STATISTICS IN TRANSITION new series, September 2016
Vol. 17, No. 3, pp. 429–447

PREDICTION OF A FUNCTION OF MISCLASSIFIED BINARY DATA

Noriah M. Al-Kandari[1], **Partha Lahiri**[2]

ABSTRACT

We consider the problem of predicting a function of misclassified binary variables. We make an interesting observation that the naive predictor, which ignores the misclassification errors, is unbiased even if the total misclassification error is high as long as the probabilities of false positives and false negatives are identical. Other than this case, the bias of the naive predictor depends on the misclassification distribution and the magnitude of the bias can be high in certain cases. We correct the bias of the naive predictor using a double sampling idea where both inaccurate and accurate measurements are taken on the binary variable for all the units of a sample drawn from the original data using a probability sampling scheme. Using this additional information and design-based sample survey theory, we derive a bias-corrected predictor. We examine the cases where the new bias-corrected predictors can also improve over the naive predictor in terms of mean square error (MSE).

Key words: binary classification, double sampling, finite population sampling, misclassification, linkage error, sampling design.

1. Introduction

In many disciplines, misclassified binary data are frequently encountered. For example, in device testing, Zhong (2002) studied the specificity and sensitivity of an inaccurate diagnostic test along with a gold standard. Stamey *et al.* (2007) proposed a Bayesian estimation of an intervention effect with pre and post misclassified binomial data. Lyles *et al.* (2004) discussed single-armed studies with misclassification of a repeated binary outcome. In epidemiology and medical studies, there are plenty of examples of misclassified binary data. For example, in studying the relationship between low level radiation exposure and cancer death rate using the Cox proportional hazard model, Krewski *et al.* (2005) noted that misclassified binary data arise in form of imperfect linkages caused by the computerized record linkage method.

Bross (1954) was probably the first to observe that classical estimators of the odds ratio can be heavily biased if the misclassification error in binary data is ignored; see Goldberg (1975) for a follow-up study. Neter *et al.* (1965) noticed that

[1]Department of Statistics and Operations Research, Kuwait University.
E-mail: noriah@stat.kuniv.edu.
[2]Joint Program in Survey Methodology, University of Maryland. E-mail: plahiri@umd.edu.

the matching errors pose an obstacle to the usefulness and correct interpretation of record checks.

There are mainly two different approaches available to correct for the bias in statistical procedures that arise from misclassified binary data. The key ingredient in both the approaches is to use additional data to deal with the identifiability problem. The first approach, pioneered by Tenenbein (1970), employs a double sampling scheme in which a training data set is collected and the binary responses are measured by an accurate instrument (in the case of a random subsample from the original data) or by both an accurate instrument and the same inaccurate instrument used to collect the original data (in the case of an independent new sample). An accurate instrument results in error-free but expensive binary data. On the other hand, an inaccurate instrument results in misclassified but relatively less expensive binary data. Tenenbein's idea is intuitive and uses both accurate and inaccurate procedures to yield not only model identifiability but also economical viability.

For the single proportion problem, when a training data is obtained using a double sampling scheme, Tenenbein (1970) proposed a maximum likelihood estimator and derived its asymptotic variance. Boese *et al.* (2006) constructed several likelihood-based confidence intervals for a proportion using data subject to only false positive misclassification. Rahardja and Zhou (2013) proposed a modification of the Wald test in presence of misclassified binary data and applied their test to traffic data. Rahardja and Yang (2015) constructed two likelihood-based confidence intervals for a binomial proportion parameter using a double-sampling scheme with misclassified binary data.

When an accurate instrument is unavailable or prohibitively expensive but certain data related to the cause of misclassification are available, one can develop an identifiable model in an attempt to correct for the misclassification bias in the estimators and predictors. For the single proportion problem using misclassified data with no training data, Gaba and Winkler (1992) and Viana *et al.* (1993) developed Bayesian approaches with highly informative priors. Bayesian inferences with informative priors were also developed for two-sample problems for two proportions. For example, see Evans *et al.* (1996) for risk difference (the difference of two proportions) and Gustafson *et al.* (2001) for odds ratios. Lahiri and Larsen (2005) used a mixture model to correct for the bias of the ordinary least square estimators of regression coefficients due to imperfect linkages.

In this paper, we assume the existence of a training sample such as the one proposed by Tenenbien (1970) and exploit a design-based sample survey approach to predict a function of misclassified binary data. Consider a set U of N units. For unit $i \in U$, we define a binary variable δ_i taking on values 0 and 1, and a $K \times 1$ vector of measurements $\mathbf{y}_i = (y_{i1}, \cdots, y_{iK})\prime$. We consider a situation when we do not observe δ_i, but instead observe a predictor $\hat{\delta}_i$ subject to a misclassification error $e_i = \hat{\delta}_i - \delta_i$ $(i \in U)$. In this paper, we are interested in the prediction of $\mathbf{Y}_\delta =$

$\sum_{i\in U}\delta_i\mathbf{y}_i$ or a non-linear function of the components of $\mathbf{Y}_\delta$, say $f(Y_{\delta 1},\cdots,Y_{\delta K})$, where $Y_{\delta k}=\sum_{i=1}^{N}\delta_i y_{ik}$, $(k=1,\cdots,K)$, based on data $\{(\hat{\delta}_i,\mathbf{y}_i),\ i\in S\subseteq U\}$.

A natural predictor of $\mathbf{Y}_\delta$ is given by $\mathbf{Y}_{\hat{\delta}}(S)=\sum_{i\in S}\hat{\delta}_i\mathbf{y}_i$. If additional data that explain the mechanism for misclassification errors e_i are available, it is possible to correct $\mathbf{Y}_{\hat{\delta}}(S)$ for bias due to the misclassification errors. The misclassification errors could arise due to a variety of reasons. For example, the data set may be obtained by merging two or more data sets using a computerized record linkage method, which may introduce misclassification errors due to incorrect linkages. There are a large number of papers available in the literature that provide valid inferences under linkage errors when data on linkage error mechanism through matching weights are available; for a review of record linkage methodology, see Felligi and Sunter (1969) and Herzog *et al.* (2007). However, in this paper, we assume that we do not have any data that explain the misclassification error for all records in $i\in S$. Thus, even for the special case when the misclassification errors is due to incorrect linkages, in this paper we deal with a situation that cannot be handled by a regular record linkage methodology such as the ones given in Scheuren and Winkler (1993) and Lahiri and Larsen (2005).

In Section 2.1, we first obtain the bias and mean squared error of the natural predictor for the scaler case $K=1$ and then generalize to $K\geq 1$. The extent of the bias clearly depends on the misclassification error distribution. When the positive and negative misclassification errors are equally likely, $Y_{\hat{\delta}}(S)\equiv Y_{\hat{\delta}}$ turns out to be an unbiased predictor of Y_δ. This is an interesting observation. We propose a method for correcting the bias in the general case by drawing a probability sample and then obtaining the misclassification errors e_i for all the units in the sample. Using this extra information, we propose a bias-corrected predictor of Y_δ. We obtain an exact expression for the MSE of the proposed bias-corrected predictor that incorporates both the sampling and misclassification errors. We also propose an estimator of the MSE of the new predictor. In Section 2.2, we discuss the estimation of relative risk estimation as an illustration of the methodology proposed in Section 2.1. In Section 2.3, we evaluate the method proposed in Section 2.1 using a numerical example. In Section 3, we consider the case $S\subseteq U$. Finally, some conclusions are presented in Section 4.

2. Prediction of $\mathbf{Y}_\delta$ when $S=U$

2.1. The Methodology

For the simplicity of exposition, we first consider the scaler case, i.e., $K=1$. We define the bias and mean squared error (MSE) of $\hat{Y}_\delta$ as follows:

$$\begin{aligned}\text{Bias}_M(\hat{Y}_\delta) &= \mathtt{E}_M(\hat{Y}_\delta-Y_\delta),\\ \text{MSE}_M(\hat{Y}_\delta) &= \mathtt{Var}_M(\hat{Y}_\delta-Y_\delta),\end{aligned}$$

Table 1. The probability distribution of the misclassification error e_i (probabilities are given within parenthesis)

	δ_i	
$\hat{\delta}_i$	0	1
0	0 (p_{i00})	-1 (p_{i01})
1	1 (p_{i10})	0 (p_{i11})

where E_M and Var_M denote the expectation and variance with respect to a misclassification model described by the two-way table given in Table 1.

For unit $i \in U$, the table displays the misclassification error distribution, where $p_{i11} + p_{i10} + p_{i01} + p_{i00} = 1$. We call p_{i10} and p_{i01} false positive and false negative probabilities, respectively. We say that we have *high*, *moderate* and *low* linkage errors if $p_{i10} + p_{i01} = p_{i;T}$ (say) is close to 1, 0.5 and 0, respectively.

Theorem 1. Under the misclassification model given in Table 1, we have

$$\begin{aligned} \text{(i) } \mathrm{Bias}_M(\hat{Y}_\delta) &= \sum_{i\in U} y_i p_{i;D} = Y_{p_D} \text{ (say)}, \\ \text{(ii) } \mathrm{MSE}_M(\hat{Y}_\delta) &= \sum_{i\in U} \left[p_{i;T} - p_{i;D}^2\right] y_i^2, \end{aligned}$$

where $p_{i;D} = p_{i10} - p_{i01}$ and $p_{i;T} = p_{i10} + p_{i01}$ $(i \in U)$.

Proof: First note that $Y_{\hat{\delta}} - Y_\delta = \sum_{i\in U} y_i e_i = Y_e$, (say). Under the misclassification model, we have $\mathrm{E}_M(e_i) = p_{i;D}$ and $\mathrm{E}_M(e_i^2) = p_{i;T}$. Thus, part (i) follows immediately. To prove part (ii), using $\mathrm{Cov}_M(e_i, e_j) = 0$ $(i \neq j \in U)$, we have

$$\begin{aligned} \mathrm{MSE}_M(Y_{\hat{\delta}}) &= \mathrm{E}_M\left(\sum_{i\in U} y_i e_i\right)^2 - \left[\mathrm{E}_M(\sum_{i\in U} y_i e_i)\right]^2 \\ &= \mathrm{E}_M\left(\sum_{i\in U} y_i^2 e_i^2 + \sum_{i\neq j} y_i y_j e_i e_j\right) - \left(\sum_{i\in U} y_i p_{i;D}\right)^2 \\ &= \sum_{i=1}^{N} y_i^2 p_{i;T} + \sum_{i\neq j} y_i y_j p_{i;D} p_{j;D} - \left(\sum_{i=1}^{N} y_i p_{i;D}\right)^2. \end{aligned}$$

Part (ii) now follows using algebra.

Throughout the paper, we assume that we do not have any additional data that explain the misclassification errors e_i for all units in U. Thus, we propose to draw a sample s_1 of size n from U, using a probability sampling scheme. For each unit in the sample, we assume that we can obtain e_i with some extra effort. Let $\pi_i = \Pr(s_1 \ni i)$ denote the first-order inclusion probability of unit $i \in U$. We propose to

estimate Y_δ by $\hat{Y} = Y_{\hat{\delta}} - \hat{Y}_{\pi^{-1}e}$, where $\hat{Y}_{\pi^{-1}e} = \sum_{i\in s_1} \pi_i^{-1} e_i y_i$. The following theorem shows that $\hat{Y}$ is an unbiased predictor of Y_δ. Moreover, the theorem provides an expression for the total MSE of $\hat{Y}$, where total MSE incorporate errors due to both the misclassification and sampling errors.

Theorem 2. Under the sampling design and misclassification model, we have

$$\begin{aligned}
&\text{(i) Bias}(\hat{Y}) = 0,\\
&\text{(ii) MSE}(\hat{Y}) = \sum_{i\in U}\sum_{j>i\in U} (\pi_i\pi_j - \pi_{ij})\psi_{ij},\\
&\text{(iii) E}\left[mse(\hat{Y})\right] = \text{MSE}(\hat{Y}),
\end{aligned}$$

where $\pi_{ij} = \mathtt{Pr}(s_1 \ni \{ij\})$, the second-order inclusion probability, $\psi_{ij} = \pi_i^{-2} y_i^2 p_{iT} + \pi_j^{-2} y_j^2 p_{jT} - 2(\pi_i\pi_j)^{-1} y_i y_j p_{i;D} p_{j;D}$, $\text{mse}(\hat{Y}) = \sum_{i\in s_1}\sum_{j>i\in s_1} \pi_{ij}^{-1}(\pi_i\pi_j - \pi_{ij})\left(\pi_i^{-1} y_i e_i - \pi_j^{-1} y_j e_j\right)^2$.

Proof: Let $\mathtt{E}_d$ and $\mathtt{Var}_d$ denote the expectation and variance with respect to the sample design. First note that $\mathtt{E}_d\left(\hat{Y}_{\pi^{-1}e}\right) = Y_e = Y_{\hat{\delta}} - Y_\delta$, since $\hat{Y}_{\pi^{-1}e}$ is the Horvitz-Thompson estimator of Y_e. To prove part (i) of Theorem 2, note that

$$\begin{aligned}
\text{Bias}(\hat{Y}) &= \mathtt{E}(\hat{Y} - Y_\delta) = \mathtt{E}(Y_{\hat{\delta}} - \hat{Y}_{\pi^{-1}e} - Y_\delta)\\
&= \mathtt{E}_M E_d(Y_{\hat{\delta}} - \hat{Y}_{\pi^{-1}e} - Y_\delta)\\
&= \mathtt{E}_M(Y_{\hat{\delta}} - Y_e - Y_\delta)\\
&= \mathtt{E}_M(0)\\
&= 0,
\end{aligned}$$

since $Y_{\hat{\delta}} - Y_\delta = 0$. To prove part (ii), we first apply the iterated formula for variance to obtain

$$\begin{aligned}
\text{MSE}(\hat{Y}) &= \mathtt{Var}(\hat{Y} - Y_\delta)\\
&= \mathtt{E}_M\mathtt{Var}_d(\hat{Y} - Y_\delta) + \mathtt{Var}_M\mathtt{E}_d(\hat{Y} - Y_\delta)\\
&= \mathtt{E}_M\mathtt{Var}_d(Y_{\hat{\delta}} - \hat{Y}_{\pi^{-1}e} - Y_\delta) + \mathtt{Var}_M\mathtt{E}_d(Y_{\hat{\delta}} - \hat{Y}_{\pi^{-1}e} - Y_\delta)\\
&= \mathtt{E}_M\mathtt{Var}_d(\hat{Y}_{\pi^{-1}e}) + \mathtt{Var}_M(Y_{\hat{\delta}} - Y_e - Y_\delta)\\
&= \mathtt{E}_M\left[\sum_{i\in U}\sum_{j>i\in U} (\pi_i\pi_j - \pi_{ij})\frac{y_i e_i y_j e_j}{\pi_i\pi_j}\right],
\end{aligned}$$

since $Y_{\hat{\delta}} - Y_e - Y_\delta = 0$. Now, part (ii) follows using $\mathtt{E}_M(e_i) = p_{i;D}$, $\mathtt{E}_M(e_i^2) = p_{i;T}$, and $\text{Cov}_M(e_i, e_j) = 0$, $(i \neq j)$, and algebra. Part (iii) follows using the iterated formula for expectation and the design-unbiasedness of the well-known Yates-Grundy estimator, Yates and Grundy (1953).

We now turn our attention to the estimation of $f(\mathbf{Y}_\delta)$, where $\mathbf{Y}_\delta = (Y_{\delta 1}, \cdots, Y_{\delta K})'$. A natural estimator is given by $f(\mathbf{Y}_{\hat{\delta}})$. Using the Taylor's series argument, it can be shown that

$$\mathrm{E}_M\left[f(\mathbf{Y}_{\hat{\delta}}) - f(\mathbf{Y}_\delta)\right] \doteq [\nabla f(\mathbf{Y}_\delta)]'\,\mathrm{E}_M(\mathbf{Y}_{\hat{\delta}} - \mathbf{Y}_\delta) = [\nabla f(\mathbf{Y}_\delta)]'\,\mathbf{Y}_{p_D},$$

where $\nabla f(\mathbf{Y}_\delta)$ is the gradient of $f(\mathbf{Y}_\delta)$, $\mathbf{Y}_{p_D} = (Y_{p_D 1}, \cdots, Y_{p_D K})'$ and $Y_{p_D k} = \sum_{i\in U} p_{i;D} y_{ik}$ $(k = 1, \cdots, K)$. Thus, $f(\mathbf{Y}_{\hat{\delta}})$ is biased for $f(\mathbf{Y}_\delta)$. A bias-adjusted estimator is given by $f(\hat{\mathbf{Y}})$, where $\hat{\mathbf{Y}} = (Y_{\hat{\delta}1} - \hat{Y}_{\pi^{-1}e1}, \cdots, Y_{\hat{\delta}1} - \hat{Y}_{\pi^{-1}eK})'$.

Theorem 3 below is useful in obtaining the total MSE of $f(\hat{\mathbf{Y}})$. Define $\Sigma \equiv ((\sigma_{kl})) = \mathtt{Var}(\mathbf{Y}_{\hat{\delta}} - \mathbf{Y}_\delta)$, the $K \times K$ covariance matrix of $\mathbf{Y}_{\hat{\delta}} - \mathbf{Y}_\delta$, where $\sigma_{kl} = \mathrm{Cov}(\hat{Y}_k - Y_{\delta k},\ \hat{Y}_l - Y_{\delta l})$, $(k, l = 1, \cdots, K)$. Note that we can write $\sigma_{kl} = (\sigma_{k+l} - \sigma_{kk} - \sigma_{ll})/2$, where σ_{k+l} is obtained from $\mathtt{Var}(\hat{Y}_k - Y_{\delta k})$ when we replace y_{ik} by $y_{ik} + y_{il}$ $(k \neq l,\ k, l = 1, \cdots, K)$.

Theorem 3. Under the misclassification model, we have

$$\mathrm{MSE}\left[f(\hat{\mathbf{Y}})\right] \approx [\nabla f(\mathbf{Y}_\delta)]'\,\Sigma\,[\nabla f(\mathbf{Y}_\delta)],$$

where

$$\begin{aligned}\sigma_{kk} &= \mathrm{E}_M\left\{\sum_{i\in U}\sum_{j>i\in U}(\pi_i\pi_j - \pi_{ij})\left(\pi_i^{-1} y_{ik} e_i - \pi_j^{-1} y_{jk} e_j\right)^2\right\} \\ &= \sum_{i\in U}\sum_{j>i\in U}(\pi_i\pi_j - \pi_{ij})\,\psi_{ij;kk},\end{aligned}$$

with

$$\psi_{ij;kk} = \pi_i^{-2} p_{i;T} y_{ik}^2 + \pi_j^{-2} p_{j;T} y_{jk}^2 - 2(\pi_i\pi_j)^{-1} p_{i;D} p_{j;D} y_{ik} y_{jk}.$$

We propose to estimate σ_{kk} by $\hat{\sigma}_{kk} = \sum_{i\in s_1}\sum_{j>i\in s_1}\pi_{ij}^{-1}(\pi_i\pi_j - \pi_{ij})\left(\pi_i^{-1} y_{ik} e_i - \pi_j^{-1} y_{jk} e_j\right)^2$, $(k = 1, \cdots, K)$. Using the property of the Yates-Grundy estimator, we have $\mathrm{E}_d(\hat{\sigma}_{kk}) = \sum_{i\in U}\sum_{j>i\in U}(\pi_i\pi_j - \pi_{ij})\left(\pi_i^{-1} y_{ik} e_i - \pi_j^{-1} y_{jk} e_j\right)^2$ and hence $\mathrm{E}(\hat{\sigma}_{kk}) = \mathrm{E}_M\mathrm{E}_d(\hat{\sigma}_{kk}) = \sigma_{kk}$. Thus, $\hat{\sigma}_{kk}$ is an unbiased estimator of σ_{kk} $(k = 1, \cdots, K)$. Thus, an unbiased estimator of σ_{kl} is given by $\hat{\sigma}_{kl} = (\hat{\sigma}_{k+l} - \hat{\sigma}_{kk} - \hat{\sigma}_{ll})/2$, $(k \neq l,\ k, l = 1, \cdots, K)$. Thus, an approximately unbiased estimator of $\mathrm{MSE}\left[f(\hat{\mathbf{Y}})\right]$ is given by

$$\mathrm{mse}\left[f(\hat{\mathbf{Y}})\right] = \left[\nabla f(\hat{\mathbf{Y}})\right]'\,\hat{\Sigma}\left[\nabla f(\hat{\mathbf{Y}})\right].$$

2.2. An illustrative example: bias adjusted SMR and relative regression coefficients in the presence of linkage errors

There has been an increasing use of computerized record linkage (CRL) method in various studies such as historical cohort mortality studies, cancer studies, political

studies and crime studies, in several countries, Howe (1985, 1998), Bennell *et al.* (2012), Giraud-Carrier *et al.* (2015). With very little effort, the method enables us to collect a large amount of data by linking records of human exposure to environmental hazards with records on health status. Since CRL utilizes already existing databases, it saves a substantial amount of money to collect new data. Various government agencies have developed sophisticated software to implement CRL, usually attaching weights reflecting the likelihood of a match to pairs of records.

Fair (1989) listed a number of health studies where environmental exposure data were linked to the Canadian Mortality Data Base (CMDB). Krewski *et al.* (2005) provided an example where National Dose Registry (NDR) of Canada has been linked to CMDB in order to study the associations between excess mortality due to cancer and occupational exposure to low levels of ionizing radiation. In Beauchamp *et al.* (2011), a sample of 2000 participants from a cohort study was linked to a statewide hospitalisations dataset in Victoria, Australia using the national health insurance (Medicare) number and demographic data as identifying variables. Kabudula *et al.* (2014) applied deterministic and probabilistic record linkage approaches to mortality records from 2006 to 2009 from the Agincourt Health and Demographic Surveillance Systems (HDSS) to those in the national civil registration (CR) in South Africa.

In a cohort mortality study, CRL method introduces two types of linkage errors. The Type I linkage error (usually called a false positive) occurs when a member of the cohort who is alive is incorrectly identified as dead. The Type II (or a false negative) error occurs when a member of the cohort who is dead is incorrectly identified as alive. Krewski *et al.* (2005) investigated the impact of linkage errors on estimates of epidemiological indicators of risk such as standardized mortality ratio (SMR) and the parameters of relative risk regression model. Their analytical and simulation results indicate that these indicators are, in general, subject to biases and additional variabilities in the presence of linkage errors.

In this subsection, we use the notation used in Krewski *et al.* (2005). In the analysis of cohort studies, mortality is usually characterized by the hazard function which relates death rate as a function of time. Denoting T the time of death, the hazard function at time u is defined as

$$\lambda(u) = \lim_{\triangle u \downarrow 0} \frac{\Pr\{u \leq T < u + \triangle u \mid T \geq u\}}{\triangle u}.$$

The corresponding survival function and the probability density function are given by $S(u) = \exp\big(-\int_0^u \lambda(t)dt\big)$ and $f(u) = \lambda(u)S(u)$, respectively. Let $\lambda_i(u)$ and $\mathbf{Z}_i(u)$ be the hazard function for a specific cause of death and the value of the vector of covariates at time u for the ith member of the cohort, $i = 1, \cdots, N$. The relative risk regression is then described as

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

where $\lambda^*(u)$, value of $\lambda_i(u)$ when $\beta = 0$, is known as the baseline hazard and γ is a positive function of the covariates and β.

Let t_i^0 and t_i^1 be the age at the time of entry into the study and time of loss to follow up for the ith member of the cohort $i = 1, \cdots, N$. Let $\delta_i = 1$ if the ith individual has died at the time of loss to follow-up and $\delta_i = 0$ otherwise. The likelihood based on the relative risk regression model is given by $L = \prod_{i=1}^N f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$. The corresponding log-likelihood is given by

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

The maximum likelihood estimate $\hat{\beta}$ of β is obtained as a solution of $\frac{\partial \log L}{\partial \beta} = 0$. Note that when $\mathbf{z}_i(u) = 1$, $\gamma\{\hat{\beta}\}$ reduces to the standardized mortality rate given by $SMR = OBS/EXP$, where $OBS = \sum_{i=1}^N \delta_i =$ observed number of death before time to follow-up and $EXP = \sum_{i=1}^N e_i =$ expected number of deaths, with $e_i = \int_{t_i^0}^{t_i^1} \lambda^*(u)du$.

Due to time-dependent covariates $\mathbf{z}_i(u)$, the integral must be re-evaluated at each iteration of the maximization process. Thus, it is very computer-intensive, specially when the cohort size is large. Breslow *et al.* (1983) simplified the likelihood by assuming $\mathbf{z}_i(u) = \mathbf{z}_j$ whenever the ith cohort member passes through the state S_j $(j = 1, \cdots, J)$. The states can be defined by cross-classification of the covariates of interest. In this case, the log-likelihood can be written as

$$\log L = \sum_{i=1}^{N} \left\{ d_j \log(\gamma\{\mathbf{z}_j'\beta\}) - \gamma\{\mathbf{z}_j'\beta\}e_j \right\},$$

where $e_j = \sum_{i=1}^N \int_{[\mathbf{z}_i(u)\in S_j]} \lambda^*(u)du$ is the contribution to the expected number of deaths from all person-years of observation in the state S_j and d_j is the total number of death in that state. The maximum likelihood estimate of β is then obtained as a solution to the following equation:

$$\sum_{i=1}^{J} \frac{\partial \Lambda_j(\beta)}{\partial \beta} \left\{ d_j - \exp\{\Lambda_j(\beta)\}e_j \right\} = 0,$$

where $\Lambda_j(\beta) = \log\left(\gamma\{\mathbf{z}_j\prime\beta\}\right)$.

Note that the results of Section 2.1 are valid for each state $S_j, j = 1, \cdots, J$. We introduce an additional suffix j to indicate that the parameter or estimator refers to the state $S_j, j = 1, \cdots, J$. For example, we shall have Y_j in place of Y. Note that with $Y_{1ij} = 1$ and $Y_{0ij} = 0$, Y_j reduces to d_j of Krewski *et al.* (2005). We have $Y_j = e_j$

if we choose $Y_{1ij} = \int_{t_{ij}^0}^{t_{ij}^1} \lambda^*(u)du$ and $Y_{0ij} = \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du$. Consider $f(X,Y) = X/Y$ with $X = \sum_{j=1}^J d_j = d$ and $Y = \sum_{j=1}^J e_j = e$. With these choices, $f(X,Y) = SMR$.

Suppose a sample of size n_j is selected using a probability sampling scheme from each state S_j , $j = 1, \cdots, J$. Let $\pi_{i(j)}$ and $\pi_{ik(j)}$ denote the first order and second order inclusion probabilities in state S_j , $j = 1, \cdots, J$. Using the results of Section 2.1, *SMR* can be adjusted for its bias due to linkage errors. The bias-corrected *SMR* is given by $\widehat{SMR} = (\hat{d}/\hat{e})$, where $\hat{d} = \sum_{j=1}^J \hat{d}_j, \hat{e} = \sum_{j=1}^J \hat{e}_j, \widehat{\Delta d_j} = \sum_{i\in s} \pi_{i(j)}^{-1} \Delta\delta_{i(j)}, \widehat{\Delta e_j} = \sum_{i\in s} \pi_{i(j)}^{-1} \Delta\delta_{i(j)} \int_{min(t_{ij}^1, t_{ij}^2)}^{t_{ij}^2} \lambda^*(u)du$, $\hat{d}_j = d_j^L - \Delta\hat{d}_j$ and $\hat{e}_j = e_j^L + \Delta\hat{e}_j$.

An application of Theorem 3 provides us with an estimator of the variance of $\widehat{SMR} - SMR$. It is given by

$$\texttt{var}(\widehat{SMR} - SMR) = \hat{e}^{-2}\left[\texttt{var}(\hat{d}-d) + \hat{d}^2\hat{e}^{-2}\texttt{var}(\hat{e}-e) + 2\hat{d}\hat{e}^{-1}\texttt{cov}(\hat{d}-d, \hat{e}-e)\right],$$

where $\texttt{var}(\hat{d}-d) = \sum_{j=1}^J v_j$ with $v_j = \sum_i^{n_j} \sum_{k>i}^{n_j} \pi_{ik(j)}{}^{-1} (\pi_{i(j)}\pi_{k(j)} - \pi_{ik(j)})(\pi_{i(j)}^{-1}\Delta\delta_{i(j)} - \pi_{k(j)}^{-1}\Delta\delta_{k(j)})^2$. We can define $\texttt{var}(\hat{e}-e)$ and $\texttt{cov}(\hat{d}-d, \hat{e}-e)$ similarly.

Let us now consider estimation of the regression coefficient β in the relative risk regression model. First, we propose to adjust the log-likelihood given in Krewski *et al.* (2005). We shall find the estimator of β as a solution, say, $\hat{\beta}$, of the following score function

$$Q = Q(\beta; (\hat{d}_j, \hat{e}_j), j = 1, \cdots, J) = \sum_{j=1}^J \frac{\partial \Lambda_j(\beta)}{\partial \beta} \left\{\hat{d}_j - \texttt{exp}\{\Lambda_j(\beta)\}\hat{e}_j\right\} = 0.$$

Since $\mathrm{E}\left\{Q(\beta; (\hat{d}_j, \hat{e}_j), j = 1, \cdots, J) - Q(\beta; (d_j, e_j), j = 1, \cdots, J)\right\} = 0$, we can expect $\hat{\beta}$ to perform well. The covariance matrix of the proposed estimator, $\hat{\beta}$, can be estimated by $\left[\frac{\partial Q}{\partial \beta}\right]^2_{\beta=\hat{\beta}}$.

2.3. Evaluation

In this subsection, we consider the prediction of $Y = \sum_{i=1}^N \delta_i$. A naive predictor of Y that ignores the misclassification errors is given by $\hat{Y} = \sum_{i=1}^N \hat{\delta}_i$. Under the misclassification error model of Section 2.1 with $p_{ikl} = p_{kl}$, $(i = 1, \cdots, N;\ k, l = 0, 1)$, the misclassification error bias of $\hat{Y}$ can be obtained as

$$\text{Bias}_M(\hat{Y}) = N(p_{10} - p_{01}) = Nb,$$

where $b = p_{10} - p_{01}$. Thus, $\hat{Y}$ is positively (negatively) biased if the false positive probability is more (less) than the false negative probability. It is interesting to

note that there is no bias in $\hat{Y}$ due to misclassification even if there is a large misclassification error as long as the false positive and false negative probabilities are identical.

Using the theory developed in Section 2.1, we can correct the misclassification bias of $\hat{Y}$ by drawing a simple random sample (SRS) of size n from U and determining the status of each sampled unit for misclassification error. We propose the misclassification bias-corrected predictor of Y as

$$\tilde{Y} = \hat{Y} - \widehat{Bias},$$

where

$$\widehat{Bias} = \frac{N}{n}\sum_{i\in s}^{n} e_i.$$

Evidently, the proposed predictor is unbiased with respect to the combined distribution of misclassification and sampling errors. But, the proposed method introduces some costs. Also, the bias correction is expected to increase the variability. Thus, we study how the above bias-correction affects the mean squared error that incorporates both the misclassification and sampling errors. We define the total mean square error of a predictor $\hat{Y}$ of Y as

$$\mathrm{MSE}(\hat{Y}) = \mathtt{E}(\hat{Y} - Y)^2,$$

where the expectation $\mathtt{E}$ is with respect to the SRS and the multinomial misclassification error model. After considerable algebra, we obtain

$$\mathrm{MSE}(\hat{Y}) = N(a + Nb^2),$$
$$\mathrm{MSE}(\tilde{Y}) = \frac{N^2}{n}(1-f)a,$$

where $a = p_{10}(1-p_{10}) + p_{01}(1-p_{01}) + 2p_{10}p_{01}$ and $f = \frac{n}{N}$.

We define the relative improvement in MSE (MSERI) as follows:

$$\mathrm{MSERI} = \frac{\mathrm{MSE}(\hat{Y}) - \mathrm{MSE}(\tilde{Y})}{\mathrm{MSE}(\tilde{Y})}.$$

It can be shown that

$$\mathrm{MSERI} = [\frac{n}{(1-f)} \times \frac{b^2}{a}] - \frac{1-2f}{1-f}, \tag{1}$$

where $a > 0$, $0 < f < 1$ and $0 < b^2 < 1$. In order for the bias-corrected predictor $\hat{Y}$ to improve on the naive predictor $\hat{Y}$ in terms of MSE, b must satisfy one of the following two conditions:

$$b > \sqrt{(\frac{1}{n} - \frac{2}{N})a} = b_2 \tag{2}$$

or

$$b < -\sqrt{(\frac{1}{n} - \frac{2}{N})a} = b_1. \tag{3}$$

In many situations, the sampling fraction f is negligible in which case MSREI $\approx n\frac{b^2}{a}$.

Define the relative bias (RB) as

$$\mathrm{RB}(\hat{Y}) = \frac{Nb}{\mathrm{E}(Y)} = \frac{Nb}{N(p_{11}+p_{01})} = \frac{p_{10}-p_{01}}{p_{11}+p_{01}}.$$

Since both MSRI and RB depend on N only through f, we arbitrarily fix N and vary f. For a numerical comparison, we fix $N = 100$. Table 2 displays $\mathrm{RB}(\hat{Y})$, $\mathrm{MSE}(\hat{Y})$, $\mathrm{MSE}(\tilde{Y})$ and MSERI for $f = 0.05,\ 0.10$ and two levels of misclassification errors (LE): High (H) and Moderate (M). Table 3 reports the results for low misclassification errors (L). First of all, we notice that the relative bias of $\hat{Y}$ depends on the configurations of false positive (p_{10}) and false negative (p_{01}) probabilities. Clearly, a high relative bias in $\hat{Y}$ is possible. In this case, the bias-corrected estimator $\tilde{Y}$ can have substantially smaller MSE than $\hat{Y}$ even when the sampling fraction f is small. When the relative bias in $\hat{Y}$ is small, one needs much higher sampling fraction for $\tilde{Y}$ to improve on $\hat{Y}$ in terms of mean squared error.

Table 2. $\mathrm{RB}(\hat{Y}^*)$, $\mathrm{MSE}(\hat{Y}^*)$, $\mathrm{MSE}(\hat{Y})$ and MSERI for high and medium misclassification errors

n	f	LE	p_{10}	p_{01}	b_2	b	b_1	$\mathrm{RB}(\hat{Y}^*)$	$\mathrm{MSE}(\hat{Y}^*)$	$\mathrm{MSE}(\hat{Y})$	MSERI
		H	0.80	0.15	0.31	0.65	-0.31	3.71	4277.75	1002.25	3.27
		H	0.75	0.15	0.31	0.60	-0.31	3.00	3654.00	1026.00	2.56
		H	0.60	0.10	0.28	0.50	-0.28	3.33	2545.00	855.00	1.98
		H	0.10	0.60	0.28	-0.50	-0.28	-0.77	2545.00	855.00	1.98
		H	0.20	0.65	0.34	-0.45	-0.34	-0.60	2089.75	1230.25	0.70
5	0.05	H	0.50	0.35	0.39	0.15	-0.39	0.38	307.75	1572.25	-0.80
		H	0.30	0.60	0.38	-0.30	-0.38	-0.49	981.00	1539.00	-0.36
		H	0.45	0.25	0.34	0.20	-0.34	0.44	466.00	1254.00	-0.63
		M	0.10	0.40	0.27	-0.30	-0.27	-0.43	941.00	779.00	0.21
		M	0.10	0.35	0.26	-0.25	-0.26	-0.45	663.75	736.25	-0.10
		M	0.35	0.20	0.31	0.15	-0.31	0.27	277.75	1002.25	-0.72
		M	0.15	0.35	0.29	-0.20	-0.29	-0.33	446.00	874.00	-0.49
		H	0.80	0.15	0.21	0.65	-0.21	3.71	4277.75	474.75	8.01
		H	0.75	0.15	0.21	0.60	-0.21	3.00	3654.00	486.00	6.52
		H	0.60	0.10	0.19	0.50	-0.19	3.33	2545.00	405.00	5.28
		H	0.10	0.60	0.19	-0.50	-0.19	-0.77	2545.00	405.00	5.28
		H	0.20	0.65	0.23	-0.45	-0.23	-0.60	2089.75	582.75	2.59
10	0.10	H	0.50	0.35	0.26	0.15	-0.26	0.38	307.75	744.75	-0.59
		H	0.30	0.60	0.25	-0.30	-0.25	-0.49	981.00	729.00	0.35
		H	0.45	0.25	0.23	0.20	-0.23	0.44	466.00	594.00	-0.22
		M	0.10	0.40	0.18	-0.30	-0.18	-0.43	941.00	369.00	1.55
		M	0.10	0.35	0.18	-0.25	-0.18	-0.45	663.75	348.75	0.90
		M	0.35	0.20	0.21	0.15	-0.21	0.27	277.75	474.75	-0.41
		M	0.15	0.35	0.19	-0.20	-0.19	-0.33	446.00	414.00	0.08

Table 3. RB$(\hat{Y}^*)$, MSE$(\hat{Y}^*)$, MSE$(\hat{Y})$ and MSERI for low misclassification errors

n	f	p_{10}	p_{01}	b_2	b	b_1	RB$(\hat{Y}^*)$	MSE$(\hat{Y}^*)$	MSE$(\hat{Y})$	MSERI
5	0.05	0.07	0.03	0.13	0.04	-0.13	0.05	25.84	186.96	-0.86
		0.07	0.13	0.19	-0.06	-0.19	-0.11	55.64	373.16	-0.86
10	0.10	0.07	0.03	0.09	0.04	-0.09	0.05	25.84	88.56	-0.71
		0.07	0.13	0.13	-0.06	-0.13	-0.11	55.64	176.76	-0.69
15	0.15	0.07	0.03	0.07	0.04	-0.07	0.05	25.84	55.76	-0.54
		0.07	0.13	0.10	-0.06	-0.10	-0.11	55.64	111.29	-0.50
30	0.30	0.07	0.03	0.04	0.04	-0.04	0.05	25.84	22.96	0.13
		0.07	0.13	0.05	-0.06	-0.05	-0.11	55.64	48.83	0.21

Suppose $p_{10} = p_{01}$, then $b = 0$. Hence, $\texttt{RB}(\hat{Y})$ will always be zero. Also, the MSERI will be a function of the sampling fraction f as shown below

$$\text{MSERI} = -\frac{1-2f}{1-f}. \tag{4}$$

Figure 1 displays the MSERI for different choices of f.

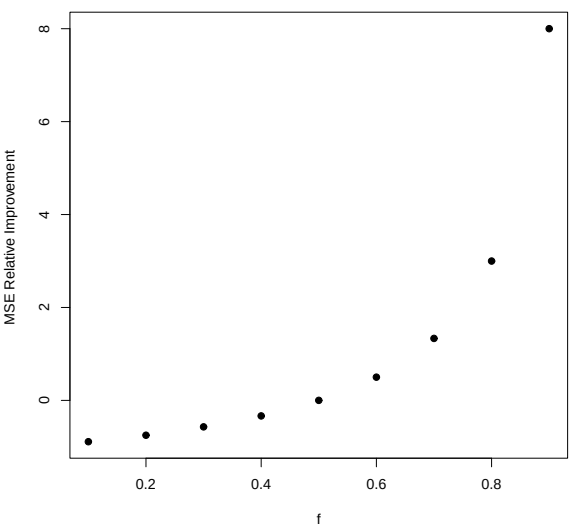


Figure 1. MSE Relative Improvement when $p_{10} = p_{01}$

3. Prediction of $\mathbf{Y}_\delta$ when $S \subset U$

In this subsection, we consider prediction of a finite population total $Y = \sum_{i=1}^{N} \delta_i$ using a sample (s_1) of size n_1 drawn by the *probability proportional to size with*

replacement (PPSWR) sampling design, where size is defined as:

$$\phi_i = \frac{x_i}{\sum_{i=1}^{N} x_i},$$

where $x_i > 0$ known, $i \in U$. A natural estimator of Y is given by $\hat{Y}^* = \sum_{i \in s_1} \omega_i \delta_i^*$, where δ_i^* is *observed* value of δ_i with misclassification error and $\omega_i = \frac{1}{n_1 \phi_i}, i \in s_1$.

Under the PPSWR sample design and the multinomial classification model of Section 2.3, a heavy algebra yields

$$\text{Bias}(\hat{Y}^*) \quad = \quad Nb$$

and

$$\text{MSE}(\hat{Y}^*) \quad = \quad \frac{1}{n_1} \sum_{i=1}^{N} \frac{1}{\phi_i} p_{1+} - \frac{N}{n_1} p_{1+}\{1 + (N-1)p_{1+}\} + N[a + b^2(N-1)],$$

where $p_{1+} = p_{11} + p_{10}$ and $p_{+1} = p_{11} + p_{01}$.

In order to correct the bias of the predictor $\hat{Y}^*$, a second sample, say s_2, of size n_2 is drawn from the first sample s_1 using a SRS design and true δ_i is measured without misclassification error. We define a bias-corrected predictor $\hat{Y}$ as follows:

$$\hat{Y} = \hat{Y}^* - \widehat{Bias},$$

where

$$\widehat{Bias} = \frac{n_1}{n_2} \sum_{i \in s_2} \omega_i (\delta_i^* - \delta_i).$$

It is easy to show that $\hat{Y}$ is an unbiased predictor of Y. Using heavy algebra, the exact MSE of $\hat{Y}$ under different sources of uncertainty is obtained as follows:

$$\begin{aligned} \text{MSE}(\hat{Y}) = \frac{1}{n_1}\{ & -2Np_{11} + \sum_{i=1}^{N} \frac{1}{\phi_i}[p_{+1} + a\frac{1}{f_2}(1-f_2)] \\ & -Np_{1+}\frac{1}{f_2}(1-f_2)[1 + p_{1+}(N-1)]\}, \end{aligned}$$

where $f_2 = \frac{n_2}{n_1}$.

It can be shown that

$$\begin{aligned} & MSE(\hat{Y}^*) - MSE(\hat{Y}) = \\ & \frac{1}{n_1}\{\sum_{i=1}^{N} \frac{1}{\phi_i}[b - \frac{a(1-f_2)}{f_2}] - Np_{1+}[1 + p_{1+}(N-1) - \frac{1}{f_2}(1-f_2)[1 + p_{1+}(N-1)]] \\ & \qquad + 2Np_{11}\} + N[a + b^2(N-1)]. \quad (5) \end{aligned}$$

The expression for MSERI is obtained by dividing Eq.(5) by $\mathrm{MSE}(\hat{Y})$. Also, we can show that

$$\mathrm{RB}(\hat{Y}^*) = \frac{Nb}{\mathrm{E}(Y)} = \frac{Nb}{N(p_{11}+p_{01})} = \frac{b}{p_{11}+p_{01}}.$$

Tables 4, 5 and 6 display $\mathrm{RB}(\hat{Y}^*)$, $\mathrm{MSE}(\hat{Y}^*)$, $\mathrm{MSE}(\hat{Y})$ and MSERI for high, medium and low misclassification probabilities given by

$$P_H = \begin{pmatrix} 0.10 \\ p_{10} \\ 0.80-p_{10} \\ 0.10 \end{pmatrix}, P_M = \begin{pmatrix} 0.25 \\ p_{10} \\ 0.50-p_{10} \\ 0.25 \end{pmatrix}, P_L = \begin{pmatrix} 0.40 \\ p_{10} \\ 0.20-p_{10} \\ 0.40 \end{pmatrix},$$

respectively.

Different configurations of the high, medium and low misclassification errors are considered by varying the false positive probability p_{10}. For each case, we consider $f_2 = 0.2$, $N = 1000$ when $n_1 = 300$, and $N = 100,000$ when $n_1 = 10,000$.

Table 4. $\mathrm{RB}(\hat{Y}^*)$, $\mathrm{MSE}(\hat{Y}^*)$, $\mathrm{MSE}(\hat{Y})$ and MSERI for high misclassification errors

n_1	p_{10}	$\mathrm{Bias}(\hat{Y}^*)$	$\mathrm{RB}(\hat{Y}^*)(\%)$	$\mathrm{MSE}(\hat{Y}^*)$	$\mathrm{MSE}(\hat{Y})$	$\mathrm{MSERI}(\%)$
	0.10	-600	-75.00	387092.90	535198.40	-27.67
	0.20	-400	-57.14	200519.40	521137.70	-61.52
	0.30	-200	-33.33	93799.32	506810.60	-81.49
300	0.40	0	0.00	66932.65	492217.10	-86.40
	0.50	200	50.00	119919.40	477357.20	-74.88
	0.60	400	133.33	252759.50	462230.90	-45.32
	0.70	600	300.00	465453.00	446838.20	4.17
	0.10	-60000	-75.00	3615422483	308209682	1073.04
	0.20	-40000	-57.14	1623101725	300300437	440.50
	0.30	-20000	-33.33	430752967	292311194	47.36
10000	0.40	0	0.00	38376209	284241951	-86.50
	0.50	20000	50.00	445971451	276092709	61.53
	0.60	40000	133.33	1653538694	267863468	517.31
	0.70	60000	300.00	3661077936	259554228	1310.53

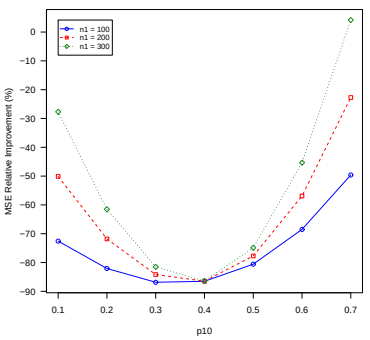
Table 5. $\mathrm{RB}(\hat{Y}^*)$, $\mathrm{MSE}(\hat{Y}^*)$, $\mathrm{MSE}(\hat{Y})$ and MSERI for medium misclassification errors

n_1	p_{10}	$\mathrm{Bias}(\hat{Y}^*)$	$\mathrm{RB}(\hat{Y}^*)(\%)$	$\mathrm{MSE}(\hat{Y}^*)$	$\mathrm{MSE}(\hat{Y})$	$\mathrm{MSERI}(\%)$
	0.05	-400	-57.14	200219.4	360416.3	-44.45
	0.15	-200	-33.33	93499.32	346089.2	-72.98
300	0.25	0	0.00	66632.65	331495.7	-79.90
	0.35	200	50.00	119619.4	316635.8	-62.22
	0.45	400	133.33	252459.5	301509.5	-16.27
	0.05	-40000	-57.14	1623071725	207789527	681.11
	0.15	-20000	-33.33	430722967	199800284	115.58
10000	0.25	0	0.00	38346209	191731041	-80
	0.35	20000	50.00	445941451	183581799	142.92
	0.45	400	133.33	1653508694	175352558	842.96

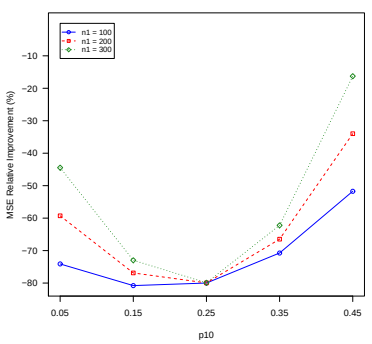
Table 6. RB($\hat{Y}^*$), MSE($\hat{Y}^*$), MSE($\hat{Y}$) and MSERI for low misclassification errors

n_1	p_{10}	Bias($\hat{Y}^*$)	RB($\hat{Y}^*$)(%)	MSE($\hat{Y}^*$)	MSE($\hat{Y}$)	MSERI(%)
	0.01	-180	-30.51	86919.25	183920.5	-52.74
	0.05	-100	-18.18	69784.31	178104.4	-60.82
300	0.10	0	0.00	66332.65	170774.4	-61.16
	0.15	100	22.22	82844.34	163377.7	-49.29
	0.199	198	49.38	118394.2	156064.4	-24.14
	0.01	-18000	-30.51	355456551	106486050	233.81
	0.05	-10000	-18.18	134508088	103264753	30.26
10000	0.10	0	0.00	38316209	99220131	-61.38
	0.15	10000	22.22	142117330	95155510	49.35
	0.199	19800	49.38	437875637	91152778	380.38

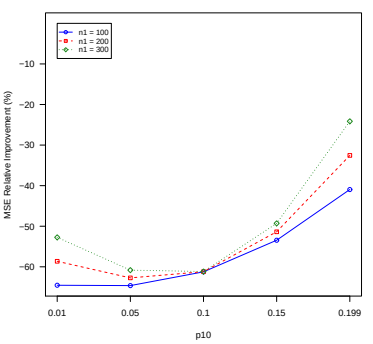
The results for MSE improvement achieved by the bias-corrected estimator over the naive estimator for different situations are plotted in Figures 2 and 3.



(a) High Misclassification Error

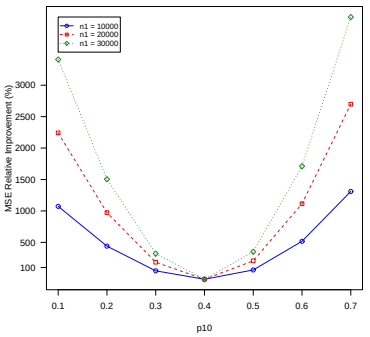


(b) Medium Misclassification Error

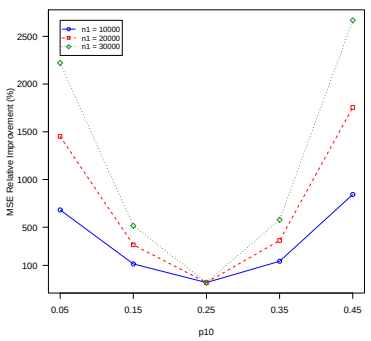


(c) Low Misclassification Error

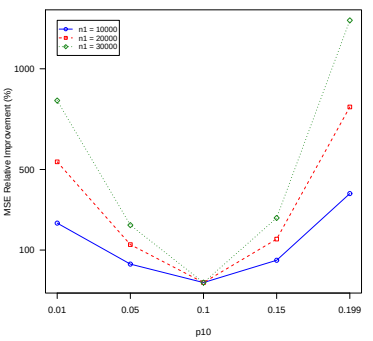
Figure 2. MSE Relative Improvement for $f_2 = 0.2$, $N = 1000$ and n_1 = 100, 200, 300



(a) High Misclassification Error



(b) Medium Misclassification Error



(c) Low Misclassification Error

Figure 3. MSE Relative Improvement for $f_2 = 0.2$, $N = 10,000$ and n_1 = 10,000, 20,000, 30,000

4. Conclusions

Our research shows that it is possible to correct bias due to misclassification in predictors by drawing a probability sample from the original data, determining the status of misclassification error for the sample and then applying the standard sample survey method. The bias-correction increases variance in the predictor, which impacts the mean squared error. The improvement depends on the distribution of the misclassification error and the sampling fraction in the drawn sample. If additional data that generates the misclassification error are available as in the record linkage literature, it may be possible to improve on the proposed method, we plan to investigate this direction in the future.

5. Acknowledgements

We are also grateful to an anonymous referee for making a number of constructive suggestions, which led to a significant improvement of our paper.

REFERENCES

BEAUCHAMP, A., TONKIN, A. M., KELSALL, H., SUNDARARAJAN, V., ENGLISH, D. R., SUNDARESAN, L., WOLFE, R., TURRELL, G., GILES, G. G., PEETERS, A., (2011). Validation of de-identified record linkage to ascertain hospital admissions in a cohort study. BMC Medical Research Methodology. 11–42.

BENNELL, C., SNOOK, B., MACDONALD, S., HOUSE, J. C., TAYLOR, P. J., (2012). Computerized crime linkage systems: a critical review and research agenda. Criminal Justice and Behavior. 39(5): 620–634.

BOESE, D. H., YOUNG, D. M., STAMEY, J. D., (2006). Confidence intervals for a binomial parameter based on binary data subject to false-positive misclassification. Computational Statistics & Data Analysis. 50: 3369–3385.

BRESLOW, N. E., LUBIN, J. H., LANGHOLZ, B., (1983). Multiplicative models and cohort analysis. Journal of the American Statistical Association. 78: 1–12.

BROSS, I., (1954). Misclassification in 2×2 tables. Biometrics. 10: 478–486.

EVANS, M., GUTTMAN, I., HAITOVSKY, Y., SWARTZ, T., (1996). Bayesian analysis of binary data subject to misclassification. In: Berry, D., Chaloner, K., Geweke, J., eds. Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner. New York: John Wiley, 67–77.

FAIR, M. E., (1989). Studies and references relating to the uses of the Canadian Mortality Data Base. Report from the Occupational and Environmental Health Research Unit, Health Division, Statistics Canada, Ottawa.

FELLIGI, I., SUNTER, A., (1969). A theory for record linkage. Journal of the American Statistical Association. 64: 1183–1210.

GABA, A., WINKLER, R. L., (1992). Implications of errors in survey data: a Bayesian model. Management Science. 38: 913–925.

GIRAUD-CARRIER, C., GOODLIFFE, J., JONES, B. M., CUEVA, S., (2015). Effective record linkage for mining campaign contribution data. Knowledge and Information Systems. 45(2): 389–416.

GOLDBERG, J. D., (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. Journal of the American Statistical Association. 70: 561–567.

GUSTAFSON, P., LE, N. D., SASKIN, R., (2001). Case-control analysis with partial knowledge of exposure misclassification probabilities. Biometrics. 57: 598–609.

HOWE, G. R., (1985). Use of computerized record linkage in follow-up studies of cancer epidemiology in Canada. National Cancer Institute Monograph. 67: 117–121.

HOWE, G., R., (1998). Use of computerized record linkage in cohort studies. Epidemiologic Reviews. 20(1): 112–121.

HERZOG, T. N., SCHEUREN, F. J., WINKLER, W. E., (2007). Data Quality and Record Linkage Techniques. Springer, New York, NY.

KABUDULA, C. W., JOUBERT, J. D., TUOANE-NKHASI, M., KAHN, K., RAO, C., GÓMEZ OLIVÉ, F. X., MEE, P., TOLLMAN, S., LOPEZ, A. D., VOS, T., BRADSHAW, D., (2014). Evaluation of record linkage of mortality data between a health and demographic surveillance system and national civil registration system in South Africa. Population Health Metrics. 12–23.

KREWSKI, D., DEWANJI, A., WANG, Y., BARTLETT, S., ZIELINSKI, J. M., MALLICK, R., (2005). The Effect of Record Linkage Errors on Risk Estimates in Cohort Mortality Studies. Survey Methodology. 31: 13–21.

LAHIRI, P., LARSEN, M. D., (2005). Regression analysis with linked data. Journal of the American Statistical Association. 100: 222–230.

LYLES, R. H., LIN, H., M., WILLIAMSON, J. M., (2004). Design and analytic considerations for single-armed studies with misclassification of a repeated binary outcome. Journal of Biopharmaceutical Statistics. 14: 229–247.

NETER, J., MAYNES, E. S., RAMANATHAN, R., (1965). The effect of mismatching on the measurement of response errors. Journal of the American Statistical Association. 60: 1005–1027.

RAHARDJA, D., YANG, Y., (2015). Maximum likelihood estimation of a binomial proportion using one-sample misclassified binary data. Statistica Neerlandica. 69(3), 272–280.

RAHARDJA, D., ZHAO, Y. D., (2013). One-way analysis of proportions for misclassified binomial data. Journal of Statistical Computation and Simulation. 1–10.

SCHEUREN, F., WINKLER, W. E., (1993). Regression Analysis of Data Files That Are Computer Matched. Survey Methodology. 19, 39–58.

STAMEY, J. D., SEAMAN, J. W., YOUNG, D. M., (2007). Bayesian estimation of intervention effect with pre- and post-misclassified binomial data. Journal of Biopharmaceutical Statistics. 17: 93–108.

TENENBEIN, A., (1970). A double sampling scheme for estimating from binomial data with misclassifications. Journal of American Statistical Association. 65(331): 1350–1361.

VIANA, M., RAMAKRISHNAN, V., LEVY, P., (1993). Bayesian analysis of prevalence from results of small screening samples. Communication Statistics Theory and Methods. 22: 575–585.

YATES, F., GRUNDY, P. M., (1953). Selection without replacement from within strata with probability proportional to size. Journal of the Royal Statistical Society: Series B. 15: 235–261.

ZHONG, B., (2002). Evaluating qualitative assays using sensitivity and specificity. Journal of Biopharmaceutical Statistics. 12: 409–424.