Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

THE EFFECT OF RECORD LINKAGE ERRORS ON STATISTICAL INFERENCE IN COHORT MORTALITY STUDIES

D. Krewski[1, 2, 4], Y. Wang[3], S. Bartlett[3], J.M. Zielinski[3] and R. Mallick[1,2]

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. 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 these indicators can be subject to bias and additional variability in the presence of linkage errors, with false links and nonlinks leading to positive and negative bias respectively in estimates of the standardized mortality ratio. Although linkage errors always increase the uncertainty in the estimates, bias can be effectively eliminated in the special case in which the false positive rate equals the false negative rate within homogeneous states defined by cross-classification of the covariates of interest.

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 *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 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.

[1]McLaughlin Center for Population Health Risk Assessment, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5

[2]School of Mathematics & Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

[3]Health Environments and Consumer Safety Branch, Health Canada, Ottawa, Ontario, Canada, K1A 0L2.

[4]To whom correspondence should be addressed.

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. 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). A comprehensive list of other health studies based on linking exposure data with the CMDB has been compiled by Fair (1989).

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 *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.

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

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

Let $\lambda_i(u)$ denote the hazard function for a specific cause of death at time *u* for individual $i=1,...,N$ in a cohort of size *N*, and let $z_i(u)$ 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

$$\lambda_i(u) = \lambda^*(u)\gamma\{\beta' z_i(u)\}, \tag{2}$$

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

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

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

$$\lambda_i(u) = \lambda^*(u)\exp\{\beta' z_i(u)\}. \tag{4}$$

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

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

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

$$\log L = \sum_{i=1}^{N}\left\{\delta_i \log\left(\gamma\{\beta' z_i(t_i^1)\}\right) - \int_{t_i^0}^{t_i^1}\gamma\{\beta' z_i(u)\}\lambda^*(u)du\right\}. \tag{6}$$

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 *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 $\{S_j; j=1,...,J\}$ such that $z_i(u) = z_j$ 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

$$\log L = \sum_{j=1}^{J} \{ d_j \log(\gamma \{\beta' z_j\}) - \gamma \{\beta' z_j\} e_j \}, \tag{7}$$

where

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

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

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

3. THE EFFECT OF LINKAGE ERRORS ON THE OBSERVED AND EXPECTED NUMBER OF DEATHS

Two 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. A false negative occurs when a deceased subject is still considered to be alive. In this section we will discuss the effect 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 records and incorrectly matched records.

3.1 Linkage errors

Let C_j denote the set of labels for those individuals who pass through state S_j. Assuming no linkage errors, let A_j and D_j represent the sets of labels of those cohort members who remain alive throughout state S_j, and those cohort members who died in state S_j, respectively. Note that $A_j \cup D_j = C_j$ and $A_j \cap D_j = \{\emptyset\}$.

Let A_j^L and D_j^L denote the sets A_j and D_j in the presence of linkage errors. We further define D_j^P and A_j^N as the index sets for false positives and false negatives in the *j*th state S_j. These sets satisfy the relations $A_j^L = (A_j - D_j^P) \cup A_j^N$ and $D_j^L = (D_j - A_j^N) \cup D_j^P$, where $A_j \cup D_j = C_j$ and $A_j^L \cap D_j^L = \{\emptyset\}$.

The effect of linkage errors on the likelihood function in (7) may be described as follows. Let t_{ij}^0, t_{ij}^1, and t_{ij}^2 denote the time at which the *i*th individual enters, dies within (should death occur), and leaves the *j*th state S_j, respectively. Using (8) and the decompositions of A_j^L and D_j^L, the expected number of deaths e_j^L in the *j*th state in the presence of linkage errors can be written as

$$
\begin{aligned}
e_j^L &= \sum_{i \in A_j^L} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du + \sum_{i \in D_j^L} \int_{t_{ij}^0}^{t_{ij}^1} \lambda^*(u)du \\
&= \left(\sum_{i \in A_j} \int_{t_{ij}^0}^{t_{ij}^2} + \sum_{i \in A_j^N} \int_{t_{ij}^0}^{t_{ij}^2} - \sum_{i \in D_j^P} \int_{t_{ij}^0}^{t_{ij}^2} \right) \lambda^*(u)du + \\
&\quad \left(\sum_{i \in D_j} \int_{t_{ij}^0}^{t_{ij}^1} + \sum_{i \in D_j^P} \int_{t_{ij}^0}^{t_{ij}^1} - \sum_{i \in A_j^N} \int_{t_{ij}^0}^{t_{ij}^1} \right) \lambda^*(u)du \qquad (10) \\
&= \sum_{i \in A_j} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du + \sum_{i \in D_j} \int_{t_{ij}^0}^{t_{ij}^1} \lambda^*(u)du + \sum_{i \in A_j^N} \int_{t_{ij}^1}^{t_{ij}^2} \lambda^*(u)du - \sum_{i \in D_j^P} \int_{t_{ij}^1}^{t_{ij}^2} \lambda^*(u)du \\
&= e_j - \Delta e_j,
\end{aligned}
$$

where

$$
e_j = \sum_{i \in A_j} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du + \sum_{i \in D_j} \int_{t_{ij}^0}^{t_{ij}^1} \lambda^*(u)du \qquad (11)
$$

and

$$
\Delta e_j = e_j^P - e_j^N \qquad (12)
$$

with

$$
e_j^P = \sum_{i \in D_j^P} \int_{t_{ij}^1}^{t_{ij}^2} \lambda^*(u)du \qquad (13)
$$

and

$$
e_j^N = \sum_{i \in A_j^N} \int_{t_{ij}^1}^{t_{ij}^2} \lambda^*(u)du. \qquad (14)
$$

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 (12) that false positives tend to reduce the expected number of deaths, and that the false negatives tend to increase the expected number of deaths.

Using the decomposition for D_j^L, the observed number of deaths d_j^L in the presence of linkage errors may be written as

$$
d_j^L = d_j + \Delta d_j, \qquad (15)
$$

where

$$
\Delta d_j = d_j^P - d_j^N. \qquad (16)
$$

Here, $d_j = \dim\left(D_j\right)$, $d_j^P = \dim\left(D_j^P\right)$, and $d_j^N = \dim\left(A_j^N\right)$, with *dim* (*A*) denoting the number of elements in the set *A*. The term Δd_j represents the bias in the observed number of deaths in the *j*th state due to linkage errors. It follows from (15) and (16) that false positives will increase the observed number of deaths, and that 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 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.

3.2 Expectations and variances of biases in 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 D_{ij} $\left(\overline{D}_{ij}\right)$ denote the event that the *i*th individual dies (does not die) in the *j*th state, and L_{ij} $\left(\overline{L}_{ij}\right)$ denote the event that the individual is labelled as having died (not died) following linkage with the mortality database. Following Fellegi and Sunter (1969), the false positive rate is defined as the conditional probability $p_{ij}^{I} = \Pr\left\{L_{ij}\middle|\overline{D}_{ij}\right\}$. Similarly, the false negative rate is $p_{ij}^{II} = \Pr\left\{\overline{L}_{ij}\middle|D_{ij}\right\}$. We assume for simplicity that the false positive and false negative rates are constant within each state $j = 1,\ldots,J$. Specifically, $p_{ij}^{I} = p_j^{I}$ and $p_{ij}^{II} = p_j^{II}$ for $i \in C_j$. This assumption will be appropriate 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.

Let ξ_{ij} be an indicator variable, with $\xi_{ij} = 1$ corresponding to a false positive related to the *i*th subject within the *j*th state and $\xi_{ij} = 0$ otherwise. Similarly, let $\psi_{ij} = 1$ denote a false negative associated with the same subject, and $\psi_{ij} = 0$ otherwise. Assuming that linkage errors related to different subjects are independent, the expectation and variance of the bias in the observed number of deaths in the *j*th state Δd_j are

$$E\left\{d_j^P - d_j^N\right\} = E\left\{\sum_{i \in C_j} \xi_{ij} - \sum_{i \in C_j} \psi_{ij}\right\} = \sum_{i \in C_j} \Delta p_{ij} = \Delta p_j N_j \qquad (17)$$

and

$$\begin{aligned}\mathrm{Var}\left\{d_j^P - d_j^N\right\} &= \sum_{i\in C_j}\mathrm{Var}\left\{\xi_{ij}\right\} + \sum_{i\in C_j}\mathrm{Var}\left\{\psi_{ij}\right\} - \sum_{i\in C_j}\mathrm{Cov}\left\{\xi_{ij},\psi_{ij}\right\} \\ &= \sum_{i\in C_j}\left\{p_{ij}^I\left(1-p_{ij}^I\right) + p_{ij}^{II}\left(1-p_{ij}^{II}\right)\right\} \\ &= \left\{p_j^I\left(1-p_j^I\right) + p_j^{II}\left(1-p_j^{II}\right)\right\}N_j,\end{aligned} \tag{18}$$

where $\Delta p_{ij} = p_{ij}^I - p_{ij}^{II}$ and $\Delta p_j = p_j^I - p_j^{II}$. Note that since a false positive and a false negative cannot occur simultaneously,

$$\mathrm{Cov}\left\{\xi_{ij},\psi_{ij}\right\} = \mathrm{E}\left\{\left(\xi_{ij}\,\psi_{ij}\right)^2\right\} - \left(\mathrm{E}\left\{\xi_{ij}\,\psi_{ij}\right\}\right)^2 = 0. \tag{19}$$

Similarly, the expectation and variance of the bias in the expected number of deaths in the *j*th state Δe_j can be expressed as

$$\mathrm{E}\left\{e_j^P - e_j^N\right\} = \Delta p_j \sum_{i\in C_j}\int_{\min\left(t_{ij}^1,t_{ij}^2\right)}^{t_{ij}^z}\lambda^*(u)\,du \tag{20}$$

and

$$\mathrm{Var}\left\{e_j^P - e_j^N\right\} = \left\{p_j^I\left(1-p_j^I\right) + p_j^{II}\left(1-p_j^{II}\right)\right\}\sum_{i\in C_j}\left\{\int_{\min\left(t_{ij}^1,t_{ij}^2\right)}^{t_{ij}^2}\lambda^*(u)\,du\right\}^2, \tag{21}$$

respectively. These results indicate that if the false positive and false negative rates are equal in every state, the observed number of deaths will, on average, be equal to the actual number of deaths. In this case, the bias in the expected number of deaths caused by linkage errors will vanish. However, both types of linkage errors will introduce additional variation into the observed and expected numbers of deaths.

Minimizing the variance of the bias in (21) is difficult since p_j^I and p_j^{II} are not functionally independent. Generally, decreasing p_j^I will result in an increase in p_j^{II} and vice versa. Nevertheless, the multiplicative factor $\left\{p_j^I\left(1-p_j^I\right) + p_j^{II}\left(1-p_j^{II}\right)\right\}$ in (21) could be minimized given the functional relationship between p_j^I and p_j^{II}. Although this factor is independent of the underlying relative risk regression model γ in (2), the mean square error obtained by combining (20) and (21) cannot be minimized without specification of the baseline hazard $\lambda^*(u)$.

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 observed and expected numbers of deaths d_j and e_j assuming no linkage errors with the observed and expected number of deaths d_j^L and e_j^L given that there are linkage errors in the expression $\mathrm{SMR} = \sum d_j / \sum e_j$. Letting SMR_L denote the standardized

mortality ratios in the presence of linkage errors, we have

$$\mathrm{SMR}_{\mathrm{L}} = \mathrm{SMR}\left[1+\frac{\sum\Delta\,\mathrm{d}_{\mathrm{j}}}{\sum\mathrm{d}_{\mathrm{j}}}\right]/\left[1-\frac{\sum\Delta\,\mathrm{e}_{\mathrm{j}}}{\sum\mathrm{e}_{\mathrm{j}}}\right]. \tag{22}$$

It follows that false positives will increase the SMR, whereas false negatives will decrease the SMR.

The difference $\Delta\mathrm{SMR} = \mathrm{SMR}_{\mathrm{L}} - \mathrm{SMR}$ can be expressed as

$$\frac{\Delta\,SMR}{SMR} = \frac{\sum\Delta\,d_j}{\sum d_j} + \frac{\sum\Delta\,e_j}{\sum e_j} + o\left(\frac{\sum\Delta\,e_j}{\sum e_j}\right). \tag{23}$$

In most cohort studies, the number of deaths is much smaller than the size of the whole cohort, in which case the error term in (23) is small. The mean and variance of the relative bias in the SMR can thus be approximated by

$$E\left\{\frac{\Delta\,SMR}{SMR}\right\} \approx \sum\Delta\,p_j\left\{\left(\sum d_j\right)^{-1} N_j + \left(\sum e_j\right)^{-1}\sum_{i\in C_j}\int_{\min\left(t_{ij}^1,t_{ij}^2\right)}^{t_{ij}^2}\lambda^*(u)du\right\} \tag{24}$$

and

$$\begin{aligned}\mathrm{Var}\left\{\frac{\Delta\mathrm{SMR}}{\mathrm{SMR}}\right\} \approx{}& \sum\left\{p_j^{\mathrm{I}}\left(1-p_j^{\mathrm{I}}\right)+p_j^{\mathrm{II}}\left(1-p_j^{\mathrm{II}}\right)\right\}\left\{\left(\sum\mathrm{d}_{\mathrm{j}}\right)^{-2}\mathrm{N}_{\mathrm{j}}+\right.\\ &\left.+2\left(\sum\mathrm{d}_{\mathrm{j}}\right)^{-1}\left(\sum\mathrm{e}_{\mathrm{j}}\right)^{-1}\sum_{\mathrm{i}\in\mathrm{C}_{\mathrm{j}}}\int_{\min\left(t_{ij}^1,t_{ij}^2\right)}^{t_{ij}^2}\lambda^*(\mathrm{u})\,\mathrm{du}+\left(\sum\mathrm{e}_{\mathrm{j}}\right)^{-2}\sum_{\mathrm{i}\in\mathrm{C}_{\mathrm{j}}}\left(\int_{\min\left(t_{ij}^1,t_{ij}^2\right)}^{t_{ij}^2}\lambda^*(\mathrm{u})\,\mathrm{du}\right)^2\right\}\end{aligned} \tag{25}$$

respectively.

Three conclusions can be drawn from (24) and (25). First, if the false positive rate is greater (less) than the false negative rate, then the SMR will, on average, be overestimated (underestimated). Second, when the false positive and false negative rates are equal in every state, the biases in the SMR caused by linkage errors effectively cancel. Third, both types of linkage errors introduce additional variation into estimates of the SMR.

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_{j} and e_{j} in the log-likelihood function (7) with the observed and expected numbers of deaths given that there are linkage errors $\mathrm{d}_{\mathrm{j}}^{\mathrm{L}}$ and $\mathrm{e}_{\mathrm{j}}^{\mathrm{L}}$, we have

$$\log\mathrm{L} = \sum_{\mathrm{j}=1}^{\mathrm{J}}\left\{\mathrm{d}_{\mathrm{j}}^{\mathrm{L}}\log\left(\gamma\left\{\beta'\,\mathbf{z}_{\mathrm{j}}\right\}\right)-\gamma\left\{\beta'\,\mathbf{z}_{\mathrm{j}}\right\}\mathrm{e}_{\mathrm{j}}^{\mathrm{L}}\right\}. \tag{26}$$

Let $\hat{\beta}$ and $\tilde{\beta}$ denote the mle's of β based on $\{d_j, e_j\}$ and $\{d_j^L, e_j^L\}$ respectively. Letting $\hat{\Lambda}_j = \Lambda_j(\hat{\beta})$, the score function (9) can be written as

$$\sum_{j=1}^{J} \frac{\partial \Lambda_j(\tilde{\beta})}{\partial \beta} \{ d_j + \Delta d_j - \exp\{\hat{\Lambda}_j + \Delta \Lambda_j\}(e_j - \Delta e_j) \} = 0 \tag{27}$$

where $\Delta \Lambda_j = \Lambda(\tilde{\beta}) - \Lambda(\hat{\beta})$. 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

$$\exp\{\hat{\Lambda}_j + \Delta \Lambda_j\} \approx \exp\{\hat{\Lambda}_j\} + \exp\{\hat{\Lambda}_j\} \frac{\partial \hat{\Lambda}_j}{\partial \beta} \Delta \beta . \tag{28}$$

Substituting (28) into (27) leads to

$$\begin{aligned} &\sum_{j=1}^{J} \frac{\partial \Lambda_j(\hat{\beta})}{\partial \beta} [d_j - \exp\{\hat{\Lambda}_j\} e_j] + \sum_{j=1}^{J} \frac{\partial \Lambda_j(\hat{\beta})}{\partial \beta} [\Delta d_j + \gamma \{\hat{\beta}' z_j\} \Delta e_j - \\ &- \gamma \{\hat{\beta}' z_j\} e_j \frac{\partial \Lambda_j'(\hat{\beta})}{\partial \beta} \Delta \beta + \gamma \{\hat{\beta}' \mathbf{z}_j\} \Delta e_j \frac{\partial \Lambda_j'(\hat{\beta})}{\partial \beta} \Delta \beta] \approx 0. \end{aligned} \tag{29}$$

Using (9), the first summation in (29) is zero. Consequently, since $\Delta e_j \Delta \beta$ is small, $\Delta \beta$ may be approximated by

$$\Delta \beta \approx \left\{ \sum \frac{\partial \hat{\Lambda}_j}{\partial \beta} \gamma \{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \hat{\Lambda}_j'}{\partial \beta} \right\}^{-1} \sum \frac{\partial \hat{\Lambda}_j}{\partial \beta} \{ \Delta d_j + \gamma \{\hat{\beta}' \mathbf{z}_j\} \Delta e_j \}. \tag{30}$$

It follows from (30) that

$$E\{\Delta \beta\} \approx \left\{ \sum \frac{\partial \Lambda_j}{\partial \beta} \gamma \{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right\}^{-1} \sum \frac{\partial \Lambda_j}{\partial \beta} \Delta p_j \alpha_j , \tag{31}$$

where

$$\alpha_j = N_j + \gamma \{\hat{\beta}' \mathbf{z}_j\} \sum_{i \in C_j} \int_{\min(t_{ij}^1, t_{ij}^2)}^{t_{ij}^2} \lambda^*(u) du. \tag{32}$$

Further,

$$\mathrm{Var}\{\Delta\beta\} \approx \left\{ \sum \frac{\partial \Lambda_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right\}^{-1} \left[\sum \frac{\partial \Lambda_j}{\partial \beta} \Theta_j \frac{\partial \Lambda_j'}{\partial \beta} \right] \left\{ \sum \frac{\partial \Lambda_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right\}^{-1}, \tag{33}$$

with

$$\Theta_j = \{p_j^I (1-p_j^I) + p_j^{II}(1-p_j^{II})\} \sum_{i \in C_j} \left\{ 1 + \gamma\{\hat{\beta}' z_j\} \int_{\min(t_{ij}^1, t_{ij}^2)}^{t_{ij}^2} \lambda^*(u) du \right\}^2. \tag{34}$$

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

$$\Delta\beta \approx (X' W X)^{-1} X' (\Delta D + \Delta W), \tag{35}$$

where $X' = (\mathbf{z}_1', ..., \mathbf{z}_J')$, $\Delta D' = (\Delta d_1, ..., \Delta d_J)$, $W = \mathrm{diag}\left(\exp(\mathbf{z}_1'\hat{\beta}) e_1, ..., \exp(\mathbf{z}_J'\hat{\beta}) e_J\right)$, and $\Delta W' = \left(\exp(\mathbf{z}_1'\hat{\beta})\Delta e_1, ..., \exp(\mathbf{z}_J'\hat{\beta})\Delta e_J\right)$. Note that the weight matrix *W* is the Fisher information matrix for $\hat{\beta}$. It follows from (35) that

$$E\{\Delta\beta\} \approx (X' W X)^{-1} X' \Pi, \tag{36}$$

where $\Pi' = (\pi_1, ..., \pi_J)$ with

$$\pi_j = \Delta p_j \left\{ N_j + \exp(\mathbf{z}_j'\hat{\beta}) \sum_{i \in C_j} \int_{\min(t_{ij}^1, t_{ij}^2)}^{t_{ij}^2} \lambda^*(u) du \right\}. \tag{37}$$

Further,

$$\mathrm{Var}\{\Delta\beta\} \approx (X' W X)^{-1} X' \Psi X (X' W X)^{-1}, \tag{38}$$

where $\Psi = \mathrm{diag}(\psi_1, ..., \psi_J)$ with diagonal elements

$$\psi_j = \{p_j^I (1-p_j^I) + p_j^{II}(1-p_j^{II})\} \sum_{i \in C_j} \left\{ 1 + \exp(\mathbf{z}_j'\hat{\beta}) \int_{\min(t_{ij}^1, t_{ij}^2)}^{t_{ij}^2} \lambda^*(u) du \right\}^2. \tag{39}$$

With a single covariate $z_i = 1$, $X' W X = e^{\hat{\beta}} \Sigma e_j$, $X' \Delta D = \Sigma d_j$ and $X' \Delta W = e^{\hat{\beta}} \Sigma \Delta e_j$. In this case,

$$\Delta\beta \approx \left(\sum \Delta d_j + e^{\hat{\beta}} \sum \Delta e_j\right) / \left(e^{\hat{\beta}} \sum e_j\right). \tag{40}$$

Since

$$\mathrm{SMR} = e^{\hat{\beta}} = \frac{\Sigma\, d_j}{\Sigma\, e_j}, \tag{41}$$

with $\Delta\beta = \Delta \mathrm{SMR} / \mathrm{SMR}$ in this case, we have

$$\Delta\beta \approx \frac{\Sigma\, \Delta d_j}{\Sigma d_j} + \frac{\Sigma\, \Delta e_j}{\Sigma e_j}. \tag{42}$$

Thus, (42) may be viewed as a special case of (23).

The preceding results indicate that when the false positive and false negative rates are equal in every state, biases in estimates of regression parameters due to linkage errors will be nearly eliminated. However, both false positives and false negatives will introduce additional variation into estimates of relative risk regression parameters.

5. CONCLUSIONS

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 also shown to introduce bias into estimates of SMRs, with false links and false nonlinks generally leading to positive and negative bias respectively. Estimation of relative risk regression coefficients are also subject to bias, although the direction of the bias 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.

Biases in estimates of the SMR can be nearly eliminated if the false positive and false negative rates are equal within each state defined by cross-classification of the covariates of interest, provided that all covariates assume constant values within each state. These are stringent conditions which are unlikely to be satisfied in practice. Nevertheless, our results suggest that it is desirable to establish linkage thresholds that balance the false positive and false negative rates separately within homogeneous states.

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 are currently conducting 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 will be used to further evaluate the impact of linkage errors on estimates of cancer risk.

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 also underway.

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.

REFERENCES

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 (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,* B34, 187-220.

Efron, B. (1982), "The Jackknife, the Bootstrap and Other Resampling Plans", *Society for Industrial and Applied Mathematics,* Philadelphia, Pennsylvania, 13-19.

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, Wiley.

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.

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.

Scheuren, F. and Winkler, W.E. (1993), "Regression Analysis of Data Files that are Computer Matched", *Survey Methodology,* 19, 39-58.

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.