The Application of the Case-Cohort Method to Data on Pulp and Paper Mill Workers in British Columbia

by

Jacqueline S. Gregory

B.Sc., University of Victoria, 2001

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

SIMON FRASER UNIVERSITY
September 2003

APPROVAL

Name: Jacqueline S. Gregory

Degree: Master of Science

Title of project: The Application of the Case-Cohort Method to Data on Pulp and Paper Mill Workers in British Columbia

Examining Committee: Boxin Tang
Chair

Dr. Randy R. Sitter
Senior Supervisor
Simon Fraser University

Dr. Brad McNeney
Simon Fraser University

Dr. Carl J. Schwarz
External Examiner
Simon Fraser University

Date Approved: _____

Abstract

There are two common methods for comparing disease incidence rates (such as cancer) in two populations (such as pulp and paper workers vs non-pulp and paper workers). In cohort studies, the two groups are followed over time and the incidence rates are directly compared. These types of studies can be inefficient for low incidence diseases when very large sample sizes are needed. Case-control methods take each incidence of disease and match it to a control. Then contributions from variables such as exposure to chemicals to the disease incidence rate can be determined. While more efficient than cohort studies, direct incidence rates cannot be computed.

This thesis used a newly proposed method, the case-cohort study, that combines features of both types of studies. Because it uses two cohorts, it uses more information than the case-control study but also gains efficiency from the matching of cases with controls.

While this method has been extensively theoretically developed in the literature, it has only be applied to simple problems or simulations. We used this new method to reanalyze a long running study conducted by the British Columbia Cancer Agency. While the new methodology did not give dramatically different results, it did yield improved precision in estimates (implying that it will be easier to detect excess disease rates). Some potential dangers in the uncritical use of this method were also identified.

Acknowledgements

I would first like to thank Dr. Nhu Le from the British Columbia Cancer Agency. Nhu provided the idea for this project and was always there to answer my many questions. I would also like to thank my supervisor, Dr. Randy Sitter, for all of his help and support during my two years at SFU. In addition, I would like to thank my examining committee, Dr. Brad McNeney and Dr. Carl Schwarz for their useful comments and criticisms of my work.

I want to thank all my friends and fellow students who made my experience at SFU enjoyable. There are a few people I would like to give special thanks to. David Beaudoin and I struggled through assignments together and were always there to keep each other sane. Simon Bonner was always there to listen to me, and he made a great dance partner. I could always count on Michael Lo to keep me company on icq and share the occasional martini. They are all great friends.

Special thanks to my parents who have always supported me, and to my sister, and best friend, Jill, who never failed to make me smile. Finally, I would like to thank Jason Cumiskey for being there through the highs and the lows.

Contents

List of Tables

Chapter 1

Introduction

In epidemiology, a *cohort* is generally used to designate a group of people who share a common experience or condition. Epidemiological studies often involve the follow-up of a large cohort of subjects, a small fraction of whom will develop a disease at an endpoint, or endpoints of interest during a prescribed follow-up period.

In 1982, an occupational cancer research program was initiated in British Columbia; one facet of this ongoing project was aimed at detecting occupational cancer risk factors. One of the studies was based on collecting lifetime occupational history from male incident cancer patients, aged 20 or older, ascertained from the British Columbia Cancer Registry between January 1, 1983 and December 31, 1989. Based on this preliminary analysis a two-phase study of British Columbian pulp and paper workers was initiated.

Chapter 2 will start with an introduction to some terminology that is common in epidemiology. The main objective of Chapter 2 is to describe the designs and the methods of analyses that are used in to analyse the data on the pulp and paper mill workers of British Columbia in Chapter 3. Section 2.1 will describe the cohort design. The objective of this section is solely to develop a background of the work that has already been done on the British Columbian pulp and paper workers. Section 2.2 begins with a brief description of the case-control design and how it compares to the cohort design. The nested case-control design is introduced in the general case. However, the focus of this section is the matched case-control design since it is used

in the application described in Chapter 3. In this section the design, the method for analysing and the problems with the design are described in detail. Finally, section 2.3 discusses the case-cohort design, which is the design of most interest for this project. The design is introduced in detail for both binary response data and time to response data, the latter being the most relevant for the application in Chapter 3. To conclude the chapter, the Cox Proportional Hazards model is introduced so that it can be included in the description of how to compute the maximum pseudolikelihood estimator of the case-cohort design.

Chapter 3 describes the two-phase study of the British Columbian pulp and paper workers. Section 3.1 and 3.2 give an overview of the chapter and a background to the two-phase study. Section 3.3 recounts Phase I which investigated the cohort's mortality and cancer incidence outcomes. In this first phase of the study of British Columbia pulp and paper workers, no attempt was made to classify workers by departments and no exposure data were obtained that might provide explanations for the differences in cancer patterns observed between workers at mills running different processes. Phase II, which is described in section 3.4, was a matched case-control study with detailed work history and exposure assessment based on mill-specific job exposure matrices. The aim of this project is to apply the case-cohort method, first proposed by Prentice (1986), to this complicated real data situation. The data collected included enough information to analyse as a case-cohort design, but part of the data was ignored so as to treat as a matched case-control design. We first re-analyse the data as a matched case-control for a single chemical, and then re-analyse it as a case-cohort design using all available information. The two analyses are compared and contrasted, in addition the ease of application and stability of the case-cohort analysis is explained.

Chapter 2

Methodology

In an observational study there is no manipulation of the study factors by the investigator. In other words, the investigator has no control over doses, treatments or exposures.

Before starting the discussion on the different observational designs, it is important to carefully define the terminology that will be used throughout this project. The *source population* (or *cohort*), though sometimes referred to as a population, is a sample which represents a hypothetical study population in which a cohort study may have been conducted; it is this hypothetical population that one wishes to make inferences about. For example, one may use the 14 paper and pulp mills in British Columbia as a source population, but it is actually viewed as a sample of the hypothetical population of all the mills where particular chemicals of interest are used. From this example, it is clear that the sample is not random, and often this is the case. Sometimes it is not possible, or too expensive (with respect to time and money), to take a random sample. The source population is treated as a random sample so that inferences can be made about the entire population.

Table 2.1 gives a depiction of this project's scenario, where the source population is represented by $A_1 + B_1 + A_0 + B_0$. Within the source population there are *subcohorts* or *groups*: an *exposed group* $(A_1 + B_1)$ and an *unexposed group* $(A_0 + B_0)$. It is possible to have more than two groups; however, for this project, we will restrict to two groups. In addition, there is the *case group* $(A_1 + A_0)$, which represents the

diseased individuals, and the *control group* ($B_1 + B_0$), which represents the non-diseased individuals.

	Disease	Non-Disease	
Exposed	A_1	B_1	$A_1 + B_1$
Unexposed	A_0	B_0	$A_0 + B_0$
	$A_1 + A_0$	$B_1 + B_0$	

Table 2.1: Depiction of the Source Population

There are two primary types of observational designs in epidemiology: the cohort design and the case-control design. Table 2.2 compares the characteristics of these designs.

Cohort	Case-Control
Begins with a defined population at risk	Generally undefined population at risk
Cases not selected but ascertained by continuous surveillance	Cases selected by investigator from an available pool of patients
Comparison group (i.e., non-cases) not selected - evolved naturally	Controls selected by investigator to resemble cases (matching on auxiliary variables)
Exposure measured before the development of disease	Exposure measured, reconstructed or recollected after development of disease
Risk or incidence of diseases and relative risk measured	Risk or incidence of disease cannot be measured directly; relative risk exposure can be estimated by odds ratio

Table 2.2: Comparison of characteristics cohort and case-control study designs

A major difference between the cohort design and the case-control design is who is being compared. The cohort design looks at exposed versus unexposed, whereas

the case-control design is interested in diseased versus non-diseased. In the cohort approach, sampling is based on exposure whereas in the case-control approach sampling is based on outcome (disease or not). A cohort study uses all individuals in the source population. In a case-control study most cases (diseased) occurring in the source population and only a random sample of the control (non-diseased) group are selected. One can view the case-control design as biased sampling, with over-sampling of cases. This makes case-control studies more efficient: one does not have to study all persons in the source population who do not develop the disease but only a small sample from them. Unfortunately, this sampling scheme hampers computing any direct measure of risk, because the resulting sample of cases and controls is not proportional to the number of cases and non-cases in the underlying source population. This is the main difference between the two designs.

Both cohort and case-control designs measure frequency, but in cohort studies the frequency of different outcomes is measured, while in case-control studies the frequency of the presumed causal factors is measured. In cohort studies, risk can be expressed as relative risk (risk ratio) and attributed risk (risk difference). In case-control studies risk is expressed as an odds ratio.

The remainder of this chapter describes and compares the designs in detail and some methods used to analyse them. Then a new design, proposed by Prentice (1986), is introduced as an alternative.

2.1 Method I: Cohort Study

In a cohort study the primary question addressed is, "What are the health effects of a given exposure?"

Long term follow-up (cohort) studies of human populations, particularly of industrial workers, have provided the most convincing evidence of the link between exposure to specific environmental agents and cancer occurrence. In epidemiology, the word cohort is often used to designate a group of people who share a common experience or condition. In other words, a cohort is simply a group of persons who

have presumed antecedent characteristics in common and who are followed throughout their experience so that one may observe the development or non-development of a given health outcome. For example: (i) all first year students in a university during a particular academic year, or (ii) all the gall-bladder patients who were operated on in a given hospital during a certain period of time.

Often, if there are two groups in the study, one of them is described as the exposed group - those individuals who have experienced the potential causal event or condition - and the other is thought of as the unexposed, or reference, group. If there are more than two groups, each may be characterised by a different level or type of exposure. For example, an occupational cohort study of chemical workers might comprise sub-cohorts of workers in a plant who work in different departments of the plant, with each sub-cohort being exposed to a different set of chemicals. The investigator measures and compares the incidence rate of the disease in each of the study groups.

Exposed and unexposed groups at one point in time are then followed to assess the differences in health outcomes between them. Follow-up from exposure to outcome is the key feature of a cohort study; it gives assurance about the sequence of events, namely the occurrence of exposure prior to outcome, a basic requirement to infer causality.

In a cohort study, the investigator controls neither the exposure conditions nor the attribution of exposure to study subjects; the subjects in the cohort are selected after exposure status has been characterised. As a result, risk factors of the health outcome are likely to be unevenly distributed between the exposed and unexposed groups leading to differences in baseline risk. To ensure relative comparability between the exposed and the unexposed subjects, the investigator can only control the selection of the unexposed group.

There are two types of cohort studies: prospective cohort studies and retrospective (historical) cohort studies. The primary difference between these two studies is the way in which the follow-up over time is conducted. The prospective cohort method assembles the cohort in the present, and follows the individuals prospectively into the future. The investigator assesses exposures in the present and watches for disease in the future. A source population is generally a “representative” sample of the

hypothetical population; this sample may be a random sample, or it may be based on something, such as exposure. The main advantage to this method is that it allows one to collect exactly the information thought to be required; however, it does have the disadvantage that many years may elapse before sufficient cases of disease have developed for analysis. In contrast, the retrospective cohort study allows one to identify a group with certain exposure characteristics, by means of historical records, at a certain defined time in the past, and then reconstruct the disease experience of the group between the defined time in the past and the present. In addition, in the retrospective cohort design, like the prospective cohort design, sampling is not based on case/disease status. The main advantage is that results are potentially available immediately, and the disadvantage is that the information available on the cohort may not be completely satisfactory, since it would most likely have been collected for other purposes or be subject to recall bias. Prospective studies, although more accurate, are costly and often impractical due to their time requirement. Retrospective studies are more frequently used as they are faster and cost less. The two types of studies have a fundamental characteristic in common: the individuals comprising the cohort are identified, and information on their exposure obtained, before their disease experience is ascertained (Breslow and Day, 1987). The goal of both studies is to compare exposed and unexposed individuals.

The design and execution of a cohort study will depend on the individual circumstances of the study, and its aim. Even though the scope and purpose of different studies may vary widely, there are a number of issues in the design and execution that require attention, irrespective of whether the study is prospective or historical. These issues are as follows:

- Inclusion rules must be clear and unambiguous.
- Dates of entry and exit must be well defined.
- Follow-up over time of the individuals enrolled in the cohort study is the essential feature of the study; thus the follow-up mechanisms to be used must be chosen carefully.

- The extent and detail of the information on exposure should reflect the relationship between exposure and excess risk that the investigator might expect. In addition one requires to know: (i) the dates at which exposure started and stopped, as well as the subject's age when exposure started, and (ii) in relation to exposure level, quantitative information is rarely available throughout the period. Thus one has to decide which summary measures are most informative.

- It is important to collect information on any auxiliary variables that may have an effect.

- The possible results the study could yield need to be investigated before substantial resources are devoted to the study. Studies that have low power for detecting realistic levels of excess risk should not be performed, unless their results can be merged with those of other studies.

2.1.1 Analysis of the Cohort Study

The following section gives the simplest form of analysis of the cohort study. The object of the section is to give a background to the application discussed in detail in Chapter 3 and not to describe all possible analysis methods.

Analysis of data from a cohort study involves estimation of the rates of cancer and other diseases of interest which occur among cohort members during the study period. Cohort studies, by recording disease occurrence in a defined group, provide measures of incidence, or mortality rates, and it is these rates that provide the basic measures of disease risk. Analysis of cohort data typically involves a comparison of the rates observed in the study group with rates for the general population. This is a useful way of identifying diseases which occur at especially high or low frequency in the cohort, so they may be studied further in relation to particular exposures.

Two measures of effect are used in cohort studies: the incidence (or mortality) rate ratio which is the incidence rate or outcome in the exposed group relative to the unexposed one; and the risk ratio or relative risk which is the proportion of the exposed cohort developing the health outcome of interest relative to the unexposed

one.

	Disease	Non-Disease	
Exposed	A_1	B_1	$A_1 + B_1$
Unexposed	A_0	B_0	$A_0 + B_0$
	$A_1 + A_0$	$B_1 + B_0$	

Table 2.3: Two by Two Contingency Table For Calculating Risk

From Table 2.3, the Relative Risk (RR) is:

$$\begin{aligned} RR &= \frac{\text{probability of disease given exposed}}{\text{probability of disease given unexposed}} \\ &= \frac{A_1/(A_1 + B_1)}{A_0/(A_0 + B_0)}. \end{aligned}$$

Furthermore, in many cohort studies, standardized mortality ratios (SMR) are used to compare the mortalities. This index is the ratio of the rate of mortality of disease among the worker group, to the rate of mortality among some reference group. Also, standardized incidence ratios (SIR) are used to compare cancer incidences. This index is the ratio of the rate of mortality and incidence of disease among the worker group, to the rate of incidence among the reference group.

More complex modelling is used when analysing cohort studies; however, since such was not done to analyse the cohort phase of the application in Chapter 3, it will not be discussed in this project.

2.2 Method II: Case-Control Study

2.2.1 Description

In a case-control study the primary question addressed is, "What are the contributing causes of a given disease?" Case-control studies are the most frequently used epidemiology study design. They examine the cause-effect relationship from a perspective opposite to that of a cohort study.

Consider the basic case-control study design. Imagine two sub-cohorts, exposed and unexposed, that can be denoted by the subscripts 1 and 0, respectively. Now, suppose that we want to study the relationship of exposure incidence rates in these populations. The disease incidence rate during a time period t (e.g. 1 year) might be expressed for the exposed group as

$$I_1 = \frac{A_1}{T_1}$$

and for the unexposed group as

$$I_0 = \frac{A_0}{T_0},$$

where A_1 and A_0 are the respective numbers of individuals in whom disease developed during time interval t, and T_1 and T_0 are the respective amounts of person-time at risk of the disease spent in the exposed and unexposed groups, and thus I_1 and I_0 are the incidence rates for the exposed and unexposed and are estimates of the rates of disease and non-disease given the exposure in the hypothetical study population.

In a cohort study, the numerator and the denominator of each rate are evaluated; doing so requires enumerating the source population and keeping it under surveillance. A case-control study attempts to observe the source population more efficiently. The efficiency of the case-control study comes from the use of a control series in place of complete assessment of the denominators of the incidence rates. The cases in a case-control study should be the same individuals who would be considered cases in a hypothetical cohort study of the same source population; using the notation above, the cases are the $A_1 + A_0$ individuals.

Case-control studies are best understood by defining a source population, a sample which represents a hypothetical study population, in which a cohort study might have been conducted. If a cohort study were undertaken, the primary tasks would be to identify the exposed and unexposed denominator experience, measured in person-time category or study cohort. In a case-control study, the cases are identified and their exposure status is determined just as in a cohort study, but denominators from which rates could be calculated are not measured. Instead a control group of study subjects

is sampled from the non-diseased sub-cohort.

The purpose of the control group is to estimate the relative (as opposed to the absolute) size of the exposed and unexposed denominators within the source population, i.e. T_0/T_1. From the estimated relative size of the denominators, the relative size of the incidence rates (or incidence proportions) can be estimated, since

$$\frac{I_1}{I_0} = \frac{A_1}{T_1} \cdot \frac{T_0}{A_0} = \frac{T_0}{T_1} \cdot \frac{A_1}{A_0},$$

and A_0, A_1 and an estimate from the sub-sample of T_0/T_1 are available (related to two-phase sampling in surveys).

Thus, case-control studies yield estimates of relative effect measures. Because the control group is used to estimate the distribution of exposure in the source population, the cardinal requirement of control selection is that the controls must be sampled independently of their exposure status.

Case-control studies first identify and select the cases and controls; these groups are then followed backward in time to assess whether their retrospective past patterns of exposure differed before the cases actually developed the health outcome. Tracking backward from outcome to antecedent is characteristic of case-control studies; it is inferred that differences in exposure patterns between cases and controls are likely a cause of the outcome.

A cohort study faces forward in time (whether collected prospectively or retrospectively), starting with a defined population and its exposure status, and observing the subsequent disease experience, whereas a case-control study faces backwards in time, starting with the disease status, and reconstructing the exposure history from which it emerged.

Usually, all cases occurring in the population of interest are included in the study, but only a fraction of the potential controls are selected. This makes case-control studies more cost effective: one does not have to study all persons in the source population who do not develop the disease but only a small sample from them. Unfortunately, this sampling scheme hampers computing any direct measure of risk, because the resulting sample of cases and controls is not proportional to the number of cases and

non-cases in the underlying source population.

As a final comment, if a case-control study is nested within a defined cohort, it is referred to as a nested case-control study. Using this definition, nested-case control studies are often used in occupational epidemiological studies.

2.2.2 Comparison with the Cohort Study

In the present section, the relative merits and drawbacks of the cohort study as compared to the case-control study are discussed. In the cohort approach a group of individuals is defined, their exposure determined and their subsequent disease experience ascertained, whereas in the case-control approach, the cases of a specific disease are identified together with a suitable comparison group, and information on exposure before disease onset obtained retrospectively. Described in this way, it would seem natural that the latter might appeal if the focus is on causation of a specific disease, and the former if interest is on the health consequences of a given exposure (Breslow and Day, 1987).

There are many reasons that (and situations for which) the cohort design is more appealing than the case-control study. One such feature is that the results of the cohort study are considered more conclusive than results from case-control studies. Another important issue is bias. The cohort study has a lower potential for bias than the case-control study. In cohort studies recall-bias and selection bias can be eliminated, whereas in case-control studies recall bias can cause major problems and selection bias is almost impossible to evaluate. Another advantage of the cohort approach is that it is good for establishing the temporal sequence and the natural history of diseases. In contrast, the case-control approach cannot assess temporal relationships because: i) it is hard to decide when a disease was actually acquired; ii) because the controls may be "at risk" longer than the cases, it is possible to obtain a nonsensical result that exposure decreases an individual's chance of being diagnosed with the disease; and iii) the case-control design misses diseases still in a latent period. A final advantage is that the cohort design can estimate overall and specific disease rates, usually incidence rates. In contrast the case-control approach cannot calculate

incidence; in addition, it cannot calculate population relative risk or attributable risk.

Case-control studies have predominated in the history of cancer epidemiology (Breslow and Day, 1987). This would suggest that there are several disadvantages to the cohort design, despite the advantages discussed thus far.

The following is a discussion of the reasons the case-control design is more appealing than the cohort design. Time and money are a very important features in designs. A major disadvantage to the cohort study is that it becomes stronger the longer the study continues. Therefore, a cohort study may lead to a commitment over many years, which can in turn be a very expensive operation. In contrast, a case-control study is inexpensive and can be accomplished quickly because events of interest have already occurred.

The case-control study is more appealing than the cohort study when the disease of interest is a rare disease. Recall that in the case-control study the proportion of cases and non-cases is not the same as in the underlying population; however, in a cohort study the proportion of the source population being diseased is the same as in the population, which may cause problems if the disease is rare. If the ratio of cases to controls is low, then the cohort will have a much higher sample size than the case-control design. This makes the latter more appealing. This is the main application and the main reason case-control studies are so popular. A few other advantages of the case-control are: (i) it can study several potential exposures at the same time and (ii) it lends itself well to hospital-based studies and outbreaks.

To conclude, it is important to mention two features that the two designs share in common. First, in both designs, inferences can be biased due to confounders. A confounder is any circumstance, other than the desired exposure, that makes one group different than another; the confounder must also be associated with disease outcome. Confounding can be protected against through random selection. Second, both allow for inference when a randomised clinical trial would be unethical. For example, if one is interested in the effect of exposure to chemicals on cancer, it is unethical to randomly assign individuals to that exposure.

2.2.3 Stratified Case-Control Study

In the general situation of the stratified case-control design the cases are divided into strata based on some auxiliary variable, e.g. age ranges. The controls are then assigned to the appropriate stratum and a stratified sample of controls is taken. The case-control study described previously is a special case of the stratified case-control design with only one stratum. The situation that we are interested in (i.e. the one used in the British Columbia pulp and paper study) involves stratifying so deeply that there is one case in each stratum and M controls called matched case-controls. As in the general situation, the M controls are matched to each case based on some auxiliary variable, such as age. In practice, it is difficult to stratify so deeply that there is only one case in each stratum. For example, if the auxiliary variable used for matching is age (in years), it is quite likely that there will be more than one case at each age. In this situation the controls that also fall in that stratum are randomly matched to cases and treated as matched.

The key element of a stratified case-control design is that the controls only need to be followed to the time that their matched case obtains the disease.

2.2.4 Analysis of the Matched Case-Control Study

Since the case-control study and general stratified case-control design were not done in the application in Chapter 3, the methods of analysis will not be discussed. For this project, we will confine ourselves to the method that was used in this application, the matched case-control study.

Matched case-control studies are typically analysed using conditional logistic regression for matched sets. Conditional logistic regression is used to investigate the relationship between an outcome and a set of prognostic factors; it is a common method for analysing a case-control study. The conditional approach is best restricted to matched case-control designs, or to similar situations involving very fine stratification where its use is in fact essential in order to avoid biased estimates of relative risk (Breslow and Day, 1980).

One design which occurs often in practice, and for which the conditional likelihood takes a particularly simple form, is the situation where each case is individually matched to one or more controls. The number of controls can be a fixed number, M, or it can vary from set to set.

Suppose that the i^{th} of I matched sets contains M_i controls in addition to the case. Denote the K-vector of exposures for the case in this set by $\boldsymbol{x}_{i0} = (x_{i01}, ..., x_{i0K})$ and the exposure vector for the j^{th} control ($j = 1, ..., M_i$) by $x_{ij} = (x_{ij1}, ..., x_{ijK})$. Now, we want to develop the the conditional likelihood.

Consider the binary dependent variable y, which indicates whether ($y = 1$) or not ($y = 0$) an individual develops the disease, and a series of independent regression variables $\boldsymbol{x} = (x_1, ..., x_K)$. The conditional probability formula for y given x is modelled as

$$Pr\{y = 1|\boldsymbol{x}\} = \Big\{1 + exp(-\alpha - \sum_{1}^{K} \beta_k x_k)\Big\}^{-1}.$$

Now, we need to take into account the matched sets as described above. In this case, the α's are allowed to vary from stratum to stratum. However, the β's remain fixed so that

$$Pr\{y = 1 \text{ in stratum } i|\boldsymbol{x}\} \;=\; \Big\{1 + exp(-\alpha_i - \sum_{1}^{K} \beta_k x_k)\Big\}^{-1}. \tag{2.1}$$

In order to account for this fact in the probability model, it is appropriate to consider the conditional probability of the retrospective data given the $M_i + 1$ sets of values for the x variables which are sampled in each stratum. More precisely, suppose it is known that $M_i + 1$ data vectors $\boldsymbol{x}_{ij}$ for $j = 0, 1, ..., M_i$ are observed in the i^{th} stratum, but it is not known which of these corresponds to the case. The conditional probability that the first vector corresponds to the case, as observed, and the remainder to the controls is

$$\frac{Pr_i\{\boldsymbol{x}_{i0}|y = 1\} \prod_{j=1}^{M_i} Pr_i\{\boldsymbol{x}_{ij}|y = 0\}}{\sum_{j=0}^{M_i} \left[Pr_i\{\boldsymbol{x}_{ij}|y = 1\} \prod_{j' \neq j}^{M_i} Pr_i\{\boldsymbol{x}_{ij'}|y = 0\}\right]}. \tag{2.2}$$

Each conditional probability $\Pr(\boldsymbol{x}|y)$ of risk factor values may be expressed as

$$pr(\boldsymbol{x}|y) = \frac{pr(y|\boldsymbol{x})pr(\boldsymbol{x})}{pr(y)}, \tag{2.3}$$

Now, substituting (2.1) and (2.3) into (2.2),

$$\frac{\frac{Pr_i(y=1|\boldsymbol{x}_{i0})Pr_i(\boldsymbol{x}_{i0})}{Pr_i(y=1)} \cdot \prod_{j=1}^{M_i} \frac{Pr_i(y=0|\boldsymbol{x}_{ij})Pr_i(\boldsymbol{x}_{ij})}{Pr_i(y=0)}}{\sum_{j=0}^{M_i} \frac{Pr_i(y=1|\boldsymbol{x}_{ij})Pr_i(\boldsymbol{x}_{ij})}{Pr_i(y=1)} \cdot \prod_{j' \neq j} \frac{Pr_i(y=0|\boldsymbol{x}_{ij'})Pr_i(\boldsymbol{x}_{ij'})}{Pr_i(y=0)}}$$

$$= \frac{Pr_i(y=1|\boldsymbol{x}_{i0})Pr_i(\boldsymbol{x}_{i0}) \cdot \prod_{j=1}^{M_i} Pr_i(y=0|\boldsymbol{x}_{ij})Pr_i(\boldsymbol{x}_{ij})}{\sum_{j=0}^{M_i} Pr_i(y=1|\boldsymbol{x}_{ij})Pr_i(\boldsymbol{x}_{ij}) \cdot \prod_{j' \neq j} Pr_i(y=0|\boldsymbol{x}_{ij'})Pr_i(\boldsymbol{x}_{ij'})}$$

$$= \frac{\frac{exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{i0k})}{1+exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{i0k})} Pr_i(\boldsymbol{x}_{i0}) \prod_{j=1}^{M_i} \frac{1}{1+exp(\alpha + \sum_{k=1}^{K} \beta_k x_{ijk})} Pr_i(\boldsymbol{x}_{ij})}{\sum_{j=0}^{M_i} \frac{exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{ijk})}{1+exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{ijk})} Pr_i(\boldsymbol{x}_{ij}) \prod_{j' \neq j} \frac{1}{1+exp(\alpha + \sum_{k=1}^{K} \beta_k x_{ij'k})} Pr_i(\boldsymbol{x}_{ij'})}$$

$$= \frac{exp(\alpha_i) exp(\sum_{k=1}^{K} \beta_k x_{ijk}) \prod_{j=0}^{M_i} \frac{1}{1+exp(\alpha + \sum_{k=1}^{K})} Pr_i(x_{ij})}{\sum_{j=0}^{M_i} exp(\alpha_i) exp(\sum_{k=1}^{K} \beta_k x_{ijk}) \prod_{j=0}^{M_i} \frac{1}{1+exp(\alpha + \sum_{k=1}^{K})} Pr_i(x_{ij})}$$

$$= \frac{exp(\sum_{k=1}^{K} \beta_k x_{i0k})}{\sum_{j=0}^{M_i} exp\big(\sum_{k=1}^{K} \beta_k x_{ijk}\big)}$$

Thus, the the conditional likelihood for all strata is:

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{I} \frac{exp(\sum_{k=1}^{K} \beta_k x_{i0k})}{\sum_{j=0}^{M_i} exp\big(\sum_{k=1}^{K} \beta_k x_{ijk}\big)} \\ &= \prod_{i=1}^{I} \frac{1}{1 + \sum_{j=0}^{M_i} exp\big(\sum_{k=1}^{K} \beta_k (x_{ijk} - x_{i0k})\big)}, \end{aligned}$$

where $\boldsymbol{\beta}^T = (\beta_1, ..., \beta_K)$. If any of the x's are matching variables, taking the same value for each member of a matched set, then their contribution to the likelihood is zero; therefore, the corresponding β cannot be estimated. This is a reminder that matched designs preclude the analysis of relative risk associated with the matching variables; however, by defining some interaction or cross-product terms involving both risk factors and matching variables, one may model how relative risk changes from one matched set to the next (see Breslow and Day, 1980).

If there is a single matched control per case, the conditional likelihood simplifies even further to

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{I} \frac{1}{1 + exp\big(\sum_{k=1}^{K} \beta_k (x_{i1k} - x_{i0k})\big)}.$$

This may be recognised as the unconditional likelihood for the logistic regression model where the sampling unit is the pair and the regression variables are the differences in the exposures for case versus control.

Familiar statistical packages, such as SAS, are available to perform conditional analysis for both matched and, more generally, stratified designs.

2.2.5 Problems with the Matched Case-Control Study

There are several reasons for considering alternatives to a matched case-control design. To begin with, the alignment of each selected control subject to its matched case seems inefficient. Why? Because the subject may also properly serve as a member of the comparison group for cases occurring at a range of other times. In addition, the strict application of the time-matched case-control approach would involve the selection of a new set of controls for each distinct disease category under study, whereas intuitively a single comparison group should suffice as in full-cohort analyses (Prentice, 1986).

The method for analysing a stratified case-control design is as follows. The individuals are followed through time, and considered "at risk", until they experience the event (diagnosed with cancer) or they are censored (leave the study or the study terminates). The cases are the individuals who are diagnosed with the disease and

the controls are the individuals who are disease-free at the end of the study. Since the controls remain in the study longer than the cases, they are exposed for longer; therefore, this would potentially cause a bias against the individuals who do not experience an event. Thus, if one were to follow the controls until the end of the follow-up period, it would be possible to obtain a nonsensical result that exposure decreases an individual's chance of being diagnosed with the disease. To avoid such a scenario, the controls are only followed until their matched case is diagnosed with the disease. There is nothing inherently wrong with the approach; however, quite often, information on the controls is available until the end of the follow-up, or it is easy to obtain. This is true of the British Columbia pulp and paper study discussed in Chapter 3. The question is, does it matter if we use this additional information or not?

2.3 Method III: Case-Cohort Designs

The difference between the matched case-control design and the case-cohort design is subtle. Recall that in the matched case-control design the controls only needed to be followed until their matched case is diagnosed with the disease. In a case-cohort design, the individuals are followed separately; therefore the controls are followed until the end of the study. Quite often, information until the end of the follow-up is available for the controls, or it does not cost much to obtain the information. Therefore, using the above definitions, quite often a case-cohort design is used and then a matched case-cohort method of analysis is applied to the data. In this case they are the same sample design with different methods of analysis; however, in this project, we will refer to them as different designs.

Before starting the discussion on the case-cohort design, it is important to define some additional terminology to bridge between the typical wording in failure time data and the Cox proportional hazard model and epidemiology studies. *Failure* will be the same as experiencing the event (i.e. being diagnosed with the disease). *Censoring* will be synonymous with an individual leaving the study non-diseased.

The case-cohort design is most useful in analyzing time to failure in a large cohort in which failure is rare. A case-cohort study viewed as failure time data consists

of a random sample, the *subcohort*, and any additional cases not in the subcohort. Covariate information is collected from all failures (i.e. cases) and a representative sample of censored observations (i.e. the subcohort of controls). Sampling is done without respect to time or disease status, and, therefore, the design is more flexible than a matched case-control design. Despite the efficiency of the methods, case-cohort designs are not often used because of perceived analytic complexity.

Designs in which a subcohort is chosen at the start of the study to constitute the control group are discussed by Prentice (1986). For failure time data, the semi-parametric Cox (1972) proportional hazards model is routinely used. Observed failures are typically more influential than censored observations in such analyses.

Relative risk is the ratio of the probability of an event in the case group to the probability of the event in the control, adjusted for covariates. This provides a natural approach to the modelling and understanding of the dependence of disease rates on aspects of the preceding covariate history. In the presence of a large cohort with infrequent disease events, the efficiency with which relative risk parameters may be estimated depends strongly on the number of subjects experiencing failure, but the marginal contribution from subjects not developing disease is small. In considering covariate sampling procedures, it is then natural to consider designs in which covariate histories are assembled for all the cases, along with an independent random sample (with replacement) of the control subjects at each distinct failure time. Although this gives rise to a partial likelihood approach to relative risk regression estimation, it leads to poorer efficiency results than does the odds ratio estimator under simple case-control sampling with unmatched controls (Self and Prentice, 1988).

Accordingly, Prentice (1986) proposed a case-cohort design to efficiently analyse cohort data when most observations are censored. Conceptually, a random sample of the cohort, or "subcohort" is designated prospectively as the source of comparison observations for the observed events. All failures are included whether in the subcohort or not, but censored observations are included only if in the subcohort. However, the potential to assess covariates for all members of the cohort must exist since one does not know in advance which individuals fail.

Prentice (1986) proposed a pseudolikelihood procedure for relative risk regression

parameter estimation. This pseudolikelihood mimics the form of partial likelihood estimation of the regression coefficient in this proportional hazards model. Also, a variance estimator was proposed that requires computation of covariances among score components that arise from the sampling design. A corresponding estimator was also given for the cumulative baseline failure rate, for which no estimation procedure currently exists under time-matched case-control sampling. Therneau and Li (1998) suggest that this model can be computed simply using the Cox Proportional Hazards function in one of the statistical packages.

The remainder of this section discusses the case-cohort design in detail. In addition, the Cox Proportional Hazards model is introduced followed by a description of how this model can be used to analyse the case-cohort design.

2.3.1 The Case-Cohort design: Binary Response

Before focusing on our main interest, relative risk estimation, it is instructive to begin with a discussion of odds ratio estimation, based on the follow-up of a cohort of size n to observe whether $D = 0$ (event does not occur in specific time period)or $D = 1$ (event occurs during a specified time period).

Suppose initially that one is interested in the dependence of failure probability on the presence, $z = 1$, or absence, $z = 0$, of some covariate. Denote $p_{ij} = Pr(D = i, z = j)$ $(i, j = 0, 1)$. If one assumes that there is no censoring, then a conventional cohort approach would involve observation of the number of failures d_0 and d_1, and the number of subjects n_0 and n_1, corresponding to $z = 0$ and $z = 1$, respectively. $n_0 - d_0$, $n_1 - d_1$, d_0 and d_1 are the counts in the cells/boxes depicted in Table 2.4.

	$z = 0$	$z = 1$	
$D = 0$	$(n_0 - d_0)$	$(n_1 - d_1)$	
$D = 1$	d_0	d_1	
	n_0	n_1	n

Table 2.4: The Counts in the Case-Cohort Design: Binary Response

Each box has a different probability (think of the boxes being bigger or smaller) and

we fix the number of balls that fall to be n; $n_0 - d_0 + n_1 - d_1 + d_0 + d_1 = n$. The probability of each box is p_{ij}, with constraint, $p_{00} + p_{01} + p_{10} + p_{11} = 1$ this is a case in which the counts are not independent. This modelled via a multinomial, with likelihood of the form

$$L(\boldsymbol{p}) = \binom{n}{(n_0 - d_0), (n_1 - d_1), d_0, d_1} {p_{00}}^{(n_0 - d_0)} {p_{01}}^{(n_1 - d_1)} {p_{10}}^{d_0} {p_{11}}^{d_1}$$

where $\boldsymbol{p} = (p_{00}, p_{01}, p_{10}, p_{11})$. It follows that the respective maximum-likelihood estimators are

$$\begin{aligned} \hat{p_{00}} &= \frac{n_0 - d_0}{n} \\ \hat{p_{01}} &= \frac{n_1 - d_1}{n} \\ \hat{p_{10}} &= \frac{d_0}{n} \\ \hat{p_{11}} &= \frac{d_1}{n}. \end{aligned}$$

Therefore, because of the invariance of the maximum likelihood estimators, the maximum likelihood estimate of the odds ratio $\lambda = p_{11}p_{00}(p_{10}p_{01})^{-1}$ is

$$\hat{\lambda} = \frac{d_1(n_0 - d_0)}{d_0(n_1 - d_1)}$$

(Prentice, 1986). The variance of the log of the odds is approximately the sum of the inverse of the counts. Thus, $\hat{\beta} = log\hat{\lambda}$ has asymptotic variance consistently estimated by ${d_0}^{-1} + {d_1}^{-1} + (n_0 - d_0)^{-1} + (n_1 - d_1)^{-1}$.

Suppose now that the entire cohort is monitored for failure as before, but that covariate values are assembled only for a randomly selected subcohort of size $m \leq n$, and for any additional failing subjects that are not in the subcohort. The counts for this scenario are depicted in Table 2.5. The total number who exposed ($z = 1$), unexposed ($z = 0$), and the grand total are the same as in the situation depicted in Table 2.4. The individuals used in the analysis are the ones that fall into the following cells: $m_0 - k_0$, $m_1 - k_1$, d_0 and d_1.

	$z=0$		$z=1$	
$D=0$	$(n_0-d_0)-(m_0-k_0)$	m_0-k_0	$(n_1-d_1)-(m_1-k_1)$	m_1-k_1
$D=1$	$(d_0)-(k_0)$	k_0	$(d_1)-(k_1)$	k_0

Table 2.5: The Counts in the Case-Cohort Design with a Subcohort: Binary Response

If one re-parameterizes such that

$$p_{00} = p\alpha \text{ and } p_{01} = p(1-\alpha)$$

and notes that $p = 1 - p_{10} - p_{11}$, the likelihood function for such case-cohort data is proportional to

$$p_{10}{}^{d_0} p_{11}{}^{d_1} (1 - p_{10} - p_{11})^{n-d} \alpha^{m_0-k_0} (1-\alpha)^{m_1-k_1}$$

(Prentice, 1986), where (m_0,k_0) and (m_1,k_1) are the numbers of subjects and cases, i.e. failures, corresponding to $z = 0$ and $z = 1$, respectively, in the randomly selected cohort, and $d = d_0 + d_1$. As before, $\hat{p}_{10} = d_0/n$ and $\hat{p}_{11} = d_1/n$. In addition, it is easy to show that $\hat{\alpha} = (m_0 - k_0)/(m - k)$, where $k = k_0 + k_1$. Now, recall that $\lambda = p_{11}p_{00}(p_{10}p_{01})^{-1}$. Re-parameterizing, one obtains:

$$\lambda = p_{11}p\alpha[p_{10}p(1-\alpha)]^{-1} = p_{11}\alpha[p_{10}(1-\alpha)]^{-1}$$

with

$$\hat{\lambda} = \frac{d_1(m_0 - k_0)}{d_0(m_1 - k_1)}.$$

Invariance of the maximum likelihood estimators then yields $\hat{\beta} = log\hat{\lambda}$. As before, this has asymptotic variance consistently estimated by ${d_0}^{-1} + {d_1}^{-1} + (m_0 - k_0)^{-1} + (m_1 - k_1)^{-1}$ (Prentice, 1986).

Prentice and Pyke (1979) show that the odds ratio estimators and their asymptotic variance matrices may be obtained by applying the original logistic regression model to the case-control study as if the data had been obtained in a prospective study. In summary, using this information, Prentice (1986) shows that asymptotic inference

on the odds ratio in a case-cohort study can be carried out by applying the binary logistic failure model

$$Pr(D|z) = \frac{exp\{(\alpha + z\beta)D\}}{\{1 + exp(\alpha + z\beta)\}} \tag{2.4}$$

directly to the $s_0 + s_1$ subjects for whom covariate data is assembled; $s_1 = d$ failing subjects and $s_0 = m - k$ subcohort members who turn out to not fail. Furthermore, Prentice also demonstrated that the case-cohort data provides a natural estimator $\hat{q} = s_1/n$ of the marginal disease probability $q = Pr(D = 1)$, though information on q is not, in itself, useful for large sample odds ratio estimation. It is possible to permit parameters in (2.1) and the subcohort selection to be stratified on baseline characteristics.

2.3.2 The Case-Cohort design: time to response data

Several generalisations of the above formulation will be simultaneously considered: the use of actual times of failure (cases); the replacement of odds ratios by relative risks; the allowance for late entry into the cohort, censorship and even intermittent exclusion from the cohort risk set; and a relaxation to allow non-exponential relative risk forms. For notational convenience, allowance for stratification on baseline covariates will be deferred.

For now, time can be thought of as the time since the beginning of the cohort study; however, in some applications, other specifications, such as age, may be more appropriate. Let $Z(t)$ denote a covariate measurement on a subject at time t. Now, let $\lambda\{t; Z(u), 0 \leq u < t\}$ denote the failure rate of interest at time t for a subject with preceding covariate history $\{Z(u), 0 \leq u < t\}$. Consider the relative risk regression model, which was introduced by Cox (1972),

$$\lambda\{t; Z(u), 0 \leq u < t\} = \lambda_0(t) r\{X(t)\beta\},$$

(see Prentice, 1986), where $r(x)$ is a fixed function with $r(0) = 1$ (e.g. $r(x) = 1 + x$ or $r(x) = e^x$); $X(t)$ is the column p-vector consisting of, possibly time-dependent,

functions of $\{Z(u), 0 \leq u < t\}$ and possibly product terms between such functions and t; β is a column p-vector of regression parameters to be estimated; and $\lambda_0(t)$ is a baseline hazard function corresponding to a standard covariate history for which the modelled regression vector $X(t) \equiv 0$.

Let $N_i(t)$, assuming it has a right-continuous sample path (Prentice, 1986), be the observed number of events for subject i up to and including time t; in other words, N_i takes the value zero prior to an observed failure on the the i^{th} subject, and the value one thereafter. Also let $Y_i(t)$, assuming it has a left-continuous sample path (Prentice, 1986), take a value of one when subject i is at risk for failure (and under observation) and zero otherwise. Now, consider a cohort of size n. Let $\{N_i(u), Y_i(u), Z_i(u); 0 \leq u < t\}$ denote counting, censoring and covariate histories for the i^{th} subject prior to time t.

Now, define the time of failure or censorship for the i^{th} subject as:

$$t_i = min\{t | Y_i(u) = 0; \quad \text{all } u > t\}$$

and the censoring indicator as:

$$\delta_i = \begin{cases} 1 & : \quad N_i(t_i) \neq N_i(t_i^-) \\ 0 & : \quad N_i(t_i) = N_i(t_i^-). \end{cases}$$

Cox (1972) defined a partial likelihood function in the following way:

$$L(\beta) = \prod_{i=1}^{n} \left(r_{ii} \Big/ \sum_{l=1}^{n} r_{li} \right)^{\delta_i}, \tag{2.5}$$

where $r_{li} = Y_l(t_i) r\{X_l(t_i)\beta\}$. This is under standard independent failure time and independent censorship assumptions and full cohort data.

Suppose now that there is a subcohort, C, of size m selected from the censored subcohort. In addition, $\{N_i, Y_i\}$ processes are available for all cohort members; however, covariate histories are only available for members of C and for subjects that fail. Now,

let $K(t) = \{i|N_i(t) = 1\}$; i.e., the set of subjects failing at or before time t. Thus, covariate histories at time t will be assumed available for subjects in $M(t) = K(t) \cup C$. Also, let $\tilde{R}(t) = D(t) \cup C$, where $D(t) = \{i|N_i(t) \neq N_i(t^-)\}$; this is empty unless a failure occurs at time t. Finally, let

$$\Delta t = \begin{cases} 1 & : \quad \tilde{R}(t) \neq C \\ 0 & : \quad \tilde{R}(t) = C \end{cases}$$

Prentice (1986) suggests that for estimation of the relative risk parameter β, using such case-cohort data, one should maximise the function

$$\tilde{L}(\beta) = \prod_{i=1}^{n} \left(r_{ii} \Big/ \sum_{l \in \tilde{R}(t_i)} r_{li} \right)^{\delta_i}. \tag{2.6}$$

The only difference between expressions (2.2) and (2.3) is that in (2.3), the i^{th} denominator factor is a sum over subjects at risk in $\tilde{R}(t_i)$ rather than over subjects at risk in the entire cohort. Since expression (2.3) does not generally have a partial likelihood interpretation (Prentice, 1986), it is termed a pseudolikelihood.

The maximum pseudolikelihood estimate, $\hat{\beta}$, is defined by a solution $U(\hat{\beta}) = 0$, where

$$U(\hat{\beta}) = \frac{\partial log\tilde{L}(\beta)}{\partial \beta}$$

is the score function. Noting that

$$log\tilde{L}(\beta) = \sum_{i=1}^{n} \delta_i \left[log(r_{ii}) - log \left\{ \sum_{l \in \tilde{R}(t_i)} r_{li} \right\} \right],$$

the score function reduces to

$$
\begin{aligned}
U(\beta) &= \sum_{i=1}^{n} \delta_i \left[\frac{Y_i(t_i) X_i(t_i) r'\{X_i t_i \beta\}}{r\{X_i t_i \beta\}} - \frac{\sum_{l \in \tilde{R}(t_i)} Y_l(t_i) X_l(t_i) r'\{X_l t_i \beta\}}{\sum_{l \in \tilde{R}(t_i)} r\{X_{t_i} \beta\}} \right] \\
&= \sum_{i=1}^{n} \delta_i \left(c_{ii} - \sum_{l \in \tilde{R}(t_i)} b_{li} / \sum_{l \in \tilde{R}(t_i)} r_{li} \right),
\end{aligned}
$$

where

$$
b_{li} = Y_i(t_i) X_i(t_i) r'\{X_i t_i \beta\} \quad \text{and} \quad c_{li} = b_{li} r^{-1}\{X_l(t_i)\beta\}.
$$

In summary, Prentice (1986) proposed a pseudolikelihood procedure for the relative risk parameter along with heuristic procedures for parameter estimation; a corresponding estimator was also given for the cumulative baseline failure rate, for which no estimation procedure existed, for time matched case-control sampling. Prentice also showed that subcohort sampling rates can be allowed to vary among strata. A pseudolikelihood function for β can be written as a product of terms (2.3) over strata. Self and Prentice (1988) developed the asymptotic distribution theory for the case-cohort maximum pseudo-likelihood estimators using a combination of martingale and finite population convergence results. They also developed corresponding asymptotic efficiency expressions for relative risk parameter estimation.

2.3.3 The Cox Proportional Hazards Model

As mentioned previously, Therneau and Li (1998) suggest that the Cox Proportional Hazards model can be used to compute the pseudolikelihood estimator of the previous section. Therefore, it is necessary to give a brief description of the Proportional Hazards model.

The hazard or risk function $h(t)$ gives the instantaneous failure rate assuming that the individual has survived to time t,

$$
h(t) = lim_{\delta \to 0} \frac{Pr(t \leq T \leq t + \delta | t \leq T)}{\delta} = \frac{f(t)}{S(t)}.
$$

In other words, the hazard or risk function $h(t)$ approximates the proportion of subjects dying or having events per unit time near time t, where $f(t)$ is the probability density function and $S(t) = Pr(T > t)$ is the survival function.

When a cohort is subdivided into two subcohorts, C_1 (exposed) and C_0 (unexposed), by the presence or absence of a certain characteristic (an exposure such as smoking), each subcohort corresponds to its own hazard or risk function and the ratio of two such functions is called the relative risk,

$$RR(t) = \frac{h(t; C_1)}{h(t; C_0)}.$$

In general, $RR(t)$, is a function of time and measures the magnitude of an effect; when it remains constant we have the proportional hazards model, which assumes that lifetime and failure time data are independently distributed with the hazard function given by

$$h[t|x(t)] = h_0(t)exp\{x(t)'\beta\},$$

where $x(t)$ is a vector of observable, possibly time dependent, covariates, and $h_0(t)$ is the hazard function at $x(t) = 0$ (or $h[t; I_0]$). This is a special case of the regression model given on page 23. The "regression coefficient", β, represents the relative risk on the log scale. One of the reasons for the model's popularity in fitting failure data is that the unknown parameter, β, can be estimated by partial likelihood without putting a parametric structure on h_0, and thus, this model is more flexible. Even though the model makes a number of assumptions which may not always be completely satisfied, fitting such models can have both descriptive and analytical value.

2.3.4 Computing the Maximum Pseudolikelihood Estimator

The Self and Prentice (1988) estimate of $\hat{\beta}$, which is nearly identical to the estimate proposed by Prentice (1986), can be computed fairly easily, using any Cox (Proportional Hazards) model program that allows for an *offset* term (Therneau and Li, 1998). If one assumes that there is a concurrent registry which can be used to identify all of the subjects who experience an event, then the goal is to collect covariate data on

only a subcohort of the subjects, randomly sampled from the cohort, and augment the sample with all of those subjects who experience an event.

Let x be a constructed variable which is equal to zero for subjects in the random subcohort and take some large negative value (e.g. -100) for subjects who have experienced the event. If there are subjects who are in both the subcohort and have experienced the event of interest, then enter them into the data as two separate observations: one with $x = 0$ and status equal to censored, and one with x equal to a large negative number and status equal to event. Now, the model is fit with $\mathit{offset}(x)$ as a term on the right hand side (Therneau and Li, 1998). The offset function is, in a sense, putting weights on the observations.

Observations which are not part of the subcohort, although formally part of the estimation of the mean, do not in actuality affect the result since they have a relative weight of $exp(x)$, which is very small, when x is a large negative number, as compared to the subcohort subjects who have a relative weight of $exp(0) = 1$ when computing the mean.

Time dependent covariates are coded by breaking each subject up into multiple observations, each over an interval (start, stop]. Each observation contains the values of the covariates that apply over that interval, along with a status variable that indicates whether the interval was terminated with an event (1-yes, 0-no).

Now, assume that we have computed the Self and Prentice (1988) estimate using this method. Because of oversampling of cases with an event, the usual estimate of variance will overstate the precision of $\hat{\beta}$ (Therneau and Li, 1998). Nevertheless, Self and Prentice (1988) proposed an asymptotically consistent estimate of $var(\hat{\beta})$; this estimate has been criticised as being overly complex for practical use (Therneau and Li, 1998). However, Therneau and Li (1998) show that $var(\hat{\beta})$ can be calculated by standard packages as

$$\hat{V} = \hat{\tau}^{-1} + (1 - \alpha) D'_{SC} D_{SC}$$

,

where $\hat{\tau}^{-1}$ is the "standard" variance estimate returned by the Cox model program and D_{SC} is the subset of the matrix of dfbeta residuals that contain only those rows from

the subcohort C; $\alpha = m/n$ is the proportion of cases sampled. The dfbeta residuals are a matrix, where the ith row gives the approximate change in the coefficients due to the addition of subject i. The dfbeta matrix contains the dfbeta residuals, with each column scaled by the standard deviation of that coefficient. For those computer packages which return dfbeta residuals, this represents a very simple calculation to correct the "standardised" variance estimate $\hat{\tau}^{-1}$.

Writing the Self and Prentice (1988) estimate in this form, gives further insight into the meaning of the estimate. Let β_p be the true coefficient for the (infinite) population at large, $\hat{\beta}_c$ the estimate for the cohort, if data were collected on all of the subjects therein, and $\hat{\beta}_{sc}$ the value for the actual study as conducted. The first term, $\hat{\tau}^{-1}$, is an estimate of var($\hat{\beta}_c$), the estimated variance that would have been obtained if all of the subjects in the cohort had been used in the computation. The second term is an estimate of the finite sample contribution var($\hat{\beta}_{sc}|$ cohort).

Another option is to treat the data as the results of a weighted random sample, as in survey methods (Barlow, 1994). Let $n(t)$ and $m(t)$ be the numbers of cohort and subcohort subjects which are at risk at time t. The subject with an event is in the sampled risk set with probability 1, but each of the other subjects with probability $\alpha(t) = m(t)/n(t)$. Then the sampling weight $w_i(t) = 1/\alpha(t)$ for the subcohort, 1 for the event at time t and 0 for the other (un-sampled) subjects.

In the case of the Pulp and Paper Mill Worker example in Chapter 3, all of the weights are equal to one. The reason for this is that the subcohort is the cohort, thus $m(t) = n(t)$.

Both the Self and Prentice (1988) and Barlow (1994) estimators will converge to the true β in large samples (Therneau and Li, 1998). If $\alpha(t)$ is constant over time, then the proposals are very similar and only differ in how much weight is given to the actual event at time t in computing the weighted mean.

Although it appears to be simple to carry out the case-cohort design with time to response data, it has only been done for very simple examples.

Chapter 3

Application of Case-Cohort Analysis method

3.1 Overview

Based on some preliminary analyses, the British Columbia Cancer Agency initiated a two-phase study of British Columbian pulp and paper mill workers. Phase I investigated the cohort's mortality and cancer incidence outcomes; Phase II was a matched case-control study (on age ranges).

The matched case-control method was analysed using conditional logistic regression with age-range matching. The cases and their matched controls were followed through time and considered "at risk" until they experienced the event (e.g. diagnosed with cancer) or they were censored (left the study or the study terminated). The controls were cutoff at the date their matched control experienced the event; therefore not all the available information was used. The B.C. Cancer Agency wished to investigate how the results would differ, if at all, if all of the available information was used.

A case-cohort design using the survival analysis method, as previously described, addresses this issue. For each individual, the time-dependent covariates are divided into intervals, such that each interval contains the values of the covariate along with a status variable that indicates whether the interval terminates with an event. This will

allow the investigator to use all of the available information without the potential bias in the matched case-control method since now each individual is examined separately. Although this method has been developed theoretically, it has only been applied to simple problems or simulations.

In this chapter, we describe the background, discuss the phase I of the study and re-perform the matched case-control analysis. In this project, we are only interested in one chemical, black liquor. So, we re-performed the matched case-control analysis for only this one chemical. Similarly we only consider one chemical in the case-cohort method. We then apply the case-cohort method using the Cox Proportional Hazards function (discussed previously) in S-Plus; the time-dependent covariate is cumulative exposure and the event is the diagnosis of prostate cancer. In order to obtain a dose-response relationship, cumulative exposure was coded as a categorical variable. When compared with the results from the matched case-control study the trends appear to be similar; however, there are some differences that suggest the case-cohort method may be more appealing. One major problem, however, is that, although the case-cohort model worked nicely for certain exposure level breakdowns, it did not converge for others. Thus, there does appear to be a problem with the stability of the estimation procedure. It is possible that this difficulty is inherent in the model formulation or it may be fixable via manipulation of the S-Plus Cox Proportional Hazards function or via creating a new computer program specific to the methodology. This stability problem requires future investigation before the case-cohort model can be used over the case-control model.

3.2 Background

The following section discusses the work from two studies completed at the British Columbia Cancer Agency (Band et al, 1997; Band et al., 2001). Pulp and paper is a major industry in British Columbia; it produces almost one third of Canada's annual pulp and paper tonnage. Wood can be converted to pulp by mechanical, semichemical and chemical processes, the most prevalent in Canada being the latter. In chemical pulping, lignin is solubilized under the following two conditions: the acidic or sulfite

process, and the alkaline, also called kraft or sulfate, process, the latter being the most common. The active chemical in the sulfite process is bisulfite salt that is usually ammonium based, whereas in the alkaline process, the active chemicals are a mixture of sodium hydroxide and sodium sulfide. After delignification, the pulp may be bleached; current practice involves use of combinations of chlorination with elemental chlorine, alkaline extraction with sodium hydroxide, and various oxidative stages using sodium or calcium hypochlorite, chlorine dioxide, or hydrogen peroxide. Thus, during chemical pulping, pulp and paper workers are exposed to known or suspected carcinogens, including organic chlorinated compounds, sulfuric acid mist, formaldehyde, and arsenic and chloroform (the last two have been previously used as antisap stain).

In 1982, an occupational cancer research program was launched in British Columbia based on the review of results of previous epidemiologic proportionate mortality, cohort, and case-control studies of pulp and paper workers. It was found that although excess risks for several cancer sites have been suggested, results were inconsistent. This is mainly because of limitations of the studies based on vital statistics or on small numbers. Although these findings relate to pulp and paper workers in general, they do not take into account the two main types of pulping processes, kraft and sulfite. Of the numerous studies that have been conducted, only five include data for these processes. Based on these studies, there is evidence of increased risk among kraft mill workers for stomach and colon cancers, lymphosarcoma, reticulum cell sarcoma, and Hodgkin's disease; in addition, there appears to be an excess risk in sulfite pulp mills for cancer of the stomach, rectum, pancreas, bladder, kidney, lymphosarcom, and reticulum cell sarcoma . Furthermore, paper mill workers were found to be at increased risk for colon, pancreas, and lung cancer; one study reported a marked excess of lung cancer among paper board workers.

One branch of the 1982 occupational cancer research program was directed towards detecting occupational cancer risk factors. The initial study was based on a death certificate analysis of all deaths in BC from 1950 to 1978, later updated to 1984. This first study revealed a statistically significant increase in the proportional mortality ratio for lymphosarcoma and reticulum cell sarcoma in pulp and paper mill workers.

The second study involved collecting lifetime occupational history from male incident cancer patients of at least 20 years of age ascertained from the British Columbia Cancer Registry between January 1, 1983 and December 31, 1988. Based on the preliminary analysis, the odds ratio for non-Hodgkin's lymphoma appeared to be significantly increased for workers in the pulp and paper industry. These findings lead to the initiation of a two-phase cohort study of British Columbia pulp and paper workers. The objective of phase I was to investigate the cohort's mortality and cancer incidence outcomes; phase II was a matched case-control study with detailed work history and exposure assessment based on mill specific job exposure matrices. In the first phase, no attempt was made to classify workers by departments and no exposure data was obtained that might provide further explanations. However, the second phase should assist in evaluating whether the excess risk for specific cancers reflects the exposure among subsets of workers.

3.3 Phase I: Cohort study

All members of the cohort were male workers with at least one year of employment in one of 14 pulp and paper mills between January 1, 1950 and December 31, 1992. The mills were included in the study if: (i) they started production in 1970 or earlier, (ii) they have had a minimum of 1000 workers ever employed, and (iii) records were available for all employees. In order to determine if workers were eligible, questionnaires were sent to management of all pulp and paper mills in British Columbia; the questionnaires requested information on the type of mill, when production began, an estimated number of total workers ever employed, and the quality and availability of records. All male workers with at least one year of employment in eligible mills on January 1, 1950 until December 31, 1992, the cut-off date for follow-up, were enrolled in the cohort. The data collection included full names and dates of birth, hire, and termination of employment. Information on tobacco smoking and other cancer risk factors related to life-style are not available.

3.3.1 Cohort Mortality Study

Phase I was divided into two studies: a cohort mortality study and a cohort cancer incidence study. The first study reported the cancer mortality experience of the chemical pulping process by type for a cohort of 30,157 pulp and paper workers in British Columbia (Band et al., 1997).

Standardized mortality ratios (SMRs) were used to compare the mortality of the cohort with that of the Canadian population. The Canadian population mortality rates were obtained from the Laboratory Centre for Disease Control, Health Canada; they were calculated by 5-year age groups and 5-year calendar periods dating back to 1950 (Band et al., 1997). The rates for the period 1985-1989 were used for the period 1990-1992. Person-years at risk were calculated from 1 year after the date of hire to December 31, 1992, or to the year of death, whichever came first. Observations were censored at the date when they were last known to be alive. Latency effects were examined using work duration and time since first employment calculated from 1 year after the date of hire; time since first employment was calculated to the last follow-up date (Band et al., 1997). Tests of significance and of the SMRs were calculated assuming that the observed number of events followed a Poisson distribution with the mean given by the expected number of events; 90 percent confidence intervals corresponding to a one-sided 5 percent significance test were used. Record linkage of the cohort with the National Mortality Database was performed at Statistics Canada using the generalised iterative record linkage method (Band et al., 1997).

Cancer risks significantly associated with work duration and time from first employment of 15 years or more were observed for cancers of the pleura, kidney and brain in the total cohort, for kidney cancer among the kraft mill workers only, for Hodgkin's disease among the sulfite mill workers only, and for esophageal cancer among the workers ever employed in both kraft and sulfite mills.

3.3.2 Cohort Cancer Incidence

Epidemiologic studies specifically designed to investigate pulp and paper workers have mainly been mortality studies with only three reporting cancer incidence results (Band

et al., 2001). Therefore, further work was done on cancer incidence outcomes of 28,278 members of the British Columbia pulp and paper cohort (Band, et al., 2001).

Details of the collection methods were previously described. Recall that the mortality study consisted of a total of 30,157 workers. Of these, 1989 were excluded from the cancer incidence cohort due to the following events which occurred prior to 1969: 1134 were lost due to follow-up, 552 died from non-cancer causes, 175 have been diagnosed with cancer. In addition, previously missing birth date information from the mortality cohort was found for 10 workers, who were added to the incidence study. Thus, 28,278 workers were included in the analysis. Of these workers, 20,041 (71%) were employed in the kraft process only, 3756 (13%) worked in the sulfite process only, and 4481 (16%) had worked in both processes. The number of workers also exposed to the paper-making process in the total cohort and in the three subcohorts was: 16,080 (56%) of all the workers, 12,647 (63%) of the workers employed in the kraft process only, 942 (25%) of the workers employed in the sulfite process only, and 2941 (56%) of the workers employed in both the kraft and sulphite processes. Over 95% of those in all the processes were successfully traced (Band et. al, 2001).

Standardized incidence ratios (SIR) were used to compare the cancer incidence of the cohort with that of the Canadian male population. A SIR of 1 means that the cancer incidence rate in the cohort and general population are the same. A SIR significantly greater than 1, indicates that the cancer rate of the cohort is greater than that of the general population. As before, the Canadian population mortality rates were obtained from the Laboratory Centre for Disease Control, Health Canada; they were calculated by 5-year age groups and 5-year calendar periods dating back to 1950 (Band et al., 1997). The rates for the period 1985-1989 were used for the period 1990-1992. Person-years at risk were calculated from 1 year after the date of hire to December 31, 1992, or to the year of death, whichever came first. Observations were censored at the date when they were last known to be alive. Latency effects (the latency period is the time when the disease is concealed, hidden, or inactive) were examined using work duration and time since first employment calculated from 1 year after the date of hire; time since first employment was calculated to the last follow-up date. A 15-year latency cutoff was selected because the person-year distribution of

all the workers with time from first exposure of $\geq$ 15 years (210,546 person-years) was equally distributed between those with $<$ 15 years of employment (110,211 or 54 %) and those with $\geq$ 15 years of employment (100,335 or 48 %) (Band et al., 2001). Tests of significance and of the SMRs were calculated assuming that the observed number of events followed a Poisson distribution with the mean given by the expected number of events; 90 percent confidence intervals corresponding to a one-sided 5 percent significance test were used. Record linkage of the cohort with the National Mortality Database was performed at Statistics Canada using the generalised iterative record linkage method. In Canada, asertainment of cancer incidence cases on a national basis dates back to 1969, hence the 1 January 1969 follow-up starting date of this study.

The cancer incidence study indicated statistically significant excess risks for work duration of 15 or more years, for the following cancer sites (Band et al., 2001):

- All workers: skin melanoma (26 cases, SIR=1.78), cancer of the pleura (6 cases, SIR=2.8), and of the prostate (175 cases, SIR=1.24)
- Workers in the kraft process: skin melanoma (25 cases, SIR=1.73)
- Workers in the sulfite process: skin melanoma (3 cases, SIR=2.65), cancer of the rectum (11 case, SIR=1.90), and of the pleura (3 cases, SIR=16.84)
- Workers employed in both the kraft and sulfite processes: cancer of the stomach (21 cases, SIR=1.55) and of the prostate (82 cases, SIR=1.44), leukimas (14 cases, SIR=1.66).

In addition, the data comparing workers exposed only to the pulping process with those exposed to the pulping and paper-making processes were analysed. These comparative analyses were carried out for all workers and also for each of the three subcohorts. The results were similar to those for the pulping and paper-making processes together and they did not reveal significant differences in the cancer risks for workers exposed to the paper-making process in addition to the pulping process.

There are several potential causes leading to the differences in cancer rates, including occupational exposure, genetic pre-disposition, lifestyle and other risk factors

(Band et al., 2001). Information on genetic predisposition on other risk factors is not available in the retrospective study dating back to 1950 (Band et al, 2001).

These findings suggest that long term work in the industry is associated with an excess risk of skin melanoma, prostate and pleural cancers. The excess risk of pleural cancer likely reflects past asbestos exposure since 90% of these cases were malignant mesotheliomas (Band et al, 2001). It should be noted that the incidence rates of prostate cancer and skin melanoma in British Columbia are high relative to Canadian rates (Band et.al, 2001). Since 94% of the pulp and paper cohort has been traced to British Columbia, the data was re-analysed using British Columbia rates. Relative risks for skin melanoma became reduced to non-significant levels, whereas the relative risks for prostate cancer remained significantly elevated among long term workers exposed to both kraft and sulfite processes but not in the total cohort (Band et al., 2001). The potential exposures that might be associated with prostate cancer risk were investigated in phase II of the study.

Additionally a significant excess risk for stomach cancer and leukemia was observed among long term workers employed in both processes, as well as for cancer of the rectum among long term workers employed in the sulfite process only. Potential exposures associated with the increased risk will be examined in a later study.

What are the different findings between the incidence study and the mortality study? The significantly increased mortality cancer risks suggested in the mortality were not confirmed in this cancer incidence study, including: a) all workers: brain and kidney cancer; b) workers in the kraft process: kidney cancer; c) workers in sulfite process: Hodgkin's disease; and d) workers in both processes: esophageal cancer. Why are there these differences? The discrepancies between cancer diagnosis listed on pathology reports and cause of death listed on death certificates caused the differences for kidney, brain and esophageal cancer; it should be emphasized that the cancer diagnosis based on pathological diagnosis is generally more accurate (Band et al., 2001).

3.4 Phase II: Matched Case-Control Study

In Canada and the United States, prostate cancer is the most common cancer in men, except for non-melanoma skin cancer (Band et al., 1997). These two countries have the highest incidence rates for prostatic cancer in the world, with the highest rates being observed among black men in the United States (Band et al., 1997). There has been an explosion of scientific interest in the epidemiology of this disease (Gallagher and Fleshner, 1998). There are still many unknowns concerning prostate cancer's etiology. A number of studies have shed light on some important risk factors: age, family history, black American ethnicity, hormonal and sexual factors, and a high consumption of animal fat and red meat (Gallagher and Fleshner, 1998). A large number of diverse occupations have also been suggested to be associated with an increased risk for prostatic cancer, including administrative, managerial, professional, health and clerical occupations; mechanics, welders, policemen, and farmers; as well as workers in metal, paint, and rubber industries (Band et al., 1997). In the study by Band et al. (1997) there is evidence of an association between prostate cancer and the pulp and paper industry.

3.4.1 Description of the Data

The general methodology of the study has been described in the previous section. Recall that the mortality study consisted of a total of 30,157 workers and the cancer incidence study consisted of 28,278 workers. Both of these studies included individuals who had been diagnosed with a variety of cancers (and other health problems), the cases. There are two types of cases: individuals whose cause of death is determined to be cancer during an autopsy, and the individuals who are diagnosed with cancer by a physician (cancer incidence cases). In this phase of the study, only the cancer incidence cases are used. Information on the incidence cases is known only for the years 1969 to 1992. All studies included the individuals who were healthy at the end of the study (controls). After selecting only the prostate cancer incidence cases and their matched controls, 1,997 unique workers remained in the analysis of the matched case-control study.

162 chemicals that are used at the pulp and paper mills were identified as potentially contributing to the development of prostate cancer. These chemicals were grouped into 23 chemical groups. Cohort members who were ever exposed to a particular chemical are considered exposed to that chemical; otherwise, they are considered non-exposed. Also, employment within the last five years of the cohort follow-up was not included in the calculation of exposure.

Table 3.1 is an example of the data set obtained in the matched case-control study. It should be noted that only the rows pertaining to the chemical of interest were used in the analysis.

3.4.2 Methodology

The matched case-control analysis was carried out by the British Columbia Cancer Agency. Recall that in 1992, a two-phase retrospective cohort study of 30,000 British Columbia pulp and paper workers was undertaken. To describe exposures of the workers for a matched case-control study within this cohort, job exposure matrices were developed. The initial stage of development included an exhaustive review of processes, job titles and chemicals coupled with a survey of each mill to evaluate equipment layout, collect hygiene data and perform interviews of employees.

Exposure information from 14 pulp mills was then organized into 90 mill-specific or period-specific matrices. Semi-quantitative exposure assessments were assigned to each combination of job title and chemical or group of chemicals. Besides an estimate of the concentration, variables describing the frequency of exposure as well as the potential for peak exposures were included,

$$Exposure = Concentration * Frequency * Duration.$$

Duration is measured in months, assuming that a work shift is 12 hours per day. In early years, workers worked 8 hours a day. Then all mills changed to 12 hour work days with fewer working days per week. Working months with 8 hours per day were converted into 2/3 equivalent months. Concentration level of exposure were evaluated based on proximity to and characteristics of the source where 0 = unexposed, 1 =

ind.	chem.	start yr.	birth yr.	...	cancer yr.	...	cum. exp.	procase	match
1	132	1950	1921	...	1985	...	0.05	1	3801
⋮								⋮	⋮
1	132	1984	1921	...	1985	...	1.75	1	3801
1	132	1985	1921	...	1985	...	1.75	1	3801
1	136	1972	1921	...	1985	...	0	1	3801
23	132	1962	1932	...	1992	...	0	0	2404
23	132	1963	1932	...	1992	...	0	0	2404
⋮								⋮	⋮
23	103	1965	1932	...	1992	...	0.6	0	2404
23	103	1966	1932	...	1992	...	1.2	0	2404
15	136	1987	1944	...	1992	...	0	0	1801
⋮								⋮	⋮
15	162	1972	1944	...	1992	...	0	0	1801
15	162	1973	1944	...	1992	...	0	0	1801
4	2	1981	1921	...	1987	...	0.5	1	3801
4	2	1982	1921	...	1987	...	0.7	1	3801
4	2	1987	1921	...	1987	...	5.5	1	3801
⋮								⋮	⋮

Table 3.1: An Example of the Matched Case-Control Data

Only the key variables are included (the original data set included 26 variables). The important variables are: chem. (a number that identifies the chemical), ind. (a number that identifies the individual), start yr. (the year of employment for that row), birth yr. (year of birth), cancer yr, (the year of cancer diagnosis or the end of the follow-up period), cum. exp. (cumulative exposure), procase (an indicator variable that equals one if the individual is a case), match (a matching variable).

low and 3 = high. Frequency duration of exposure was broken into levels where 0 = never exposed, 1 = less than one hour per work shift, 2 = 1 to 3 hours per work shift and 3 = greater than 3 hours per work shift. The total exposure amount for lifetime is the sum of all exposures for the same chemical.

A matched case-control analysis method was used. Cases comprised all 287 workers who were diagnosed with prostate cancer; controls comprised 1,710 healthy workers at the end of the follow-up. The controls were matched to cases based on age (year of birth). The matching is based on age since the individuals would then likely have worked in the mills around the same time; this is important since the degree of exposure in, say 1950, is different than the exposure in, say 1988. The controls were followed until their matched case experienced an event.

Conditional logistic regression for matched sets data was carried out using SAS; test of significance of the adjusted odds ratios (ORs) and 95% confidence intervals were calculated. Analyses were performed for each of the 162 chemicals individually. Each analysis was done for 3 different levels of exposure (and of course the baseline level of no exposure). The 3 exposure levels were chosen such that there was approximately the same number of controls in each level.

3.4.3 Results

For this project, only the results for the chemical black liquor are of interest (Table 3.2). We re-performed the matched case-control analysis to verify that we obtained the same results as the British Columbia Cancer Agency. These results will be compared to the results from the case-cohort study in the following section.

Exposure	Cases	OR	95% CI
Non-Exp	247	1.00	-
≤ 2.92	14	2.65	1.58 - 5.08
2.92- 12.0	12	1.93	0.96 - 3.87
> 12.0	14	1.96	1.04 - 3.71

Table 3.2: Results from the Matched Case-Controls Method

The odds of an individual with exposure ≥ 2.92 developing prostate cancer is

2.65 times that of the odds of an individual with no exposure. Similarly, the odds of an individual with exposure within the range 2.92 to 12.0 and an individual with exposure > 12 developing prostate cancer are, respectively, 1.93 and 1.96 times that of the odds of an individual with no exposure. Therefore, the odds ratios indicate that the exposed individuals are significantly more likely to be diagnosed with prostate cancer than the un-exposed individuals. When we look at just the point estimate, the odds of being diagnosed with prostate cancer does not appear to increase when exposure increases. However, if we look at the confidence intervals, it is hard to draw a conclusion.

3.5 Case-Cohort Method

3.5.1 Description of the Data

In this section, we describe our re-analysis of the phase II data using the case-cohort method described in Chapter 2. As in the matched case-control study, the data for the analyses included information on 1,997 unique workers with at least one year of employment in one of 14 pulp and paper mills between January 1, 1950 and December 31, 1992. Recall that, information on the cancer incidence cases is known only for the years 1969 to 1992. The data included 287 individuals who have been diagnosed with prostate cancer (cancer incidence cases) and 1,710 individuals who were healthy at the end of the study (controls).

162 chemicals that are used at the pulp and paper mills were identified as potentially contributing to the development of prostate cancer. Each row of this data set represents one individual's exposure to one chemical at one job for one year; information is given for several factors, such as cumulative exposure, in each row. Recall that

$$Exposure = Concentration * Frequency * Duration.$$

For this project, we were only interested in an individual's exposure history to one chemical (black liquor). Therefore, the data set needed to be altered from its

original format. Tables 3.3 and 3.4 give and example of the original format and the new format, respectively.

The remainder of this section describes how the data was converted from the original format to the format used in the case-cohort analysis. First, the data set was separated into the individuals who were at some point exposed to black liquor (373 individuals; 80 cases and 293 controls) and the individuals who were never exposed to black liquor (1,624 individuals; 207 cases and 1,417 controls).

The following was done to the individuals who were exposed to black liquor (chemical 132). First, the rows that did not pertain to the chemical of interest were removed (373 individuals; 80 cases and 293 controls). There were some individuals who were exposed to the same chemical at two jobs in the same year; therefore, there were two rows for that year for that individual. When this situation arose, one of the two rows was removed; this removed 13.3% of the rows (no individuals were removed). Finally, rows were added to each individual from the termination date (last year of work) until the diagnostic date, or the end of study (1992); once again, no individuals were removed. Note that the cumulative exposure for these added years is the cumulative exposure for last year of work. To illustrate, consider individual 1. In the original data (Table 3.3) individual 1 was exposed to two chemicals, 132 (black liquor) and 136 (some other chemical). The row that pertained to chemical 136 was removed. There are two rows for year 1975 for chemical 132, so one of these rows was removed. Finally, rows were added for the years 1981 (since 1980 was the last year of work) through 1985 (the diagnostic date).

Next, the data individuals who were not exposed to black liquor was re-formatted. Since all of the information in this portion of the data set did not pertain to the chemical of interest, only one row for each individual who was not exposed was kept; this row contained all of the important information such as age, year of diagnosis (or year censored) and whether the individual was a case or a control. Certain values of some of the variables had to be replaced. Cumulative exposure was set to be zero for all individuals; the starting year was set to be 1950 (the first year of the study). Finally a row was added for each individual for each year from 1951 until the diagnostic date, or the end of study (1992); this left all 1,624 individuals (207 cases

ind.	chem.	start yr.	birth yr.	...	cancer yr.	...	cum. exp.	procase	match
1	132	1950	1921	...	1985	...	0.05	1	3801
⋮								⋮	⋮
1	132	1975	1921	...	1985	...	1.75	1	3801
1	132	1975	1921	...	1985	...	1.75	1	3801
1	132	1980	1921	...	1985	...	1.75	1	3801
1	136	1972	1921	...	1985	...	0	1	3801
23	132	1962	1932	...	1992	...	0	0	2404
⋮								⋮	⋮
23	103	1965	1932	...	1992	...	0.6	0	2404
23	103	1966	1932	...	1992	...	1.2	0	2404
15	136	1987	1944	...	1992	...	0	0	1801
⋮								⋮	⋮
15	162	1972	1944	...	1992	...	0	0	1801
15	162	1973	1944	...	1992	...	0	0	1801
4	2	1981	1921	...	1987	...	0.5	1	3801
4	2	1981	1921	...	1987	...	0.5	1	3801
4	2	1987	1921	...	1987	...	5.5	1	3801
⋮								⋮	⋮

Table 3.3: An Example of the Original Format

ind.	start yr.	birth yr.	...	cancer yr.	...	cum. exp.	procase	status	cov
1	1950	1921	...	1985	...	0.05	1	0	1
⋮								⋮	⋮
1	1985	1921	. ...	1985	...	1.75	1	1	1
23	1962	1932	. ...	1992	...	0	0	0	0
23	1963	1932	...	1992	...	0	0	0	0
⋮								⋮	⋮
23	1992	1932	...	1992	...	1.2	0	0	1
15	1950	1944	...	1992	...	0	0	0	0
⋮								⋮	⋮
15	1992	1944	...	1992	...	0	0	0	0
4	1950	1927	...	1987	...	0	1	0	1
4	1951	1927	...	1987	...	0	1	0	1
⋮								⋮	⋮
4	1987	1927	...	1987	...	0	1	1	2
⋮								⋮	⋮

Table 3.4: An Example of the Case-Cohort Data

Key variables are the same as those in Table 3.1. New ones introduced are: status (indicator variable that equals one if the interval was terminated with an event) and cov (a categorical covariate).

and 1,417 controls). As an example, consider individual 4. This individual was never exposed to chemical 132, so all rows, but one, were removed. Then, the cumulative exposure was set to zero and the starting year was set to 1950. This row was repeated for all years until 1987 (the diagnostic date). The only variable that changed from row to row was start yr. (the starting year).

The rows for the exposed and unexposed individuals were combined to form a data set with all 1,997 individuals (287 cases and 1710 controls) remaining in the data set to be analysed.

Recall that in the model proposed by Prentice (1986) the time dependent covariates were coded by breaking each subject up into multiple observations, each over an interval (start, stop]. Each observation contains the value of the covariates that apply over that interval, along with a status variable that indicates whether the interval was terminated with an event (i.e. diagnosis of cancer). Therefore, a status indicator variable was created; it took the value one for all the rows when the individual experienced an event (i.e. when the starting year of the row and the diagnosis date were the same) and zero otherwise. Once again, consider individual 4. This individual was diagnosed with prostate cancer in 1987, so the status variable is equal to 0 for all years except this year.

In addition, the time dependent covariate, cumulative chemicals exposure, had to be coded as a categorical variable rather than continuous. This would give some sense of the dose-response relationship.

Cumulative exposure has a minimum value of 0 and a maximum value of 534.14. Several approaches were used to come up with cut-points for the cumulative exposure. The question was how to come up with the cut-points and how many categorical levels would be the best. All of the methods that are described below were done for a different number of categorical levels. Initially the categorical levels of exposure were created by simply dividing the range of cumulative exposure into groups of equal size. A major problem with this method is that it creates empty cells; in other words, there will be some levels that have no controls (or no cases). This will lead to a failure of maximum likelihood estimation procedure. Therefore, a different way of dividing the cumulative exposure was required. From looking at the data, it is obvious that there

are a large number of zeros present, and not so many values in the upper limit. Out of the 69,960 records, 65,317 have cumulative exposure values equal to zero. One possible method to choose the cut-points was to divide the data so that there were an equal number of exposed individuals in each interval. However, what one really desires is to have approximately the same number of events (cases who are diagnosed with cancer) in each interval. This is how the intervals were selected (with the end-points rounded to the closest integer). Table 3.5 shows the categorical exposure levels we selected, plus the cut-points that were used in the matched case-control analysis. As a final note, it may be of future interest to look at design methods for selecting the cut-points, rather than just an ad-hoc method as was used in this project.

Exposure Levels	Exposure	Records	Events
0	0	62846	247
1	> 0	4271	40
0	0	62846	247
1	$(0, 7]$	2280	21
2	> 7	1991	19
0	0	62846	247
1	$(0, 2]$	1276	14
2	$(2, 12]$	1581	12
3	> 12	1414	14
0	0	62846	247
1	$(0, 1]$	920	10
2	$(1, 4]$	961	8
3	$(4, 17]$	1328	10
4	> 17	1062	12
0	0	247	62846
1	(0, 2.92]	1596	14
2	(2.92, 12.0]	1261	12
3	> 12.0	1414	14

Table 3.5: Exposure Levels

For the analysis, we used the above cut-points. The latter was done so that a direct comparison could be made.

3.5.2 The Analysis

As described in detail in Section 2.3.4, the Self and Prentice (1988) estimate of $\hat{\beta}$, which is nearly identical to the estimate proposed by Prentice (1986), can be computed fairly easily using any Cox (Proportional Hazards) model; the *coxph* function S-Plus was used for this analysis. Relative risks and 95% confidence intervals were calculated.

Exposure	Events	RR	95% CI	cluster bound 95% CI
Non-Exp	247	1.00	-	-
≤ 7	21	2.04	1.31 - 3.19	1.33 - 3.12
> 7	19	1.61	1.01 - 2.57	0.97 - 2.66
Non-Exp	247	1.00	-	-
≤ 2	14	2.46	1.43 - 4.21	1.51 - 4.00
2 - 12	12	1.61	0.90 - 2.88	0.87 - 2.98
> 12	14	1.57	0.91 - 2.68	0.88 - 2.77
Non-Exp	247	1.00	-	-
≤ 1	10	2.44	1.29 - 4.61	1.33 - 4.48
1- 4	8	1.89	0.92 - 3.77	0.98 - 3.55
4- 17	10	1.35	0.72 - 2.55	0.69 - 2.67
> 17	12	1.90	1.07 - 3.40	1.03 - 3.52
Non-Exp	247	1.00	-	-
≤ 2.92	15	2.05	1.25 - 3.55	1.27 - 3.31
2.92- 12.0	11	1.82	0.99 - 3.35	0.95 - 3.50
> 12.0	14	1.56	0.91 - 2.36	0.88 - 2.77

Table 3.6: Results from the Case-Cohort Method

In this example, all of the weights are equal to 1. Also, since the cases are known, the subcohort consists of only the controls.

The model proposed by Prentice (1986) was fit separately to different groups of dummy variables (for each different range of cumulative exposure; the baseline being cumulative exposure being equal to zero), with age included in all of the models. First the model was fit with no cluster function, and then it was fit with cluster function. The cluster function identifies correlated groups of observations. In this example, there are multiple rows for each individual, so by using the cluster function, this is accounted for by adjusting the standard error. The results of these two models are sumarised in Tables 3.6. The only difference in the two models is the addition of a

robust standard error and therefore different confidence intervals.

In all of the category groups above, the risk for developing prostate cancer is higher for the exposed individuals than the un-exposed individuals. Based on the point estimates, there also does not appear to be more of a risk for the more exposed individuals; this is the same result as found in the matched case-control analysis.

The models that are of interest are the ones that incorporate the cluster function. In particular we are most interested in the model with the cluster function and the same cut-points as the matched case-control analysis.

The risk of an individual with exposure ≥ 2.92 developing prostate cancer is 2.05 times that of the risk of an individual with no exposure. Similarly, the odds of an individual with exposure within the range 2.92 to 12.0 and an individual with exposure > 12 developing prostate cancer are, respectively, 1.82 and 1.56 times that of the risk of an individual with no exposure. The results of this model will be compared to the results of the matched case-control analysis.

3.5.3 Comparison of the Results

One of the major objectives of this project is to compare the results of the matched case-control method with the results from the case-cohort method. As was expected, they both indicate that the chance of being diagnosed with prostate cancer is much higher for the exposed individuals than the unexposed individuals. Another similarity is that the risk (or odds) of developing prostate cancer does not appear to increase when the level of exposure increases.

The noticable difference between the results of the two models is that at each level the case-cohort model has lower risk values and shorter confidence intervals than the case-control model. One possible reason for this difference is due to the extra information that is used in the case-cohort study but not in the matched case-control. Consider just this extra information. The reason the risks are lower in the case-cohort is that there are more controls with more exposure, which reduces the the relative risk. Recall the partitioning of the source population, re-displayed in Table 3.7,
and the Relative Risk (RR)

	Disease	Non-Disease	
Exposed	A_1	B_1	$A_1 + B_1$
Unexposed	A_0	B_0	$A_0 + B_0$
	$A_1 + A_0$	$B_1 + B_0$	

Table 3.7: Two by Two Contingency Table For Calculating Risk

$$\begin{aligned} RR &= \frac{\text{probability of disease given exposed}}{\text{probability of disease given unexposed}} \\ &= \frac{A_1/(A_1 + B_1)}{A_0/(A_0 + B_0)}. \end{aligned}$$

If there are more controls who are exposed, then the value of B_1 will increase and the value of B_0 will decrease. In other words, if during this period, there is an increase in controls who fall into the high levels of exposure, the relative risk will decrease. This is a possible explanation for why the risks are lower in the case-cohort study when compared to the matched case-control study. Thus, if the proportion of exposed and un-exposed controls does not change even when this additional information is included, the two methods should give similar risks. Similarly, if there is a shift towards fewer exposed controls, one would expect the risks to be higher in the case-cohort study than the matched case-control study. Therefore, it depends on the situation which method will show higher risks. However, assuming this explains the differences in the risks, it can be argued that the case-cohort method gives a more accurate interpretation of what is going on. Thus, it is more appealing. It would be possible to determine if the controls fall into the high levels of exposure during this period of time; however, this would require complicated work involving linking two data sets; thus, it was not done for this project.

Furthermore, in the case-cohort method, we used more information than in the matched case-control study; therefore, in a sense we have a larger sample size to calculate the estimated confidence intervals. Therefore, we would expect to obtain smaller confidence intervals, which suggests that the case-cohort method is the more accurate of the two. However, since the methods of analysis are very different, one should be

cautious about comparing the two estimated confidence intervals and drawing any conclusions from them.

3.5.4 Problems with the Stability of the Model

Although the model worked nicely for the exposure levels given below, there does appear to be a problem with the stability of the method; the estimation method (i.e. maximization) does not always converge. Clearly, there will be problems with the model when there are intervals with missing cells (i.e. no cases or no controls). If there is a cell with no count, the estimate will be zero or infinity. Therefore, when such a situation arose, it made sense that the estimation procedure did not converge. In addition, it is desirable to not have low numbers of individuals in each cell since small counts can lead to convergence problems. However there was one situation that evoked suspicion (Table 3.8). When these cut-points were used, the estimation method failed to converge, but there were no problems in a situation that was very similar (Table 3.9). The two situations have the same number of levels with approximately the same number of events in each interval. Therefore, the fact that the estimation procedure failed to converge for one and not the other needs to be investigated further.

In the above two situations, the number of events in the exposure levels 1 and 2 are different. It may be useful to examine the case that is switching levels when the cut-points are changed. It may, somehow, be contributing to the convergence problem. Another cause of the problem may be that the estimates are going off to positive or negative infinity, or it may be finding a local maximum. In order to determine if this is what is happening, one could look at the value of the estimate at each interval of the maximizing procedure. This was not done in this project, but it is a possible avenue for future work.

Initially, we looked into using different initial values in the S-Plus Cox Proportional Hazards function. Although this changed the output, it did not change whether or not the method converged.

Next, we changed the length of ranges in the situation where the estimation procedure failed. When only the range $(0, 0.45]$ was changed to $(0, 1]$, the estimation

Exposure Levels	Range	Records	Events
0	0	62846	247
1	$(0, 0.45]$	668	6
2	$(0.45, 1.54]$	450	6
3	$(1.54, 3.92]$	761	6
4	$(3.92, 10.37]$	779	6
5	$(10.37, 19.06]$	676	9
6	> 19.06	937	7

Table 3.8: Exposure Levels that did not Converge

Exposure Levels	Range	Records	Events
0	0	62846	247
1	$(0, 0.56]$	668	6
2	$(0.56, 2.33]$	441	5
3	$(2.33, 6.62]$	755	6
4	$(6.62, 12.10]$	779	6
5	$(12.10, 26.98]$	680	10
6	> 26.98	937	7

Table 3.9: Exposure Levels that did Converge

procedure converged. Based on this, we tried upper cut-points between 0.45 and 1.0 to determine the minimum value of the upper end that could be used so that the method converged. It was found that with 0.5 the method did not converge, but with 0.6 the method did converge. Using bisection, we concluded that at 0.56 the estimation procedure did not converge, but at 0.57 the estimation procedure did converge. The next step would be to use the estimate from the estimation procedure that did converge as the initial value in the situation we were initially concerned about (Table 3.9). If the method then converged, it may be a matter of changing the initial value. However, if it still did not converge, this would suggest a more serious problem. The problem may be with the algorithm that the function in S-Plus is implementing. If this is the case, using the Cox Proportional Hazards function in S-Plus may not be satisfactory.

Chapter 4

Conclusion

Pulp and paper is a major industry in British Columbia. During the pulping process, pulp and paper mill workers are exposed to known or suspected carcinogens. In 1982, an occupational cancer research program was launched in British Columbia. One branch of this research program was directed towards detecting occupational cancer risk factors. Based on preliminary findings, a two-phase study of British Columbia pulp and paper mill workers was launched by the British Columbia Cancer Agency.

Phase I was a cohort study that was divided into 2 sub-studies: a cohort cancer mortality study and a cohort cancer incidence study. The former reported the cancer mortality of 30,157 pulp and paper workers in British Columbia. This study reported cancer risks significantly associated with work duration and time from first employment of 15 years or more were observed for cancers of the pleura, kidney and brain in the total cohort, for kidney among the kraft mill workers only, for Hodgkin's disease among the sulfite mill workers only, and for esophageal cancer among the workers employed in both kraft and sulfite mills.

The cohort cancer incidence study used 28,278 members of the British Columbia pulp and paper cohort. This study found that long term work in the industry is associated with an excess risk of skin melanoma, prostate and pleural cancers. The excess risk of pleural cancer was explained by past exposure to asbestos. Since the incidence rates of prostate cancer and skin melanoma in British Columbia are high relative to Canadian rates, the data was re-analysed using British Columbian rates rather than

Canadian rates. It was found that skin melanoma was no longer significant, whereas the relative risks for prostate cancer remained significantly elevated among long term workers. The potential exposure that might be associated with prostate cancer were investigated in phase II of the study.

In the first phase of the study of British Columbia pulp and paper workers, no attempt was made to classify workers by departments and no exposure data were obtained that might provide explanations for the difference in cancer patterns observed between the kraft only and sulfite only workers. Thus, mill-specific and period-specific job exposure matrices were developed for a matched case-control study with detailed exposure assessment by title. Therefore, Phase II, which is the matched case-control study, should enable one to evaluate whether the excess risk for prostate cancer reflects the exposure among subsets of workers.

The matched case-control study was comprised of 287 cases (workers who were diagnosed with prostate cancer) and 1,710 controls (workers who were healthy at the end of the follow-up). In this project we focused on the results for one chemical, black liquor. The results of this study indicated that the exposed individuals are significantly more likely to be diagnosed with prostate cancer than the un-exposed individuals. However, the odds of being diagnosed with prostate cancer does not appear to increase when exposure increases.

The aim of this project was to apply the case-cohort method to the pulp and paper worker data in order to determine if this method is more appealing than the matched case-control method. We were successful in computing the estimates for the case-cohort model by using the Cox Proportional Hazards model in S-Plus. As we expected, the results from this model were similar to those found in the matched case-control analysis. We found that the risk of developing prostate cancer is much higher for the exposed individuals than the unexposed individuals. However, the risk does not appear to increase when exposure increases.

Although the trends are similar for the matched case-control, the risks are consistently lower in the case-cohort model. Recall the key difference between the two methods is that in the matched case-control study the controls are only followed until

their matched case experiences the event, whereas in the case-cohort study the controls are followed until the end of the follow-up period. Therefore, in the case-cohort method we have this additional information. The following comments are based on hypotheses rather than fact since the work required to verify the hypotheses was complicated and not done in this project. Now, consider only the information on the controls that is included in the case-cohort method, but not the matched case-control method. If during this period, there is an increase in controls who fall into the high levels of exposure, the relative risk will decrease. This is a possible explanation for why the risks are lower in the case-cohort study when compared to the matched case-control study. In addition, the confidence intervals are smaller in the case-cohort analysis than in the matched case-control analysis. This is another indicator that the case-cohort may be more appealing method.

Although the case-cohort method appears to be more appealing than the matched case-control method, future work must be done on the stability of the estimation procedure. As was mentioned in Chapter 3, the procedure does not always converge. There are certain situations, such as when one of the category levels contains either no events or no controls (or a small number of either), where the estimates will be undefined. However, the method still failed to converge in one situation where it would seem that it should not have had a problem. It is possible that the cut-points that were chosen for one of the categories were not appropriate for some reason (e.g. the range within the cut-points may have been too small). How cut-points should be chosen needs to be examined more rigorously. In addition, this stability problem requires further investigation before the case-cohort model can be recommended over the case-control model.

As a final note, thus far the case-cohort model has only included one chemical (i.e one time-dependent covariate). Since the workers were, in general, exposed to more than one chemical, it is quite conceivable that more than one chemical, as well as possible interactions between the chemicals, needs to be accounted for in the model. How this can be done is one potential avenue of future work.

Bibliography

[1] Band, P.R., et al. (1997). "Cohort Mortality of Pulp and Paper Mill Workers in British Columbia, Canada." *American Journal of Epidemiology*, **146**, 186-94.

[2] Band, P.R., et al. (2001). "Cohort Cancer Incidence among Pulp and Paper Mill Workers in British Columbia." *Scandinavian Journal of Work and Environmental Health*, **27(2)**, 113-19.

[3] Barlow, W.E. (1994). "Robust Variance Extimation for the Case-Cohort Design." *Biometrics*, **50**, 1064-72.

[4] Breslow, N.E. and Breslow, N.E. (1980). "Statistical Mehtods in Cancer Research. Volume I: The Analysis of Case-Control Studies." Oxford University Press, Oxford.

[5] Breslow, N.E. and Breslow, N.E. (1987). "Statistical Mehtods in Cancer Research. Volume II: The Design and Analysis of Cohort Studies." Oxford University Press, Oxford.

[6] Cox, D.R. (1972). "Regression Models and Life Tables (with discussion)." *Journal of the Royal Statistical Society. Series B*, **34**, 187-220.

[7] Cox, D.R. (1975). "Partial Likelihood." *Biometrika*, **62**, 269-76.

[8] Gallagher, R.P., Fleshner N. (1998). "Prostate cancer: 3. Individual risk factors." *Canadian Medical Association Journal*, **159(7)**, 807-13.

[9] Prentice, R.L. (1986). "A Case-Cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials." *Biometrika*, **73**, 1-11.

[10] Prentice, R.L. and Pyke, R.L. (1979). "Logistic Disease Incidence Models and Case-Control Studies." *Biometrika*, **66**, 403-11.

[11] Rothman, K.J. and Greenland, S. (1998). "Modern Epidemiology." Lippincott-Raven Publishers, USA.

[12] Self, S.G. and Prentice, R.L. (1988). "Asymptotic Distribution Theory and Efficiency Results for Case-Cohort Studies." *Annals of Statistic*, **16**, 64-81.

[13] Therneau, T.M. and Li, H. (1998). "Computing the Cox Model for Case Cohort Designs." *Techincal Report Series: Section of Biostatistics*, 1-25.