INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI[®]

Assessment of Record Linkage and Measurement Error in Cohort Mortality Studies

by

RANJEETA MALLICK, M.Sc.

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements of the degree of
Doctor of Philosophy

School of Mathematics and Statistics
Ottawa-Carleton Institute for Mathematics and Statistics
Carleton University
Ottawa, Ontario, Canada

April 10, 2005

©2005, Ranjeeta Mallick

Library and
Archives Canada

Bibliothèque et
Archives Canada

0-494-06739-X

Published Heritage
Branch

Direction du
Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

Epidemiological studies are designed to investigate disease causation through observational study of human population groups. The overall statistical goal of an epidemiological study is accuracy and precision in estimation; that is, to estimate the value of the parameter of interest with as little error as possible. In addition to random errors, epidemiologic studies may involve systematic error which undermines their validity. Record linkage error and measurement error are two common errors that occur in epidemiological studies.

In this thesis, we study the impact of linkage errors on estimates of various epidemiological indicators of risk such as relative risk regression model parameters and standard mortality ratios. We describe a method to adjust the biases in the indicators of risk in the presence of linkage errors. To explore the impact of record linkage errors in cohort mortality studies in more detail, we conduct a simulation study based on actual data from the National Dose Registry of Canada being linked with the Canadian Mortality Data Base.

The proportional hazards regression model is commonly used to model survival data as a function of covariates. Typically, these covariates are measured with error and are not directly observable. The Cox proportional hazards model is applied in the Harvard Six Cities Study to estimate the mortality rate ratios for exposure to different air pollutants in individuals, with and without other risk factors as covariates. There are several methods in the literature for adjusting for measurement errors. In this thesis, we consider two recent methods, namely RCAL and SIMEX, and evaluate their effectiveness in correcting measurement errors in the Cox proportional hazard model, using real data from the Harvard Six Cities Study. We use computer simulation to evaluate the accuracy and precision of both RCAL and SIMEX, and to assess the robustness of RCAL to mis-specification of the measurement error distribution.

Acknowledgement

First and foremost, I wish to express my sincere gratitude and appreciation to my thesis supervisors, Professor Daniel Krewski, Professor J.N.K. Rao, and Dr. Jan M. Zielinski whose continued guidance, support and patience made this thesis a manageable effort.

I am very much grateful to Professor Anup Dewanji of Indian Statistical Institute for all the time he has spent working with me throughout my thesis work and for his valuable guidance. I would also like to thank Dr. Kevin Brand of McLaughlin Centre for Population Health Risk Assessment, University of Ottawa and Professor Karen Fung of University of Windsor for their advice on many related issues.

Finally, I am grateful to my family members who showed their support in my efforts with continued interest and encouragement.

Contents

1	Introduction	1
1.1	Epidemiological Studies	1
1.1.1	Cohort Mortality Studies	3
1.1.2	Measures of Effect in Epidemiological Studies	5
1.2	Sources of Error	5
1.3	Objective of Thesis	7
1.4	Outline of Thesis	10
2	Impact of Record Linkage Errors in Epidemiological Studies	11
2.1	Introduction	11
2.2	Record Linkage Errors	14
2.3	Relative Risk Regression Model	20
2.4	Effect of Linkage Errors on the Observed and Expected Number of Deaths	22
2.4.1	Linkage errors	23
2.4.2	Expectations and variances of differences between the observed and expected numbers of deaths	25
2.5	Effect on Linkage Errors on Estimates of SMRs	28
2.6	Effect on Linkage Errors on Relative Risk Regression Parameters	32
2.7	A Simulation Study Based on Real Data	36
2.7.1	Description of Data	36
2.7.2	Simulation Model	38
2.7.3	Simulation Results	42
2.8	Conclusions	46

3	Adjustment for Record Linkage Errors	50
3.1	Introduction	50
3.2	General Method for Adjusting Statistical Estimates	51
3.2.1	The Sampling Scheme	52
3.3	Adjustment of the Observed Number of Death and Expected Number of Deaths	55
3.3.1	Unbiasedness and Precision of Adjusted Counts	56
3.4	Adjusted SMRs	65
3.5	Adjusted Relative Risk Regression Parameters	67
3.5.1	Accounting for Overdispersion	67
3.6	Real-Life Data Analysis using Generalized Method	70
3.6.1	Simulation Model	70
3.6.2	Simulation Results	72
3.7	Conclusions	80
4	Evaluation of RCAL and SIMEX Methods for Correcting Measure- ment Errors in Cox Proportional Hazard Model	82
4.1	Introduction	82
4.2	Measurement Error Correction Methods	84
4.2.1	Quasi-likelihood Estimation	85
4.2.2	Regression Calibration Algorithm	86
4.2.3	Simulation Extrapolation	88
4.3	Measurement Error Adjustment in the Cox Proportional Hazard Model	90
4.3.1	Harvard Six Cities Study	90
4.3.2	Error Distributions	92
4.3.3	Cox Proportional Hazard Model	95
4.3.4	Application of RCAL to Six Cities Study	95
4.3.5	Application of SIMEX to Six Cities Study	101
4.3.6	Results	102
4.3.7	Discussions	120
5	Summary and Conclusions	124
5.1	Summary of Main Contributions	124

5.2 Areas for Future Work	126
-------------------------------------	-----

List of Figures

4.1	Box plots of estimates of beta using different methods of estimation when true measurement error is known ($\tau = 1$)	107
4.2	Box plots of estimates of beta using different methods of estimation when true measurement error is known ($\tau = 2$)	108
4.3	Box plots of estimates of beta using different methods of estimation when true measurement error is known ($\tau = 4$)	109
4.4	Box plots of estimates of beta using different methods of estimation when true measurement error is known ($\tau = 8$)	110
4.5	Box plots of estimates of beta using different methods of estimation when true measurement error is known ($\tau = 16$)	111
4.6	Box plots of estimates of beta using different methods of estimation when true measurement error is known ($\tau = 32$)	112
4.7	Summary of Box plots of estimates of beta using different methods of estimation when true measurement error is known.	113
4.8	Box plots of estimates of beta using Naive, SIMEX, RCAL(fixed), RCAL (dynamic) methods of estimation when measurement error is misspecified (true measurement error $\tau = 1$)	114
4.9	Box plots of estimates of beta using Naive, SIMEX, RCAL(fixed), RCAL (dynamic) methods of estimation when measurement error is misspecified (true measurement error $\tau = 2$)	115
4.10	Box plots of estimates of beta using Naive, SIMEX, RCAL(fixed), RCAL (dynamic) methods of estimation when measurement error is misspecified (true measurement error $\tau = 4$)	116

4.11	Box plots of estimates of beta using Naive, SIMEX, RCAL(fixed), RCAL (dynamic) methods of estimation when measurement error is misspeci- fied (true measurement error $\tau = 8$)	117
4.12	Box plots of estimates of beta using Naive, SIMEX, RCAL(fixed), RCAL (dynamic) methods of estimation when measurement error is misspeci- fied (true measurement error $\tau = 16$)	118
4.13	Box plots of estimates of beta using Naive, SIMEX, RCAL(fixed), RCAL (dynamic) methods of estimation when measurement error is misspeci- fied (true measurement error $\tau = 32$)	119

List of Tables

2.1	Relative Bias of SMR (%) due to Record Linkage Errors	44
2.2	Coefficient of Variation of SMR (%) due to Record Linkage Errors . . .	44
2.3	RMSE of SMR (%) due to Record Linkage Errors	45
2.4	Relative Bias of ERR due to Record Linkage Errors	45
2.5	Coefficient of Variation of ERR (%) due to Record Linkage Errors . . .	46
2.6	RMSE of ERR (%) due to Record Linkage Errors	46
3.1	Relative Bias of SMR (%) unadjusted, adjusted with sample size 300 (in parenthesis), adjusted with sample size 600 (<i>italic</i>) and adjusted with sample size 1500 (bold)	74
3.2	Coefficient of Variation of SMR (%) unadjusted, adjusted with sample size 300 (in parenthesis), adjusted with sample size 600 (<i>italic</i>) and adjusted with sample size 1500 (bold)	75
3.3	Root Mean Square Error of SMR (%) unadjusted, adjusted with sample size 300 (in parenthesis), adjusted with sample size 600 (<i>italic</i>) and adjusted with sample size 1500 (bold)	76
3.4	Relative Bias of ERR (%) unadjusted, adjusted with sample size 300 (in parenthesis), adjusted with sample size 600 (<i>italic</i>) and adjusted with sample size 1500 (bold)	77
3.5	Coefficient of variation of ERR (%) unadjusted, adjusted with sample size 300 (in parenthesis), adjusted with sample size 600 (<i>italic</i>) and adjusted with sample size 1500 (bold)	78
3.6	Root Mean Square Error of ERR (%) unadjusted, adjusted with sample size 300 (in parenthesis), adjusted with sample size 600 (<i>italic</i>) and adjusted with sample size 1500 (bold)	79

4.1	Mean and Standard Deviation of PM2.5 Measurements in Selected North American Cities, Averaged Over Monitoring Periods of Varying Duration ($\mu\text{g}/\text{m}^3$)	91
4.2	Mean, Standard Deviation (in parenthesis), and MSE-EME (bold) of Estimates of $\beta = 0.268$ Using Different Methods of Estimation when the Measurement Error Variance τ is Known	104
4.3	Mean, Standard Deviation (in parenthesis), and MSE-EME (bold) of Estimates of $\beta = 0.268$ for RCAL(fixed) when the Measurement Error Variance is Mis-specified	105
4.4	Mean, Standard Deviation (in parenthesis), and MSE-EME (bold) of Estimates of $\beta = 0.268$ for RCAL(dynamic) when the Measurement Error Variance τ^2 is Mis-specified	106

Chapter 1

Introduction

1.1 Epidemiological Studies

An experiment is a set of observations, conducted under controlled circumstances, in which the scientist manipulates the conditions to ascertain what effect such manipulations have on the observations (Rothman and Greenland, 1998). The ideal experiment would create sets of circumstances across which only one factor affecting the outcomes of interest would vary. In biologic experimentation, the ideal of creating duplicate sets of circumstances in which only one relevant factor varies is unrealistic. Instead, the experimenter is satisfied if the variability of the extraneous factors other than the key study variables was too small to affect the outcome under study to an important extent.

Epidemiological studies are designed to non-experimentally investigate disease causation through observational study of human population groups. They are usually concerned with identifying or measuring the effects of risk factors or exposures. When epidemiologic experiments are feasible their design is guided by principles that reduce variation by extraneous factors in comparison with the study factors.

Epidemiologic experiments include clinical trials (with patients as subjects), field trials (with interventions assigned to individual community members), and community intervention trials (with interventions assigned to the whole community). When experiments are not feasible, epidemiologists design non-experimental studies to simulate the conditions if an experiment has been conducted. Non-experimental studies include cohort studies in which subjects are classified (and possibly selected) according to their exposure status and followed over time to ascertain disease incidence; case-control studies, in which the subjects are selected according to their disease status and further classified according to their exposure status; proportional mortality studies, which are best viewed as a type of case-control study; cross-sectional studies, including prevalence studies; and ecologic studies in which the units of observation are groups of people.

Case-control studies are started with the identification of persons with a disease of interest and a control (comparison, referent) group without the disease. The relationship of an attribute to the disease is examined by comparing diseased and non-diseased persons with regard to the frequency or levels of the attribute in each group. In cross-sectional studies, the presence or absence of disease or other health-related variable is determined for each member of the study population, or in a representative sample at one particular time. This contrasts with longitudinal studies in which subjects are followed over a period of time. Cohort studies are defined populations which, as a whole, are followed in an attempt to determine distinguishing subgroup characteristics. There the individuals or populations are followed to assess the outcome of exposures, procedures, or effects of a characteristic (e.g., occurrence of disease). In analyzing these studies, subsets of a defined population are identified. These groups may or may not be exposed to factors hypothesized to influence the probability of the occurrence of a

particular disease or other outcome.

A retrospective study deals with the present and past and does not involve studying future events. In contrast, a prospective study involves observation of a population for a sufficient number of persons over a sufficient number of years to generate incidence or mortality rates subsequent to the selection of the study group. The value of a prospective study lies in the uniqueness of its perspective. A retrospective study depends on memory of past events, whereas a prospective study records the events as they happen. Sometimes it is very difficult to get some information that we may need while doing a retrospective study. When we use death certificates, for example we usually need to interview family members in order to get other relevant information. Even if the subjects are alive, we have to rely on them to recall things that has happened many years ago. As memory is a selective thing, it can introduce many kinds of biases. A prospective study can be slow: the subjects may be uncooperative and the disease of interest may take long time to appear, making the study extremely expensive.

1.1.1 Cohort Mortality Studies

The word “cohort” is derived from the similar Latin word, which identified one of the 10 divisions in a Roman legion (Adami *et al.*, 2002). A soldier, once becoming a member of the division of 500-600 soldiers, remained within the group until death, upon which another did not replace him. In epidemiology, cohorts are groups of individuals that can be followed over time. In cohort studies, individuals are classified according to their exposure and are observed until death or occurrence of the disease.

Cohort studies have the defining characteristic that they are exposure-based. The groups to be studied are selected on the basis of exposure. In general, population cohort groups which can be easily followed up are initially chosen and classified according

to their exposure categories. In special exposure cohorts, the groups are chosen on the basis of a particular exposure. A special exposure cohort becomes necessary when rare exposure needs to be studied, such as those that may occur in occupational settings. Cohort studies have the other defining characteristic that they are patently or conceptually longitudinal; the study groups are observed over a period of time to determine the frequency of disease occurrence among them.

Methodologically, there are two types of cohort studies: closed or fixed cohorts and open or dynamic cohorts. A closed cohort is composed of a set of individuals who are followed for a defined period of time. After becoming a member of the cohort, an individual remains in the cohort until the end of the study or until death. An open cohort is a cohort of individuals whose membership changes over time, with people entering or exiting based on defining criteria.

There are many cohort mortality studies that are being conducted till the present time that started many years before. For example, a UK prospective study of smoking habits and death among doctors started in 1951 and has continued for 50 years (Doll *et al.*, 2004). In this landmark study, Doll and colleagues prospectively followed the cohort and collected updated information on multiple exposures, particularly smoking, over several decades (Adami *et al.*, 2002). Another notable cohort is the Nurses Health study, which began in 1976 with over 121,000 U.S. registered nurses. The cohort was initially supposed to prospectively evaluate the effect of oral contraceptives on the risk of breast cancer. Later, diet and other exposures have been studied in relation to the risk of cancer, as well as other chronic health conditions. Indeed, the British doctors study was among the first one to demonstrate convincingly the role of smoking in the etiology of lung cancer (Doll, 1956). The Nurses' Health Study has provided insight

into the role of estrogens in the etiology of breast cancer (Adami *et al.*, 2002).

1.1.2 Measures of Effect in Epidemiological Studies

There are many ways to assess the association or lack of exposure in a study. Absolute effect measures, which are the differences in incidences or rates between exposed and unexposed group, are easy to interpret. But the common measures are expressed as ratios that are collectively called relative risks. This term includes the risk ratio, rate ratio, odds ratio, standardized mortality ratio, and standardized incidence ratio. The risk ratio is simply the cumulative incidence of disease among the exposed divided by the cumulative incidence risk among the unexposed. The rate ratio is a ratio of rates of disease among the exposed and unexposed. The odds ratio is the ratio of the odds of the disease among the exposed to that of the unexposed. The standardized mortality ratio or standardized incidence ratio is a ratio of the observed number of cases in a cohort divided by the expected number of cases in the general population usually stratified by age and gender characteristic. The relative risk measures provide an estimate of the magnitude of the effect of exposure relative to an unexposed group (Adami *et al.*, 2002).

1.2 Sources of Error

The overall goal of an epidemiologic study is the accuracy in estimation; that is, to estimate the value of the parameter of interest with as little error as possible. Sources of error in estimation can be classified as either random or systematic. Random error is that part of our experience that we cannot predict. Whether or not the sources of error that we cannot explain are actually chance phenomena makes little difference; we treat such errors as random until we can explain them and reduce their impact.

In epidemiologic studies, random error comes from many sources. A major contributor is the process of selecting the specific study subjects, which leads to sampling error. Case-control studies involve a sampling process, whereas cohort studies do not. However, it is standard practice to treat all the epidemiologic studies, including cohort studies, as having sampling error. Even if all the individuals in a population were included in a study, the study subjects are viewed as a sample of the potential biologic experience of an even broader conceptual population. A reduction in chance related uncertainty can be achieved by enrolling progressively larger number of individuals in a study.

In addition to the random errors, epidemiologic studies face problems of systematic error which undermines their validity. The validity of a study is usually divided into two parts. One is internal validity, which is the validity of the inferences pertaining to the source population. The other is external validity or generalization, which is the validity of the inferences pertaining to people outside that population. Out of many types of biases affecting the internal validity the most important ones are selection bias, confounding, and information bias. Selection biases are the distortions that result from procedures used to select subjects and factors that influence the study preparation. The concept of confounding is a central one in modern epidemiology. The apparent effect of the exposure of interest is distorted because the effect of an extraneous factor is mistaken for or mixed with the actual exposure effect (which may be null). The distortion introduced by a confounding factor can be large, and it can lead to overestimation or underestimation of an effect, depending on the direction of the association that the confounding factor has with exposure and disease. Confounding can even change the apparent direction of an effect (Adami *et al.*, 2002). Information bias can occur whenever there are errors in the measurement of variables of inter-

est. For discrete variables, measurement error is usually called classification error or misclassification. Classification error that depends on values of outcome is called differential misclassification, whereas the classification error that does not depend on other variables is called non-differential misclassification. Non-differential misclassification occurs when the proportion of subjects misclassified on exposure does not depend on disease status or when the proportion of subjects misclassified on disease does not depend on exposure. Misclassification can occur simultaneously in both directions. If a confounding variable is misclassified, the ability to control confounding is hampered.

1.3 Objective of Thesis

This thesis focuses on two types of errors in cohort studies. The first is record linkage error, which may be classified as non-differential misclassification. The second is measurement error, which is due to faulty measurement of the variable of interest.

Record linkage is the process of bringing together two or more separately recorded pieces of information pertaining to the same individual (Bartlett *et al.*, 1993). For example, this involves linking records of human exposure to environmental hazards with records on health status, often using computerized methods for matching individual records from different databases. Procedures for computerized record linkage (CRL) have become highly refined, using sophisticated algorithms to evaluate the likelihood of a correct match between two records (Hill, 1988; Newcombe, 1988). The advent of computerized record linkage methodology has facilitated the conduct of cohort mortality studies in which exposure data in one database are electronically linked with mortality data from another database. This also introduces linkage errors due to mismatch of individuals in different databases. In a cohort mortality study, CRL can introduce two

types of error. A false positive occurs when a member of the cohort, who is alive until the end of follow-up, is incorrectly identified as dead during the follow-up period. A false negative occurs when a deceased subject is still considered to be alive.

With regards to record linkage error, our first objective is to investigate the impact of record linkage errors on estimates of epidemiological indicators of risk such as the standardized mortality ratio (SMR) and the parameters of relative risk regression models. This will be done analytically and through simulation to see how these indicators, are in general, subject to biases and additional variability in the presence of linkage errors. Our second objective relating to record linkage is to present a general method for adjusting the biases in the indicators of risk in the presence of linkage errors.

Measurement error is quite common in epidemiological studies. In its most general form, measurement error refers to imperfect information on the values of risk factors or other covariates used to predict health outcomes. Such errors can arise because the variables under study are often subjective, or of biological variability and laboratory errors. Errors in historical variables can occur when respondents are asked to recall events that occurred in the distant past. In many cases, the true value of the variable of interest is simply impossible to measure accurately. Considerable research has been carried out on the development of methods to correct for such errors (Greenland, 1980, Carroll *et al.*, 1995). The spectrum of methods in the literature reflects different approaches to this problem, and different assumptions about the relationship between the true and the observed exposure variables.

The Harvard Six Cities Study (Dockery *et al.*, 1993) was the first large-scale cohort study to demonstrate an association between long-term exposure to fine particulate matter less than 2.5 microns in aerodynamic diameter (PM_{2.5}) and mortality in ur-

ban centres in the United States. Because of the pivotal role of this study in the establishment of the first U.S. national ambient air quality objective for PM_{2.5} in 1997 (Greenbaum *et al.*, 2001), the results of this study were subjected to an independent detailed re-analysis to test the robustness of the findings to alternative analytic methods (Krewski *et al.*, 2000), including an assessment of the effect of exposure measurement error on estimates of risk based on the Cox proportional hazards model. It is well-known that random measurement error leads to downward bias in estimates of risk, and overstatement of the precision of such estimates.

With respect to measurement error correction, we focus on two methods for adjusting for measurement error that have received particular attention in the literature: regression calibration (RCAL) method, and simulation extrapolation (SIMEX) method. The RCAL method was suggested by Carroll and Stefanski (1990). The SIMEX method was proposed by Cook and Stefanski (1994), and later revised by Carroll *et al.* (1996) and Stefanski and Cook (1995). Our objective is to evaluate the effectiveness of RCAL and SIMEX for correcting measurement errors in Cox proportional hazards model, using data from the Harvard Six Cities Study that focuses on the health effects of chronic exposure to airborne pollutants. The Cox proportional hazards model was used in the Harvard Six Cities Study to estimate the mortality rate ratios for exposure to different levels of ambient air pollutants, taking into account the modifying effects of other relevant covariates. We will use computer simulation to evaluate the accuracy and precision of both RCAL and SIMEX, and to assess the robustness of RCAL to mis-specification of the measurement error distribution.

1.4 Outline of Thesis

In Chapter 2, we study the impact of linkage errors on estimates of various epidemiological indicators of risk such as the relative risk regression model parameters and the standardized mortality ratios (SMRs). To explore the impact of record linkage errors in cohort mortality studies in more detail, we conduct a simulation study based on actual data from the National Dose Registry of Canada linked with the Canadian Mortality Data Base (CMDB).

In Chapter 3, we present a method to reduce the biases in the indicators of risk in the presence of linkage errors. The unbiasedness of the estimate proposed and variance of the estimate is also derived. In order to evaluate the effectiveness of the method under real life conditions, the data in Chapter 2 are again used. Using the proposed method of adjustment, we study the reduction in bias of estimates of both standardized mortality ratio and excess relative risk parameters.

In Chapter 4, we evaluate the effectiveness of RCAL and SIMEX for correcting measurement errors in Cox proportional hazard model, using real data from the Harvard Six Cities Study that focuses on the effect of chronic exposure to airborne pollutants. The Cox proportional hazard model was used in the Harvard Six Cities Study to estimate the mortality rate ratios for exposure to different air pollutants in individuals, with and without including other risk factors as covariates. Computer simulation was used to evaluate the accuracy and precision of both RCAL and SIMEX, and to assess the robustness of RCAL to mis-specification of the measurement error distribution.

Finally, Chapter 5 summarizes the main contributions of this thesis, and suggests directions for future research.

Chapter 2

Impact of Record Linkage Errors in Epidemiological Studies

2.1 Introduction

In recent years, a number of historical cohort studies have been carried out in environmental epidemiology using existing administrative databases as information sources (Howe and Spasoff, 1986; Carpenter and Fair, 1990). In general terms, this involves linking records of human exposure to environmental hazards with records on health status, often using computerized methods for matching individual records from different databases. In a cohort mortality study, the vital status of each cohort member is determined by linkage with mortality records maintained by government agencies. Excess mortality within the cohort relative to the general population may be due to exposures experienced by the cohort members.

In specific terms, record linkage is the process of bringing together two or more separately recorded pieces of information pertaining to the same entity (Bartlett *et al.*, 1993). Procedures for computerized record linkage (CRL) have become highly refined, using sophisticated algorithms to evaluate the likelihood of a correct match between

two records (Hill, 1988; Newcombe, 1988). Statistics Canada has developed a CRL system called CANLINK which is capable of handling both single file linkages and linkages between two separate files (Howe and Lindsay, 1981; Smith and Silins, 1981). In this system, weights reflecting the likelihood of a match are attached to pairs of records. Two thresholds are set: potential matches with linkage weights above the upper threshold are considered to be links whereas potential matches with weights below the lower threshold are considered to be non-links. Potential matches with weights between the upper and lower thresholds are resolved using additional information when available. Otherwise, a single threshold is selected to discriminate between links and non-links.

The confidentiality of records protected under the Statistics Act is strictly maintained in any study in which record linkage is employed. All studies requiring linkage with protected data bases must satisfy a rigorous review and approval process prior to implementation, following well-established procedures for data confidentiality (Singh *et al.*, 2002). All linked files with identifying information remain in the custody of Statistics Canada (Labossière, 1986).

Computerized record linkage methods have been used to link environmental exposure data to the Canadian Mortality Data Base (CMDB). For example, a study of Canadian farm operators was initiated to investigate possible relationships between causes of death in over 326,000 farm operators in Canada and various socio-demographic and farming variables, particularly pesticide use (Jordan-Simpson *et al.*, 1990). In this study, the CMDB was linked with the 1971 Census of Population and the 1971 Census of Agriculture. Another ongoing large-scale study is based on the National Dose Registry (NDR) of Canada (Ashmore and Grogan, 1985, Ashmore and Davies, 1989). The

NDR contains information on occupational exposures to ionizing radiation experienced by over 400,000 Canadians dating back to 1950. The NDR has recently been linked to the CMDB to investigate associations between excess mortality due to cancer and occupational exposure to low levels of ionizing radiation (Ashmore *et al.*, 1997, 1998). More recently, the NDR has been linked to the Canadian Cancer Incidence Database (Sont *et al.*, 2001). Joint analyses of incidence and mortality have also been conducted in specific occupational groups such as dentists (Zielinski, *et al.*, 2004). A comprehensive list of other health studies based on linking exposure data with the CMDB has been compiled by Fair (1989).

The success of record linkage studies depends on the quality of databases being linked (Roos *et al.*, 2001). Using population based longitudinal administrative data, Roos *et al.* examined data quality issues in studies of health and health care. Ardal and Ennis (2001) considered systematic errors in administrative databases involved in secondary analysis of health information. Although record linkage studies will benefit from the use high quality data, limitations in data quality may be offset to a certain extent by the large sample sizes found in many administrative data bases.

Record linkage studies have several advantages over traditional epidemiological studies. By using existing administrative databases, the need to collect new data for health studies is circumvented, and large sample sizes can often be achieved with relatively little effort. Depending on the nature of the databases utilized, record linkage provides an inexpensive way of exploring many possible associations in epidemiological studies. Record linkage also has certain disadvantages. There is generally little control over the information collected, and there can be appreciable loss to follow-up. Another disadvantage of record linkage is the occurrence of linkage errors, which is the focus of

this thesis. Inevitably, some records that should match will fail to be linked, and other non-matching records will be incorrectly linked.

Relatively little work has been done to determine the impact of these linkage errors on statistical inference. Neter *et al* (1965) used a simple linear regression model to analyze the impact of errors introduced during the matching process. Their results indicate that linkage errors inflate the residual variance and introduce bias into the estimated slope parameter. Winkler and Scheuren (1991) derived an expression for the bias in estimates of linear regression coefficients due to linkage errors. Advances in the estimation of linkage error rates by Belin and Rubin (1991) enabled Scheuren and Winkler (1993) to implement an improved bias adjustment procedure. Linear regression methods for the analysis of computer matched data files are further discussed by Scheuren and Winkler (1997).

In this chapter, we study the impact of linkage errors on estimates of various epidemiological indicators of risk such as the relative risk regression model parameters and the standard mortality ratios (SMRs). To explore the impact of record linkage errors in cohort mortality studies in more detail, we conduct a simulation study based on actual data from the National Dose Registry of Canada being linked with the Canadian Mortality Data Base.

2.2 Record Linkage Errors

When conducting an epidemiological study, record linkage techniques can be used to combine large databases to achieve large sample sizes without recruiting a large number of subjects. However, linking large databases together can lead to many errors. Often, names are misspelled, or addresses are input incorrectly. These errors can considerably

reduce the number of correct matches in the affected records. In case of files being matched on the basis of unique identifiers, such as the Social Insurance Number or a health care number, a linked file with few mismatches may be possible if the rate of input errors for the identifiers is low. However, when less exact identifiers are used, it is easy to find many errors in a linked file. Some errors that may occur are: records are not matched when a match should exist, records in one file is incorrectly matched to records in another file, or several records in one file seem to match to one record in a second file when logically only one record can do so. Barlett *et al.* (1993) give a description of setting up a record linkage study and the steps necessary for conducting further analysis.

Consider two files, *A* and *B*, that contain information that is required for an epidemiology study. For example, file *A* might contain data on health outcomes such as cancer incidence or mortality and file *B* might contain some kind of relevant exposure data. If the goal of the epidemiologic study is to associate exposure with health outcome, then the two files need to be merged in such a way that the health data for the subjects in file *A* are matched with the corresponding exposure data in file *B*. The files are linked together using identifier variables such as name, address, date of birth, sex, or social insurance number.

When two such files are linked together, the information in the identifier variables may not be adequate to perfectly match data for all subjects in file *A* with the corresponding data in file *B*. Reasons for this may be coding errors, inconsistent coding, spelling mistakes or variations in spelling of names or addresses, or missing data in identifier variables. If a unique identifier such as the Social Insurance Number is available in all files being linked, then there will be a high probability of making the correct

links when the files are merged. However, matching on identifiers such as name or address could cause problems due to spelling errors, variations in spelling, or a change of address. For an identifier such as the subject's name, parsing methods are available to reduce mismatched records due to spelling errors (Newcombe, 1988).

One way to link two records is to determine the probability that a record, or observation, in one file matches a record on another file (Newcombe, 1998). This can be done by comparing the probability of a match given the identifiers to the relative odds of not matching given the identifiers:

$$\frac{P(M|I_1 I_2 \dots I_d)}{P(\overline{M}|I_1 I_2 \dots I_d)} = \frac{P(I_1|M)P(I_2|M) \dots P(I_d|M)P(M)}{P(I_1|\overline{M})P(I_2|\overline{M}) \dots P(I_d|\overline{M})P(\overline{M})}, \quad (2.1)$$

where M and $\overline{M}$ are the events that two records match and do not match, respectively, and I_1 , I_2 and I_d are the outcomes of the identifier variables.

A weight based on the conditional probability in (2.1) is

$$w = \log \frac{P(M|I_1 I_2 \dots I_d)}{P(\overline{M}|I_1 I_2 \dots I_d)} = W_1 + W_2 + \dots + W_d + W,$$

where $W_i = \log\{P(I_i|M)/P(I_i|\overline{M})\}$ ($i = 1, \dots, d$) and $W = \log\{P(M)/P(\overline{M})\}$. Note that without knowing the true matches and non-matches, only the relative probabilities can be calculated.

Several authors have developed methods for calculating weights, notably are Fellegi and Sunter (1969) and Newcombe (1988). In particular, Fellegi and Sunter (1969) proposed a method that takes into account the probability of error for each identifier using a likelihood ratio. The weight is given by $w = \sum_i w_i$ where

$$w_i = \begin{cases} \log(m_i/u_i), & \text{if identifier } i \text{ of a record pair agrees} \\ \log[(1 - m_i)/(1 - u_i)], & \text{if identifier } i \text{ of a record pair disagrees.} \end{cases}$$

Here, $m_i = P(\text{identifier } i \text{ agrees} | \text{record pair} \in M)$, $u_i = P(\text{identifier } i \text{ disagrees} | \text{record pair} \in U)$, M is the true set of matched records, and U is the set of unmatched pairs.

After weights have been assigned to a set of the matched pairs, they must be evaluated to see if there is sufficient evidence to declare the matched pairs a link or not. Fellegi and Sunter (1969) suggest the following scheme for assessing the status of matched records:

$$\text{Potential link} = \begin{cases} \text{link,} & \text{if } w \geq w_u \\ \text{possible link,} & \text{if } w_l < w < w_u \\ \text{non-link,} & \text{if } w \leq w_l, \end{cases}$$

where w_l and w_u are the lower and upper bounds, respectively, for determining links and non-links. Note that w_l and w_u may be equal, say $w_l = w_u = w_s$, in which case the only possible outcomes for a match would be a link and a non-link.

Selecting values for w_l and w_u or w_s is not always an easy task. These values may have to be established by manually linking a small portion of the sample files and calculating the false positive and false negative link rates. A false positive occurs when two records are linked by the computer but not linked manually; a false negative link is when two records are linked manually but are not linked by the computer. This assumes that the person doing the manual linking makes only correct matches or non-matches. The value of w_s is when the weight at which the rates of false positives and false negatives are equal.

The values of m_i and u_i can be estimated by using the EM algorithm, as described by Jaro (1989). Let d be the number of identifiers and N be the number of pairs matched from the two files. Let $\gamma_i^j = 1$ if identifier i agrees for pair j and $\gamma_i^j = 0$ if identifier i does not agree for pair j ($i = 1, \dots, d, j = 1, \dots, N$). Also, let $\gamma^j = (\gamma_i^j)_i$ be the vector of ones and zeroes showing the agreement of identifiers for pair j . Assuming that the identifiers are independent,

$$P(\gamma^j|M) = \prod_{i=1}^d m_i^{\gamma_i^j} (1 - m_i^{\gamma_i^j}) \quad \text{and} \quad P(\gamma^j|U) = \prod_{i=1}^d u_i^{\gamma_i^j} (1 - u_i^{\gamma_i^j}).$$

Using initial values for m_i and u_i , alternating between the estimation and maximization steps of the EM algorithm will give final estimates of m_i and u_i . Included in the same algorithm, Jaro (1989) also described how to estimate the proportion of matched pairs, $p = |M|/|M \cup U|$ in the sample.

Linkage error rates can be estimated using a previous computer matching algorithm (Belin and Rubin, 1991). First, the weights from the matches and the non-matches are transformed to a normal distribution using the Box-Cox transformation (Box and Cox, 1964). Let the distribution of the transformed matched and non-matched weights N_M and N_U have distributions with means μ_M and μ_U and standard deviations σ_M and σ_U , respectively. The transformed weights w_{ti} are assumed to come from the mixture distribution

$$\lambda N_M \left(\frac{w_{ti} - \mu_M}{\sigma_M} \right) + (1 - \lambda) N_U \left(\frac{w_{ti} - \mu_U}{\sigma_U} \right)$$

Next, the mixture coefficient λ is estimated using the previous computer matching. Linkage error rates can be estimated from the above model using the transformed weights and the previous computer matching experience. The standard error of the estimated error rate can also be estimated using the EM algorithm.

Scheuren and Winkler (1993) conducted a simulation study to examine how well the method of Belin and Rubin (1991) performs when linking two files. They used two files with approximately 10,000 records per file for which the true match status was known. To evaluate the linkage algorithm, they altered the files so that the identifier variables had varying degrees of error and then linked the files to see how many records were correctly matched and what weight of the matched and unmatched records were.

The identifiers in the files were altered according to three scenarios : good, mediocre, and poor. For the good scenario, a large number of identifiers were used for matching

and matching probabilities were selected to be optimal. In the mediocre scenario, fewer identifiers were used for matching the good scenario, and minor typographical errors were introduced independently into one-seventh of the records. Matching probabilities were not optimal, but were still quite good. For the poor scenario, even fewer identifier variables were used than the mediocre scenario. Minor typographical errors were introduced into one-fifth of the observations on some of the variables, and into one-third of the observations on other variables. The matching probabilities chosen were considerably less than optimal.

The good scenario generally had a reasonable probability of matching records that matched, with the weight for the matched records generally larger than the weights of the non-matches. For the mediocre scenario, there was more overlap in the weights of the matches and the non-matches. For the poor scenario, the only matches and non-matches that could be distinguished were the matches with the highest weights and the non-matches with the lowest weights. The results also showed that there was a wider range of matching weights for the files in the good scenario than for the files in the mediocre and poor scenarios.

While the results of the simulation are not surprising, Scheuren and Winkler (1993) demonstrate to what extent the quality of identifier variables and prior information can make when linking two files. If the quality of the identifiers is good, then there is a high probability of correctly matching records; for files with poor identifier variables, the probability of correctly matching records is greatly reduced.

2.3 Relative Risk Regression Model

Statistical methods for the analysis of cohort mortality studies are well established (Breslow and Day, 1987). The primary objective of such analysis is to determine if the exposure to the agent of interest increases the mortality rate among cohort members. Mortality is characterized by the hazard function, which specifies the death rate as a function of time. Letting T denote the time of death, the hazard function at time u is formally defined as

$$\lambda(u) = \lim_{\Delta u \downarrow 0} \frac{Pr\{u \leq T < u + \Delta u | T \geq u\}}{\Delta u}. \quad (2.2)$$

Let $\lambda_i(u)$ denote the hazard function for a specific cause of death at time u for individual $i = 1, \dots, N$ in a cohort of size N , and let $\mathbf{z}_i(u)$ represent a corresponding vector of covariates specific to that individual. We assume that the effect of these covariates is to modify the baseline hazard $\lambda^*(u)$ in accordance with the relative risk regression model

$$\lambda_i(u) = \lambda^*(u)\gamma\{\beta' \mathbf{z}_i(u)\}, \quad (2.3)$$

where γ is a positive function of the covariates and β is a vector of regression parameters.

Two special cases of the general relative risk regression model of particular interest are the multiplicative and additive risk regression models. Define the function γ in (2.3) by

$$\log \gamma(z) = \frac{(1+z)^\rho - 1}{\rho}. \quad (2.4)$$

When $\rho = 1$, the general relative risk regression model reduces to be the multiplicative risk regression model

$$\lambda_i(u) = \lambda^*(u)\exp\{\beta' \mathbf{z}_i(u)\}. \quad (2.5)$$

This proportional hazards model was introduced by Cox (1972), and is widely used in the analysis of mortality data (Kalbfleisch and Prentice, 1980). The additive risk regression model

$$\lambda_i(u) = \lambda^*(u) + \beta' \mathbf{z}_i(u) \quad (2.6)$$

occurs as a limiting case as $\rho \rightarrow 0$.

Let t_i^0 and t_i^1 be the age at the time of entry into the study, and the age at the time of loss to follow-up (due to withdrawal from the study, termination of the study, or death) for the i th subject of the cohort. Let $\delta_i = 1$ or 0 , according to whether the i th individual has or has not died at the time of loss to follow-up. The log-likelihood function based on the relative risk model (2.3) may be written as

$$\log L = \sum_{i=1}^N \{ \delta_i \log(\gamma \{ \beta' \mathbf{z}_i(t_i^1) \}) - \int_{t_i^0}^{t_i^1} \gamma \{ \beta' \mathbf{z}_i(u) \} \lambda^*(u) du \}. \quad (2.7)$$

When there is a single covariate $z_i(u) \equiv 1$, the maximum likelihood estimate of $\theta = \exp\{\beta\}$ reduces to the standardized mortality ratio $\text{SMR} = \text{OBS}/\text{EXP}$, where $\text{OBS} = \sum_{i=1}^N \delta_i$ and $\text{EXP} = \sum_{i=1}^N e_i$ are the observed and expected numbers of deaths, respectively, with $e_i = \int_{t_i^0}^{t_i^1} \lambda^*(u) du$.

Maximization of the likelihood function (2.7) can be computationally burdensome with large sample sizes. Breslow *et al.* (1983) simplify the likelihood by assuming that the covariates take on constant values within states through which a subject passes during the course of the study. The states are defined by cross-classification of the covariates of interest. Specifically, suppose that there are J such states $\{S_j; j = 1, \dots, J\}$ such that $\mathbf{z}_i(u) = \mathbf{z}_j$ whenever the i th subject is in S_j at time u . These states are mutually exclusive and exhaustive, so that at any given time u , each member of the cohort will fall into one and only one state. The log-likelihood function (2.7) may then

be written as

$$\log L = \sum_{j=1}^J \{d_{jj} \log(\gamma\{\beta' \mathbf{z}_j\}) - \gamma\{\beta' \mathbf{z}_j\} e_j\}, \quad (2.8)$$

where

$$e_j = \sum_{i=1}^N \int_{\{\mathbf{z}_i(u) \in S_j\}} \lambda^*(u) du \quad (2.9)$$

is the contribution to the expected number of deaths from all person-years of observation in the state S_j , and d_{jj} denotes the total number of deaths in that state. Letting $\Lambda_j(\beta) = \log(\gamma\{\beta' \mathbf{z}_j\})$, the maximum likelihood estimate $\hat{\beta}$ of β is obtained as the solution to the score equation

$$\frac{\partial \log L}{\partial \beta} = \sum_{j=1}^J \frac{\partial \Lambda_j(\hat{\beta})}{\partial \beta} \{d_{jj} - \exp\{\Lambda_j(\hat{\beta})\} e_j\} = 0. \quad (2.10)$$

2.4 Effect of Linkage Errors on the Observed and Expected Number of Deaths

Two principal types of errors can occur when linking data files in CRL (Fellegi and Sunter, 1969). A false positive occurs when a member of the cohort, who is alive until the end of follow-up, is incorrectly identified as dead during the follow-up period. A false negative occurs when a deceased subject is still considered to be alive.

False positives and false negatives in the present context represent a special case of false classification and false misclassification into a specified category (Anderson, 1974). The classification problem arises when an individual has to be classified into one of several categories on the basis of a number of measurements. If an individual is classified as one category wrongfully, then it is a false classification for that category. Similarly, we have a false misclassification if the individual is not classified in the correct category.

In this section we will discuss the effect of these linkage errors on the observed and expected numbers of deaths, respectively. To do this, we first define sets of indices within states which will be used to represent sets of correctly matched records and incorrectly matched records.

2.4.1 Linkage errors

Let A_j and D_j denote the set of labels for those individuals in the cohort who remain alive throughout state S_j , and those who are dead in S_j , respectively. Write D_{jj} as the subset of D_j corresponding to those individuals who have died in S_j . Let A_j^L , D_j^L and D_{jj}^L denote the corresponding sets in the presence of linkage errors. We further define D_j^P as the set of labels of those alive in S_j (that is, in A_j) but labelled as dead in S_j corresponding to the false positives in S_j . Similarly, A_j^N is the set of those dead in S_j (that is, in D_j) but labelled as alive in S_j corresponding to the false negatives in S_j . Let us also write D_{jj}^P as the subset of D_j^P corresponding to those who are labelled to have died in S_j and, similarly, A_{jj}^N as the subset of A_j^N who have died in S_j (that is, in D_{jj}). These sets satisfy the relations $A_j^L = (A_j - D_j^P) \cup A_j^N$, $D_j^L = (D_j - A_j^N) \cup D_j^P$, and $D_{jj}^L = (D_{jj} - A_{jj}^N) \cup D_{jj}^P$.

The effect of linkage errors on the likelihood function in (2.8) may be described as follows. Let t_{ij}^0 denote the time at which the i th individual enters, actually or by linkage error, the j th state S_j . Similarly, t_{ij}^1 denotes the time of death (if it occurs, actually or by linkage error) for the i th individual in S_j and t_{ij}^2 the time of leaving S_j , actually or by linkage error. Note that, if t_{ij}^1 exists, it is less than or equal to t_{ij}^2 . Let us, for the sake of simplicity, assume that t_{ij}^1 , if exists, is equal to t_{ij}^0 ; that is, all the deaths in a state occur at the corresponding entry times in that state. This will underestimate the expected number of deaths. This simplifying assumption is expected

to lead to a reasonable estimate of bias that can occur due to record linkage errors. Assuming all the deaths to occur at the times of leaving the corresponding states also offers similar simplification. Using (2.9) and the decomposition of A_j^L , the expected number of deaths e_j^L in S_j in the presence of linkage errors can be written as

$$\begin{aligned} e_j^L &= \sum_{i \in A_j^L} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \\ &= \sum_{i \in A_j} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du + \sum_{i \in A_j^N} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du - \sum_{i \in D_j^P} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \\ &= e_j - \Delta e_j, \end{aligned} \quad (2.11)$$

where

$$\begin{aligned} e_j &= \sum_{i \in A_j} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \quad \text{and} \\ \Delta e_j &= e_j^P - e_j^N, \end{aligned} \quad (2.12)$$

with

$$e_j^P = \sum_{i \in D_j^P} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \quad \text{and} \quad e_j^N = \sum_{i \in A_j^N} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du. \quad (2.13)$$

For notational convenience, let us write $T_\lambda(i, j)$ for $\int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du$ in what follows. The term Δe_j represents the bias in the expected number of deaths in the j th state due to linkage errors. It follows from (2.11) and (2.12) that the false positives tend to reduce the expected number of deaths and the false negatives tend to increase the expected number of deaths.

Using the decomposition for D_{jj}^L , the observed number of deaths d_{jj}^L in the presence of linkage errors may be characterized as

$$d_{jj}^L = d_{jj} + \Delta d_{jj}, \quad (2.14)$$

where

$$\Delta d_{jj} = d_{jj}^P - d_{jj}^N, \quad (2.15)$$

with d_{jj} , d_{jj}^P and a_{jj}^N denoting the number of individuals in the sets D_{jj} , D_{jj}^P and A_{jj}^N , respectively. The term Δd_{jj} represents the difference between the observed number of deaths in the j th state due to linkage errors. It follows from (2.14) and (2.15) that the false positives will increase the observed number of deaths and the false negatives will reduce the observed number of deaths.

Vital status is often determined by linkage with the CMDDB, which is generally much larger than that of the cohort of interest. When the exposure records of a live individual are incorrectly associated with those of a dead person, the deceased individual usually does not belong to the cohort. Thus, the person-years at risk contributed by the person remaining alive will end prematurely in the year of presumed death; the lost person-years at risk correspond to the time period from the year of presumed death until the end of the follow-up. On the other hand, when the exposure records of a dead individual are incorrectly associated with those of a live person, the person-years at risk contributed by this individual will include an extra period from the actual death-year to the end of the follow-up. Thus, false positives will deflate the number of person-years at risk and false negatives will inflate the number of person-years at risk in the cohort.

2.4.2 Expectations and variances of differences between the observed and expected numbers of deaths

The effect of linkage errors on the observed and expected numbers of deaths depends on the false positive and false negative rates. Let p_j^P and p_j^N denote the false positive and false negative rates, respectively, in S_j ($j = 1, \dots, J$), which are assumed to be constant within S_j and same for all the individuals in A_j and D_j , respectively. This assumption is reasonable whenever individuals in the same state are highly homogeneous, particularly

with respect to attributes such as the quality of personal identifiers that influence linkage error rates. Although this idealized assumption is unlikely to be fully satisfied in practice, it affords considerable simplification in the subsequent evaluation of the effects of linkage errors. Formally, p_j^P (p_j^N) is the conditional probability that an individual in A_j (D_j) is labelled dead (alive) in S_j . That is, $p_j^P = P[i \in D_j^P | i \in A_j]$ and $p_j^N = P[i \in A_j^N | i \in D_j]$.

Let a_j , d_j , a_j^N and d_j^P represent the number of individuals in A_j , D_j , A_j^N and D_j^P , respectively. Then, note that, d_j^P follows a *Binomial* (a_j, p_j^P) distribution and a_j^N follows a *Binomial* (d_j, p_j^N) distribution. Also, d_{jj}^P follows a *Binomial* (a_j, p_{jj}^P) distribution, where p_{jj}^P is the conditional probability that an individual in A_j is labelled as having died in S_j . That is, $p_{jj}^P = P[i \in D_{jj}^P | i \in A_j]$. Clearly, $p_{jj}^P \leq p_j^P$. Similarly, a_{jj}^N follows a *Binomial* (d_{jj}, p_{jj}^N) distribution, where p_{jj}^N is the conditional probability that an individual in D_{jj} is labelled as alive in S_j . That is, $p_{jj}^N = P[i \in A_{jj}^N | i \in D_{jj}]$. Although there is no simple relationship between p_j^N and p_{jj}^N , it is reasonable to assume $p_j^N = p_{jj}^N$ in the present context.

Assuming that linkage errors related to different individuals are independent, the expectation and variance of the difference in the observed number of deaths in S_j , given by Δd_{jj} in (2.15), are

$$E[\Delta d_{jj}] = E[d_{jj}^P] - E[a_{jj}^N] = a_j p_{jj}^P - d_{jj} p_{jj}^N \quad (2.16)$$

and

$$V[\Delta d_{jj}] = V[d_{jj}^P] + V[a_{jj}^N] = a_j p_{jj}^P (1 - p_{jj}^P) + d_{jj} p_{jj}^N (1 - p_{jj}^N). \quad (2.17)$$

Since A_j and D_{jj} consist of different sets of individuals, d_{jj}^P and a_{jj}^N are independent.

Similarly, the expectation and variance of the difference in the expected number of deaths in S_j , given by Δe_j in (2.12), can be calculated as follows. For this purpose,

it is convenient to write e_j^P and e_j^N in terms of the following indicator variables. For $i \in A_j$, define $\xi_{ij} = I\{i \in D_j^P\}$ and $\xi_{ijj} = I\{i \in D_{jj}^P\}$. Also, for $i \in D_j$, define $\psi_{ij} = I\{i \in A_j^N\}$. Then, from (2.13) and the definitions of D_j^P and A_j^N , we have

$$e_j^P = \sum_{i \in A_j} \xi_{ij} T_\lambda(i, j) \quad (2.18)$$

and

$$e_j^N = \sum_{i \in D_j} \psi_{ij} T_\lambda(i, j). \quad (2.19)$$

In particular, one can write $d_{jj}^P = \sum_{i \in A_j} \xi_{ijj}$ and $a_{jj}^N = \sum_{i \in D_{jj}} \psi_{ijj}$, which are useful in deriving (2.16) and (2.17). From (2.18) and (2.19), we have

$$E[\Delta e_j] = E[e_j^P] - E[e_j^N] = p_j^P \sum_{i \in A_j} T_\lambda(i, j) - p_j^N \sum_{i \in D_j} T_\lambda(i, j), \quad (2.20)$$

and

$$V[\Delta e_j] = V[e_j^P] + V[e_j^N] = p_j^P(1 - p_j^P) \sum_{i \in A_j} T_\lambda^2(i, j) + p_j^N(1 - p_j^N) \sum_{i \in D_j} T_\lambda^2(i, j), \quad (2.21)$$

since A_j and D_j consist of different sets of individuals.

The results in (2.16)-(2.17) and (2.20)-(2.21) indicate that record linkage errors will lead to bias and additional variation in the observed and expected number of deaths. Minimizing the variance terms in (2.17) and (2.21) is difficult since the two error rates p_j^P and p_j^N are not functionally independent. Generally, decreasing p_j^P will result in an increase in p_j^N and vice versa (see Section 2.5 for more discussion of this point). Although these error rates are independent of the underlying relative risk regression model γ in (2.3), the mean square error obtained by combining the expectation and variance terms cannot be minimized without specification of the baseline hazard $\lambda^*(u)$, which appears in T_λ .

2.5 Effect on Linkage Errors on Estimates of SMRs

To determine the effect of linkage errors on the SMR, we replace the actual observed and expected numbers of deaths d_{jj} and e_j by the observed and expected number of deaths d_{jj}^L and e_j^L in the presence of linkage errors in the expression $SMR = \sum d_{jj} / \sum e_j$. Letting SMR_L denote the standardized mortality ratios in the presence of linkage errors, we have

$$SMR_L = SMR \left[1 + \frac{\sum \Delta d_{jj}}{\sum d_{jj}} \right] / \left[1 - \frac{\sum \Delta e_j}{\sum e_j} \right]. \quad (2.22)$$

It follows, from (2.11)-(2.15), that the false positives will increase the SMR, whereas the false negatives will decrease the SMR.

By Taylor series expansion of SMR_L about SMR , we get

(i.e., $f(\hat{x}, \hat{y}) \cong f(x, y) + (\hat{x} - x)f_x(x, y) + (\hat{y} - y)f_y(x, y)$),

$$\begin{aligned} SMR_L &= \frac{\sum d_{jj}^L}{\sum e_j^L} \cong \frac{\sum d_{jj}}{\sum e_j} + \frac{\sum \Delta d_{jj}}{\sum e_j} + (\sum \Delta e_j) \frac{\sum d_{jj}}{(\sum e_j)^2} \\ &= SMR + \frac{\sum \Delta d_{jj}}{\sum e_j} + \sum d_{jj} \frac{(\sum \Delta e_j)}{(\sum e_j)^2}. \end{aligned}$$

So, the difference $\Delta SMR = SMR_L - SMR$ can be expressed as

$$\begin{aligned} \frac{SMR_L - SMR}{SMR} &\cong \frac{\sum \Delta d_{jj}}{\sum e_j} \frac{\sum e_j}{\sum d_{jj}} + \sum d_{jj} \frac{(\sum \Delta e_j)}{(\sum e_j)^2} \frac{\sum e_j}{\sum d_{jj}} \\ &= \frac{\sum \Delta d_{jj}}{\sum d_{jj}} + \frac{(\sum \Delta e_j)}{\sum e_j}. \end{aligned} \quad (2.23)$$

Then, the mean and variance of the relative difference in the SMR can be approximated by

$$E \left[\frac{\Delta SMR}{SMR} \right] \simeq \frac{\sum_j E[\Delta d_{jj}]}{\sum_j d_{jj}} + \frac{\sum_j E[\Delta e_j]}{\sum_j e_j} \quad (2.24)$$

and

$$V \left[\frac{\Delta SMR}{SMR} \right] \simeq \left(\sum_j d_{jj} \right)^{-2} V \left[\sum_j \Delta d_{jj} \right] + \left(\sum_j e_j \right)^{-2} V \left[\sum_j \Delta e_j \right]$$

$$+2 \left(\sum_j d_{jj} \right)^{-1} \left(\sum_j e_j \right)^{-1} \text{Cov} \left[\sum_j \Delta d_{jj}, \sum_j \Delta e_j \right], \quad (2.25)$$

respectively. The right hand side of (2.24) can be easily calculated by using (2.16) and (2.20). In order to calculate the right hand side of (2.25), note that

$$V \left[\sum_j \Delta d_{jj} \right] = \sum_j V [\Delta d_{jj}] + 2 \sum_{j < j'} \text{Cov} [\Delta d_{jj}, \Delta d_{j'j'}], \quad (2.26)$$

$$V \left[\sum_j \Delta e_j \right] = \sum_j V [\Delta e_j] + 2 \sum_{j < j'} \text{Cov} [\Delta e_j, \Delta e_{j'}], \text{ and} \quad (2.27)$$

$$\text{Cov} \left[\sum_j \Delta d_{jj}, \sum_j \Delta e_j \right] = \sum_j \text{Cov} [\Delta d_{jj}, \Delta e_j] + \sum_{j \neq j'} \text{Cov} [\Delta d_{jj}, \Delta e_{j'}]. \quad (2.28)$$

Without loss of generality, let us assume, for $j < j'$, that $t_{ij}^0 \leq t_{ij'}^0$, for the same individual i (alive or dead) in S_j and $S_{j'}$; that is, the entry time in S_j is same or earlier than that in $S_{j'}$. We then have, for $j < j'$,

$$\begin{aligned} \text{Cov} [\Delta d_{jj}, \Delta d_{j'j'}] &= \text{Cov} \left[\sum_{i \in A_j} \xi_{ijj} - \sum_{i \in D_{jj}} \psi_{ij}, \sum_{i \in A_{j'}} \xi_{ij'j'} - \sum_{i \in D_{j'j'}} \psi_{ij'} \right] \\ &= \text{Cov} \left[\sum_{i \in A_j} \xi_{ijj}, \sum_{i \in A_{j'}} \xi_{ij'j'} \right] - \text{Cov} \left[\sum_{i \in D_{jj}} \psi_{ij}, \sum_{i \in A_{j'}} \xi_{ij'j'} \right] \\ &\quad - \text{Cov} \left[\sum_{i \in A_j} \xi_{ijj}, \sum_{i \in D_{j'j'}} \psi_{ij'} \right] + \text{Cov} \left[\sum_{i \in D_{jj}} \psi_{ij}, \sum_{i \in D_{j'j'}} \psi_{ij'} \right] \\ &= \text{Cov} \left[\sum_{i \in A_j} \xi_{ijj}, \sum_{i \in A_{j'}} \xi_{ij'j'} \right] - \text{Cov} \left[\sum_{i \in A_j} \xi_{ijj}, \sum_{i \in D_{j'j'}} \psi_{ij'} \right] \\ &= \sum_{i \in A_j \cap A_{j'}} \text{Cov} [\xi_{ijj}, \xi_{ij'j'}] - \sum_{i \in A_j \cap D_{j'j'}} \text{Cov} [\xi_{ijj}, \psi_{ij'}] \\ &= \sum_{i \in A_j \cap A_{j'}} [E(\xi_{ijj} \xi_{ij'j'}) - E(\xi_{ijj}) E(\xi_{ij'j'})] \\ &\quad - \sum_{i \in A_j \cap D_{j'j'}} [E(\xi_{ijj} \psi_{ij'}) - E(\xi_{ijj}) E(\psi_{ij'})] \\ &= - \left(\sum_{i \in A_j \cap A_{j'}} p_{jj}^P p_{j'j'}^P + \sum_{i \in A_j \cap D_{j'j'}} p_{jj}^P p_{j'j'}^N \right). \quad (2.29) \end{aligned}$$

Since one subject who is labelled to die in j cannot be labelled to die in j' again, $\xi_{ijj} \xi_{ij'j'} = 0$ and $E(\xi_{ijj} \xi_{ij'j'}) = 0$. Also, as somebody who is labelled to die in j cannot

be labelled as alive in j' with $j < j'$, $\xi_{ijj}\psi_{ij'} = 0$ and $E(\xi_{ijj}\psi_{ij'}) = 0$. We thus have

$$\begin{aligned}
Cov[\Delta e_j, \Delta e_{j'}] &= Cov \left[\sum_{i \in A_j} \xi_{ij} T_\lambda(i, j) - \sum_{i \in D_j} \xi_{ij} T_\lambda(i, j), \sum_{i \in A_{j'}} \xi_{ij'} T_\lambda(i, j') - \sum_{i \in D_{j'}} \xi_{ij'} T_\lambda(i, j') \right] \\
&= Cov \left[\sum_{i \in A_j} \xi_{ij} T_\lambda(i, j), \sum_{i \in A_{j'}} \xi_{ij'} T_\lambda(i, j') \right] \\
&\quad - Cov \left[\sum_{i \in A_j} \xi_{ij} T_\lambda(i, j), \sum_{i \in D_{j'}} \xi_{ij'} T_\lambda(i, j') \right] \\
&\quad - Cov \left[\sum_{i \in D_j} \xi_{ij} T_\lambda(i, j), \sum_{i \in A_{j'}} \xi_{ij'} T_\lambda(i, j') \right] \\
&\quad + Cov \left[\sum_{i \in D_j} \xi_{ij} T_\lambda(i, j), \sum_{i \in D_{j'}} \xi_{ij'} T_\lambda(i, j') \right] \\
&= \sum_{i \in A_j \cap A_{j'}} T_\lambda(i, j) T_\lambda(i, j') [E(\xi_{ij} \xi_{ij'}) - E(\xi_{ij}) E(\xi_{ij'})] \\
&\quad - \sum_{i \in A_j \cap D_{j'}} T_\lambda(i, j) T_\lambda(i, j') [E(\xi_{ij} \psi_{ij'}) - E(\xi_{ij}) E(\psi_{ij'})] \\
&\quad + \sum_{i \in D_j \cap D_{j'}} T_\lambda(i, j) T_\lambda(i, j') [E(\psi_{ij} \psi_{ij'}) - E(\psi_{ij}) E(\psi_{ij'})] \\
&= \sum_{i \in A_j \cap A_{j'}} p_j^P (1 - p_{j'}^P) T_\lambda(i, j) T_\lambda(i, j') \\
&\quad + \sum_{i \in A_j \cap D_{j'}} p_j^P p_{j'}^N T_\lambda(i, j) T_\lambda(i, j') \\
&\quad + \sum_{i \in D_j \cap D_{j'}} p_{j'}^N (1 - p_j^N) T_\lambda(i, j) T_\lambda(i, j'), \tag{2.30}
\end{aligned}$$

where the third covariance term is zero since $D_j \cap A_{j'} = \emptyset$.

[Note that $E(\xi_{ij}) E(\psi_{ij'}) = 0$ since $\xi_{ij} \psi_{ij'} = 0$. Also, $E(\xi_{ij}) E(\xi_{ij'}) = p_j^N$ since $\xi_{ij} \xi_{ij'} = 1$ only if $\xi_{ij} = 1$. Finally, $E(\psi_{ij}) E(\psi_{ij'}) = p_{j'}^N$ since $\psi_{ij} \psi_{ij'} = 1$ only if $\psi_{ij'} = 1$.] We

further calculate

$$\begin{aligned}
Cov[\Delta d_{jj}, \Delta e_j] &= Cov \left[\sum_{i \in A_j} \xi_{ijj} - \sum_{i \in D_{jj}} \psi_{ij}, \sum_{i \in A_j} \xi_{ij} T_\lambda(i, j) - \sum_{i \in D_j} \psi_{ij} T_\lambda(i, j) \right] \\
&= \sum_{i \in A_j} [E(\xi_{ijj} \xi_{ij}) - E(\xi_{ijj}) E(\xi_{ij})] T_\lambda(i, j) \\
&\quad + \sum_{i \in D_{jj}} [E(\psi_{ij} \psi_{ij}) - E(\psi_{ij}) E(\psi_{ij})] T_\lambda(i, j) \\
&= \sum_{i \in A_j} p_{jj}^P (1 - p_j^P) T_\lambda(i, j) + \sum_{i \in D_{jj}} p_j^N (1 - p_j^N) T_\lambda(i, j), \tag{2.31}
\end{aligned}$$

and

$$\begin{aligned}
Cov[\Delta d_{jj}, \Delta e_{j'}] &= Cov \left[\sum_{i \in A_j} \xi_{ijj} - \sum_{i \in D_{jj}} \psi_{ij}, \sum_{i \in A_{j'}} \xi_{ij'} T_\lambda(i, j') - \sum_{i \in D_{j'}} \psi_{ij'} T_\lambda(i, j') \right] \\
&= \sum_{i \in A_j \cap A_{j'}} T_\lambda(i, j') [E(\xi_{ijj} \xi_{ij'}) - E(\xi_{ijj}) E(\xi_{ij'})] \\
&\quad - \sum_{i \in A_j \cap D_{j'}} T_\lambda(i, j') [E(\xi_{ijj} \psi_{ij'}) - E(\xi_{ijj}) E(\psi_{ij'})] \\
&\quad + \sum_{i \in D_{jj} \cap D_{j'}} T_\lambda(i, j') [E(\psi_{ij} \psi_{ij'}) - E(\psi_{ij}) E(\psi_{ij'})] \\
&= \sum_{i \in A_j \cap A_{j'}} p_{jj}^P (1 - p_{j'}^P) T_\lambda(i, j') + \sum_{i \in A_j \cap D_{j'}} p_{jj}^P p_{j'}^N T_\lambda(i, j') \\
&\quad + \sum_{i \in D_{jj} \cap D_{j'}} p_{j'}^N (1 - p_j^N) T_\lambda(i, j'), \text{ and} \tag{2.32}
\end{aligned}$$

[Note that $\xi_{ijj} \xi_{ij'} = 1$ only if $\xi_{ijj} = 1$. Hence, $E(\xi_{ijj} \xi_{ij'}) = E(\xi_{ijj}) = p_{jj}^P$. Similarly, $\psi_{ij} \psi_{ij'} = 1$ only if $\psi_{ij'} = 1$. Hence, $E(\psi_{ij} \psi_{ij'}) = E(\psi_{ij'}) = p_{j'}^N$. Also $\xi_{ijj} \psi_{ij'} = 0$ and $E(\xi_{ijj} \psi_{ij'}) = 0$ since whoever is identified as dead in state j cannot be identified as alive in j' .] Finally, we have

$$\begin{aligned}
Cov[\Delta d_{j'j'}, \Delta e_j] &= Cov \left[\sum_{i \in A_{j'}} \xi_{ij'j'} - \sum_{i \in D_{j'j'}} \psi_{ij'}, \sum_{i \in A_j} \xi_{ij} T_\lambda(i, j) - \sum_{i \in D_j} \psi_{ij} T_\lambda(i, j) \right] \\
&= \sum_{i \in A_{j'} \cap A_j} T_\lambda(i, j) [E(\xi_{ij'j'} \xi_{ij}) - E(\xi_{ij'j'}) E(\xi_{ij})] \\
&\quad - \sum_{i \in D_{j'j'} \cap A_j} T_\lambda(i, j) [E(\psi_{ij'} \xi_{ij}) - E(\psi_{ij'}) E(\xi_{ij})] \\
&= - \sum_{i \in A_j \cap A_{j'}} p_j^P p_{j'j'}^P T_\lambda(i, j) + \sum_{i \in A_j \cap D_{j'j'}} p_j^P p_{j'}^N T_\lambda(i, j). \tag{2.33}
\end{aligned}$$

[Since $A_{j'} \cap D_j = \emptyset$ and $D_{j'j'} \cap D_j = \emptyset$, $\xi_{ij'j'} \xi_{ij} = 0$ as a subject who is labelled as dead by j falsely cannot again die in $j' > j$. Similarly, $\psi_{ij'} \xi_{ij} = 0$ as a subject cannot be a false positive (dead) in j and a false negative (alive) in $j' > j$.]

Using (2.26)-(2.33), the variance of the relative difference $\Delta \text{SMR}/\text{SMR}$ can be approximated by the right hand side of (2.25). Two conclusions can be drawn from (2.24) and (2.25). First, linkage errors can lead to bias in the estimate of SMR. Second,

both types of linkage errors introduce additional variation into estimates of the SMR. Note that the first term in (2.33) is dominated by the first term in (2.30) for $p_{j'}^P < 0.5$, and the negative covariance term (2.29) is dominated by the calculation of the variance in (2.26). Therefore, the additional variance (2.25) is strictly positive, since the false positive and false negative rates are positive.

2.6 Effect on Linkage Errors on Relative Risk Regression Parameters

We now consider the effect of linkage errors on regression parameter estimates in the general relative risk regression model (2.3). Replacing the observed and expected numbers of deaths d_{jj} and e_j in the log-likelihood function (2.8) with the observed and expected numbers of deaths in the presence of linkage errors d_{jj}^L and e_j^L , we have

$$\log L = \sum_{j=1}^J \{d_{jj}^L \log(\gamma\{\beta' \mathbf{z}_j\}) - \gamma\{\beta' \mathbf{z}_j\} e_j^L\}. \quad (2.34)$$

Let $\hat{\beta}$ and $\tilde{\beta}$ denote the maximum likelihood estimates of β based on $\{d_{jj}, e_j\}$ and $\{d_{jj}^L, e_j^L\}$, respectively. Letting $\hat{\Lambda}_j = \Lambda_j(\hat{\beta})$, the score equation (2.10) can be written as

$$\sum_{j=1}^J \frac{\partial \Lambda_j(\tilde{\beta})}{\partial \beta} \{d_{jj} + \Delta d_{jj} - \exp\{\hat{\Lambda}_j + \Delta \Lambda_j\} (e_j - \Delta e_j)\} = 0, \quad (2.35)$$

where $\Delta \Lambda_j = \Lambda_j(\tilde{\beta}) - \Lambda_j(\hat{\beta})$. Assuming that $\Delta \beta = \tilde{\beta} - \hat{\beta}$ is small, a first order expansion of $\exp\{\hat{\Lambda}_j + \Delta \Lambda_j\}$ around $\hat{\beta}$ gives

$$\exp\{\hat{\Lambda}_j + \Delta \Lambda_j\} \simeq \exp\{\hat{\Lambda}_j\} + \exp\{\hat{\Lambda}_j\} \frac{\partial \hat{\Lambda}_j}{\partial \beta} \Delta \beta. \quad (2.36)$$

Substituting (2.36) into (2.35) leads to

$$\begin{aligned} & \sum_{j=1}^J \frac{\partial \Lambda_j(\hat{\beta})}{\partial \beta} [d_{jj} - \exp\{\hat{\Lambda}_j\} e_j] + \sum_{j=1}^J \frac{\partial \Lambda_j(\hat{\beta})}{\partial \beta} [\Delta d_{jj} + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j - \\ & - \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \hat{\Lambda}_j(\hat{\beta})}{\partial \beta} \Delta \beta + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j \frac{\partial \hat{\Lambda}_j(\hat{\beta})}{\partial \beta} \Delta \beta] \simeq 0. \end{aligned} \quad (2.37)$$

Using (2.10), the first summation in (2.37) is zero. Consequently, since $\Delta e_j \Delta \beta$ is small, $\Delta \beta$ may be approximated by

$$\Delta \beta \simeq \left(\sum_j \frac{\partial \hat{\Lambda}_j}{\partial \beta} \gamma \{ \hat{\beta}' \mathbf{z}_j \} e_j \frac{\partial \hat{\Lambda}'_j}{\partial \beta} \right)^{-1} \sum_j \frac{\partial \hat{\Lambda}_j}{\partial \beta} \{ \Delta d_{jj} + \gamma \{ \hat{\beta}' \mathbf{z}_j \} \Delta e_j \}. \quad (2.38)$$

It follows from (2.38) that

$$E[\Delta \beta] \simeq \left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma \{ \hat{\beta}' \mathbf{z}_j \} e_j \frac{\partial \Lambda'_j}{\partial \beta} \right)^{-1} \sum_j \frac{\partial \Lambda_j}{\partial \beta} \alpha_j, \quad (2.39)$$

where $\alpha_j = E[\Delta d_{jj}] + \gamma \{ \hat{\beta}' \mathbf{z}_j \} E[\Delta e_j]$, which can be calculated from (2.16) and (2.20).

We now find $\frac{\partial \Lambda_j}{\partial \beta}$. Since $\gamma \{ \beta' \mathbf{z}_j \} = 1 + \beta \omega_j$, we have $\Lambda_j(\beta) = \log(\gamma \{ \beta' \mathbf{z}_j \}) = \log(1 + \beta \omega_j)$.

Hence,

$$\frac{\partial \Lambda_j(\beta)}{\partial \beta} = \frac{\omega_j}{1 + \beta \omega_j}$$

So that

$$\begin{aligned} E[\Delta \beta] &\simeq \left\{ \sum_j \left(\frac{\omega_j}{1 + \beta \omega_j} \right)^2 (1 + \beta \omega_j) e_j \right\}^{-1} \left\{ \sum_j \frac{\omega_j}{1 + \beta \omega_j} E[\Delta d_{jj}] + (1 + \beta \omega_j) E[\Delta e_j] \right\} \\ &= \left\{ \sum_j \frac{\omega_j^2 e_j}{1 + \beta \omega_j} \right\}^{-1} \left[\sum_j \frac{\omega_j}{1 + \beta \omega_j} \{ E(\Delta d_{jj}) + (1 + \beta \omega_j) E(\Delta e_j) \} \right] \\ &= \left\{ \sum_j \frac{\omega_j^2 e_j}{1 + \beta \omega_j} \right\}^{-1} \left[\sum_j \frac{\omega_j}{1 + \beta \omega_j} E(\Delta d_{jj}) + \sum_j \omega_j E(\Delta e_j) \right]. \end{aligned} \quad (2.40)$$

Further,

$$V[\Delta \beta] \simeq \left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma \{ \hat{\beta}' \mathbf{z}_j \} e_j \frac{\partial \Lambda'_j}{\partial \beta} \right)^{-1} \left(\sum_j \sum_{j'} \frac{\partial \Lambda_j}{\partial \beta} \Theta_{jj'} \frac{\partial \Lambda'_{j'}}{\partial \beta} \right) \left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma \{ \hat{\beta}' \mathbf{z}_j \} e_j \frac{\partial \Lambda'_j}{\partial \beta} \right)^{-1} \quad (2.41)$$

where

$$\Theta_{jj'} = \text{Cov} [\Delta d_{jj} + \gamma \{ \hat{\beta}' \mathbf{z}_j \} \Delta e_j, \Delta d_{j'j'} + \gamma \{ \hat{\beta}' \mathbf{z}_{j'} \} \Delta e_{j'}],$$

which can be obtained using (2.17), (2.21) and (2.29)-(2.33). Specifically,

$$\begin{aligned}
\Theta_{jj'} &= \text{Cov} [\Delta d_{jj} + \gamma \{\hat{\beta}' z_j\} \Delta e_j, \Delta d_{j'j'} + \gamma \{\hat{\beta}' z_{j'}\} \Delta e_{j'}] \\
&= \text{Cov} [\Delta d_{jj}, \Delta d_{j'j'}] + (1 + \hat{\beta} \omega_{j'}) \text{Cov} [\Delta d_{jj}, \Delta e_{j'}] \\
&\quad + (1 + \hat{\beta} \omega_j) \text{Cov} [\Delta e_j, \Delta d_{j'j'}] + (1 + \hat{\beta} \omega_j)(1 + \hat{\beta} \omega_{j'}) \text{Cov} [\Delta e_j, \Delta e_{j'}]. \quad (2.42)
\end{aligned}$$

Now,

$$\begin{aligned}
\sum_j \sum_{j'} \text{Cov} [\Delta d_{jj}, \Delta d_{j'j'}] &= \sum_j V[\Delta d_{jj}] + 2 \sum_{j < j'} \text{Cov} [\Delta d_{jj}, \Delta d_{j'j'}] \\
\sum_j \sum_{j'} \text{Cov} [\Delta e_j, \Delta e_{j'}] &= \sum_j V[\Delta e_j] + 2 \sum_{j < j'} \text{Cov} [\Delta e_j, \Delta e_{j'}]. \quad (2.43)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\sum_j \sum_{j'} (1 + \hat{\beta} \omega_{j'}) \text{Cov} [\Delta d_{jj}, \Delta e_{j'}] &= \sum_j (1 + \hat{\beta} \omega_j) \text{Cov} [\Delta d_{jj}, \Delta e_{j'}] \\
&\quad + \sum_{j < j'} (1 + \hat{\beta} \omega_{j'}) \text{Cov} [\Delta d_{jj}, \Delta e_{j'}] + \sum_{j > j'} (1 + \hat{\beta} \omega_{j'}) \text{Cov} [\Delta d_{jj}, \Delta e_{j'}] \\
&= \sum_j (1 + \hat{\beta} \omega_j) \text{Cov} [\Delta d_{jj}, \Delta e_{j'}] \\
&\quad + \sum_{j < j'} (1 + \hat{\beta} \omega_{j'}) \text{Cov} [\Delta d_{jj}, \Delta e_{j'}] \\
&\quad + \sum_{j < j'} (1 + \hat{\beta} \omega_j) \text{Cov} [\Delta d_{j'j'}, \Delta e_j] \quad (2.44)
\end{aligned}$$

and

$$\begin{aligned}
\sum_j \sum_{j'} (1 + \hat{\beta} \omega_j) \text{Cov} [\Delta d_{j'j'}, \Delta e_j] &= \sum_j (1 + \hat{\beta} \omega_j) \text{Cov} [\Delta d_{jj}, \Delta e_j] \\
&\quad + \sum_{j < j'} (1 + \hat{\beta} \omega_j) \text{Cov} [\Delta d_{j'j'}, \Delta e_j] \\
&\quad + \sum_{j < j'} (1 + \hat{\beta} \omega_{j'}) \text{Cov} [\Delta d_{jj}, \Delta e_{j'}]. \quad (2.45)
\end{aligned}$$

Therefore,

$$\begin{aligned}
V(\Delta \beta) &\cong \left[\sum_j \frac{\omega_j^2 e_j}{1 + \beta \omega_j} \right]^{-2} \sum_j \sum_{j'} \left[\frac{\omega_j \omega_{j'}}{(1 + \beta \omega_j)(1 + \beta \omega_{j'})} \right] \\
&\quad \times [\text{Cov} [\Delta d_{jj}, \Delta d_{j'j'}] + (1 + \hat{\beta} \omega_{j'}) \text{Cov} [\Delta d_{jj}, \Delta e_{j'}]
\end{aligned}$$

$$\begin{aligned}
& + (1 + \hat{\beta}\omega_j)Cov[\Delta e_j, \Delta d_{j'j'}] + (1 + \hat{\beta}\omega_j)(1 + \hat{\beta}\omega_{j'})Cov[\Delta e_j, \Delta e_{j'}]] \\
= & \left[\sum_j \frac{\omega_j^2 e_j}{1 + \hat{\beta}\omega_j} \right]^{-2} \left[\sum_j \frac{\omega_j^2}{(1 + \hat{\beta}\omega_j)^2} \{ V(\Delta d_{jj}) + (1 + \hat{\beta}\omega_j)Cov[\Delta d_{jj}, \Delta e_j] \right. \\
& + (1 + \hat{\beta}\omega_j)Cov[\Delta e_j, \Delta d_{jj}] + (1 + \hat{\beta}\omega_j)^2 V(\Delta e_j) \} \\
& + \sum_{j < j'} \frac{\omega_j \omega_{j'}}{(1 + \hat{\beta}\omega_j)(1 + \hat{\beta}\omega_{j'})} \{ 2Cov[\Delta d_{jj}, \Delta d_{j'j'}] + (1 + \hat{\beta}\omega_{j'})Cov[\Delta d_{jj}, \Delta e_{j'}] \\
& + (1 + \hat{\beta}\omega_j)Cov[\Delta d_{j'j'}, \Delta e_j] + (1 + \hat{\beta}\omega_j)Cov[\Delta d_{j'j'}, \Delta e_{j'}] \\
& + (1 + \hat{\beta}\omega_{j'})Cov[\Delta d_{jj}, \Delta e_{j'}] \\
& \left. + 2(1 + \hat{\beta}\omega_j)(1 + \hat{\beta}\omega_{j'})Cov[\Delta e_j, \Delta e_{j'}] \} \right] \\
= & \left[\sum_j \frac{\omega_j^2 e_j}{1 + \hat{\beta}\omega_j} \right]^{-2} \left[\sum_j \frac{\omega_j^2}{(1 + \hat{\beta}\omega_j)^2} \{ V(\Delta d_{jj}) + 2(1 + \hat{\beta}\omega_j)Cov[\Delta d_{jj}, \Delta e_j] \right. \\
& + (1 + \hat{\beta}\omega_j)^2 V(\Delta e_j) \} \\
& + \sum_{j < j'} 2 \frac{\omega_j \omega_{j'}}{(1 + \hat{\beta}\omega_j)(1 + \hat{\beta}\omega_{j'})} \{ Cov[\Delta d_{jj}, \Delta d_{j'j'}] + (1 + \hat{\beta}\omega_{j'})Cov[\Delta d_{jj}, \Delta e_{j'}] \\
& + (1 + \hat{\beta}\omega_j)Cov[\Delta d_{j'j'}, \Delta e_j] + (1 + \hat{\beta}\omega_j)(1 + \hat{\beta}\omega_{j'})Cov[\Delta e_j, \Delta e_{j'}] \} \left. \right] \\
& (2.46)
\end{aligned}$$

In the special case of the multiplicative risk model (2.5), the difference $\Delta\beta$ due to linkage errors may be approximated by

$$\Delta\beta \simeq (X'WX)^{-1}X'(\Delta D + \Delta W), \quad (2.47)$$

where $X' = (\mathbf{z}'_1, \dots, \mathbf{z}'_J)$, $\Delta D' = (\Delta d_{11}, \dots, \Delta d_{JJ})$, $W = \text{diag}(\exp(\mathbf{z}'_1 \hat{\beta})e_1, \dots, \exp(\mathbf{z}'_J \hat{\beta})e_J)$, and $\Delta W' = (\exp(\mathbf{z}'_1 \hat{\beta})\Delta e_1, \dots, \exp(\mathbf{z}'_J \hat{\beta})\Delta e_J)$. Note that the weight matrix W is the Fisher information matrix for $\hat{\beta}$. It follows from (2.39) that

$$E[\Delta\beta] \simeq (X'WX)^{-1}X'\Pi, \quad (2.48)$$

where $\Pi' = (\pi_1, \dots, \pi_J)$ with π_j being same as α_j , but $\gamma\{\hat{\beta}'\mathbf{z}_j\}$ replaced by $\exp(\mathbf{z}'_j \hat{\beta})$.

Further,

$$V[\Delta\beta] \simeq (X'WX)^{-1}X'\Psi X(X'WX)^{-1}, \quad (2.49)$$

where Ψ is the matrix of $\Theta_{jj'}$'s with $\gamma\{\hat{\beta}'z_j\}$ replaced by $\exp(z_j'\hat{\beta})$. Note that (2.47)-(2.49) are special cases of (2.38)-(2.41), respectively, written in matrix notation.

With a single covariate $z_i = 1$, $X'WX = e^{\hat{\beta}} \sum_j e_j$, $X'\Delta D = \sum_j d_{jj}$ and $X'\Delta W = e^{\hat{\beta}} \sum_j \Delta e_j$.

In this case,

$$\Delta\beta \simeq \frac{\sum_j \Delta d_{jj} + e^{\hat{\beta}} \sum_j \Delta e_j}{e^{\hat{\beta}} \sum_j e_j}. \quad (2.50)$$

Since $\text{SMR} = e^{\hat{\beta}} = \sum_j d_{jj} / \sum_j e_j$, with $\Delta\beta = \Delta \text{SMR} / \text{SMR}$. In this case, we have

$$\Delta\beta \simeq \frac{\sum_j \Delta d_{jj}}{\sum_j d_{jj}} + \frac{\sum_j \Delta e_j}{\sum_j e_j}. \quad (2.51)$$

Thus, (2.51) may be viewed as a special case of (2.23).

The preceding results indicate that both false positives and false negatives will introduce bias and additional variation into the estimates of relative risk regression parameters. The only negative contribution to this additional variance (2.41) is through $\text{Cov}[\Delta d_{jj}, \Delta d_{j'j'}]$, given by (2.29), and the first term in (2.33) (see $\Theta_{jj'}$). Using the same argument as that used at the end of section 2.4.1, it follows that this additional variance is strictly positive.

2.7 A Simulation Study Based on Real Data

2.7.1 Description of Data

The National Dose Registry of Canada (NDR) is a centralized registry of records of occupational exposure to ionizing radiation. This registry contains records of individual exposure for mainly industrial, medical, and dental workers monitored in Canada for

occupational radiation exposure. The registry, maintained by the Radiation Protection Bureau of Health Canada, contains historical summary records of exposure to x-rays, gamma rays, neutrons, beta particles, and radon daughters. It currently has records on over 500,000 individuals from over 24,000 organizations dating back to 1950. Few workers occupationally exposed to radiation prior to 1951 have records in the NDR.

A cohort mortality study of occupational radiation exposure was conducted based on the records of the NDR. The study cohort consisted of 206,620 individuals. The cohort data were first linked internally so that fragmented records could be combined prior to linkage with the Canadian Mortality Database (CMDB). Individual records in the NDR were then matched to those in the CMDB using computerized record linkage methods, using separate linkages conducted for males and females. Surnames were phonetically coded, and each record in NDR was compared with all records having the same phonetic code in the CMDB. A weight was assigned to each pair of records to reflect the likelihood of a match between individuals in these two databases. These weights were computed on the basis of the agreement, partial agreement, or discrepancy between the surname, the first and second given names, and the date of birth. To minimize potential bias due to linkage errors, thresholds were set at the point where the rates for false positive and false negative links were equal (Howe and Lindsey, 1981; Smith and Silins, 1981). This computerized record linkage (CRL) procedure, called CANLINK, has been developed in Statistics Canada, which is capable of handling both single file linkages and linkages between two separate files.

The NDR has been linked to several databases for different studies. It has recently been linked to the CMDB to investigate associations between excess mortality due to cancer and occupational exposure to low levels of ionizing radiation (Ashmore *et al.*,

1997, 1998). More recently, the NDR has been linked to the Canadian Cancer Incidence Database (Sont *et al.*, 2001). Our simulation study, described in the following section, uses the linkage between NDR and CMDB to introduce errors of both types.

The data used in the simulation are based on a subset of the NDR involving male employees of Atomic Energy of Canada Limited (AECL). The cohort consisted of 6,140 full-time male employees employed by AECL at some time between 1951 and 1987. Using CANLINK to link this cohort to the CMDB, 498 cohort members were identified as having died from some type of cancer. Assuming this to be the true number of cancer deaths, false positives and false negatives were then introduced by computer simulation. We focus on cancer mortality (instead of total mortality) for this study on impact of linkage errors.

Following Gribbin *et al.* (1993), person-years at risk were calculated for each cohort member. The period of observation for each individual commenced in 1951 or the year in which he or she was first employed by AECL, whichever was later. Follow-up terminated in 1987 for those employees still alive at that time, or in the year of death. Person-years at risk for the entire cohort were calculated for states defined by age (0-44, 44-55, 55-65, ..., ≥ 85 years), calendar year (1951-1960, 1961-1970, 1971-1980, 1981-1987), years since first employment (0-9, 10-20, ≥ 20 years), and cumulative exposure to gamma radiation (0, 1-50, 51-100, ≥ 100 mSv), excluding any exposures occurring in the last ten years of observation to allow for a latent period in cancer induction.

2.7.2 Simulation Model

The expected number of deaths in each state was computed using (2.9) and the calendar year, age and sex-specific death rates $\lambda^*(u)$ obtained from Statistics Canada for the Canadian population. The SMR for the AECL cohort was calculated as the ratio

of the observed number of deaths to these expected number of deaths summed over the different states. The association between cancer mortality and radiation exposure was examined using the general relative risk regression model given in (2.3), with $z_i(u) = z_j$ whenever the i th cohort member is in state S_j at time u . For this data set, z_j is the cumulative exposure to gamma radiation for an individual in the j th state. Following Gilbert *et al.* (1990), we consider the linear relative regression model given by $\gamma\{\beta'z_j\} = 1 + \beta'z_j$, with the parameter β representing the excess relative risk (ERR) per mSv attributable to radiation exposure. The maximum likelihood estimate $\hat{\beta}$ of β is obtained by solving (2.10).

Note that, from the description of data in the previous section, the index j for the different states is a function of the four classification variables, namely the period (p), age group (a), years in occupation (y), and cumulative exposure (c), taking altogether $4 \times 6 \times 4 \times 3 = 288$ values. Although this function can be written in different ways, we write

$$\begin{aligned} j = j(p, a, y, c) &= 4 \times 6 \times 3 \times (c - 1) + 4 \times 6 \times (y - 1) + 4 \times (a - 1) + p \\ &= 72c + 24y + 4a + p - 100. \end{aligned}$$

For a fixed value of j , we also write the corresponding period $p = p(j)$ and age group $a = a(j)$.

The set of alive individuals A_j and the set of cancer deaths D_{jj} in a state S_j can be easily identified from the data. Note that for a fixed period p , the union of A_j 's over all j with $p(j) = p$ is the set of living members of the cohort at the end of period p . The entry times t_{ij}^0 's in a state S_j (for all individuals in A_j) are at the beginning of the corresponding period $p(j)$. The deaths in S_j are assumed to occur at this entry time. The exit times from S_j (for those in A_j) are at the end of the corresponding period

$p(j)$. The set of dead individuals at the end of a period p is, as with the A_j 's, the union of D_j 's over all j with $p(j) = p$. Denoting this union as $D^{(p)}$, for $p = 1, 2, 3, 4$, a typical D_j for a j with $p(j) = p$ is given by

$$D_j = D^{(p-1)} \cup D_{jj} \cup D_{jo} ,$$

where D_{jo} is the set of non-cancer deaths in S_j and $D^{(0)}$ is empty.

Since the entry and exit times (t_{ij}^0 and t_{ij}^2) in a state S_j are the same for all individuals i , the calculation of expected number of deaths e_j in S_j is straightforward. The death rates $\lambda^*(u)$, as obtained from Statistics Canada are constant within a state S_j , depending only on the corresponding period $p(j)$ and age group $a(j)$. Let us denote this rate by $\lambda_{p(j),a(j)}^*$. Then, since the deaths are assumed to occur at the entry times, it follows from (2.9) that

$$e_j = a_j \lambda_{p(j),a(j)}^* (t_{p(j)}^2 - t_{p(j)}^0) , \quad (2.52)$$

where a_j is the number of individuals in A_j , and $t_{p(j)}^0$ and $t_{p(j)}^2$ are the entry and exit times in S_j , respectively, depending only on the period $p(j)$, as discussed before. The observed number of cancer deaths d_{jj} are the number of individuals in D_{jj} .

False positives and false negatives are introduced randomly in the following simple manner. At the end of study in 1987 (that is, $p = 4$), a proportion α^+ of all the alive individuals (the union of all A_j 's with $p(j) = 4$, as discussed previously) are randomly selected to be labelled as dead (false positives) in the corresponding S_j 's. These false positives are then randomly assigned a false age (age group, to be more specific) at death following the distribution $\{p_l / \sum_{a \leq a(j)} p_a, l = 1, \dots, a(j)\}$, where p_l 's are the age group-specific death probabilities for the Canadian population, obtained from Statistics Canada. Specifically, p_l is the proportion of deaths in the l th age group ($l = 1, \dots, 6$),

with $\sum_{l=1}^6 p_l = 1$. That is, an individual in A_j with $p(j) = 4$, after being selected for false positive, is labelled as having died in the age-group $l \leq a(j)$ with probability $p_l / \sum_{a \leq a(j)} p_a$. This individual is then labelled to have died from cancer with probability p_{lc}/p_l , where p_{lc} is the probability of cancer death in the l th age group. Note that this conditional probability p_{lc}/p_l is the proportion of cancer deaths in the l th age group. In order to introduce the false negatives, a proportion α^- of deaths (due to cancer or other causes) in S_j are selected randomly to remain alive till the end of study, for all $j = 1, \dots, 288$. After introducing false positives and false negatives in the linked data, we calculate SMR and $\hat{\beta}$ for each simulated data set.

The false positive rate is given by the ratio of the number of false positives to the size of alive people in the AECL cohort (5,642). The false negative rate is given by the ratio of the number of false negatives to the size of dead people in the AECL cohort (498). Both false positive and false negative error rates take values of 0, 0.1%, 0.5%, 1%, 5%, and 10%, leading to $6 \times 6 = 36$ combinations of these two error rates. For each such combination, the relative biases in the estimators of the SMR and excess relative risk were estimated on the basis of 500 simulations. The true SMR and the relative risk regression parameter β was 0.2818 and 0.03156, respectively, calculated from the original data set.

The relative bias (RB) was estimated by

$$RB = \frac{1}{500} \sum_{i=1}^{500} \left(\frac{\hat{\theta}_L^{(i)} - \hat{\theta}}{\hat{\theta}} \right), \quad (2.53)$$

where $\hat{\theta}$ denotes the estimate of the SMR (assuming no linkage error) and $\hat{\theta}_L^{(i)}$ is the estimate based on i th simulation run (including linkage errors). The coefficient of variation (CV) was estimated by

$$CV = \sqrt{\frac{1}{500} \sum_{i=1}^{500} \left(\frac{\hat{\theta}_L^{(i)} - \bar{\theta}}{\hat{\theta}} \right)^2}. \quad (2.54)$$

where $\bar{\theta}$ is the mean of $\hat{\theta}_L^{(i)}$. The root mean square error (RMSE) was defined by

$$RMSE = \sqrt{RB^2 + CV^2}. \quad (2.55)$$

2.7.3 Simulation Results

We observed that the relative bias in the SMR changed with the magnitude of the false positive and false negative rates (see Table 2.1). For a given false positive rate, the relative bias in the SMR decreased as the false negative rate increased. Similarly, for a given false negative rate, the relative bias in the SMR increased with the increase in the false positive rate.

We also observed that the relative bias in the ERR changed with false positive and false negative rates (see Table 2.4). However, the changes in the relative bias in the ERR did not follow a pattern. When the ERR is calculated, cumulative gamma exposure is the covariate used in the model. When false positive or false negative occurs, the cumulative gamma exposure in each strata is affected. The direction of this dependency is not fixed. For example, a false negative in a stratum may increase the cumulative gamma exposure or may not change it at all. Similarly, a false positive may reduce the cumulative gamma exposure or may not change it at all. But at the same time, the number of observed and expected deaths change in a certain direction with a fixed change pattern in false positive and false negative rates. Hence, the relative bias of the estimates of ERR does not follow any direction pattern even though they are sensitive to the change in false positive and false negative rates.

We also observed that the coefficient of variation(CV) of both the SMR and ERR increases with increasing false negative and false positive error rates (see Tables 2.2 and 2.5). The RMSE of the ERR is much larger than that of the SMR, primarily due to the larger CV of the ERR (see Tables 2.3 and 2.6).

Table 2.1: Relative Bias of SMR (%) due to Record Linkage Errors

False Non-link Rate	False Link Rate					
	0%	0.1%	0.5%	1%	5%	10%
0%	0.00	0.53	2.81	5.71	29.00	58.99
0.1%	-0.09	0.43	2.96	5.70	28.90	58.74
0.5%	-0.54	-0.01	2.52	5.25	28.45	58.28
1%	-1.02	-0.48	2.05	4.78	27.97	57.78
5%	-5.20	-4.67	-2.15	0.57	23.68	53.39
10%	-10.44	-9.91	-7.40	-4.69	18.32	47.90

Table 2.2: Coefficient of Variation of SMR (%) due to Record Linkage Errors

False Non-link Rate	False Link Rate					
	0%	0.1%	0.5%	1%	5%	10%
0%	0.00	0.87	2.00	2.75	6.64	9.51
0.1%	0.37	0.91	2.12	2.93	6.82	9.18
0.5%	0.88	1.23	2.27	3.04	6.87	9.22
1%	1.16	1.46	2.46	3.14	6.93	9.29
5%	2.40	2.59	3.11	3.82	7.29	9.57
10%	3.57	3.71	4.13	4.67	7.97	10.04

Table 2.3: RMSE of SMR (%) due to Record Linkage Errors

False Non-link Rate	False Link Rate					
	0%	0.1%	0.5%	1%	5%	10%
0%	0.00	1.02	3.45	6.34	29.75	59.76
0.1%	0.38	1.01	3.64	6.40	29.70	59.45
0.5%	1.03	1.23	3.39	6.07	29.27	59.00
1%	1.54	1.54	3.20	5.72	28.81	58.52
5%	5.73	5.34	3.78	3.86	24.78	54.21
10%	11.03	10.58	8.47	6.62	19.98	48.94

Table 2.4: Relative Bias of ERR due to Record Linkage Errors

False Non-link Rate	False Link Rate					
	0%	0.1%	0.5%	1%	5%	10%
0%	0.00	-5.24	-23.21	32.16	298.39	612.01
0.1%	0.84	22.21	34.41	80.15	274.08	510.36
0.5%	14.47	35.94	46.21	93.23	284.56	521.16
1%	15.86	36.94	46.75	92.81	283.74	522.44
5%	-8.71	11.79	24.36	66.25	259.77	496.44
10%	-27.69	-7.69	6.46	49.50	247.26	489.43

Table 2.5: Coefficient of Variation of ERR (%) due to Record Linkage Errors

False Non-link Rate	False Link Rate					
	0%	0.1%	0.5%	1%	5%	10%
0%	0.00	215.56	521.33	741.61	1468.55	2017.62
0.1%	156.81	307.04	612.85	786.02	1660.87	2119.40
0.5%	335.49	438.63	662.14	857.33	1690.85	2142.91
1%	460.50	544.34	740.30	911.47	1722.18	2171.56
5%	962.82	999.25	1148.08	1217.63	1893.82	2280.25
10%	1447.83	1466.94	1569.61	1602.23	2174.12	2474.73

Table 2.6: RMSE of ERR (%) due to Record Linkage Errors

False Non-link Rate	False Link Rate					
	0%	0.1%	0.5%	1%	5%	10%
0%	0.00	215.62	521.85	742.31	1498.55	2108.4
0.1%	156.81	307.84	613.82	790.09	1683.33	2179.98
0.5%	335.81	440.1	663.75	862.38	1714.63	2205.37
1%	460.77	545.59	741.78	916.19	1745.39	2233.52
5%	962.86	999.318	1148.33	1219.43	1911.56	2333.66
10%	1448.09	1466.94	1569.63	1602.99	2188.13	2522.66

2.8 Conclusions

Record linkage is now a well-established technique in epidemiological studies of population health risks. By linking information on individual exposures from one database to information on health outcomes in another database, it is possible to construct large-scale informative databases on risks to health of populations and population subgroups. The success of such studies will depend to a large extent on the quality of the

two databases being linked, including the amount of information on individual identifiers used to link individuals in the two databases. In most studies, the accuracy of the linkage is examined by estimating the false link (false positive) and false non-link (false negative) rates associated with the linkage process. In practice, this is usually done by drawing a sample of linked and non-linked records, and determining the accuracy of the linkages in the sample using auxiliary information drawn from other sources.

Although CRL has been used for some time in cohort mortality studies, the impact of linkage errors on the reliability of statistical inferences drawn from such studies has not been subjected to detailed investigation. The theoretical results presented in this chapter address this issue. These results show that in addition to inflating the observed number of deaths, false positives will tend to deflate the expected number of deaths. Conversely, false negatives inflate the expected numbers of deaths and deflate the observed number of deaths. Linkage errors were shown to introduce bias into estimates of SMRs. Relative risk regression coefficients are also subject to bias, the direction of which depends on the nature of the regression coefficient. In addition to these biases, linkage errors introduce additional uncertainty into estimates of both SMRs and regression coefficients.

Although we make the simplifying assumption of $t_{ij}^1 = t_{ij}^0$, one can derive the relevant expressions for bias and increased variability without this assumption; however, the expressions are too complex to offer additional insight into the effects of linkage errors. This is also true of the assumption that $p_{jj}^N = p_j^N$. There is a technical issue with the definition of A_j for the state(s) corresponding to the last age interval, which is usually open up to ∞ on the right hand side. In such state(s), the assumption that $t_{ij}^1 = t_{ij}^0$ will be problematic if the probability of dying in this last interval is apprecia-

ble. This problem may be circumvented by assuming the human life span to have a finite upper limit.

As discussed at the end of section 2.4.1, false positives occur primarily when an individual who is alive at the end of the follow-up period is incorrectly linked with a dead person. However, a person who died in one of the states S_j may be falsely linked with another person with an earlier death time. This leads to a false positive which persists until the actual time of death; the analysis in section 2.4 allows for this type of error. Similarly, a dead person may be falsely linked with another person dying at a later time, who is not alive at the end of follow-up. This case is treated as a false negative only up to the false death time. At this false time of death, this will contribute incorrectly to the number of deaths, an error which has not been considered in section 2.4. However, this type of error would not normally be detected in typical record linkage studies in which a simplified manual check is used to identify false positives and false negatives. Since this type of error is likely to be rare, the effect is expected to be small.

In order to further explore the potential impact of linkage errors, let τ_j be the upper age limit for the j th state S_j . (Note that some of the τ_j 's may be equal.) Then, letting α denote the probability of a linkage error (of either type), the false positive and negative rates, p_{jj}^P and p_{jj}^N , may be written as $\alpha P[T \leq \tau_j]$ and $\alpha P[T > \tau_j]$, respectively. In particular, $p_{jj}^P = \alpha P[\tau_{j-1} < T \leq \tau_j]$, where τ_{j-1} is the lower age limit for the j th state, and $p_{jj}^N = p_j^N$. Therefore, the false positive rates may be greater than the false negative rates in the older age groups, with the reverse happening in the younger age groups. Assuming a similar pattern in the size of the D_j 's and A_j 's, some cancellation of terms may take place in the calculation of $E[\Delta e_j]$ in (2.20) and $E[\Delta d_{jj}]$ in (2.16).

This cancellation effect will reduce the expected bias in the SMR and the relative risk regression parameters given in (2.24) and (2.39), respectively.

Although we have considered only all-cause mortality in this chapter, cause-specific mortality can be examined by simple modifications of the definitions of D_{jj} , D_{jj}^L and D_{jj}^P . These sets should then consider only those deaths from the specific cause of interest. Consequently, d_{jj} and e_j should denote, respectively, the observed and expected number of deaths of the specific type in S_j . The hazard function in (2.2) and (2.3) should relate to the specific type of death, with $\lambda^*(u)$ being the corresponding baseline cause-specific hazard rate. Finally, the indicator δ_i in section 2.3 should indicate the specific type of death.

While the preceding analytical results shed considerable light on the effects of linkage errors in cohort mortality studies, it is important to investigate such effects under conditions as close as possible as may be encountered in practice. To this end, we conducted a computer simulation study based on actual data from the National Dose Registry of Canada, in which the introduction of false links and false non-links with known probabilities have been used to further evaluate the impact of linkage errors on estimates of cancer risk (Mallick *et al.*, 2002). These simulation results corroborate the theoretical findings.

While the results reported here may help to clarify the impact of linkage errors on statistical inference, methods that take such errors into account in the statistical analysis remain to be developed. Such methods may be based on response error models employed in survey sampling, used in conjunction with traditional statistical methods for analysis of cohort mortality data. Research in this area is underway. This issue is addressed in the next chapter.

Chapter 3

Adjustment for Record Linkage Errors

3.1 Introduction

In cohort mortality studies, CRL can introduce two types of errors. A false positive occurs when a member of the cohort, who is alive until the end of follow-up, is incorrectly identified as dead during the follow-up period. A false negative occurs when a deceased subject is still considered to be alive. In Chapter 2, we investigated the impact of record linkage errors on estimates of epidemiological indicators of risk such as standardized mortality ratio(SMR) and the parameters of relative risk regression model. The analytical and simulation results demonstrate that these indicators, are in general, subject to bias and additional variability in the presence of linkage errors.

In this chapter, we present a method to adjust the biases in the indicators of risk in the presence of linkage errors. In order to correct for the bias, additional information is required about the distribution of the linkage errors. Specifically, a sample from the population cohort is selected. Each unit in the sample is matched manually and by computerized record linkage method. The manual matching is considered as the true

matching.

In Section 3.2, the general method for adjustment is described. Section 3.3 describes the calculation for the adjusted counts (i.e., adjusted expected deaths and observed deaths). In Section 3.4, the bias and the variance of the adjusted counts are given. Section 3.5 calculates the adjusted estimate of the SMR. In this section the bias and variance of the adjusted estimate is calculated, and an estimate of the variance of the adjusted estimate is also provided. Section 3.6 deals with relative regression parameter. Section 3.6.1 focuses on the situation of simple linear regression based on the adjusted counts. Section 3.6.2 provides an estimate of the regression parameters accounting for over-dispersion using quasi-likelihood functions. Section 3.6.3 deals with the dependence of adjusted counts. Section 3.6.4 gives the simulation results. A discussion of the results of this chapter are given in Section 3.7.

3.2 General Method for Adjusting Statistical Estimates

We will now describe a general method for adjusting statistical estimates in the presence of linkage errors. In order to do that, we introduce the following notations. For each subject $i = 1, \dots, N$ of the cohort, we define the following indicator variables:

$$\delta_{i11} = \begin{cases} 1 & \text{if subject } i \text{ has died and is labelled as having died} \\ 0 & \text{otherwise,} \end{cases}$$

$$\delta_{i10} = \begin{cases} 1 & \text{if subject } i \text{ is alive but is labelled as having died} \\ 0 & \text{otherwise,} \end{cases}$$

$$\delta_{i01} = \begin{cases} 1 & \text{if subject } i \text{ has died but is labelled as being alive} \\ 0 & \text{otherwise, and} \end{cases}$$

$$\delta_{i00} = \begin{cases} 1 & \text{if subject } i \text{ is alive and is labelled as being alive} \\ 0 & \text{otherwise.} \end{cases}$$

We assume that the random vector $(\delta_{i11}, \delta_{i10}, \delta_{i01}, \delta_{i00})$ has a multinomial distribution with parameters $(p_{i11}, p_{i10}, p_{i01}, p_{i00})$. We can only observe the status of subject i in the presence of linkage errors; that is, we can observe $\hat{\delta}_i = \delta_{i11} + \delta_{i10}$ but not $\delta_i = \delta_{i11} + \delta_{i01}$, which is the true status of the subject after the two files have been linked.

Define $Y = \sum_i \delta_i Y_{1i} + \sum_i (1 - \delta_i) Y_{0i}$, where Y_{1i} and Y_{0i} are known constants. When linkage errors are present, we can estimate Y by Y^L , where Y^L is the same as Y with δ_i replaced with $\hat{\delta}_i$ and E_δ is the expected value with respect to the distribution of the δ_i 's. Also, define $\Delta Y = \sum_i U_i \Delta \delta_i$, where $U_i = Y_{1i} - Y_{0i}$ and $\Delta \delta_i = \delta_{i10} - \delta_{i01}$; note that $Y^L = Y + \Delta Y$.

Suppose that a sample s is drawn from the cohort and that the exposure and mortality data are linked both manually and with CRL. We assume that the manual linking is accurate and gives the correct links. Letting $\pi_i = P(i \in s)$ denote the probability that subject i is a member of the sample, we can then estimate the bias of Y_L by

$$\widehat{\Delta Y} = \sum_{i \in s} \pi_i^{-1} U_i \Delta \delta_i.$$

We also estimate $E_\delta(Y)$ by

$$\hat{Y} = Y_L - \widehat{\Delta Y}.$$

3.2.1 The Sampling Scheme

In order to estimate the bias and construct unbiased estimators, we implement the following sampling scheme. Select a simple random sample (SRS) s of size n from N .

Let N_j be the number of people who enter S_j and $N_{j'}$ be the number of people who don't. Then, $N = N_j + N_{j'}$ and $n = n_j + n_{j'}$, where n_j is the number of people selected from the j -th stratum. It follows that

$$P(n_j = k) = \frac{\binom{N_j}{k} \binom{N_{j'}}{n-k}}{\binom{N}{n}}$$

($k = 0, 1, \dots, N_j$). If $n_j \leq c$ (a predetermined fixed number), choose $(c - n_j)$ from $(N_j - n_j)$, so that the effective sample size in j -th stratum $n_j^e = c$. If $n_j > c$, then $n_j^e = n_j$. Thus,

$$\begin{aligned} Pr(n_j^e = c) &= \sum_{k=0}^c \frac{\binom{N_j}{k} \binom{N-N_j}{n-k}}{\binom{N}{n}}, \\ Pr(n_j^e = l) &= \frac{\binom{N_j}{l} \binom{N-N_j}{n-l}}{\binom{N}{n}} \quad (l = c + 1, \dots, \min(n, N_j)), \text{ and} \\ Pr(n_j^e < c) &= 0 = Pr(n_j^e > N_j). \end{aligned}$$

At this point, we may make the following observation. When we select n_j units from j -th stratum, where $n_j \geq c$, then each of the N_j units were equally likely to be selected. Hence, s_j is a SRS of size n_j from c population of size N_j . But when $n_j < c$, then we select $c - n_j$ units from $N_j - n_j$ units using SRS.

A composition of two simple random samples, the first of n_1 drawn from the original population and the second of size n_2 drawn from the subpopulation of $(N - n_1)$ units not covered by the first sample, constitutes a simple random sample of size $n_1 + n_2$ from the original population (Hedayat and Sinha, 1991). So, the sample we get by sampling and resampling is also a SRS of size c from N_j . We can then say s_j is a SRS from N_j whether or not $n_j > c$.

Given n_j^e , the inclusion probability of i -th individual being sampled from the j -th stratum is

$$[\pi_{i(j)} | n_j^e] = \frac{n_j^e}{N_j}.$$

Thus, the unconditional probability of $i \in s_j$ is

$$\begin{aligned}
\bar{\pi}_{i(j)} &= E(\pi_{i(j)}) = E[E(\pi_{i(j)}|n_j^e)] \\
&= Pr(n_j^e > c)[\pi_{i(j)}|n_j^e > c] + Pr(n_j^e = c)[\pi_{i(j)}|n_j^e = c] \\
&= \sum_{n_j^e=c+1}^{\min(n, N_j)} \frac{n_j^e}{N_j} \frac{\binom{N_j}{n_j^e} \binom{N-N_j}{n-n_j^e}}{\binom{N}{n}} + \left(\frac{c}{N_j}\right) \sum_{k=0}^c \frac{\binom{N_j}{k} \binom{N-N_j}{n-k}}{\binom{N}{n}}.
\end{aligned} \tag{3.1}$$

Given n_j^e , we next find the conditional second order inclusion probability of i and i' ,

$(\pi_{ii'}(j)|n_j^e) = P[\text{both } i \text{ and } i' \text{ are in } s_j | n_j^e]$:

$$[\pi_{ii'}(j)|n_j^e] = \frac{n_j^e(n_j^e - 1)}{N_j(N_j - 1)}.$$

The unconditional probability of both i and i' being included in s_j is thus

$$\begin{aligned}
\bar{\pi}_{ii'}(j) &= E[\pi_{ii'}(j)] = E_{n_j^e}(\pi_{ii'}|n_j^e) \\
&= \sum_{n_j^e=c+1}^{\min(n, N_j)} \frac{n_j^e(n_j^e - 1)}{N_j(N_j - 1)} \frac{\binom{N_j}{n_j^e} \binom{N-N_j}{n-n_j^e}}{\binom{N}{n}} \\
&\quad + \left[\frac{c(c-1)}{N_j(N_j - 1)}\right] \left[\sum_{k=0}^c \frac{\binom{N_j}{k} \binom{N-N_j}{n-k}}{\binom{N}{n}}\right].
\end{aligned} \tag{3.2}$$

We now calculate $\pi_i(jj') = P[i \text{ is included in both } s_j \text{ and } s_{j'}]$. Let $N_{jj'} = |S_j \cup S_{j'}|$

be the number of units in the union of S_j and $S_{j'}$. Then,

$$\begin{aligned}
\pi_i(jj') &= \frac{\binom{N-1}{n-1}}{\binom{N}{n}} + \sum_{l=0}^{c-1} \sum_{l'=0}^{c-1} \sum_{u=0}^{l \wedge l'} \frac{\binom{N_{jj'}-1}{l+l'-u} \binom{N-N_{jj'}}{n-l-l'+u}}{\binom{N}{n}} \frac{\binom{N_j-l-1}{c-l-1}}{\binom{N_j-l}{c-l}} \frac{\binom{N_{j'}-l'-1}{c-l'-1}}{\binom{N_{j'}-l'}{c-l'}} \\
&= \frac{n}{N} + \sum_{l=0}^{c-1} \sum_{l'=0}^{c-1} \sum_{u=0}^{l \wedge l'} \frac{\binom{N_{jj'}-1}{l+l'-u} \binom{N-N_{jj'}}{n-l-l'+u}}{\binom{N}{n}} \frac{c-l}{N_j-l} \frac{c-l'}{N_{j'}-l'}.
\end{aligned} \tag{3.3}$$

Similarly, $\pi_{i(j) i'(j')_{i \neq i', j \neq j'}} = P[i \text{ is included in } s_j \text{ and } i' \text{ is included in } s_{j'} \text{ where } i \neq i', j \neq j']$ is calculated as follows:

$$\begin{aligned}
\pi_{i(j) i'(j')_{i \neq i', j \neq j'}} &= \frac{\binom{N-2}{n-2}}{\binom{N}{n}} + \frac{n}{N} \left[\sum_{l=0}^{c-1} \left(\frac{\binom{N_j-1}{l} \binom{N-N_j}{n-l}}{\binom{N}{n}} \frac{c-l}{N_j-l} + \frac{\binom{N_{j'}-1}{l} \binom{N-N_{j'}}{n-l}}{\binom{N}{n}} \frac{c-l}{N_{j'}-l} \right) \right] \\
&\quad + \sum_{l=0}^{c-1} \sum_{l'=0}^{c-1} \sum_{u=0}^{l \wedge l'} \frac{\binom{N_{jj'}-2}{l+l'-u} \binom{N-N_{jj'}}{n-l-l'+u}}{\binom{N}{n}} \frac{c-l}{N_j-l} \frac{c-l'}{N_{j'}-l'}.
\end{aligned} \tag{3.4}$$

3.3 Adjustment of the Observed Number of Death and Expected Number of Deaths

In order to apply the general theory given in Section 3.2 we redefine the δ_i 's for the state j as follows.

$$\delta_{i11}(j) = \begin{cases} 1 & \text{if subject } i \text{ has died and is labelled as having died in state } j \\ 0 & \text{otherwise,} \end{cases}$$

$$\delta_{i10}(j) = \begin{cases} 1 & \text{if subject } i \text{ is alive but is labelled as having died in state } j \\ 0 & \text{otherwise,} \end{cases}$$

$$\delta_{i01}(j) = \begin{cases} 1 & \text{if subject } i \text{ has died but is labelled as being alive in state } j \\ 0 & \text{otherwise, and} \end{cases}$$

$$\delta_{i00}(j) = \begin{cases} 1 & \text{if subject } i \text{ is alive and is labelled as being alive in state } j \\ 0 & \text{otherwise.} \end{cases}$$

From Eqn. 2.14 in Chapter 2, the observed number of deaths is given by

$$d_{jj}^L = d_{jj} + \Delta d_{jj}. \quad (3.5)$$

We define a two-phase difference estimate for the adjusted observed number of deaths as

$$\hat{d}_{jj} = d_{jj}^L - \Delta \hat{d}_{jj} = d_{jj} + (\Delta d_{jj} - \Delta \hat{d}_{jj}), \quad (3.6)$$

where

$$\Delta d_{jj} = \sum_{i=1}^{N_j} \Delta \delta_i(j) = \sum_{i=1}^{N_j} [\delta_{i10}(j) - \delta_{i01}(j)] \quad (3.7)$$

and

$$\Delta \hat{d}_{jj} = \sum_{i \in s_j} \frac{\delta_{i10}(j) - \delta_{i01}(j)}{\pi_i(j)} = \sum_{i=1}^{N_j} \frac{a_i(s_j)}{\bar{\pi}_i(j)} \Delta \delta_i(j), \quad (3.8)$$

where

$$a_i(s_j) = \begin{cases} 1 & \text{if } i \in s_j \\ 0 & \text{otherwise.} \end{cases} \quad (3.9)$$

Similarly, the expected number of deaths in each state with linkage error is

$$e_j^L = e_j - \Delta e_j. \quad (3.10)$$

We define the adjusted expected number of deaths as

$$\hat{e}_j = e_j^L + \Delta \hat{e}_j = e_j + (\Delta \hat{e}_j - \Delta e_j), \quad (3.11)$$

where, assuming that only death times are at the beginning of each cell's entry time,

$$\Delta e_j = \sum_{i=1}^{N_j} \left[\int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \right] \Delta \delta_i(j) \quad (3.12)$$

and

$$\begin{aligned} \Delta \hat{e}_j &= \sum_{i \in s_j} \left[\int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \right] \frac{\Delta \delta_i(j)}{\pi_i(j)} \\ &= \sum_{i=1}^{N_j} \frac{\Delta \delta_i(j) a_i(s_j)}{\bar{\pi}_i(j)} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \end{aligned} \quad (3.13)$$

3.3.1 Unbiasedness and Precision of Adjusted Counts

In order to show the unbiasedness of the adjusted observed number of deaths ($\hat{d}_{jj}$), the adjusted expected number of deaths $\hat{e}_j$, we first calculate the expectation of $a_i(s_j)$ (of Eqn. 3.9):

$$\begin{aligned} E[a_i(s_j)] &= 1 \times Pr(i \in s_j) + 0 \times Pr(i \notin s_j) \quad (\text{by definition of } a_i(s_j)) \\ &= \bar{\pi}_i(j) \end{aligned} \quad (3.14)$$

It follows that

$$E[\Delta \hat{d}_{jj}] = \sum_{i=1}^{N_j} \Delta \delta_i(j) E \left[\frac{a_i(s_j)}{\bar{\pi}_i(j)} \right] \quad (\text{from Eqn. 3.8})$$

$$\begin{aligned}
&= \sum_{i=1}^{N_j} \Delta \delta_i(j) \quad (\text{since } E[a_i(s_j)] = \bar{\pi}_i(j)) \\
&= \Delta d_{jj} \quad (\text{from Eqn. 3.7}).
\end{aligned}$$

From Eqn. 3.6, $E[\hat{d}_{jj}] = d_{jj}$. Similarly,

$$\begin{aligned}
E[\Delta \hat{e}_j] &= \sum_{i=1}^{N_j} \frac{\Delta \delta_i(j)}{\bar{\pi}_i(j)} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \cdot E[a_i(s_j)] \quad (\text{from Eqn. 3.13}) \\
&= \sum_{i=1}^{N_j} \Delta \delta_i(j) \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \cdot \frac{\bar{\pi}_i(j)}{\bar{\pi}_i(j)} \quad (\text{from Eqn. 3.14}) \\
&= \Delta e_j
\end{aligned}$$

and $E[\hat{e}_j] = e_j$.

We now calculate variance of the adjusted counts. Let

$$d_i(s_j) = \frac{a_i(s_j)}{\bar{\pi}_i(j)} \quad (3.15)$$

So that

$$E[d_i(s_j)] = \frac{E[a_i(s_j)]}{\bar{\pi}_i(j)} = \frac{\bar{\pi}_i(j)}{\bar{\pi}_i(j)} = 1. \quad (3.16)$$

Then,

$$E[d_i(s_j)d_{i'}(s_j)] = \frac{E[a_i(s_j)a_{i'}(s_j)]}{\bar{\pi}_i(j)\bar{\pi}_{i'}(j)}. \quad (3.17)$$

If $i \neq i'$, it follows from the definition of $a_i(s_j)$ that

$$a_i(s_j)a_{i'}(s_j) = \begin{cases} 1 & \text{if both } i \text{ and } i' \text{ are in } s_j \\ 0 & \text{otherwise} \end{cases}$$

So that

$$\begin{aligned}
E[a_i(s_j)a_{i'}(s_j)] &= 1 \times \text{Pr}(i \text{ and } i' \text{ are in } s_j) \\
&= \bar{\pi}_{ii'}(j) \quad (i \neq i').
\end{aligned} \quad (3.18)$$

Therefore,

$$E [d_i(s_j) d_{i'}(s_j)] = \frac{\bar{\pi}_{ii'}(j)}{\pi_i(j) \pi_{i'}(j)} \quad (i \neq i'). \quad (3.19)$$

But if $i = i'$ then

$$a_i(s_j) a_{i'}(s_j) = \begin{cases} 1 & \text{if } i \in s_j \\ 0 & \text{otherwise} \end{cases}$$

So that

$$E [a_i(s_j) a_{i'}(s_j)] = 1 \times Pr(i \in s_j) = \pi_i(j) \quad (i = i'). \quad (3.20)$$

Therefore,

$$E [d_i(s_j) d_{i'}(s_j)] = \frac{\pi_i(j)}{\pi_i(j) \pi_i(j)} = \frac{1}{\pi_i(j)} \quad (i = i'). \quad (3.21)$$

We can now write

$$E [d_i(s_j) d_{i'}(s_j)] = \begin{cases} \frac{1}{\pi_i(j)} & \text{if } i = i' \\ \frac{\bar{\pi}_{ii'}(j)}{\pi_i(j) \pi_{i'}(j)} & \text{otherwise} \end{cases} \quad (3.22)$$

It follows that

$$\begin{aligned} Cov [d_i(s_j), d_{i'}(s_j)] &= E [d_i(s_j) d_{i'}(s_j)] - E [d_i(s_j)] E [d_{i'}(s_j)] \\ &= E [d_i(s_j) d_{i'}(s_j)] - 1 \quad (\text{by Eqn. 3.16}) \\ &= \begin{cases} \frac{1}{\pi_i(j)} - 1 & \text{if } i = i' \text{ (by Eqn. 3.22)} \\ \frac{\bar{\pi}_{ii'}(j)}{\pi_i(j) \pi_{i'}(j)} - 1 & \text{otherwise} \end{cases} \\ &=: \begin{cases} d_i(j) & \text{if } i = i' \\ d_{ii'}(j) & \text{otherwise} \end{cases} \end{aligned} \quad (3.23)$$

Similarly, for $j \neq j'$,

$$E [d_i(s_j) d_{i'}(s_{j'})] = \frac{E [a_i(s_j) a_{i'}(s_{j'})]}{\pi_i(j) \pi_{i'}(j')}.$$

If $i \neq i'$, by definition of $a_i(s_j)$,

$$a_i(s_j) a_{i'}(s_{j'}) = \begin{cases} 1 & \text{if } i \in s_j \text{ and } i' \in s_{j'}, i \neq i', j \neq j' \\ 0 & \text{otherwise,} \end{cases}$$

So that

$$\begin{aligned} E[a_i(s_j)a_{i'}(s_{j'})] &= 1 \times \Pr(i \in s_j \text{ and } i' \in s_{j'}, i \neq i', j \neq j') \\ &= \pi_{i(j)i'(j')} \quad (i \neq i', j \neq j'). \end{aligned} \quad (3.24)$$

Then,

$$E[d_i(s_j)d_{i'}(s_{j'})] = \frac{\pi_{i(j)i'(j')}}{\pi_{i(j)}\pi_{i'(j')}} \quad (i \neq i'). \quad (3.25)$$

If $i = i'$, then

$$\begin{aligned} E[a_i(s_j)a_{i'}(s_{j'})] &= 1 \times \Pr(i \in s_j \text{ and } s_{j'}) \\ &= \pi_i(jj') \quad (i = i', j \neq j'). \end{aligned} \quad (3.26)$$

Then

$$E[d_i(s_j)d_i(s_{j'})] = \frac{\pi_i(jj')}{\pi_i(j)\pi_i(j')} \quad (i = i', j \neq j'). \quad (3.27)$$

Then, for $j \neq j'$, we can write

$$E[d_i(s_j)d_{i'}(s_{j'})] = \begin{cases} \frac{\pi_i(jj')}{\pi_i(j)\pi_i(j')} & \text{if } i = i' \\ \frac{\pi_{i(j)i'(j')}}{\pi_{i(j)}\pi_{i'(j')}} & \text{otherwise.} \end{cases} \quad (3.28)$$

We then have

$$\begin{aligned} \text{Cov}[d_i(s_j), d_{i'}(s_{j'})] &= E[d_i(s_j)d_{i'}(s_{j'})] - E[d_i(s_j)] \cdot E[d_{i'}(s_{j'})] \\ &= E[d_i(s_j)d_{i'}(s_{j'})] - 1 \quad (\text{by Eqn. 3.16}) \\ &= \begin{cases} \frac{\pi_i(jj')}{\pi_i(j)\pi_i(j')} - 1 & \text{if } i = i' \quad (\text{by Eqn. 3.28}) \\ \frac{\pi_{i(j)i'(j')}}{\pi_{i(j)}\pi_{i'(j')}} - 1 & \text{otherwise} \end{cases} \\ &= \begin{cases} d_i(jj') \quad (\text{say}) & \text{if } i = i' \\ d_{i(j)i'(j')} \quad (\text{say}) & \text{otherwise.} \end{cases} \end{aligned} \quad (3.29)$$

Now from Eqn. 3.6,

$$\hat{d}_{jj} - d_{jj} = \Delta d_{jj} - \Delta \hat{d}_{jj}$$

$$\begin{aligned}
&= \sum_{i=1}^{N_j} \left[1 - \frac{a_i(s_j)}{\bar{\pi}_i(j)} \right] \Delta \delta_i(j) \\
&= \sum_{i=1}^{N_j} [1 - d_i(s_j)] \Delta \delta_i(j). \tag{3.30}
\end{aligned}$$

Then,

$$\begin{aligned}
V[\hat{d}_{jj} - d_{jj}] &= V \left[\sum_{i=1}^{N_j} [1 - d_i(s_j)] \Delta \delta_i(j) \right] \\
&= E \left[\sum_{i=1}^{N_j} [1 - d_i(s_j)] \Delta \delta_i(j) \right]^2 \quad (\text{since } E \left[\sum_{i=1}^{N_j} [1 - d_i(s_j)] \Delta \delta_i(j) \right] = 0) \\
&= \sum_{i=1}^{N_j} \sum_{i'=1}^{N_j} \{E[1 - d_i(s_j)](1 - d_{i'}(s_j))\} \Delta \delta_i(j) \Delta \delta_{i'}(j) \\
&= \sum_{i=1}^{N_j} \sum_{i'=1}^{N_j} \{E[d_i(s_j)d_{i'}(s_j)] - 1\} \Delta \delta_i(j) \Delta \delta_{i'}(j) \quad (\text{since } E[d_i(s_j)] = 1) \\
&= \sum_{i=1}^{N_j} \sum_{i'=1}^{N_j} \text{Cov}[d_i(s_j)d_{i'}(s_j)] \Delta \delta_i(j) \Delta \delta_{i'}(j) \\
&= \sum_{i=1}^{N_j} \sum_{i'=1, i \neq i'}^{N_j} \left(\frac{\bar{\pi}_{ii'}(j)}{\bar{\pi}_i(j)\bar{\pi}_{i'}(j)} - 1 \right) \Delta \delta_i(j) \Delta \delta_{i'}(j) \\
&\quad + \sum_{i=1}^{N_j} \left(\frac{1}{\pi_i(j)} - 1 \right) (\Delta \delta_i(j))^2 \quad (\text{by Eqn. 3.23}) \\
&= \sum_{i=1}^{N_j} \sum_{i'=1, i \neq i'}^{N_j} d_{ii'}(j) \Delta \delta_i(j) \Delta \delta_{i'}(j) + \sum_{i=1}^{N_j} d_i(j) (\Delta \delta_i(j))^2. \tag{3.31}
\end{aligned}$$

An unbiased estimator of $V[\hat{d}_{jj} - d_{jj}]$ is thus given by:

$$\begin{aligned}
\hat{V}[\hat{d}_{jj} - d_{jj}] &= \sum_{i \in s_j} \sum_{i' \in s_j, i' \neq i} \frac{d_{ii'}(j) \Delta \delta_i(j) \Delta \delta_{i'}(j)}{\bar{\pi}_{ii'}(j)} + \sum_{i \in s_j} d_i(j) (\Delta \delta_i(j))^2 \\
&= \sum_{i=1}^{N_j} \sum_{i'=1, i' \neq i}^{N_j} d_{ii'}(s_j) \Delta \delta_i(j) \Delta \delta_{i'}(j) + \sum_{i=1}^{N_j} d_i(s_j) d_i(j) (\Delta \delta_i(j))^2, \tag{3.32}
\end{aligned}$$

where

$$d_{ii'}(s_j) = \frac{d_{ii'}(j)}{\bar{\pi}_{ii'}(j)} a_{ii'}(s_j)$$

and

$$a_{ii'}(s_j) = \begin{cases} 1 & \text{if both } i \text{ and } i' \text{ are in } s_j \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$\begin{aligned}
E[d_{ii'}(s_j)] &= \frac{d_{ii'}(j)}{\bar{\pi}_{ii'}(j)} E[a_{ii'}(s_j)] \\
&= \frac{d_{ii'}(j)}{\bar{\pi}_{ii'}(j)} \pi_{ii'}(j) = d_{ii'}(j).
\end{aligned} \tag{3.33}$$

Hence,

$$\begin{aligned}
E[\hat{V}[\hat{d}_{jj} - d_{jj}]] &= \sum_{i=1}^{N_j} \sum_{i'=1, i' \neq i}^{N_j} E[d_{ii'}(s_j) \Delta \delta_i(j) \Delta \delta_{i'}(j)] + \sum_{i=1}^{N_j} E[d_i(s_j) d_i(j) (\Delta \delta_i(j))^2] \\
&= \sum_{i=1}^{N_j} \sum_{i'=1, i' \neq i}^{N_j} d_{ii'}(j) \Delta \delta_i(j) \Delta \delta_{i'}(j) + \sum_{i=1}^{N_j} d_i(j) (\Delta \delta_i(j))^2 \\
&\quad (\text{since } E[d_{ii'}(s_j)] = d_{ii'}(j) \text{ and } E[d_i(s_j)] = 1) \\
&= V[\hat{d}_{jj} - d_{jj}].
\end{aligned} \tag{3.34}$$

Similarly, we have

$$\begin{aligned}
\hat{e}_j - e_j &= \Delta \hat{e}_j - \Delta e_j \\
&= \sum_{i=1}^{N_j} [d_i(s_j) - 1] \Delta \delta_i(j) \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du
\end{aligned}$$

by Eqn. 3.11. Then,

$$\begin{aligned}
V[\hat{e}_j - e_j] &= V \left[\sum_{i=1}^{N_j} [d_i(s_j) - 1] \Delta \delta_i(j) \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \right] \\
&= E \left[\sum_{i=1}^{N_j} [d_i(s_j) - 1] \Delta \delta_i(j) \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \right]^2 \\
&= \sum_{i=1}^{N_j} \sum_{i'=1}^{N_j} [E d_i(s_j) d_{i'}(s_j) - 1] \Delta \delta_i(j) \Delta \delta_{i'}(j) \cdot \left(\int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \right) \left(\int_{t_{i'j}^0}^{t_{i'j}^2} \lambda^*(u) du \right) \\
&= \sum_{i=1}^{N_j} \sum_{i'=1}^{N_j} Cov[d_i(s_j) d_{i'}(s_j)] \Delta \delta_i(j) \Delta \delta_{i'}(j) \cdot \left(\int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \right) \left(\int_{t_{i'j}^0}^{t_{i'j}^2} \lambda^*(u) du \right) \\
&= \sum_{i=1}^{N_j} \sum_{i'=1, i' \neq i}^{N_j} d_{ii'}(j) \Delta \delta_i(j) \Delta \delta_{i'}(j) \cdot \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) d(u) \cdot \int_{t_{i'j}^0}^{t_{i'j}^2} \lambda^*(u) d(u) \\
&\quad + \sum_{i=1}^{N_j} d_i(j) (\Delta \delta_i(j))^2 \cdot \left(\int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) d(u) \right)^2
\end{aligned} \tag{3.35}$$

by Eqn. 3.23. The unbiased estimator of $V[\hat{e}_j - e_j]$ is given by

$$\begin{aligned}\hat{V}_s[\hat{e}_j - e_j] &= \sum_{i=1}^{N_j} \sum_{i'=1, i' \neq i}^{N_j} d_{ii'}(s_j) \Delta \delta_i(j) \Delta \delta_{i'}(j) \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \cdot \int_{t_{i'j}^0}^{t_{i'j}^2} \lambda^*(u) du \\ &\quad + \sum_{i=1}^{N_j} d_i(s_j) d_i(j) (\Delta \delta_i(j))^2 \cdot \left(\int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \right)^2.\end{aligned}\quad (3.36)$$

The unbiasedness of $\hat{V}_s[\hat{e}_j - e_j]$ is thus proved by Eqns. 3.33 and 3.16.

For $j \neq j'$,

$$\begin{aligned}\text{Cov}[\hat{d}_{jj} - d_{jj}, \hat{d}_{jj'} - d_{jj'}] &= \text{Cov} \left[\sum_{i=1}^{N_j} [1 - d_i(s_j)] \Delta \delta_i(j), \sum_{i'=1}^{N_{j'}} [1 - d_{i'}(s_{j'})] \Delta \delta_{i'}(j') \right] \\ &= \sum_{i=1}^{N_j} \sum_{i'=1}^{N_{j'}} \Delta \delta_i(j) \Delta \delta_{i'}(j') \text{Cov}[d_i(s_j), d_{i'}(s_{j'})] \\ &\quad (\text{since } \text{Cov}[1 - d_i(s_j), 1 - d_{i'}(s_{j'})] = \text{Cov}[d_i(s_j), d_{i'}(s_{j'})]) \\ &= \sum_{i=1}^{N_j} \sum_{i'=1}^{N_{j'}} d_{i(j)i'(j')} \Delta \delta_i(j) \Delta \delta_{i'}(j') \\ &\quad + \sum_{i \in S_j \cap S_{j'}} d_i(jj') \Delta \delta_i(j) \Delta \delta_{i'}(j')\end{aligned}\quad (3.37)$$

by Eqn. 3.29. The unbiased estimator of $\text{Cov}[\hat{d}_{jj} - d_{jj}, \hat{d}_{jj'} - d_{jj'}]$ is thus given by

$$\begin{aligned}\widehat{\text{Cov}}[\hat{d}_{jj} - d_{jj}, \hat{d}_{jj'} - d_{jj'}] &= \sum_{i \in S_j} \sum_{i' \in S_{j'}, i' \neq i} \frac{d_{i(j)i'(j')}}{\bar{\pi}_{i(j)i'(j')}} \Delta \delta_i(j) \Delta \delta_{i'}(j') + \sum_{i \in S_j \cap S_{j'}} \frac{d_i(jj')}{\bar{\pi}_i(jj')} \Delta \delta_i(j) \Delta \delta_i(j') \\ &= \sum_{i=1}^{N_j} \sum_{i'=1}^{N_{j'}} d_{i(s_j)i'(s_{j'})} \Delta \delta_i(j) \Delta \delta_{i'}(j') + \sum_{i=1}^{N_j} d_i(s_j s_{j'}) \Delta \delta_i(j) \Delta \delta_i(j'),\end{aligned}\quad (3.38)$$

where

$$d_i(s_j s_{j'}) = \begin{cases} \frac{d_i(jj')}{\bar{\pi}_i(jj')} & \text{if } i \in s_j \text{ and } i' \in s_{j'} \\ 0 & \text{otherwise.} \end{cases}$$

In other words,

$$d_i(s_j s_{j'}) = \frac{d_i(jj')}{\bar{\pi}_i(jj')} a_i(s_j, s_{j'}),$$

where

$$a_i(s_j s_{j'}) = \begin{cases} 1 & \text{if } i \in s_j \text{ and } s_{j'} \\ 0 & \text{otherwise.} \end{cases}$$

It follows that

$$E[a_i(s_j s_{j'})] = \bar{\pi}_i(j j'),$$

and

$$E[d_i(s_j s_{j'})] = d_i(j j'). \quad (3.39)$$

Similarly,

$$\begin{aligned} d_{i(s_j) i'(s_{j'})} &= \begin{cases} \frac{d_{i(j) i'(j')}}{\bar{\pi}_{i(j) i'(j')}} & \text{if } i \in s_j \text{ and } i' \in s_{j'}, i \neq i', j \neq j' \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{d_{i(j) i'(j')}}{\bar{\pi}_{i(j) i'(j')}} a_{i(s_j) i'(s_{j'})}, \end{aligned}$$

where

$$a_{i(s_j) i'(s_{j'})} = \begin{cases} 1 & \text{if } i \in s_j \text{ and } i' \in s_{j'}, i \neq i', j \neq j' \\ 0 & \text{otherwise.} \end{cases}$$

Again, $E[a_{i(s_j) i'(s_{j'})}] = \bar{\pi}_{i(j) i'(j')}$, so that

$$E[d_{i(s_j) i'(s_{j'})}] = d_{i(j) i'(j')} \quad (3.40)$$

Thus,

$$E[\widehat{Cov}[\hat{d}_{jj} - d_{jj}, \hat{d}_{j'j'} - d_{j'j'}]] = Cov[\hat{d}_{jj} - d_{jj}, \hat{d}_{j'j'} - d_{j'j'}]. \quad (3.41)$$

Similarly,

$$Cov[\hat{e}_j - e_j, \hat{e}_{j'} - e_{j'}]$$

$$\begin{aligned} &= \sum_{i=1}^{N_j} \sum_{i'=1, i \neq i'}^{N_{j'}} d_{i(j) i'(j')} \Delta \delta_{i(j)} \Delta \delta_{i'(j')} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \cdot \int_{t_{i'j'}^0}^{t_{i'j'}^2} \lambda^*(u) du \\ &\quad + \sum_{i \in S_j \cap S_{j'}} d_{i(j j')} \Delta \delta_{i(j)} \Delta \delta_{i(j')} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \cdot \int_{t_{ij'}^0}^{t_{ij'}^2} \lambda^*(u) du, \end{aligned} \quad (3.42)$$

and the unbiased estimator of $Cov [\hat{e}_j - e_j, \hat{e}_{j'} - e_{j'}]$ is given by

$$\begin{aligned} \widehat{Cov} [\hat{e}_j - e_j, \hat{e}_{j'} - e_{j'}] &= \sum_{i=1}^{N_j} \sum_{i'=1, i' \neq i}^{N_j} d_{i(s_j)i'(s_{j'})} \Delta \delta_i(j) \Delta \delta_{i'}(j') \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \cdot \int_{t_{i'j'}^0}^{t_{i'j'}^2} \lambda^*(u) du \\ &\quad + \sum_{i=1}^{N_j} d_i(s_j s_{j'}) \Delta \delta_i(j) \Delta \delta_i(j') \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \cdot \int_{t_{ij'}^0}^{t_{ij'}^2} \lambda^*(u) du. \end{aligned} \quad (3.43)$$

Now, $Cov [\hat{d}_{jj} - d_{jj}, \hat{e}_j - e_j]$

$$\begin{aligned} &= Cov \left[\sum_{i=1}^{N_j} (1 - d_i(s_j)) \Delta \delta_i(j), \sum_{i'=1}^{N_j} (d_{i'}(s_j) - 1) \Delta \delta_{i'}(j) \int_{t_{i'j}^0}^{t_{i'j}^2} \lambda^*(u) du \right] \\ &= \sum_{i=1}^{N_j} \sum_{i'=1}^{N_j} \Delta \delta_i(j) \Delta \delta_{i'}(j) \int_{t_{i'j}^0}^{t_{i'j}^2} \lambda^*(u) du \{-Cov[d_i(s_j), d_{i'}(s_j)]\} \\ &= - \sum_{i=1}^{N_j} \sum_{i'=1, i' \neq i}^{N_j} d_{ii'}(j) \Delta \delta_i(j) \Delta \delta_{i'}(j) \int_{t_{i'j}^0}^{t_{i'j}^2} \lambda^*(u) du \\ &\quad - \sum_{i=1}^{N_j} d_i(j) (\Delta \delta_i(j))^2 \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \end{aligned} \quad (3.44)$$

from Eqn. 3.23. The unbiased estimator is given by

$$\begin{aligned} \widehat{Cov} [\hat{d}_{jj} - d_{jj}, \hat{e}_j - e_j] &= - \sum_{i=1}^{N_j} \sum_{i'=1}^{N_j} d_{ii'}(s_j) \Delta \delta_i \Delta \delta_{i'}(j) \int_{t_{i'j}^0}^{t_{i'j}^2} \lambda^*(u) du \\ &\quad - \sum_{i=1}^{N_j} d_i(s_j) d_i(j) (\Delta \delta_i(j))^2 \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du. \end{aligned} \quad (3.45)$$

When $j \neq j'$,

$$\begin{aligned} Cov [\hat{d}_{jj} - d_{jj}, \hat{e}_{j'} - e_{j'}] &= Cov \left[\sum_{i=1}^{N_j} (1 - d_i(s_j)) \Delta \delta_i(j), \sum_{i'=1}^{N_{j'}} (d_{i'}(s_{j'}) - 1) \Delta \delta_{i'}(j') \right] \\ &= \sum_{i=1}^{N_j} \sum_{i'=1}^{N_{j'}} \Delta \delta_i(j) \Delta \delta_{i'}(j') \int_{t_{i'j'}^0}^{t_{i'j'}^2} \lambda^*(u) du (-Cov[d_i(s_j), d_{i'}(s_{j'})]) \\ &= - \sum_{i=1}^{N_j} \sum_{i'=1, i' \neq i}^{N_{j'}} d_{i(j)i'(j')} \Delta \delta_i(j) \Delta \delta_{i'}(j') \int_{t_{i'j'}^0}^{t_{i'j'}^2} \lambda^*(u) du \\ &\quad - \sum_{i \in S_j \cap S_{j'}} d_i(jj') \Delta \delta_i(j) \Delta \delta_i(j') \int_{t_{ij'}^0}^{t_{ij'}^2} \lambda^*(u) du \end{aligned} \quad (3.46)$$

from Eqn. 3.29, and the unbiased estimator is

$$\begin{aligned} \widehat{Cov} [\hat{d}_{jj} - d_{jj}, \hat{e}_{j'} - e_{j'}] \\ = - \sum_{i=1}^{N_j} \sum_{i'=1, i' \neq i}^{N_j} d_{i(s_j)i'(s_{j'})} \Delta \delta_i(j) \Delta \delta_{i'}(j') \int_{t_{i'j'}^0}^{t_{i'j'}^2} \lambda^*(u) du \\ - \sum_{i=1}^{N_j} d_{i(s_j s_{j'})} \Delta \delta_i(j) \Delta \delta_{i'}(j') \int_{t_{ij'}^0}^{t_{ij'}^2} \lambda^*(u) du \end{aligned} \quad (3.47)$$

3.4 Adjusted SMRs

By definition of the standardized mortality ratio,

$$SMR = \frac{\Sigma d_{jj}}{\Sigma e_j}. \quad (3.48)$$

We propose the following adjusted estimate of SMR:

$$\widehat{SMR} = \frac{\Sigma \hat{d}_{jj}}{\Sigma \hat{e}_j} \quad (3.49)$$

By Taylor series expansion (i.e., $f(\hat{x}, \hat{y}) \cong f(x, y) + (\hat{x} - x)f_x(x, y) + (\hat{y} - y)f_y(x, y)$),

$$\widehat{SMR} = \frac{\Sigma \hat{d}_{jj}}{\Sigma \hat{e}_j} \cong \frac{\Sigma d_{jj}}{\Sigma e_j} + \frac{\Sigma \hat{d}_{jj} - \Sigma d_{jj}}{\Sigma e_j} - (\Sigma \hat{e}_j - \Sigma e_j) \frac{\Sigma d_{jj}}{(\Sigma e_j)^2}. \quad (3.50)$$

We have shown that

$$E[\hat{d}_{jj} - d_{jj}] = E[\hat{e}_j - e_j] = 0,$$

so that

$$E[\Sigma \hat{d}_{jj} - \Sigma d_{jj}] = E[\Sigma \hat{e}_j - \Sigma e_j] = 0.$$

Hence,

$$\begin{aligned} E[\widehat{SMR}] &= E \left[\frac{\Sigma \hat{d}_{jj}}{\Sigma \hat{e}_j} \right] \cong \left[\frac{\Sigma d_{jj}}{\Sigma e_j} + \frac{\Sigma \hat{d}_{jj} - \Sigma d_{jj}}{\Sigma e_j} - (\Sigma \hat{e}_j - \Sigma e_j) \frac{\Sigma d_{jj}}{(\Sigma e_j)^2} \right] \\ &= \frac{\Sigma d_{jj}}{\Sigma e_j} = SMR, \end{aligned} \quad (3.51)$$

which proves that $\widehat{SMR}$ is an approximately unbiased estimator of SMR.

The variance of the $\widehat{SMR}$ is given by

$$\begin{aligned}
V[\widehat{SMR}] &\cong \frac{V[\Sigma \hat{d}_{jj} - \Sigma d_{jj}]}{(\Sigma e_j)^2} + V\left(\Sigma \hat{e}_j - \Sigma e_j\right) \times \left[\frac{\Sigma d_{jj}}{(\Sigma e_j)^2}\right]^2 \\
&\quad - 2 \left(\frac{\Sigma d_{jj}}{(\Sigma e_j)^3}\right) Cov\left[\Sigma(\hat{d}_{jj} - d_{jj}), \Sigma(\hat{e}_j - e_j)\right] \\
&= \frac{1}{(\Sigma e_j)^2} \left[\sum_j V[\hat{d}_{jj} - d_{jj}] + \sum_{j \neq j'} Cov[\hat{d}_{jj} - d_{jj}, \hat{d}_{j'j'} - d_{j'j'}] \right] \\
&\quad + \left[\frac{\Sigma d_{jj}}{(\Sigma e_j)^2}\right]^2 \left[\sum_j V[\hat{e}_j - e_j] + \sum_{j \neq j'} Cov[\hat{e}_j - e_j, \hat{e}_{j'} - e_{j'}] \right] \\
&\quad - 2 \left(\frac{\Sigma d_{jj}}{(\Sigma e_j)^3}\right) \left[\sum_j Cov[\hat{d}_{jj} - d_{jj}, \hat{e}_j - e_j] + \sum_{j \neq j'} Cov[\hat{d}_{jj} - d_{jj}, \hat{e}_{j'} - e_{j'}] \right].
\end{aligned} \tag{3.52}$$

$$\tag{3.53}$$

The terms $V[\hat{d}_{jj} - d_{jj}]$, $Cov[\hat{d}_{jj} - d_{jj}, \hat{d}_{j'j'} - d_{j'j'}]$, $V[\hat{e}_j - e_j]$, $Cov[\hat{e}_j - e_j, \hat{e}_{j'} - e_{j'}]$, $Cov[\hat{d}_{jj} - d_{jj}, \hat{e}_j - e_j]$, and $Cov[\hat{d}_{jj} - d_{jj}, \hat{e}_{j'} - e_{j'}]$ have been calculated in Eqns. 3.31, 3.37, 3.35, 3.42, 3.44, 3.46, respectively. Using these results, we have

$$\begin{aligned}
V[\Sigma(\hat{d}_{jj} - d_{jj})] &= \sum_j V[\hat{d}_{jj} - d_{jj}] + \sum_{j \neq j'} Cov[\hat{d}_{jj} - d_{jj}, \hat{d}_{j'j'} - d_{j'j'}] \\
&= \sum_{j=1}^J \sum_{i=1}^{N_j} \sum_{i'=1}^{N_j} d_{ii'}(j) \Delta \delta_{i(j)} \Delta \delta_{i'(j)} \\
&\quad + \sum_{j=1, j \neq j'}^J \sum_{i=1}^{N_j} \sum_{i'=1, i' \neq i}^{N_{j'}} d_{i(j)i'(j')} \Delta \delta_{i(j)} \Delta \delta_{i'(j')} \\
&\quad + \sum_{j=1, j \neq j'}^J \sum_{i \in C_j \cap C_{j'}} d_{i(jj')} \Delta \delta_{i(j)} \Delta \delta_{i'(j)}.
\end{aligned} \tag{3.54}$$

Similarly,

$$\begin{aligned}
V(\Sigma(\hat{e}_j - e_j)) &= \sum_j V[\hat{e}_j - e_j] + \sum_{j \neq j'} Cov[\hat{e}_j - e_j, \hat{e}_{j'} - e_{j'}] \\
&= \sum_{j=1}^J \sum_{i=1}^{N_j} \sum_{i'=1}^{N_j} d_{ii'}(j) \Delta \delta_{i(j)} \Delta \delta_{i'(j)} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \int_{t_{i'j}^0}^{t_{i'j}^2} \lambda^*(u) du
\end{aligned}$$

$$\begin{aligned}
& + \sum_{j=1, j \neq j'}^J \sum_{i=1}^{N_j} \sum_{i'=1, i' \neq i}^{N_{j'}} d_{i(j)i'(j')} \Delta \delta_{i(j)} \Delta \delta_{i'(j')} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \int_{t_{i'j'}^0}^{t_{i'j'}^2} \lambda^*(u) du \\
& + \sum_{j \neq j'}^J \sum_{i \in C_j \cap C_{j'}} d_i(jj') \Delta \delta_{i(j)} \Delta \delta_{i'(j)} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \int_{t_{i'j'}^0}^{t_{i'j'}^2} \lambda^*(u) du, \tag{3.55}
\end{aligned}$$

and $Cov [\Sigma(\hat{d}_{jj} - d_{jj}), \Sigma(\hat{e}_j - e_j)]$

$$\begin{aligned}
& = \sum_j Cov [\hat{d}_{jj} - d_{jj}, \hat{e}_j - e_j] + \sum_{j \neq j'} Cov [\hat{d}_{jj} - d_{jj}, \hat{e}_{j'} - e_{j'}] \\
& = - \sum_{j=1}^J \sum_{i=1}^{N_j} \sum_{i'=1}^{N_{j'}} d_{ii'}(j) \Delta \delta_{i(j)} \Delta \delta_{i'(j)} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \\
& \quad - \sum_{j \neq j'} \sum_{i \in C_j \cap C_{j'}} d_i(jj') \Delta \delta_{i(j)} \Delta \delta_{i'(j')} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du \\
& \quad - \sum_{j \neq j'} \sum_{i=1}^{N_j} \sum_{i'=1, i' \neq i}^{N_{j'}} d_{i(j)i'(j')} \Delta \delta_{i(j)} \Delta \delta_{i'(j')} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u) du. \tag{3.56}
\end{aligned}$$

In order to estimate $V(\widehat{SMR})$, we use the unbiased estimators of $V[\hat{d}_{jj} - d_{jj}]$, $V[\hat{e}_j - e_j]$, $Cov[\hat{d}_{jj} - d_{jj}, \hat{d}_{j'j'} - d_{j'j'}]$, $Cov[\hat{e}_j - e_j, \hat{e}_{j'} - e_{j'}]$, $Cov[\hat{d}_{jj} - d_{jj}, \hat{e}_j - e_j]$, and $Cov[\hat{d}_{jj} - d_{jj}, \hat{e}_{j'} - e_{j'}]$ that were calculated in Eqns. 3.32, 3.36, 3.38, and 3.43.

3.5 Adjusted Relative Risk Regression Parameters

3.5.1 Accounting for Overdispersion

Since d_{jj} follows the Poisson distribution

$$d_{jj} \sim \text{Poisson}(\lambda_j(\beta)e_j), \tag{3.57}$$

we have

$$\begin{aligned}
E(\hat{d}_{jj}) & = E_{d_{jj}} [E[\hat{d}_{jj}|d_{jj}, e_j]] \\
& = E_{d_{jj}}(d_{jj}) \quad (\text{since } E[\hat{d}_{jj}] = d_{jj}) \\
& = \lambda_j(\beta)e_j \\
& =: \mu,
\end{aligned}$$

and

$$\begin{aligned} V(\hat{d}_{jj}) &= E_{d_{jj}} [V(\hat{d}_{jj}|d_{jj}, e_j)] + V_{d_{jj}} [E(\hat{d}_{jj}|d_{jj}, e_j)] \\ &= g(N_j) + V_{d_{jj}}(d_{jj}). \end{aligned}$$

By Eqn. 3.31, $V(\hat{d}_{jj}|d_{jj}, e_j)$ does not involve e_j and d_{jj} and only depends on N_j , the j -th state-size under linkage error. We thus have

$$\begin{aligned} V(\hat{d}_{jj}) &= g(N_j) + \lambda_j e_j \\ &=: V_j. \end{aligned}$$

Define

$$\begin{aligned} \hat{V}_j &= g(N_j) + \lambda_j(\beta) \hat{e}_j \\ &= g(N_j) + \lambda_j(\beta) (\hat{e}_j - e_j). \end{aligned} \tag{3.58}$$

Assume now that $\hat{d}_{jj}$ are independent. If V_j and e_j were known, the quasi-likelihood estimator $\tilde{\beta}$ would be the solution of the score-like equation.

$$S(\beta) = \sum_{j=1}^J \frac{\partial \mu_j}{\partial \beta} V_j^{-1} (\hat{d}_{jj} - \lambda_j(\beta) e_j) = 0, \tag{3.59}$$

which implies that

$$\sum_{j=1}^J \frac{\partial \lambda_j}{\partial \beta} e_j V_j^{-1} (\hat{d}_{jj} - \lambda_j(\beta) e_j) = 0, \tag{3.60}$$

Since $\frac{\partial \mu_j}{\partial \beta} = \frac{\partial \lambda_j}{\partial \beta} e_j$.

We now have

$$\frac{\partial S(\beta)}{\partial \beta} = \sum_{j=1}^J \frac{\partial}{\partial \beta} \left[\frac{\partial \lambda_j}{\partial \beta} e_j V_j^{-1} \right] [\hat{d}_{jj} - \lambda_j(\beta) e_j] - \sum_{j=1}^J \frac{\partial \lambda_j}{\partial \beta} e_j V_j^{-1} e_j \frac{\partial \lambda_j}{\partial \beta}, \tag{3.61}$$

so that

$$-E \left[\frac{\partial S(\beta)}{\partial \beta} \right] = \sum_{j=1}^J \frac{\partial \lambda_j}{\partial \beta} e_j V_j^{-1} e_j \frac{\partial \lambda_j}{\partial \beta} =: V_T. \tag{3.62}$$

The sequence of estimates generated by the Newton-Raphson method with Fisher scoring is

$$\begin{aligned}\beta_1 &= \beta_0 + \left[-E \frac{\partial}{\partial \beta} S(\beta) \right]_{\beta=\beta_0}^{-1} S(\beta) \\ &= \beta_0 + \left[\sum_{j=1}^J \frac{\partial \lambda_j}{\partial \beta} e_j V_j^{-1} e_j \frac{\partial \lambda_j}{\partial \beta} \right]_{\beta=\beta_0}^{-1} \left[\sum_{j=1}^J \frac{\partial \lambda_j}{\partial \beta} e_j V_j^{-1} (\hat{d}_{jj} - \lambda_j(\beta) e_j) \right]_{\beta=\beta_0} \quad (3.63)\end{aligned}$$

where β_0 is an arbitrary value. However, e_j and V_j are unknown quantities, which are being estimated by $\hat{e}_j$ and $\hat{V}_j$ respectively. Since $E[\hat{e}_j] = e_j$, we have

$$E[\hat{V}_j] = E[V_j + \lambda_j(\hat{e}_j - e_j)] = V_j. \quad (3.64)$$

Assuming a Taylor series expansion, we have

$$\hat{V}_j^{-1} \approx V_j^{-1} - \frac{1}{V_j^2} (\hat{V}_j - V_j) = V_j^{-1} - \frac{\lambda_j(\hat{e}_j - e_j)}{V_j^2}, \quad (3.65)$$

so that

$$E[\hat{V}_j^{-1}] \approx V_j^{-1}. \quad (3.66)$$

Hence, e_j and V_j^{-1} may be replaced by $\hat{e}_j$ and $\hat{V}_j^{-1}$ in the estimating algorithm (1) to arrive at the following algorithm.

We obtain the quasi-likelihood estimate $\tilde{\beta}$ by iterating the following sequence of estimates until convergence is obtained:

$$\beta_1 = \beta_0 + \left[\sum_{j=1}^J \frac{\partial \lambda_j}{\partial \beta} \hat{e}_j \hat{V}_j^{-1} \hat{e}_j \frac{\partial \lambda_j}{\partial \beta} \right]_{\beta=\beta_0}^{-1} \left[\sum_{j=1}^J \frac{\partial \lambda_j}{\partial \beta} \hat{e}_j \hat{V}_j^{-1} (\hat{d}_{jj} - \lambda_j(\beta) \hat{e}_j) \right]_{\beta=\beta_0}. \quad (3.67)$$

Similarly, by Taylor series expansion of $S(\tilde{\beta})$ around the value β , we get

$$\tilde{\beta} - \beta \cong -V_T^{-1} S(\beta) \quad (3.68)$$

with

$$V(\tilde{\beta} - \beta) \cong -V_T^{-1} V(S(\beta)) V_T^{-1} \quad (3.69)$$

The asymptotic variance of $\tilde{\beta}$ is given by

$$V_A(\tilde{\beta} - \beta) \cong -V_T^{-1} V(S(\beta))_{asymptotic} V_T^{-1}, \quad (3.70)$$

with

$$\begin{aligned} V(S(\beta))_{asymptotic} &\cong \sum_{j=1}^J \frac{\partial \lambda_j}{\partial \beta} e_j V_j^{-1} V_j V_j^{-1} e_j \frac{\partial \lambda_j}{\partial \beta} \\ &= \sum_{j=1}^J \frac{\partial \lambda_j}{\partial \beta} e_j V_j^{-1} e_j \frac{\partial \lambda_j}{\partial \beta} \\ &= V_T \end{aligned}$$

Here, under the usual limiting condition on eigenvalues of $-E \left[\frac{\partial S(\beta)}{\partial \beta} \right]$, the asymptotic covariance matrix of $\tilde{\beta}$ is

$$V(\tilde{\beta} - \beta)_{asymptotic} = V_T^{-1} V_T V_T^{-1} = V_T^{-1} \quad (3.71)$$

We suggest the following estimator of the covariance matrix:

$$\hat{V}_A(\tilde{\beta}) = V_T^{-1}|_{\beta=\tilde{\beta}}. \quad (3.72)$$

3.6 Real-Life Data Analysis using Generalized Method

3.6.1 Simulation Model

The data and the model used for analysis are described in detail in Sections 2.7.1 and 2.7.2, and the effects of record linkage were evaluated. In this section the methods described to reduce the bias of record linkage error are investigated by computer simulation. The same data and the same methods of stratifying the data are incorporated into the present analysis.

As in Chapter 2, we assume the existing data to be the correct data, and introduce false positive and false negatives. To identify linkage errors, a sample of 300 is taken

from the whole population. Since the population size is 6,140, almost 5% of the population is sampled. As the minimum size of the strata size was 1, we had to fix c to be equal to one. So, each stratum where no individual was sampled was resampled, and one individual was sampled from it. The adjusted observed number of death and the adjusted number of expected death for each strata are calculated using Eqns. 3.6 and 3.11. The adjusted estimate of SMR is calculated using Eqn. 3.49. Similarly, following the methods of Section 3.5, an adjusted estimate of β is calculated. In order to determine the effect of the sample size on reduction of bias, the same procedure is also repeated for an initial size of 600, which is almost 10% of the population, as well as a sample size of 1500, which is almost 25% of the population.

As in Chapter 2, both false positive and false negative error rates take values of 0%, 0.1%, 0.5%, 1%, 5%, 10%, leading to 36 combinations of these two error rates. For each such combination, the relative biases in the estimators of the SMR and the excess relative risk were estimated on the basis of 200 simulations. (Even though in Chapter 2 we did the simulation 500 times, in this chapter we reduced the number to 200, since our correction requires an enormous amount of computer time.)

The relative bias, the coefficient of variation, and the root mean square error were calculated for both SMR and ERR after adjustment for record linkage errors.

The relative bias (RB) is estimated by

$$RB = \frac{1}{200} \sum_{i=1}^{200} \left(\frac{\hat{\theta}_L^{(i)} - \hat{\theta}}{\hat{\theta}} \right),$$

where $\hat{\theta}$ denotes the estimate of the SMR (assuming no linkage error) and $\hat{\theta}_L^{(i)}$ is the estimate based on i th simulation run (including linkage errors). The coefficient of

variation (CV) is estimated by

$$CV = \sqrt{\frac{1}{200} \sum_{i=1}^{200} \left(\frac{\hat{\theta}_L^{(i)} - \bar{\hat{\theta}}}{\hat{\theta}} \right)^2}.$$

where $\bar{\hat{\theta}}$ is the mean of $\hat{\theta}_L^{(i)}$. The root mean square error (RMSE) is defined by

$$RMSE = \sqrt{RB^2 + CV^2}.$$

3.6.2 Simulation Results

In Chapter 2, we observed through the simulation that false positives and false negatives introduce bias into the estimates of SMR and ERR. For a fixed false negative error rate, an increase in the false positive rate resulted in an increase in the relative bias of the SMR. For a fixed false positive rate, an increase in the false negative error rate resulted in a decrease in the relative bias of the SMR. In Table 3.1, we show the relative bias in the estimates of SMR due to linkage errors, after adjusting for linkage errors using samples size of 300, 600 and 1500.

We observe that the adjusted estimates have lower relative bias. We also notice that the bias decreases with increasing sample size. However, when the relative bias was negligibly small (for example, in case of a false positive rate at 0.1% and false negative rate at 0.5%, the relative bias is 0.001%), the adjustment method did not work as intended. It produced estimates with higher bias, as we tried to correct an estimate which had very little error in the first place. When the relative bias was high, as is the case where the false positive rate was 5% or 10%, the adjustment method performs much better. We also notice in some of the cases (specially those involving lower error rates) that the direction of bias changed, even though the absolute relative bias decreased.

In Table 3.2, we observe that the coefficient of variation of the adjusted estimates is higher than that of the estimates of SMR with uncorrected linkage errors. The coefficient decreased with the increasing sample size, but was always higher than that of the case where the linkage error was present. As a result, the RMSE of the estimates after adjustment was less than the unadjusted estimates rate cases (see Table 3.3). The greatest reduction occurred with higher error rates. With lower error rates, some of the RMSEs were higher than those of the unadjusted estimates. In such cases, the reduction in bias was not great, but coefficient of variation was increased.

In Chapter 2, we observed that the relative bias of the estimates of the ERR was sensitive to the magnitude of the false positives and false negative error rates. The direction of bias was not clear as in case of the SMR. From Table 3.4, we see that the adjustment method is not effective in reducing the bias, and in fact introduces more bias at lower error rates. However, it works as intended in the cases where the false rates were high. When false positive rates are varied from 0% to 0.5% and false negative rates are varied from 0% to 1%, the method is not useful (in fact has the opposite effect). As the error rates were increased and the bias was higher, the method is able to reduce bias, even though in some cases the direction of the bias is changed.

From Table 3.5, we observe that the coefficient of variation is less in the adjusted cases. It also decreases with the increasing sample sizes. This results in a decreased RMSE of the estimates (see Table 3.6).

Table 3.1: Relative Bias of SMR (%) unadjusted,
adjusted with sample size 300 (in parenthesis),
adjusted with sample size 600 (*italic*)
and adjusted with sample size 1500 (**bold**)

False Non-link Rate	False Link Rate					
	0%	0.1%	0.5%	1%	5%	10%
0%	0.00	0.53	2.98	5.87	28.84	58.53
	(0.00)	(0.50)	(2.24)	(4.78)	(21.68)	(40.30)
	<i>0.00</i>	<i>0.25</i>	<i>2.03</i>	<i>3.70</i>	<i>16.67</i>	<i>29.28</i>
	0.00	0.12	1.17	2.19	8.15	13.30
.1%	-0.12	0.42	2.87	5.75	28.72	58.40
	(0.03)	(0.41)	(2.27)	(4.61)	(21.88)	(40.99)
	<i>-0.04</i>	<i>0.09</i>	<i>1.76</i>	<i>3.41</i>	<i>15.58</i>	<i>28.57</i>
	0.06	0.21	0.99	2.00	7.8	13.48
0.5%	-0.54	-0.001	2.45	5.33	28.30	57.97
	(0.22)	(0.43)	(2.67)	(4.92)	(22.60)	(39.63)
	<i>-0.09</i>	<i>0.32</i>	<i>1.60</i>	<i>3.94</i>	<i>16.37</i>	<i>29.95</i>
	0.008	0.21	0.94	2.24	7.55	14.19
1%	-1.02	-0.48	1.97	4.85	27.81	57.46
	(0.08)	(0.009)	(2.61)	(4.40)	(21.64)	(41.04)
	<i>0.009</i>	<i>0.006</i>	<i>1.35</i>	<i>3.46</i>	<i>15.81</i>	<i>31.24</i>
	0.19	0.21	1.25	2.01	7.71	14.32
5%	-5.13	-4.59	-2.15	0.73	23.6	53.16
	(2.56)	(1.93)	(2.72)	(0.56)	(22.24)	(41.48)
	<i>.31</i>	<i>0.96</i>	<i>1.95</i>	<i>0.32</i>	<i>16.04</i>	<i>31.21</i>
	0.18	-0.26	1.40	0.25	7.08	15.09
10%	-10.69	-10.15	-7.72	-4.86	17.91	47.34
	(0.11)	(1.93)	(5.66)	(5.26)	(23.07)	(43.15)
	<i>1.10</i>	<i>1.45</i>	<i>3.14</i>	<i>3.42</i>	<i>16.06</i>	<i>30.65</i>
	-0.03	-0.76	1.68	2.78	8.39	14.71

Table 3.2: Coefficient of Variation of SMR (%) unadjusted,
adjusted with sample size 300 (in parenthesis),
adjusted with sample size 600 (*italic*)
and adjusted with sample size 1500 (**bold**)

False Non-link Rate	False Link Rate					
	0%	0.1%	0.5%	1%	5%	10%
0%	0.00	0.88	2.02	2.84	6.71	9.11
	(0.00)	(0.91)	(3.18)	(3.84)	(9.27)	(14.49)
	<i>0.00</i>	<i>1.65</i>	<i>3.11</i>	<i>4.14</i>	<i>9.50</i>	<i>13.92</i>
	0.00	1.25	2.56	3.52	7.56	11.57
0.1%	0.39	0.95	2.10	2.85	6.73	9.08
	(2.11)	(0.95)	(3.96)	(4.72)	(9.30)	(14.81)
	<i>1.02</i>	<i>2.17</i>	<i>3.36</i>	<i>4.22</i>	<i>10.51</i>	<i>14.86</i>
	0.86	1.24	2.75	3.62	8.29	10.97
0.5%	0.84	1.20	2.19	3.02	6.76	9.11
	(4.48)	(3.59)	(4.99)	(6.57)	(10.68)	(15.18)
	<i>2.45</i>	<i>2.94</i>	<i>4.36</i>	<i>4.81</i>	<i>9.98</i>	<i>14.02</i>
	1.61	1.92	3.15	3.88	8.56	11.1
1%	1.14	1.42	2.36	3.15	6.84	9.21
	(5.30)	(3.93)	(6.59)	(5.41)	(10.80)	(16.56)
	<i>3.41</i>	<i>3.73</i>	<i>4.77</i>	<i>6.16</i>	<i>10.13</i>	<i>13.14</i>
	1.90	2.37	3.13	4.32	8.50	11.88
5%	2.29	2.49	3.14	3.82	7.06	9.46
	(14.24)	(13.09)	(12.02)	(14.23)	(14.71)	(19.06)
	<i>8.32</i>	<i>8.74</i>	<i>8.63</i>	<i>8.20</i>	<i>12.55</i>	<i>15.38</i>
	5.03	4.26	5.27	5.91	9.55	12.62
10%	3.59	3.68	4.20	4.64	7.78	10.11
	(17.16)	(12.79)	(20.34)	(18.13)	(20.27)	(24.68)
	<i>12.42</i>	<i>17.97</i>	<i>8.34</i>	<i>12.34</i>	<i>15.81</i>	<i>19.23</i>
	7.37	6.80	6.90	8.58	9.89	12.75

Table 3.3: Root Mean Square Error of SMR (%) unadjusted,
adjusted with sample size 300 (in parenthesis),
adjusted with sample size 600 (*italic*)
and adjusted with sample size 1500 (**bold**)

False Non-link Rate	False Link Rate					
	0%	0.1%	0.5%	1%	5%	10%
0%	0.00	1.03	3.60	6.52	29.61	59.23
	(0.00)	(1.04)	(3.89)	(6.14)	(23.58)	(42.82)
	<i>0.00</i>	<i>1.67</i>	<i>3.71</i>	<i>5.55</i>	<i>19.18</i>	<i>32.42</i>
	0.00	1.26	2.81	4.15	11.12	17.63
0.1%	0.41	1.04	3.56	6.42	29.5	59.10
	(2.11)	(1.03)	(4.57)	(6.60)	(23.77)	(43.58)
	<i>1.02</i>	<i>2.17</i>	<i>3.79</i>	<i>5.43</i>	<i>18.80</i>	<i>32.20</i>
	0.86	1.26	2.93	4.14	11.38	17.38
0.5%	1.00	1.20	3.28	6.13	29.09	58.68
	(4.49)	(3.62)	(5.66)	(8.21)	(25.00)	(42.44)
	<i>2.45</i>	<i>2.96</i>	<i>4.65</i>	<i>6.21</i>	<i>19.17</i>	<i>33.07</i>
	1.61	1.94	3.29	4.48	11.41	18.01
1%	1.53	1.5	3.07	5.79	28.64	58.20
	(5.31)	(3.93)	(7.09)	(6.98)	(24.21)	(44.26)
	<i>3.41</i>	<i>3.73</i>	<i>4.96</i>	<i>7.06</i>	<i>18.78</i>	<i>33.89</i>
	1.91	2.38	3.37	4.76	11.47	18.60
5%	5.62	5.22	3.80	3.90	24.63	53.99
	(14.47)	(13.23)	(12.32)	(5.43)	(26.67)	(45.65)
	<i>8.33</i>	<i>8.79</i>	<i>8.85</i>	<i>4.13</i>	<i>20.37</i>	<i>34.79</i>
	5.03	4.26	5.46	5.92	11.89	19.68
10%	11.28	10.80	8.79	6.72	19.53	48.40
	(17.16)	(12.94)	(21.11)	(18.88)	(30.71)	(50.43)
	<i>12.46</i>	<i>18.03</i>	<i>8.91</i>	<i>12.80</i>	<i>22.54</i>	<i>36.37</i>
	7.37	6.84	7.10	9.02	12.96	19.47

Table 3.4: Relative Bias of ERR (%) unadjusted,
adjusted with sample size 300 (in parenthesis),
adjusted with sample size 600 (*italic*)
and adjusted with sample size 1500 (**bold**)

False Non-link Rate	False Link Rate					
	0%	0.1%	0.5%	1%	5%	10%
0%	0.00	17.31	46.04	94.01	309.96	490.10
	(0.00)	(88.54)	(103.67)	(83.58)	(-98.59)	(88.95)
	<i>0.00</i>	<i>86.41</i>	<i>-73.59</i>	<i>78.49</i>	<i>-16.53</i>	<i>44.20</i>
	0.00	85.87	-69.75	60.02	-13.08	43.61
0.1%	-0.83	15.97	44.56	92.45	310.69	491.20
	(-91.90)	(88.10)	(-87.60)	(84.72)	(-54.51)	(113.67)
	<i>-88.98</i>	<i>87.17</i>	<i>-83.87</i>	<i>76.78</i>	<i>13.20</i>	<i>61.61</i>
	-82.67	86.79	-77.67	55.70	-12.49	32.06
0.5%	37.27	54.17	80.78	128.42	343.10	522.18
	(-92.92)	(-90.06)	(80.99)	(-74.02)	(-64.25)	(58.29)
	<i>-80.90</i>	<i>-85.98</i>	<i>71.85</i>	<i>-70.68</i>	<i>-20.04</i>	<i>50.62</i>
	-80.62	-82.68	63.60	-65.17	-15.00	43.07
1%	32.89	49.33	76.77	122.10	334.58	517.10
	(-83.38)	(-87.07)	(79.61)	(-100.19)	(-53.39)	(70.19)
	<i>-81.72</i>	<i>-82.35</i>	<i>71.42</i>	<i>-29.35</i>	<i>52.76</i>	<i>63.40</i>
	-76.28	-77.52	62.73	-7.58	-31.86	-42.61
5%	95.97	111.70	140.08	181.60	397.98	560.18
	(96.67)	(-59.37)	(-58.64)	(84.25)	(103.95)	(260.09)
	<i>-58.58</i>	<i>-44.93</i>	<i>18.76</i>	<i>58.10</i>	<i>49.13</i>	<i>121.83</i>
	-44.27	25.81	4.69	18.75	-33.09	53.76
10%	-53.24	-39.89	-4.78	35.25	279.34	458.50
	(108.88)	(57.81)	(183.43)	(30.79)	(192.80)	(214.55)
	<i>75.02</i>	<i>39.77</i>	<i>127.84</i>	<i>-17.63</i>	<i>95.00</i>	<i>190.75</i>
	2.74	38.28	44.92	2.78	-21.95	-7.13

Table 3.5: Coefficient of variation of ERR (%) unadjusted,
adjusted with sample size 300 (in parenthesis),
adjusted with sample size 600 (*italic*)
and adjusted with sample size 1500 (**bold**)

False Non-link Rate	False Link Rate					
	0%	0.1%	0.5%	1%	5%	10%
0%	0.00	252.93	592.76	791.30	1678.17	2098.09
	(0.00)	(42.70)	(124.78)	(152.04)	(321.37)	(458.68)
	<i>0.00</i>	<i>74.99</i>	<i>92.33</i>	<i>123.84</i>	<i>309.49</i>	<i>413.95</i>
	0.00	29.44	123.83	130.88	304.41	413.58
0.1%	184.57	307.16	618.03	814.48	1710.22	2117.22
	(42.83)	(82.45)	(116.86)	(139.59)	(287.43)	(113.67)
	<i>20.69</i>	<i>34.67</i>	<i>113.90</i>	<i>154.66</i>	<i>317.08</i>	<i>61.61</i>
	58.88	39.53	116.47	269.22	279.82	32.06
0.5%	318.74	413.70	674.40	897.66	1727.66	522.18
	(64.83)	(81.31)	(119.31)	(159.52)	(314.42)	(414.29)
	<i>118.23</i>	<i>60.20</i>	<i>169.58</i>	<i>246.55</i>	<i>323.91</i>	<i>417.31</i>
	62.37	111.50	235.79	438.66	302.37	468.11
1%	406.07	487.71	742.39	1195.36	1729.17	2143.88
	(79.14)	(93.75)	(181.93)	(901.37)	(293.71)	(511.50)
	<i>143.48</i>	<i>311.95</i>	<i>133.67</i>	<i>233.34</i>	<i>278.53</i>	<i>432.88</i>
	325.81	105.51	321.77	402.42	392.57	514.58
5%	888.15	916.88	1127.08	1599.11	1948.91	2193.54
	(1041.17)	(204.47)	(232.49)	(1408.32)	(582.48)	(981.22)
	<i>347.68</i>	<i>395.93</i>	<i>730.35</i>	<i>377.57</i>	<i>613.23</i>	<i>595.35</i>
	224.77	730.01	463.99	583.74	367.78	465.28
10%	1451.19	1457.02	1609.04	1678.17	2256.86	2471.34
	(926.65)	(1041.19)	(1163.27)	(321.37)	(922.86)	(765.20)
	<i>258.25</i>	<i>305.79</i>	<i>679.70</i>	<i>309.49</i>	<i>701.28</i>	<i>965.51</i>
	594.42	592.28	344.95	304.41	468.64	500.19

Table 3.6: Root Mean Square Error of ERR (%) unadjusted,
adjusted with sample size 300 (in parenthesis),
adjusted with sample size 600 (*italic*)
and adjusted with sample size 1500 (**bold**)

False Non-link Rate	False Link Rate					
	0%	0.1%	0.5%	1%	5%	10%
0%	0.00	253.52	594.55	796.87	1706.55	2154.57
	(0.00)	(98.30)	(162.23)	(173.50)	(336.15)	(467.23)
	<i>0.00</i>	<i>114.42</i>	<i>118.07</i>	<i>146.62</i>	<i>309.93</i>	<i>416.31</i>
	0.00	90.78	142.12	143.98	304.69	415.87
0.1%	184.57	307.57	619.63	819.71	1738.21	2173.46
	(101.39)	(120.66)	(146.05)	(163.28)	(292.55)	(429.70)
	<i>91.36</i>	<i>93.82</i>	<i>141.45</i>	<i>172.66</i>	<i>317.36</i>	<i>475.33</i>
	101.50	95.37	139.49	274.93	280.10	398.49
0.5%	320.92	417.24	679.22	878.89	1761.39	2173.46
	(113.30)	(121.34)	(144.20)	(151.20)	(320.92)	(429.70)
	<i>143.26</i>	<i>104.96</i>	<i>184.18</i>	<i>175.34</i>	<i>324.53</i>	<i>475.33</i>
	101.93	138.81	244.22	165.52	302.74	398.49
1%	407.40	490.20	746.35	905.93	1761.24	220.19
	(114.96)	(127.95)	(198.59)	(188.38)	(298.52)	(418.37)
	<i>165.12</i>	<i>322.64</i>	<i>151.55</i>	<i>248.29</i>	<i>283.48</i>	<i>420.37</i>
	247.84	130.92	327.82	438.73	393.86	470.09
5%	893.32	923.65	1135.75	1209.08	1989.13	2205.36
	(1045.65)	(212.91)	(239.77)	(905.30)	(591.69)	(516.29)
	<i>351.63</i>	<i>398.47</i>	<i>730.59</i>	<i>240.46</i>	<i>615.20</i>	<i>437.50</i>
	229.08	730.47	464.02	402.86	369.26	516.34
10%	1452.17	1457.57	1609.05	1599.50	2274.08	2263.94
	(933.03)	(1042.79)	(1177.65)	(1441.59)	(942.78)	(1015.10)
	<i>594.43</i>	<i>308.37</i>	<i>691.62</i>	<i>377.75</i>	<i>707.69</i>	<i>607.69</i>
	268.93	593.51	347.86	583.75	469.15	500.25

3.7 Conclusions

Record linkage is the method of bringing together two or more separately recorded pieces of information regarding the same individual. In general terms, it involves linking records of human exposure to environmental hazards with records on health status, often using computerized methods for matching individual records from different databases. However, this linkage process introduces false link and false non-links. In Chapter 2, we explored the effect of the false links and non-links on the observed number of deaths, on the expected number of cases, on estimates of the SMR, and on estimates of the ERR. It was established that false links and false non-links introduced bias in the estimates of the SMR and ERR.

In this chapter, a general method to reduce the bias of linkage error on the estimates of SMR and ERR was suggested. A simple random sample was drawn from the study cohort. It was matched manually, and the manual match was taken as the right link. Comparing that match with the computerized match, a correction was made to the observed death, expected death, and exposure level. The bias of the estimates of SMR and ERR based on those adjusted values were then calculated.

We observed that the method was able to reduce bias when the error rates were high. We also established that the method was sensitive to the sample size. In the case of SMR, the bias was reduced from 60% to almost 13% when the sample size was approximately 25%. When the bias was less than 10%, the larger sample size did not make much difference. When the error rates were smaller, the sample size did not seem to have any effect on the estimates of ERR; rather, it appeared to introduce bias. Again, in the case of estimates of ERR, our method performed well when the error rates were high and the bias was large.

When the bias was not very large or the error rates were not high, the method failed to perform well in the case of ERR. Although it gave encouraging results for the situations involving higher rates of error (especially false positives), it requires some improvement in case of smaller error rates. It is not practical to sample 25% of a large cohort due to resource constraints. However, when it is possible to sample a large fraction of the cohort, this method works quite well. In conclusion, we recommend the use of the suggested method when the error rate is high (i.e., over 0.5%).

Chapter 4

Evaluation of RCAL and SIMEX Methods for Correcting Measurement Errors in Cox Proportional Hazard Model

4.1 Introduction

Measurement error is quite common in epidemiological studies (Armstrong *et al.*, 1992). In its most general form, measurement error refers to imperfect information on the values of risk factors or other covariates used to predict health outcomes. Such errors can arise because the variables under study are often subjective, or as a result of biological variability and laboratory errors. Errors in historical variables can occur when respondents are asked to recall events that occurred in the distant past. In many cases, the true value of the variable of interest is simply impossible to measure accurately. Considerable research has been carried out on the development of methods to correct for such errors (Fuller, 1987, Carroll *et al.*, 1995). The spectrum of methods in the literature reflects different approaches to this problem, and different assumptions

about the relationship between the true and the observed exposure variables.

The proportional hazards regression model (Cox, 1972) is commonly used to model survival data as a function of relevant covariates. Typically, the covariates are measured with error and are not directly observable. A naive approach to statistical inference is to use the observed values of the covariates in the Cox model, which usually produces biased estimates. An alternative strategy is to take into account error in covariate measurement when fitting linear and nonlinear regression models (cf. Stefanski (1985), Stefanski and Carroll (1985), Fuller (1987), Carroll (1989), Carroll and Stefanski (1990), Carroll *et al.* (1995)). Different methods have been proposed for estimating the parameters in the Cox model when the covariates are measured with error (Prentice, 1982; Pepe *et al.*, 1989; Nakamura, 1992; Carroll *et al.*, 1995; Hu *et al.*, 1998; Fung and Krewski, 1999a, 1999b; Huang and Wang, 2000).

Recently, two computer-based simulation methods (RCAL and SIMEX) have appeared in the literature for correcting impact of measurement errors. The RCAL (regression calibration) method was originally suggested by Carroll and Stefanski (1990). The SIMEX (simulation extrapolation) method was proposed by Cook and Stefanski (1994) and later revised by Stefanski and Cook (1995) and Carroll *et al.* (1996).

In this chapter, we will evaluate the effectiveness of RCAL and SIMEX for correcting measurement errors in Cox proportional hazard model, using real data from the Harvard Six Cities Study (Dockery *et al.*, 1993) that focuses on the effect of chronic exposure to airborne pollutants. The Cox proportional hazard model was used in the Harvard Six Cities Study to estimate the mortality rate ratios for exposure to different air pollutants in individuals, with and without other risk factors as covariates.

The Harvard Six Cities Study was the first large-scale cohort study to demonstrate

an association between long-term exposure to fine particulate matter less than 2.5 microns in aerodynamic diameter (PM_{2.5}) and mortality in urban centres in the United States. Because of the pivotal role of this study in the establishment of the first U.S. national ambient air quality objective for PM_{2.5} in 1997 (Greenbaum *et al.*, 2001), the results of this study were subjected to an independent detailed re-analysis to test the robustness of the findings to alternative analytic methods (Krewski *et al.*, 2000, 2003), including an assessment of the effect of exposure measurement error on estimates of risk based on the Cox proportional hazards model. It is well-known that random measurement error leads to downward bias in estimates of risk, and overstatement of the precision of such estimates.

Data from the Harvard Six Cities Study were used to evaluate the potential impact of measurement error on estimates of risk. After introducing a known amount of measurement error into the original data, estimates of risk were calculated using two methods for adjusting for measurement error: RCAL and SIMEX. With RCAL, the observed value of PM_{2.5} is replaced by its expected value with respect to the measurement error distribution. SIMEX adjusts for measurement error by adding progressively larger errors to the data and then extrapolating back to the case of no measurement error. Computer simulation was used to evaluate the accuracy and precision of both RCAL and SIMEX, and to assess the robustness of RCAL to mis-specification of the measurement error distribution.

4.2 Measurement Error Correction Methods

Let Y be a dichotomous response variable that is a function of two predictors, Z and X , in which Z is assumed to be observed without error, and the value of X is not

known with certainty due to the presence of measurement error. Independent replicate observations $W_1, W_2, \dots, W_n$ ($n \geq 2$) of the value of X are assumed to be available, thereby permitting an assessment of the measurement error distribution.

A fundamental prerequisite for analyzing measurement error is the specification of an underlying model for the measurement error model. Carroll *et al.* (1995) group the error models into two general categories: classical error models and regression calibration models. In classical measurement error models, the conditional distribution of W given (Z, X) is modelled, whereas in regression calibration models, the conditional distribution of X given (Z, W) is modelled.

4.2.1 Quasi-likelihood Estimation

Recently, a set of computer programs (QVF) have been developed by Schmiediche *et al.* (1997) for quasi-likelihood estimation in regression models with measurement error in the predictors. With quasi-likelihood estimation, we assume that the mean and the variance function of the response variate Y are known, although the full distribution of Y does not need to be specified. A general regression model specifies the relationship between the expected value of Y and the predictors (X, Z)

$$E(Y|Z, X) = f(Z, X; \beta),$$

where β is the regression parameter vector. The variance of Y is given by

$$Var(Y|Z, X) = \sigma^2 g^2(Z, X; \beta, \theta),$$

where $g(\cdot)$ is the variance function with parameter θ and σ^2 is the scale parameter. This general class of models includes linear and nonlinear regression, as well as generalized linear models such as the logistic, Poisson and gamma regression models.

The QVF routine computes estimates of the parameters of these models assuming that there is no measurement error in the predictors. In other words, the variables measured with and without error are treated identically. In the presence of measurement error, the QVF provides naive or unadjusted estimates of the unknown regression model parameters. The algorithm is based on the conditional mean and variance of Y as functions of the data X and Z . The conditionally unbiased estimating function for β is given by the sum of the terms

$$\frac{Y - f(Z, X; \beta)}{\sigma^2 g^2(Z, X; \beta, \theta)} \cdot \frac{\partial}{\partial \beta} f(Z, X; \beta)$$

over the entire set of data. This set of quasi-likelihood equations is set to zero and can be solved by weighted least squares or other nonlinear regression algorithms. Estimates of θ may be based on true replicates, if they exist, or on functions of squared residuals.

4.2.2 Regression Calibration Algorithm

In regression calibration, X is replaced by the regression of X on (Z, W) , and then a standard analysis is performed. RCAL, the usual correction for attenuation in linear regression, is consistent for the slope in models with mean that is an exponential function of a linear combination of the predictors and is very nearly consistent for slopes in most applications for generalized linear models. The RCAL estimates are usually easy to compute, have straightforward standard error estimates (Carroll *et al.*, 1995), and are amenable to bootstrap analysis.

The RCAL algorithm can be described in the following steps:

1. Using replication or other validation data, estimate the regression of X on (Z, W) , $m(Z, W, \gamma)$, depending on parameters γ which are estimated by $\hat{\gamma}$.

2. Replace the unobserved X by its estimate $m(Z, W, \hat{\gamma})$, and then run a standard analysis (QVF) to obtain parameter estimates.
3. Adjust the resulting standard errors to account for the estimation of β , using either bootstrap or asymptotic methods.

For example, suppose that the mean of Y given (X, Z) can be described by $f(Z, X, \beta)$ for some unknown parameter β . The replacement of X by its estimated value in effect produces a modified model for the observed data, namely

$$E(Y|Z, W) \approx f\{Z, m(Z, W, \beta), \beta\}.$$

The regression calibration model is thus an approximate working model for the observed data. The simplest form of RCAL is the “correction for attenuation” used in linear regression. It can be described as follows:

- (a) X is a scalar,
- (b) the measurement error is additive ($W = X + U$) with error variance σ_u^2 estimated by $\hat{\sigma}_u^2$,
- (c) the covariates (X, Z, W) are jointly normally distributed, and
- (d) as in logistic regression and generalized linear models, the response is affected only by a linear combination of predictors, namely $\beta_0 + \beta_x X + \beta_z^T Z$.

The RCAL estimator estimating the effect of X , namely β_x , is obtained as follows. Let $\hat{\beta}_x(\text{naive})$ be the naive estimator formed by ignoring the measurement error and let $\hat{\sigma}_{w|z}^2$ be the regression mean squared error from a linear regression of W on Z . Then, the RCAL estimator is $\hat{\beta}_x(\text{naive})\hat{\sigma}_{w|z}^2/(\hat{\sigma}_{w|z}^2 - \sigma_u^2)$.

Schafer (1992) proposed an adaptation of the RCAL method for the multiplicative error model $W = XU$. If we let $R = (1, Z^t, W)^t$, then for the linear model he noted that

$$E(RX) = \begin{pmatrix} 1/E(U) \\ E(ZW)/E(U) \\ E(W^2)E(U)/E(U^2) \end{pmatrix}.$$

If one knows or has estimates of the first two moments of U , then one estimates $E(RX)$ by replacing $E(ZW)$ and $E(W^2)$ by their sample averages. One potential difficulty with the multiplicative model is that the efficiency will likely deteriorate when the moments of U are estimated (Carroll *et al.*, 1996).

RCAL is a simple and generally applicable approximate estimation method that is especially well suited to problems in which validation or replication data are available for modelling the calibration function $E(X|W)$.

4.2.3 Simulation Extrapolation

An alternative to RCAL is simulation extrapolation, which relies on computer simulation to generate parameter estimates. SIMEX is a simulation-based method of estimating and reducing bias due to measurement error. This technique was proposed by Cook and Stefanski (1994) and further developed by Carroll *et al.* (1996) and Stefanski and Cook (1995). An integral component of SIMEX is a self-contained simulation study resulting in graphical displays that illustrate the effect of measurement error on parameter estimates and the need for bias correction.

Consider a linear regression model $Y = \beta_1 + \beta_x X + \epsilon$ with additive measurement error $W = X + U$, where U is independent of (Y, X) and has mean zero and variance σ_u^2 . The ordinary least squares estimate of β_x , denoted by $\hat{\beta}_{x,\text{naive}}$, consistently estimates $\beta_x \sigma_x^2 / (\sigma_x^2 + \sigma_u^2)$, but not β_x .

Using SIMEX, we can determine through simulation the effect of measurement error on an estimator. Suppose that in addition to the original data used to calculate $\hat{\beta}_{x,\text{naive}}$, there are $M - 1$ additional data sets available, each with successively larger measurement error variances, say $(1 + \lambda_m)\sigma_u^2$, where $0 = \lambda_1 < \lambda_2 < \dots < \lambda_M$. The least squares estimate of slope from the m th data set, $\hat{\beta}_{x,m}$, consistently estimates $\beta_x \sigma_x^2 / (\sigma_x^2 + (1 + \lambda_m)\sigma_u^2)$. We can think of this problem as a nonlinear regression model with dependent variable $\hat{\beta}_{x,m}$ and independent variable λ_m , having a mean function of the form

$$g(\lambda) = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + (1 + \lambda)\sigma_u^2},$$

where $\lambda \geq 0$. The parameter of interest, β_x , is obtained from $g(\lambda)$ by extrapolating to $\lambda = -1$.

The SIMEX procedure involves the following steps.

- Successive additional independent measurement errors are added to the original data, thereby creating data sets with successively larger measurement error variances.
- Naive estimates of the parameters are obtained from each of these data sets by the ordinary Cox regression method.
- The first two steps above are repeated a large number of times and the mean of the estimates is calculated for each level of contamination by the additional errors.
- A quadratic function is then fit to these averages using regression and the SIMEX estimate is obtained by extrapolating back to the case of no measurement error.

SIMEX is widely applicable in a number of situations and assumes only that one has a measurement error structure to which additional error can be added. A disadvantage is that the procedure is computationally intensive.

4.3 Measurement Error Adjustment in the Cox Proportional Hazard Model

4.3.1 Harvard Six Cities Study

The original Harvard Six Cities Study (Dockery *et al.*, 1993) is a cohort study that focuses on the effects of long-term exposure to airborne pollutants. In this study, persons were selected randomly from six U.S cities (Watertown, Massachusetts; Harriman/Kingston, Tennessee; St. Louis, Missouri; Steubenville, Ohio; Portage, Wisconsin; Topeka, Kansas) that had a wide range of levels of ambient particles and gaseous pollutants. This study began in 1974 and ended in 1991. The main focus of the investigation was to determine whether mortality rates are higher in cities with higher levels of air pollution. The study sample was restricted to a sub-cohort of 8,111 Caucasian subjects who were 25-74 years of age at the time of enrollment. Information on age, sex, weight, height, education level, smoking history, occupational exposure, and medical history were obtained from a series of questionnaires administered at baseline and at subsequent intervals (3, 6, and 12 years after enrollment). Levels of ambient air pollution were measured to estimate concentrations of particulate and gaseous air pollutants, using centrally located monitors in each city. The exposure estimates for persons from each of the six cities were based on mean pollution levels of each pollutant from centrally sited monitors, calculated for periods that were consistent and comparable among the six cities.

The statistical analysis of all-cause mortality and cause-specific mortality was based on the Cox proportional hazards risk models. The results of the study were summarized in terms of mortality ratios, with or without adjustments for other risk factors (e.g., smoking status, sex, age, education, and body mass index). All-cause mortality was found to be associated with inhalable particles, fine particles, and sulfate particles. The excess relative risk was about 26% comparing the city with the highest level of particles (Steubenville) with the lowest (Portage). Cardiopulmonary mortality, but not lung cancer mortality, was found to be associated with the level of air pollution. The excess relative risks were largely invariant with respect to smoking status, and occupational exposure to dusts, gases, or fumes. Similar results were found in subsequent analysis of another large cohort assembled by the American Cancer Society conducted by Pope *et al.* (1995,2002), although the most recent follow-up of this cohort revealed an association between PM_{2.5} and lung cancer mortality. Moreover, positive associations with mortality were also found with sulfur dioxide and acid aerosols, but not with ozone. The study estimated that annual average all-cause mortality increased in association with an increase in fine particles (less than 2.5 μm in median aerodynamic diameter (PM_{2.5})).

This study came under intense scrutiny when the Environmental Protection Agency used the results to support new National Ambient Air Quality Standards for fine particles (Greenbaum *et al.*, 2001) and to maintain the standards for particles less than 10 μm (PM₁₀) already in effect. To address this concern, an independent reanalysis was conducted by Krewski *et al.* (2000) on behalf of the Health Effects Institute.

The reanalysis included a set of new variables, all the gaseous copollutants, and the first attempts to apply spatial analytical methods to test the validity of the data

and the results from the Six Cities Study. The reanalysis considered alternative Cox proportional hazards of different specifications for the covariates, as well as covariates not considered by the original investigators. Furthermore, mortality from other causes, including respiratory diseases, cardiovascular diseases, and many forms of cancers other than lung were considered. The reanalysis confirmed the quality of the original data, replicated the original results, and tested those results against alternative risk models and analytic approaches without substantially altering the original findings of an association between indicators of particulate matter air pollution and mortality.

In this chapter, the magnitude of possible measurement error in the exposure variable, PM_{2.5}, will be assessed and the effect on the conclusions drawn will be investigated. In addition, we will attempt to adjust for measurement errors using RCAL and SIMEX method.

4.3.2 Error Distributions

In the Harvard Six Cities Study, individual exposure to air pollution was gauged using PM_{2.5} concentrations obtained from the single fixed site monitor located within each of the six cities. Since pollution levels vary somewhat within urban areas (Jerrett *et al.*, 2002), the use of a central monitor will lead to exposure measurement error. Pollution levels also vary temporally as well as spatially, demonstrating considerable daily as well as seasonal variation (Lin *et al.*, 2002). However, since mortality is dependent on long-term exposure to air pollution, we will focus on the average PM_{2.5} concentrations in the six cities over the nearly 20 year period of follow-up (Villeneuve *et al.*, 2002).

4.3.2.1 The Measurement Error Model

Let y_{it} be the personal exposure to a pollutant of person i on day t , and let the ambient pollution level from a centrally located monitor in that city on day t be x_t . If ambient pollution levels are used instead of personal exposures,

$$y_{it} = x_t + (y_{it} - x_t) = x_t + e_{it},$$

where the measurement error incurred is given by $e_{it} = y_{it} - x_t$. Averaging over the number of days, T , in the period of interest, the measurement error associated with long-term average exposure is

$$\bar{e}_i = \bar{y}_i - \bar{x},$$

where $\bar{y}_i = \sum_{t=1}^T y_{it}/T$ and $\bar{x} = \sum_{t=1}^T x_t/T$. The variance of $\bar{e}_i$ then reflects the variance of the measurement error distribution. Assuming individuals tend to spend most of their time in the same part of the city, spatial variation in long-term average air pollution levels provides an indication of the variance of the measurement error distribution.

4.3.2.2 Error Distribution of PM2.5

There are many sources of errors in measuring the exposure of air pollutant of individuals, including ambient vs personal exposure, variation among monitors placed in different location of the city, and mobility of individuals. A potentially important source of exposure measurement error in the Six Cities Study is that data from fixed-site monitors rather than personal dosimeters necessarily were used to evaluate individual exposures to ambient air pollutants. We will, therefore, consider averaged errors of using fixed-site monitor exposure levels instead of levels of individuals who are located in different parts of the city.

To see how measurements of PM2.5 vary in different parts of the city, we used the database at the web site (URL: <http://capita.wustl.edu/datawarehouse/Datasets/CAP-ITA/NAMPM25/Data/NAPM25.html>) for cities that have more than one monitor measuring PM2.5 concurrently. The cities we considered were Philadelphia, Mingo Junction, Detroit, New York, Winston-Salem, Vancouver and Windsor. These cities have long history of recordings that is particularly useful, since the Six Cities Study uses the mean pollution level of PM2.5 over an extended period of time. Only data recorded on the same days from all monitors in these city were included in the analysis. To measure variability, the standard deviation of the mean PM2.5 over the duration of time among monitors within each city is calculated as shown in Table 4.1.

Table 4.1: Mean and Standard Deviation of PM2.5 Measurements in Selected North American Cities, Averaged Over Monitoring Periods of Varying Duration ($\mu\text{g}/\text{m}^3$).

City	No. of Monitoring Sites	No. of Days of Monitoring/ Duration of Monitoring	Standard Deviation	Mean
Philadelphia	15	15 days over a month	5.33	22.7
Mingo Junction	3	35 days over a year	4.05	21.7
Detroit	2	164 days over 5 years	2.75	18.3
New York	2	363 days over 8 years	8.54	29.6
Winston-Salem	2	146 days over 4 years	0.12	21.4
Vancouver	2	164 days over 7 years	0.11	9.2
Windsor	2	198 days over 5 years	0.13	14.8

From Table 4.1, we see that the standard deviations vary from 0.1 to 8.5. Closer examination reveals that the standard error depends on where the monitors are located (close or far apart), and not just on how many days we are averaging. Based on the data set considered here, we will assume that the error distribution for PM2.5 is approximately normal, with standard deviation ranging from 0.1 to 8.0.

4.3.3 Cox Proportional Hazard Model

In the original Harvard Six Cities Study, the Cox proportional hazard model (Cox, 1972) was used to estimate the mortality rate ratios for exposure to different air pollutants in individuals, with and without including other risk factors as covariates. The proportional hazard model assumes that the hazard rate is the product of a baseline hazard function $\lambda_0(t)$ and a scalar which depends on the regression coefficients and the exposure variables and/or covariates. The model is given by

$$\lambda(t; \tilde{x}) = \lambda_0(t) \exp(\tilde{\beta}' \tilde{x}),$$

where $\tilde{\beta} = (\beta_1, \dots, \beta_p)'$ is the vector of regression coefficients.

4.3.4 Application of RCAL to Six Cities Study

4.3.4.1 Method

The earliest application of regression calibration method in survival analysis was discussed by Prentice (1982). Others have also proposed modifications to the original idea (cf. Pepe *et al.*, 1989, Clayton, 1991, Nakamura, 1982 and Hughes, 1993). The basis of this method is to replace the covariate of interest X by the regression of X on W in the standard Cox regression analysis. This method provides a first-order correction to

the bias in the estimated regression coefficients in the Cox model, but is known to be inconsistent.

Let us denote the true exposure history of an individual at age t by $x(t)$, and let $W(t)$ represent inaccurate or incomplete measure(s) of that exposure history. The Cox proportional hazard regression model representation of disease risk and true exposure may be expressed as a relative risk model,

$$\lambda[t|x(t)] = \lambda_0(t)\theta[x(t); \beta]. \quad (4.1)$$

It is necessary to assume that, given the true exposure history $x(t)$, the inaccurate measurements $W(t)$ are uninformative concerning disease outcome, which is given by

$$\lambda[t|x(t), W(t)] = \lambda[t|x(t)]. \quad (4.2)$$

With these assumptions (Prentice, 1982), the induced relationship between disease risk and measured exposure, $W(t)$, may be written as

$$\lambda[t|W(t)] = \lambda_0(t)\bar{\theta}[W(t); \beta], \quad (4.3)$$

where

$$\bar{\theta}[W(t); \beta] = E\{\theta[x(t); \beta] \mid W(t), T \geq t\}. \quad (4.4)$$

The notation $T \geq t$ in the conditional expectation means that the expectation is conditional on the individual not having succumbed to the disease under study before attaining age t . Formally, T represents the age of onset of the disease of interest. This condition enters into the expectation because of the definition of an age-specific incidence rate - the probability per unit time of first occurrence of the disease in individuals free of disease until a given age. Real cohort studies differ from this ideal

in that individuals enter the study at some age, say E , that is in general greater than zero. Thus, had they already succumbed to disease, they would not have been included in the study cohort. This is termed “truncation”, and its implication is that real epidemiologic studies measure incidence rates in survivor populations. Likewise, in real studies, there are many reasons for termination of follow-up other than occurrence of the single disease under study. These are collectively termed “censoring” and have the implication that real studies measure incidence rates in individuals whose follow-up has not been censored. Again, we must assume that this will not affect our conclusions.

Writing C for the age at which follow-up would be censored were disease occurrence not to intervene, we can express these assumptions more formally as follows

$$\lambda[t|x(t), T > E, C > t] = \lambda[t|x(t)]. \quad (4.5)$$

This equation merely states that the incidence rates we can observe in real studies are the same as the ideal theoretical incidence rates. These considerations lead us to replace the conditioning on $T \geq t$ in (4.3) and (4.4) by $(T \geq t, T > E, C > t)$, the condition that the individual be free of disease and under study at age t . We will represent this condition by $R(t)$.

Even if Eqn.(4.3) is of the form same as (4.1), the occurrence of $\{T \geq t\}$, the conditioning event, will depend to some extent on the baseline hazard function $\lambda_0(\cdot)$. In order to see this, let $f(x|w)$ be the probability density function for x given w . Then,

$$f(x|T \geq t, w) = Kp(T \geq t|x)f(x|w) = K \exp\left\{-\int_0^t e^{x^*(u)\beta} \lambda_0(u) du\right\} f(x|w), \quad (4.6)$$

where K is a constant of integration.

The corresponding relative risk (4.4) is given by

$$\int_0^t e^{x^*(t)\beta} f(x|T \geq t, w) dx, \quad (4.7)$$

which may be a complicated function of β and $\lambda_0(\cdot)$.

Prentice (1982) suggested that the “posterior” distribution of true exposure history $x(t)$ given $W(t)$ and $R(t)$ will be independent of β and $\lambda_0(\cdot)$ whenever $p(T \geq t|z) = p(T \geq t|w)$. This equality holds if $\beta = 0$ and if $f(x|w)$ is approximately concentrated at a single point, or if $p(T \geq t|x) \approx 1$ over the follow-up period. The later condition is quite likely to be satisfied in large cohort studies like Harvard Six Cities Study.

Assuming that the dependence of the covariate error distribution on β and $\lambda_0(\cdot)$ is negligible, Clayton (1991) suggested a modified regression calibration method to estimate β . In this approach, the regression calibration method is used to estimate true value x given w within each risk set.

Suppose the true exposure X is normally distributed with mean μ_X and variance σ_X^2 , and measurement errors U are normally distributed with zero mean and variance σ_U^2 . Assuming an additive error model

$$W = X + U \quad (4.8)$$

where U is uncorrelated with X , it can be shown that

$$E(X|W) = \mu_X + \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} \right) (W - \mu_W). \quad (4.9)$$

When the parameters are unknown, one can use the estimators $\hat{\mu}_X = \hat{\mu}_W = \bar{W}$, the average of the W 's, and $\hat{\sigma}_X^2 = \hat{\sigma}_W^2 - \sigma_U^2$, where $\hat{\sigma}_W^2$ is the sample variance of the observed W 's. The regression calibration estimate for β in the Cox proportional model is then found by replacing X by

$$\hat{E}(X|W) = \bar{W} + \frac{\hat{\sigma}_X^2}{\hat{\sigma}_X^2 + \sigma_U^2}(W - \bar{W}) \quad (4.10)$$

in the usual Cox regression program.

Clayton (1991) modified the regression calibration method by applying it dynamically within each risk set. At each time t_i at which an event occurs, the risk set $\mathfrak{R}_i$ consists of members of the study cohort for whom an event has not occurred and were still under observation just prior to t_i . For each member of the i^{th} risk set, the estimate of the true covariate X is

$$\hat{X} = \hat{\mu}_i + \frac{\hat{\sigma}_X^2}{\hat{\sigma}_X^2 + \sigma_U^2}(W - \hat{\mu}_i), \quad (4.11)$$

where $\hat{\mu}_i$ is the sample mean of the W 's in the i^{th} risk set. Thus, we can estimate the parameter of the model by fitting a proportional hazard model using the product $\hat{W}$ as covariate. This and the previous regression calibration method will be referred to as RCAL(dynamic) and RCAL(fixed), respectively.

Based on the above results, the simulation study can be done as follows:

Algorithm for the RCAL Method of Estimation

The simulation study involved the following steps. Details are presented for RCAL (dynamic) only; with RCAL (fixed), adjustment of the observed values is done only once, rather than each time the risk set changes.

- [R1] The Harvard Six Cities Study data are taken as representing the true values of $X = \text{PM}_{2.5}$. Thus, μ and σ^2 are given by the mean and variance of the $\text{PM}_{2.5}$ values in Harvard Six Cities Study.
- [R2] Generate u from $N(0, \tau^2)$. The dispersion parameter τ^2 will be varied to reflect different measurement error scenarios. Specifically, we consider $\tau = 1, 2, 4, 8$,

16, and 32. Values of $\tau < 1$ suggested by the data in section 2.2 are considered somewhat unrealistic; values of $\tau > 8$ are included to explore the effects of extreme measurement error.

[R3] Calculate the observed $W = X + u$.

[R4] Divide the data set into risk sets. At each time t_i ($i = 1, \dots, k$) at which a death occurs, define the risk set R_i as those members of data set for whom death has not occurred and who were still under study just prior to t_i .

To expedite the calculation in the present study, we will not consider all risk sets at which a death occurs. Instead, we will consider five risk sets with 0%, 25%, 50%, 75%, and 100% occurrence of deaths, respectively.

[R5] If W_i is the covariate associated with the i th event (death), then the estimate of the true covariate is

$$\hat{X} = \hat{\mu}_i + \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \tau^2}(W - \hat{\mu}_i), \quad (4.12)$$

where $\hat{\mu}_i$ is the mean calculated in the i th risk set.

[R6] After estimating $\hat{X}$, we fit the Cox proportional hazards model with time-dependent covariates to estimate β using $\hat{X}$. Here, the time-dependent covariate PM2.5 changes in each of the risk sets considered.

With RCAL(fixed), we consider only one risk set (with 0% deaths) in step R4. In this case, the observations are corrected just once using the equation in step R5, where μ_i is replaced by the mean of the W 's.

4.3.5 Application of SIMEX to Six Cities Study

4.3.5.1 Method

Correction for measurement error can also be done using the simulation extrapolation method (SIMEX) proposed by Cook & Stefanski (1994). SIMEX is a simulation based method for estimating and reducing bias due to measurement error. The application of SIMEX to the Six Cities Study involves the following steps.

- [S1] The Harvard Six Cities Study data is again taken as the reference data set, assumed to be free of measurement error.
- [S2] Generate u from $N(0, \sigma_u^2)$, with σ_u^2 varied as τ^2 above, i.e., we consider $\sigma_u = 1, 2, 4, 8, 16$, and 32 .
- [S3] Calculate the observed values $W = X + u$.
- [S4] A simulation step in which additional data sets with increasingly large measurement error $(1 + \lambda)\sigma_u^2$ is created. Specifically, for any $\lambda \geq 0$, we generate

$$W_{b,i}(\lambda) = W_i + \sqrt{\lambda}U_{b,i}$$

($i = 1, 2, \dots, n$; $b = 1, 2, \dots, B$), where $\{U_{b,i}\}_{i=1}^n$ are mutually independent, identically distributed, normal variables with mean 0 and variance σ_u^2 , and W_i is the observed data. For our simulation, σ_u will be varied from 1 to 32 as in the RCAL simulation, the number of simulations $B = 100$, and λ ranges from 0.25 to 2 in steps of 0.25.

- [S5] For each of the eight λ 's, 100 naive estimates are obtained using Cox proportional hazards model, and their mean $\hat{\beta}_\lambda$ is calculated.

[S6] Correction using SIMEX is based on a quadratic model $\hat{\beta}_\lambda = \alpha_1 + \alpha_2\lambda + \alpha_3\lambda^2$ for λ . The SIMEX estimate is then obtained by extrapolating to $\lambda = -1$. As with RCAL, steps S1-S6 are repeated 100 times.

4.3.6 Results

The mean, standard deviation and MSE(mean square error) of the estimates of the regression coefficient $\beta = 0.268$ in the Cox regression model based on the naive (Cox regression) estimator when standard deviation of the measurement error distribution is $\tau = 1, 2, 4, 8, 16$, or 32 are given in Table 4.2, along with the corresponding estimates using SIMEX and RCAL (fixed and dynamic) to adjust for measurement error. When the measurement error is large ($\tau > 4$), the naive estimator is notably lower than the true value of β , reflecting the well-known bias towards the null hypothesis in the presence of measurement error. SIMEX reduces this bias towards the null somewhat, although the SIMEX estimator of β remains subject to an appreciable negative bias when τ is large. Both versions of RCAL (the first using the original risk set only to adjust for measurement error, and the second dynamically updating this adjustment each time the risk set changes) perform better than SIMEX, yielding much less biased estimates when τ is large.

The distribution of estimates (both naive and adjusted) are shown in Figures 4.1-4.6 for $\tau = 1, 2, 4, 8, 16$ and 32 respectively, with a summary of results for all values of τ given in Figure 4.7. The Box plots indicate the median of the distribution of the estimates (shown by the central white bar), the interquartile range (shown by the black rectangle centered around the median), the minimum and the maximum (shown by the first set of horizontal lines above and below the interquartile range), and outlying values (shown by the second set of horizontal lines, whenever an outlier

is present in the simulated data). These graphical results confirm the indications of central tendency shown in Table 4.2. Since RCAL(fixed) demonstrates more variability than RCAL (dynamic), the later version of RCAL appears preferable with respect to precision. SIMEX exhibits surprisingly little variation, but fails to adequately adjust for the presence of measurement error in the data.

In order to explore the robustness of the above methods of adjustment when the measurement error variance is misspecified, we conducted a further simulation for values of τ other than the true value. In light of the poor performance of SIMEX, we will confine our attention in what follows to RCAL. The results in Table 4.3 and Figures 4.8-4.13 indicate that RCAL is highly sensitive to the choice of τ , seriously overestimating β when τ is assumed to be much larger than it actually is, and underestimating β when τ is assumed to be much smaller than the true value. As in the case when τ is correctly specified, RCAL(fixed) demonstrates more variation than RCAL(dynamic).

Table 4.2: Mean, Standard Deviation (in parenthesis), and MSE-EME¹ (**bold**) of Estimates of $\beta = 0.268$ Using Different Methods of Estimation when the Measurement Error Variance τ is Known.

τ	Naive Est.	SIMEX	RCAL(fixed)	RCAL(dynamic)
1	0.263 (0.012) 0.0002	0.271 (0.001) 0.00001	0.269 (0.013) 0.0002	0.269 (0.011) 0.0001
2	0.248 (0.023) 0.0009	0.273 (0.002) 0.00003	0.268 (0.025) 0.0006	0.268 (0.022) 0.0005
4	0.198 (0.037) 0.0063	0.260 (0.004) 0.0008	0.266 (0.052) 0.0027	0.263 (0.043) 0.0019
8	0.110 (0.041) 0.0266	0.174 (0.005) 0.0089	0.261 (0.097) 0.0095	0.254 (0.084) 0.0072
16	0.040 (0.029) 0.0528	0.070 (0.004) 0.0392	0.263 (0.188) 0.0354	0.247 (0.162) 0.0267
32	0.012 (0.016) 0.0658	0.022 (0.002) 0.0605	0.281 (0.369) 0.1363	0.249 (0.318) 0.1015

¹Mean Square Error due to exposure measurement error

Table 4.3: Mean, Standard Deviation (in parenthesis), and MSE-EME² (**bold**) of Estimates of $\beta = 0.268$ for RCAL(fixed) when the Measurement Error Variance is Mis-specified.

Assumed τ	True τ					
	1	2	4	8	16	32
1	0.269	0.253	0.202	0.112	0.041	0.012
	(0.013)	(0.024)	(0.038)	(0.042)	(0.03)	(0.016)
	0.0002	0.0008	0.0058	0.0261	0.0524	0.0658
2	0.286	0.268	0.215	0.119	0.044	0.013
	(0.014)	(0.025)	(0.041)	(0.044)	(0.031)	(0.017)
	0.0005	0.0006	0.0045	0.0241	0.0511	0.0653
4	0.354	0.332	0.266	0.148	0.054	0.016
	(0.017)	(0.031)	(0.052)	(0.055)	(0.039)	(0.022)
	0.0077	0.0051	0.0027	0.0174	0.0473	0.0640
8	0.625	0.588	0.47	0.261	0.096	0.029
	(0.030)	(0.056)	(0.089)	(0.097)	(0.069)	(0.038)
	0.1283	0.1055	0.0487	0.0072	0.0343	0.0586
16	1.71	1.608	1.287	0.715	0.263	0.08
	(0.081)	(0.152)	(0.243)	(0.266)	(0.188)	(0.104)
	2.0859	1.8187	1.0974	0.2706	0.0267	0.0586
32	6.056	5.69	4.56	2.53	0.931	0.281
	(0.287)	(0.539)	(0.860)	(0.94)	(0.664)	(0.369)
	33.5833	29.6995	19.1180	0.0002	0.8805	0.1015

²Mean Square Error due to exposure measurement error

Table 4.4: Mean, Standard Deviation (in parenthesis), and MSE-EME³ (**bold**) of Estimates of $\beta = 0.268$ for RCAL(dynamic) when the Measurement Error Variance τ^2 is Mis-specified.

Assumed τ	True τ					
	1	2	4	8	16	32
1	0.269 (0.011) 0.0001	0.252 (0.021) 0.0007	0.2 (0.033) 0.0057	0.109 (0.360) 0.1549	0.039 (0.025) 0.0531	0.011 (0.014) 0.0662
2	0.286 (0.012) 0.0005	0.268 (0.022) 0.0005	0.213 (0.035) 0.0043	0.116 (0.038) 0.0254	0.041 (0.027) 0.0523	0.012 (0.015) 0.0658
4	0.354 (0.015) 0.0076	0.332 (0.027) 0.0048	0.263 (0.043) 0.0019	0.144 (0.047) 0.0176	0.051 (0.033) 0.0482	0.015 (0.019) 0.0644
8	0.626 (0.026) 0.1288	0.586 (0.048) 0.1034	0.465 (0.077) 0.0447	0.254 (0.084) 0.0072	0.090 (0.059) 0.0352	0.057 (0.033) 0.0012
16	1.715 (0.070) 2.0987	1.604 (0.132) 1.8023	1.273 (0.210) 1.0541	0.695 (0.229) 0.2348	0.247 (0.162) 0.0267	0.070 (0.09) 0.0473
32	6.068 (0.249) 33.7020	5.677 (0.466) 29.4744	4.504 (0.744) 18.4972	2.459 (0.812) 5.4598	0.875 (0.573) 0.6968	0.0249 (0.319) 0.1015)

³Mean Square Error due to exposure measurement error

Figure 4.1: Box plots of estimates of beta using different methods of estimation when true measurement error is known ($\tau = 1$).

Figure 4.2: Box plots of estimates of beta using different methods of estimation when true measurement error is known ($\tau = 2$).

Figure 4.3: Box plots of estimates of beta using different methods of estimation when true measurement error is known ($\tau = 4$).

Figure 4.4: Box plots of estimates of beta using different methods of estimation when true measurement error is known ($\tau = 8$).

Figure 4.5: Box plots of estimates of β using different methods of estimation when true measurement error is known ($\tau = 16$).

Figure 4.6: Box plots of estimates of beta using different methods of estimation when true measurement error is known ($\tau = 32$).

Figure 4.7: Summary of Box plots of estimates of beta using different methods of estimation when true measurement error is known.

Figure 4.8: Box plots of estimates of beta using Naive, SIMEX, RCAL(fixed), RCAL(dynamic) methods of estimation when measurement error is misspecified (true measurement error $\tau = 1$).

Figure 4.9: Box plots of estimates of beta using Naive, SIMEX, RCAL(fixed), RCAL(dynamic) methods of estimation when measurement error is misspecified (true measurement error $\tau = 2$).

Figure 4.10: Box plots of estimates of beta using Naive, SIMEX, RCal(fixed), RCal(dynamic) methods of estimation when measurement error is misspecified (true measurement error $\tau = 4$).

Figure 4.11: Box plots of estimates of beta using Naive, SIMEX, RCAL(fixed), RCAL(dynamic) methods of estimation when measurement error is misspecified (true measurement error $\tau = 8$).

Figure 4.12: Box plots of estimates of beta using Naive, SIMEX, RCAL(fixed), RCAL(dynamic) methods of estimation when measurement error is misspecified (true measurement error $\tau = 16$).

Figure 4.13: Box plots of estimates of beta using Naive, SIMEX, RCAL(fixed), RCAL(dynamic) methods of estimation when measurement error is misspecified (true measurement error $\tau = 32$).

4.3.7 Discussions

Most epidemiologic studies involve a certain degree of measurement error, with random measurement error leading to both greater imprecision and a bias towards the null value of regression model parameters, including those in the Cox regression model. If the measurement error distribution is known, measurement error adjustment techniques such as regression calibration can be used to increase precision and reduce bias. However, mis-specification of the measurement error distribution can lead to further bias when using regression calibration, compounding the original problem posed by inherent error in measured covariates. Consequently, there is a need for measurement error adjustment methods that are robust against mis-specification of the measurement error distribution.

Our first simulation study involved a comparison of Cox model parameter estimates obtained using both RCAL and SIMEX to adjust for measurement error, using plausible measurement error distributions defined on the basis of observed spatial variation in particulate concentrations within selected cities in the North America. SIMEX has the advantage of not having to specify the functional form of the measurement error distribution, but is computationally intensive because of the need to simulate a large number of replicate datasets during the course of the analysis.

The unadjusted estimates demonstrated both bias and imprecision in the presence of moderate to large measurement error in particulate levels. Both RCAL and SIMEX produced estimates with less bias and greater precision than the unadjusted estimates when the measurement error distribution was correctly specified, with RCAL outperforming SIMEX in the presence of large measurement error. Neither RCAL nor SIMEX method performed well when the degree of measurement error was grossly misspecified.

Although the present simulation study was restricted to a model involving only a single covariate (PM_{2.5}), additional research on models involving multiple covariates is also desirable. Zidek *et al.* (1998) showed that when two covariates are measured with different degrees of error in Poisson regression, the predictive effect of a causal variable measured with error may be transferred to a surrogate variable due to multicollinearity. Similar results were reported by Fung & Krewski (1999a). Fung & Krewski (1999b) considered the case of two correlated covariates in logistic regression when one of the predictors is subject to additive measurement error. Liu *et al.* (2001) considered interactions between continuous and categorical covariates measured with and without measurement error, respectively, in the analysis of data on time to recurrence of depression, although the interaction was not statistically significant.

In this study, exposure measurement error derives largely from the use of fixed site ambient monitors rather than personal dosimeters to gauge individual exposure to particulate air pollution. Although most epidemiologic studies of air pollution and population health rely on ecologic exposure data (NRC, 2001), a few studies involving personal dosimeters have been conducted (Janssen *et al.*, 1998; Ozkaynak *et al.*, 1996; Stieb *et al.*, 1998). When available, individual exposure data can be used to construct more complicated measurement error models than those considered here (Dominici *et al.*, 2000).

Variation among air pollution levels recorded at different locations within the same urban area was used to gauge the variance of the exposure measurement error distribution, assuming that spatial variation in air pollution levels reflects inter-individual variation in exposure. However, measurement error associated with spatial variation in long-term average PM_{2.5} concentrations will be reduced to some extent by population

mobility within urban areas (Warner *et al.*, 1996). Intra-urban spatial autocorrelation in air pollution levels (Burnett *et al.*, 2001), which has not been explicitly considered in this investigation, may also affect estimates of exposure measurement error based on the observed variation in pollution concentration among fixed site monitors.

One of our original objectives was to investigate the extent to which SIMEX, as a non parametric method of measurement error adjustment, may be more robust against mis-specification of the measurement error distribution than a parametric method such as RCAL. Since SIMEX performed considerably less well than RCAL when the measurement error distribution was correctly specified, the robustness of SIMEX is of limited interest. Further work on the robustness of other methods of adjusting for measurement error when the error distribution is mis-specified is desirable, including the new methods proposed by Hu *et al.* (1998) and Huang and Wang (2000). Although the search for a robust method of measurement error adjustment may prove elusive, the identification of such a method would be of considerable practical utility, since the distribution of exposure measurement error is generally imperfectly specified.

Although RCAL performs well when the measurement error distribution is normal with moderate variance, additional research is needed on measurement error adjustment methods that may be applied with non-normal error distributions. Hu *et al.* (1998) have developed more complex methods that can be applied with non-normal error distributions which appear promising. Although this method was not included in the present simulation study because of the lack of readily available software, a Fortran program (NLMIX) now available in STATLIB may serve to promote the use of such methods in the future.

In the absence of multiple monitors or personal dosimeters in the Harvard Six Cities

Study, it is difficult to assess the impact of exposure measurement error on estimates of risk in that study. The results in Table 1 may provide some indication of the plausible range of values for the variance of the measurement error distribution, although the range is quite wide. Such differences may be attributed to differences in population size (most of these cities are larger than those in the Harvard Six Cities Study), geography and climate, and both local and distant sources of particulate air pollution. Population mobility may also lead to overestimation of measurement error variance. Because the preferred measurement error adjustment method (RCAL) is not robust against misspecification of the measurement error variance, it is important that the measurement error distribution be well-characterized before this method is employed. Since our main objective is to compare the performance of two well known measurement error adjustment methods, RCAL and SIMEX, our results should be treated as broadly indicative rather than definitive with respect to the possible effect of measurement error on risk estimates in the Harvard Six Cities Study.

Chapter 5

Summary and Conclusions

5.1 Summary of Main Contributions

The main contributions of this thesis are the following.

1. Although CRL has been used for some time in cohort mortality studies, the impact of linkage errors on the reliability of statistical inferences drawn from such studies has not been subjected to detailed investigation. We presented theoretical results to address this issue. In particular, we investigated the impact of record linkage errors on estimates of epidemiological indicators of risk such as the standardized mortality ratio and the parameters of relative risk regression models. A simulation study was also conducted to see how these indicators are subject to biases and additional variability in the presence of linkage errors.

Our results show that in addition to inflating the observed number of deaths, false positives tend to deflate the expected number of deaths. Conversely, false negatives inflate the expected numbers of deaths and deflate the observed number of deaths. Linkage errors introduce bias into estimates of SMRs. Relative risk regression coefficients are also subject to bias, the direction of which depends on

the nature of the regression coefficient. In addition to these biases, linkage errors introduce additional uncertainty into estimates of both SMRs and regression coefficients.

2. We presented a general method for adjusting the biases in the indicators of risk in the presence of linkage errors. The method works as follows. A sample drawn from the population is matched manually and the manual match is taken as the right link. Comparing manual match with the computerized match, a correction is made to the observed death, expected death, and exposure level. The bias of the estimates of SMR and ERR based on those adjusted values are then calculated. Our method, although sensitive to sample size, was effective in reducing the bias when the error rates were high. When the bias was not very high or the error rates were not high, the method performed poorly, especially in case of the ERR. While this gives encouraging results for the case of higher rates of error (especially false positives), the method needs some improvement in case of smaller error rates.
3. If the measurement error distribution is known, measurement error adjustment techniques such as regression calibration can be used to increase precision and reduce bias. However, mis-specification of the measurement error distribution can lead to further bias when using regression calibration, compounding the original problem posed by inherent error in measured covariates. Consequently, there is a need for measurement error adjustment methods that are robust against mis-specification of the measurement error distribution. We evaluated the effectiveness of two computer-based simulation methods RCAL and SIMEX for correcting measurement errors in Cox proportional hazard model, using real data from the Harvard Six Cities Study that focuses on the effect of chronic exposure

to airborne pollutants.

We compared Cox model parameter estimates obtained using both RCAL and SIMEX to adjust for measurement error, using plausible measurement error distributions defined on the basis of observed spatial variation in particulate concentrations within selected cities in the North America. SIMEX was found to have the advantage of not having to specify the functional form of the measurement error distribution, but is computationally intensive because of the need to simulate a large number of replicate datasets during the course of the analysis. The unadjusted estimates demonstrated both bias and imprecision in the presence of moderate to large measurement error in particulate levels. Both RCAL and SIMEX produced estimates with less bias and greater precision than the unadjusted estimates when the measurement error distribution was correctly specified, with RCAL outperforming SIMEX in the presence of large measurement error. Neither RCAL nor SIMEX method performed well when the degree of measurement error was grossly misspecified.

5.2 Areas for Future Work

Record linkage is a powerful tool for efficiently conducting cohort mortality studies based on existing databases. However, it can introduce error which affects different health indicators used to study the effect of exposure on mortality. We need to find ways to reduce the bias of the estimates calculated using the linked data. The method proposed in Chapter 3 did well in reducing the bias in estimates of SMR. It also was able to reduce the bias in the estimate of ERR, even though it introduced more bias when the false positive rate was low. The bias reduction also depended on the

size of the sample drawn to assess the incorrect matches. Although better results were achieved with larger sample sizes, bigger sample size will require more time and money. Therefore, we need to explore to find ways which can reduce the bias effectively even with smaller sample sizes and to determine what should be the optimal sample size for a specified level of precision.

We assessed two methods RCAL (fixed and dynamic) and SIMEX methods to account for measurement error with Cox proportional hazards model. One of our original objectives was to investigate the extent to which SIMEX, as a nonparametric method of measurement error adjustment may be more robust to mis-specification of the measurement error distribution than a parametric method such as RCAL. Since SIMEX performed considerably less well than RCAL when the measurement error distribution was correctly specified, the robustness of SIMEX becomes of less interest. Further work on the robustness of other methods of adjusting for measurement error when the error distribution is mis-specified is desirable, including the new methods proposed by Hu *et al.* (1998) and Huang and Wang (2000). Although the search for a robust method of measurement error adjustment may prove elusive, the identification of such a method would be of considerable practical utility, since the distribution of exposure measurement error is generally imperfectly specified.

Bibliography

- [1] Adami, H., Hunter David and Trichopoulos, D. (2002). *Textbook of Cancer Epidemiology*, Oxford University Press, New York.
- [2] Anderson, T.W. (1974). *An Introduction to Multivariate Statistical Analysis*, New York, Wiley.
- [3] Ardal, S. and Ennis, S. (2001). Data Detectives: Uncovering Systematic Errors in Administrative Databases. *Proc. of Statistics Canada Symposium (Achieving Data Quality in a Statistical Agency: A Methodological Perspective)*.
- [4] Armstrong, B.K., White, E. and Saracci, R. (1992). *Principles of Exposure Measurement in Epidemiology*, Oxford University Press.
- [5] Ashmore, J.P. and Davies, B.D. (1989). The National Dose Registry: A Centralized Record Keeping System for Radiation Workers in Canada. *Applications of Computer Technology to Radiation Protection. IAEA-SR-136/58, J. Stephan Institute, Ljublyua*, 505-520.
- [6] Ashmore, J.P. and Grogan, D. (1985). The National Dose Registry of Canada. *Radiation Protection Dosimetry* 11, 95-100.

- [7] Ashmore, J.P., Krewski, D. and Zielinski, J.M. (1997). Protocol for a Cohort Mortality Study of Occupational Radiation Exposure Based on the National Dose Registry of Canada. *European Journal of Cancer* 33, S10-S21.
- [8] Ashmore, J.P., Krewski, D., Zielinski, J.M., Jiang, H., Semenciw, R., and Band, P.R. (1998). First Analysis of Mortality and Occupational Radiation Exposure based on the National Dose Registry of Canada. *American Journal of Epidemiology* 148, 564-574.
- [9] Bartlett, S., Krewski, D., Wang, Y. and Zielinski, J.M. (1993). Evaluation of error rates in large scale computerized record linkage studies. *Survey Methodology* 19, 3-12.
- [10] Belin, T.R. and Rubin, D.B. (1991). Recent Developments in Calibrating Error Rates for Computer Matching. Proceedings of the 1991 Annual Research Conference, U.S. Bureau of the Census, 657-668.
- [11] Box, G.E.P. and Cox, D.R. (1964). An Analysis of Transformations, *Journal of The Royal Statistical Society*, 211-243.
- [12] Breslow, N.E., Lubin, J.H. and Langholz, B. (1983). Multiplicative Models and Cohort Analysis. *Journal of the American Statistical Association* 78, 1-12.
- [13] Breslow, N.E. and Day, N.E., (1987). Statistical Methods in Cancer Research (Vol. II): The Design and Analysis of Cohort Studies. International Agency for Research on Cancer Scientific Publications, Lyon, France.
- [14] Burnett, R.T., Ma, R., Jerrett, M., Goldberg, M. S., Cakmak, S., Pope, A., and Krewski, D. (2001). The spatial association between community air pollution

and mortality: A new method of analyzing correlated geographic cohort data. *Environmental Health Perspectives* 109, 375-380.

- [15] Carpenter, M. and Fair, M.E. (Eds.) (1990). *Canadian Epidemiology Research Conference - 1989: Proceedings of Record Linkage Sessions & Workshop*, Ottawa Select Printing, Ottawa.
- [16] Carroll, R.J. (1989). Covariance analysis in generalized linear measurement error models. *Statistics in Medicine* 8, 1075-1093.
- [17] Carroll, R.J., Freedman, L.S., Kipnis, V. and Li, L. (1998). A new class of measurement-error models, with applications to dietary data. *The Canadian Journal of Statistics* 26, 1-11.
- [18] Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). *Nonlinear Measurement Error Models*. Chapman and Hall, New York.
- [19] Carroll, R.J., Kuchnhoff, H., Lombard, F. and Stefanski, L.A. (1996). Asymptotics for the SIMEX estimator in structural measurement error models. *Journal of American Statistical Association* 91, 242-250.
- [20] Carroll, R.J. and Stefanski, L.A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of American Statistical Association* 85, 652-663.
- [21] Clayton D.G. (1991). Models for the analysis of cohort and case-control studies with inaccurately measured exposures. *Statistical Models for Longitudinal Studies of Health*, Dwyer, J.H., Feinleib, M., Lipsert, P., et al., editors, Oxford University Press, New York, 301-331.

- [22] Cook, J. and Stefanski, L.A. (1994). A simulation extrapolation method for parametric measurement error models. *Journal of American Statistical Association* 89, 1314-1328.
- [23] Cox, D.R (1972). Regression Models with Life Tables (with Discussion). *Journal of Royal Statistical Society* 34, 187-220.
- [24] De Klerk, N. H., English, N. H. and Armstrong, B. K. (1989). A Review of the Effects of Random Measurement Error in Relative Risk Estimates in Epidemiological Studies. *International Journal of Epidemiology* 18(3), 705-712.
- [25] Dockery, D.W., Pope, A., Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, B.G. and Speizer, F.E. (1993). An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine* 329, 1753-1759.
- [26] Doll, R. and Hill, A.B. (1956) Lung Cancer and other causes of death in relation to smoking. *British Medical Journal* 2, 1071-1081.
- [27] Doll R, Peto R, Boreham J, Sutherland I. (2004) Mortality in relation to smoking: 50 years' observation on male British doctors. *British Medical Journal* 328, 1519-33.
- [28] Dominici, F., Zeger, S., Samet, J.M. (2000) A measurement error model for time-series studies of air pollution and mortality. *Biostatistics* 1, 157-175.
- [29] Fair, M.E. (1989). Studies and References Relating to Uses of the Canadian Mortality Data Base, Report from the Occupational and Environmental Health Research Unit, Health Division, Statistics Canada, Ottawa.

- [30] Fellegi, I.P. and Sunter, A. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association* 64, 1183-1210.
- [31] Fuller, W. (1987). *Measurement Error Models*, Wiley, New York.
- [32] Fuller, W.A. (1995). Estimation in the Presence of Measurement Error. *International Statistical Review* 63 (2), 121-147.
- [33] Fung, K.Y. and D. Krewski (1999a). On Measurement Error Adjustment Methods in Poisson Regression. *Environmetrics* 10, 213-224.
- [34] Fung, K.Y. and D. Krewski (1999b). Evaluation of regression calibration and SIMEX methods in logistic regression when one of the predictors is subject to additive measurement error, *Journal of Epidemiology and Biostatistics* 4, 65-74.
- [35] Gilbert, E.S., Fry, S.A., Wiggs, L.D., Voelz, G.L., Cragle, D.L. and Petersen, G.R. (1990). Methods for analyzing combined data from studies of workers exposed to low doses of radiation, *American Journal of Epidemiology* 131, 917-927.
- [36] Greenbaum, D.S., Bachmann, J.D., Krewski, D., Samet, J.M., White, R., and Wyzga, R.E. (2001) Case study: Particulate air pollution standards and morbidity and mortality. *American Journal of Epidemiology* 154, 78S-90S.
- [37] Greenland, S. (1980). The effect of misclassification in the presence of covariates. *American Journal of Epidemiology* 116, 402-406.
- [38] Gribbin, M.A., Weeks, J.L. and Howe, G. R. (1993). Cancer Mortality (1956-1985) among Male Employees of Atomic Energy of Canada Limited Respect to Occupational Exposure to External Low-Linear-Energy-Transfer Ionizing Radiation, *Radiation Research* 133, 365-380.

- [39] Hedayat, A.S. and Sinha, B.K. (1991). *Design and Inference in Finite Population Sampling*, Wiley, New York.
- [40] Hill, T. (1988). Generalized Iterative Record Linkage System: GIRLS Strategy (Release 2.7), Report from Research and General System, Informatics Services and Development Division, Statistics Canada, Ottawa.
- [41] Howe, G.R. and Lindsay, J. (1981). A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies, *Computers and Biomedical Research* 14, 327-340.
- [42] Howe, G.R. and Spasoff, R.A. (Eds.) (1986). *Proceeding of the Workshop on Computerized Linkage in Health Research.*, University of Toronto Press, Toronto.
- [43] Hu, P., Tsiatis, A.A., and Davidian, M. (1998). Estimating the Parameters in the Cox Model When Covariate Variables are Measured with Error. *Biometrics* 54, 1407-1419.
- [44] Huang, Y. and Wang, C.Y. (2000). Cox Regression with Accurate Covariates Unascertainable: A Nonparametric Correction Approach. *Journal of the American Statistical Association* 95, 1209-1219.
- [45] Hughes, M.D. (1993). Regression dilution in the proportional hazards model. *Biometrics* 49, 1056-1066.
- [46] Janssen, N., Hoek, G., Brunekreef, B., Harssema, H., Mensink, I., and Zuidhof, A. (1998). Personal sampling of particles in adults: relation among personal, indoor, and outdoor air concentrations. *American Journal of Epidemiology* 147, 537-547.

- [47] Jaro, M.A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* 84, 414-420.
- [48] Jerrett, M., Burnett, R., Kanaroglou, P., Finkelstein, N., Giovis, C., and Brook, J. (2002) Geographies of susceptibility in the air pollution-mortality association: a zonal time series from Hamilton, Canada. *Epidemiology* 13, S152.
- [49] Jordan-Simpson, D.A., Fair, M.E., and Poliquin, C. (1990). Canadian Farm Operator Study: Methodology. *Health Reports* 2, 141-155.
- [50] Kalbfleish, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*, New York, Wiley.
- [51] Kelly P. (1986). Robustness of the Census Bureaus Record Linkage System. *Proceedings of the Section on Survey Research Methods, American statistical Association*, 620-624.
- [52] Krewski, D., Burnett, R.T., Goldberg, M.S., Hoover, K., Siemiatycki, J., Jerrett, M., Abrahamowicz, M., and White, W.H. (2000). Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality – Special Report, Health Effects Institute, Cambridge, MA, USA.
- [53] Krewski, D., Dewanji, A., Wang, Y., Bartlett, S., Zielinski, J.M., and Mallick, R. (2002). The Effect of Record Linkage Errors on Statistical Inference in Cohort Mortality Studies, Submitted to *Survey Methodology*.
- [54] Krewski, D., Burnett, R., Goldberg, M.S., Hoover, K., Siemiatycki, J., Jerrett, M., Abrahamowicz, M., and White, W.H. (2003). Overview of the re-analysis

of the Harvard six cities study and American Cancer Society study of particulate air pollution and mortality. *Journal of Toxicology and Environmental Health* 66:1507-1552.

- [55] Labossière, G. (1986). Confidentiality and Access to Data: The Practice at Statistics Canada. *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, University of Toronto Press, Toronto.
- [56] Lin, M., Chen, Y., Burnett, R.T., Villeneuve, P., and Krewski, D. (2002) The influence of ambient coarse particulate matter on asthma hospitalization in children: case-crossover and time series analyses. *Environmental Health Perspectives* 110, 575-581.
- [57] Liu, K., Mazumdar, S., Stone, R.A., Dew, M.A., Houck, P.R., and Reynolds, C.F. III (2001). Accounting for covariate measurement error in a Cox model analysis of recurrence of depression. *Journal of Psychiatric Research* 35, 177-185.
- [58] Mallick, R., Krewski, D., Dewanji, A., and Zielinski, J.M. (2002). A Simulation Study of the Effect of Record Linkage Errors in Cohort Mortality Data", *Proceedings of the International Conference in Recent Advances in Survey Sampling*, Carleton University.
- [59] Nakamura, T. (1992) Proportional hazards model with covariates subject to measurement error. *Biometrics* 48, 829-838.
- [60] National Research Council (2001) *Research Priorities for Airborne Particulate Matter III: Early Research Progress*. National Academy Press, Washington, DC.
- [61] Nelder, J.A. (1972). Discussion of Paper by Lindley and Smith, *Journal of the Royal Statistical Society* 24, 1-41.

- [62] Neter, J., Maynes, E.S., and Ramanathan, R. (1965). The Effect of Mismatching on the Measurement of Response Errors. *Journal of the American Statistical Association* 60, 1005-1027.
- [63] Newcombe, H.B. (1988). Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business, Oxford Medical Publications, Oxford.
- [64] Newcombe, H.B. (1998). Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business. Oxford University Press, Oxford, United Kingdom.
- [65] Ozkaynak, H., Xue, J., Spengler, J., Wallace, L., Pellizzari, E., and Jenkins, P. (1996). Personal exposure to airborne particles and metals: results from the particle team study in Riverside, California. *Journal of Exposure Analysis and Environmental Epidemiology* 6, 57-78.
- [66] Pepe, M.S., Self, S.G., and Prentice, R.L. (1989). Further results in covariate measurement errors in cohort studies with time to response data. *Statistics in Medicine* 8, 1167-1178.
- [67] Pope, C.A. III, Thun, M.J., Namboodiri, M.M., Dockery, D.W., Evans, J.S., Speizer, F.E., and Heath, C.W. (1995). Particulate air pollution as a predictor of mortality in a prospective study of US adults *American Journal of Respiratory and Critical Care Medicine* 151, 669-674
- [68] Pope, A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., and Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality and long-term

- exposure to fine particulate air pollution. *Journal of American Medical Association* 287, 1132-1141.
- [69] Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 69, 331-342.
 - [70] Roos, L.L., Soodeen, R., Jebamani, L. (2001). An Information-Rich Environment: Linked-Record Systems and Data Quality in Canada. *Proc. of Statistics Canada Symposium (Achieving Data Quality in a Statistical Agency: A Methodological Perspective)*.
 - [71] Rothman, K.J. and Greenland, S. (1998). *Modern Epidemiology*, Lippincott Williams and Wilkins, Philadelphia.
 - [72] Schafer, D. (1992). Replacement Methods for Measurement Error Models, Preprint.
 - [73] Scheuren, F. and Winkler, W.E. (1997). Regression Analysis of Data Files that are Computer Matched - Part II. *Survey Methodology* 23, 157-165.
 - [74] Scheuren, F. and Winkler, W.E. (1993). Regression Analysis of Data Files that are Computer Matched. *Survey Methodology* 19, 39-58.
 - [75] Schlesselman, J.J. (1982). Case-Control Studies: Design, Conduct, Analysis. Oxford University Press, Oxford.
 - [76] Schmiediche, H., Carroll, R.J., and Newton, H.J. (1997). QVF Program for Quasilikelihood Estimation with Measurement Error in Predictors, Dept. of Statistics, Texas A & M University.

- [77] Singh, A.C., Feder, M., Dunteman, G., Yu, F. (2001). Protecting Confidentiality while Preserving Quality of Public Use Micro Data. *Proc. of Statistics Canada Symposium (Achieving Data Quality in a Statistical Agency: A Methodological Perspective)*.
- [78] Smith, M.E. and Silins, J. (1981). Generalized Iterative Record Linkage System, *Social Statistics Section, Proceedings of the American Statistical Association*, 128-137.
- [79] Sont, W.N., Zielinski, J.M., Ashmore, J.P., Jiang, H., Krewski, D., Fair, M.E., Band, P. and Létourneau, E. (2001). First Analysis of Cancer Incidence and Occupational Radiation Exposure Based on the National Dose Registry of Canada, *American Journal of Epidemiology* 153, 309-318.
- [80] Stefanski, L.A. (1985). The effect of measurement error on parameter estimation. *Biometrika* 72, 583-592.
- [81] Stefanski, L.A. and Carroll, R.J. (1985). Covariate measurement error in logistic regression. *Annals of Statistics* 13, 1335-1351.
- [82] Stefanski, L.A. and Cook, J. (1996). Simulation Extrapolation: The Measurement Error Jackknife, *Journal of American Statistical Association*.
- [83] Stieb, D.M., Brook, J.R., Broder, I., Judek, S., Burnett, R.T., and Beveridge, R. C. (1998). Personal exposure of adults with cardiorespiratory disease to particulate acid and sulfate in Saint John, New Brunswick, Canada. *Applied Occupational Environmental Hygiene* 13, 461-468.

- [84] Villeneuve, P.J., Krewski, D., Burnett, R.T., Goldberg, M.S., and Chen, Y. (2002). Fine particulate air pollution and all-cause mortality with the Harvard Six-cities Study: Variations in risk by period of exposure. *Epidemiology*, In press.
- [85] Warner, K.E., Fulton, G.A., Nicolas, P., and Grimes, D.R. (1996). Toward a more realistic appraisal of the lung cancer risk from radon: the effects of residential mobility. *American Journal of Public Health* 86, 1222-1227.
- [86] Winkler, W.E. and Scheuren, F. (1991). How Computer Matching Error effect Regression Analysis: Exploratory and Confirmatory Analysis. Technical Report, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- [87] Zidek, J.V., Le, N.D., Wong, H., and Burnett, R. (1998). Including structural measurement errors in the nonlinear regression analysis of clustered data. *Canadian Journal of Statistics* 26, 1-12.
- [88] Zielinski, J.M., Garner, M.J., Krewski, D., Ashmore, J.P., Band, P.R., Fair, M.E., Jiang, H., Letourneau, E.G., Semenciw, R., and Sont, W.N. (2005). Decreasing trends in occupational exposure to ionizing radiation among dental workers in Canada. *Journal of the Canadian Dental Association*, to appear.