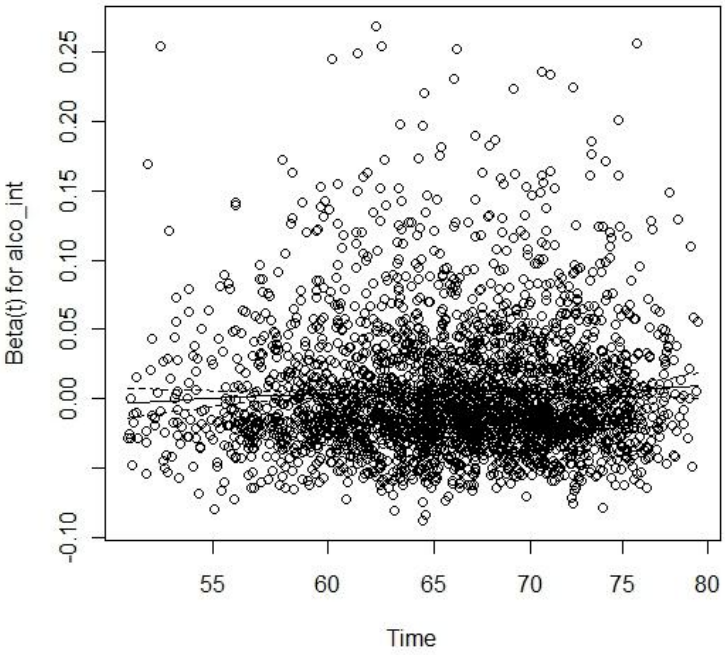

Figure 28: Plots of scaled Schoenfeld residuals against log-transformed time scale (age) for Alcohol intake

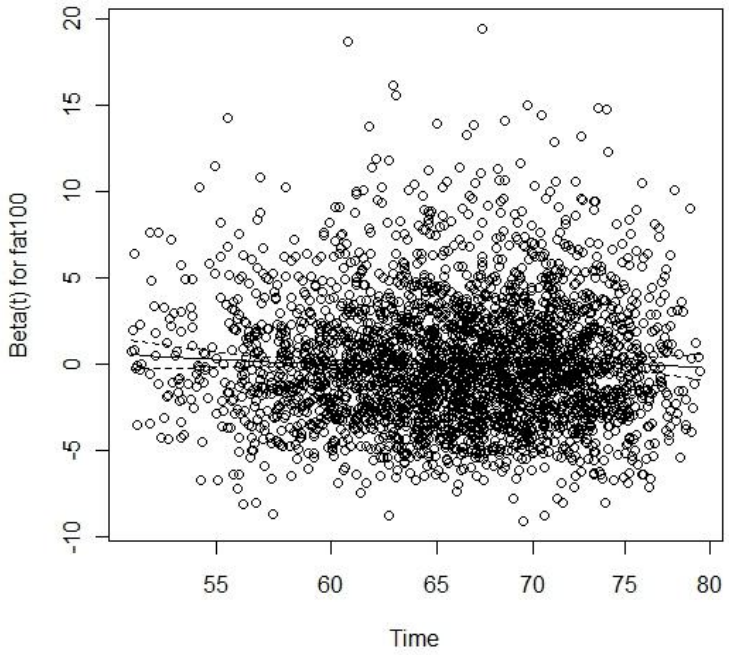

Figure 29: Plots of scaled Schoenfeld residuals against log-transformed time scale (age) for Fat intake

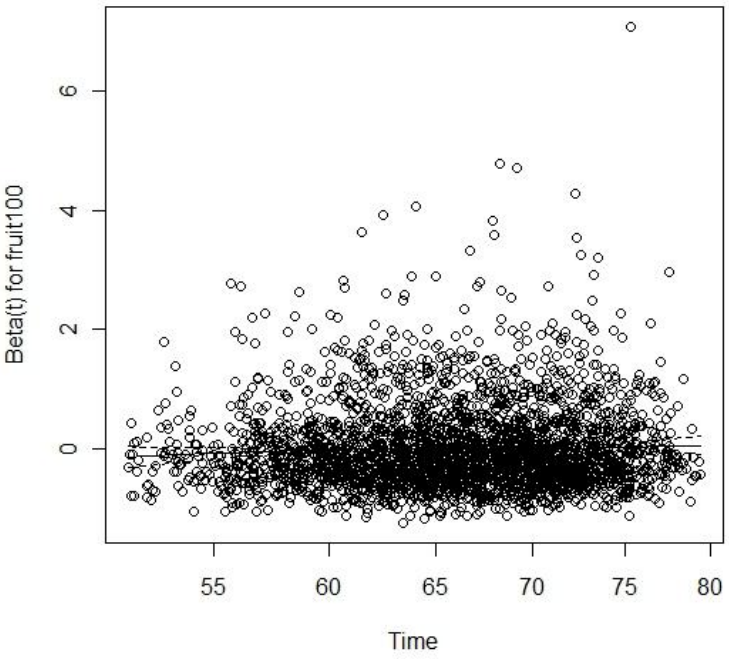

Figure 30: Plots of scaled Schoenfeld residuals against log-transformed time scale (age) for fruit intake

Appendix D

R programming

D.1 Data preparation

```
#### DCH COHORT DATA

#Questionnaire
spskema = read.csv("D:/Pneumonia/Data/spskema.csv", header = TRUE, sep = ",")

##Smoking status
smoking = read.csv("D:/Pneumonia/Data/smokingsta.csv", header = TRUE, sep = ",")

#Environmental tobacco smoke (ETS)
ets <- read.csv("D:/Pneumonia/Data/ets.csv", header = TRUE, sep = ",")

#Additionally excluded for cancer before baseline
eksklud571 = read.csv("D:/Pneumonia/Data/eksklud571.csv", header = TRUE, sep = ",")

#Vital status in 2006
vitalsta2006 = read.csv("D:/Pneumonia/Data/vitalsta2006.csv", header = TRUE, sep = ",")

#Merge data sets - spskema, smoking, ets, sport
data1 <- merge(spskema, smoking, by.x = "knr", by.y = "id", all = FALSE)
data2 <- merge(mm1, ets, by = "knr", all = TRUE)
data3 <- merge(mm2, sport, by = "knr", all = TRUE)


#Remove eksklud571 from the data
exclude  <- which(data3[, 1] %in% eksklud571[,1])
data  <- data3[-exclude,]

#####################################################

###Potential confounders and vital status

data_pne <- data.frame(cbind(knr = data $knr, age = mm$age, gender = mm$kqn, bmi = mm$bmi,
                      education = mm$s56x01n, alcohol = mm$noalko, alco_int = mm$alko,
                      smoking = mm$rygning, smo_duration = mm$varighed, smo_int = mm$forbrug,
                       ets = mm$ets, occup_exp = mm$etsenrol, sport = mm$sport, sport_int = mm$tsportint,
                      fat = mm$fedt, fruit = mm$fruit))
```

```
data_pne <- merge(data_pne, vital2006[,2:4], by = "knr", all = TRUE)

#Extract relevant diagnoses from lpr data
#LPR register for DCH cohort members
lprdata <- read.csv("I:/Pneumonia/lprdata.csv", header = TRUE, sep = ",")

#Pneumonia diagnosis in ICD-10 and ICD-8
pne10 <- c("DJ120", "DJ121", "DJ122", "DJ123", "DJ124", "DJ125", "DJ126", "DJ127", "DJ128", "DJ129",
           "DJ130", "DJ131", "DJ132", "DJ133", "DJ134", "DJ135", "DJ136", "DJ137", "DJ138", "DJ139",
           "DJ140", "DJ141", "DJ142", "DJ143", "DJ144", "DJ145", "DJ146", "DJ147", "DJ148", "DJ149",
           "DJ150", "DJ151", "DJ152", "DJ153", "DJ154", "DJ155", "DJ156", "DJ157", "DJ158", "DJ159",
           "DJ160", "DJ161", "DJ162", "DJ163", "DJ164", "DJ165", "DJ166", "DJ167", "DJ168", "DJ169",
           "DJ170", "DJ171", "DJ172", "DJ173", "DJ174", "DJ175", "DJ176", "DJ177", "DJ178", "DJ179",
           "DJ180", "DJ181", "DJ182", "DJ183", "DJ184", "DJ185", "DJ186", "DJ187", "DJ188", "DJ189")
pne8 <- c(48000:48699)

#Ornithosis diagnosis in ICD-10 and ICD-8
orn10 <- c("DA4810", "DA4811", "DA4812", "DA4813", "DA4814", "DA4815", "DA4816", "DA4817",
"DA4818", "DA4819")
orn8 <- c(07300:07399)

#Leginellosis diagnosis in ICD-10 and ICD-8
leg10 <- c("DA7090", "DA7091", "DA7092", "DA7093", "DA7094", "DA7095", "DA7096", "DA7097",
"DA7098", "DA7099")
leg8 <- c(47100:47199)

diag <- c(pne10, pne8, orn10, orn8, leg10, leg8)

#All pneumonia hospitalizations
PNEdiag <- which(lprdata$c_diag %in% diag)

#LPR subset for pneumonia
lprPNE <- lprdata[PNEdiag,]

#Subset of pneumonia cases before baseline
lprPbef <- subset(lprPNE,lprPNE$d_inddt <= lprPNE$mdato)
#No.of individuals hospotalized for the first time before baseline
#length(unique(lprPbef$knr))                    #746
```

```
#Subset of no pneumonia before baseline
lprP <-  subset(lprPNE,lprPNE$d_inddt > lprPNE$mdato)

bef <- which(data_pne$knr %in% unique(lprPbef$knr))
#data_Pbef <- data_pne[bef,]
data_pne <- data_pne[-bef,]

#########
#Pneumonia status variable
pne_sta <- matrix(NA, nrow=dim(data_pne)[1], ncol=2)
colnames(pne_sta) <- c("knr", "pnesta")
#1stcol - knr
pne_sta[,1] <- data_pne$knr

#2ndcol - status: 1 = pne, 0 = nopne
for(i in 1:length(pne_sta[,2])) {
        if(pne_sta[i,1] %in% individuals) pne_sta[i,2] <- 1
        else pne_sta[i,2] <- 0
        }

data_pne <- merge(data_pne, pne_sta, by = "knr", all = FALSE)
```

```
###################################
#####   Time to 1st event!!!
###################################

#Censoring variable (status)
censor <- matrix(NA, nrow=dim(data_pne)[1], ncol=3)
colnames(censor) <- c("knr", "censorsta", "censordate")
#1stcol - knr
censor[,1] <- data_pne$knr

#2ndcol - status: 1 = censored, 0 = not censored
#3rdcol - censoring date
```

```
for(i in 1:length(censor[,2])) {
	if(data_pne$status2006[i]==0 & data_pne$pnesta[i]==0) censor[i,2] <- 0
	if(data_pne$status2006[i]==0 & data_pne$pnesta[i]==1) censor[i,2] <- 1 & censor[i,3] <-
lprdata$d_inddto[i]
	if(data_pne$status2006[i]==1 & data_pne$pnesta[i]==0) censor[i,2] <- 0
	if(data_pne$status2006[i]==1 & data_pne$pnesta[i]==1) censor[i,2] <- 1 & censor[i,3] <-
lprdata$d_inddto[i]

	if(data_pne$status2006[i]==3) censor[i,2] <- 0 & censor[i,3] <- 6/27/2006
	if(data_pne$status2006[i]==5) censor[i,2] <- 0 & censor[i,3] <- 6/27/2006

	if(data_pne$status2006[i]==60) censor[i,2] <- 0 & censor[i,3] <- data_pne$sdato2006[i]
	if(data_pne$status2006[i]==70) censor[i,2] <- 0 & censor[i,3] <- data_pne$sdato2006[i]

	if(data_pne$status2006[i]==80 & data_pne$pnesta[i]==0) censor[i,2] <- 0 & censor[i,3] <-
data_pne$sdato2006[i]
	if(data_pne$status2006[i]==80 & data_pne$pnesta[i]==1) censor[i,2] <- 1 & censor[i,3] <-
lprdata$d_inddto[i]
	if(data_pne$status2006[i]==90 & data_pne$pnesta[i]==0) censor[i,2] <- 0 & censor[i,3] <-
data_pne$sdato2006[i]
	if(data_pne$status2006[i]==90 & data_pne$pnesta[i]==1) censor[i,2] <- 1 & censor[i,3] <-
lprdata$d_inddto[i]
}

#Remove dates after the follow up
for(i in 1:dim(censor)[1])
{
	ifelse(censor[i,3] >= cens_date, censor[i,3] <- cens_date, censor[i,3] <- censor[i,3])
}

data_pne <- merge(data_pne, censor, by="knr")

###################
#Age of cohort members at the entry and the end of the study
data_pne$agestart <- (data_pne$entrydate - data_pne$bdate)/365.25
data_pne$ageend <- (data_pne$end_date - data_pne$bdate)/365.25

###################################
#Recurrent events  (at least 30days between two hospitalizations)
lprP1 <- subset(lprP, lprP$count==1)
lprMulti <- lprP[which(lprP$count!=1),]
```

```
dim(lprMulti)                                                    # 3624    3

lprMulti$multista <- rep(NA, length(lprMulti[,1]))

for(i in 1:length(unique(lprMulti$knr)) ) {
        mp <- which(lprMulti$knr==unique(lprMulti$knr)[i])

        for(j in 1:length(mp)) {
                lprMulti$multista[mp[1]] <- 1
                ifelse(lprMulti$d_inddto[mp[j+1]] >= (lprMulti$d_inddto[mp[j]]+31),
                lprMulti$multista[mp[j+1]] <- 1, lprMulti$multista[mp[j+1]] <- 0)
        }
}

#Hospitalizations with at least 30days between
lprMulti1 <- lprMulti[which(lprMulti$multista==1),-4]

###LPR for multiple event!!!
lprP1 <- merge(lprP1, lprMulti1, all=TRUE)

for(j in 1:length(lprP1$knr)) {
        lprP1$count1[j] <- sum(lprP1$knr==lprP1$knr[j])
        }

#Indiv. with multiple events
length(unique(lprP1$knr[which(lprP1$count1!=1)]))       # 626

#########
aPl <- which(airPne$knr %in% lprP1$knr)
air1 <- airPne[aPl,]

air1$multipne <- rep(NA, length(air1$knr))

for( i in 1:length(unique(air1$knr))) {
        a1 <- which(air1$knr==unique(air1$knr)[i])
        l1 <- which(lprP1$knr==unique(air1$knr)[i])

        for( j in 1:length(l1)) {
                for( k in 1: length(a1)) {
                ifelse((lprP1$d_inddto[l1[j]] >= air1$riskstart[a1[k]] &
                        lprP1$d_inddto[l1[j]] <= air1$riskend[a1[k]]),
```

```
                                air1$multipne[a1[k]] <- 1, air1$multipne[a1[k]] <- 0 )
        }}
}

#########

for(l in 1:length(lprP$knr)) {
fdd <- ifelse(lprP$knr[l]==airPne$knr & lprP$d_inddto[l] >= airPne$riskstart &
        lprP$d_inddto[l] <= airPne$riskend, 1, 0)
}

##################
```

D.2 Testing potential confounders – Univariate Cox regression

```
#########################
##### CATEGORIES

#Two age groups - under and over 56 years old
#mean(data_pne$age)  #56.17682
#mean(data_pne$agestart)        #56.67808
data_pne$age_cat <- cut(data_pne$age, c(0,56,1000),1:2)

#BMI categories
data_pne$bmi3 <- cut(data_pne$bmi,c(0,20,30,1000), c('underweight', 'normal', 'obese'))
data_pne$obese <- cut(data_pne$bmi,c(0,30,1000),c('normal','obese'))

#Smoking categories
smoint_cat <- cut(data_pne$smo_int,c(0,15,25,1000),c("low","med","high"))
current <- which(data_pne$smoking == 3)

air$smo_cat <- replace(air$smoking, current, as.character(smoint_cat[current]))

#Alcohol categories
alc <- air$alco_int/12                                  # drinks per day
alc_w <- alc * 7                                        # drinks per week
air$alc_cat <- cut(alc_w, c(0,1,20,1000), c('no drinks','1-20drinks','21+drinks'))
```

```
#Physical activity
sportint_cat <- cut(data_pne$sport_int,c(0,3.5,1000),c('low','high'))

phactive <- which(data_pne$sport == 1)
data_pne$sport_cat <- replace(data_pne$sport, phactive, as.character(sportint_cat[phactive]))

##############################################
data_pne$age_cat <- as.factor(data_pne$age_cat)
data_pne$gender <- as.factor(data_pne$gender)
data_pne$bmi_cat <- as.factor(data_pne$bmi_cat)
data_pne$obese <- as.factor(data_pne$obese)
data_pne$education <- as.factor(data_pne$education)
data_pne$alcohol <- as.factor(data_pne$alcohol)
data_pne$smoking <- as.factor(data_pne$smoking)
data_pne$smo_cat <- as.factor(data_pne$smo_cat)
data_pne$ets <- as.factor(data_pne$ets)
data_pne$occup_exp <- as.factor(data_pne$occup_exp)
data_pne$sport <- as.factor(data_pne$sport)
data_pne$sport_cat <- as.factor(data_pne$sport_cat)

#alco_int => continuous 100g/day
#data_pne$smo_duration => continuous
#data_pne$fedt & data_pne$fruit => continuous 100g/day

#####
attach(data_pne)

############################
####### UNI COX MODEL #########
library(survival)

#Age
mod1 <- coxph(Surv(agestart, ageend, censorsta)~age_cat, data_pne)
#Check proportional hazard assumption
#cox.zph(mod1, transform="log")

#Gender
mod2 <- coxph(Surv(agestart, ageend, censorsta)~gender, data_pne)
cox.zph(mod2, transform="log")
```

```
#BMI
mod3 <- coxph(Surv(agestart, ageend, censorsta)~bmi_cat, data_pne)
cox.zph(mod3, transform="log")

#Education
mod4 <- coxph(Surv(agestart, ageend, censorsta)~education, data_pne)
cox.zph(mod4, transform="log")

#Alcohol
mod5 <- coxph(Surv(agestart, ageend, censorsta)~alc_cat, data_pne)
cox.zph(mod5, transform="log")

#Smoking
mod6 <- coxph(Surv(agestart, ageend, censorsta)~smo_cat, data_pne)
cox.zph(mod6, transform="log")

#ETS
mod7 <- coxph(Surv(agestart, ageend, censorsta)~ets, data_pne)
cox.zph(mod7, transform="log")

#Occupational exposure
mod8 <- coxph(Surv(agestart, ageend, censorsta)~occup_exp, data_pne)
# cox.zph(mod8, transform="log")

#Physical activity
mod9 <- coxph(Surv(agestart, ageend, censorsta)~sport_cat, data_pne)
# cox.zph(mod9, transform="log")

#Fruit and Fat
#Mean fruit and fat consumption (100g/day)
fat100 <- data_pne$fat/100
fruit100 <- data_pne$fruit/100

mod10 <- coxph(Surv(agestart, ageend, censorsta)~fat100, data_pne)
mod11 <- coxph(Surv(agestart, ageend, censorsta)~fruit100, data_pne)
# cox.zph(mod10, transform="log")
# cox.zph(mod11, transform="log")
```

```
#Income
kincome <- data_pne$Komincome/100000
mod12 <- coxph(Surv(agestart, ageend, censorsta)~kincome, data_pne)
#mod12_ph <- cox.zph(mod12, transform="log")

####################
#EXPOSURE

mod_no2 <- coxph(Surv(agestart, ageend, censorsta)~no2, data_pne)
mod_nox <- coxph(Surv(agestart, ageend, censorsta)~nox, data_pne)

mod_logno2 <- coxph(Surv(agestart, ageend, censorsta)~log(no2,2), data_pne)
mod_lognox <- coxph(Surv(agestart, ageend, censorsta)~log(nox,2), data_pne)

cox.zph(mod_logno2, transform="log")
cox.zph(mod_lognox, transform="log")

#Checking functional form
### SPLINES
library(Design)
d <- datadist(data_pne)
options(datadist="d")

#Age
fit1 <- cph(Surv(agestart, ageend, censorsta) ~ rcs(age,3), data_pne)
#BMI
fit3 <- cph(Surv(agestart, ageend, censorsta) ~ rcs(bmi,3), data_pne)
#Smoking
fit4 <- cph(Surv(agestart, ageend, censorsta) ~ rcs(smo_duration,3), data_pne)
#fit4a <- cph(Surv(agestart, ageend, censorsta) ~ rcs(smo_cat,3), data_pne)

fit10 <- cph(Surv(agestart, ageend, censorsta) ~ rcs(fedt,3), data_pne)
fit11 <- cph(Surv(agestart, ageend, censorsta) ~ rcs(fruit,3), data_pne)
```

D.3 Modeling the exposure to air pollution

```
```

```
#Total population with 1st event
air <- read.csv("D:/Pneumonia/Data/airKKHlong.csv", header = TRUE, sep = ",")
length(unique(air$knr))             # 53239

#Total population with multiple events
air <- read.csv("D:/Pneumonia/Data/airKKHmulti.csv", header = TRUE, sep = ",")
length(unique(air$knr))             # 53239

###Compute cumulative mean NO2 and NOx exposure
#No.of days at risk
air$d <- as.vector(unlist(tapply(air$daysrisk, air$knr, cumsum)))

#NO2
air$deno2 <- as.vector(unlist(tapply(air$daysrisk * air$avno2_mod, air$knr, cumsum)))
air$no2 <- air$deno2/air$d
#NOx
air$denox <- as.vector(unlist(tapply(air$daysrisk * air$avnox_mod, air$knr, cumsum)))
air$nox <- air$denox/air$d

#From '71 - entire exposure history (up to 40years)
air71 <- air
###Compute cumulative mean NO2 and NOx exposure
#No.of days at risk
air71$d <- as.vector(unlist(tapply(air71$daysrisk, air71$knr, cumsum)))

#NO2
air71$deno2 <- as.vector(unlist(tapply(air71$daysrisk * air71$avno2_mod, air71$knr, cumsum)))
air71$no2 <- air71$deno2/air71$d
#NOx
air71$denox <- as.vector(unlist(tapply(air71$daysrisk * air71$avnox_mod, air71$knr, cumsum)))
air71$nox <- air71$denox/air71$d

#From 1981, and 1991
air81 <- subset(air, air$year >= 1981)
air91 <- subset(air, air$year >= 1991)
###Compute cumulative mean NO2 and NOx exposure
#No.of days at risk
air81$d <- as.vector(unlist(tapply(air81$daysrisk, air81$knr, cumsum)))
air91$d <- as.vector(unlist(tapply(air91$daysrisk, air91$knr, cumsum)))

#NO2
```

```
air81$deno2 <- as.vector(unlist(tapply(air81$daysrisk * air81$avno2_mod, air81$knr, cumsum)))
air81$no2 <- air81$deno2/air81$d

air91$deno2 <- as.vector(unlist(tapply(air91$daysrisk * air91$avno2_mod, air91$knr, cumsum)))
air91$no2 <- air91$deno2/air91$d

#NOx
air81$denox <- as.vector(unlist(tapply(air81$daysrisk * air81$avnox_mod, air81$knr, cumsum)))
air81$nox <- air81$denox/air81$d

air91$denox <- as.vector(unlist(tapply(air91$daysrisk * air91$avnox_mod, air91$knr, cumsum)))
air91$nox <- air91$denox/air91$d

####
#Cut off the data before baseline
airPne <- subset( air, air$riskstart!=air$riskend)
airPne71 <- subset( air71, air71$riskstart!=air71$riskend)
airPne81 <- subset( air81, air81$riskstart!=air81$riskend)
airPne91 <- subset( air91, air91$riskstart!=air91$riskend)

length(unique(air$knr))                 # 53239
length(unique(airPne71$knr))            # 53239
length(unique(airPne81$knr))            # 53239
length(unique(airPne91$knr))            # 53239


write.table(airPne, "D:/Pneumonia/Data/airMultifromBaseline.csv", sep=",", row.names=FALSE,
col.names=TRUE, quote = FALSE)

write.table(airPne71, "D:/Pneumonia/Data/air71fromBaseline.csv", sep=",", row.names=FALSE,
col.names=TRUE, quote = FALSE)
write.table(airPne81, "D:/Pneumonia/Data/air81fromBaseline.csv", sep=",", row.names=FALSE,
col.names=TRUE, quote = FALSE)
write.table(airPne91, "D:/Pneumonia/Data/air91fromBaseline.csv", sep=",", row.names=FALSE,
col.names=TRUE, quote = FALSE)
###################
```

D.4 Ordinary Cox regression models - time to first pneumonia

```
#Modeled exposure to NO2 and NOx
```

```
#Read data <- cummean exposure from '71, '81 or '91 (from baseline)
airPne <- read.csv("D:/Pneumonia/Data/air71FromBaseline.csv", sep=",", header=T)
airPne <- read.csv("D:/Pneumonia/Data/air81FromBaseline.csv", sep=",", header=T)
airPne <- read.csv("D:/Pneumonia/Data/air91FromBaseline.csv", sep=",", header=T)

## Spearman's rho - NO2 and NOx correlation
rho <- cor(airPne$avno2_mod, airPne$avnox_mod, method = "spearman")
rh1 <- cor(airPne$no2, airPne$nox, method = "spearman")
rh2 <- cor(airPne$logno2, airPne$lognox, method = "spearman")

####FACTORS
airPne$gender <- as.factor(airPne$gender)
airPne$bmi_cat <- as.factor(airPne$bmi_cat)
airPne$education <- as.factor(airPne$education)
airPne$alcohol <- as.factor(airPne$alcohol)
#airPne$alc_cat <- as.factor(airPne$alc_cat)
airPne$smoking <- as.factor(airPne$smoking)
airPne$smo_cat <- as.factor(airPne$smo_cat)
#smo_duration => continuous
airPne$ets <- as.factor(airPne$ets)
airPne$occup_exp <- as.factor(airPne$occup_exp)
airPne$sport <- as.factor(airPne$sport)
airPne$sport_cat <- as.factor(airPne$sport_cat)
#fat & fruit => continuous 100g/day
#data_pne$komincome <-  sesincome$Komincome

##############################
#Cubic splines

fitTOT_no2 <- cph(Surv(ageriskstart, ageriskend, pne) ~ rcs(no2,4), airPne)
fitTOT_nox <- cph(Surv(ageriskstart, ageriskend, pne) ~ rcs(nox,4), airPne)

fitTOT_logno2 <- cph(Surv(ageriskstart, ageriskend, pne) ~ rcs(logno2,4), airPne)
fitTOT_lognox <- cph(Surv(ageriskstart, ageriskend, pne) ~ rcs(lognox,4), airPne)
```

```
par(mfrow = c(2, 2))
plot(fitTOT_no2, xlab="NO2")
plot(fitTOT_logno2, xlab="logNO2")
plot(fitTOT_nox, xlab="NOx")
plot(fitTOT_lognox, xlab="logNOx")
################################

################################
#NO2 and NOx categories - quartiles
airPne$logno2_cat <- cut(airPne$logno2, c(0, 3.8, 4, 4.3, 100), c(1:4))
airPne$lognox_cat <- cut(airPne$lognox, c(0, 4, 4.6, 5, 100), c(1:4))

################################
### COX proportional hazard ###
################################
library(survival)

#### FULL COHORT #####
attach(airPne)

#Cubic splines
library(Design)
d <- datadist(airPne)
options(datadist="d")

#### Exposure to NO2 ######
#Cox PH adjusted for age
m1TOTlogno2 <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2, airPne)
#cox.zph(m1TOTlogno2 , transform="log")

#Cox PH adjusted for age, smoking, ets and occupational exposure
m2TOTlogno2 <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                         smo_cat + occup_exp + ets, airPne)
#cox.zph(m2TOTlogno2 , transform="log")

#Cox PH fully adjusted
m3TOTlogno2 <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                         gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                         education + occup_exp + sport_cat + fruit100 + fat100, airPne)
```

```
#Test hp assumption
m3TOTzph <- cox.zph(m3TOTlogno2 , transform="log")

####Exposure to NOx ######
#Cox PH adjusted for age
m1TOTlognox <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox, airPne)
#cox.zph(m1TOTlognox , transform="log")

#Cox PH adjusted for age, smoking, ets and occupational exposure
m2TOTlognox <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                              smo_cat + occup_exp + ets, airPne)
#cox.zph(m2TOTlognox , transform="log")

#Cox PH fully adjusted
m3TOTlognox <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                              gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                              education + occup_exp + sport_cat + fruit100 + fat100, airPne)
m3xTOTzph <- cox.zph(m3TOTlognox , transform="log")
plot(m3xTOTzph[1])

####################
##Plots of the scaled Schoenfeld residuals against logNO2 exposure
plot(m3TOTzph[1])
par(mfrow=c(1,3))
#gender
plot(m3TOTzph[2])
#bmi
plot(m3TOTzph[3])
#Occupational exposure
plot(m3TOTzph[14])

par(mfrow=c(1,2))
plot(m3TOTzph[1])
plot(m3xTOTzph[1])
##### ###############
```

```
#NO2 categories
no2_cat <- cut(no2, c(0, 13.6, 17.3, 19.5, 1000), c(1:4))
airPne$logno2_cat <- cut(logno2, c(0, 3.8, 4, 4.3, 100), c(1:4))

m1TOTlogno2_cat <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2_cat, airPne)
m2TOTlogno2_cat <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2_cat +
                         smo_cat + ets + occup_exp, airPne)
m3TOTlogno2_cat <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2_cat +
                         gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                         education + occup_exp + sport_cat + fruit100 + fat100, airPne)



################################
# Traffic proxy status variables – 1year mean exposure at baseline
################################

#Read data
traffic = read.csv("D:/Pneumonia/Data/trafficproxy.csv", header = TRUE, sep = ",")

traffic$i50m10000 <- as.factor(traffic$i50m10000)
traffic$i100m10000 <- as.factor(traffic$i100m10000)
traffic$i50m5000 <- as.factor(traffic$i50m5000)
traffic$i100m5000 <- as.factor(traffic$i100m5000)

#Desc.Analysis
sum(traffic $i50m5000); sum(traffic $i50m10000)
sum(traffic $i100m5000); sum(traffic $i100m10000)

#Data for DCH cohort members
data_pneTR  <- traffic

####FACTORS
data_pneTR$gender <- as.factor(data_pneTR$gender)
data_pneTR$bmi_cat <- as.factor(data_pneTR$bmi_cat)
data_pneTR$education <- as.factor(data_pneTR$education)
data_pneTR$alcohol <- as.factor(data_pneTR$alcohol)
#airPne$alc_cat <- as.factor(data_pneTR$alc_cat)
data_pneTR$smoking <- as.factor(data_pneTR$smoking)
data_pneTR$smo_cat <- as.factor(data_pneTR$smo_cat)
data_pneTR$ets <- as.factor(data_pneTR$ets)
```

```
data_pneTR$occup_exp <- as.factor(data_pneTR$occup_exp)
data_pneTR$sport <- as.factor(data_pneTR$sport)
data_pneTR$sport_cat <- as.factor(data_pneTR$sport_cat)

####### COX MODEL #########
library(survival)

#Presence of major road- 10000 cars/day- within 50m radius at residential address
mod1t <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$i50m10000, data_pneTR)
mod1ta <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$i50m10000 +
                smo_cat + occup_exp + ets, data_pneTR)
mod1tb <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$i50m10000 +
                gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                education + occup_exp + sport_cat + fruit100 + fat100, data_pneTR)

#Presence of major road- 10000 cars/day- within 100m radius at residential address
mod2t <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$i100m10000, data_pneTR)
mod2ta <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$i100m10000 +
                smo_cat + occup_exp + ets, data_pneTR)
mod2tb <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$i100m10000 +
                gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                education + occup_exp + sport_cat + fruit100 + fat100, data_pneTR)

#Presence of major road- 5000 cars/day- within 50m radius at residential address
mod3t <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$i50m5000, data_pneTR)
mod3ta <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$i50m5000 +
                smo_cat + occup_exp + ets, data_pneTR)
mod3tb <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$i50m5000 +
                gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                education + occup_exp + sport_cat + fruit100 + fat100, data_pneTR)

#Presence of major road- 5000 cars/day- within 100m radius at residential address
mod4t <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$i100m5000, data_pneTR)
mod4ta <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$i100m5000 +
                smo_cat + occup_exp + ets, data_pneTR)
mod4tb <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$i100m5000 +
                gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                education + occup_exp + sport_cat + fruit100 + fat100, data_pneTR)
```

```
#Traffic loads
data_pneTR$strint2 <- data_pneTR$strint200m/1000000
data_pneTR$strint1 <- data_pneTR$strint100m/1000000

mod5t <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$strint1, data_pneTR)
mod5ta <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$strint1 +
                    smo_cat + occup_exp + ets, data_pneTR)
mod5tb <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$strint1 +
                    gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                    education + occup_exp + sport_cat + fruit100 + fat100, data_pneTR)

mod6t <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$strint2, data_pneTR)
mod6ta <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$strint2 +
                    smo_cat + occup_exp + ets, data_pneTR)
mod6tb <- coxph(Surv(ageriskstart, ageriskend, pne)~data_pneTR$strint2 +
                    gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                    education + occup_exp + sport_cat + fruit100 + fat100, data_pneTR)
##############################
```

D.4 Extended Cox regression models – recurrent pneumonias

```
#Read data – DCH with recurrent pneumonias
airPne <- read.csv("D:/Pneumonia/Data/recurrentData.csv", sep=",", header=T)

####FACTORS
airPne$gender <- as.factor(airPne$gender)
airPne$bmi_cat <- as.factor(airPne$bmi_cat)
airPne$education <- as.factor(airPne$education)
airPne$alcohol <- as.factor(airPne$alcohol)
#airPne$alc_cat <- as.factor(airPne$alc_cat)
airPne$smoking <- as.factor(airPne$smoking)
airPne$smo_cat <- as.factor(airPne$smo_cat)
#smo_duration => continuous
airPne$ets <- as.factor(airPne$ets)
airPne$occup_exp <- as.factor(airPne$occup_exp)
airPne$sport <- as.factor(airPne$sport)
airPne$sport_cat <- as.factor(airPne$sport_cat)
```

```
################################
### COX proportional hazard ###
################################
library(survival)

#NO2 categories
#no2_cat <- cut(no2, c(0, 13.6, 17.3, 19.5, 1000), c(1:4))
airPne$logno2_cat <- cut(airPne$logno2, c(0, 3.8, 4, 4.3, 100), c(1:4))
airPne$lognox_cat <- cut(airPne$lognox, c(0, 4, 4.6, 5, 100), c(1:4))
attach(airPne)

#Cubic splines
library(Design)
d <- datadist(airPne)
options(datadist="d")

#### NO2 ######
m3logno2 <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                          gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                          education + occup_exp + sport_cat + fruit100 + fat100,
                          data= subset(airPne, count <=1) )

m3logno2cat <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2_cat +
                          gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                          education + occup_exp + sport_cat + fruit100 + fat100,
                          data= subset(airPne, pne_noci==1 & count <=1 ) )

#PH assumption check
m3firstzph <- cox.zph(m3logno2, transform="log"); m3xfirstzph <- cox.zph(m3lognox, transform="log")

##Plots of the scaled Schoenfeld residuals against log-transformed age as underlying time-scale
#logno2 and lognox exposure
par(mfrow=c(1,2))
plot(m3firstzph[1])
plot(m3xfirstzph[1])

#gender
plot(m3firstzph[2])
#Occupational exposure
plot(m3firstzph[14])
```

```
################################
#Recurrent models
################################

#Intensity-based model
#Cox PH adjusted for age
m1logno2IB <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2, airPne)
#Cox PH adjusted for age smoking, ets and occup.exposure
m2logno2IB <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                            smo_cat + ets + occup_exp, airPne)
#Cox PH fully adjusted
m3logno2IB <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                            gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                            education + occup_exp + sport_cat + fruit100 + fat100, airPne)
cox.zph(m1logno2IB, transform="log")
cox.zph(m2logno2IB, transform="log")
cox.zph(m3logno2IB, transform="log")

#Andersen - Gill (variance-corrected) model
#Cox PH adjusted for age
m1logno2AG <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 + cluster(knr), airPne)

#Cox PH adjusted for age smoking, ets and occup.exposure
m2logno2AG <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                            smo_cat + ets + occup_exp + cluster(knr), airPne)

#Cox PH fully adjusted
m3logno2AG <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                            gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                            education + occup_exp + sport_cat + fruit100 + fat100 + cluster(knr), airPne)
cox.zph(m1logno2AG, transform="log")
cox.zph(m2logno2AG, transform="log")
cox.zph(m3logno2AG, transform="log")

#Conditional AG
m1logno2AGc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 + cluster(knr) + strata(enum),
                             airPne)
m2logno2AGc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                            smo_cat + ets + occup_exp + cluster(knr) + strata(enum), airPne)

m3logno2AGc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
```

```
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100 +
                        cluster(knr) + strata(enum), airPne)
cox.zph(m1logno2AGc, transform="log")
cox.zph(m2logno2AGc, transform="log")
cox.zph(m3logno2AGc, transform="log")

#Frailty (random effect) model
m1logno2f <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 + frailty(knr), airPne)

m2logno2f <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                        smo_cat + occup_exp + ets + frailty(knr), airPne)

m3logno2f <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100 +
                        frailty(knr), airPne)

cox.zph(m1logno2f , transform="log")
cox.zph(m2logno2f , transform="log")
cox.zph(m3logno2f , transform="log")

#Conditional Frailty model
m1logno2f <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 + frailty(knr)+strata(enum), airPne)



m2logno2f <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                        smo_cat + occup_exp + ets + frailty(knr) )+strata(enum), airPne)

m3logno2f <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100 +
                        frailty(knr) )+strata(enum), airPne)

cox.zph(m1logno2f , transform="log")
cox.zph(m2logno2f , transform="log")
cox.zph(m3logno2f , transform="log")

##################################
#### NOx ######
```

```
#Intensity-based model
m1lognoxIB <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox, airPne)
m2lognoxIB <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        smo_cat + ets + occup_exp, airPne)
m3lognoxIB <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100, airPne)

#Andersen - Gill (variance-corrected) model
m1lognoxAG <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox + cluster(knr), airPne)
m2lognoxAG <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        smo_cat + ets + occup_exp + cluster(knr), airPne)
m3lognoxAG <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100 + cluster(knr), airPne)

#Conditional AG
m1lognoxAGc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox + cluster(knr) + strata(enum),
                        airPne)
m2lognoxAGc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        smo_cat + ets + occup_exp + cluster(knr) + strata(enum), airPne)
m3lognoxAGc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100 +
                        cluster(knr) + strata(enum), airPne)

#Frailty model
m1lognoxf <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox + frailty(knr), airPne)
m2lognoxf <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        smo_cat + ets + occup_exp + frailty(knr), airPne)
m3lognoxf <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100 + frailty(knr), airPne)

#Conditional Frailty
m1lognoxfc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox + frailty(knr) + strata(enum), airPne)

m2lognoxfc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        smo_cat + ets + occup_exp + frailty(knr) + strata(enum), airPne)
m3lognoxfc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
```

```
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100 +
                        frailty(knr) + strata(enum), airPne)

##### ####################
### Survival(KM) and Cumulative hazard (NA)####

# Null Cox model
tsurv <- survfit(Surv(ageriskstart, ageriskend, pne) ~ 1, data=subset(airPne, count <=1))
par(mfrow=c(1,2))
plot(tsurv, xlim=c(50,80), xlab="Age", ylab="Survival", conf.int=T, mark.time=F)
plot(tsurv, xlim=c(50,80), xlab="Age", ylab="Cumulative Hazard", fun="cumhaz", conf.int=T,
              mark.time=F)

######
#Cumulative hazard function plots by use of the Nelson-Aalen estimator
fitGender <- survfit ( Surv(ageriskstart, ageriskend, pne) ~ gender, data =airPne)
fitSmo <- survfit ( Surv(ageriskstart, ageriskend, pne) ~ smo_cat, data =airPne)
fitEts <- survfit ( Surv(ageriskstart, ageriskend, pne) ~ ets, data =airPne)
fitAlco <- survfit ( Surv(ageriskstart, ageriskend, pne) ~ alc_cat, data =airPne)
fitOccup <- survfit ( Surv(ageriskstart, ageriskend, pne) ~ occup_exp, data =airPne)
fitEduc <- survfit ( Surv(ageriskstart, ageriskend, pne) ~ education, data =airPne)
fitSport <- survfit ( Surv(ageriskstart, ageriskend, pne) ~ sport_cat, data =airPne)
fitBMI <- survfit ( Surv(ageriskstart, ageriskend, pne) ~ bmi_cat, data =airPne)

fitEv <- survfit ( Surv(ageriskstart, ageriskend, pne) ~ events, data=subset(airPne, events!=0))

#########
#Indiv. with history of Pneumonia - #485

m1clogno2IB <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2, data=subset(airPne, pne_noci==1))
m2clogno2IB <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                        smo_cat + ets + occup_exp, data=subset(airPne, pne_noci==1))
m3clogno2IB <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100, data=subset(airPne,
                        pne_noci==1))

#Andersen - Gill (variance-corrected) model
m1clogno2AG <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 + cluster(knr), data=subset(airPne,
                        pne_noci==1))
```

```
m2clogno2AG <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                        smo_cat + ets + occup_exp + cluster(knr), data=subset(airPne, pne_noci==1))
m3clogno2AG <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100 + cluster(knr),
                        data=subset(airPne, pne_noci==1))

#Conditional  AG
m1clogno2AGc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 + cluster(knr) + strata(enum),
                        data=subset(airPne, pne_noci==1))
m2clogno2AGc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                        smo_cat + ets + occup_exp + cluster(knr) + strata(enum), data=subset(airPne,
                        pne_noci==1))
m3clogno2AGc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100 +
                        cluster(knr) + strata(enum), data=subset(airPne, pne_noci==1))

#Frailty model
m1clogno2f <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 + frailty(knr), data=subset(airPne,
                        pne_noci==1))
m2clogno2f <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                        smo_cat + ets + occup_exp + frailty(knr), data=subset(airPne, pne_noci==1))
m3clogno2f <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100 + frailty(knr),
                        data=subset(airPne, pne_noci==1))

#Conditional Frailty
m1clogno2fc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 + frailty(knr) + strata(enum),
                        data=subset(airPne, pne_noci==1))
m2clogno2fc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                        smo_cat + ets + occup_exp + frailty(knr) + strata(enum), data=subset(airPne,
                        pne_noci==1))
```

```
m3clogno2fc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ logno2 +
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100 +
                        frailty(knr) + strata(enum), data=subset(airPne, pne_noci==1))

#############
### log NOx
m1clognoxIB <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox, data=subset(airPne, pne_noci==1))
m2clognoxIB <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        smo_cat + ets + occup_exp, data=subset(airPne, pne_noci==1))
m3clognoxIB <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100, data=subset(airPne,
                        pne_noci==1))

#Andersen - Gill (variance-corrected) model
m1clognoxAG <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox + cluster(knr), data=subset(airPne,
pne_noci==1))
m2clognoxAG <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        smo_cat + ets + occup_exp + cluster(knr), data=subset(airPne, pne_noci==1))
m3clognoxAG <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100 + cluster(knr),
                        data=subset(airPne, pne_noci==1))

#Conditional AG
m1clognoxAGc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox + cluster(knr) + strata(enum),
                        data=subset(airPne, pne_noci==1))


m2clognoxAGc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        smo_cat + ets + occup_exp + cluster(knr) + strata(enum), data=subset(airPne,
                        pne_noci==1))
m3clognoxAGc <- coxph(Surv(ageriskstart, ageriskend, pne) ~ lognox +
                        gender + rcs(bmi,3) + smo_cat + ets + alcohol + alco_int +
                        education + occup_exp + sport_cat + fruit100 + fat100 +
                        cluster(knr) + strata(enum), data=subset(airPne, pne_noci==1))
#####
```

Bibliography

[1] Leon Gordis, *Epidemiology*, Philadelphia, Pennsylvania: Elsevier Inc. (USA), 2004.

[2] www.who.int, "The World Health Organization."

[3] W.C. Willett, "Diet and Cancer," *The Oncologist*, 2000.

[4] R.C. Duilio Divisi, Sergio Di Tommaso, Salvatore Salvemini, Margherita Garramone, "Diet and Cancer," *ACTA BIOMED*, 2006.

[5] B.P. Salvi SS, "Chronic obstuctive pulmonary disease in non-smokers," *Lancet*, 2009.

[6] von M.E. Eder W, Ege MJ, "The asthma epidemic," *N Engl J Med*, 2006.

[7] J. Xu, K.D. Kochanek, and S.L. Murphy, "National Vital Statistics Reports Deaths : Final Data for 2007," *Statistics*, vol. 58, 2010.

[8] H.R. Fry AM, Shay DK, "Trends in hospitalizations for pneumonia among persons aged 65 years or older in the United States," *JAMA*, 2005.

[9] G.R. Trotter CL, Stuart JM, "Increasing hospital admissions for pneumonia, England," *Emerg Infect Dis*, 2008.

[10] S. Explained, "Causes of death statistics -EU," *Main*, 2011.

[11] Z.J. Andersen, K. Bønnelykke, M. Hvidberg, S.S. Jensen, M. Ketzel, S. Loft, M. Sørensen, A. Tjønneland, K. Overvad, and O. Raaschou-nielsen, "Long-term exposure to air pollution and asthma hospitalizations in elderly adults: a cohort study," *Thorax - in press*, 2011.

[12] Z. J. Andersen, A. Tjønneland, K. Overvad, and O. Raaschou-nielsen, "Chronic Obstructive Pulmonary Disease and Long-Term Exposure to Traffic-related Air Pollution A Cohort Study," *American Journal of Respiratory and Critical Care Medicine*, 2011.

[13] www.lungeforening.dk, "Danmarks Lungeforening."

[14] N.B. Van Dieren S, Beulens JW, van der Schouw YT, Grobbee DE, "The global burden of diabetes and its complications: an emerging pandemic," *Eur J Cardiovasc Prev Rehabil*, 2010.

[15] Steenland K; Stavitz DA, “Topics in Environmental Epidemiology,” *New York: Oxford University Press*, 1997.

[16] Z.J. Andersen, “Short-Term Health Effects of Air Pollution in Copenhagen,” University of Copenhagen, Faculty of Health Sciences, 2007.

[17] Brook RD, Pope CA, “Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association,” *Circulation*, vol. 3rd, 2010.

[18] B.M. Dominici F, Peng RD, “Fine Particulate Air Pollution and Hospital Admission for cardiovascular and respiratory diseases,” *JAMA*, 2006.

[19] K.N. Nawrot TS, Perez L, “Public Health importance of triggers of myocardial infarction: a comparative risk assessment,” *Lancet*, 2011.

[20] Z.J. Andersen, T.S. Olsen, K.K. Andersen, S. Loft, M. Ketzel, and O. Raaschou-Nielsen, “Association between short-term exposure to ultrafine particles and hospital admissions for stroke in Copenhagen, Denmark.,” *European heart journal*, vol. 31, Aug. 2010, pp. 2034-40.

[21] H. JC and van E. S, “Pulmonary and systemic responce to atmospheric pollution,” *Respirology*, 2009.

[22] Ms. Mark Loeb MD, B.N. MSc, S.D.W. PhD, R.H. PhD, S.C.C. PhD, D.L. PhD, P.K. PhD, A.E.S. MD, L.N. MD, and T.J.M. MD, “Environmental Risk Factors for Community-Acquired Pneumonia Hospitalization in Older Adults,” *Journal of the American Geriatrics Society*, 2009.

[23] M.W. Frampton, J. Boscia, N.J. Roberts, M. Azadniv, A. Torres, C. Cox, P.E. Morrow, J. Nichols, D. Chalupa, L.M. Frasier, F.R. Gibb, D.M. Speers, Y. Tsai, M. J, M.J. Utell, M. W, N.J. Rob-, and M. Frasier, “Nitrogen dioxide exposure : effects on airway and blood cells Nitrogen dioxide exposure : effects on airway and blood cells,” *Cell*, 2011.

[24] B. Neupane, M. Jerrett, R.T. Burnett, T. Marrie, A. Arain, and M. Loeb, “Long-Term Exposure to Ambient Air Pollution and Risk of Hospitalization with Community-acquired Pneumonia in Older Adults,” *Critical Care Medicine*, 2009.

[25] O.K. Tjønneland A, Olsen A, Boll K, Stripp C, Christensen J, Engholm G, “Study design, exposure variables, and socioeconomic determinants of participation in Diet, Cancer and Health: a population-based prospective cohort study of 57,053 men and women in Denmark.,” *Scandinavian Journal of Public Health*, 2007.

[26] “www.en.wikipedia .org.”

[27] V. Degroot, H. Beckerman, G. Lankhorst, and L. Bouter, “How to measure comorbidity, a critical review of available methods,” *Journal of Clinical Epidemiology*, vol. 56, Mar. 2003, pp. 221-229.

[28] U. Stab, J. Dahl, C. Østergaard, K. Oren, N. Frimodt-møller, and H. Carl, “Recurrent bacteraemia : A 10-year regional population-based study of clinical and microbiological risk factors,” *Infection*, 2010.

[29] J.B. Kornum, R.W. Thomsen, and A. Riis, “Diabetes , Glycemic Control and Risk of Hospitalization with Pneumonia : A Population-based Case-control Study,” *Diabetes Care*, 2008, pp. 1-11.

[30] B. Carstensen, J.K. Kristensen, P. Ottosen, and K. Borch-Johnsen, “The Danish National Diabetes Register: trends in incidence, prevalence and mortality.,” *Diabetologia*, vol. 51, Dec. 2008, pp. 2187-96.

[31] W.G.M. Chen, Tze-Ming MD; Gokhale, Janaki MD; Shofer, Scott MD, PhD; Kuschner, “Outdoor Air Pollution: Nitrogen Dioxide, Sulfur Dioxide, and Carbon Monoxide Health Effects,” *American Journal of the Medical Science*, 2007.

[32] M. Ketzel, P. Wåhlin, a Kristensson, E. Swietlicki, R. Berkowicz, O.J. Nielsen, and F. Palmgren, “Particle size distribution and particle mass measurements at urban,near-city and rural level in the Copenhagen area and Southern Sweden,” *Atmospheric Chemistry and Physics*, vol. 4, Feb. 2004, pp. 281-292.

[33] C. van Ertbruggen, C. Hirsch, and M. Paiva, “Anatomically based three-dimensional model of airways to simulate flow and particle transport using computational fluid dynamics.,” *Journal of applied physiology (Bethesda, Md. : 1985)*, vol. 98, Mar. 2005, pp. 970-80.

[34] C.A. Donalson K, Stone V, “Ultrafine particles,” *Occup Endiron Med*, 2001.

[35] H.O. Jensen SS, Berkowicz R, Hansen SH, “A Danish decision-support GIS tool for management of urban air quality and human exposures,” *Transportation Research Part D: Transport and Environment*, 2001.

[36] A.U. SS, Jensen, Roskilde, National Environmental Research Institute, “Background concentrations for use in the Operational Street Pollution Model (OSPM),” *NERI Technical Reports*, 1998.

[37] R.-N.O. Berkowicz R, Ketzel M, Jensen SS, Hvidberg M, “Evaluation and application of OSPM for traffic pollution assessment for large number of street locations.,” *Environ Model Software*, 2008.

[38] B. R., “OSPM - A parameterised street pollution model,” *Environ Monit Assess*, 2000.

[39] A.U. Jensen SS, Hvidberg M, Pedersen J, et al. Roskilde, National Environmental Research Institute, "GIS-based national street and traffic data base 1960-2005," *NERI Technical Reports*, 2009.

[40] K.M. Berkowitz R, Winther M, "Traffic pollution modeling and emission data," *Environ Model Software*, 2006.

[41] R.-N.O. Ketzel M, Berkowicz R, Hvidberg M, Jensen SS, "Evaluation of AirGIS - a GIS-based air pollution and human exposure modelling system," *International Journal for Environment and Pollution*, 2011.

[42] Kunzli N, Liu LJ, "Swiss Cohort Study on Air Pollution and Lung Diseases in Adults. Traffic-related air pollution correlates with adult-onset asthma among never-smokers," *Thorax*, 2009.

[43] D. G. Kleinbaum and M. Klein, *Survival Analysis A Self-Learning Text*, Springer - Statistics for Biology and Health, 2005.

[44] H.G. Odd Aalen, Ornulf Borgdan, *Surival and Event History Analysis A process Point of View*, Springer - Statistics for Biology and Health, 2008.

[45] A. C. M. Thiebaut and J. Benichou, "Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study," *Statistics in medicine*, 2004.

[46] Therneau, *Modeling Survival Data Extending the Cox Model*, Springer - Statistics for Biology and Health, 2000.

[47] T. Martinussen and T. H. Scheike, *Dynamic Regression Models for Survival Data*, Springer - Statistics for Biology and Health, 2006.

[48] M. Miloslavsky, M. J. van der Laan, S. Keles, S. Butler, and B. University of California, "Recurrent Events Analysis in the Presence of Time Dependent Covariates and Dependent Censoring Recurrent Events Analysis in the Presence of Time Dependent Covariates and Dependent Censoring," *Biostatistics*, 2002.

[49] J.M. Box-Steffensmeier and S. De Boef, "Repeated events survival models: the conditional frailty model.," *Statistics in medicine*, vol. 25, Oct. 2006, pp. 3518-33.

[50] P.K. Andersen, O. Borgan, R.D. Gill, and N. Keiding, *Statistical Models Based on Counting Processes*, Springer, 1993.

[51] R.W. Thomsen, A. Riis, J. Jacobsen, S. Christensen, and C.J. Mcdonald, "Rising incidence and persistently high mortality of hospitalized pneumonia : a 10-year population-based study in Denmark," *Journal of Internal Medicine*, 2006, pp. 410-417.

[52] S.R. Cole and M.-G. Hall, “Time Scale and Adjusted Survival Curves for Marginal Structural Cox Models,” *American Journal of Epidemiology Advance Accesse*, 2010.

[53] H.U. Wegmann M, Fehrenbach A, Heimann S, Fehrenbach H, Renz H, Garn H, “NO2-induced airway inflammation is associated with progressive airflow limitation and development of emphysema-like lesions in C57bl/6 mice,” *Exp Toxicol Pathol*, 2005.

[54] S.L. David W. Hosmer, *Applied survival analysis - regression modeling of time to event*, JohnWiley and Sons, 1st edition, 1999.

[55] T.M. Thomas H. Scheike, *Dynamic Regression Models for Survival Data*, Springer, 2006.

[56] L.S. Raaschou-Nielsen O, Bak H, Sørensen M, Jensen SS, Ketzel M, Hvidberg M, Schnohr P, Tjønneland A, Overvad K, “Air pollution from traffic and risk for lung cancer in three Danish cohorts,” *Cancer Epidemiol Biomarkers*, 2010.

[57] Z. J.Andersen, L. C. Kristiansen, K. K. Andersen, T. S. Olsen, A. Tjønneland, K. Overvad, and O. Raaschou-nielsen, “Stroke and long-term exposure to air pollution: a cohort study.”

[58] K.J. Jensen SS, Larson T, Deepti KC, “Modeling traffic air pollution in street canyons in New York City for intra-urban exposure assessment in the US multi-ethnic study of atherosclerosis and air pollution,” *Atmos Environ*, 2009.