ANALYSIS OF NICKEL CONTAMINANTS IN SEVERAL GROUPS OF CASE-CONTROL SETS GENERATED WITHIN THE FALCONBRIDGE MINING COHORT.

ANALYSIS OF NICKEL CONTAMINANTS IN SEVERAL GROUPS OF CASE-CONTROL SETS GENERATED WITHIN THE FALCONBRIDGE MINING COHORT.

BY

SERGIO R. ESCOBEDO, B.Sc., M. en C.

A Project

Submitted to the School of Graduate Studies
In Partial Fulfilment of the Requirements for the

Degree
Master of Science

McMaster University

MARCH 1992

MASTER OF SCIENCE (1992) MCMASTER UNIVERSITY
(Statistics) Hamilton, Ontario

TITLE: ANALYSIS OF NIKEL CONTAMINANTS IN SEVERAL GROUPS OF CASE-CONTROL SETS GENERATED WITHIN THE FALCONBRIDGE MINING COHORT.

AUTHOR: SERGIO R. ESCOBEDO ROMO

SUPERVISOR: DR. HARRY H. SHANNON

NUMBER OF PAGES: vi, 79

ABSTRACT

The Falconbridge study was a historical prospective mortality study conducted at a nickel company in Sudbury Ontario. The study included all the men that worked there for at least 6 months between 1950 and 1976 with an update in 1984. Nearly 11500 subjects were included in the cohort. From here all the subjects identified with lung cancer were selected and matched with different number of healthy subjects (1:1, 1:2, 1:4, 1:6, 1:8, 1:10).

Information was generated of the time-intensity of different contaminants normally found in the mining atmosphere and eight of such contaminants were evaluated for the different sets of case:control. Given the fact that some contaminants present a time delay effect, an assesment of this factor was also made using the subject initial and final time of exposure to each contaminant.

Finally a variable selection was performed for the three data sets generated in the project.

The results showed that some of the postulated contaminants did have an effect in the final outcome, however, because of the complex effect between the different chemical compounds a definite conclusion is not given.

ACKNOWLEDGEMENTS

I wish to express my sincere thankfulness to my supervisor Dr. Harry Shannon for his time and patience during the project; to Dr. Peter Macdonald and Dr. Mary Lesperance for their assistance and understanding.1

Unlimited thanks to McMaster University and the Mathematics and Statistics Department for financial support.

To my wife Laura for her constant encouragement and tolerance.

To Richard and David Escobedo with love.

To my parents.

TABLE OF CONTENTS

CHAPTER I

1.1 INTRODUCTION

1.1.1 Project overview

The principal aim of the present project is to explore the relationship between the different contaminants that exist in nickel and processing plants in Falconbridge and the suspected increase in lung cancer that has been reported (Shannon et al, 1984, Shannon, 1990 to be published). The Falconbridge study was carried out as a prospective study involving a long follow-up period (1950-1984) and using a cohort of 11,567 subjects. The Standardized Mortality Ratio (SMR) was used to evaluate total mortality by selected causes, by length of exposure, exposure category, sector category and others. The SMR is the ratio of the observed number of deaths attributed to a particular cause divided by the number of deaths that would have been expected in an equivalently aged group.

A retrospective study is set up that selects those subjects who have been detected as having lung cancer matching them with a different number of controls in a similar way as described by Liddell (Liddell, 1977).[1]

---

[1]To decrease the possible effect of birthday and initial working date, the matching of the controls with the cases was made keeping those two variables as equal as possible.

Two files were used as a primary source of demographic information and working history. These files were derived from the same Falconbridge study. The first file, includes the first working day, the final working date and the subject birthdate, among other variables. The second file is the personal working history of all workers, including each department where the subject worked and the time spent in each.

Different estimates of contaminants were made based on various actual measurements and given the assumption that most of the departments in the mine had a more or less constant environment. A mean estimate of the different nickel compounds for each subject was made. While this assumption might not be strictly true, it is an indication of the relative levels of different contaminats between the different working areas.

Once the different exposure indices were calculated for each subject, they were multiplied by the number of days spent in each department. This produced a data set that was analyzed using conditional logistic regression in an attempt to asses the impact of each variable on the final outcome.

Although this project is statistically oriented, a description of the methodology used is also given.

1.2.1 Nickel and its correlation with cancer.

The first evidence linking nickel workers with an increased risk of lung and nasal cancer was published in 1933 (Bridge, 1933) from refinery workers at Clydach, Wales. Hill (1939) in an unpublished report, describes the first epidemiological study of nickel workers finding a great increase in the risk of dying of cancer, compared with the population of England and Wales. Other reports (Doll, 1970) provide strong evidence relating lung and nasal sinus cancers with specific parts of the refining process.

Traditionally in Canada, there has been a major interest in respiratory cancer and the nickel exposed populations. Large excesess of lung and nasal cancer had been detected in Port Colborne Ontario (Chovil A. and Sutherland R., 1981), particularly in the sinter plant. Although cancer in sites other than the lung and nasal sinuses, (stomach and kidney) have also been found exceeding the expected numbers, these results have not been replicated and are thus doubtful.

One approach in the studies to date has been to analyze isolated groups of workers exposed a number of years to nickel.

The results showed that workers at high risk include individuals involved in roasting, calcining and sintering the nickel ore (Report of the International Committee on Nickel Carcinogenesis in Man, 1990).

At this point there are still doubts of the identity of the possible carcinogens, but most of the studies tend to point to the soluble forms of nickel especially in the Kristiansand refinery in Norway (Magnus K, et al, 1982), where soluble nickel appeared to be the primary lung cancer hazard even though there was evidence that oxidic nickel may also be responsible for the nasal cancer found there, with a concentration range 1-5 mg Ni/m^3.

Other nickel forms which seem to play a role promoting lung and nasal cancer risk included: sulfidic nickel and oxidic nickel.

1.3.1. The Falconbridge Study.

In 1984 Shannon et al published the results of a historical prospective mortality study conducted at the Falconbridge Nickel mines in the Sudbury area of Ontario, that includes a description of the process with the different mining phases. In a historical prospective study subjects are selected from registers according to whether or not exposure to the putative casual factor has occuurred or not. The exposure is not under the control of the investigator and as a result the groups may not be comparable and the outcome is present at the time of sampling. The extended follow-up found one nasal cancer and the SMR was significantly high overall in the miners.

Detailed analysis by date of first mining and mining

duration did not show a consistent trend with an occupational etiology tendency. In general there were no significant increases in lung and nasal neoplasms.

Also, the same extended follow up explored the relationship that may exist with the environmental conditions including the person-years observations by age and time since first employment, the distribution of length of exposure, total mortality by major cause of death, total mortality beyond 15 years from first exposure by major cause of death groups, mortality beyond 15 years from first exposure by selected causes, mortality by sector category for major cause of death groups, mortality from a priori causes of interest by sector category beyond 15 years from first exposure in that sector, and many more.

Smoking habits data were not gathered nevertheless, in view of the absence of respiratory cancer, it was speculated, that tobacco consumption may even been less than in the general population.

As indicated in the study, except for mining no discrepancy was found between the observed and the expected number of subjects with the disease, however no further evaluation was made for the different occupations inside the different departments.

The general procedure for the assessment involved the use of SMR which takes into account the expected compared against the observed value. This is the classical way to analyze a

cohort.

1.3.2 The Falconbridge Mine Workplace Environment

The reliability of an occupational epidemiology study depends in large measure on the amount, specificity and precision of the exposure data (Checkoway 1989). Thus the estimation of the dose-response relationship depends on the exposures and it is worth describing a few of them.

Exposure in the occupational environment setting is defined as "the presence of a substance in the environment external to the worker". Exposure levels are usually evaluated in accordance to the intensity of the possible contaminant and the duration-time of the contact. According to Shannon et al (1991) two senior company personnel, reviewed and summarised the environment data using konimeter counts; measuring dust in particles per cubic centimetre on the different workplaces of the mine. These measurements were taken occasionally before 1960 and later as part of a regular semi-annual survey from 1960 to 1984. Gravimetric sampling, measuring total dust in milligrams per cubic meter were conducted from 1978 and onwards. Although dust concentration changes from time to time, it is frequently used in epidemiological studies using the average concentration to represent the intensity of the contaminants.

A so called side-by-side program of sampling was instituted to compare the two methods of measurement in each

of the main sectors of the mine-mining, milling and smelting. Regression lines were obtained to predict an estimate from the counts to gravimetric measures. For periods when data was not available an estimate was made taking into account what was known of work practice, ventilation and production (Shannon et al 1991). Again averages of the contaminants concentrations were used for the whole period. While this procedure has its limitations because of the logical changing conditions of any work setting, the exposure is usually identified by both the concentration of the suspected species and the duration of the exposure. These two combine into a measure known as cumulative exposure, which is the sum of the concentration over time. Thanks to the work history file, the levels of the different nickel contaminants were used to calculate a time-exposure measurement known as cumulative exposure, that is an integration of the concentration in time. Changes of workplace was also translated as a change in concentration and duration of exposure.

An important term in epidemiology is dose. Dose is defined as the "amount of a substance that remains at the biological target during some specified time "interval" In the simplest model it is assumed that a linear relation between cumulative exposure and dose exists producing the basis for the association and possible causality. However this is not always the situation and thus more complex models are needed that involved patterns of association, retention,

detoxification routes and in general more physiological information that may assist in improving the approximation of the doses (Checkoway, 1989).

1.4.1 Study Objetives

i) The present study selects several controls (1,2,4,8, and 10) for each case (lung and nasal cancer) using subjects from the same cohort and matching them with its corresponding case.

ii) Exposure indexes were calculated using the work history file for each individual and the effect of each nickel compound was assessed through the conditional logistic regression. An estimate of the precision was made using different numbers of controls.

A univariate analysis was performed with the purpose of detecting each contaminat effect. Upon completation variables were selected for a multivariate analysis. The initial and final working dates effects were also evaluated but only in a univariable way.

1.5.1 General outline of the study

This study is basically divided into five chapters. Chapter one includes an outline of the project and its objectives. In chapter II, a review of the statistical methodology related with the project is given. These methods

are important in epidemiological work specially in the case-control context. The main emphasis in the chapter is on the logistic regression model which arises in the context of the proportional hazards model proposed by Cox (1972). Most of the theoretical concepts are simply described and because of the nature of the project there are no mathematical demonstrations or proofs. Chapter III presents a description of the data files both the ones provided by the Falconbridge study and the ones generated by this project. These last data files were used to evaluate the effect of the different chemical species on the outcome (lung carcinoma). Using various numbers of controls for the comparison permits an evaluation of the precision in the estimate.

It is known that diverse chemical compounds exercise their effect at different times. With this in mind the initial and final dates of exposure for every chemical species were generated for each individual. The variables were assesed through the conditional logistic model, using the initial and final dates as variables in the univariate model. Chapter IV provides the results. The discussion and general conclusions are included in the final chapter.

CHAPTER II

2.1 General Methodology of Case and Control Studies.

The purpose of this chapter is to present a brief review of some concepts related with the selection of the cases and controls for the present project.

Several well known books and references have an in depth discussion of the topic including the specific situation of a case-control within a cohort (Schlesselman 1982, Breslow and Day 1980, Liddell et al 1977 and Checkoway et al 1989).

2.1.1 Historical Development

Breslow and Day (1980) document the first case-control study by Lane-Claypon of the role of reproductive experience in the etiology of breast cancer .The report gives a description of the methods for selecting matched hospital controls. An earlier report that also uses the same case-control approach was published by Broders 1920, (In Breslow and Day 1980) related with the development of squamous cell epithelioma of the lip.

The final conclusion was although the percentage of tobacco use was approximately the same, among the cases nearly 80% smoked pipes while among the controls only about 40% were

pipe-smokers, hinting that pipe smoking was an important factor in the etiology of the disease.

In other fields like social science this same methodology has been in practice for nearly the same time. But it was after World War II that this kind of study started being used in a more systematic and rational way. A turning point was the paper presented by Cornfield (1951), where he developed a measure linking the exposure frequencies of cases and controls and simply transforming it into the ratio of the frequency of disease among exposed individuals compared to those that are non-exposed in other words the relative risk. If the risk of the occurrence of the event D when E is present is taken as the rate of D's occurrence specific to the presence of E, and likewise for the risk of D when E is absent, then the relative risk is simply the ratio of these two risks

$$\mathbf{R} = \frac{P(D|E)}{P(D|\bar{E})}$$

The relative risk can be estimated from a prospective study (a cohort for example) or approximated in a retrospective study (case-control). In the last case (retrospective study) we are actually estimating the ratio of the probability of E given the event D and the probability of E given the absence of D. An estimator of the relative risk in retrospective studies follows straight from Bayes theorem and the low

incidence assumption i.e. the P(D)→0. The final estimator is called the odds ratio (Fleiss 1981) :

$$\Psi = \frac{P(E/D)\,/\,P(\bar{E}/D)}{P(E/\bar{D})\,/\,P(\bar{E}/\bar{D})}$$

Mantel and Haenszel (1959) showed how to estimate the relative risk on a pooled population and test it via chi-square to summarize the data in several strata. This paper is still considered one of the foundations of the design of case-control studies and is also one of the most widely cited references in the medical literature (Bailar and Anthony, 1977).

As an example of case-control study, we might refer to the study of leukemia in the American rubber industry (Wolf et al 1981). In this study a group of 72 cases of leukemia occurring among the employees of four rubber and manufacturing companies during the period from 1964 to 1973. The aim was determine if certain environmental factors were related with the leukemia cases . Earlier studies in one company hinted the association of lymphatic leukemia with a possible work history of solvent exposure. All the leukemia deaths were identified from death certificates from the life insurance records of the four companies. A definition of a case was given as "any active or retired hourly rubber worker who died of leukemia in that period from 1964 to 1973". So all the cases were identified by reference on the death certificate to the Eighth

revision by ICDA (International Classification of Diseases) code 204-207 as it was assigned by a nosologist. In this way all 72 cases were identified for the study period. A brief description is given of the exclusion subjects since it is known that all the excluded potential cases have an impact on the final analysis, thus on the conclusions. The matching factors were race, sex, rubber plant and date of birth for two of the four controls and because of this were named "loose" controls.

The two controls left were additionally matched by date of hiring and named "tight" controls. According to the authors this method of matching allows examination total duration of total duration of employment using the "loose" controls. Comparing the cases with the "tight" controls allows to explore specific workplace differences. However it was not possible to control other possible confounders such as exposure to external sources of radiation, use of drugs or exposure to other non occupational chemical leukemogens. The results of the study showed no significant elevation in the odds ratio for all the companies combined or for the each company taken individually. Also no difference was found for general services or any other occupational group investigated.

2.1.2 Cohorts and Case-Control Studies.

The aim of most epidemiological studies is to establish whether some exposure represents a health hazard

(Breslow and Day, 1987). The case-control study is retrospective in nature implying with this, that the individual or subjects with a certain condition are selected and compared with a series of individuals or subjects for which the condition is absent. Cases and controls as a group are then compared with respect to past exposures that might be important in the etiology of the disease .

From what has been said, the most important difference between a case-control and a cohort study happens because of the very nature of the study, the cohort study follows subjects who are in theory free of the disease and "follows them" over a large period of time developing rates of the disease if the group has been or not exposed to a certain risk factor. On the other hand, the case-control protocol chooses subjects depending on the presence or absence of the study disease relating it with the possible exposure factor. As has been mentioned it is possible to make an assessment of the risk on both study types via the relative risk parameter (Breslow and Day 1980).

Because of the nature of the study it is important to keep a balance between the possible generalization of the study with its validity. Validity refers to the fact that cases and controls should be as much alike as possible, in other words all possible factors that might have an influence in the outcome (disease) and that is present in a differential manner on both groups will produced a confounded effect with

the risk factor. It should be as similar as possible in the group of cases as in the controls. On the other hand generalization can be achieved when all the cases with the disease and match them with a certain numbers of controls from that same population. If subjects are extremely heterogeneous, then the random variation will prevent any possible difference, that may exist between cases and controls. This in turn will produce a lack of validity and so on.

2.1.3 Advantages and Disadvantages of Case-Control Methods.

Some of the principal strengths and weakness of this method are presented in table 1 (Schlesselman, 1982).

TABLE 1

CASE-CONTROL STUDIES

ADVANTAGES	DISADVANTAGES
Well suited to the study of rare diseases or those with long latency.	Relies on recall or records for information on past exposure.
Relatively quick to mount and conduct.	Validation of information is difficult or sometimes impossible.
Relatively inexpensive.	Control of extraneous variables may be incomplete.
Requires comparatively few subjects.	Selection of an appropriate comparison group may be difficult.
Existing records can occasionally be used.	Rates of disease in exposed and unexposed individual cannot be determined.
No risk to subjects.	Method relatively unfamiliar to medical community and difficult to explain.
Allows study to multiple potential causes of a disease.	Detailed study of mechanism is rarely possible.

Matching is another important point in the design of cases and controls. It deals with the pairing of one or M controls for each case, based on their similarity with some selected variables (Schlesselman,1982). The principal objective of matching is to permit the use of efficient analytical methods (regression, standardization, etc.) to control confounding by the factors matched for. A possible confounder is a factor that has the following characteristics (Anderson et al, 1980):

1) The risk groups differ on the factor

2) The factor itself influences the outcome.

This factor can be wholly or partially responsible for the apparent effect of the study exposure or mask an underlying true association (Schlesselman, 1982).

Several examples from the literature can exemplify the mechanisms of how a confounder works and the different relations that it has with the disease and the risk factor or exposure (Shapiro et al 1979, Steckel 1976). However the most important factor that limits the use of cases and controls as a research methodology lies in the fact that they are highly susceptible to several bias types particularly selection and recall bias, difficult to avoid due to the presumption that cases tend to consider more carefully the causes of their disease than controls do (rumination) (Sackett, 1979). The problem is that the information on exposures relies on a subjective source.

2.1.4 Matching

Matching is always a possibility to be considered when dealing with case-control studies. During the design phase of the study careful consideration should be given to how many, and which factors will be used for matching. Otherwise the project could be jeopardized, since matching is by far the most popular way of controlling possible confounders.

However if an excessive number of factors (sex,age,race,

etc) had been used, the case and control groups would be so similar that they could provide no useful information with the consequent loss of time and resources.

Some considerations that should be taken into account when matching are (Breslow, 1982):

i) Matching is only justified for factors which are known to confound the association being tested.

ii) Matching may also be justified for those factors which could interact with the exposure risk of interest in producing the disease.

iii) It is usually possible to justify the costs in time and money matching for age, sex and nominal scale variables with a large number of categories (sibship, neighbourhood).

iv) The matching should be as close as possible.

This last point is especially important when children or young adults are part of the study given the fact that one or two years of difference may have impact because its relatively greater proportion than it is in middle or young age.

2.1.5. Case-control studies generated within a cohort.

A cohort is a group of subjects selected with a specific reason. Although the subgroup used in this project was generated using a large cohort it is not the intention here to discuss the usefulness of such study types.

Breslow (1987) gives a full description of this methodology. However not using all the information generated in a cohort will be a waste of resources, so one way to generate answers to specific hypothesis is to generate a comparison between the small group of cases with the disease of interest and a number of controls. One way to address the question on how to select the controls is to form a subcohort from the beginning. Prentice (1986) discusses such a possibility. This has been termed a case-control design. A few obvious advantages of sampling in this way are: saves time and resources compared with the cohort studies with relatively small loss of precision. They have the advantage of enabling the study of an exposure when the cohort group cannot be enumerated feasibly. For studying subjects and exposures that have occurred in scattered small workplaces and finally the evaluation of several different risk factors is permitted. On the other hand case-control studies are more susceptible to bias than cohort studies, but in the specific case of a cohort-based case-control study they are no more prone to bias than cohorts.

In the Falconbridge study we are dealing with a historical cohort so no such possibility exists. Only the retrospective study is feasible and from this point of view a case-control study may be used to clarify an initial hypothesis that requires further information or even generate several new hypothesis through an exploration of some

variables. Some examples may clarify this point. Jansen (1979) in a study of brewery workers found an excess of oesophageal cancer. The question of the possible relation between beer and the development of oesophageal cancer was subsequently answered when a case-control study showed the relation between heavy beer drinking and the increase risk of oesophageal cancer (Breslow and Day 1987).

Once the question is established and the cases have been identified, it is time to select the controls. The most recommended procedure and the one that is used in this project is:

First, select from the risk set (R_i) all the cases (d_i) that develop or die from the disease of interest at time t_i.

Second, select the controls (m_i), at random and without replacement from among the (g_i) members of (R_i), who do not have the disease at that time.

The total cases (d_i) and sample of controls constitutes a reduced risk set (R_i*).

The theoretical argument that supports this sampling scheme is that the number of potential controls (g_i) is infinite. This implies that there will be no overlap between the controls sampled from different risks sets, which is not strictly true especially with restricted groups such as the elderly.

Lubin and Gail (1984) suggested that this approach is invalid because of the fact that some controls might appear

several times not yielding in this way more "new" information.

One alternative proposed by the same authors is to exclude controls that have previously been chosen, but include them as cases if they developed the disease. Naturally the original cohort should be very large for this approach to be valid otherwise the sample is biased (Robins et al, 1986 (a)). Other sampling schemes have been described trying to reduce this problem.

2.1.6. Number of controls.

Urey (1975) developed a simple way to estimate the theoretical efficiency of a 1:M case-control using a relative risk of 1. He found that a 1:1 ratio of case-control is 50% efficient compared with the total population (100% efficient), while matching by 4 controls brings the efficiency to 80% Using 5-10 controls brings a repid decrease in the efficiency.

2.2 Statistical Methodology

2.2.1 Introduction.

The most popular and versatile model that relates the disease probability with the corresponding levels of exposure of the risk factors is the linear logistic model also known as the logit model. What characterizes this model from the linear regression model is that the outcome in the linear logistic is binary or dichotomous. Once this difference is understood the principles that rule the linear model also govern the logit

in the logistic model. Despite the logistic model other appealing probability distribution functions have been proposed. Cox (1970) examine several models. There are two main reasons (Hosmer 1988) to choose the logistic model: (1) It is extraordinarily flexible from a mathematical point of view and (2) the results are biologically explicit.

The specific form of the logistic model using Cox (1970) notation is as follows. Let $Y_1,\ldots,Y_n$ be an independent dichotomous random variable, then the linear logistic model is:

$$\lambda_i = \log\left(\frac{\Theta_i}{1-\Theta_i}\right) = \boldsymbol{x}_i\beta = \sum_{s=1}^{p} x_{is}\beta_s. \qquad (1)$$

where $\{x_{is}\}$ (i=1,...,n; s=1,...,p) are known constants $\beta_1,\ldots,\beta_p$ are unknown parameters and $\Theta_i=P(Y_i=1|x_i)$.

The likelihood of an observed binary sequence $y_1,\ldots,y_n$:

$$prob(Y_1 = y_1,\ldots,Y_n = y_n) = \frac{\exp\left(\sum_{s=1}^{p}\beta_s t_s\right)}{\prod_{i=1}^{n}\left(1+e^{\boldsymbol{x}_i\beta}\right)}. \qquad (2)$$

where

$$t_s = \sum_{i=1}^{n} x_{is}y_i \qquad (3)$$

is the observed value of the random variable

$$T_s = \sum_{i=1}^{n} x_{is} Y_i$$

2.2.2 Interpretation of the Regression Coefficients.

The logit transformation of Θ_i produces a linear model . The logit is linear in the parameters and may be fitted by a iterative method namely the maximum likelihood. Once the parameters estimates are computed, the next step involves the interpretation of the coefficients β_i. The interpretation of any fitted model requires the ability to generate practical answers to the research question or questions. The estimated coefficients for the independent variables x_i in the model represent the rate of change of the logit of the probability that $Y_i=1$ per unit of the independent variable.

The first step in finding the functional relationship between of the dependent and the independent variables is to find out what relation generates a linear model. In the case of observations with normal distribution the canonical link function (McCullagh and Nelder 1984) is the identity function (y=y in the general linear model context). In the logistic regression model case the canonical link function is the logit transformation.

The most simple way of explaining the parameters in a logistic regression model is via a dichotomous independent variable. Suppose that the independent variable x_i is coded

either zero or 1. The associated probabilities are summarized in the following table

TABLE 2

Dependent Variable Y	Independent Variable X	
	X=1	X=0
Y=1	$P(Y=1\|X=1)=\theta_1 = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$P(Y=1\|X=0)=\theta_0 = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
Y=0	$P(Y=0\|X=1)=1-\theta_1 = \frac{1}{1+e^{\beta_0+\beta_1}}$	$P(Y=0\|X=0)=1-\theta_0 = \frac{1}{1+e^{\beta_0}}$

The odds that Y is 1 for individuals with x=1 is defined as $\theta_1/[1-\theta_1]$. Likewise, the odds that Y is 0 for individuals with x=0 is defined as $\theta_0/[1-\theta_0]$. The log of the odds is called logit so in our context:

$$\lambda_1 = \ln\left[\frac{\theta_1}{1-\theta_1}\right] \qquad (5)$$

and the complement

$$\lambda_0 = \ln\left[\frac{\theta_0}{1-\theta_0}\right] \qquad (6)$$

From here the odds ratio is easily defined as the ratio of the odds for x=1 to the odds for x=0.

$$\psi = \frac{\theta_1/[1-\theta_1]}{\theta_0/[1-\theta_0]} \qquad (7)$$

The log of the odds ratio is known as log-odds and defined as

$$\ln(\psi) = \ln\left[\frac{\theta_1/[1-\theta_1]}{\theta_0/[1-\theta_0]}\right] \qquad (8)$$

Using the expressions on table 2 and inserting them on equation (7)

$$\psi = \frac{\left[\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right]\left[\frac{1}{1+e^{\beta_0}}\right]}{\left[\frac{e^{\beta_0}}{1+e^{\beta_0}}\right]\left[\frac{1}{1+e^{\beta_0+\beta_1}}\right]} \qquad (9)$$

and simplifying

$$\psi = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1} \qquad (10)$$

So the logit difference or log-odds is

$$\ln(\psi) = \ln(e^{\beta_1}) = \beta_1 \qquad (11)$$

In this way the direct interpretation of the coefficients is one of the fundamentals reasons of its use and has turned the logistic regression model into a powerful research tool in epidemiology.

The interpretation of the odds ratio is also based on the fact that it is approximately equal to another quantity called the relative risk. Oftentimes the event Y_i represents the presence of the disease (D) in individual i and x_i the event that s/he is exposed (E) to a suspected cancerigen. If the risk of occurrence of an event say D when E is present is P(D/E), and likewise the risk of D when E is absent, P(D/E) then the relative risk is simply the ratio of this two risks

$$R = \frac{P(D/E)}{P(D/\overline{E})} \qquad (12)$$

R may be estimated only in a prospective study and it can be demonstrated that the odds ratio is an approximation of R when the incidence of the disease is low (Fleiss 1981).

2.2.3 Statistical Adjustment

Even though in the last two sections the main emphasis has been to discuss the logistic model only in terms of a single variable, in practice multiple variables models are very common. One of the goals of doing this is to statistically adjust the estimated effects of each variable in the model and look for differences in the distribution of the other variables. Allaying this idea to the multivariate logistic model it means that each estimated coefficient provides an estimate of the log odds adjusted for, or taking into account all the other variables included in the model. For example a situation that frequently happens in

epidemiology is, when there are two variables one continuous and the other dichotomous. Both variables are important but the dichotomous is of particular importance and we would like to know the effect of this variable adjusted by the other variable. The situation is similar to the analysis of covariance in the linear regression situation. Armitage (1971) gives a full description.

2.2.4 Fitting the Logistic Regression Model

Suppose that we have a sample of n independent observations of the pair(x_i,y_i), where i=1,2,...,n, where y_i denotes the value of a dichotomous dependent variable and x_i is the vecotr of the independent variable for the i_{th} subject. Furthermore assume that the independent variable can take the values 0 or 1 only. To fit the logistic model from a data set it is required that a iterative method be implemented to find the values for the unknown parameters $\beta_0,....,\beta_p$. Unfortunately because of the very nature of the model (the variance is not constant producing errors that are not normally distributed), the method of least squares yields estimators that are not unbiased.

The general method of estimation that leads to the least square method under the linear regression model is called maximum likelihood. The method produces values for the unknown parameters which maximize the probability of obtaining the observed data set. To apply this method the likelihood function is constructed first . This function expresses the

probability of the observed data as a function of the unknown parameters.

Let P(Y=1/x) denote the probability of y=1 given a set of covariates x and the conditional probability that Y=0 given x by P(Y=0/x). So the likelihood for each pair (x_i, y_i) may be expressed by

$$\zeta_i = \theta_i^{y_i}[1-\theta_i]^{1-y_i} \qquad (13)$$

Where θ_i is a function of x. Assuming independence

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \zeta_i \qquad (14)$$

The method of maximum likelihood uses the log of the last expression namely the log likelihood, defined as

$$L(\beta) = \ln[\mathcal{L}(\beta)] = \sum_{i=1}^{n} (y_i \ln\theta_i + (1-y_i)\ln[1-\theta_i]) \qquad (15)$$

From here the values that maximizes L(ß) may be found by taking derivatives of the equations with respect to $\beta_1, \beta_2, \ldots, \beta_p$ setting them equal to zero and solving.

For a formal treatment of the maximum likelihood method (Cox 1970) we can start from equation (2) in which case the log likelihood is

$$L(\beta) = \sum_{s=1}^{p} \beta_s t_s - \sum_{i=1}^{n} \log(1+e^{x_i\beta}) \qquad (16)$$

where $t_s=\Sigma x_{is}Y_i$ therefore

$$\frac{\partial L(\beta)}{\partial \beta_s} = t_s - \sum_{i=1}^{n} \frac{x_{is}e^{x_i\beta}}{1+e^{x_i\beta}} \qquad (17)$$

and

$$I_{s_1 s_2}(\beta) = E\left(-\frac{\partial^2 L(\beta)}{\partial \beta_{s_1}\partial \beta_{s_2}}\right) = \sum_{i=1}^{n} \frac{x_{is_1}x_{is_2}e^{x_i\beta}}{(1+e^{x_i\beta})^2} \qquad (18)$$

equation (17) involves the Y's only through T's and does not depend on the Y's. So in particular equation (18) gives the expected and the observed value of the second derivates of L(ß).

The maximum likelihood estimate of ß satisfies the systems of equations

$$\left[\frac{\partial L(\beta)}{\partial \beta_s}\right]_{\beta=\hat{\beta}} = 0 \qquad (19)$$

and its asymptotic covariance matrix is the inverse matrix $\{I^{s_1 s_2}(\beta)\}$ to equation (19) and its consistently estimated by $\{I^{s_1 s_1}(\beta)\}$.

From here tests and confidence intervals for the parameters follow well developed theory of the same maximum likelihood estimation (Cox and Hinkley 1974).

The method of maximum likelihood can be applied to the cases and controls sampling scheme to obtain a logistic regression model in which the dependent variable is the outcome variable of interest to the investigator. The key steps in the development are the use of two conditional

probabilities involving the Bayes theorem. Since the likelihood was developed on subjects selected we need a variable that keeps status record of each subject in the population. Let the variable s denote the selection , s=1, or the nonselection, s=0, of a subject. The total likelihood for a sample of size n_1 cases (y=1) and n_0 controls (y=0) is given by

$$\prod_{i=1}^{n_1} P(\boldsymbol{x}_i | y_i=1, s_i=1) \prod_{i=1}^{n_0} P(\boldsymbol{x}_i | y_i=0, s_i=1) \qquad (20)$$

For an individual term in the likelihood function shown in equation (20) the application of the Bayes theorem produces

$$P(\boldsymbol{x}_i | y_i, s_i=1) = \frac{P(y_i | \boldsymbol{x}_i, s_i=1)\, P(\boldsymbol{x}_i | s_i=1)}{P(y_i | s_i=1)} \qquad (21)$$

Assuming that the selection of cases and controls is independent of the covariates with respective probabilities δ_1 and δ_0 then

$$\delta_1 = P(s_i=1 | y_i=1, \boldsymbol{x}_i) = P(s_i=1 | y_i=1) \qquad (22)$$

and

$$\delta_0 = P(s_i=1 | y_i=0, \boldsymbol{x}_i) = P(s_i=1 | y_i=0) \qquad (23)$$

The substitution of δ_1 and δ_0 in the logistic regression model, Θ_i for $P(y_i=1 / x_i,)$ into equation (21) produces

$$P(y_i=1 | \boldsymbol{x}_i, s_i=1) = \frac{\delta_1}{\delta_0[1-\Theta_i] + \delta_1 \Theta_i} \qquad (24)$$

If we divide the numerator and the denominator of the

expression on the right hand of equation (24) by $\delta_0[1-\theta_i]$ the result is a logistic regression model with intercept term $\ln(\delta_1/\delta_0)+\beta_0$.

Let θ^*_i the right hand of equation of equation (24) and P(x) the probability distribution of the covariates. The general term in equation (21) then becomes for $y_i=1$

$$P(\boldsymbol{x}_i|y_i=1, s_i=1) = \frac{\theta^*_i P(\boldsymbol{x}_i)}{P(y_i=1|s_i=1)} \qquad (25)$$

A similar term for y=0 can be obtained replacing θ^*_i by $[1-\theta^*_i]$ in the numerator and $P(y_i=1/s_i=1)$ by $P(y_i=0/s_i=1)$ in the denominator of equation (25). Let

$$L^*(\beta) = \prod_{i=1}^{n} (\theta^*_i)^{y_i}[1-(\theta^*_i)^{1-y_i}] \qquad (26)$$

Then the likelihood becomes

$$L^*(\beta)\prod_{i=1}^{n}\left[\frac{P(\boldsymbol{x}_i)}{P(y_i|s_i)}\right] \qquad (27)$$

The first term in equation (27), $L^*(\beta)$ is the likelihood obtained when we pretend the case control data was collected as a cohort study.

If we assume that the probability distribution of x, P(x) contains no information about the coefficients in the logistic regression model then maximization of the full likelihood with respect to the parameters in the logistic model θ^*_i is only subject to the restriction $P(y_i=1/s_i=1)=n_1/n$ and $P(y_i=0/s_i=1)=n_o/n$ (Hosmer 1989).

The likelihood equation obtained by differentiating with respect to the parameters β^*_0 assures that this condition is met. Thus the maximization procedure needs only to consider the portion of the likelihood that looks like a cohort study. The consequences of this are well known to biostatisticians. The analysis of data from case-control studies via logistic regression may proceed in the same way using the same computer program as cohort studies. Natural inferences about the intercept parameter β_0 are not possible without any prior knowledge of the sampling fractions δ_1 and δ_0.

2.2.5. Logistic Regression for Matched Case-Control Studies.

2.2.5.1 Properties of Conditional Distributions.

For the purpose of the following discussion we will need the distribution of the random variable $T_1,\ldots,T_p$. This follows from equation (2) summing all the binary sequence that generate the particular values $t_1,\ldots,t_p$;

$$P(t_i=t_i,\ldots,T_p=t_p)=\frac{c(t_1,\ldots,t_p)\exp\left(\sum_{s=1}^{p}\beta_s t_s\right)}{\prod_{i=1}^{n}(1+e^{x_i\beta})} \qquad (28)$$

where $c(t_1,\ldots,t_p)$ is the number of different binary sequences that yield the specified values $t_1,\ldots,t_p$.

To find a conditional distribution of T_p (Cox 1970) given $T_1=t_1\ldots,T_{p-1}=t_{p-1}$ where T_s are simple sufficient statistics we

have that $P(T_p=t_p/T_1=t_1,\ldots,T_{p-1}=t_{p-1})=$

$$\frac{P(T_1=t_1,\ldots,T_p=t_p)}{P(T_1=t_1,\ldots,T_{p-1}=t_{p-1})} \qquad (29)$$

The numerator is given by equation 2 and the denominator by summing over all possible t_p. When the ratio in equation (29) is formed, several factors cancel and the conditional probability is

$$\frac{c(t_1,\ldots,t_p)\,e^{\beta_p t_p}}{\sum_u c(t_1,\ldots,t_{p-1},u)\,e^{\beta_p u}} \qquad (30)$$

Equation (30) does not involve $\beta_1,\ldots,\beta_{p-1}$.

The distribution can be written

$$P_T(t;\beta)=\frac{c(t_{t-1},t)\,e^{\beta t}}{\sum_u c(t_{p-1},u)\,e^{\beta u}} \qquad (31)$$

An important case of equation (31) (which has already been discussed) corresponds to $\beta=0$

$$p_T(t;0)=\frac{c(t_{p-1},t)\,e^{\beta t}}{\sum_u c(t_{p-1},u)} \qquad (32)$$

Cox (1970) also describes the optimum properties, considering the problem of testing the null hypotheses $\beta=\beta_0$ against the alternative $\beta=\beta'$considering the conditional distribution of equation (32) applying Neyman-Pearson lemma to this equation form a critical area for those sample points having values of

the likelihood ratio

$$\frac{p_T(t;\beta')}{p_T(t;\beta_0)} \propto e^{(\beta'-\beta_0)t} \qquad (33)$$

The factor of proportionality being independent of t. Thus for all $\beta_0 < \beta'$, the critical region should consist of the upper tail of values of t and the resulting procedure is uniformly most powerful. Likewise upper and lower limits can be obtained.

The optimal properties of the conditional maximum likelihood, derived by letting the sample size become large, hold only when the number of parameters remains fixed. In any 1-M matched study this is not the case. With a fully stratified analysis, the number of parameters increases at the same rate as the sample size. For example for a model that contains only one dichotomous variable it can be shown that the bias of the estimate is large when analyzing a matched 1-1 design with the unconditional logistic regression. The method considers one nuisance parameter for each strata and separate the ß's to yield maximum likelihood in the logistic model. The following is a summary of the conditional likelihood applied to the matched design (Hosmer 1989) .

Suppose that there are K strata with n_{1k} cases and n_{0k} controls in stratum k,with k=1,2,...,K. The conditional likelihood for the kth stratum is obtained as the probability of the observed data conditional on the stratum total and the total number of cases observed, the sufficient statistics (as

we saw in the previous section) for the nuisance parameter. In this situation it is the probability of the observed outcome relative to the probability of the data for all the possible assignments of the cases and the controls.

There are several possible assignments of case status to n_{1k} among the n_k subjects of the stratum. Let j be anyone of these assignments. Let subject 1 to n_{1k} be assigned to the cases and $n_{1k}+1$ to n_k to the controls. This will be indexed by i for the observed and i_j for the possible assignments. The conditional likelihood may thus be expressed as

$$l_k(\beta) = \frac{\prod_{i=1}^{n_{1k}} P(\boldsymbol{x}_i|y=1) \prod_{i=n_{1k}+1}^{n_k} P(\boldsymbol{x}_i|y=0)}{\sum_j \left[\prod_{i_j=1}^{n_{1k}} P(\boldsymbol{x}_{ji_j}|y=1) \prod_{i_j=n_{1k}+1}^{n_k} P(\boldsymbol{x}_{ji_j}|y=0)\right]} \qquad (34)$$

where the summation over the j in the denominator is over the n_k choose n_{1k} combinations. The full conditional likelihood is the product of the $l_k(\beta)$ over the k strata;

$$l(\beta) = \prod_{k=1}^{K} l_k(\beta) \qquad (35)$$

Let the logit in the kth stratum be $g_k(x)=\alpha_k+\beta'x$ where α_k denotes the contribution to the logit of all the terms constant within the stratum as we saw in the mathematical development, the stratification, the application of the

conditional probability to each stratum yields:

$$l_k(\beta) = \frac{\prod_{i=1}^{n_{1k}} e^{\beta' x_i}}{\sum_j \prod_{i_j=1}^{n_{1k}} e^{\beta' x_{j i_j}}} \qquad (36)$$

It is important to notice that the terms of the form

$$\frac{e^{\alpha_k}}{(1+e^{\alpha_k+\beta' x})} \qquad (37)$$

appear equally on both sides of the numerator and denominator of equation (37) and cancel out, leaving the equation depending only on the ß's.

CHAPTER III

DATA SET

3.1 The Falconbridge Study

As noted earlier, the data use for this project was obtained from a large cohort study, that was designed to examine the mortality of the workers at Falconbridge Nickel Mines, LTD.

The population of study was defined as all men who had worked at the mine from January 1, 1950 and with at least 6 months service by December 31, 1976. A second phase amplifies the study to include follow-up until 1984. It was estimated that the six months minimum would exclude subjects with a very short period of work, with low exposure index and difficult to trace.

Two data files were generated and a brief description is pertinent.

File one which was called personal information file contained some demographic data like full name, birth place, death place, first work date, last work date and other items; These records were obtained from company records. This file has 11,567 lines, one for each subject and this file was used to generate the cases and the corresponding matched controls. The most important record for this project was the vital status of each subject.

The strategy to ascertain this status was using the CRL technic, essentially a matching between a list of the Falconbridge employees and the National Mortality Data Base which contains a list of all the people who died in Canada and the United States while residents of Canada. A complementary follow up was made via personal contact, newsletters and other methods. Confidentiality was protected by statistics Canada by removing personal identifiers from the file that had appended to it causes and dates of death. The causes were coded according to the ICD (International Classification of Diseases).

Birth date and date death are important variables that ought to be mentioned. The first one was complete for all the subjects that comprise the cohort and it was used for structuring the case and control file as a matching factor given the fact that subjects born in different years are more likely to behave unlike in relevant behaviours than subjects born in the same year.

The birth date has the format mm-dd-yy. The death date had a special importance when calculating the exposure value of the different contaminants because it is essential that for controls the exposures beyond the dates of death of the corresponding cases are not included.

In some instances the last working date corresponds with the death date and it is clear that these subjects died while actively working. The first working date acted as a second

matching variable and it was incorporated into the computer program so that any subject selected as a possible control also had the same working date as the corresponding case.

All the cases had an ICD coding of 162, if the underlying cause of death was lung cancer. The program used for this project detected this number and create a list of all the cases. Once the cases were identified the selection of the controls could start. Any subject that matches the case with respect to birthdate and first work date was considered a possible control. Because of the fact that using two matching criteria reduces the number of possible controls, up to one year of difference in the matching date was allowed to fulfil the necessary quota of controls for the specific case.

Because this was what actually happened it was subsequently increased but in any situation was never greater or smaller than 5 years. It was presumed that this period was not wide enough to affect the precision of the analysis while allowing the formation of several matched controls, when generated inside the same cohort.

One thing important to notice is that any subject may be selected that is free of the disease at the time when the case acquired the illness, in other words a future case could served as a control for another case.

The second file is an account of the different job positions that each subject held while working for the company.It includes a serial number that is unique and is

equal for the same person on both files.

From here a serial of individual dates and events are linked to this number, representing the different job positions held for members of the cohort. This history file has more than 32,000 lines, showing internal job movement (promotions , for example).

Each department has a code number which identifies it (appendix 1). On the other hand Shannon et al have calculated a factor for 8 different contaminants that might be present in the different working settings (Shannon et al 1992).

From here it was possible to establish a length-intensity factor for all the workers.

These "weights" were then used to evaluate the individual effect of each contaminant on the final outcome, these weights are also included in appendix 1. The program was then designed to generate the exposure file for each individual. This program used the information previously described. Once the cases and controls were assembled using the unique identifier, the program goes to work with the work history file and subtracts the second work date from the first work date; then it identifies the first work place that corresponds to such period. Since the program identified the matrix of departments it selected the corresponding workplace and matched it with the corresponding sets of weights for each contaminant. The algorithm repeats until the final workdate was read. The program then adds the length-intensity effect for each

contaminant and stores the information into a new file. The program then goes to look for a new subject. This continues until the last subject was evaluated. At the end, each subject had one column with the effect of the first contaminant, the second column that corresponds to the second contaminant and so on. Given the fact that the sequence of numbers that the program follows had the structure of the cases and corresponding controls, the final file has this same sequence. At the end the program was run for each set of cases and controls and so 6 different files were output and analyzed. Only the file for the 1:10 (1 case:10 controls) was included in the project.

In the initial and final date of exposure files a similar procedure was followed. First the subject was identified. Using the workplace matrix each position was associated with the corresponding contaminants which in turn had an initial (and final) date of exposure. The program generated two files with the initial and final date of exposure of each single contaminant. One characteristic found was that the dates of exposure were given as the number of days starting with a reference value. And again only the data files for the 1:10 were included in this project (appendix 2).

CHAPTER IV

4. RESULTS

4.1. Exposure. Effect of different contaminants on the cases and controls.

Once the cases and controls were selected according to the matching criteria 3 exposure files were created for each set (1:1, 1:2, 1:4, 1:6, 1:8, 1:10), for simplicity only the files for the 1:10 set were included in appendix I with a table that describes the variables used in the study. This file contains the exposure that each subject incurred during his work period in the company. The first 4 columns were identification columns, the next 9 were the cumulative exposures to the contaminants, combining intensity of exposure with duration.

Each control has been matched by age and initial working date. The second exposure file provides the initial date on which each subject was exposed to the different contaminants, so again the first four columns are indication columns and the last 9 are initial exposure dates to the several contaminants.

The idea was to asses if there was a stronger effect of any contaminant if the subject's exposure was recent or occurred some time before.

The third exposure file is similar to the second except that the final exposure date was registered for each subject

to each contaminant. Table #2 A produced a description of 3 demographic variables of the case-control sets. None of these variables seem to be very different for each set studied, in fact the agreement between both groups is very good.

4.2 Results of the contaminant data using a GLIM program.

As has been said earlier, to analyze the data a program developed by Adena and Wilson (1982) was used. It was adapted for use with the files described.

The highlights of the program are that it can handle the case-control sets, so the observations will correspond to a 1 for the observed case and M zeros for the number of controls which can be considered as counts. The controls for each set were generated using a random procedure but as described in chapter II for the last two sets (1:8 and 1:10) the intervals for the matching variables were increased up to ± 5 years.

The errors will distribute as a Poisson variate and the link function is hence a logarithmic function.

The first goal using this general linear model is to evaluate the effect of each of the exposure variables on the outcome. Two approaches were used to evaluate the effect of each contaminant. First, each variable was assessed individually giving information of the effect of the different number of controls on the estimated parameters. The results are presented on tables #3 to #10. Each table represents the contribution of a single contaminant for each different set of cases and controls and as already noted, a measure of the

precision of the estimated parameters will be available.

In formal terms we have the following general model:

$$\lambda_i = logit(\theta_i) = \beta_0 + \beta_1 x_i$$

This implies that for each variable evaluated a different slope will be generated and from there a different effect in the final outcome. If the exposure to the different contaminants was not related to the respiratory cancer we would expect the same exposure index for the case and its corresponding controls. In other words the respiratory cancer will not be related to the exposure value.

From the tables #3 to #10 it is worth noticing that all the contaminants produce a small change in deviance regardless of the set of case-controls.

For the first variable (Ni3S2) in the table #3 for the sets 1:1, 1:2, 1:4 and 1:6 the response was not estimable because a lack of observations. The GLIM system defines such variables as "aliased".

This is explained by the fact that only one subject in the whole cohort was exposed to the contaminant and this subject was selected in the 1:8 and 1:10 sets. The reason for this is that it was not selected for the first sets for being relatively "apart" in terms of the matching with respect to other controls. The intervals produced are of almost the same width as they have similar change in deviance thus indicating that the difference in sets makes no impact on the width of the interval.

Table #4 shows the effect of the different sets upon the estimates of the chemical (Ni3Fe)S using the logistic regression and assessing only the impact of this chemical on the model.

Even though there is not a big change in the deviance (from 0.23 to 2.15) and they are not significant, the maximum is reached for the 1:4 set of cases and controls. Table #4 shows the different length of the intervals. These intervals tend to lack a clear patterns of response going from a minimum value for the 1:4 set to the maximum that corresponds to 1:1, however the difference in this two extremes is very small.

For NiSO4 the change in deviance is a maximum between the sets 1:2 and 1:4.

The intervals seem to be smaller for 1:4 and 1:6 but the width is relatively small. A similar behaviour is shown by Ni in pentlandite. For this species the response is shown in table #6.

The contaminant Ni2FeO produces higher change in deviance. Strangely in tables #7 the set 1:1 is the one that produces the interval with less variability and although there is not an immediate answer to this fact it seems related to the contaminant itself because of the fact that not many controls were exposed to it.

Table #8 present the results of Nickel in pyrrhotite, the greatest change in deviance is produced by the 1:8 set and exactly for this set, the smallest confidence interval is

produce.

Total nickel in tables #9 show a good change in the response. The change in deviance goes up to 14.52 for the 1:4 set and the smallest interval is registered in here. However the next closest interval namely the one corresponding to the 1:1 set has the smallest width of all the intervals on this group.

Total dust exhibited the strongest change in deviance up to values of 41.08 and 12.24 the smallest change.

The confidence intervals are so small that there seem to be no noticeable difference between them.

In summary although there does not seem to be a difference in the set that produces the best estimate i.e. that one with the smallest variance, there seems to be a tendency for the 1:4 set to produce the highest change in deviance (thus a factor which explains better the phenomena in terms of model building).

4.3. Variable selection and model building.

The second approach for evaluating the global effect of the variables involves a multiple variable procedure.
A subjective way to tackle the problem was initially used selecting those variables that seemed to produce a significant amount of difference in the deviance.

The initial variables were:Ni2FeO, Ni in pyrrhotite, total Ni and total Dust, variables that showed the maximum change of

deviance across the different sets.

Several models were tested and the change in deviance was used to evaluate each one of them.

Table #11 shows the result of such selection. The deviance is a measure that distributes like a chi-square so it may be used with the corresponding degrees of freedom to assign a p value.

All variables are significant when added on the first step ($p<0.001$).

It is clear from this table that total dust and nickel in pyrrhotite should be included, however total nickel produces the least significant change.

Observing the models that have two main effects those that include total nickel had always the least change in deviance, indicating that this variable is not as important as the others.

This is confirmed by the model that include contaminants and that left out total nickel, has also the highest change in deviance. So the model that includes Ni in pyrrhotite, total dust and Ni2FeO seems to give the best fit even though is not the model with the smallest deviance. The final model with its corresponing parametrs is presented in the following table;

TERM	ESTIMATE (xE+06)	STANDARD ERROR (xE+06)
INTERCEPT	-0.8732	0.1672
NIKEL IN PYRRHOTITE	-0.007244	0.007413
TOTAL DUST	-0.0000298	0.0000437
$(Ni_2Fe)O$	-0.003570	0.002503

The decision to keep this model is based on the fact that for the previous more simple models that included total nickel, always produced the smallest change in deviance. However it is noticeable that any of the models on table #11 with three variables in it could have been selected.

4.4 Effect of the initial date of exposure of each contaminant

Table #12 shows the effect of the contaminants depending on the initial date of exposure for the 1:4 set. The model that was used in this section was different than the previous ones because of the fact that dates instead of contaminant index, were used for the evaluation.

$$\theta_i = \beta_0 + \beta_1 x_{ij}$$

where x_{ij} represents the different initial date of exposure for each single contaminant. The conditional logistic regression model was used in all the variables. The designation of each contaminant ($N_{i2}FeO$, Ni in Pyrrhotite, and so on) implies the final date effect of that contaminant in the case-control. The set 1:4 was again selected for the same reasons previously described. No variable shows a statistically significant result except for Ni2FeO which shows a marginal significance ($0.1<p<0.05$) testing the change of deviance like a chi-square.

Due to the complexity of the interpretation and to the lack of deviance change no model was sought for the main effects of the initial exposure.

4.5. Effect of the final date of exposure of each contaminant

Table #13 shows the effect of the contaminant depending on the final date of exposure again for the 1:4 set. As in the initial evaluation date the conditional logistic model is of the type

$$\theta_i = \beta_0 + \beta_1 x_{ij}$$

Where x_{ij} represent the diferent contaminat fine exposure date for the cases and controls.

Starting with Ni2FeO but especially total nickel and total dust, showed very strong effects ($p<0.001$) when this variables are analyzed one at a time. So there seems to be an effect when subjects are exposed late to the total dust and total

nickel.

The possible implications of this will be referred in the discussion chapter.

DISCUSSION

Cornfield (1951) demonstrated that the relative risk may be estimated from retrospective studies (case-control) studies using the odds ratio. When the logistic regression model contains independent variables continous variablesthe interpretation will depend on the specific units of that variable.

If we want to compare the exposed population with a reference population this implies that the cases should consist of all the subjects with the outcome of interest or at least a representative part of this group. In the same way the controls should form a random group from the total exposed population that may developed the disease of interest.

Thus in the Falconbridge mine study those subjects with six or less than six months of work were excluded because their relative short period of exposure producing consequently a better estimation of a dose-response pattern. In this study the selection of the controls was limited to those individuals with a minimum exposure to nickel and nickel derivates even though some potential controls might develop the disease, but such subjects did not suffer a long enough exposure.

The design study of this project comes close to what

Breslow (1982) considered the ideal case-control architecture that in which all cases and controls are generated via a "population-based" selection.

Each of the cases was clearly identified as having a the disease, while controls were matched by initial working date and age. However severe limitations may occur because of the very nature of the design. This will of course have an effect on the final estimators. Mortality for the full cohort might not be complete, some of the cases may die without having been reported, however there is unlikely to be a complete follow-up. Some factors that influence this lack of completeness are among others: cost involved per subject, subjects not traceable, change of residence outside Canada or the United States.

Because of this some demographic variables relating the cases with the controls were calculated.It might be argued that if any of these potential variables produce a strong effect for each of the corresponding sets of cases and controls then it might be claimed that a factor is present in a differential manner . Table 2A showed that this was not the case, given the fact that most of the sets showed consistently the same ages (initial, final and birthdate).

It is accepted that a sample of the study base, the cohort,will draw valid conclusions if it has been

properly generated and that there is usually little loss of precision (Checkoway 1989). Cohort based studies also called nested case-control studies like the present project, offer the possibility of reaching conclusions for the entire cohort and in doing so reducing costs.

Another advantage of this kind of studies is the possibility of matching those subjects and their possible controls with one or more confounding factors thus increasing the precision of the estimators generated (Liddel 1977).

The utilisation of more than one control per case has been widely advocated (Miettinen, 1969) and the same conclusion is reached in the present project, however even thou it is not clear what number of controls is the optimum the evidence seem to point around 4. The confidence intervals for the estimates were not strongly different between them hinting that none of the sets is of more practical importance than the others and this could be the reason why the confidence intervals did not decrease with the increase number of controls.

None of the chemical contaminants were selected using the step-wise regression methodology. A different approach was used evaluating each contaminant at a time. Of the models tested Ni in pyrrhotite, total dust and Ni2FeO seem to give the best fit, in terms of the

calculated deviance. However this way of selecting the contaminants did not exclude the possibility that other species may have a possible effect in the induction and possible development of the neoplasm.

The inclusion of the first and final date of exposure is related with the known lag between the exposure date and the clinical manifestations of the disease. The implications of Ni_2FeO having a marginal effect on the initial date of exposure are unknown.

Total nickel and total dust are statistically related with the final exposure date even though the relationship might not be clear. It may be argued that total nickel is really not aporting new information to the outcome since is a linear combination of all the chemical species that include nickel in its formula, hence should reflect the individual contributions. Similar argument can be used for the case of dust since includes all the suspended particles in the air. However way it is detected in the final date of exposure is uncertain.

It could be argued that at some point in time the effect of any particular species has a more profound effect depending on the final date that the subject has, however to ascertain at what time this is happening will require development of intervals of the same length (0-5, 5-10, 10-15,etc) and exploration of the possibility of a trend in time for each variable.

Linearity was intrinsicaly assumed during the whole modeling process and even thou never statistically tested no evidence was found of the contrary. It is a common practice to assume it and especially in the situation of ascertaining if a variable should be in the model. It was the intention of the project only to find a dose-response effect in which case a linear, quadratic, cubic or any other truly monotonic relation will produce a significant effect. Plotting the data is one way to asscertain this even partitioning the data in smaller subgroups. This simple as may sound might require a transformation of the response scale with the dificult consequense of its interpretation.

The usual approach to model procedures indicates the need to explore the possibilty of interactions between the main effects selected. It is likely that more than two main effects will produce a significant interaction. For example total nickel and total dust may interact between them and produce a different respone in the outcome producing a different biological meaning, that will depend on the time that the measuremente was performed. However the intention of this project was to asses only the contaminants on an individual basis.

Ussing the deviance as the sole criteria for the selection of the models offers some limitations. However the analisis of residuals is most of the time the diagnostic tool for the purpose of rejecting a model. The residuals in

particular Pearsons residuals are a way to discriminate the validity of the assumptions of the model. Several summary statistcs besides the deviance have been proposed (Hosmer and Lemeshow 1989). Among them the most important is the leaverage. This quantity is defined as the j_{th} element of the hat matrix. Their importance resides in that they are calculated according to their distance from the mean. The farther and frequently the points are the more suspicious we are of the model. The hat matrix for the logistic regression is defined as (Hosmer and Lemmeshow 1989) :

$$H = V^{1/2}X(X'V\,X)^{-1}X'V^{1/2}$$

Where X is the design matrix and V is a jxj diagonal matrix with general element $v_j=m_j\Theta_j[1-\Theta_j]$, where j is the number of distinct values of x observed. If some subjects have the same value of x then j<n if m_j denotes the number of subjects with $x=x_j$. Let y_j denotes the number of positive responses y=1, among the m_j subjects with $x=x_j$ then $\Sigma y_j=n_1$. No further assessment analysis was sought due to the objective of only fit main effects.

REFERENCES

Adena M. A. and Wilson S.R. (1982). Generalized Linear Models in Epidemiological Research. Case-Control Studies. The Instat Foundation for Statistical Data Analysis. Sydney.

Anderson S.,Auquier A., Walter W.H., Oakes D., Vandaele W., Herbert I. W. (1980). Statistical Methods for Comparative Studies. John Wiley and Sons. New York.

Armitage P. (1971). Statistical Methods in Medical Research. Blackwell Scientific Publications. Oxford.

Bailar J.C. and Anthony G. B. (1977). Most Cited Papers of the Journal of the National Cancer Institute 1962-75. J. Natl.Can.Inst. 59:709-714.

Breslow N. (1982). Design and Analysis of Case-Control Studies. Ann.Rev.Pub.Heal. 3:29-54.

Breslow N.E. (1976). Regression Analysis of the log odds ratio. A Method for Retrospective Studies. Biometrics 32:409-416.

Breslow N.E. and Day N.E. (1980) Statistical Methods in Cancer Research. The Design and Analysis of Case-Control Studies. International Agency for Research on Cancer.

Breslow N.E. and Day N.E. (1987). Statistical Methods in Cancer Research. The Design and Analysis of Cohort Studies. International Agency for Research on Cancer.

Checkoway H., Pearce N. and Crawford-Brown D.J. (1989). Research Methods in Occupational Epidemiology. Oxford University Press. New York.

Cornfield J. (1951). A Method of Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast and Cervix. J.Nat.Can.Ins. 11:1269-1275.

Cox D.R. (1970). Analysis of Binary Data. Chapman and Hall. New York.

Cox D.R. (1972). Regression Models and Life Tables (with discussion). J.R. Statis. Soc.(B). 74:187-220.

Cox D.R. and Hinkley D.V. (1974). Theoretical Statistics. Chapman and Hall. London.

Cox D.R. and Oakes D. (1984). Analysis of Survival Data. Chapman and Hall. New York.

Doll R., Morgan L.G., Speizer F.E. (1970). Cancers of the Lung and Nasal Sinuses in Nickel Workers. Br.J. Ind. Med. 38:327-333.

Fleiss J., L., (1981). Statistical Methods for Rates and Proportions. John Wiley and Sons. New York.

Gail M., Williams R., Byar D.P. and Brown C. (1976). How many controls? J.Chron.Dis. 29:723-731.

Hosmer D.W. and Lemeshow S. (1989). Applied Logistic Regression. John Wiley and Sons. New York.

Jensen O.M. (1979). Cancer Morbidity and Causes of Death Among Danish Brewery Workers. Int. J. Can. 23:454-463.

Kupper L.L., McMichael A.J. and Spirtas A. (1975). A Hybrid Epidemiological Study Design Useful in Estimating Relative Risk. J.Amer.Stat.Assoc. 70:(351)524-528.

Lidell F.D.K., McDonald J.C. and Thomas D.C. (1977). Methods of Cohort Analysis: Appraisal by Application to Asbestos Mining. J.R. Stat.Soc.Ser.A, 140:469-491.

Lubin J.H. and Gail M.H. (1984). Biased Selection of Controls for Case-Control Analysis of Cohort Studies. Biometrics 40:63-75.

McCullagh P. and Nelder J.A. (1984). Generalized Linear Models. Chapman and Hall Ed. London.

Magnus K., Andersen A., Hogetueit A.C. (1982). Cancer of Respiratory Organs Among Wrokers at a Nickel Refinery in Norway. Int.J.Can. 30:681-685.

Mantel N. and Haenzel W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. J.Nat. Can. Inst.22(4):719-748.

Prentice R.L. and Breslow N.E. (1978). Retrospective Studies and Failure Time Models. Biometrics 34:153-158.

Prentice R.L. (1986). A Case Cohort Design for Epidemiological Cohort Studies and Disease Prevention Trials. Biometrics 42:1-12.

Robins J., Gail M.H. and Lubin J.H. (1986). More on Biased Selection of Controls. Biometrics 42:293-299.

Report of the International Committee on Nickel Carcinogenesis in Man. Scandinavian Journal of Work, Environment and Health (1990);

(special issue).

Roberts R.S., Julian J.A. et al. (1989). A Study of Mortality in Workers Engaged in the Mining, Smelting and Refining of Nickel-I: Methodology and Mortality by Major Cause Groups. Tox. and Indu. Heal. 5(6):957-974.

Roberts R.S., Julian J.A. et al. (1989). A Study of Mortality in Workers Engaged in the Mining, Smelting and Refining of Nickel-II: Mortality from Cancer of the Respiratory Tract and Kidney. Tox. and Indu. Heal. 5(6):975-993.

Sackett D.L. (1979). Bias in Analytical Research. J.Chron.Dis 32:52-63.

Schelesselman J.J. (1982). Case-Control Studies: Design, Conduct and Analysis. Oxford University Press.

Shannon H.S., Julian J.A. and Roberts R.S. (1984). A Mortality Study of 11,500 Nickel Workers. J.Nat.Can.Inst. 73:1251-1258.

Shannon H.S., Walsh C., Jadon N., Julian J.A., Weglo J.K., Thornhill P.G. and Cecutti A.G.(1991). Mortality of 11,500 Nickel Workers. Extended Follow-up and relationship to Environmental Conditions. Toxicol. Ind. Health. 7(4):277-294.

Shapiro S. et al. (1979). Oral Contraceptive Use in Relation to Myocardial Infarction. Lancet 1:743-747.

Steckel R.J. (1976). Re: "Estrogens and Endometrial Carcinoma" (letter to the Editor). New Engl. J.Med. 294:848.

Urey H.K. (1975). Efficiency of Case-Control Studies with Multiple Controls per Case: Continuous or Dichotomous Data. Biometrics 31:643-649.

Vessey M.P., McPherson L., Villard-Mackintosh, Yeates D. (1989). Oral Contraceptives and Breast Cancer: Latest Findings in a Large Cohort Study. Br.J. Can. 59:613-617.

Wolf P.H, Andjelkovich D., Smith Allan and Tyroler H. A Case-Control Study of Leukemia in the U.S. Rubber Industry. J.Occup.Med. 23(2):103-108.

APPENDIX #1

This appendix includes the contamination levels for the different departments or working areas. These indexes were used to calculate the total exposure to each nickel contaminant (Shannon et al 1991).

The second table gives a number to each of the working areas of the mine.

APPENDIX

CONTAMINATION LEVELS OF THE DIFFERENT NICKEL *

COMPOUNDS BY DEPARTMENT (mg Ni/m3)

Group/ Department	Ni3S2	(Ni3Fe)S	NiSO4	Ni in Pentlan-dite	(Ni2Fe)O	Ni in Pyrrrho-tite	Total Ni	Total Dust
Mining:	0	0	0	0.01	0	0.01	0.02	2
Ore dressing	0	0	0	0.03	0	0.01	0.04	3.8
Concentrator	0	0	0	0.02	0	0.01	0.03	3.2
Feed Preparation:								
Concentrate recieving	0	0	0.01	0.15	0	0.01	0.17	4.9
Pellet plant	0	0	0.01	0.13	0	0.01	0.15	4.3
Briquetting	0	0	0.01	0.13	0	0.01	0.15	4.3
Slurry-Filtering-Drying	0	0	0	0.03	0	0	0.03	0.9
Smelter:								
Sinter plant	0	0.07	0	0.12	0.02	0.01	0.22	6.3
Blast furnace	0	0.04	0	0.07	0.01	0.07	0.13	3.7
Settlers	0	0.03	0	0.05	0.01	0.05	0.09	3
Converters	0	0.03	0	0.01	0.01	0.01	0.05	1
Matte room	0.09	0	0	0	0	0	0.1	1.2
Pyrrhotite plant:	0	0	0.01	0	0	0	0.01	0.9
Nickel-Iron Refinery:	0	0	0	0	0.01	0	0.01	0.5
Maintenance:								
Repair crew	0	0.08	0	0.08	0.01	0.08	0.18	5
Welders	0	0	0	0	0.05	0	0.05	3.7
Shops	0	0	0	0	0.03	0	0.03	2.2
Miscellaneous	0	0.04	0	0.05	0	0.05	0.09	2.5
Surface:								
Unexposed	0	0	0	0	0	0	0	0

* Modify from Shannon et al (1992)

APPENDIX

RELATION OF THE DIFFERENT JOB POSSITION BY DEPARTMENT

Group/ Department	Codes
Mining	014 017 019 020 022 023 024 025 027 029 031 032 035 037 040 102 TO 108 185 187 190 194 195 203 204 205 302 303 305 306 308 403 404 405 407 408 431 472 473 477 508 514 526 527 535 536 540 550 602 603 604 608 610 702 703 704 708 790 802 803 804 890 901 902 903 904 919 990 149(13)
Milling	
Ore Dressing Concentrator	123 149(17) 149(22) 623 823 723 739 122 124 126 128 129 131 191 525 526 622 624 625 631 722 724 726 728 731 824 826 830 831
Feed Preparation	
Concentrate recieving	143 143(00) 149(16) 149(20) 160
Pellet plant	143(03) 143(04)
Briquetting	143(10)
Slurry-Filtering-Drying	143(11)
Smelter	
Sinter plant	143(02) 143(05) 143(06) 143(08)
Blast furnace	143(01) 143(12) 143(24) 144(01)
Settlers	-
Converters	143(13) 143(14) 144(03) 144(12) 144(13) 144(14) 144(21) 144(23)
Matte room	145
Pyrrhotite plant	149(15) 163 166
Nickel-Iron Refinery	242 243 245 247 258 249 259 246
Maintenance	
Repair crew	146 147 148 192
Welders	174 490 572 773
Shops	673 873
Miscellaneous	010 011 026 111 120 149(23) 170 171 172 173 175 177 178 179 180 181 182 184 187 470 471 479 503
Surface	
Unexposed	001 TO 007 012 015 016 030 033 041 045 TO 049 050 055 TO 100 109 110 119 139 142(25) 143(23) 149(18) 149(21) 150 151 155 159 162 176 183 186 188 193 196 197 198 199 219 310 402 409 TO 428 442 455 480 481 482 483 TO 489 500 542 555 556 560 565 575 576 580 583 TO 598 619 639 680 719 780 819 839 880 892 999
Missing	-99

APPENDIX # 2

For this study I generated Three files using the Falconbridge original information. As an example, only the set for the relation 1:10 is included in this project. The file #1, contaminant type effect includes the total exposure to each nickel contaminant. The first two columns are identifiers of each subject and the remaining columns the time-effect of each of the eight contaminats in the following order Ni3S2, (Ni3Fe)S, NiSO4, Ni in Pentlandite, (Ni2Fe)O, Ni in Pyrrhotite, Total Nickel and Total Dust in sequential order for each column.

Files #2 and #3 include the initial and final exposure date respectively. Columns one to four are identifiers. Column one is a sequential number, column two the set of case-control, column three 1=case and any other number is a control, the rest eight columns correspond to the initial and final exposure dates of each contaminants the same order as in file #1.

File No 1

CONTAMINANT TYPE EFFECT*

c-c	ID	Ni_3S_2	$(Ni_3Fe)S$	$NiSO_4$	Ni in pentlandite	$(Ni_2Fe)O$	Ni in pyrrhotite	Total Ni	Total Dust
1	46	0.00	238.29	8.21	428.95	66.72	12.58	788.31	21561.00
2	299	0.00	0.00	0.00	99.37	0.00	99.37	198.74	19874.00
3	704	0.00	0.00	0.00	0.00	259.00	0.00	259.00	18100.40
4	956	0.00	21.75	3.45	81.10	490.35	3.45	600.10	39407.90
5	1214	0.00	0.00	0.00	79.41	8.19	79.41	167.01	16291.50
6	1799	0.00	0.00	0.00	0.00	447.45	0.00	447.45	32753.70
7	1884	0.00	0.00	0.81	82.96	36.42	71.62	191.81	16379.90
8	5355	0.00	0.00	0.00	95.13	0.00	95.13	190.26	19026.00
9	3685	0.00	0.00	0.00	125.78	61.75	83.86	271.39	26201.50
10	8305	0.00	0.00	2.93	143.60	0.00	102.58	249.11	21365.70
11	8910	0.00	0.00	0.00	0.00	259.41	0.00	259.41	19023.40
1	250	0.00	93.05	13.26	335.98	119.50	19.65	581.44	18879.10
2	644	0.00	139.86	27.90	584.46	129.53	27.90	930.71	28366.70
3	778	0.00	0.00	0.00	24.08	126.09	11.78	161.95	10316.00
4	1281	0.00	0.00	0.00	99.60	0.00	99.60	199.20	19920.00
5	9750	0.00	0.00	54.78	734.04	191.28	65.73	1045.83	41086.61
6	9251	0.00	0.00	53.18	829.50	69.33	122.26	1074.27	50057.20
7	679	0.00	0.00	0.00	148.40	1.16	146.06	295.62	29480.60
8	3617	0.00	0.00	0.00	0.00	418.83	0.00	418.83	30752.40
9	8880	0.00	0.00	0.00	150.51	0.00	150.51	301.02	30102.00
10	5101	0.00	0.00	0.00	95.24	138.39	95.24	328.87	28601.50
11	7450	0.00	0.00	11.55	158.40	413.34	14.30	597.59	36323.10
1	262	0.00	0.00	0.00	104.64	5.25	104.64	214.53	21313.00
2	1311	0.00	0.00	0.00	101.61	0.00	101.61	203.22	20322.00
3	1405	0.00	0.00	0.00	107.39	0.00	107.39	214.78	21478.00
4	2095	0.00	295.72	13.62	657.13	97.52	13.62	1079.85	34196.10
5	2199	0.00	367.93	22.00	891.46	99.15	42.71	1426.74	43066.90
6	4127	0.00	0.00	0.00	107.04	0.00	107.04	214.08	21408.00
7	5091	0.00	0.00	0.00	111.19	0.00	111.19	222.38	22238.00
8	5188	0.00	0.00	0.00	74.47	11.86	74.47	163.35	15525.40
9	5434	0.00	0.00	0.00	103.24	0.00	103.24	206.48	20648.00
10	5480	0.00	0.00	0.00	0.00	541.00	0.00	541.00	40034.00
11	5882	0.00	0.00	0.00	54.24	24.69	54.24	133.17	12658.60
1	273	0.00	0.00	0.00	111.64	0.00	111.64	223.28	22328.00
2	391	0.00	0.00	0.00	108.11	0.00	108.11	216.22	21622.00
3	1107	0.00	61.71	40.08	716.58	47.35	66.91	934.07	34657.40
4	1704	0.00	330.40	1.26	437.60	4.90	2.96	779.12	22062.00
5	1831	0.00	311.72	4.27	586.53	103.46	8.96	1015.19	34299.50
6	9790	0.00	0.00	0.00	35.61	152.47	35.61	223.69	17663.80
7	195	0.00	404.01	2.16	478.94	20.87	3.78	909.76	24871.70
8	4006	0.00	0.00	0.00	118.31	0.00	118.31	236.62	23662.00
9	8908	0.00	0.00	0.00	0.00	0.00	0.00	12.63	0.00
10	5635	0.00	0.00	0.00	106.59	0.00	106.59	213.18	21318.00
11	1100	0.00	0.00	6.05	175.79	24.38	91.09	297.31	21191.50
1	314	0.00	0.00	0.00	99.36	188.35	99.36	387.07	33809.90
2	383	0.00	0.00	34.40	648.57	0.00	166.97	849.94	43370.00
3	644	0.00	139.86	27.90	584.46	129.53	27.90	930.71	28366.70
4	778	0.00	0.00	0.00	24.08	126.09	11.78	161.95	10316.00
5	1281	0.00	0.00	0.00	99.60	0.00	99.60	199.20	19920.00
6	8960	0.00	0.00	0.00	148.40	0.00	148.40	296.80	29680.00
7	213	0.00	0.00	0.00	349.52	10.39	141.82	501.73	49854.70
8	593	0.00	0.00	0.00	150.55	0.00	150.55	301.10	30110.00
9	5394	0.00	0.00	1.20	33.47	66.94	16.67	118.28	7029.00
10	7160	0.00	0.00	0.00	373.05	2.65	149.80	525.50	52539.10
11	8880	0.00	0.00	0.00	150.51	0.00	150.51	301.02	30102.00

FILE N° 2
INITIAL EXPOSURE DATE*

seq	set	c-c	ID	Ni_3S_2	$(Ni_3Fe)S$	$NiSO_4$	Ni in pentlandite	$(Ni_2Fe)O$	Ni in pyrrhotite	Total Ni	Total Dust
1	1	1	46	-99	27315	26882	26882	27315	26882	26882	26882
2	1	2	299	-99	-99	-99	25886	-99	25886	25886	25886
3	1	3	704	-99	-99	-99	-99	27678	-99	27678	27678
4	1	4	956	-99	27841	27496	27496	27841	27496	27496	27496
5	1	5	1214	-99	-99	-99	27671	33662	27671	27671	27671
6	1	6	1799	-99	-99	-99	-99	27265	-99	27265	27265
7	1	7	1884	-99	-99	27272	26896	34586	26896	26896	26896
8	1	8	5355	-99	-99	-99	26918	-99	26918	26918	26918
9	1	9	3685	-99	-99	-99	26843	26810	26843	26810	26810
10	1	10	8305	-99	-99	26173	26173	-99	26173	26173	26173
11	1	11	8910	-99	-99	-99	-99	27693	-99	27693	27693
12	2	1	250	-99	22552	21226	21226	22552	21226	21226	21226
13	2	2	644	-99	23832	21046	21046	23832	21046	21046	21046
14	2	3	778	-99	-99	-99	21230	23131	21230	21230	21230
15	2	4	1281	-99	-99	-99	21472	-99	21472	21472	21472
16	2	5	9750	-99	-99	27393	21822	24912	21822	21822	21822
17	2	6	9251	-99	-99	21376	21376	33602	21376	21376	21376
18	2	7	679	-99	-99	-99	21315	21016	21315	21016	21016
19	2	8	3617	-99	-99	-99	-99	21304	-99	21304	21304
20	2	9	8880	-99	-99	-99	21197	-99	21197	21197	21197
21	2	10	5101	-99	-99	-99	21190	27761	21190	21190	21190
22	2	11	7450	-99	-99	25232	25232	21223	25232	21223	21223
23	3	1	262	-99	-99	-99	27768	37113	27768	27768	27768
24	3	2	1311	-99	-99	-99	28029	-99	28029	28029	28029
25	3	3	1405	-99	-99	-99	28028	-99	28028	28028	28028
26	3	4	2095	-99	28251	27799	27799	28251	27799	27799	27799
27	3	5	2199	-99	27741	27605	27605	27741	27605	27605	27605
28	3	6	4127	-99	-99	-99	28072	-99	28072	28072	28072
29	3	7	5091	-99	-99	-99	27761	-99	27761	27761	27761
30	3	8	5188	-99	-99	-99	27881	30610	27881	27881	27881
31	3	9	5434	-99	-99	-99	28311	-99	28311	28311	28311
32	3	10	5480	-99	-99	-99	-99	27986	-99	27986	27986
33	3	11	5882	-99	-99	-99	28043	37009	28043	28043	28043
34	4	1	273	-99	-99	-99	23947	-99	23947	23947	23947
35	4	2	391	-99	-99	-99	23457	-99	23457	23457	23457
36	4	3	1107	-99	24306	24227	24227	24306	24227	24227	24227
37	4	4	1704	-99	26821	26707	24675	26559	24675	24675	24675
38	4	5	1831	-99	23891	23879	23879	23891	23879	23879	23879
39	4	6	9790	-99	-99	-99	23434	28396	23434	23434	23434
40	4	7	195	-99	25201	24941	24430	26639	24430	24430	24430
41	4	8	4006	-99	-99	-99	23412	-99	23412	23412	23412
42	4	9	8908	-99	-99	-99	-99	-99	-99	23225	-99
43	4	10	5635	-99	-99	-99	23961	-99	23961	23961	23961
44	4	11	1100	-99	-99	23901	23901	24506	23901	23901	23901
45	5	1	314	-99	-99	-99	24761	20588	24761	20588	20588
46	5	2	383	-99	-99	20460	19734	-99	19734	19734	19734
47	5	3	644	-99	23832	21046	21046	23832	21046	21046	21046
48	5	4	778	-99	-99	-99	21230	23131	21230	21230	21230
49	5	5	1281	-99	-99	-99	21472	-99	21472	21472	21472
50	5	6	8960	-99	-99	-99	21048	-99	21048	21048	21048
51	5	7	213	-99	-99	-99	20712	20602	20712	20602	20602
52	5	8	593	-99	-99	-99	20768	-99	20768	20768	20768
53	5	9	5394	-99	-99	20649	20649	28549	20649	20649	20649
54	5	10	7160	-99	-99	-99	21024	20600	21024	20600	20600
55	5	11	8880	-99	-99	-99	21197	-99	21197	21197	21197

FILE N 3
FINAL EXPOSURE DATE,

seq	set	c-c	ID	Ni_3S_2	$(Ni_3Fe)S$	$NiSO_4$	Ni pentlandite	$(Ni_2Fe)O$	Ni pyrrhotite	Ni	Dust
1	1	1	46	-99	33832	33803	33832	33906	31443	33906	33906
2	1	2	299	-99	-99	-99	31528	-99	31528	31528	31528
3	1	3	704	-99	-99	-99	-99	31970	-99	31970	31970
4	1	4	956	-99	28029	27496	28029	36614	27496	36614	36614
5	1	5	1214	-99	-99	-99	35073	35735	35073	35735	35735
6	1	6	1799	-99	-99	-99	-99	29644	-99	29644	29644
7	1	7	1884	-99	-99	27272	33670	34586	33670	34586	34586
8	1	8	5355	-99	-99	-99	35965	-99	35965	35965	35965
9	1	9	3685	-99	-99	-99	33950	32565	33950	33950	33950
10	1	10	8305	-99	-99	26173	36098	-99	36098	36098	36098
11	1	11	8910	-99	-99	-99	-99	34757	-99	34757	34757
12	2	1	250	-99	24044	21226	24044	33774	22552	33774	33774
13	2	2	644	-99	26732	25573	26732	29314	25573	29314	29314
14	2	3	778	-99	-99	-99	23889	35518	23889	35518	35518
15	2	4	1281	-99	-99	-99	33090	-99	33090	33090	33090
16	2	5	9750	-99	-99	28854	28854	34971	28854	34971	34971
17	2	6	9251	-99	-99	30527	30527	33602	30527	33602	33602
18	2	7	679	-99	-99	-99	28771	21016	28771	28771	28771
19	2	8	3617	-99	-99	-99	-99	33367	-99	33367	33367
20	2	9	8880	-99	-99	-99	32871	-99	32871	32871	32871
21	2	10	5101	-99	-99	-99	30625	33578	30625	33578	33578
22	2	11	7450	-99	-99	25232	26387	33967	26387	33967	33967
23	3	1	262	-99	-99	-99	37528	37113	37528	37528	37528
24	3	2	1311	-99	-99	-99	38352	-99	38352	38352	38352
25	3	3	1405	-99	-99	-99	38912	-99	38912	38912	38912
26	3	4	2095	-99	38936	31338	38936	38936	31338	38936	38936
27	3	5	2199	-99	38806	31817	38806	38806	31854	38806	38806
28	3	6	4127	-99	-99	-99	38912	-99	38912	38912	38912
29	3	7	5091	-99	-99	-99	38880	-99	38880	38880	38880
30	3	8	5188	-99	-99	-99	36442	38933	36442	38933	38933
31	3	9	5434	-99	-99	-99	38806	-99	38806	38806	38806
32	3	10	5480	-99	-99	-99	-99	38806	-99	38806	38806
33	3	11	5882	-99	-99	-99	36574	37832	36574	37832	37832
34	4	1	273	-99	-99	-99	26179	-99	26179	26179	26179
35	4	2	391	-99	-99	-99	28352	-99	28352	28352	28352
36	4	3	1107	-99	25905	29785	29785	33340	29785	33340	33340
37	4	4	1704	-99	28868	27776	28868	27788	27776	28868	28868
38	4	5	1831	-99	26538	25040	34733	26538	34733	34733	34733
39	4	6	9790	-99	-99	-99	26905	32565	26905	32565	32565
40	4	7	195	-99	34818	25152	34818	28524	25152	34818	34818
41	4	8	4006	-99	-99	-99	34575	-99	34575	34575	34575
42	4	9	8908	-99	-99	-99	-99	-99	-99	28874	-99
43	4	10	5635	-99	-99	-99	34995	-99	34995	34995	34995
44	4	11	1100	-99	-99	23901	32943	33861	32943	33861	33861
45	5	1	314	-99	-99	-99	28090	20854	28090	28090	28090
46	5	2	383	-99	-99	20460	36098	-99	36098	36098	36098
47	5	3	644	-99	26732	25573	26732	29314	25573	29314	29314
48	5	4	778	-99	-99	-99	23889	35518	23889	35518	35518
49	5	5	1281	-99	-99	-99	33090	-99	33090	33090	33090
50	5	6	8960	-99	-99	-99	34925	-99	34925	34925	34925
51	5	7	213	-99	-99	-99	31610	34894	31610	34894	34894
52	5	8	593	-99	-99	-99	28363	-99	28363	28363	28363
53	5	9	5394	-99	-99	22401	27002	28549	27002	28549	28549
54	5	10	7160	-99	-99	-99	30040	20600	30040	30040	30040
55	5	11	8880	-99	-99	-99	32871	-99	32871	32871	32871

APPENDIX # 3

This appendix includes all the tables for the effect of the different contaminants, the model selection and the confidence intervals for the regression parameters.

TABLE # 2A

DEMOGRAPHIC DATA OF ALL THE SETS

OF CASES AND CONTROLS

Set	Birth date mean (standard deviation)	Initial working age mean (standard deviation)	Final working age mean (standard deviation)
cases	16.973 (10.381)	30.815 (7.386)	56.249 (10.07)
one control	16.982 (10.368)	31.054 (7.794)	60.036 (6.297)
two controls	17.018 (10.309)	31.045 (7.712)	60.099 (6.344)
four controls	17.027 (10.242)	31.108 (7.656)	59.525 (6.937)
six controls	17.083 (10.157)	31.054 (7.619)	58.754 (7.682)
eight controls	17.2 (10.024)	30.894 (7.443)	57.84 (8.643)

TABLE # 3

EFFECT OF DIFFERENT SETS OF CASES AND CONTROLS ON THE ESTIMATION OF Ni_3S_2 USING THE LOGISTIC REGRESSION MODEL.

case controls set	Deviance	Degrees of Freedom	Reduction of the deviance*	Estimate	95% Confidence Interval
1 : 1	153.88	221	-	-	-
1 : 2	243.89	332	-	-	-
1 : 4	357.3	554	-	-	-
1 : 6	431.99	776	-	-	-
1 : 8	487.78	997	0.21	-.03040	0.14767 -0.20847
1 : 10	532.33	1219	0.17	-0.02866	0.15705 -0.20703

*with respect to the model with one parameter.

TABLE # 4

EFFECT OF DIFFERENT SETS OF CASES AND CONTROLS ON THE ESTIMATION OF $(Ni_3Fe)S$ USING THE LOGISTIC REGRESSION MODEL.

case controls set	Deviance	Degrees of Freedom	Reduction of the deviance*	Estimate	95% Confidence Interval**
1 : 1	152.54	220	1.34	-.001187	9.4 -33.2
1 : 2	241.54	331	1.94	-.001395	7.0 -34.9
1 : 4	355.15	553	2.15	-.001448	6.5 -35.4
1 : 6	430.12	775	1.16	-.001339	7.3 -34.1
1 : 8	486.99	997	0.79	-.000866	67.1 -106.7
1 : 10	532.1	1219	.23	-.000473	6.5 -35.4

*with respect to the model with one parameter.

** (X1000)

TABLE # 5

EFFECT OF DIFFERENT SETS OF CASES AND CONTROLS ON THE ESTIMATION OF $NiSO_4$ USING THE LOGISTIC REGRESSION MODEL.

case controls set	Deviance	Degrees of Freedom	Reduction of the deviance*	Estimate	95% Confidence Interval**
1 : 1	153.21	220	0.07	-.004851	-128.0 118.3
1 : 2	242.83	331	1.06	-.006644	-201.4 68.6
1 : 4	356.37	553	0.93	-.006133	-158.7 90.1
1 : 6	431.04	775	0.88	-.006303	-198.0 72.0
1 : 8	487.3	997	0.48	-.004298	-170.6 84.6
1 : 10	532.02	1219	0.31	-.003430	-192.4 69.7

*with respect to the model with one parameter.

**(X1000)

TABLE # 6

EFFECT OF DIFFERENT SETS OF CASES AND CONTROLS ON THE ESTIMATION OF Ni IN PENTLANDITE USING THE LOGISTIC REGRESSION MODEL.

case controls set	Deviance	Degrees of Freedom	Reduction of the deviance*	Estimate	95% Confidence Interval**
1 : 1	151.56	220	2.32	-.000395	-13.4 2.1
1 : 2	240.1	331	3.79	-.000752	-15.7 0.62
1 : 4	353.02	553	4.27	-.000792	-12.5 2.9
1 : 6	427.07	775	3,83	-.000820	-16.5 0.92
1 : 8	485.07	997	2.71	-.000622	-14.2 1.7
1 : 10	530.68	1219	1.65	-.000479	-16.1 0.29

*with respect to the model with one parameter.

**(X1000)

TABLE # 7

EFFECT OF DIFFERENT SETS OF CASES AND CONTROLS ON THE ESTIMATION OF Ni_2FeO USING THE LOGISTIC REGRESSION MODEL.

case controls set	Deviance	Degrees of Freedom	Reduction of the deviance*	Estimate	95% Confidence Interval**
1 : 1	146.27	220	7.61	-.002451	-72.7 -26.6
1 : 2	227.83	331	16.06	-.003712	-59.9 -14.6
1 : 4	342.4	553	14.9	-.003787	-49.8 -8.3
1 : 6	420.5	775	17.1	-.003289	-351.8 -306.6
1 : 8	478.39	997	9.39	-.002955	-51.6 -8.2
1 : 10	522.91	1219	9.42	-.002869	-61.4 -14.7

*with respect to the model with one parameter.

**(X1000)

TABLE # 8

EFFECT OF DIFFERENT SETS OF CASES AND CONTROLS ON THE ESTIMATION OF Ni IN PYRRHOTITE USING THE LOGISTIC REGRESSION MODEL.

case controls set	Deviance	Degrees of Freedom	Reduction of the deviance*	Estimate	95% Confidence Interval**
1 : 1	150.43	220	3.45	-.003990	82.9 -2.9
1 : 2	238.35	331	5.54	-.005041	-93.3 -8.0
1 : 4	348.12	553	9.18	-.006518	-108.6 -22.0
1 : 6	419.52	775	5.92	-.007665	-122.1 -32.0
1 : 8	475.55	997	12.23	-.007533	-111.8 -25.0
1 : 10	522.07	1219	10.26	-.006867	-100.0 -14.0

*with respect to the model with one parameter.

**(X1000)

TABLE # 9

EFFECT OF DIFFERENT SETS OF CASES AND CONTROLS ON THE ESTIMATION OF TOTAL NIKEL USING THE LOGISTIC REGRESSION MODEL.

case controls set	Deviance	Degrees of Freedom	Reduction of the deviance*	Estimate	95% Confidence Interval**
1 : 1	146.43	220	7.45	-.000795	-14.2 -2.0
1 : 2	229.62	331	14.27	-.001185	-18.8 -5.0
1 : 4	342.78	553	14.52	-.001211	-85.2 -72.0
1 : 6	413.67	775	13.53	-.001148	-18.5 -4.0
1 : 8	473.76	997	9.02	-.000916	-98.4 -85.0
1 : 10	525.56	1219	6.77	-.000786	-29.3 -15.0

*with respect to the model with one parameter.

**(X1000)

TABLE # 10

EFFECT OF DIFFERENT SETS OF CASES AND CONTROLS ON THE ESTIMATION OF TOTAL DUST USING THE LOGISTIC REGRESSION MODEL.

case controls set	Deviance	Degrees of Freedom	Reduction of the deviance*	Estimate ***	95% Confidence Interval**
1 : 1	135.56	220	18.32	-3.284	-0.50 -0.16
1 : 2	208.43	331	35.46	-4.756	-0.64 -0.31
1 : 4	316.22	553	41.08	-5.375	-0.60 -0.26
1 : 6	391.93	775	38.92	-5.351	-0.71 -0.36
1 : 8	455.91	997	31.87	-4.729	-2.2 -1.3
1 : 10	505.48	1219	12.24	-4.257	-0.71 -0.36

*with respect to the model with one parameter.

**(X1000)

***(EX10-05)

TABLE# 11

EFFECT OF SELECT CONTAMINANTS ON THE CASE-CONTROL SET 1:4
EVALUATION OF MULTIVARIATE MODELS.*

MODEL	DEVIANCE	CHANGE OF DEVIANCE**	DEGREES OF FREEDOM	CHANGE OF DEGREES OF FREEDOM
1	343.89	-	555	-
Ni2FeO	327.83	-16.06	554	1
Ni in Pyrrhotite	324.32	-19.57	554	1
Total Nickel	329.62	-14.27	554	1
Total Dust	308.43	-35.46	554	1
Ni2FeO+Pyrrhotite	308.19	-35.70	553	1
Ni2FeO+Total Nickel	320.22	-23.67	553	2
Ni2FeO+Total Dust	305.40	-38.49	553	2
Ni in Pyrrhotite+Total Ni	324.32	-19.57	553	2
Ni in Pyrrhotite+Total Dust	308.39	-35.50	553	2
Total Ni+Total Dust	306.41	-37.48	553	2
Ni2FeO+Ni in Pyrrhotite +Total Ni	304.88	-39.81	552	3
Ni2FeO+Ni in Pyrrhotite + total Dust	303.45	-40.23	552	3
Ni2FeO+Total Ni +Total Dust	303.42	-40.08	552	3
Ni in Pyrrhotite+Total Ni +Total Dust	304.84	-39.05	552	3
Ni2FO+Ni in Pyrrhotite +Total Ni+Total Dust	303.42	-40.08	552	3

* all the models are statistically significant ($p<0.01$)

** Change of deviance with respect of the model adjusted by the mean.

TABLE# 12

EFFECT OF SOME NICKEL CONTAMINANTS IN THE FALCONBRIDGE MINE.

INITIAL DATE OF EXPOSURE FOR THE 1:4 CASE:CONTROL SET.*

MODEL**	ESTIMATE (xE+06)	STD. ERROR OF ESTIMATE (xE+06)	DEVIANCE	DEGREES OF FREEDOM	CHANGE OF DEVIANCE
1" (a)	-2.398	9475	532.33	554	-
1+Ni3S2	-114.3	36.2	532.16	553	-0.17
1+(Ni3Fe)S	0.09345	0.87	532.32	553	-0.01
1+NiSO4	-0.6948	0.853	531.65	553	-0.68
1+Ni in Pentlandite	-0.766	0.957	531.72	553	-0.61
1+Ni2FeO	-1.24	0.675	528.95	553	-3.38
1+Ni in Pyrrhotite	-0.9024	0.936	531.44	553	-0.89
1+Total Ni	-0.3273	2.57	532.32	553	-0.01
1+Total dust	-0.247	1.92	530.8	553	-1.53

* none of the models are statistically signifiant ($p<0.01$).

** terms are initial date of exposure

TABLA# 13

EFFECT OF SOME NICKEL CONTAMINANTS IN THE FALCONBRIDGE MINE

FINAL DATE OF EXPOSURE FOR THE 1:4 CASE:CONTROL SET.*

MODEL**	ESTIMATE (xE+06)	STD. ERRO OF ESTIMATE (xE+06)	DEVIANCE	DEGREES OF FREEDOM	CHANGE OF DEVIANCE *
1"	-2.398	94750	532.33	554	-
1+Ni3S2	-113.8	36.09	532.16	553	-0.17
1+(Ni3Fe)S	-0.0873	7.69	532.33	553	0
1+NiSO4	-6.101	7.671	531.68	553	-0.65
1+Ni in Pentlandite	-13.37	7.235	529.22	553	-3.11
1+Ni2FeO	-11.59	5.88	528.43	553	-3.9
1+Ni in Pyrrhotite	-13.39	7.28	529.2	553	-3.13
1+Total Ni	-113.1	21.74	505.91	553	-26.42
1+Total Dust	-55.6	11.38	514.32	553	-18.01

* statistically significant ($p<0.01$)

** terms are final date of exposure