The Effect of Record Linkage Errors on Risk Estimates in Cohort Mortality Studies

D. Krewski, A. Dewanji, Y. Wang, S. Bartlett, J.M. Zielinski and R. Mallick [1]

Abstract

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, however, introduces linkage errors due to mismatching an individual from one database with a different individual from the other database. In this article, the impact of linkage errors on estimates of epidemiological indicators of risk such as standardized mortality ratios and relative risk regression model parameters is explored. It is shown that the observed and expected number of deaths are affected in opposite direction and, as a result, these indicators can be subject to bias and additional variability in the presence of linkage errors.

Key Words: Cohort study; Computerized record linkage; Linkage errors; Linkage threshold weight; Poisson regression; Relative risk regression; Standardized mortality ratio.

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, Krewski, Wang and Zielinski 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 nonlinks. 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 nonlinks.

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, Feder, Dunteman and Yu 2001). 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, Fair and Poliquin 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

1. D. Krewski, McLaughlin Centre for Population Health Risk Assessment, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5. School of Mathematics & Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6. To whom correspondence should be addressed; A. Dewanji, Applied Statistics Unit, Indian Statistical Institute, Kolkata, India; Y. Wang, Healthy Environments and Consumer Safety Branch, Health Canada, Ottawa, Ontario, Canada, K1A 0L2; S. Bartlett, Healthy Environments and Consumer Safety Branch, Health Canada, Ottawa, Ontario, Canada, K1A 0L2; J.M. Zielinski, Healthy Environments and Consumer Safety Branch, Health Canada, Ottawa, Ontario, Canada, K1A 0L2; R. Mallick, McLaughlin Centre for Population Health Risk Assessment, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5. School of Mathematics & Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

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, Krewski and Zielinski 1997; Ashmore, Krewski, Zielinski, Jiang, Semenciw and Létourneau 1998). More recently, the NDR has been linked to the Canadian Cancer Incidence Database (Sont, Zielinski, Ashmore, Jiang, Krewski, Fair, Band and Létourneau 2001). 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, Soodeen and Jebamani 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 paper. Inevitably, some records that match will fail to be linked, and other nonmatching records will be incorrectly linked.

Relatively little work has been done to determine the impact of these linkage errors on statistical inferences. Neter, Maynes and Ramanathan (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).

The purpose of this paper is to explore the impact of linkage errors on statistical inferences in cohort mortality studies. Relative risk regression models employed in the analysis of data from such studies are described in section 2, and expressions for the observed and expected numbers of deaths based on these models developed. The impact of linkage errors on the observed and expected number of deaths and person-years at risk is discussed in section 3. An analysis of the impact of linkage errors on estimates of standardized mortality ratios (SMRs) and relative risk regression parameters is given in section 4. Both types of errors can cause bias and additional variability in estimates of these parameters. Our conclusions are presented in section 5.

2. Relative Risk Regression Models

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

$$\lambda(u) = \lim_{\Delta u \downarrow 0} \frac{\Pr\{u \le T < u + \Delta u \,|\, T \ge u\}}{\Delta u}. \qquad (1)$$

Let $\lambda_i(u)$ denote the hazard function for a specific cause of death at time u for individual $i = 1, \ldots, N$ 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

$$\lambda_i(u) = \lambda^*(u)\gamma\{\beta' \mathbf{z}_i(u)\}, \qquad (2)$$

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) by

$$\log\gamma(z) = \frac{(1+z)^\rho - 1}{\rho}. \qquad (3)$$

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

$$\lambda_i(u) = \lambda^*(u)\exp\{\beta' \mathbf{z}_i(u)\}, \qquad (4)$$

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

$$\lambda_i(u) = \lambda^*(u) + \beta' \mathbf{z}_i(u) \tag{5}$$

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

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, respectively. 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) may be written as

$$\log L = \sum_{i=1}^{N} \left\{ \begin{array}{l} \delta_i \log(\gamma\{\beta' \mathbf{z}_i(t_i^1)\}) \\ \\ \quad - \int_{t_i^0}^{t_i^1} \gamma\{\beta' \mathbf{z}_i(u)\} \lambda^*(u) du \end{array} \right\}. \tag{6}$$

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 SMR = OBS/EXP, where OBS = $\sum_{i=1}^{N} \delta_i$ and 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 (6) can be computationally burdensome with large sample sizes. Breslow, Lubin and Langholz (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 $\{S_j;\ j = 1, \ldots, J\}$ 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 (6) may then be written as

$$\log L = \sum_{j=1}^{J} \{d_{jj} \log(\gamma\{\beta' \mathbf{z}_j\}) - \gamma\{\beta' \mathbf{z}_j\} e_j\}, \tag{7}$$

where

$$e_j = \sum_{i=1}^{N} \int_{[\mathbf{z}_i(u) \in S_j]} \lambda^*(u) du \tag{8}$$

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

$$\frac{\partial \log L}{\partial \beta} = \sum_{j=1}^{J} \frac{\partial \Lambda_j(\hat{\beta})}{\partial \beta} \{d_{jj} - \exp\{\Lambda_j(\hat{\beta})\} e_j\} = 0. \tag{9}$$

3. The Effect of Linkage Errors on the Observed and Expected Numbers 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 is incorrectly identified as dead, and a false negative occurs when a deceased member is considered to be alive. More specifically, for the mathematical development to follow, a false positive occurs in a particular state when an individual who remains alive throughout this state is incorrectly labelled as dead in this state. Similarly, a false negative occurs in a particular state when a member, who died before or during the sojourn in this state, is considered to be alive throughout this state. Within a particular state, false positives and false negatives thus represent special cases of misclassification error discussed by Anderson (1974, chapter 6.2.1). In this section, we will discuss the effect of these two types of 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 and incorrectly matched records.

3.1 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 labeled 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 labeled 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 labeled to have died in S_j and, similary, 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 (7) 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. Although this will underestimate the expected number of deaths, for the

purpose of studying bias, it may not be that objectionable. Assuming all the deaths to occur at the times of leaving the corresponding states also offers similar simplification. Using (8) and the decomposition of A_j^L, the expected number of deaths e_j^L in S_j the presence of linkage errors can be written as

$$\begin{aligned} e_j^L &= \sum_{i\in A_j^L} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du \\ &= \sum_{i\in A_j} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du + \sum_{i\in A_j^N} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du \\ &\qquad - \sum_{i\in D_j^P} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du \\ &= e_j - \Delta e_j, \end{aligned} \tag{10}$$

where

$$e_j = \sum_{i\in A_j} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du, \text{ and } \Delta e_j = e_j^P - e_j^N \tag{11}$$

with

$$e_j^P = \sum_{i\in D_j^P} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du \text{ and } e_j^N = \sum_{i\in A_j^N} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du. \tag{12}$$

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 (10) and (11) 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 $d_{\bar{j}\bar{j}}^L$ in the presence of linkage errors may be written as

$$d_{\bar{j}\bar{j}}^L = d_{\bar{j}\bar{j}} + \Delta d_{\bar{j}\bar{j}}, \tag{13}$$

where

$$\Delta d_{\bar{j}\bar{j}} = d_{\bar{j}\bar{j}}^P - a_{\bar{j}\bar{j}}^N, \tag{14}$$

with $d_{\bar{j}\bar{j}}$, $d_{\bar{j}\bar{j}}^P$ and $a_{\bar{j}\bar{j}}^N$ denoting the number of individuals in the sets $D_{\bar{j}\bar{j}}$, $D_{\bar{j}\bar{j}}^P$ and $A_{\bar{j}\bar{j}}^N$, respectively. The term $\Delta d_{\bar{j}\bar{j}}$ represents the difference between the observed number of deaths in the j^{th} state due to linkage errors. It follows from (13) and (14) 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 CMDB, which is generally much larger than 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.

3.2 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, for $j = 1, \ldots, J$, 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 labeled dead (alive) in S_j. That is, $p_j^P = P[i \in D_j^P \mid i \in A_j]$ and $p_j^N = P[i \in A_j^N \mid i \in D_j]$.

Let us write a_j, d_j, a_j^N and d_j^P as 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, $d_{\bar{j}\bar{j}}^P$ follows a $Binomial(a_j, p_{\bar{j}\bar{j}}^P)$ distribution, where $p_{\bar{j}\bar{j}}^P$ is the conditional probability that an individual in A_j is labeled to have died in S_j. That is, $p_{\bar{j}\bar{j}}^P = P[i \in D_{\bar{j}\bar{j}}^P \mid i \in A_j]$. Clearly, $p_{\bar{j}\bar{j}}^P \le p_j^P$. Similarly, $a_{\bar{j}\bar{j}}^N$ follows a $Binomial(d_{\bar{j}\bar{j}}, p_{\bar{j}\bar{j}}^N)$ distribution, where $p_{\bar{j}\bar{j}}^N$ is the conditional probability that an individual in $D_{\bar{j}\bar{j}}$ is labeled as alive in S_j. That is, $p_{\bar{j}\bar{j}}^N = P[i \in A_{\bar{j}\bar{j}}^N \mid i \in D_{\bar{j}\bar{j}}]$. Although there is no trivial relationship between p_j^N and $p_{\bar{j}\bar{j}}^N$ in general, it is reasonable to assume $p_j^N = p_{\bar{j}\bar{j}}^N$ in this context of linkage errors.

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 $\Delta d_{\bar{j}\bar{j}}$ in (14), are

$$E[\Delta d_{\bar{j}\bar{j}}] = E[d_{\bar{j}\bar{j}}^P] - E[a_{\bar{j}\bar{j}}^N] = a_j\, p_{\bar{j}\bar{j}}^P - d_{\bar{j}\bar{j}}\, p_j^N \tag{15}$$

and

$$\begin{aligned} V[\Delta d_{\bar{j}\bar{j}}] &= V[d_{\bar{j}\bar{j}}^P] + V[a_{\bar{j}\bar{j}}^N] \\ &= a_j\, p_{\bar{j}\bar{j}}^P (1 - p_{\bar{j}\bar{j}}^P) + d_{\bar{j}\bar{j}}\, p_j^N (1 - p_j^N). \end{aligned} \tag{16}$$

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 (11), 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 (12) and the definitions of D_j^P and A_j^N, we have

$$e_j^P = \sum_{i \in A_j} \xi_{ij'} T_\lambda(i, j) \tag{17}$$

and

$$e_j^N = \sum_{i \in D_j} \psi_{ij'} T_\lambda(i, j). \tag{18}$$

In particular, one can write $d_{jj'}^P = \sum_{i \in A_j} \xi_{ijj'}$ and $a_{jj'}^N = \sum_{i \in D_{jj'}} \psi_{ij'}$, which are useful to derive (15) and (16). From (17) and (18), we have

$$\begin{aligned} E[\Delta e_j] &= E[e_j^P] - E[e_j^N] \\ &= p_j^P \sum_{i \in A_j} T_\lambda(i, j) - p_j^N \sum_{i \in D_j} T_\lambda(i, j), \end{aligned} \tag{19}$$

and

$$\begin{aligned} V[\Delta e_j] &= V[e_j^P] + V[e_j^N] \\ &= p_j^P (1 - p_j^P) \sum_{i \in A_j} T_\lambda^2(i, j) \\ &\quad + p_j^N (1 - p_j^N) \sum_{i \in D_j} T_\lambda^2(i, j), \end{aligned} \tag{20}$$

since A_j and D_j consist of different sets of individuals.

The results (15) – (16) and (19) – (20) 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 (16) and (20) 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 5 for further discussion of this point). Although these error rates are independent of the underlying relative risk regression model γ in (2), 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.

4. The Effect of Linkage Errors on Estimates of SMRs and Regression Coefficients

4.1 Standardized Mortality Ratios

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 $\text{SMR} = \sum d_{jj'} / \sum e_j$. Letting SMR_L denote the standardized mortality ratios in the presence of linkage errors, we have

$$\text{SMR}_L = \text{SMR}\left[1 + \frac{\sum \Delta d_{jj'}}{\sum d_{jj'}}\right] \Big/ \left[1 - \frac{\sum \Delta e_j}{\sum e_j}\right]. \tag{21}$$

It follows, from (10) – (14), that the false positives will increase the SMR, whereas the false negatives will decrease the SMR.

By using a first order Taylor series approximation of SMR_L about SMR, the difference $\Delta\text{SMR} = \text{SMR}_L - \text{SMR}$ can be expressed as

$$\frac{\Delta\text{SMR}}{\text{SMR}} = \frac{\sum_j \Delta d_{jj'}}{\sum_j d_{jj'}} + \frac{\sum_j \Delta e_j}{\sum_j e_j}. \tag{22}$$

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

$$E\left[\frac{\Delta\text{SMR}}{\text{SMR}}\right] \simeq \frac{\sum_j E[\Delta d_{jj'}]}{\sum_j d_{jj'}} + \frac{\sum_j E[\Delta e_j]}{\sum_j e_j} \tag{23}$$

and

$$\begin{aligned} V\left[\frac{\Delta\text{SMR}}{\text{SMR}}\right] &\simeq \left(\sum_j d_{jj'}\right)^{-2} V\left[\sum_j \Delta d_{jj'}\right] \\ &\quad + \left(\sum_j e_j\right)^{-2} V\left[\sum_j \Delta e_j\right] \\ &\quad + 2\left(\sum_j d_{jj'}\right)^{-1}\left(\sum_j e_j\right)^{-1} \text{Cov}\left[\sum_j \Delta d_{jj'}, \sum_j \Delta e_j\right], \end{aligned} \tag{24}$$

respectively. The right hand side of (23) can be easily calculated by using (15) and (19). In order to calculate the right hand side of (24), note that

$$\begin{aligned} V\left[\sum_j \Delta d_{jj'}\right] &= \sum_j V[\Delta d_{jj'}] \\ &\quad + 2\sum_{j<j'} \text{Cov}[\Delta d_{jj'}, \Delta d_{jj'}], \end{aligned} \tag{25}$$

$$V\left[\sum_j \Delta e_j\right] = \sum_j V[\Delta e_j] + 2\sum_{j<j'} \text{Cov}[\Delta e_j, \Delta e_{j'}], \tag{26}$$

and

$$\begin{aligned} &\text{Cov}\left[\sum_j \Delta d_{jj'}, \sum_j \Delta e_j\right] \\ &= \sum_j \text{Cov}[\Delta d_{jj'}, \Delta e_j] + \sum_{j \neq j'} \text{Cov}[\Delta d_{jj'}, \Delta e_{j'}]. \end{aligned} \tag{27}$$

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

$$\mathrm{Cov}[\Delta d_{ij}, \Delta d_{ij'}] = -\left(\sum_{i \in A_j \cap A_{j'}} p_{ij}^P \, p_{ij'}^P + \sum_{i \in A_j \cap D_{j'}} p_{ij}^P \, p_{j'}^N \right), \quad (28)$$

$$\begin{aligned}\mathrm{Cov}[\Delta e_j, \Delta e_{j'}] &= \sum_{i \in A_j \cap A_{j'}} p_j^P (1 - p_{j'}^P) \, T_\lambda(i, j) \, T_\lambda(i, j') \\ &+ \sum_{i \in A_j \cap D_{j'}} p_j^P \, p_{j'}^N \, T_\lambda(i, j) \, T_\lambda(i, j') \\ &+ \sum_{i \in D_j \cap D_{j'}} p_{j'}^N (1 - p_j^N) \, T_\lambda(i, j) \, T_\lambda(i, j'), \quad (29)\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}[\Delta d_{ij}, \Delta e_j] &= \sum_{i \in A_j} p_{ij}^P (1 - p_j^P) \, T_\lambda(i, j) \\ &+ \sum_{i \in D_j} p_j^N (1 - p_j^N) \, T_\lambda(i, j), \quad (30)\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}[\Delta d_{ij}, \Delta e_{j'}] &= \sum_{i \in A_j \cap A_{j'}} p_{ij}^P (1 - p_{j'}^P) \, T_\lambda(i, j') \\ &+ \sum_{i \in A_j \cap D_{j'}} p_{ij}^P \, p_{j'}^N \, T_\lambda(i, j') \\ &+ \sum_{i \in D_{ij} \cap D_{j'}} p_{j'}^N (1 - p_j^N) \, T_\lambda(i, j'), \text{ and} \quad (31)\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}[\Delta d_{ij'}, \Delta e_j] &= - \sum_{i \in A_j \cap A_{j'}} p_j^P \, p_{ij'}^P \, T_\lambda(i, j) \\ &+ \sum_{i \in A_j \cap D_{j'}} p_j^P \, p_{j'}^N \, T_\lambda(i, j). \quad (32)\end{aligned}$$

Using (25) – (32), the variance of the relative difference $\Delta\mathrm{SMR}/\mathrm{SMR}$ can be approximated by the right hand side of (24). Two conclusions can be drawn from (23) and (24). First, linkage errors can lead to bias in the estimate of the SMR. Second, both types of linkage errors introduce additional variation into estimates of the SMR. Note that the first term in (32) is dominated by the first term in (29) for $p_{j'}^P < 0.5$, and the negative covariance term (28) is dominated in the calculation of the variance in (25). Therefore, the additional variance (24) is strictly positive, since both the false positive and false negative rates are positive.

4.2 Relative Risk Regression Parameters

To determine the effect of linkage errors on regression parameter estimates, consider first the general relative risk regression model (2). Replacing the observed and expected numbers of deaths d_{ij} and e_j in the log-likelihood function (7) with the observed and expected numbers of deaths in the presence of linkage errors d_{ij}^L and e_j^L, we have

$$\log L = \sum_{j=1}^{J} \{ d_{ij}^L \log(\gamma\{\beta' \mathbf{z}_j\}) - \gamma\{\beta' \mathbf{z}_j\} e_j^L \}. \quad (33)$$

Let $\hat{\beta}$ and $\tilde{\beta}$ denote the maximum likelihood estimates of β based on $\{d_{ij}, e_j\}$ and $\{d_{ij}^L, e_j^L\}$, respectively. The score equation (9) can be written as

$$\sum_{j=1}^{J} \frac{\partial \Lambda_j(\tilde{\beta})}{\partial \beta} [d_{ij} + \Delta d_{ij} - \exp\{\Lambda_j(\tilde{\beta})\}(e_j - \Delta e_j)] = 0. \quad (34)$$

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

$$\exp\{\Lambda_j(\tilde{\beta})\} \simeq \exp\{\hat{\Lambda}_j\} + \exp\{\hat{\Lambda}_j\} \frac{\partial \hat{\Lambda}_j}{\partial \beta} \Delta\beta, \quad (35)$$

where $\hat{\Lambda}_j = \Lambda_j(\hat{\beta})$ and $\partial\hat{\Lambda}_j/\partial\beta$ is $\partial\Lambda_j/\partial\beta$ evaluated at $\beta = \hat{\beta}$. Substituting (35) into (34) leads to

$$\begin{aligned}&\sum_{j=1}^{J} \frac{\partial \hat{\Lambda}_j}{\partial \beta} [d_{ij} - \exp\{\hat{\Lambda}_j\} e_j] \\ &+ \sum_{j=1}^{J} \frac{\partial \hat{\Lambda}_j}{\partial \beta} \left[\begin{array}{l} \Delta d_{ij} + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j \\ \quad - \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \dfrac{\partial \hat{\Lambda}_j'}{\partial \beta} \Delta\beta \\ \quad + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j \dfrac{\partial \hat{\Lambda}_j'}{\partial \beta} \Delta\beta \end{array} \right] \simeq 0. \quad (36)\end{aligned}$$

Using (9), the first summation in (36) is zero. Consequently, since $\Delta e_j \Delta\beta$ is small, $\Delta\beta$ may be approximated by

$$\Delta\beta \simeq \left(\sum_j \frac{\partial \hat{\Lambda}_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \hat{\Lambda}_j'}{\partial \beta} \right)^{-1} \sum_j \frac{\partial \hat{\Lambda}_j}{\partial \beta} \{ \Delta d_{ij} + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j \}. \quad (37)$$

It follows from (37) that

$$E[\Delta\beta] \simeq \left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right)^{-1} \sum_j \frac{\partial \Lambda_j}{\partial \beta} \alpha_j, \quad (38)$$

where $\alpha_j = E[\Delta d_{ij}] + \gamma\{\hat{\beta}' \mathbf{z}_j\} E[\Delta e_j]$, which can be calculated from (15) and (19). Further,

$$\begin{aligned}V[\Delta\beta] \simeq &\left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right)^{-1} \\ &\left(\sum_j \sum_{j'} \frac{\partial \Lambda_j}{\partial \beta} \Theta_{jj'} \frac{\partial \Lambda_{j'}'}{\partial \beta} \right) \\ &\left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right)^{-1} \quad (39)\end{aligned}$$

with

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

which can also be easily obtained using (16), (20) and (28)–(32).

In the special case of the multiplicative risk model (4), the difference $\Delta\beta$ due to linkage errors may be approximated by

$$\Delta\beta \simeq (X'WX)^{-1}X'(\Delta D + \Delta W), \tag{40}$$

where $X' = (\mathbf{z}_1', ..., \mathbf{z}_J')$, $\Delta D' = (\Delta d_{11}, ..., \Delta d_{JJ})$, $W = \mathrm{diag}(\exp(\mathbf{z}_1'\hat{\beta})e_1, ..., \exp(\mathbf{z}_J'\hat{\beta})e_J)$, and $\Delta W' = (\exp(\mathbf{z}_1'\hat{\beta})\Delta e_1, ..., \exp(\mathbf{z}_J'\hat{\beta})\Delta e_J)$. Note that the weight matrix W is the Fisher information matrix for $\hat{\beta}$. It follows from (38) that

$$E[\Delta\beta] \simeq (X'WX)^{-1}X'\Pi, \tag{41}$$

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

Further,

$$V[\Delta\beta] \simeq (X'WX)^{-1}X'\Psi X(X'WX)^{-1}, \tag{42}$$

where Ψ is the matrix of $\Theta_{jj'}$'s with $\gamma\{\hat{\beta}'\mathbf{z}_j\}$ replaced by $\exp(\mathbf{z}_j'\hat{\beta})$. Note that (40)–(42) are special cases of (37)–(39), 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,

$$\Delta\beta \simeq \frac{\sum_j \Delta d_{jj} + e^{\beta}\sum_j \Delta e_j}{e^{\beta}\sum_j e_j}. \tag{43}$$

Since the $\mathrm{SMR} = e^{\hat{\beta}} = \sum_j d_{jj}/\sum_j e_j$, with $\Delta\beta = \Delta\mathrm{SMR}/\mathrm{SMR}$ in this case, we have

$$\Delta\beta \simeq \frac{\sum_j \Delta d_{jj}}{\sum_j d_{jj}} + \frac{\sum_j \Delta e_j}{\sum_j e_j}. \tag{44}$$

Thus, (44) may be viewed as a special case of (22).

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 (39) is through $\mathrm{Cov}[\Delta d_{jj}, \Delta d_{jj'}]$, given by (28), and the first term in (32) (see $\Theta_{jj'}$). Using the same argument as in section 4.1, it follows that this additional variance is strictly positive.

5. Conclusions

Record linkage is now a well-established technique in epidemological 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 nonlink (false negative) rates associated with the linkage process. In practice, this is usually done by drawing a sample of linked and nonlinked 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 paper 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 $t_{jj}^1 = t_{jj}^0$ will be problematic if the probability of dying in this last interval is appreciable. This problem may be circumvented by assuming the human life span to have a finite upper limit.

As discussed at the end of section 3.1, 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 3 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 3. 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_j^P and p_j^N, may be written as $\alpha P[T \le \tau_j]$ and $\alpha P[T > \tau_j]$, respectively. In particular, $p_{jj}^P = \alpha P[\tau_{j-1} < T \le \tau_j]$, 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 (19) and $E[\Delta d_{jj}]$ in (15). This cancellation effect will reduce the expected bias in the SMR and the relative risk regression parameters given in (23) and (38), respectively.

Although we have considered only all-cause mortality in this article, 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 (1) and (2) 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 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 nonlinks with known probabilities have been used to further evaluate the impact of linkage errors on estimates of cancer risk (Mallick, Krewski, Dewanji and Zielinski 2002). These simulation results corroborate the theoretical findings of this paper.

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 analyses 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 analyses of cohort mortality data. Research in this area is underway.

6. Acknowledgements

This research was supported in part by a grant from the National Science and Engineering Research Council of Canada to D. Krewski, who currently holds the NSERC/SSHRC/McLaughlin Chair in Population Health Risk Assessment at the University of Ottawa. Preliminary versions of this paper were presented at the Annual Joint Meeting of the American Statistical Association in San Francisco, August 8-12, 1993, and the Annual Meeting of the Statistical Society of Canada, Montreal, July 10-16, 1995. The final draft was presented in the session in honour of J.N.K. Rao at the Statistics Canada Symposium 2001 held in Ottawa on October 18, 2001. The first author (D. Krewski) is particularly grateful to have been invited to speak in the session in honour of J.N.K. Rao, who served as his doctoral thesis supervisor many years ago. This work was completed while A. Dewanji was a Visiting Scholar at the McLaughlin Centre for Population Health Risk Assessment in the summer of 2002 and 2003.

References

Anderson, T.W. (1974). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, Inc.

Ardal, S., and Ennis, S. (2001). Data detectives: Uncovering systematic errors in administrative databases. *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada, Ottawa.

Ashmore, J.-P., and Grogan, D. (1985). The national dose registry of Canada. *Radiation Protection Dosimetry*, 11, 95-100.

Ashmore, J.-P., and Davies, B.D. (1989). The national dose registry: A centralized record keeping system for radiation workers in Canada. In *Applications of Computer Technology to Radiation Protection*, IAEA-SR-136/58, J. Stephan Institute, Ljublyua, 505-520.

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.

Ashmore, J.-P., Krewski, D., Zielinski, J.M., Jiang, H., Semenciw, R. and Létourneau, E. (1998). First analysis of occupational radiation mortality based on the national dose registry of Canada. *American Journal of Epidemiology*, 148, 564-574.

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.

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.

Breslow, N.E., Lubin, J.H. and Langholz, B. (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Association*, 78, 1-12.

Breslow, N.E., and Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. 2: *The Design and Analysis of Cohort Studies*. IARC scientific publication No. 82, international agency for research on cancer, Lyon, France.

Carpenter, M., and Fair, M.E. (Eds.) (1990). *Canadian Epidemiology Research Conference – 1989: Proceedings of Record Linkage Sessions & Workshop*. Ottawa Select Printing, Ottawa.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of Royal Statistical Society*, B, 34, 187-220.

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.

Fellegi, I., and Sunter, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

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.

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.

Howe, G.R., and Spasoff, R.A. (Eds.) (1986). *Proceeding of the Workshop on Computerized Linkage in Health Research*. University of Toronto Press, Toronto.

Jordan-Simpson, D.A., Fair, M.E. and Poliquin, C. (1990). Canadian farm operator study: Methodology. *Health Reports*, 2, 141-155.

Kalbfleish, J.D., and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons, Inc.

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.

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 International Conference in Recent Advances in Survey Sampling*. Carleton University, Ottawa, to appear.

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.

Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford Medical Publications. Oxford.

Roos, L.L., Soodeen, R. and Jebamani, L. (2001). An information-rich environment: Linked-record systems and data quality in Canada. *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Statistics Canada, Ottawa.

Scheuren, F., and Winkler, W.E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39-58.

Scheuren, F., and Winkler, W.E. (1997). Regression analysis of data files that are computer matched–Part II. *Survey Methodology*, 23, 157-165.

Singh, A.C., Feder, M., Dunteman, G. and Yu, F. (2001). Protecting confidentiality while preserving quality of public use micro data. *Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective*. Statistics Canada, Ottawa.

Smith, M.E., and Silins, J. (1981). Generalized iterative record linkage system. *Social Statistics Section, Proceedings of the American Statistical Association*, 128-137.

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.

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.