ANALYSIS OF OCCUPATIONAL COHORT DATA USING EXPOSURE AS A CONTINUOUS TIME-DEPENDENT VARIABLE

by

Maria Lorenzi
B.Sc., Simon Fraser University, 2000

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FO THE DEGREE OF
MASTER OF SCIENCE
in the Department
of
Statistics and Actuarial Science

SIMON FRASER UNIVERSITY
Summer 2005

APPROVAL

Name: Maria Lorenzi

Degree: Master of Science

Title of thesis: Analysis of Occupational Cohort Data using Exposure as a Continuous Time-Dependent Variable

Examining Committee: Dr. Michael Stephens, Professor Emeritus
Simon Fraser University
Chair

Dr. Richard Lockhart, Professor
Simon Fraser University
Senior Supervisor

Dr. John Spinelli, Adjunct Professor
Simon Fraser University
Co-Supervisor

Dr. Nhu Le, Senior Biostatistician, British Columbia Cancer Agency
External Examiner

Date Approved: August 03, 2005

SIMON FRASER UNIVERSITY

PARTIAL COPYRIGHT LICENCE

Abstract

In occupational cohort studies, a group of workers is followed over time, and disease and work history information are collected for each individual in order to determine whether exposure to a particular substance is linked to differences in mortality or disease incidence rates. These studies are typically analysed by treating cumulative exposure as a categorical variable and then comparing disease or mortality rates between different exposure groups. A main shortfall of such analyses is a heavy dependence on the choice of these exposure categories, as certain choices may mask or exaggerate important features of the dose-response curve. In this project, an extension to the Cox proportional hazards model is used to treat cumulative exposure as a continuous variable and model the dose-response curve nonparametrically for a study of aluminium smelter workers conducted by the British Columbia Cancer Agency and compare the results to the categorical analyses.

Acknowledgments

First of all, I'd like to thank my supervisor, Dr. John Spinelli for all his patience, advice, and the wonderful opportunity to work at the BC Cancer Agency for the better part of the last two years. Special thanks also go to Dr. Nhu Le for asking me so many questions and making me go find the answers.

To the staff and all my former students at New Westminster Secondary School, I am extremely grateful for the three years that I spent with you, and particularly for your understanding when I decided that teaching and going to school at the same time was a really *great* idea. My thoughts are with you often.

These past three years would not have been nearly as enjoyable without the company of all the rest of the grad students in our department. Our retreats, dinners, and bowling nights have always been fun. And I rather enjoyed our communal 450/801 stress-out sessions. In particular, thanks to Karey Shumansky for stressing with me at BC Cancer for the last year and to Amy Summers, thanks for getting me to the gym, for lending me SATC and for our numerous movie/shopping trips.

To my dear friends Tricia, Leslie, and Pam, thanks for your understanding all those times that I couldn't come out to play. Even more thanks for the good times that we did have when I managed to pry myself away from my computer. Knowing that you ladies were in my corner has been a great comfort.

Very special thanks go to Jason Nielsen, who can make me laugh even when I'm sad and has always been there for me with a hug and a cold beer when needed. Thanks for insisting on our Jeopardy dates even when I was stressed, and for always enthusiastically eating whatever cooking experiment I put in front of you! I cannot thank you enough for your support and especially your patience.

Finally, my deepest gratitude goes to my mother Beate and my sister Lucia. Without

their encouragement, humour and neverending support over the years, this degree would never have been possible. They are both unbelievable inspirations to me. Mum, all the work and worry that you invested in me paid off! Thank you for having such high expectations of me, and for being behind me no matter what.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

There are many known and suspected carcinogens present in today's world. Asbestos, cigarette smoke, and radiation are three examples of agents that have definitively been shown to be linked to increased mortality and cancer occurrence [3]. A large number of carcinogenic relationships have been proven through studies incorporating long-term follow-up on large numbers of individuals; these studies are called *cohort* studies. By starting off with a sizable number of healthy individuals and tracking their exposure and disease histories over a prolonged period of time, cohort studies allow one to measure the effect of differential exposure on mortality and disease rates. It is possible to compare overall mortality or incidence of the cohort to some external population (usually the general population of the region/country in which the cohort is located), or compare different subgroups within the group to identify groups with higher rates of disease. These subgroups can be based on levels of exposure, age, job description, or many other classifications.

Typically, the analysis of cohort studies involves grouping cumulative exposure into a fairly small number of categories and then using either non-parametric methods or Poisson regression to compare disease or mortality rates between the different groups. A possible downfall of these methods is that certain choices of categories may mask or overemphasize certain features of the dose-response relationship. Recommendations have been made to select exposure categories *a priori* in various ways in order to reduce bias, but the potential for missing important aspects still remains [8]. In this project, the counting process form of the Cox proportional hazards model will be used to model cumulative exposure as a continuous variable in the hope of removing the problems associated with category selection.

In Chapter 2, cohort studies will be described, along with methods to analyze cohort

data. Section 2.1 will include an overview of the terminology and main design and implementation features of cohort studies. Sections 2.2 and 2.3 will describe the two common methods used to analyse cohort data: non-parametric analysis and Poisson regression. Section 2.4 begins with an overview of Cox proportional hazards models, then describes how the counting-process formulation of the Cox PH model can be used to analyse exposure as a continuous variable, as opposed to a categorical variable as used in the two previously described methods. The chapter concludes with a discussion about issues specific to occupational cohort studies and how they affect the implementation of the three analysis methods described.

Chapter 3 describes the application of the three analysis methods to data collected from a study of workers at an alumimum smelter in British Columbia. Sections 3.1 and 3.2 give an overview of the goals of the study and the data used. Section 3.3 briefly summarizes some results for the overall cohort analysis. In Section 3.4, the data is analysed by treating exposure as a categorical variable and using non-parametric methods and Poisson regression to quantify the relationship between exposure and incidence of both bladder cancer and non-Hodgkins lymphoma. Section 3.5 covers the Cox proportional hazards analysis, including the extensive data manipulation needed to obtain the proper data format needed in order to use the counting process formulation. The final section in Chapter 3 is a discussion comparing the results of the three analysis methods.

Chapter 2

Methodology

2.1 Overview of Cohort Studies

2.1.1 Description

Cohort studies hold a fundamental place in epidemiological research. By following a defined group of individuals over a long period of time, it is possible to study the association between a given exposure and one or more disease outcomes and possibly identify a causal link. Many of the studies that definitely established the carcinogenicity of various agents were cohort studies [3].

First, a cohort is defined by identifying a group of individuals. Often these individuals share some kind of experience or condition, such as employment or residence in the same location/region, but many other criteria can be used to assemble a suitable group of individuals. By following this group over an extended period of time and recording exposure and disease information, rates of disease incidence can be obtained, and relationships between exposure and incidence can be quantified by comparing the rates of disease outcome in exposed and unexposed groups. Incidence rates can be compared between subgroups of the same cohort (for example low exposure vs high exposure groups), or between the cohort and some reference population.

There are two types of cohort studies, which differ in how follow-up is conducted, but not in analysis: In *prospective* cohort studies, the cohort is put together in the present, and then individuals are followed into the future. In *retrospective* cohort studies, historical records are used to assemble a cohort at some well-defined time

in the past, and then exposure information and disease occurence information are collected up to the present time. With prospective studies, results will not be available for many years, but researchers have a lot of control over the quality of follow-up and the type of information collected. Retrospective studies offer potentially more immediate results; however, one must make do with the information that is already available, which may not always be the exact type of data which the researcher would like.

The basic goal of a cohort study is to compare the risk of death or cancer incidence between groups of individuals. More specifically, one can compare the disease risk of subgroups within the cohort (say unexposed versus exposed individuals) or between the cohort and some external population (usually the general population in the region/country where the cohort is located). This is accomplished by comparing the rates of the outcome of interest (death or cancer incidence) between the groups.

Naturally, rates of death and cancer incidence vary widely with age and also with calendar period, so if two comparison groups differ in composition with respect to either or both of these two variables, their raw disease rates will most likely also be different, regardless of whether the exposure experience differs between the groups or not. In a cohort study, we are not directly interested in the effects of age or calendar time (although we may want to study their interaction with exposure), so measures must be taken to control for these two variables. This is done by stratification and/or inclusion of age and calendar year as covariates in a regression model

Further analysis of the patterns of disease risk within a cohort can be done by dose-response modelling: by categorizing the degree of exposure (even in a crude manner) one can ascertain which groups are at highest risk and assess the shape of the dose-response curve. For example, one can see whether there is increasing risk with increasing levels of exposure, and quantify the association. It is also of interest to assess the temporal relationship between exposure and disease. Most cancers do not manifest themselves until years after a person first experiences a given exposure, therefore it is of interest to quantify the delay of onset. Also, increased duration of exposure (as opposed to the mere presence) is often associated with increased risk of disease, and should be studied. If may be of further interest to assess how risk changes after an individual is no longer exposed to a given agent.

2.1.2 Design and Implementation of Cohort Studies

Although the specific circumstances and goals of each cohort study will be different, there are a number of fundamental issues that must be clearly resolved before the study can begin [3]:

- *Inclusion rules for the study:* From the very name of the type of study, it is obvious that a clear definition of the study cohort is of vital importance. Often, cohorts are defined by either geographic location (eg. all residents of a certain city at a particular time), or by occupation (eg. all individuals who worked at a certain factory, held a certain profession between two predetermined dates, or were employed at one specific date).

- *Entry and exit dates:* Since each individual in the cohort contributes person-years of observation time to the study, it is crucial to know exactly when observation time begins and ends. The date of entry into the cohort is the first date that the individual is considered to be at risk. This is not necessarily the date of first exposure, as many cohort studies are concerned only with individuals who have accumulated a specified amount of time past their first exposure. The date of exit for a particular individual is either the date of disease occurence (or death) or their his or her date of follow-up where disease/vital status was known.

- *Follow-up:* Since the whole purpose of a cohort study is to follow individuals over time and record exposure and disease information, a high quality of follow-up information is essential. One would like to capture as much observation time per person as possible, so mechanisms must be in place to reliably obtain vital status and/or disease information after individuals have ceased to be exposed.

- *Information on disease:* There must be unambiguous coding of disease for all incidences. A common coding for this is the International Classification of Diseases (ICD).

- *Exposure information:* The type and detail of exposure information must adequately support the aims of the study. Begin and end dates of exposure must

be unambiguous, and, as much as possible, exposure information must be quantifiable and available at the individual level.

- *Information on other exposures:* Quite often an exposure other than the ones being explicitly studied can be a confounding factor in the analysis, and information on any known or suspected confounders should be collected if there are resources available to do so. This will help increase confidence that any dose-response found after analysis is indeed directly related to the exposure in question instead of being an artifact of some other unmeasured behaviour or exposure. A typical confounder is smoking.

- *The power of the study:* Since a cohort study requires a large commitment of resources, it is vital to know at the outset what kind of results one could expect to find. If the study is too small to actually detect any potential excess risk, there is little point in undertaking it.

2.2 Dose-Response Modelling in an Occupational Cohort Setting

2.2.1 Importance of Dose-Response Modelling

Many early retrospective cohort studies focused only on discovering the presence or absence of a link between exposure and mortality or cancer incidence. As new methodology was developed, the focus shifted towards determining how different *patterns* of exposure led to changes in cancer risk. This led to the consideration of dose-response curves, which describe the change in relative risk with increasing exposure.

Ideally, a simple form would be found for the dose-response curve, since it is believed that the true underlying relationship between exposure and response is indeed simple in nature [8]. Of course, the data at hand will not fit any model perfectly, so there will inevitably be a tradeoff between model simplicity and goodness-of-fit, no matter which model is eventually chosen.

2.2.2 Quantifying and Measuring Exposure

In order to study how relative risk changes with increasing exposure, one must first decide how to quantify and measure eposure.

"Exposure" here can refer to a specific chemical, environmental, or lifestyle agent, or it could refer to a quantity such as "time employed" which could be regarded as a surrogate for some (possibly unknown) physical exposure. In either case, measurements of exposure must be obtained for each individual throughout the study so that the association between exposure and disease can be quantified.

Ultimately, the quality of an occupational cohort study relies heavily on the quality of the exposure measurement techniques, so great care must be taken to obtain the most accurate measurements possible given the situation. In terms of collecting reliable exposure information, a prospective cohort study is clearly superior because one can decide at the outset what type of information to collect and how best to go about measuring and recording it, within the resources available to the researchers. Depending on the nature of the exposure agent, one can, for instance, attach individual measurement devices to individuals, or measure exposures in specific work areas for periods of time. With a retrospective cohort study, on the other hand, one is limited by the type of exposure information that has already been collected. A typical situation is to have work history records for each individual that document the job title, department, and physical area of workplace where the job was performed, as well as start and stop dates for each job. A retrospective exposure assessment is then needed to assign mean daily exposure levels to the different job title/area/time classifications. There are many methods available for performing such an exposure assessment, and advancements to these techniques have been made in recent years [4]. In particular, statistical modelling and extrapolation are being used instead of expert-based assessment where direct measurements are not available.

There are various ways to quantify and categorize exposure for use in dose-response modelling. The most commonly used measures are time since first exposure, duration of exposure, and cumulative exposure. The choice between them can depend on the type of data available, the aims of the study, and the type of dose-response that one hopes to detect. To analyse time since first exposure, one simply needs a single date

for each individual corresponding to the first time that they were exposed to the given agent (in addition to the data needed for mortality or incidence analysis). For duration of exposure, one needs to know whether the individual was exposed or not for every time interval considered in the study. For cumulative exposure, a quantitative level of exposure for each individual must be ascertained for each section of time. One advantage of looking at time since first exposure or duration is that there is no ambiguity as to how to measure time: as long as the dates in the work history records are correct, it is easy to determine how long an individual has been exposed to a given agent, and how long ago the first exposure took place. To analyze cumulative exposure, more care is needed, especially in retrospective cohort studies where one cannot go back in time to collect exposure measurements. One must decide on "low" or "high" levels of exposure, how exposure categories are going to be chosen, and what numeric value is going to represent each of these categories. Also, great care must be taken to assign mean exposure levels for jobs or areas where no direct measurements exist in order to avoid potential bias.

2.2.3 Latency

Quite often a lag time is used, which entails calculating the cumulative exposure at a point which is a predetermined amount of time before the person-year in question. Incorporating a lag time can help to minimize bias that arises from the 'healthy worker effect'. This effect occurs because a worker who is accumulating exposure (and crossing over category boundaries) is necessarily still employed and therefore at a lesser apparent risk of death. The lag can also help account for a possible latency period, which is the delay between exposure and death from the disease (or incidence of the disease). Often, various lag times are analyzed to determine the most reasonable latency period for the cause of death or cancer site in question.

2.2.4 Complexity of Occupational Cohort Data

In contrast to most counting process survival analyses, data from an occupational cohort study is often very complicated. Instead of a single binary variable that changes once for an individual (eg. transplant/no transplant), or a repeated measurement at

regular intervals (eg. blood pressure at yearly checkups), the quantity of interest is usually cumulative exposure, which changes continuously but is not measured continually on an individual basis. In order to quantify exposure, a job exposure matrix is used. The job exposure matrix gives a mean daily exposure level for each job/time period combination, and covers every job held during the study period. There are many issues that arise in the creation of a job exposure matrix; especially in a retrospective cohort study exposure levels from the past must be assessed as accurately as possible.

For each individual, the work history records can be matched to the job exposure matrix to obtain mean daily exposure levels for each job that was held. Multiplying the duration of the job by the exposure level gives the cumulative exposure received during that job, and these cumulative exposures can be summed to get the cumulative exposure of an individual at any point in time.

A practical difficulty with occupational cohort datasets is that there are multiple records for each individual corresponding to each job, and while cumulative exposure is readily available for endpoints of each record, a dataset in this form does not accurately record cumulative exposure at the failure times in the cohort.

2.3 Non-Parametric Analysis of Cohort Studies

The most natural study question for a cohort is whether or not disease rates differ between people in groups that are differently exposed. In order to assess this, one must first take into account changes in disease rates that are caused by differences in ages and calendar time. It seems obvious that age would have a significant effect on an individual's instantaneous likelihood of developing cancer or dying. Calendar time is also important, since disease rates in the general population often fluctuate over time, and changes in this background rate must be accounted for.

The simplest way to control for age and calendar time is to stratify by age and calendar period. After the desired age intervals and calendar periods have been chosen (typically, both are 5-year intervals), each age/calendar period combination is

regarded as a stratum. Assume that there are J strata in total. For the cohort, observed numbers of deaths d_j and total person-years of observation time n_j are calculated for each of the strata. Then stratum-specific rates are straightforwardly calculated as:

$$\hat{\lambda}_j = \frac{d_j}{n_j}$$

Here, the $\hat{\lambda}_j$ are viewed as estimates of the underlying true disease rate in the stratum, λ_j. The goal of a cohort study is to determine whether these true rates differ between groups (ie. between the cohort and the general population) and how the rates are affected by differing exposure.

2.3.1 External comparisons

Overall cohort

If comparison with an external population is desired, the stratum-specific rates for this reference population are needed, and are calculated in the same manner as the cohort stratum-specific rates. The external population's death rate for stratum j will be denoted λ_j^*.

To calculate the number of deaths that would be expected in a given stratum j if the cohort had the same stratum-specific rates as the reference population, the external rate in stratum j is multiplied by the number of person-years in the j^{th} stratum of the cohort. Denoting the expected number of deaths in stratum j as E_j^* we have

$$E_j^* = n_j \lambda_j^*$$

The most common quantity used to compare the rate of death between a cohort and an external population is the standardized mortality ratio (SMR). When disease incidence is of interest, the corresponding quantity is the standardized incidence ratio (SIR), which is calculated in an identical fashion. The SMR is defined as the ratio between the observed deaths in the cohort and the expected number of deaths, which is calculated by applying the external population's rates to the cohort's age structure.

Letting D denote the total number of observed deaths in the cohort and E^* denote the total expected number of deaths:

$$SMR = \frac{\sum_{j=1}^{J} d_j}{\sum_{j=1}^{J} n_j \lambda_j^*} = \frac{D}{\sum_{j=1}^{J} E_j^*} = \frac{D}{E^*}$$

An SMR that is greater than 1 indicates an excess of risk in the cohort; an SMR that is less than 1 indicates a deficit of risk. As with any statistical quantity, a measure of significance is required in order to test the hypothesis that the SMR is not equal to 1 (ie. that the risk of disease differs between the cohort and the reference population.) This is accomplished by assuming that under the null hypothesis the observed number of deaths, D has a Poisson distribution with mean and variance E^*. P-values can be calculated using either exact tables of the Poisson distribution, or by various approximations which refer to normal or chi-square tables [3]. Confidence intervals for D can also be found using either exact methods or approximations, analagously to the calculation of p-values. To obtain exact or approximate confidence intervals for the SMR, the upper and lower endpoints of the confidence interval for D are simply divided by the observed deaths and multiplied by the SMR.

The reliability of the SMR depends on the assumption that the rate ratios are constant across all age categories. If this assumption does not hold, the SMR can be severely biased, although this is rare in practice.

Comparison of cohort subgroups

Regardless of whether a significantly large or small SMR is found, there may still be a dose-response relationship if mortality rates differ among cohort subgroups. Typically, causes of death that are of interest *a priori* or those with a significantly large or small SMR are selected for a subgroup analysis. The simplest type of subgroup comparison is to determine whether disease rates are the same in different cumulative exposure categories or not.

After a set of categories has been chosen, the person-years in the cohort must be correctly assigned to those categories. For each person-year, one needs to calculate which exposure category the individual would fall into if death occurred during that year, rather than just use the cumulative exposure for the individual at the end of

their observation time. Once the person-years have been assigned to the appropriate exposure category (and categories for any stratification variables that are being used), one can then calculate SMRs for each exposure category. Assume there are K exposure groups. Then the expected number of deaths for each category is calculated as:

$$E_k^* = \sum_j n_{jk}\lambda_j^*$$

and the SMRs for each category are calculated as:

$$SMR_k = \frac{O_k}{E_k^*}$$

The overall SMR can be represented as O_+/E_+^* where O_+ and E_+^* are the total number of observed deaths and expected deaths in the cohort.

In occupational cohort studies, differences in risk between exposure groups within the cohort are of greater interest than differences between the exposure groups and the external population. A large SMR may suggest that individuals in the cohort are receiving some sort of exposure that is increasing their risk for that particular cause of death. In order to demonstrate causality, however, one must show that an *increase* to a specific exposure leads to an increase in risk of death.

Using the lowest exposed group as a baseline, relative risks ψ_k can be calculated for the other exposure groups by calculating the ratio of SMRs of group k to group 1. Pairwise significance tests can be carried out by using the binomial distribution, and Pearson's χ^2 test can be used to assess the null hypothesis that none of the relative risks are different than 1. The test statistic for the χ^2 test is as follows:

$$\chi^2_{K-1} = \sum_{k=1}^{K} \frac{(O_k - \tilde{E}_k^*)^2}{\tilde{E}_k^*}$$

In the above formula, $\tilde{E}_k^* = O_+(E_k^*/E_+^*)$ is the "adjusted expected value" [3] calculated under the hypothesis that all the $SMR_k s$ are equal.

In order to compare differences in risk between the K exposure categories and the baseline ($k = 1$) category, the ratio of SMR_ks can be used. The relative risk of exposure group k compared to exposure group 1 is:

$$\hat{\psi}_k = \frac{SMR_k}{SMR_1}$$

Values of ψ_k greater than 1 denote an excess of risk in exposure group k compared to exposure group 1, while values less than 1 denote a deficit of risk. Obtaining p-values and confidence intervals for the ψ_ks requires regarding the distribution of the observed deaths as Poisson. Assume, for simplicity, that we have only two exposure categories ($k = 2$). Then $O_1 \sim Poisson(\theta_1 E_2^*)$ and $O_2 \sim Poisson(\theta_2 E_2^*)$. Now denote θ_1 as θ and set $\psi = \theta_2/\theta_1$. Then $\theta_2 = \psi\theta$. Now we are interested in the distribution of O_2 conditional on the sum $O_+ = O_1 + O_2$. This distribution is binomial with parameter $\pi = \frac{\psi E_2^*}{E_1^* + \psi E_2^*}$. Solving for ψ we get:

$$\psi = \frac{\pi E_2^*}{(1-\pi)E_2^*}$$

It can be straightforwardly shown that the maximum likelihood estimate (mle) of π is $\hat{\pi} = O_2/O_+$. Then, by substitution, the mle of ψ is

$$\hat{\psi} = \frac{\hat{\pi}\tilde{E}_1^*}{(1-\hat{\pi})\tilde{E}_2^*} = \frac{O_2 E_1^*}{O_1 E_2^*}$$

This is just the ratio of the SMRs as described above. In order to obtain confidence limits for ψ, one must find confidence limits for π and then transform them. Breslow & Day [3] give exact $100(1-\alpha)\%$ CIs for π as:

$$\begin{aligned} \pi_L &= \frac{O_2}{O_2 + (O_1+1)F_{\alpha/2}(2O_1+2, 2O_2)} \\ \pi_U &= \frac{(O_2+1)F_{\alpha/2}(2O_2+2, 2O_1)}{O_1 + (O_2+1)F_{\alpha/2}(2O_2+2, 2O_1)} \end{aligned}$$

Here, $F_{\alpha/2}(\nu_1, \nu_2)$ represents the $100_{\alpha/2}$ percentile of the $F(\nu_1, \nu_2)$ distribution.

When looking for a dose-response effect, it is not of particular interest to ascertain pairwise differences between the various exposure categories and the baseline. Rather, it is desirable to test whether there is a monotonic trend in the SMRs with increasing exposure. To do this, the following Poisson trend statistic with 1 degree of freedom can be used:

$$\chi_1^2 = \frac{(\sum_{k=1}^K x_k(O_k - \tilde{E}_k^*))^2}{\sum_{k=1}^K x_k^2 \tilde{E}_k^* - (\sum_{k=1}^K x_k \tilde{E}_k^*)^2/O_+}$$

Here x_k is a quantitative dose level representing the k^{th} exposure category (often the midpoint). If no quantified exposure is available and the categories are just ordered, simply using $x_k = k$ would work.

One great disadvantage of categorizing cumulative exposure for analysis is that the results can be very sensitive to the choice of cutpoints for the exposure variable. In particular, the choice of the baseline comparison group can have a large effect on the magnitude and pattern of the resulting relative risks. Also, the number of exposure categories needs to be carefully considered. Too few categories could mask any particularities in the pattern of dose-response; too many categories could lead to too many parameters and would cause problems if the data is sparse (eg. if the disease in question is fairly rare). Some recommended strategies for cutpoint selection are a.) to decide *a priori* on some set of cutpoints, or b.) to use quantiles of the exposure distribution of either subjects or healthy members of the cohort [8]. Larger datasets are more robust to different choices of the number and placement of cutpoints.

2.3.2 Internal comparisons

So far, any comparisons between cohort subgroups have been done using the SMR_ks which are dependent on external rates. We have already mentioned the possible bias for any individual SMR_k but there is also a possible problem with the ratio of two SMRs: if age or strata distributions differ significantly between the two groups, the resulting SMR ratio could be severely biased, to the point of changing the sign of the effect in severe cases [3].

In order to compare subgroups without reliance on external data, internal standardization can be used. This is done by combining all the exposure groups together, calculating the stratum-specific rates and then comparing each subgroup to this internal standard. The stratum-specific death rates are then:

$$\lambda j = \frac{D_j}{N_j}$$

The expected number of deaths in stratum k is then calculated as follows:

$$E_k = \sum_{j=1}^{J} n_{jk} D_j / N_j$$

Internally standardized mortality ratios can be calculated by dividing the number of observed deaths by the number of expected deaths in each stratum, and then relative risks are obtained by taking the ratio of these internal SMRs analagously to the external comparisons. As with the adjusted expected values $\tilde{E}_k^*$, the sum of the E_k is equal to the sum of observed deaths in the cohort. These expected values can be used in place of the $\tilde{E}_k^*$ in the formulas to carry out tests for homogeneity and trend. These tests are not exact and tend to be rather conservative.

Both the external and internal comparison methods are simple to carry out and are usually viewed as just a preliminary glimpse into the data that motivates further analysis using parametric or semiparametric methods such as those described in the following sections. Using modelling techniques can allow us to incorporate covariates into the analysis, and also better account for the effect of age or strata membership by estimating the coefficients associated with these variables. Furthermore, we can look for an adequate functional form for the dose-response curve, which will hopefully provide a clearer picture of the relationship between exposure and death or cancer risk.

2.4 Analysis for Grouped Cohort Data

In this project, we are particularly concerned with modelling the relationship between the degree of exposure and rates of cancer death or incidence. We want to be able to separate the effects of exposure from those of other factors such as age and calendar period. To this end, we assume that we have data grouped into J age/calendar period strata and K exposure categories. If there are d_{jk} deaths and n_{jk} person-years of observation time in the j^{th} stratum and k^{th} exposure category, then the observed death or incidence rate is denoted as:

$$\hat{\lambda}_{jk} = \frac{d_{jk}}{n_{jk}}$$

The observed death counts d_{jk} are assumed to have a Poisson distribution with mean and variance $\lambda_{jk} n_{jk}$ and the person-years denominators are assumed to be fixed.

The most common model that is used for dose-response analysis is the multiplicative model. The basic model equation is:

$$\lambda_{jk} = \theta_j \phi_k$$

where θ_j corresponds to stratum j and ϕ_k is the relative risk for disease for exposure level k compared to the lowest exposure category ($k = 1$). Taking the logarithmic transform, this model can be restated as:

$$\log \lambda_{jk} = \alpha_j + \beta_k$$

where $\alpha_j = \log \theta_j$ and $\beta_k = \log \phi_k$. In order to be able to include other regression coefficients into the model, the multiplicative model can be generalized as follows:

$$\log \lambda_{jk} = \alpha_j + \mathbf{x}'_{jk} \beta_k$$

Here, the $\mathbf{x}_{jk}$ are p-dimensional row vectors of regression variables, and the β_k are the corresponding coefficients. The regression variables in this model represent fixed covariates. Since the exposure information is now contained in the vector of covariates, the exposure groups can be represented in various ways, depending on how much quantitative information is available about the level of exposure:

1. As a set of $k-1$ dummy variables indicating membership of a particular category or a single variable with k values treated as a factor
2. As a single numerical variable taking values $1, \ldots, k$, imposing an order on the categories and forcing a linear fit
3. A variable that takes on values that reflect the actual level of exposure in each category (often the midpoint), again using a linear fit

Since the d_{jk} are considered to have a Poisson distribution with $E(d_{jk}) = n_{jk}\lambda_{jk}$, we can rewrite the general multiplicative model to get a Poisson regression model:

$$\log E(d_{jk}) = \log n_{jk} + \alpha_j + \mathbf{x}_{jk} \beta_k$$

The logarithm of person-years n_{jk} is included as an offset term that has a known coefficient of 1. This means that we are only modelling the number of deaths in

each stratum/exposure category, ignoring the fact that if we had observed a different number of deaths, we would have also obtained a different number of person-years.

For the purposes of this project, we will consider age and calendar period as two separate stratification variables, and use exposure category as a factor and so we will rewrite the above model to reflect this. Assuming that we have I age strata and J calendar periods, define indicator variables that take value 1 for person-years occurring in the stratum in question and 0 otherwise.

We can then model the log expected number of deaths in each combination of exposure category/strata as follows:

$$\log E(d_{ijk}) = \log n_{ijk} + \alpha_i age_i + \beta_j calendar_j + \mathbf{x}_k \beta$$

Once the model has been fit, the relative risks comparing exposure groups can be calculated by exponentiating the estimated regression coefficients $\hat{\beta}_k$ which correspond to the exposure categories. Similarly, the relative risks of the various strata to the baseline group are also obtained by exponentiating the appropriate coefficients.

As with the analysis of SMRs, it is desirable to conduct a test for trend to assess whether there is an increasing or decreasing rate of disease with an increase in exposure. This is done by fitting the model with a single continuous variable x_k representing the level of exposure for each category. This term is then included as a linear term in the model fit. If the exposure categories represent intervals of some quantitative measurement, then one can use the midpoint of each interval as the value of the new covariate. If exposure is instead represented as a set of k ordered categories that do not have a readily available physical quantitative measure, simply using the category number $x_k = k$ is sufficient. The model fit is then:

$$\log E(d_{jk}) = \log n_{jk} + \alpha_j + x_k \beta$$

Small p-values for the effect of x are evidence that there is indeed a trend that is linear in x.

2.5 Analysis for Continuous Cohort Data with Time-Dependent Covariates

In order to get a clearer picture of the effect of exposure without the potential hazards of category selection, it is desirable to model exposure as a continuous, time-dependent variable. This can be done using an extension of the Cox proportional hazards model. In this section, we will briefly introduce the Cox proportional hazards (PH) model, then describe the counting process formulation and explain its use in the analysis of occupational cohort data.

2.5.1 Introduction to the Cox Proportional Hazard Model

Assume we have n individuals in a study group and we are interested in how differing values of various covariates affects survival of these individuals. Let $h_i(t)$ denote the hazard function for individual i, i.e. the instantaneous probability of failure at time t. In the context of cohort studies, failure represents death of a particular cause or diagnosis of a specified cancer. For an individual i, the Cox model specifies the hazard as:

$$h_i(t) = h_0(t)g(X_i(t), \beta)$$

Here, $X_i(t)$ is the set of covariate values for individual i at time t. These covariates can be either fixed or time-dependent. The nonnegative function $h_0(t)$ is the baseline hazard function, and is left unspecified. β is a vector of coefficients corresponding to $X_{(t)}$. $g(X_i(t), \beta)$ is some known function. A common choice for this function is $g(X_i(t), \beta) = e^{X_i(t)\beta}$. For two individuals with fixed covariate vectors, the ratio of their hazards is given by

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)g(X_i(t), \beta)}{h_0(t)g(X_j(t), \beta)} = \frac{g(X_i(t), \beta)}{g(X_j(t), \beta)}$$

This shows that the hazards of the two individuals are proportional, and that a change in covariate values affects the hazard function in a multiplicative fashion.

In order to examine the likelihood function, we will use $g(X_i(t), \beta) = e^{X_i(t)\beta}$. Assume that for each of the n individuals we have observed the pair (t_i, δ_i) where t_i is either a lifetime or a censoring time and δ_i is an indicator variable that takes the value 1 if t_i is a lifetime and 0 if t_i is a censoring time. Furthermore, assume that we have covariate information $\mathbf{x}_i$ for each individual and time t, where $\mathbf{x}_i$ is a vector of covariate values. Let K be the number of distinct failure times in the data set, and then define R_k as the risk set at time $t_{(k)}$, ie. the set of individuals that are alive and at risk just prior to failure time $t_{(k)}$. Then the likelihood can be written as

$$L(\beta) = \prod_{k=1}^{K} \frac{e^{\beta' \mathbf{x}_{(k)}}}{\sum_{l \in \mathbf{R}_i} e^{\beta' \mathbf{x}_{(l)}}}$$

The above equation is not a likelihood in the usual sense, but its use for inference on β has been justified through its formulation as both a marginal and partial likelihood [5]. Score and information functions can be straightforwardly calculated, and the resulting estimator for β has been shown to be consistent and asymptotically normal.

2.5.2 The Counting Process Formulation

In order to take into account time-dependent covariates, the counting process form of the Cox model is used. This form is very versatile and allows for many useful extensions of the Cox model: multiple events per subject, left truncation, and time-dependent strata and covariates [9].

In the standard Cox proportional hazards model described in the previous section, each individual has observed data consisting of the pair of variables (t_i, δ_i). In the counting process formulation, this data is replaced with a pair of functions $(N_i(t), Y_i(t))$, where

$$N_i(t) = \text{number of events that are observed for individual } i \text{ in the interval } [0, t]$$

$$Y_i(t) = \begin{cases} 1 & \text{if individual } i \text{ is in the risk set at time } t \\ 0 & \text{otherwise} \end{cases}$$

The above formulation clearly generalizes to include multiple events per individual, but we will restrict our interest to the case where each individual can have at most one event, ie. $N_i(t) \leq 1$ for all i, t.

The likelihood for this more general Cox proportional hazards model is very similiar to that for the standard model discussed in section 2.5.1:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{Y_i(t_i) e^{\beta' \mathbf{x}_i(t_i)}}{\sum_{l=1}^{n} Y_l(t_i) e^{\beta' \mathbf{x}_l(t_i)}} \right)^{\delta_i}$$

In order to simplify the likelihood, define the *risk score* for individual i as $r_i(\beta, t) = \exp(\beta' \mathbf{x}_i(t_i))$ [9]. The likelihood can now be written as:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{Y_i(t_i) r_i(t)}{\sum_{l=1}^{n} Y_l(t_i) r_l(t)} \right)^{\delta_i}$$

In the above likelihood function δ_i represents the number of events that occur at time t_i, and this it is obvious that this formulation can easily accomodate tied lifetimes (when this occurs, $\delta_i > 1$ for the value of i in question). The likelihood can also take into account left-truncation, which involves redefining the indicator variable $Y_i(t)$. Letting u_i denote the value that the lifetime of individual i is known to exceed, define $Y_i(t)$ as follows:

$$Y_i(t) = I(u_i \leq t \leq t_i)$$

With this redefinition, the above likelihood can be used as written.

The partial log-likelihood is:

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left[\log \left(Y_i(t_i) r_i(t) \right) - \log \sum_{l=1}^{n} Y_l(t_i) r_i(t) \right]$$

In the fixed covariate setting, the $\mathbf{x}_i(t)$ terms are simply replaced by vectors $\mathbf{x}_i$ containing the covariate values for each individual. The inclusion of time-dependent covariates adds a significant amount of complexity over the fixed case because it must be possible to calculate the values of the covariates for each individual at all failure times t_i where the individual is in the risk set. In the case of a binary variable (eg. a variable that indicates whether or not an individual has had a certain treatment), this is fairly straightforward. If the covariate is continuously varying (eg. cumulative

exposure to a chemical agent), the calculation can be significantly more difficult. A method for calculating cumulative exposure at a given time for an occupational cohort dataset will be discussed in 2.5.3.

Once covariate information is available for all the k risk times, the likelihood function can be used for parameter estimation. The maximum likelihood estimate $\hat{\beta}$ is found by solving $U(\hat{\beta}) = 0$, where

$$U(\hat{\beta}) = \frac{\partial l(\beta)}{\partial \beta}$$

2.5.3 Counting Process Form for Occupational Cohort Data

Use of the counting process form to accomodate time-dependent covariates is fairly straightforward from a data layout perspective. A typical dataset for the basic Cox model contains one observation per individual with survival or censoring time, status (censoring time or lifetime), and other variables indicating membership of various strata and values of covariates. The data file would typically appear as follows:

```
id   time   vitstat   x1   x2   ...
```

where `id` is an optional variable that serves as an identification variable for each individual, `vitstat` indicates vital status or disease status and `time` is the time to "failure". The variables `x1`, `x2`, etc. represent stratum and covariate information.

For the generalized counting process formulation, a dataset would contain a *set* of observations for each individual, and each observation would include a begin time, an end time, and status, strata and covariate information as follows:

```
id   begin   end   vitstat   x1   x2   ...
```

It is easily seen that the usual data format is just a special case of the generalized formulation, since the "start" time for each record is simply assumed to be 0 on the time scale in question. The structural difference between the two file formats is minimal, and both are straighforward. Unfortunately, occupational cohort data is rarely collected in this form, however, so an extensive data manipulation phase is necessary in order to be able to do a proportional hazards analysis.

A typical cohort study dataset would have two data files. One would contain demographic information and fixed covariates such as information on smoking for each individual in the cohort. This information would typically include at least date

of birth, gender, date of death, date of last follow-up and dates and types of cancer diagnoses, as well as an id variable to uniquely identify each individual. The other file would be a work history file which would contain one record for each job that each individual held, including starting and ending dates, and either an exposure level, or an identifying code that can later be used to link to a job exposure matrix (JEM) which gives the mean daily exposure level for each distinct job type in various time periods. These files are then usually run through one of several available computer programs that are meant to specifically analyze cohort studies. Most of the variables that we would need for survival analysis are internally calculated by these programs, and so are not part of the data file. These include cumulative exposure, time since first exposure, and time-dependent stratification variables like age group and calendar interval, which are calculated by classifying each person-year into the appropriate stratum. In order to do survival analysis, all these variables need to be explicitly calculated for each work history record and then formatted as required. With a large cohort, this can amount to signicant amount of computing power and time.

There is a feature of the Cox proportional hazards model that can greatly simplify the programming, and this can be seen by examining the likelihood: one only needs to know covariate information at the k distinct failure times, and only for those individuals who are in that particular risk set. So each individual's complete work history needs to be subdivided into intervals corresponding to the k failure times. Once this is done, cumulative exposure can be calculated for each worker at each failure time, and the start and stop dates for each records can be converted to other time variables.

Since an individual who stops work and then later resumes work is still considered at risk during the hiatus, these gaps must be filled in with dummy work history records that are assigned zero exposure. Likewise, the time between the end of an individual's work history and their date of last follow-up must also be accounted for by including another dummy work history record.

Choice of time variable

In an occupational cohort study, all time references in the dataset are to calendar time, although other time-related variables are often included as stratification variables in

order to match the format of available population-based rates. However, calendar time is not necessarily the time variable in which we are primarily interested. There are three choices of time variable: age, calendar time and time since first exposure. The time variable selected will be the basis for the estimation of the baseline hazard function, which is left completely unspecified in Cox proportional hazard modelling. Once the basic time variable has been selected, the other two time variables are typically included as regression variables in order to control for them. Age is commonly selected for the basic time variable since it is known to be an important determinant in cancer or death rates. Time since first exposure is also a frequent choice, but care must be taken since cumulative exposure can be highly correlated with time since first exposure, and so the effect of exposure could be lost in the estimate of the baseline hazard function [3].

2.5.4 Analysis

Once the dataset is in the correct format for the counting process form of the Cox proportional hazards model, there are many options available for analysis. The goal of all of them is to relate increases in cumulative exposure to changes in the risk of death or incidence of a disease. The curve that relates these two quantities is called the dose-response curve, and the objective is to be able to describe this curve as well as possible.

linear Model

The simplest analysis is to include cumulative exposure as a single, fixed, continous variable x_i with corresponding coefficient β. This entails modelling the logarithm of the rate ratio as a linear function of x_i, ie.

$$\log RR = x_i\hat{\beta}$$

logarithmic model

Experience has shown that in occupational cohort studies the rate ratio increases in a linear fashion at lower cumulative exposure levels, but then plateaus or even decreases

at higher levels [8]. A linear model would clearly not capture a plateau. By modelling the natural logarithm of cumulative exposure, however, this plateau effect can be incorporated. This is called the "logarithmic" model.

$$\log RR = \log x_i \hat{\beta}$$

If there are non-exposed individuals in the cohort ($x_i = 0$), then a small constant k must be added to all the x_i in order to avoid taking the logarithm of 0.

Exposure as a Factor

In order to avoid imposing a specific structure on the dose-response curve, one could recode cumulative exposure into K categories and model this new variable as a set of $K-1$ dummy variables $x_2, \ldots, x_K$ with corresponding coefficients $\beta_2, \ldots \beta_K$. Then the relative risk of exposure category k could be estimated with respect to the baseline category by simply calculating $e^{x_k \hat{\beta}_k}$. While this is still a simplistic method, it could be used to compare results of the Cox proportional hazards model with those obtained from SMR analysis or Poisson regression. Naturally, modelling exposure as a factor suffers from the same category choice problem as SMR and Poisson analysis.

Other Parametric Fits for Continuous Exposure

Another common model choice is to include the square root of cumulative exposure as a covariate. Like the logarithmic model, this model would capture the plateau effect which is common among occupational cohort studies. It is also possible to incorporate higher-order terms to model a polynomial fit. This allow for more flexibility than the linear or logarithmic model while still ensuring a smooth fit. However, this fit is not local, so a few data points could have a major effect on the resulting model, especially those that fall in the upper tails of the distribution of cumulative exposure.

Spline Fit

A far more flexible method for revealing the form of the dose-response is to use splines within the Cox proportional hazards model. By fitting data locally, instead of trying to fit one simplistic model to the data as a whole, splines can take into account a wide

variety of features in the data. Splines consist of a set of knot points and a continuous function made up of a series of segments that can be linear or have a higher degree. These segments are formed by taking a linear combination of a set of polynomial basis functions. There are several different types of splines, determined by how the number and location of knot points are chosen, and on the degree of the polynomial basis. Cubic splines (with continuous first and second derivatives) are usually fit because they have been found to provide a fairly reasonable compromise between adequate fit and computational simplicity.

With *regression* and *natural* splines the number of knot points is chosen in advance. One can either explicitly choose the knot locations, or simply select the degrees of freedom, which then automatically places the knots: one knot is placed at each of the endpoints, and the rest are equally distributed throughout the range of the data. So for example, if there were four knot points, there would be knots at the 33^{rd} and 67^{th} percentile of the data, as well as the two at the endpoints. For either natural or regression splines, the degrees of freedom (df) represent the number of basis functions used for the fit. The difference between the two methods lies in the number of knots and the treatment of the fit beyond the outer knots. Regression splines use $df - 1$ knots, natural splines use $df + 1$ knots and also impose a linearity constraint outside the outer knots.

Depending on the data at hand, the arbitrary choice of knot location in regression spline smoothing can potentially exaggerate or mask important features of the data. An alternative is to use *smoothing* splines, which optimize knot position for the given degrees of freedom. This way, the algorithm can allocate more knots to more "bumpy" sections of the data, and therefore detect jumps and dips that may be lost by simply placing knots at quantiles.

In the context of Cox proportional hazards modelling, splines are used to estimate the log hazard ratio, which can be interpreted as the log relative risk. Confidence bands can be constructed to assess whether changes in the log hazard ratio show a statistically significant departure from a straight line or not.

2.5.5 Model Checking

After the Cox proportional hazards model has been fit, it is prudent to assess whether the model was indeed appropriate for the data at hand. Schoenfeld residuals are particularly useful for this purpose. If there are K distinct failure times, the Schoenfeld residual at the kth failure time is:

$$s_k = X_{(k)} - \bar{x}(\hat{\beta}, t_k)$$

Here, $X_{(k)}$ is the covariate vector for the individual failing at time $t_{(k)}$. The quantity $\bar{x}(\hat{\beta}, t_k)$ is a weighted average of covariate values for those individuals at risk just prior to time t_k, with $Y_i(t)e^{X_i(t)\beta}$ as the weights. If there are tied failure times, Schoenfeld residuals are calculated as:

$$s_k = \int_{t_{k-1}}^{t_k} \sum_i [X_i - \bar{x}(\hat{\beta}, s)] dN_i(s)$$

where $dN_i(s)$ is the number of failures occurring at time s.

When doing any sort of proportional hazards analysis, it is important to check that the proportionality assumption actually holds. For fixed covariates, this means that for any two subjects i and j, the relative hazard between them is:

$$\frac{e^{X_i\beta}}{e^{X_j\beta}}$$

This relationship should be independent of the time scale, in this case age. If we have a fixed categorical covariate with just a few levels, this relationship can be visually verified by looking at the estimated log survival curves for each covariate group. They should appear proportional to each other. Clearly, if there are many levels to the covariate, or if the covariate is continuous, it is more difficult to confirm proportionality. It is preferable to have a statistical test that allows one to detect non-proportionality. One such test is called the *Z:ph* test. Using a selected transformation of the time scale $g(t)$ (typical choices for $g(t)$ include $log(t)$ and $1 - KM(t-)$, where $KM(t)$ is the Kaplan-Meier survival curve), one calculates the Pearson product-moment correlation between the Schoenfeld residuals and $g(t_k)$ for each covariate and then tests whether this correlation is equal to zero using the a χ^2 test developed to deal with the various choices of time scale [?]. Small p-values indicate the presence of non-proportionality.

In addition to the *Z:ph* test, most computer packages can also produce plots of scaled Schoenfeld residuals s_k^* versus the transformed time. Scaled Schoenfeld residuals are calculated by multiplying the S_k by the inverse of the estimated weighted variance of X at time k, $V(\hat{\beta}, t)$. This variance is calculated as:

$$V(\hat{\beta}, t) = \frac{\sum_i Y_i(t) r_i(t)[X_i(t) - \bar{x}(\hat{\beta}, t)]'[X_i(t) - \bar{x}(\hat{\beta}, t)]}{\sum_i Y_{i(t) r_i(t)}}$$

Then the scaled Schoenfeld residuals are $S_k^* = V^{-1}(\hat{\beta}, t_k) s_k$. If proportionality holds, the line fitted to the plot of the s_k^* should have a slope of 0. The *Z:ph* plots in *R* are augmented with a smoothing spline fit of the residuals, along with a ± 2 standard error confidence band which makes it easy to visually assess the validity of the proportionality assumption.

Chapter 3

Application to Aluminum Smelter Data

3.1 Background

In 1989, the Cancer Control Agency of British Columbia conducted a cohort study of workers at an aluminum smelter in British Columbia [6]. The main objectives of the study were to determine if workers at the plant had an excess risk of cancer incidence or mortality, or mortality from non-cancer causes. It was also of interest to study the relationship between exposure to coal tar pitch volatiles (CTPV) and mortality or cancer incidence at specific sites. Data on smoking habits was also collected, since smoking was a potential confounder for various cancers.

The mortality and incidence of the cohort was compared to that of the general B.C. population using standardized mortality ratios (SMRs) and standardized incidence ratios (SIRs). A significant excess in mortality from brain cancer was observed. As for incidence, significant excess risk of bladder cancer incidence was found, and this risk was found to be significantly related to increased exposure to CTPV. There were also elevated rates of brain and testicular cancer, although they were not significant.

In 2000, the British Columbia Cancer Agency undertook an update and expansion of the study by repeating the analysis with the now much larger cohort of workers that have worked at the smelter for at least three years, with the study end taken to be December 31, 1999. Beyond extending the incidence and mortality studies and the

dose-response analysis of the old study, the aims of this new study were to update the dataset with new personal and work history information, to update the exposure assessment for CTPV by using new methods of retroactive exposure assessment, and to include benzo[a]pyrene (BaP) in the exposure assessment. In particular, the larger cohort and longer study duration would hopefully allow for better quantification of the dose-response relationships between exposure and cancer mortality or incidence.

3.2 Aluminum Smelter Data

The original cohort included all workers who had worked at the Alcan smelter for at least 5 years between January 1, 1954 and October 15, 1985. Personal information was collected from the Alcan records and included full names, dates of birth, and gender. In addition, complete work histories were obtained for each individual. These included job title, department, start date and stop date for each job held. Active follow-up was conducted to locate and ascertain vital status information from any cohort members that were not actively employed at Alcan at the study end date. A total of 4503 individuals were included in the cohort, and the successful trace percentage was 91.8 %.

The cohort for the new study consisted of all workers with at least three years of employment at the Alcan smelter between 1954 and December 31, 1999, the cutoff date for follow-up. There were 7007 workers enrolled in the cohort, but 15 of these had no listed birthdate so the final cohort consisted of 6992 individuals. Of these, 6395 were males and 597 were females. Because of the relatively small number of females (with only 51 total deaths) only males will be considered in the analysis that follows.

Mortality and cancer information was obtained through linkage with the National Mortality Database at Statistics Canada. This information consisted of date of death or cancer diagnosis and the corresponding International Classification of Diseases (ICD) code, which indicates cause of death or the cancer site.

No active follow-up took place in this new phase of the study, but reasonable follow-up information was obtained by combining several pieces of information. When the original study was done, both the date and location of last follow-up had been

recorded. For the updated study, individuals were linked through the Medical Services Plan (MSP) registry, which provides medical insurance for residents of British Columbia. This linkage provided entry and exit dates to the MSP program, indicating whether or not an individual was a resident of BC at the time. All non-deceased individuals were censored at study end (December 31, 1999) unless the individual a.) was known to be out of the country at the time of last contact, or b.) was last known to be employed at Alcan prior to 1985 and was not successfully linked to the MSP registry. Individuals in these two categories were censored at their last date of contact.

In the original study, a job exposure matrix (JEM) was created to assess the effects of CTPV exposure measured as benzene soluble materials (BSM). In order to construct this matrix, the study period was first broken down into 13 smaller intervals corresponding to union contract periods. Then employee records were used to determine all distinct jobs that were held at the plant. Each combination of job title and time interval was placed into one of four exposure categories: no exposure, low exposure ($< 0.2\ mg/m^3$ BSM), moderate exposure (0.2-1.0 mg/m^3 BSM), and high exposure ($> 1.0\ mg/m^3$ BSM).

In the years since the original study was conducted, the methodology for retrospective exposure assessment has advanced, and this allowed the research team to not only extend the JEM to include the years from 1986 onward, but also to refine and improve exposure assessment for the job/time period combinations that were included in the original JEM. The original JEM was created using expert-based assessment, which involved a team of union and company employees who assigned job/time period combinations to the four exposure categories. Subsequent work has shown that this type of assessment can be much improved by using quantitative methods [4]. The new exposure assessment for the updated study included direct measurement of mean daily exposure levels, statistical modelling, and extrapolation to obtain quantitative exposure levels for jobs with no direct measurements. The resulting JEM contains 78 distinct job identifiers (referred to as "plant" codes) and was assessed over 9 time intervals. Each job/time period combination was assigned to one of 7 exposure categories: unexposed, 0.01-0.1, 0.1-0.2, 0.2-0.4, 0.4-1, 1-2, and $> 2\ mg/m^3$, and the midpoint of these categories was used to calculate individual cumulative exposure.

For the highest exposure category, 2.5 mg/m^3 was used for calculation. The default level of exposure used was 0.0001 mg/m^3 BSM per day.

In addition, a new JEM was created to capture the exposure assessment for benzo[a]pyrene, which is thought be more indicative of cancer risk than BSM [1]. This JEM also lists exposure levels for 78 different job identifiers, but divides the study period into 13 intervals instead of 9. Like the BSM JEM, personal exposure measurements were used wherever possible, and modelling and extrapolation were used to estimate mean daily exposure levels for jobs with no direct measurements. Seven categories were chosen: unexposed, 0.05-0.5, 0.5-1, 1-3, 3-7, 7-14, and > 14 $\mu g/m^3$. Again, the midpoints of these intervals were used for calculations, with 18 $\mu g/m^3$ being used for the highest category.

For the overall cohort analysis, a 3 year lag time was used in order to help control for the healthy worker effect.

3.3 Overall Cohort Analysis

3.3.1 Mortality Study

The mortality of the cohort was compared to that of the British Columbia population by calculating standardized mortality ratios (SMRs). The Laboratory Center for Disease Control division of Health Canada provided population mortality rates, calculated in 5 year age groups and 5 year calendar intervals from 1950 to 1999. For each individual, person years at risk were calculated from their first date of hire until their date of last follow-up or death.

Expected values of deaths and SMRs were calculated using the BC rates and significance tests were performed. It was assumed that the number of observed deaths followed a Poisson distribution with mean equal to the number of expected deaths. P-values and 95% confidence intervals were calculated for each cause of death.

Overall mortality for males was found to be significantly less than that for the general population of British Columbia (SMR=0.87). No significant excesses were found, but elevated rates of brain, pancreatic, stomach, and bladder cancer were detected.

3.3.2 Incidence Study

Since population-based cancer rates for British Columbia only exist from 1970 on, the mortality cohort had to be suitably redefined for the incidence study. Of the 6395 male workers included in the mortality cohort, 642 were excluded because they had no person-years at risk after 1969. These individuals either died or were lost to follow-up before 1970. The final cohort for the incidence study therefore consisted of 5781 males. Of these, 662 were diagnosed with cancer at least once, 287 were lost to follow-up, and 4832 had no diagnosed cancer prior to death or the end of the study.

A significant excess of bladder, pancreas and stomach cancer was found, along with non-significant excesses in brain, lung and mouth cancer.

3.4 Preliminary Dose Response Analysis and Poisson Regression

A preliminary dose-response analysis of the ratios of SIRs was performed on seven cancers of interest: bladder including in-situ, lung, kidney, stomach, brain, pleura, and non-Hodgkins lymphoma.

For BSM, the following cumulative exposure categories were chosen: 0-0.05, 0.05-2, 2-4, 4-8, 8-16, and 16+ BSM-years (in mg/m^3-year). 0.05 BSM-years represents approximately one year at the default exposure level (0.0001 mg/m^3 of average daily exposure), so workers with less than this amount of exposure can be considered unexposed. 2 BSM-years is equal to 10 years of exposure at the threshold limit value (TLV) for BSM, which is 0.02 mg/m^3 of average daily exposure. The TLV is the amount of exposure below which no adverse health effect is expected. The remaining cutpoints are equivalent to 20, 40, and 80 years of exposure at the TLV. The ratio of BaP to BSM ($\mu g : mg$) is roughly 10:1, so the cutpoints for BaP were directly calculated from the BSM cutpoints, giving the exposure categories as: 0-0.5, 0.5-20, 20-40, 40-80, and 80+ BaP-years, measured in $\mu g/m^3$-years (the highest two categories were combined due to low numbers of observed events and person-years). Table 3.1 shows the age categories and calendar periods that were used as stratification variables in the preliminary analyses and as covariates in Poisson regresson.

Table 3.1: Categories used for Poisson regression

Age Categories	Calendar Periods
≤ 50	1970-1974
50-54	1975-1979
55-59	1980-1984
60-64	1985-1989
65-69	1990-1994
70-74	1995-1999
75-79	
80+	

A variety of latency periods were also analysed for the cancers of interest in order to determine which lag time resulted in the strongest linear relationship. Poisson regression was conducted with 3, 5, 10, 15, and 20 year lag times. The latency period that showed the strongest linear trend was selected as the "optimal lag time".

Of the cancers analyzed, bladder cancer, non-Hodgkins lymphoma and lung cancer were selected to perform a more detailed analysis including Cox proportional hazards. Only these three sites analysed with their optimal lag will be considered for the remainder of the project (20 years for bladder and lung cancer, 10 years for non-Hodgkins lymphoma).

3.4.1 Non-Parametric Dose-Response Analysis

Bladder Cancer Incidence

Using British Columbia cancer rates from LCDC, the number of expected bladder cancer incidences E_k^* was calculated for each of the k exposure categories. The SIR_ks and relative risks ψ_k were then obtained, along with adjusted expected values $\tilde{E}_k^*$. Tests for homogeneity and trend in the SIR_ks were carried out. The results are shown in Table 3.2.

This first analysis of the data shows a relative deficit of risk in the second exposure category and then an excess in the highest three categories. Despite the "dip" and rise in the relative risk which is clearly nonlinear, the SIR test for trend is still highly significant, which suggests that there is indeed an increase in risk with increased

Table 3.2: External Comparisons for Bladder Cancer

BaP years	0-0.5	0.5-20	20-40	40-80	80+
O_k	26	18	16	17	13
E_k^*	17.23	14.71	7.09	6.93	4.01
SIR	1.51	1.22	2.26	2.45	3.24
95 % CI	0.99-2.21	0.73-1.93	1.29-3.67	1.43-3.93	1.73-5.55
$\hat{\psi}_k$	1.00	0.81	1.50	1.63	2.15
95 % CI		0.42-1.54	0.75-2.90	0.83-3.11	1.01-4.34
$\tilde{E}_k^*$	31.04	26.50	12.77	12.48	7.22
Test for homog. of SIR	10.630	p-value	0.031		
Test for trend in SIR	9.13	p-value	0.003		

exposure. For the trend test, the category midpoints were used as the coefficients x_k. For the highest exposure category, 100 BaP-years was used as the coefficient, as it is approximately the mean of the cumulative exposures in that category. It is interesting to note that only the relative risk in the highest category is statisticall different than 1 at the 0.05 level, as shown by the confidence intervals.

Next, an internal comparison of exposure categories was conducted by combining all exposure categories to get stratum-specific inicidence rates and calculating the expected incidences E_k. The results are shown in Table 3.3. The SIR_ks and ψ_ks were then calculated using these expected numbers and the tests for homogeneity and trend were repeated. The relative risks show a similar pattern to those obtained through external comparison, with a deficit in the second exposure category, and a linear increase in the highest three categories. Again, both the test for homogeneity and trend are significant, which confirms the dose-response relationship shown by the relative risks.

Non-Hodgkins Lymphoma Incidence

Just as for bladder cancer incidence, external and internal comparisons were done for non-Hodgkins lymphoma incidence. The results for the external comparison are in Table 3.4 and those for the internal comparison are in Table 3.5. Immediately one notices a much stronger dose-response effect than for bladder cancer, with the risk

Table 3.3: Internal Comparisons for Bladder Cancer

BaP years	0-0.5	0.5-20	20-40	40-80	80+
O_k	26	18	16	17	13
E_k	29.45	26.74	13.20	13.45	7.16
SIR	0.88	0.67	1.21	1.26	1.82
95 % CI	0.58-1.29	0.40-1.06	0.69-1.97	0.74-2.02	0.97-3.10
$\hat{\psi}_k$	1.00	0.76	1.37	1.43	2.06
95 % CI		0.39-1.44	0.69-2.66	0.73-2.74	0.97-4.15
Test for homogeneity of SIR	9.557	p-value	0.049		
Test for trend in SIR	7.751	p-value	0.005		

in the highest categories being about 6 times those in the baseline category. In both analyses, the relative risk for the 20-40 BaP-years category is higher than would be expected if the dose-response relationship was truly linear, but the overall trend is clear.

Table 3.4: External Comparisons for Non-Hodgkins Lymphoma

BaP years	0-0.5	0.5-20	20-40	40-80	80+
O_k	3	5	7	4	8
E_k^*	7.21	9.51	4.61	4.44	3.26
SIR	0.42	0.53	1.52	0.90	2.45
95 % CI	0.09-1.22	0.17-1.23	0.61-3.13	0.25-2.31	1.06-4.84
$\hat{\psi}_k$	1.00	1.26	3.65	2.17	5.90
95 % CI		0.25-8.14	0.83-21.87	0.37-14.78	1.42-34.52
$\tilde{E}_k^*$	6.71	8.84	4.29	4.13	3.03
Test for homogeneity of SIR	13.574	p-value	0.009		
Test for trend in SIR	10.073	p-value	0.002		

As would be expected by the reasonably monotonic increase of the SIR_k and the very high relative risks in the highest categories, both the tests for homogeneity and trend in the SIR_ks have very small p-values for both internal and external standardization. As with bladder, only the highest category has a relative risk significantly greater than 1, even though the magnitude of the relative risks is much greater. This

Table 3.5: Internal Comparisons for Non-Hodgkins Lymphoma

BaP years	0-0.5	0.5-20	20-40	40-80	80+
O_k	3	5	7	4	8
E_k	7.19	8.62	4.09	3.95	3.14
SIR	0.42	0.58	1.71	1.01	2.55
95 % CI	0.09-1.22	0.19-1.35	0.69-3.52	0.28-2.59	1.10-5.02
$\hat{\psi}_k$	1.00	1.39	4.10	2.43	6.11
95 % CI		0.27-8.95	0.94-24.57	0.41-16.56	1.47-35.77
Test for homog. of SIR	13.568	p-value	0.009		
Test for trend in SIR	10.136	p-value	0.001		

is due to the small number of cases (27), which causes the extreme width of the confidence intervals.

Lung Cancer Incidence

Tables 3.6 and 3.7 show the external and internal comparisons for lung cancer. The results of the two analyses are almost identical to each other, and show a strong linear trend with the highest exposure category having 2 times the risk of the baseline category. Both tests for SIR homogeneity fail to reject the null hypothesis, but the trend tests are both significant at the 0.05 level.

Table 3.6: External Comparisons for Lung Cancer

BaP years	0-0.5	0.5-20	20-40	40-80	80+
O_k	41	43	23	22	18
E_k^*	46.77	39.06	19.41	19.02	10.16
SIR	0.88	1.10	1.19	1.16	1.77
95 % CI	0.63-1.19	0.80-1.48	0.75-1.78	0.73-1.75	1.05-2.80
$\hat{\psi}_k$	1.00	1.26	1.35	1.32	2.02
95 % CI		0.80-1.98	0.77-2.31	0.75-2.29	1.09-3.60
$\tilde{E}_k^*$	51.15	42.71	21.23	20.80	11.11
Test for homogeneity of SIR	6.503	p-value	0.165		
Test for trend in SIR	5.138	p-value	0.023		

Table 3.7: Internal Comparisons for Lung Cancer

BaP years	0-0.5	0.5-20	20-40	40-80	80+
O_k	41	43	23	22	18
E_k	50.98	43.23	20.85	20.79	11.15
SIR	0.80	0.99	1.10	1.06	1.61
95 % CI	0.58-1.09	0.72-1.34	0.70-1.66	0.66-1.60	0.96-2.55
$\hat{\psi}_k$	1.00	1.24	1.37	1.32	2.01
95 % CI		0.79-1.95	0.79-2.34	0.75-2.26	1.09-3.57
Test for homog. of SIR	6.456	p-value	0.168		
Test for trend in SIR	5.167	p-value	0.023		

3.4.2 Poisson Regression

After the preliminary non-parametric analysis, Poisson regression was used to model the dose-response relationship. The same age groups and calendar periods used as strata for the SIR analyses were included as covariates in the models, and the exposure categories were kept the same.

Let $exposure_k, k = 1, \ldots, 5$ be a series of indicator variables for the 5 exposure categories. Similarly, let $age_i, i = 1, \ldots, 8$ and $calendar_j, j = 1, \ldots, 6$ represent indicator variables for the age groups and calendar intervals respectively. Note that it is individual person-years, not workers, that are categorized into these groups. We now want to model the log expected number of deaths in each combination of exposure category/strata as follows:

$$\log E(d_{ijk}) = \log n_{ijk} + \sum_{i=1}^{8} \alpha_i age_i + \sum_{j=1}^{6} \beta_j calendar_j + \sum_{k=1}^{5} \gamma_k exposure_k$$

Since the number of person-years in each stratum/exposure category combination is assumed to be known constant, $\log n_{ijk}$ is included in the model as an offset term with no estimated coefficient. The exponentiated cofficients that resulted from this model fit are the estimated relative risks, and can be easily compared to those obtained with the non-parametric analyses conducted above.

Bladder Cancer

Table 3.8 shows the relative risks for the 5 exposure categories obtained from Poisson regression with the above model, along with 95% confidence intervals.

Table 3.8: Poisson Regression for Bladder Cancer

Exposure Category	Relative Risk	95% CI
0-0.5	1.00	-
0.5-20	0.84	0.45-1.58
20-40	1.53	0.79-2.95
40-80	1.65	0.85-3.18
80+	2.36	1.14-4.89

The estimates of relative risk obtained from Poisson regression were very similar to those obtained in the external comparison, although some of the values are a little higher. We again notice that only the coefficient for the 80+ BaP-years category is significantly different than 1 at the 0.05 level.

The residual deviance for this model fit was 162.98 on 295 degrees of freedom, which suggests that the model may be overfitting the data. This is likely due to the sparseness of data within the cells. A Poisson test for trend was performed and the resulting p-value was 0, further confirming the monotonic increasing trend noticed in the preliminary analysis.

Non-Hodgkins Lymphoma

With the same model and covariates as for bladder cancer, Poisson regression was performed for non-Hodgkins lymphoma as well. The estimated relative risks and 95% confidence intervals are in Table 3.9.

The relative risks obtained were higher than those produced by the non-parametric analysis. We also note that the confidence intervals are extremely wide. As in the internal and external comparisons, the estimated relative risk for the 20-40 BaP-years category is much higher than that for the 40-80 BaP-years category, which suggests a non-linear dose-response curve. However, the small number of cases suggests that this may simply be an artifact of the data, rather than evidence of a truly non-linear

Table 3.9: Poisson Regression for Non-Hodgkins Lymphoma

Exposure Category	Relative Risk	95% CI
0-0.5	1.00	-
0.5-20	1.43	0.34-6.05
20-40	4.58	1.16-18.01
40-80	2.94	0.63-13.64
80+	8.21	1.99-33.78

relationship between cumulative exposure and risk of non-Hodgkins lymphoma.

The residual deviance of this model fit was 99.13 on 384 degrees of freedom, and the p-value for the Poisson test for trend was 0.00002.

Lung Cancer

Table 3.10 shows the relative risks and 95% confidence intervals for the Poisson regression analysis of lung cancer.

Table 3.10: Poisson Regression for Lung Cancer

Exposure Category	Relative Risk	95% CI
0-0.5	1.00	-
0.5-20	1.34	0.85-2.09
20-40	1.48	0.87-2.53
40-80	1.45	0.83-2.52
80+	2.26	1.24-4.12

Here we see a very similar pattern to the non-parametric comparisons, with the estimated relative risks for the 20-40 and 40-80 BaP-years categories nearly identical. The magnitude of the relative risks are all larger than the previous estimates, but still only the highest exposure category has a relative risk that differs significantly from 1.

For this model fit, the residual deviance was 223.35 on 301 degrees of freedom. The p-value for the Poisson test for trend was 0.

3.4.3 Discussion of Preliminary Analyses

For all three cancer sites, the trend test shows a significant monotonic increasing trend in the SIRs. However, the pattern of relative risks for all three sites suggest that a linear fit may not be the most adequate descriptor of the dose-response relationship. Bladder cancer shows a category that has a *deficit* of risk compared to the baseline group, and the three analyses of lung cancer show the 20-40 and 40-80 BaP-year categories to have almost identical risk, which does not conform to a linear model. Also, all analyses of non-Hodgkins showed that the relative risk for the 20-40 BaP-year was as much as nearly twice that of the 40-80 BaP-year category. This suggests that perhaps other parametric or semi-parametric models could be useful to better describe the relationship between cumulative exposure and risk.

3.5 Cox Proportional Hazards

3.5.1 Data Manipulation

In order to perform Cox PH modelling on the cohort data, a significant amount of data manipulation was necessary. In this section, the steps in the manipulation will briefly be described. Three data files were created, one for each cancer site.

The original work history file had already been significantly modified in order to do the preliminary analyses. Since follow-up time ends when the individual is diagnosed with cancer even though the person may have continued working, each individual's set of records needed to be truncated on their diagnosis date (if they have one). In order to avoid creating multiple files, indicator variables were created to identify whether a record should be included in the analysis of a certain cancer, and new end date variables were created to truncate records that spanned the date of diagnosis. It is this modified work history file that was further manipulated in order to perform Cox PH modelling.

For each record in the work history file, the following additional variables needed to be calculated:

- Time in days since first exposure at the beginning and end of the record

- Age at beginning and end of the record (calculated as days since Jan 1, 1890)
- Birthdate of the individual (calculated as days since Jan 1, 1890)
- Calendar dates of the beginning and end of the record (as days since Jan 1, 1890)

In addition, extra records needed to be created to represent the time between the last date worked and the date of last follow-up. Obviously, no CTPV exposure is accrued during this time, so a dummy plant code was created to take this into account. The next step was to create an "event" variable to indicate whether the individual in question was diagnosed with cancer at the end of the time interval or not. At this point, the data file was imported into R, along with vectors of failure times (both in terms of age and time since first exposure).

In order to assign exposure to each work history record, the job exposure matrix needed to be manipulated as well. The original JEM file listed average exposure levels for each of 78 plant codes during different intervals, with varying numbers of intervals covered per plant. The function *survSplit* was used here to split any records in the JEM that covered more than one interval in order to obtain one row for every possible combination of time interval and plant. From this file, a 78×13 matrix was created with each row corresponding to a plant code and each column to a time interval. Each cell contained the mean daily exposure level for the corresponding plant/interval combinations, and any empty cells were filled with zeros for ease of computation.

As discussed in section 2.5.3, the likelihood for the Cox proportional hazards model requires one to know the value of all covariates at each of the failure times. For this study, that requires the calculation of every individual's cumulative exposure at each of these failure times. This was accomplished in four steps:

- splitting each record that spanned a failure time into two intervals (to obtain a record that ends on the failure time)
- further splitting each of these records into "equal exposure intervals" corresponding to cells in the JEM

- multiplying record duration by the mean daily exposure level in the appropriate cell in the JEM
- calculating the cumulative exposure for each individual by appropriately summing the total exposure for each record

The record splitting was easily accomplished using the R function *survSplit* from the *survival* package and using the vector of failure times as cutpoints.

3.5.2 Time Variable

Both commonly used time variables were considered in the analysis: age and time since first exposure (TSFE). Breslow and Day [3] recommend using age since it is known to have a highly significant effect on background cancer rates. The authors also caution against using TSFE since it is highly correlated with cumulative exposure, and may therefore obscure some of the dose-response effect. For the sake of comparison, both time metrics were used in the analysis, but the results were very similar, so only those with age as the time variable will be presented in detail here.

Models

For each cancer site, three simple semiparametric models were fit with cumulative exposure as a continuous variable:

- Linear: $logRR = \sum_{j=1}^{6} \beta_j calendar_j + x_i(t)\hat{\gamma}$
- Logarithmic: $logRR = \sum_{j=1}^{6} \beta_j calendar_j + log(x_i(t) + 1)\hat{\gamma}$
- Square root: $logRR = \sum_{j=1}^{6} \beta_j calendar_j + sqrt(x_i(t))\hat{\gamma}$

In the logarithmic model, 1 was added to the cumulative exposure in each record in order to avoid taking the logarithm of zero. Also, using 1 instead of a smaller constant ensured that all of the logged values were greater than 0.

Next, cumulative exposure was categorized into the same groups as for the Poisson regression analysis, and the data was analyzed with and without an adjustment for calendar time. The adjustment was done by calculating the calendar year for

each work history record and grouping the result into 5 year intervals, which were subsequently treated as a factor.

For a more flexible dose-response curve, three types of smoothing models were fit:

- Polynomial fits of varying degrees
- Regression splines with knots at cumulative exposure interval endpoints
- Smoothing splines with varying degrees of freedom

After all the models had been fit, Schoenfeld residuals were calculated and plotted to assess the validity of the proportional hazards assumption.

3.6 Results

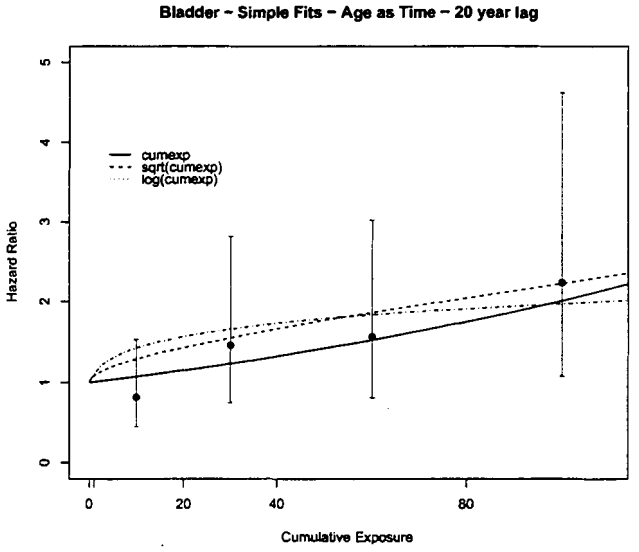
3.6.1 Bladder Cancer

First, the four simplest models were fit to the bladder cancer data: linear, logarithmic, square root, and categorical. In each case, calendar interval was included as a factor to help adjust for changes in background cancer rates. The results are shown in Table 3.11. The hazard ratios are the exponentiated coefficients and represent the multiplicative increase in risk for an increase of one unit of the coefficient in question. The p values from the likelihood ratio tests are included to assess the fit of each of the models.

Table 3.11: Simple Cox PH Models for Bladder

Fit	Covariate	Hazard Ratio	95% CI	LRT p-value
Linear	cumexp	1.007	1.002-1.012	0.0298
Logarithmic	Log(cumexp+1)	1.158	1.014-1.324	0.050
Square root	Sqrt(cumexp)	1.083	1.019-1.152	0.0285
Categorical	0-0.5 BaP-years	1.00	-	0.0532
	0.5-20 BaP-years	0.82	0.44-1.53	
	20-40 BaP-years	1.46	0.75-2.82	
	40-80 BaP-years	1.56	0.81-3.02	
	80+ BaP-years	2.24	1.08-4.62	

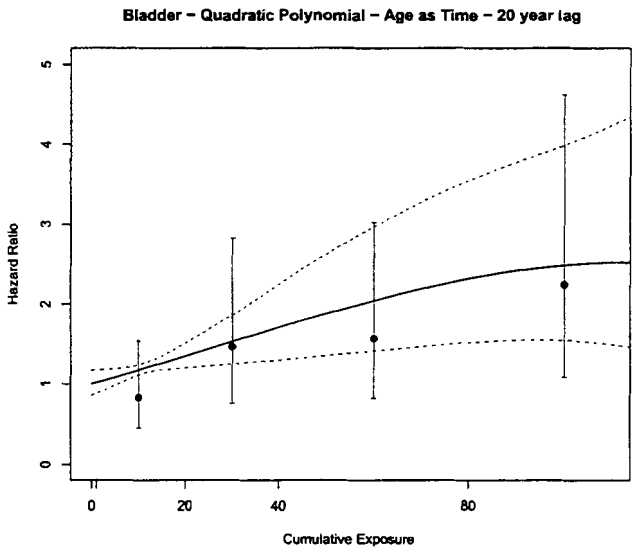
Figure 3.1: Bladder - Simple Models of Relative Risk as a Function of Cumulative Exposure



The graphs of the three continuous fits are shown in Figure 3.1. The vertical bars represent the point estimates and confidence intervals for the relative risks of the four non-zero exposure categories from the categorical fit. The linear fit appears to be the closest fit to the categorical estimates, while the logarithmic and square root fits overestimate the relative risk for individuals in all but the highest exposure category. Also none of these models can account for the apparent deficit of risk in the second exposure category, since they are all monotone increasing functions of cumulative exposure. The poor fit of the logarithmic model in particular is quite evident, and is confirmed by the large p-value from the likelihood ratio test.

In order to allow a bit more flexibility in the model, a series of polynomial fits with varying degrees of freedom were fit to the data. Fits with 2, 3, 4, 5, and 8 degree polynomials were computed. Since this forms a series of nested models, a series of likelihood ratio tests were performed to determine whether each added degree improved the fit. For bladder cancer, it was shown that the addition of the quadratic term improved the fit over the linear model, but adding more terms did not produce any more statistically significant improvements in fit. The quadratic polynomial is

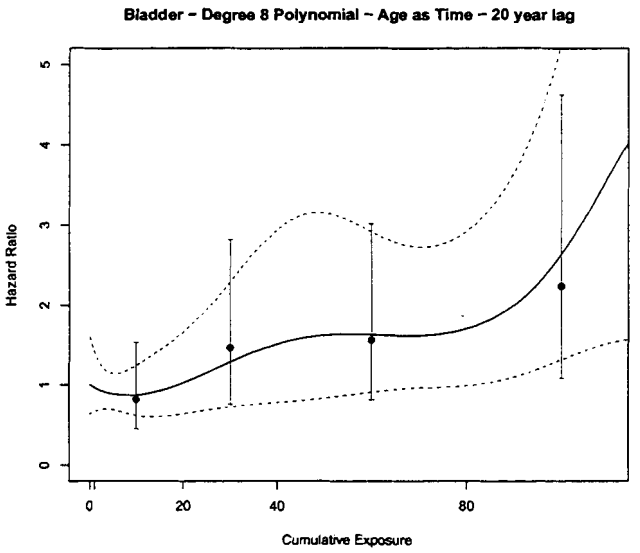
Figure 3.2: Bladder - Quadratic Model of Relative Risk as a Function of Cumulative Exposure



shown in Figure 3.2, with vertical bars to show the point estimates and 95% CIs from the categorical fit for reference. With degree 2, the fit does not pick up the deficit in risk in the 0.5-20 BaP-years exposure category. It also seems to slightly overestimate the hazard ratio for the midpoints of each exposure category. It is also interesting to note that the confidence bands at the lower exposure levels are much narrower than those given by the categorical estimates.

For comparison, the degree 8 polynomial is included in Figure 3.3. This more flexible curve does seem to be better match the hazard ratio estimates from the categorical Cox PH analysis, and the confidence bands nearly match the confidence intervals obtained from the categorical analysis as well. However the resulting shape of the dose-response curve is much more complex and after 80 BaP-years, the hazard ratio rises far more steeply than suggested in any of the other previous analyses. This is likely due to a few observations that are having a disproportionately large influence on the polynomial fit and are therefore distorting the shape of the overall dose-response curve.

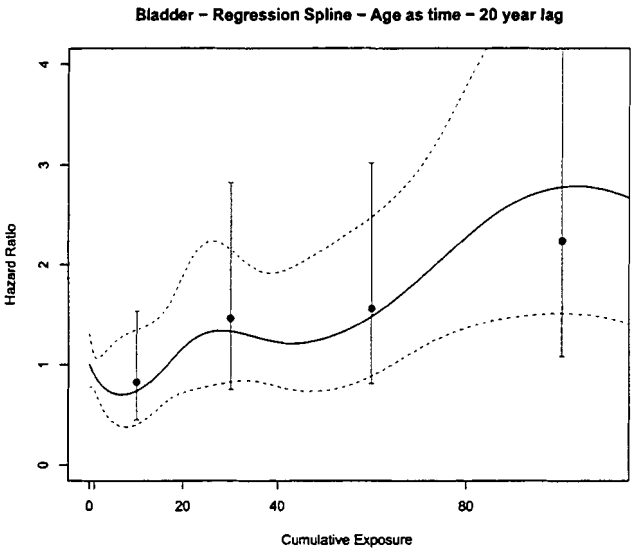
Figure 3.3: Bladder - Degree 8 Model of Relative Risk as a Function of Cumulative Exposure



Next, a regression spline was used to model the relationship between the log-hazard and cumulative exposure, with calendar period included as a factor. The knots were placed at category cutpoints, and the resulting spline, along with 95% confidence bands is shown in Figure 3.4. The plot reveals several of the features noticed in earlier analyses: a deficit of risk between 0 and 20 BaP-years and a rise in relative risk to over 2. The categorical point estimates and confidence intervals correspond well with those obtained by the regression spline fit. At about 100 BaP years, the graph seems to indicate a decrease of risk for those most highly exposed, which is due to a small number of individuals with high cumulative exposure who were never diagnosed with bladder cancer. Note also the confidence bands that become extremely wide after 80 BSM-years due to the small number of observations in that range.

In order to remove the somewhat arbitrary constraints on the placement of knots, a series of smoothing splines with different degrees of freedom was fit. 2, 3, 4, and 6 degrees of freedom were used. None of the smoothing spline fits captured the risk deficit that was detected by the categorical analyses, showing instead an almost monotonic increasing relationship between cumulative exposure and risk. The spline

Figure 3.4: Bladder - Regression Spline of Relative Risk as a Function of Cumulative Exposure



with 4 df is shown in Figure 3.5.

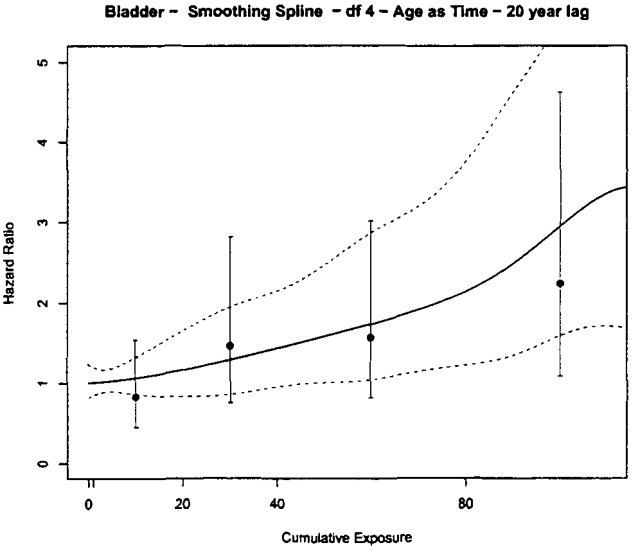
In order to check whether the proportionality assumption actually held for this data, the *cox.zph* function in *R*'s survival package was used. This function calculates tests of proportionality of hazards for each variable in the model as well as a global test, and the resulting object can be used to plot the Schoenfeld residuals against time. The resulting p-values for all the variables were non-significant, and the p-value for the overall test of non-proportionality was 0.961, which indicates no serious departure from the proportionality assumption.

3.6.2 Non-Hodgkins Lymphoma

The same three simple models that were fit to the bladder data were fit to the non-Hodgkins data. The estimated relative risks, along with 95% confidence intervals for the 4 fits are in Table 3.12.

It is immediately apparent that the confidence intervals here are much wider than those for bladder cancer, which is due to the small number of cases (27, compared to

Figure 3.5: Bladder - Smoothing Spline of Relative Risk as a Function of Cumulative Exposure



90 for bladder cancer).

Figure 3.6 shows the dose-response curves for the four simple models. The linear and logarithmic models fit rather poorly to the categorical model relative risks, with the linear underestimating all 4 relative risks, and the log model overestimating the hazard ratios except that for the highest exposure category. The square root model is clearly the best of the three, judging by the graph and by the p-values from the likelihood ratio tests in Table 3.12.

The five polynomial fits all resulted in a monotone increasing dose-response curve which severely overestimates the relative risk of the 40-80 BaP-year category as compared to the categorical Cox PH and Poisson regression fits. As with bladder, likelihood ratio tests were performed to determine whether adding higher-order terms lead to better fit; again, no significant improvements were found after the quadratic model. Figure 3.7 shows the quadratic polynomial, which fails to capture the drop in risk which seems to occur in the 40-80 BaP-year range.

Next, a regression spline was used to model the relationship between the log-hazard

Figure 3.6: NHL - Simple Models of Relative Risk as a Function of Cumulative Exposure

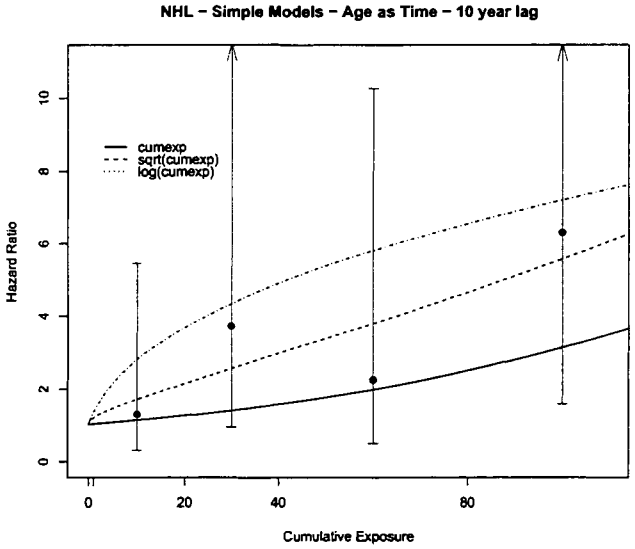


Figure 3.7: NHL - Quadratic Model of Relative Risk as a Function of Cumulative Exposure

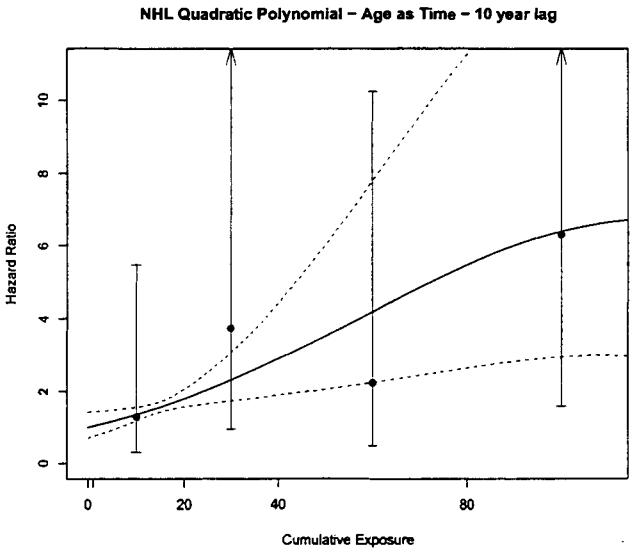


Table 3.12: Simple Cox PH Models for Non-Hodgkins Lymphoma

Fit	Covariate	Hazard Ratio	95% CI	LRT p-value
Linear	cumexp	1.001	0.994-1.008	0.0167
Logarithmic	Log(cumexp+1)	1.537	1.157-2.043	0.0066
Square root	Sqrt(cumexp)	1.189	1.072-1.319	0.0065
Categorical	0-0.5 BaP-years	1.00	-	0.0279
	0.5-20 BaP-years	1.29	0.30-5.46	
	20-40 BaP-years	3.73	0.95-14.64	
	40-80 BaP-years	2.24	0.49-10.26	
	80+ BaP-years	6.31	1.59-25.07	

and cumulative exposure, with calendar period included as a factor. The knots were placed at category cutpoints, and the resulting spline, along with 95% confidence bands is shown in Figure 3.8. The confidence bands are noticeably wider than for the polynomial fits, particularly for lower levels of cumulative BaP exposure.

The smoothing spline fits were very similar to the regression spline and polynomial models, except for an attenuation of risk that is shown for high levels of cumulative exposure. In the df=4 plot in Figure 3.9, the relative risk seems to plateau at about 100 BaP years and then begins to decline. This is caused by a small number of highly exposed individuals who were never diagnosed with NHL, just as for bladder cancer. This phenomenon is further discussed in Section 3.7. The *zph* p-value for this spline was 0.998.

3.6.3 Lung

The results of the four simple model fits are shown in Table 3.13. Here, the relationship between cumulative exposure and relative risk seems to be non-linear, although the confidence intervals are wide enough to allow the possibility of a linear dose-response curve. The fits shown in Figure 3.10 show that all the simple models severly underestimate the relative risk in the highest exposure category. None of the simple models appear to be a particularly good fit to the nonlinear categorical hazard ratio estimates, but all three fitted lines are contained within the four confidence bands.

A series of polynomials models was fit and once again, the quadratic polynomial

Figure 3.8: NHL - Regression Spline of Relative Risk as a Function of Cumulative Exposure

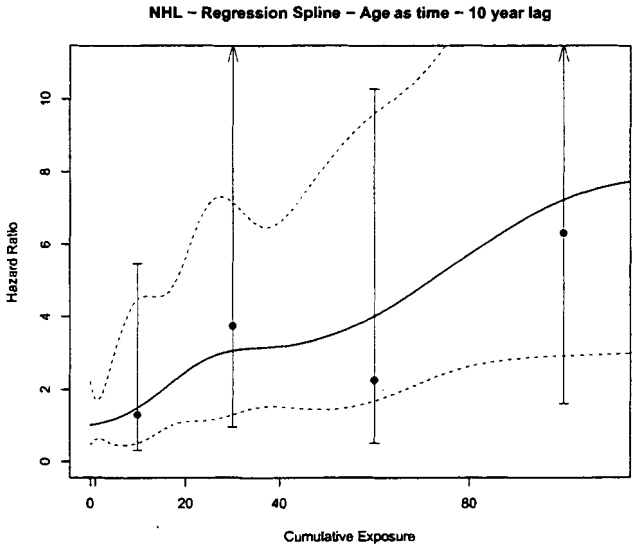


Figure 3.9: NHL - Smoothing Spline of Relative Risk as a Function of Cumulative Exposure

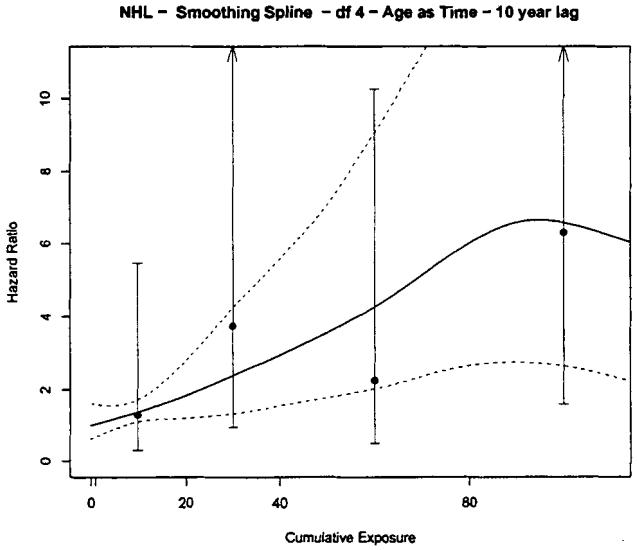


Table 3.13: Simple Cox PH Models for Lung Cancer

Fit	Covariate	Hazard Ratio	95% CI	LRT p-value
Linear	cumexp	1.006	1.004-1.009	0.00251
Logarithmic	Log(cumexp+1)	1.116	1.058-1.178	0.00258
Square root	Sqrt(cumexp)	1.063	1.034-1.093	0.00241
Categorical	0-0.5 BaP-years	1.00	-	0.0059
	0.5-20 BaP-years	1.39	0.92-2.11	
	20-40 BaP-years	1.42	0.80-2.53	
	40-80 BaP-years	1.32	0.73-2.42	
	80+ BaP-years		2.54	1.30-4.97

was chosen based on likelihood ratio tests, and is shown in Figure 3.11. This model underestimates the estimated hazard ratio in the 0.5-20 and 20-40 BaP-years categories, and has much narrower confidence intervals than these estimates as well.

The regression spline in Figure 3.12 captures the non-linearity of the categorical fit, although it slightly underestimates the relative risk in the 0.5-20 and 20-40 BaP-year exposure groups. Here the confidence bands are congruent with the confidence intervals obtained from the categorical estimates. The regression spline shows an attenuation of risk at about 100 BaP-years: there is a plateau at a relative risk of about 2.5, while the estimated risk in the quadratic polynomial continues to rise above 2.

The smoothing spline in Figure 3.13 differs from the regression spline in that it remains fairly flat, only beginning to rise after about 60 BaP-years. The spline is nearly monotonic though, which more closely correlates to the expected shape of the dose-response curve. The attenuation of risk that was noticed in the regression spline model is present here too, plateauing at about 100 BaP-years. The p-value for the *zph* test was 0.991.

3.7 Discussion

For the dose-response analysis by cumulative exposure categories, the Cox proportional hazard results were very similar to those obtained from Poisson regression. The largest discrepancies were observed for non-Hodgkins lymphoma, but these were

Figure 3.10: Lung - Simple Models of Relative Risk as a Function of Cumulative Exposure

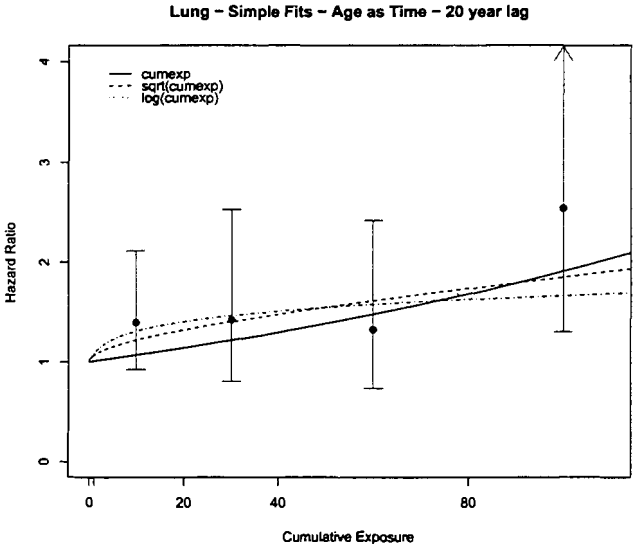


Figure 3.11: Lung - Quadratic Model of Relative Risk as a Function of Cumulative Exposure

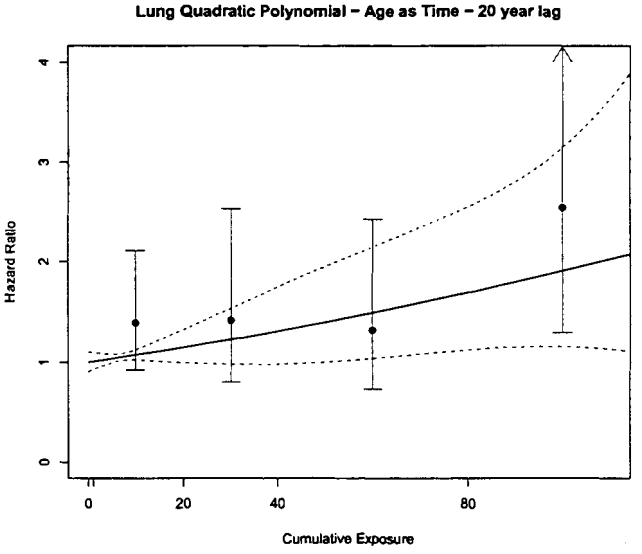


Figure 3.12: Lung - Regression Spline of Relative Risk as a Function of Cumulative Exposure

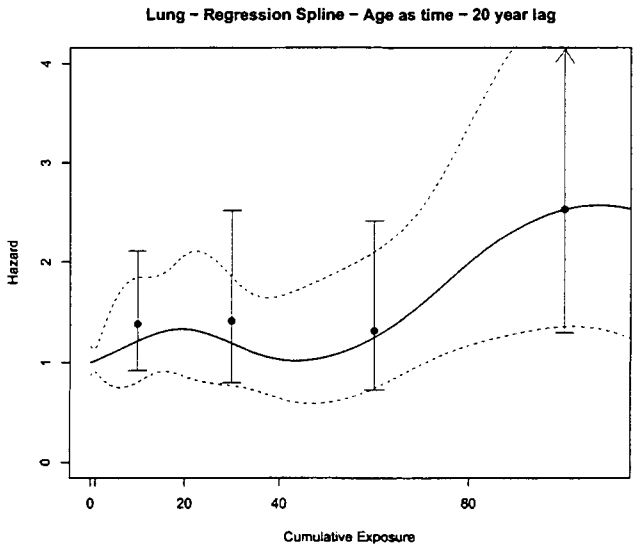
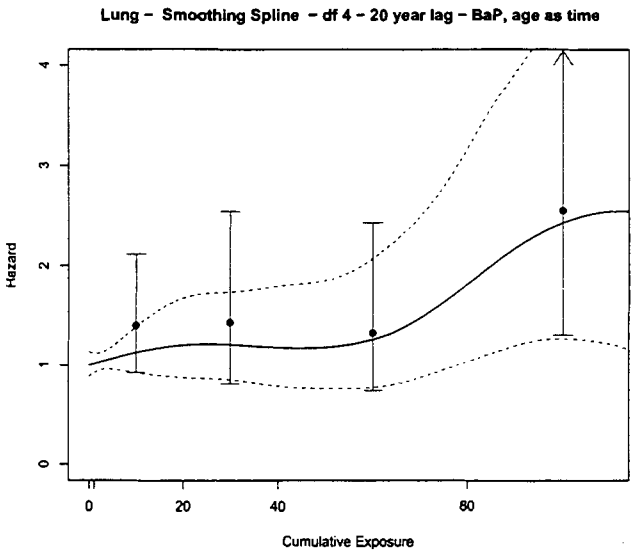


Figure 3.13: Lung - Smoothing Spline of Relative Risk as a Function of Cumulative Exposure



likely due to the small number of cases (27) which made confidence intervals for all analysis methods extremely wide. A comparison of the hazard ratios for the Poisson and Cox proportional hazards methods is presented in Table 3.14. Five cumulative BaP exposure values were chosen to obtain point estimates: 0, 10, 30, 60, and 100 BaP years. These roughly correspond to the midpoints of the cumulative exposure categories. For the three cancer sites of interest, the estimated hazard ratios for each BaP-year category compared to the baseline group are shown for the Poisson regression model and all seven Cox PH models that were shown earlier.

Table 3.14: Comparison of Hazard Ratio Estimates

Site	Model	Exposure Level				
		0	10	30	60	100
Bladder	Poisson regression	1.00	0.84	1.53	1.65	2.36
	Cox PH - Categorical	1.00	0.82	1.46	1.56	2.24
	Cox PH - linear	1.00	1.07	1.23	1.52	2.01
	Cox PH - Square Root	1.00	1.29	1.55	1.86	2.23
	Cox PH - Log-log	1.00	1.42	1.66	1.83	1.97
	Cox PH - Quadratic	1.00	1.17	1.52	2.04	2.47
	Cox PH - Regression Spline	1.00	0.77	1.41	1.56	2.92
	Cox PH - Smoothing Spline	1.00	1.12	1.20	1.26	2.42
NHL	Poisson Regression	1.00	1.43	4.58	2.94	8.21
	Cox PH - Categorical	1.00	1.29	3.73	2.24	6.31
	Cox PH - linear	1.00	1.12	1.41	1.98	3.11
	Cox PH - Square Root	1.00	1.73	2.58	3.81	5.63
	Cox PH - Log-log	1.00	2.80	4.38	5.85	7.27
	Cox PH - Quadratic	1.00	1.36	2.30	4.18	6.41
	Cox PH - Regression Spline	1.00	1.48	3.04	4.00	7.21
	Cox PH - Smoothing Spline	1.00	1.37	2.36	4.27	6.60
Lung	Poisson regression	1.00	1.34	1.48	1.45	2.26
	Cox PH - Categorical	1.00	1.39	1.42	1.32	2.54
	Cox PH - linear	1.00	1.07	1.21	1.47	1.91
	Cox PH - Square Root	1.00	1.21	1.40	1.61	1.85
	Cox PH - Log-log	1.00	1.30	1.46	1.57	1.66
	Cox PH - Quadratic	1.00	1.07	1.23	1.49	1.91
	Cox PH - Regression Spline	1.00	1.22	1.19	1.25	2.54
	Cox PH - Smoothing Spline	1.00	1.12	1.20	1.26	2.42

For the Cox proportional hazards analyses with cumulative exposure as a continuous variable, square root fit seemed to be the most reasonable of the simple models for all three sites of interest, giving the smallest LRT p-value for all three cancer sites. For bladder and lung, most of the hazard ratio estimates are similar for all models. There is much more variation in the estimates for non-Hodgkins lymphoma, likely due to the small number of cases. The estimates obtained from the logarithmic model differ greatly from the other models, with an estimated hazard ratio of 2.80 at 10 BaP-years. All other models estimate the hazard ratio at this exposure level as being no greater than 1.73.

The models that incorporated smoothing all captured the various features that had been suggested by the categorical fits, if enough degrees of freedom were used. This is shown in Figure 3.3, where the polynomial fit captures two "dips", or areas of negative slope. However, as the degrees of freedom increased, the polynomials and splines showed an attenuation of risk at high levels of cumulative exposure, and even a dramatic *drop* in risk for the most highly exposed individuals, which does not conform with any reasonable dose-response model. There are several possible reasons for the effect. There could be a group of individuals who are less susceptible to cancer (or diseases in general) for reasons not accounted for in the study; these individuals could accumulate very high amounts of exposure without developing the cancer in question or contracting some other disease which would cause them to stop working (and thereby no longer being exposed). There could also be a "saturation of effect" [7] which means that there is some threshold beyond which additional exposure carries no increase in risk of the cancer in question. Also, there is the possibility that some other unknown or unmeasured risk factor is affecting the highly exposed individuals differently than the less-exposed workers and is therefore altering the dose-response curve. Regardless of the reason, the small number of observed person-years with extremely high cumulative exposure are highly influential when fitting models, and can greatly distort the resulting dose-response curves. It is possible, of course, to refit the models and ignore the observations with extremely high cumulative exposure, but deciding on a cut-off point would be fairly arbitrary, and one could argue that it should not be necessary to ignore data in order to get a "satisfactory" model fit, so this option was not considered in this project.

It was also of interest to determine, for each model, the level of cumulative BaP exposure that would give an estimated hazard ratio (HR) of 2. This level is often considered when discussing matters of compensation, as a hazard ratio of 2 corresponds to an attributable risk of 50 %. This means that a cancer case arising at this level of cumulative exposure has a 50 % chance of having been caused by the exposure, rather than whatever other risk factors usually contribute to the development of the cancer. Table 3.15 shows the cumulative BaP levels corresponding to a hazard ratio of 2 for all three cancer sites and the six continuous Cox proportional hazards models. For bladder cancer, all models give a cumulative BaP exposure estimate of over 70 except the quadratic model. The results for non-Hodgkins lymphoma are fairly inconsistent, with widely varying estimates of the cumulative BaP level giving an HR of 2. Ignoring the linear and log fits, which seem to be an unreasonable fit for this data, the BaP-level giving a HR of 2 is estimated to be between 15 and 25 BaP-years. For lung cancer, all models give estimates that are within the highest exposure category, except for the log model: the estimated relative risk for this model does not rise above 2 in the observed cumulative exposure range.

Table 3.15: Cumulative Exposure Levels Giving a Relative Hazard of 2

	Model					
Site	Linear	Square Root	Log	Quadratic	Regr. Spline	Smooth. Spline
Bladder	99.28	75.10	110.05	58.01	71.29	85.14
NHL	61.45	16.15	4.05	24.42	15.55	23.28
Lung	107.92	128.97	NA	107.92	80.38	85.14

Both the Poisson regression and Cox proportional hazards models are based on the assumption that cumulative exposure and other factors work in a multiplicative manner. There are of course other possible models that may be a better fit to the data, notably additive and power models [3]. A logical next step would be to write programs that can fit a greater variety of models to occupational cohort data with time-dependent covariates and then compare the various methods, perhaps through a simulation study. However, in order for these future steps to be feasible, some of the data processing issues need to be resolved in a more time-efficient matter. The sheer amount of programming and computing time required to prepare the data for Cox

proportional hazards analysis is, at this point, out of proportion with the amount of information that is gained.

Chapter 4

Conclusion

Ultimately, the goal of all the methods of analysis in the project is to gain insight into the dose-response relationship. The categorical methods all let the data speak for itself to some extent, with various constraints and adjustment for factors such as age and calendar time. The major downfall of these methods is that they depend heavily on the choice of the category cutpoints and the number of categories, so important features of the dose-response curve may be obscured or exaggerated. Of course, the true underlying dose-response relationship is unknown, so it is difficult to assess how well the categorical analyses are capturing the relationship. However, because a shape is not imposed on the dose-response curve, these methods are commonly used as a preliminary analysis since they are usually quick to run and may reveal striking features that may not be picked up if one simply jumped straight to a parametric analysis.

The simple Cox proportional-hazards models impose a parametric form on the dose-response curve, providing interpretable parameters, an easily understandable functional form, and simple graphs that are monotonic. Of course, their usefulness is limited by how well they fit the data, as with any parametric model; if the imposed shape greatly differs from the observed data, any inferences made from the model could be misleading.

The Cox proportional-hazards polynomial and spline methods allow for a much more flexible fit to the data, but there is a danger that the features they reveal are more an artifact of the particular data set rather than a clear picture of the underlying

dose-response effect. They also require more computing time than the simple models to perform analysis. In addition, sparse data points in the upper end of the cumulative exposure range can have an exaggerated effect on the estimated curves in this area and can therefore distort the apparent dose-response effect. The spline methods have the disadvantage of not providing interpretable parameters, making inference and testing more difficult. However, their flexibility does allow them to reveal interesting features without the risks implicit in categorizing cumulative exposure. All of the smoothing models essentially served to confirm the dose-response relationships observed from the Poisson regression analysis. With few degrees of freedom, the resulting models were very similar to the simple parametric models; as the amount of smoothing increased, the small number of observations with very high exposure began to distort the upper regions of the graph, and the resulting graphs began to look less reasonable as dose-response models to base inference on.

Ultimately, the most desirable outcome of an occupational cohort study is a fairly simple parametric model that adequately captures the relationship between exposure and cancer risk. Simple models are easy to interpret, and can easily be communicated to others in the form of parameter estimates and graphs. It is also generally believed that, in most cases, the "true" underlying dose-response relationship is something inherently simple. Therefore the most reasonable course of action for this type of data seems to be to first use a categorical analysis with *a priori* categories (Poisson regression or Cox proportional hazards) as a first assessment of the data, and then fit a variety of simple parametric models or smoothing models with very low degrees of freedom. Breslow [2] advocates exploring a class of models that are reasonable and consistent given the data, and to avoid selecting one "best" model unless understanding of the underlying process leads to a definitive choice. In this project, a wide variety of models have been presented and compared for one particular dataset, with no one model being clearly superior to the others. The collection of analyses revealed more about the true underlying relationship than any one model could have shown, and some reasonably consistent conclusions could be drawn.

Bibliography

[1] B. G. Armstrong et al. Estimating the relationship between exposure to tar volatiles and the incidence of bladder cancer in aluminum smelter workers. *Scandinavian Journal of Work and Environmental Health*, 12(5):486–93, 1986.

[2] N. Breslow. Biostatistics and Bayes. *Statistical Science*, 5:269–298, 1990.

[3] N.E. Breslow and N.E. Day. *Statistical Methods in Cancer Research, Volume II: The Design and Analysis of Cohort Studies.* Oxford University Press, Oxford, 1987.

[4] M.C. Friesen et al. From expert-based to quantitative retrospective exposure assessment at a soderberg aluminum smelter. *to appear*, 2004.

[5] J. F. Lawless. *Statistical Models and Methods for Lifetime Data, Second Edition.* Wiley Series in Probability and Statistics. John Wiley & Sons, New Jersey, 2003.

[6] J.J. Spinelli et al. Mortality and cancer incidence in aluminum reduction plant workers. *Journal of Occupational Medicine*, 33(11):1150–1155, 1991.

[7] L. Stayner et al. Attenuation of exposure-response curves in occupational cohort studies at high exposure levels. *Scandinavian Journal of Work and Environmental Health*, 29(4):317–324, 2003.

[8] K. Steenland and J. Deddens. A practical guide to dose-response analyses and risk assessment in occupational epidemiology. *Epidemiology*, 15(1):63–70, 2004.

[9] T.M. Therneau and P.M. Grambsch. *Modeling Survival Data: Extending the Cox Model.* Statistics for biology and Health. Springer, New York, 2000.