

# **Regression analysis of biased case-control data with application to spontaneous adverse drug reaction reports**

**Technical Report No. ASD/2013/1  
Dated: 9 April 2013**

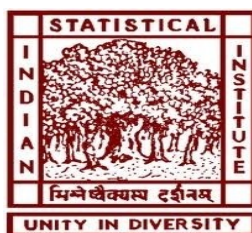
Palash Ghosh

And

Anup Dewanji

Applied Statistics Unit  
Indian Statistical Institute  
Kolkata 700108

[palash\\_r@isical.ac.in](mailto:palash_r@isical.ac.in), [dewanjia@isical.ac.in](mailto:dewanjia@isical.ac.in)



# Regression analysis of biased case-control data with application to spontaneous adverse drug reaction reports

Palash Ghosh and Anup Dewanji

Applied Statistics Unit, Indian Statistical Institute, Kolkata 700108

emails: *palash\_r@isical.ac.in, dewanjia@isical.ac.in*

## Abstract

The data obtained from case-control sampling may suffer from selection or reporting bias, resulting in biased estimation of the parameter(s) of interest by standard analysis of case-control data. In this work, the problem of reporting bias is dealt with by introducing the concept of reporting probability. Then, considering a reference sample from the source population, we obtain unbiased estimate of the population parameters by fitting a pseudo likelihood. This is an extension of the work by Scott & Wild (1997) and Lee et al. (2006) who, without the problem of reporting bias, have considered a profile likelihood approach by estimating the exposure distribution in the population nonparametrically. The procedure has been shown to be semi-parametrically fully efficient. We motivate the need for such methodology by considering the analysis of spontaneous adverse drug reaction (ADR) reports in presence of reporting bias.

Key words: Reporting bias, Response-selective sampling, Spontaneous reporting database, Semiparametric estimation, Pseudo likelihood

## 1 Introduction

Prentice & Pyke (1979) have proved that a prospective logistic regression model can be used for the analysis of case-control data. Since then, substantial research has been done resulting in such advantage in various modifications depending on practical requirements. In this context, Hsieh et al. (1985), Scott & Wild (1997) and Lee et al. (2006) have considered augmentation of case and/or control data by different extent of information to improve the overall efficiency in estimation of the parameter(s) of interest. All the works, in this regard, have the basic assumption that the case and/or control samples are representative of the corresponding source population. In many situations, this assumption may not be true. Then, the existing methodologies lead to biased estimates. In this work, we develop methodologies to deal with such situations.

Breslow (1996) has discussed the limitations and the challenges in case-control study including the problem of selection-bias due to high rates of non-participation and confounding leading to distortion of the response-exposure relationship. Prentice & Breslow (1978) have dealt with the problem of selection bias by considering inclusion probabilities for the individuals of the source population in the corresponding case or control samples. They have assumed these probabilities to be independent of the explanatory variables, resulting in the corresponding conditional likelihood being free from those inclusion probabilities. There may, however, be situations when both case and control samples suffer from selection bias the extent of which may depend on the exposure and the case-control status as well. In fact, most of the hospital-based or registry-based case-control studies are subject to selection bias, which is often ignored. In pharmacovigilance studies, one objective

is to detect alarming signal regarding adverse drug reaction (ADR) from a drug of interest (Bate et al., 1998). Information on drug related ADRs can be found in spontaneous reporting (SR) databases, where clinicians and/or health professionals report the suspected ADRs after the drug is in the market. As the reporting of ADRs is not mandatory and sometimes the ADRs may not be recognized easily, SR data represents a biased case-control sample of the corresponding counterparts in the source population, defined by the collection of individuals suffering from a particular disease. Individuals of the source population experiencing the ADR of interest are considered as case, while those not experiencing the ADR of interest are considered to be members of the control population. The controls in the SR database experience some other ADRs and are, therefore, clearly subject to selection bias in addition to reporting bias. It is likely that the extent of this selection/reporting bias in both case and control samples from SR database depends on both exposure to the particular drug and case-control status.

This work intends to address this problem of biased case-control sample with the help of additional information from a reference sample. Lee et al. (2006) have considered the use of reference sample augmented by a sample of cases only, assumed to represent the population of cases without having any selection or reporting bias. Ghosh & Dewanji (2011) explicitly incorporates the reporting bias by means of some reporting parameters to deal with biased case-control data with binary exposure with the help of a reference sample while analysing ADR data from SR database. In this paper we generalize the work of Ghosh & Dewanji (2011) to address the regression problem when the exposure can be a continuous variable. In what follows, we also consider multiple response categories, in which the term ‘case-control sampling’ may be replaced by ‘response-selective sampling’ (Lee & Hirose, 2010), or choice-based sampling (Cosslett, 1981). In Section 2, we describe the model along with the selection or reporting probabilities. Section 3 considers semiparametric estimation of the model parameters in the sense that the exposure distribution remains nonparametric. A special case of the proposed methodology are considered in Section 4. Section 5 presents a simulation study to investigate the properties of the estimates, while Section 6 illustrates the method through analysis of an example.

## 2 Modeling and Likelihood

Consider a categorical random variable  $Y$  having categories  $j = 0, \dots, J$ , the distribution of which depends on the vector of covariates  $X$  of dimension  $p \geq 1$  through a prospective model of the form

$$P(Y = j|X = x) = p_j(x, \boldsymbol{\beta}), \quad (1)$$

for  $j = 0, \dots, J$ , with  $\sum_0^J p_j(x, \boldsymbol{\beta}) = 1$ , where  $\boldsymbol{\beta}$  is the corresponding vector of parameters. In response-selective sampling, the observation on  $X$  is obtained conditional on the response  $Y$ . A particular individual from a response category may be selected (reported) with certain probability depending even on  $X$  as given by the model

$$P(R = 1|Y = j, X = x) = \mu_j(x, \boldsymbol{\gamma}), \text{ say,} \quad (2)$$

where  $R$  is the binary random variable taking values 1 or 0 representing selection in the sample or not and  $\boldsymbol{\gamma}$  is the corresponding vector of parameters.

Suppose there are  $n_j$  individuals in the sample with  $Y = j$  having the observed  $X$ -values as  $x_{ji}$ , for  $i = 1, \dots, n_j$ , and  $j = 0, \dots, J$ . Also, let  $g(x)$  denote the marginal density of  $x$ . Then, the

retrospective likelihood is given by

$$\prod_{j=0}^J \prod_{i=1}^{n_j} P(X = x_{ji} | R = 1, Y = j) = \prod_{j=0}^J \prod_{i=1}^{n_j} \frac{\mu_j(x_{ji}, \boldsymbol{\gamma}) p_j(x_{ji}, \boldsymbol{\beta}) g(x_{ji})}{\int \mu_j(x, \boldsymbol{\gamma}) p_j(x, \boldsymbol{\beta}) g(x) dx}. \quad (3)$$

If the selection probabilities  $\mu_j(x, \boldsymbol{\gamma})$ 's do not depend on the covariate vector  $x$ , then (3) becomes the likelihood of response-selective sampling (See Lee & Hirose, 2010; Scott & Wild, 1997). As noted in Ghosh & Dewanji (2011), and is also evident from individual terms in (3), we have  $P(X = x | R = 1, Y = j) = P(X = x | Y = j)$  in such case so that the selected (reported) sample can be taken as a representative sample from the corresponding population counterpart in the context of response-selective sampling. It is to be noted that, even in such situations without any selection or reporting bias, the model parameters are not identifiable (Cosslett, 1981), more so when such bias is present, unless some additional information is utilized. As indicated in Cosslett (1981), Scott & Wild (1997) and Lee et al. (2006), we consider only the exposure information from a reference sample of size  $n_{J+1}$ , say, drawn randomly from the source population, which augments with the  $(J + 1)$  response-selective samples for further analysis. This reference sample may be thought of as a random sample from the exposure distribution  $G(x)$ , say. In this work, we consider  $G(x)$  to be arbitrary and unspecified. In practice, in the context of pharmacovigilance studies, this may be of great advantage since information on drug use in the source population can be easily available in a random sample from prescription database (Mann, 1998), while information on ADR status is difficult to obtain.

The combined log-likelihood of  $(J + 1)$  response-selective samples along with the reference sample from  $G(x)$  can be written as

$$l(\boldsymbol{\phi}, G) = \sum_{j=0}^J \sum_{i=1}^{n_j} \log \frac{\mu_j(x_{ji}, \boldsymbol{\gamma}) p_j(x_{ji}, \boldsymbol{\beta}) g(x_{ji})}{\int \mu_j(x, \boldsymbol{\gamma}) p_j(x, \boldsymbol{\beta}) g(x) dx} + \sum_{i=1}^{n_{J+1}} \log \{g(x_i)\}, \quad (4)$$

where  $\boldsymbol{\phi} = (\boldsymbol{\gamma}, \boldsymbol{\beta})$  and  $x_i$ , for  $i = 1, \dots, n_{J+1}$ , are the observed exposure values in the reference sample. Following Gilbert et al. (1999), the likelihood (4) is a  $(J + 2)$ -sample selection bias model (See eq. (2.2) of Gilbert et al., 1999) with  $J \geq 1$  and the weight functions given by  $w_j(x, \boldsymbol{\phi}) = \mu_j(x, \boldsymbol{\gamma}) p_j(x, \boldsymbol{\beta})$ , for  $j = 0, \dots, J$ , and  $w_{J+1} = 1$ , independent of the model parameters. Then, by Theorem 2 of Gilbert et al. (1999), the model is identifiable if and only if  $w_j(x, \boldsymbol{\phi})$  and  $w_j(x, \boldsymbol{\phi}')$ , with  $\boldsymbol{\phi} \neq \boldsymbol{\phi}'$ , are linearly independent as functions of  $x$ , for at least one  $j = 0, \dots, J$ . Identifiability of the binary regression models with case-augmented samples, considered by Lee et al. (2006) for example, follows from this Theorem. The consequence of allowing selection or reporting bias through the reporting probabilities  $\mu_j(\cdot, \boldsymbol{\gamma})$ 's of (2) is the focus of this work. With  $(J + 1)$  response-selective samples subject to reporting bias and the reference sample, it is likely that there is at least one  $j \in \{0, \dots, J\}$  such that  $w_j(x, \boldsymbol{\phi})$  and  $w_j(x, \boldsymbol{\phi}')$ , for  $\boldsymbol{\phi} \neq \boldsymbol{\phi}'$ , are linearly independent. When, however,  $\mu_j(x, \boldsymbol{\gamma})$  is independent of  $x$  for all  $j$ , then  $w_j(x, \boldsymbol{\phi})$  and  $w_j(x, \boldsymbol{\phi}')$  are not linearly independent with  $\boldsymbol{\phi} = (\mu_j, \boldsymbol{\beta})$  and  $\boldsymbol{\phi}' = (\mu'_j, \boldsymbol{\beta})$ , where  $\mu_j \neq \mu'_j$ . The model parameters are not identifiable in such case. However, as remarked before, if the  $\mu_j$ 's are not of interest, the regression parameters in  $\boldsymbol{\beta}$  can be estimated from the response-selective samples without the need of the reference sample; also, as is evident from (3) and (4), the parameter  $\boldsymbol{\beta}$  can be estimated using the method of Lee et al. (2006) from the response-selective samples along with the reference sample. In the following section, we develop a semiparametric estimation procedure without making any assumption regarding the functional forms of  $\mu_j(\cdot, \boldsymbol{\gamma})$  and  $p_j(\cdot, \boldsymbol{\beta})$ , and address the identifiability issue for a specific form in Section 4.

### 3 Semiparametric estimation of model parameters

Note that the log-likelihood (4) is a function of  $\phi = (\gamma, \beta)$ , which is the parameter vector of interest, and of  $G(x)$ , the covariate distribution, which is the infinite dimensional nuisance parameter. In particular, the  $\beta$ -component of  $\phi$  is the quantity of interest. Nevertheless, in order to find the maximum likelihood estimate of  $\phi$ , we consider the profile log-likelihood  $l_p(\phi)$  of  $\phi$  obtained by maximizing the full log-likelihood (4) with respect to the nuisance parameter  $G(x)$  for fixed  $\phi$ . In this semiparametric framework, when  $G(x)$  is completely unspecified, the nonparametric maximum likelihood estimate of  $G(x)$  for fixed  $\phi$  is discrete with all its mass concentrated on the observed covariate values (Scott & Wild, 1997). We, therefore, work with the discrete distribution of  $X$  taking values in  $\{x_{01}, \dots, x_{0K}\}$ , say, which is the set of all observed distinct  $X$  values.

It is convenient to write  $A_j$  as the set of all  $X$  values with  $Y = j$ , for  $j = 0, \dots, J$ , and  $A_{J+1}$  as the set of  $X$  values in the reference sample. Also, let  $\delta_i$  denote the probability mass of  $X$  at  $x_{0i}$ , for  $i = 1, \dots, K$ , with  $\sum_{i=1}^K \delta_i = 1$ . Then, the log-likelihood (4) can be written in terms  $\delta = (\delta_1, \dots, \delta_K)$  as

$$l(\phi, \delta) = \sum_{j=0}^J \left[ \sum_{i \in A_j} n_{ji} \log(\mu_{ji} p_{ji} \delta_i) - n_j \log \sum_{k=1}^K \mu_{jk} p_{jk} \delta_k \right] + \sum_{i \in A_{J+1}} n_{J+1,i} \log \delta_i, \quad (5)$$

where  $\mu_{ji} = \mu_j(x_{0i}, \gamma)$ ,  $p_{ji} = p_j(x_{0i}, \beta)$  and  $n_{ji}$  is the frequency of  $x_{0i}$  with  $Y = j$ , for  $j = 0, \dots, J+1$ , with  $Y = J+1$  denoting the reference sample. The profile log-likelihood  $l_p(\phi)$  can be obtained from (5) as  $l(\phi, \hat{\delta}(\phi))$ , where  $\hat{\delta}(\phi)$  is the maximum likelihood estimate of  $\delta$  for given  $\phi$ . The semiparametric maximum likelihood estimate of  $\phi$  that maximizes the log-likelihood  $l(\phi, G)$  or  $l(\phi, \delta)$  can be obtained by maximizing the profile log-likelihood  $l_p(\phi)$ . As noted by Lee & Hirose (2010), this profile log-likelihood depends on a vector of arbitrary parameters  $(\rho_1, \dots, \rho_J)$ , where  $\log\{\int \mu_j(x, \gamma) p_j(x, \beta) g(x) dx / \int \mu_0(x, \gamma) p_0(x, \beta) g(x) dx\} = \rho_j$ , for  $j = 1, \dots, J$  (See also Scott & Wild, 2001). In this case, estimating all the parameters requires additional information on sampling fraction in each response category. In our context, however, such information is not available. Since  $\beta$  is the parameter of interest, we suggest a dimension-reduction approach, where a single offset parameter  $\rho$ , independent of  $j$ , is required instead of the arbitrary vector  $(\rho_1, \dots, \rho_J)$  (See Appendix A). Following Scott & Wild (1997), the estimator  $\hat{\phi}$  can be obtained as the solution of the pseudo log-likelihood equation  $\partial l^*(\psi) / \partial \psi = 0$ , where  $\psi = (\phi, \rho)$  and

$$l^*(\psi) = \sum_{j=0}^J \sum_{i \in A_j} n_{ji} \log \left( \frac{e^\rho p_{ji}^*}{1 + \sum_{l=0}^J e^\rho p_{li}^*} \right) + \sum_{i \in A_{J+1}} n_{J+1,i} \log \left( \frac{1}{1 + \sum_{l=0}^J e^\rho p_{li}^*} \right), \quad (6)$$

with  $p_{ji}^* = \mu_{ji} p_{ji}$  and  $\rho$  being the scalar nuisance parameter. This  $l^*(\psi)$  is a pseudo log-likelihood, derived from the log-likelihood (5) using profile log-likelihood approach. See the Appendix A for details. This pseudo log-likelihood (6) may be treated as that of a prospective sample of size  $n^* = \sum_{j=0}^{J+1} n_j$  from a multinomial distribution with  $(J+2)$  cells with the cell probabilities given by  $e^\rho p_{ji}^* / (1 + \sum_{l=0}^J e^\rho p_{li}^*)$ ,  $j = 0, \dots, J$ , and  $1 / (1 + \sum_{l=0}^J e^\rho p_{li}^*)$ , as function of the covariate value  $x_{0i}$ . The estimate  $\hat{\psi} = (\hat{\phi}, \hat{\rho})$  can be obtained by maximizing (6). Note that the model parameter  $\psi$  is non-identifiable from the likelihood (6) if and only if, for all  $x$ , for any two different values of  $\psi$ , the quantity  $e^\rho p_{ji}^*$  evaluated at these two values of  $\psi$ , are equal, for all  $j = 0, 1, \dots, J$ . Therefore, if this equality is violated for at least one  $j$ ,  $\psi$  becomes identifiable from (6).

Note that, from the derivation in Appendix A, the offset parameter  $\rho$  can be expressed in the form  $\rho = -\log(n_{J+1}/N)$ , where  $N$  is the size of the source population, which is unknown in our context. This has been verified in our simulation study also. However, this leads to an alternative estimation procedure in case the size  $N$  of the source population is known. In such case,  $\rho$  can be estimated by  $-\log(n_{J+1}/N)$ , which may be substituted in the pseudo log-likelihood (6). This requires one less parameter to be estimated. Our simulation study indicates that the resulting estimate may be more efficient.

Following Lee et al. (2006) and using multi-sample representation of Hirose (2005), the estimator  $\hat{\phi}$ , when suitably normalized, follows asymptotically a normal distribution under the standard regularity conditions. The asymptotic variance matrix of  $\hat{\phi}$  is estimated by the corresponding partition of the inverse of the observed information matrix  $-\partial^2 l^*(\phi, \rho)/\partial\phi\partial\rho$  evaluated at  $\phi = \hat{\phi}, \rho = \hat{\rho}$  (See Appendix B for a sketch of the proof). The estimate  $\hat{\phi}$  is also semiparametrically fully efficient in the sense that the asymptotic variance matrix of  $\hat{\phi}$  coincides with the corresponding semiparametric efficiency bound B, say. See Appendix C for details.

## 4 A Special case

A special case with  $J = 1$  considers the biased case and control samples, corresponding to  $j = 1$  and 0, respectively, along with a reference sample from the source population. As discussed in Section 1, this has application in pharmacovigilance studies in which the objective is to investigate strength of association between the drug of interest and the ADR of concern based on the SR database screened for those suffering from a particular disease for which the drug (exposure) is taken. The reference sample corresponding to  $j = J + 1 = 2$  is drawn randomly from the source population consisting of individuals suffering from the particular disease. The case sample ( $j = 1$ ) consists of those in the SR database reporting the ADR of concern, while the control sample ( $j = 0$ ) consists of those reporting other ADRs. Let us consider the modeling as given by the commonly used logit forms

$$\mu_j(x, \gamma) = \frac{e^{\gamma_j + \gamma x}}{1 + e^{\gamma_j + \gamma x}}, \text{ for } j = 0, 1, \quad (7)$$

$$\text{and } p_1(x, \beta) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}, \quad p_0(x, \beta) = 1 - p_1(x, \beta), \quad (8)$$

where  $\gamma = (\gamma_0, \gamma_1, \gamma)$  and  $\beta = (\alpha, \beta)$  are the parameters of interest with  $\phi = (\gamma, \beta)$ . Note that the regression parameter  $\gamma$  in the reporting probability  $\mu_j(x, \gamma)$ 's is the same for  $j = 0, 1$ , although the intercept parameters  $\gamma_0$  and  $\gamma_1$  are different.

As discussed in Section 2, the weight functions are given by

$$w_0(x, \phi) = \frac{e^{\gamma_0 + \gamma x}}{1 + e^{\gamma_0 + \gamma x}} \times \frac{1}{1 + e^{\alpha + \beta x}},$$

$$\text{and } w_1(x, \phi) = \frac{e^{\gamma_1 + \gamma x}}{1 + e^{\gamma_1 + \gamma x}} \times \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}},$$

with  $w_2(x, \phi) = 1$ . It can be verified that these weight functions satisfy the identifiability condition of Theorem 2 of Gilbert et al. (1999). Therefore, the model including exposure distribution  $G(x)$  is identifiable from the likelihood (3) and (4). As argued in Section 3, this also ensures identifiability of the model parameters  $\phi$  from pseudo log-likelihood (6).

In this special case, the pseudo log-likelihood (6) takes the form

$$\begin{aligned}
l^*(\boldsymbol{\psi}) &= \sum_{j=0}^1 \sum_{i \in A_j} n_{ji} [\rho + \gamma_j + j\alpha + (\gamma + j\beta)x_{0i}] + \sum_{j=0}^1 \sum_{i \in A_j} n_{ji} \log[1 + \exp(\gamma_{1-j} + \gamma x_{0i})] \\
&+ \sum_{i \in A_{J+1}} n_{ji} \log \left\{ [1 + \exp(\gamma_0 + \gamma x_{0i})][1 + \exp(\gamma_1 + \gamma x_{0i})][1 + \exp(\alpha + \beta x_{0i})] \right\} \\
&- \sum_{j=0}^2 \sum_{i \in A_j} n_{ji} \log \left[ 1 + \exp(\gamma_1 + \gamma x_{0i}) + \exp(\alpha + \beta x_{0i}) + \exp(\gamma_0 + \alpha + (\gamma + \beta)x_{0i}) \right. \\
&\left. + (1 + e^\rho) \left\{ \exp(\gamma_0 + \gamma x_{0i}) + \exp(\gamma_0 + \gamma_1 + 2\gamma x_{0i}) + \exp(\gamma_1 + \alpha + (\gamma + \beta)x_{0i}) \right. \right. \\
&\left. \left. + \exp(\gamma_0 + \gamma_1 + \alpha + (2\gamma + \beta)x_{0i}) \right\} \right]. \tag{9}
\end{aligned}$$

This can be maximized by using some numerical maximization procedure (for example, *optim* in R). This is found to have better numerical stability than a software to analyze multinomial logistic regression model as indicated by (6). In the next section, we carry out a simulation study to investigate properties of the estimator in which the data is generated from the model of this section and the pseudo log-likelihood (9) is maximized.

## 5 Simulation

For the purpose of the simulation, let us assume that the size  $N$  of the source population is 20,000. The exposure distribution  $G(x)$  is assumed to be exponential with mean 2. The case-control status of each individual is determined by using (8) with  $\alpha = -2.5$  and  $\beta = 0.5$ . The SR database is constructed by applying the reporting probability (7). We consider two sets of values for  $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma)$  as  $(-3.5, -4.5, 1.5)$  and  $(-3.5, -2.5, 1)$  to study the impact of reporting on the estimate of  $\beta$ , the parameter of interest, based on the combined data using the proposed method of Section 4 and only the SR database as well. These two choices of  $\boldsymbol{\gamma}$  reflect the two different scenarios in which the method based on only the SR database over- and under-estimates  $\beta$ , respectively. The average number of cases and controls in the simulated SR database are (2208, 4625) and (2518, 2882), respectively, for the two different choices of  $\boldsymbol{\gamma}$ . A reference sample of size  $n_{J+1}$  is drawn from the source population and the corresponding exposure values are recorded. The exposure values of the cases and the controls in the SR database along with those in the reference sample form a simulated dataset which is used to construct the pseudo log-likelihood (9). This is then maximized to obtain the semiparametric maximum likelihood estimates of the model parameters. The *Optim* procedure with BFGS method of the statistical software R has been used for this purpose. The corresponding standard errors are also obtained and a check is performed if the asymptotic 95% confidence intervals based on normal approximation contain the corresponding true parameter values. The estimate  $\hat{\beta}$  is also obtained along with its standard error from only SR database ignoring reporting bias and using the standard case-control analysis.

This process is repeated 5000 times. Since the objective is to investigate the relationship between exposure  $x$  and the response probability (8), the parameter of interest is  $\beta$ . In Table 1, the average of the estimates of  $\beta$  over the 5000 simulations along with its standard error in parentheses are presented. The average standard error (ASE) and sample standard error (SSE) obtained by the standard deviation of the 5000 estimates are found to be similar in each setting, as expected. The estimated coverage probabilities (CP) based on 5000 simulations are also presented.

Table 1: Simulation results on the estimate of  $\beta$  using SR database alone, SR database with the reference sample and the combined data with known source population size  $N = 20,000$ . True value of  $\beta$  is 0.5. Corresponding standard errors are in parentheses.

$n_{J+1}$	$(\gamma_0, \gamma_1, \gamma)$	SR Data alone $\hat{\beta}$ (SSE)	Combined data $\hat{\beta}$ (SSE, ASE, CP)	Combined data with known $N$ $\hat{\beta}$ (SSE, ASE, CP)
200	-3.5, -4.5, 1.5	0.633 ( 0.018 )	0.498 ( 0.037, 0.037, 0.951 )	0.499 ( 0.035, 0.036, 0.956 )
300		0.633 ( 0.018 )	0.500 ( 0.036, 0.037, 0.942 )	0.500 ( 0.035, 0.035, 0.951 )
500		0.633 ( 0.018 )	0.500 ( 0.036, 0.036, 0.951 )	0.501 ( 0.033, 0.034, 0.954 )
800		0.633 ( 0.018 )	0.500 ( 0.035, 0.035, 0.944 )	0.501 ( 0.033, 0.034, 0.956 )
1000		0.633 ( 0.018 )	0.499 ( 0.034, 0.035, 0.946 )	0.501 ( 0.034, 0.034, 0.947 )
1500		0.633 ( 0.019 )	0.501 ( 0.035, 0.034, 0.942 )	0.500 ( 0.033, 0.033, 0.945 )
2000		0.633 ( 0.018 )	0.500 ( 0.034, 0.034, 0.952 )	0.501 ( 0.033, 0.033, 0.953 )
200	-3.5, -2.5, 1	0.363 ( 0.014 )	0.510 ( 0.058, 0.057, 0.947 )	0.503 ( 0.046, 0.046, 0.951 )
300		0.363 ( 0.014 )	0.507 ( 0.054, 0.054, 0.948 )	0.503 ( 0.045, 0.045, 0.946 )
500		0.363 ( 0.014 )	0.507 ( 0.050, 0.050, 0.950 )	0.501 ( 0.044, 0.044, 0.949 )
800		0.363 ( 0.014 )	0.505 ( 0.047, 0.047, 0.952 )	0.501 ( 0.043, 0.043, 0.944 )
1000		0.363 ( 0.014 )	0.503 ( 0.045, 0.046, 0.952 )	0.501 ( 0.042, 0.043, 0.949 )
1500		0.363 ( 0.014 )	0.503 ( 0.043, 0.045, 0.945 )	0.500 ( 0.041, 0.042, 0.949 )
2000		0.363 ( 0.014 )	0.502 ( 0.043, 0.044, 0.946 )	0.500 ( 0.041, 0.042, 0.948 )

Depending on the value of  $\gamma$  as (-3.5, -4.5, 1.5) and (-3.5, -2.5, 1), the method based on only the SR data over- and under-estimates the parameter  $\beta$ , respectively, while the method based on combined data produces unbiased estimate in each setting. As expected, the standard error decreases with  $n_{J+1}$ . The estimated coverage probabilities are also close to 0.95. All this give evidence for consistency and asymptotic normality for the estimate obtained by the proposed method using combined data. When the source population size  $N$  is known, the parameter  $\rho$  can be replaced by  $-\log(n_{J+1}/N)$  in the pseudo log-likelihood (9), as remarked in Section 3, to reduce the number of parameters to be estimated. For this, we assume  $N = 20,000$  to be known and carry out the analysis with 5000 simulated datasets. The results are presented in Table 1, which are similar in nature with slight improvement in efficiency as expected.

In case the coefficient  $\gamma$  in reporting probability (7) is zero making the reporting probability independent of  $x$ , as discussed at the end of Section 2, identifiability condition is violated. In other words, when data is obtained through the model with  $\gamma = 0$ , or nearly zero, the proposed methodology may not give satisfactory result. However, when the size  $N$  of the source population is known, one can replace  $\rho$  by  $-\log(n_{J+1}/N)$  and the other parameters become identifiable from pseudo log-likelihood (9) with  $\gamma = 0$ . Nevertheless, since there is no reporting bias in such case, the parameter  $\beta$  can be estimated using the standard case-control analysis from only the SR database. This has been observed in our simulation study also (See Table 2).



Table 2: Simulation results on the estimate of  $\beta$ , when  $\gamma = 0$ , using SR database alone, SR database combined with the reference sample and known source population size  $N = 20,000$ . True value of  $\beta$  is 0.5. Corresponding standard errors are in parentheses.

$n_{J+1}$	$(\gamma_0, \gamma_1) = (-1, 0.4)$		$(\gamma_0, \gamma_1) = (0.3, -0.1)$	
	SR Data alone $\hat{\beta}$ (SSE)	Combined data with known $N$ $\hat{\beta}$ (SSE, ASE, CP)	SR Data alone $\hat{\beta}$ (SSE)	Combined data with known $N$ $\hat{\beta}$ (SSE, ASE, CP)
200	0.500 ( 0.016 )	0.500 ( 0.016, 0.016, 0.954 )	0.500 ( 0.014 )	0.500 ( 0.014, 0.014, 0.947 )
300	0.500 ( 0.016 )	0.501 ( 0.016, 0.016, 0.955 )	0.500 ( 0.014 )	0.500 ( 0.014, 0.014, 0.952 )
500	0.500 ( 0.015 )	0.500 ( 0.016, 0.016, 0.951 )	0.500 ( 0.014 )	0.500 ( 0.014, 0.014, 0.954 )
800	0.500 ( 0.015 )	0.500 ( 0.015, 0.016, 0.957 )	0.500 ( 0.014 )	0.500 ( 0.014, 0.014, 0.951 )
1000	0.500 ( 0.016 )	0.500 ( 0.016, 0.016, 0.950 )	0.500 ( 0.014 )	0.500 ( 0.014, 0.014, 0.953 )
1500	0.500 ( 0.016 )	0.500 ( 0.016, 0.016, 0.946 )	0.500 ( 0.014 )	0.500 ( 0.014, 0.014, 0.951 )
2000	0.501 ( 0.016 )	0.501 ( 0.016, 0.016, 0.945 )	0.500 ( 0.014 )	0.500 ( 0.014, 0.014, 0.947 )

## 6 An Example

We illustrate the methodology developed in this paper through the analysis of a spontaneous reporting data from the adverse event reporting system (AERS) maintained by the Food and Drug Administration (FDA) in US. The data contains those reported to AERS from all over the world between the time period 2006 to the first quarter of 2011 (<http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm082193.htm>). The source population consists of all the patients who are suffering from the disease ‘blood cholesterol increased’ and using the drug simvastatin. Our objective is to investigate the relationship among the doses of simvastatin with occurrence of the ADR myalgia, better known as muscle pain. There are 1011 reports obtained from AERS database during that period, among which 161 are suffering from the ADR myalgia considered as cases and other 850 individuals having some other ADRs considered as controls. Both the case and the control samples suffer from reporting bias, assumed to depend on doses of the drug due to spontaneous nature of reporting, as discussed in Section 1. Figure 1 (a) (b) illustrate the dose distribution of the drug simvastatin among the case and the control samples. It is clear that there is difference in dose distribution among the two samples, based on the reported data to AERS database. While the mode lies at 20 mg in the case sample, it is at 40 mg in the control sample. The distribution in the source population may be different.

We consider the reporting and the prospective model of (7) and (8), respectively, for this illustration. To apply the methodology, we need a reference sample having exposure information from the same source population. We do not have any such reference sample from the corresponding source population of AERS database. For the purpose of illustration, we consider a reference sample of size 4898, consisting of patients taking simvastatin, from The Norwegian Prescription Database (See Hartz et al., 2007) with the assumption that it represents the source population of this example. Figure 1 (c) gives the corresponding dose distribution which seems to be different from those in the case and control samples in Figure 1 (a) (b).

Using likelihood (9), the estimated model parameter  $\hat{\beta}$  is  $-0.051$  with corresponding standard error 0.0046. While ignoring the reporting bias, analysis based on only AERS database gives  $\hat{\beta} = -0.0054$  with corresponding standard error 0.0047. Therefore, it appears that the probability

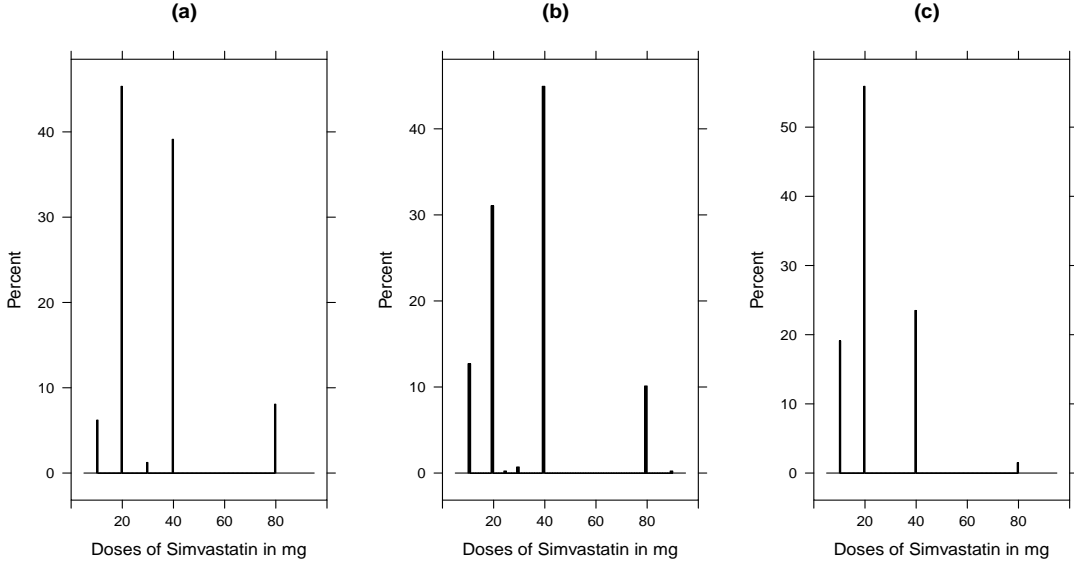


Figure 1: Dose distribution of the drug simvastatin among (a) 161 cases (b) 850 controls (c) reference sample of size 4898

of the ADR myalgia decreases significantly with higher doses of simvastatin, while ignoring the reporting bias makes this effect insignificant. Although this analysis indicates the importance of adjusting for reporting bias through a reference sample, the result should be interpreted with some caution. Ideally, a reference sample needs to be drawn from the source population under study. This may not have been satisfied in this illustration and the ‘unexpected’ result may be a consequence of this.

## Appendix A

### Pseudo Log-Likelihood

The ‘pseudo log-likelihood’ (6) is obtained from the log-likelihood (5). In order to obtain the profile likelihood, as discussed in Section 3, the log-likelihood (5) is maximized over  $\delta$  for fixed  $\phi$ . Introducing the Lagrange multiplier  $\lambda$  to take care of the constraint  $\sum_{i=1}^K \delta_i = 1$  and equating the derivative of the log-likelihood (5) with respect to  $\delta_i$  to zero, we get

$$\sum_{j=0}^J \left\{ \frac{n_{ji}}{\delta_i} - \frac{n_j \mu_{ji} p_{ji}}{\sum_{k=1}^K \mu_{jk} p_{jk} \delta_k} \right\} + \frac{n_{J+1,i}}{\delta_i} + \lambda = 0. \quad (10)$$

Multiplying (10) by  $\delta_i$  and summing over  $i$ , we have  $\lambda = -n_{J+1}$ . Using this value of  $\lambda$  in (10), the expression for  $\delta_i$  can be written as

$$\delta_i = \frac{n_{J+1,i} + \sum_{j=0}^J n_{ji}}{n_{J+1} \left[ 1 + \sum_{j=0}^J \frac{n_j}{n_{J+1}} \frac{\mu_{ji} p_{ji}}{\sum_{k=1}^K \mu_{jk} p_{jk} \delta_k} \right]}. \quad (11)$$

From (11), after setting an offset parameter  $\rho$  as

$$e^\rho = n_j / (n_{J+1} \sum_{i=1}^K \mu_{ji} p_{ji} \delta_i), \text{ for } j = 0, 1, \dots, J, \quad (12)$$

we have  $\delta_i = (n_{J+1,i} + \sum_{j=0}^J n_{ji}) / (n_{J+1}(1 + \sum_{j=0}^J e^{\rho} \mu_{ji} p_{ji}))$ , which is substituted in (5) to get the pseudo log-likelihood (6).

To justify this offset parameter  $\rho$  being independent of  $j$ , consider  $n_0/n_j$  as a consistent estimator of  $P(R=1, Y=0)/P(R=1, Y=j)$  (Scott & Wild, 1997) so that

$$\frac{n_0}{n_j} = \frac{P(R=1, Y=0)}{P(R=1, Y=j)} + o_p(1), \text{ for } j = 1, \dots, J. \quad (13)$$

However,  $n_0/n_j \rightarrow \omega_0/\omega_j$ , where  $n_j/n \rightarrow \omega_j$  as  $n \rightarrow \infty$  with  $n = \sum_{l=0}^{J+1} n_l$ , resulting in

$$\frac{\omega_0}{\omega_j} = \frac{P(R=1, Y=0)}{P(R=1, Y=j)},$$

which leads to

$$\frac{\omega_0}{\omega_{J+1}P(R=1, Y=0)} = \frac{\omega_j}{\omega_{J+1}P(R=1, Y=j)}, \text{ for } j = 1, \dots, J, \quad (14)$$

the population counter-part of (12). The implicit dependence of  $\rho$  on  $\phi$ , written as  $\rho = \rho(\phi)$ , is clear from the above description.

## Appendix B

### Asymptotics

The asymptotic properties of the estimator  $\hat{\psi} = (\hat{\phi}, \hat{\rho})$  obtained by maximizing the pseudo likelihood (6), may be accomplished by considering the multi-sample representation of Hirose (2005) and Lee et al. (2006). Let  $E_j$  denote the expectation with respect to the conditional distribution of exposure  $X$ , given  $Y = j$ , having density  $f_j(x, \phi, g) = \mu_j(x, \gamma) p_j(x, \beta) g(x) / \pi_j$  with  $\pi_j = \int \mu_j(x, \gamma) p_j(x, \beta) g(x) dx$ , for  $j = 0, \dots, J$ , and  $E_{J+1}$  denote the expectation with respect to the unconditional distribution of  $X$  having density  $f_{J+1}(x, \phi, g) = g(x)$ .

As in Lee et al. (2006), the estimating equation from (6), the pseudo likelihood equation, can be written as

$$\frac{\partial l^*(\psi)}{\partial \psi} = \sum_{j=0}^{J+1} \sum_{i=1}^{n_j} \frac{\partial \log Z_j(x_{ji}, \psi)}{\partial \psi} = 0, \quad (15)$$

where

$$\begin{aligned} Z_j(x, \psi) &= \left( \frac{e^{\rho} \mu_j(x, \gamma) p_j(x, \beta)}{1 + \sum_{l=0}^J e^{\rho} \mu_l(x, \gamma) p_l(x, \beta)} \right), \text{ for } j = 0, \dots, J, \text{ and} \\ Z_{J+1}(x, \psi) &= \left( \frac{1}{1 + \sum_{l=0}^J e^{\rho} \mu_l(x, \gamma) p_l(x, \beta)} \right). \end{aligned}$$

Note that, the  $x_{ji}$ , for  $i = 1, \dots, n_j$ , are independent random variables with common density  $f_j(x, \phi, g)$ , for  $j = 0, 1, \dots, J+1$ , as mentioned above. Then, we have

$$\begin{aligned} E_j \left( \frac{\partial \log Z_j}{\partial \psi} \right) &= \int \frac{1}{Z_j} \frac{\partial Z_j}{\partial \psi} \frac{\mu_j(x, \gamma) p_j(x, \beta) g(x)}{\pi_j} dx \\ &= \int \left( 1 + \sum_{l=0}^J e^{\rho} \mu_l(x, \gamma) p_l(x, \beta) \right) \frac{\partial Z_j}{\partial \psi} \frac{g(x)}{e^{\rho} \pi_j} dx \\ &= \frac{1}{\omega_j} E_X \left[ \omega_{J+1} \left( 1 + \sum_{l=0}^J e^{\rho} \mu_l(X, \gamma) p_l(X, \beta) \right) \frac{\partial Z_j}{\partial \psi} \right], \quad (16) \end{aligned}$$

for  $j = 0, 1, \dots, J$ , using (12) and (14).

$$\text{Similarly, } E_{J+1} \left( \frac{\partial \log Z_{J+1}}{\partial \boldsymbol{\psi}} \right) = \frac{1}{\omega_{J+1}} E_X \left[ \omega_{J+1} \left( 1 + \sum_{l=0}^J e^{\rho} \mu_l(X, \boldsymbol{\gamma}) p_l(X, \boldsymbol{\beta}) \right) \frac{\partial Z_{J+1}}{\partial \boldsymbol{\psi}} \right]. \quad (17)$$

Hence,

$$\sum_{j=0}^{J+1} \omega_j E_j \left[ \frac{\partial \log Z_j}{\partial \boldsymbol{\psi}} \right] = 0, \text{ since } \sum_{j=0}^{J+1} Z_j = 1. \quad (18)$$

Then, under some suitable regularity conditions (See Hirose, 2005, p67-79), the asymptotic distribution of  $\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})$  is multivariate normal with mean zero and variance-covariance matrix given by

$$I(\boldsymbol{\psi})^{-1} \Sigma I(\boldsymbol{\psi})^{-1}, \quad (19)$$

where

$$\begin{aligned} I(\boldsymbol{\psi}) &= \sum_{j=0}^{J+1} \omega_j E_j \left[ -\frac{\partial^2 \log Z_j}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T} \right] \text{ and} \\ \Sigma &= \sum_{j=0}^{J+1} \omega_j E_j \left[ \left\{ \frac{\partial \log Z_j}{\partial \boldsymbol{\psi}} - E_j \left( \frac{\partial \log Z_j}{\partial \boldsymbol{\psi}} \right) \right\} \left\{ \frac{\partial \log Z_j}{\partial \boldsymbol{\psi}} - E_j \left( \frac{\partial \log Z_j}{\partial \boldsymbol{\psi}} \right) \right\}^T \right]. \end{aligned} \quad (20)$$

In our context, it can be shown that the variance-covariance matrix of (19) has the form

$$I(\boldsymbol{\psi})^{-1} - \begin{bmatrix} 0 & 0 \\ 0^T & H \end{bmatrix}, \quad (21)$$

where  $H$  is a scalar element. The resulting variance-covariance matrix of  $\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})$  is  $[I_{\boldsymbol{\phi}\boldsymbol{\phi}} - I_{\boldsymbol{\phi}\rho} I_{\rho\rho}^{-1} I_{\rho\boldsymbol{\phi}}]^{-1}$ , where  $I(\boldsymbol{\psi})$  is partitioned as

$$I(\boldsymbol{\psi}) = \begin{bmatrix} I_{\boldsymbol{\phi}\boldsymbol{\phi}} & I_{\boldsymbol{\phi}\rho} \\ I_{\rho\boldsymbol{\phi}}^T & I_{\rho\rho} \end{bmatrix}. \quad (22)$$

Note that  $nI(\boldsymbol{\psi})$  can be consistently estimated by  $-\partial^2 l^*(\boldsymbol{\psi})/\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T$  evaluated at  $\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}$ . From (20),  $\Sigma$  can be written as

$$\begin{aligned} & \sum_{j=0}^{J+1} \omega_j E_j \left[ \left( \frac{\partial \log Z_j}{\partial \boldsymbol{\psi}} \right) \left( \frac{\partial \log Z_j}{\partial \boldsymbol{\psi}} \right)^T \right] - \sum_{j=0}^{J+1} \omega_j E_j \left( \frac{\partial \log Z_j}{\partial \boldsymbol{\psi}} \right) E_j \left( \frac{\partial \log Z_j}{\partial \boldsymbol{\psi}} \right)^T \\ &= I(\boldsymbol{\psi}) - \sum_{j=0}^{J+1} \omega_j E_j \left( \frac{\partial \log Z_j}{\partial \boldsymbol{\psi}} \right) E_j \left( \frac{\partial \log Z_j}{\partial \boldsymbol{\psi}} \right)^T. \end{aligned} \quad (23)$$

To establish (21) by using (19), we need to show that the second term of (23) is (See Neuhaus et al., 2002)

$$I(\boldsymbol{\psi}) \begin{bmatrix} 0 & 0 \\ 0^T & H \end{bmatrix} I(\boldsymbol{\psi}). \quad (24)$$

Note that,

$$E_j \left( \frac{\partial \log Z_j}{\partial \boldsymbol{\psi}} \right) \left( \frac{\partial \log Z_j}{\partial \boldsymbol{\psi}} \right)^T = \frac{1}{\omega_j} E_X \left[ \omega_{J+1} \left( 1 + \sum_{l=0}^J e^{\rho} \mu_l(X, \boldsymbol{\gamma}) p_l(X, \boldsymbol{\beta}) \right) \frac{1}{Z_j} \frac{\partial Z_j}{\partial \boldsymbol{\psi}} \frac{\partial Z_j}{\partial \boldsymbol{\psi}^T} \right], \quad (25)$$

for  $j = 0, \dots, J+1$ . Using (16) and (17), the second term of (23) becomes

$$\sum_{j=0}^{J+1} \frac{1}{\omega_j} E_X \left[ \omega_{J+1} T(X) \frac{\partial Z_j}{\partial \boldsymbol{\psi}} \right] E_X \left[ \omega_{J+1} T(X) \frac{\partial Z_j}{\partial \boldsymbol{\psi}} \right], \quad (26)$$

where  $T(X) = \left( 1 + \sum_{l=0}^J e^{\rho} \mu_l(X, \boldsymbol{\gamma}) p_l(X, \boldsymbol{\beta}) \right)$ . Using (25), the information matrix can be written as

$$I(\boldsymbol{\psi}) = \sum_{j=0}^{J+1} E_X \left[ \omega_{J+1} T(X) \frac{1}{Z_j} \frac{\partial Z_j}{\partial \boldsymbol{\psi}} \frac{\partial Z_j}{\partial \boldsymbol{\psi}^T} \right] = E_X \left[ \omega_{J+1} T(X) \sum_{j=0}^{J+1} \frac{1}{Z_j} \frac{\partial Z_j}{\partial \boldsymbol{\psi}} \frac{\partial Z_j}{\partial \boldsymbol{\psi}^T} \right]. \quad (27)$$

Now,

$$\begin{aligned} I_{\phi\phi} &= E_X \left[ \omega_{J+1} T(X) \sum_{j=0}^{J+1} \frac{1}{Z_j} \frac{\partial Z_j}{\partial \boldsymbol{\phi}} \frac{\partial Z_j}{\partial \boldsymbol{\phi}^T} \right] \\ I_{\rho\phi} &= E_X \left[ \omega_{J+1} T(X) \sum_{j=0}^{J+1} \frac{1}{Z_j} \frac{\partial Z_j}{\partial \rho} \frac{\partial Z_j}{\partial \boldsymbol{\phi}^T} \right] \\ &= E_X \left[ \omega_{J+1} T(X) \left\{ \sum_{j=0}^J Z_{J+1} \frac{\partial Z_j}{\partial \boldsymbol{\phi}^T} - (1 - Z_{J+1}) \frac{\partial Z_{J+1}}{\partial \boldsymbol{\phi}^T} \right\} \right] \\ &= E_X \left[ \omega_{J+1} T(X) \left\{ -Z_{J+1} \frac{\partial Z_{J+1}}{\partial \boldsymbol{\phi}^T} - (1 - Z_{J+1}) \frac{\partial Z_{J+1}}{\partial \boldsymbol{\phi}^T} \right\} \right] \\ &= -E_X \left[ \omega_{J+1} T(X) \frac{\partial Z_{J+1}}{\partial \boldsymbol{\phi}^T} \right] \\ I_{\rho\rho} &= E_X \left[ \omega_{J+1} T(X) \left\{ \sum_{j=0}^J Z_j Z_{J+1}^2 - Z_{J+1} (1 - Z_{J+1})^2 \right\} \right] \\ &= E_X \left[ \omega_{J+1} T(X) \left\{ -(1 - Z_{J+1}) Z_{J+1}^2 - Z_{J+1} (1 - Z_{J+1})^2 \right\} \right] \\ &= -E_X \left[ \omega_{J+1} T(X) \frac{\partial Z_{J+1}}{\partial \rho} \right], \end{aligned}$$

using the results,

$$\frac{\partial Z_j}{\partial \rho} = Z_j Z_{J+1} \quad \text{and} \quad \frac{\partial Z_{J+1}}{\partial \rho} = -Z_{J+1} (1 - Z_{J+1}). \quad (28)$$

Since the last column of  $I(\boldsymbol{\psi})$  is  $-E_X \left[ \omega_{J+1} T(X) \frac{\partial Z_{J+1}}{\partial \boldsymbol{\psi}} \right]$ , it can be checked that (See Neuhaus et al., 2002)

$$-I(\boldsymbol{\psi})^{-1} E_X \left[ \omega_{J+1} T(X) \frac{\partial Z_{J+1}}{\partial \boldsymbol{\psi}} \right] = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}. \quad (29)$$

From the relationship

$$\sum_{j=0}^J E_X \left[ \omega_{J+1} T(X) \frac{\partial Z_j}{\partial \boldsymbol{\psi}} \right] = -E_X \left[ \omega_{J+1} T(X) \frac{\partial Z_{J+1}}{\partial \boldsymbol{\psi}} \right], \quad (30)$$

we claim that

$$E_X \left[ \omega_{J+1} T(X) \frac{\partial Z_j}{\partial \boldsymbol{\psi}} \right] = -\tau_j(\boldsymbol{\psi}) E_X \left[ \omega_{J+1} T(X) \frac{\partial Z_{J+1}}{\partial \boldsymbol{\psi}} \right], \quad (31)$$

where  $\sum_{j=0}^J \tau_j(\boldsymbol{\psi}) = 1$ . Using (29) and (31), we have

$$I(\boldsymbol{\psi})^{-1} E_X \left[ \omega_{J+1} T(X) \frac{\partial Z_j}{\partial \boldsymbol{\psi}} \right] = \tau_j(\boldsymbol{\psi}) \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}. \quad (32)$$

From (29) and (32),

$$\sum_{j=0}^{J+1} \frac{1}{\omega_j} I(\boldsymbol{\psi})^{-1} E_X \left[ \omega_{J+1} T(X) \frac{\partial Z_j}{\partial \boldsymbol{\psi}} \right] E_X \left[ \omega_{J+1} T(X) \frac{\partial Z_j}{\partial \boldsymbol{\psi}^T} \right] I(\boldsymbol{\psi})^{-1} = \left( \frac{1}{\omega_{J+1}} + \sum_{j=0}^J \tau_j(\boldsymbol{\psi}) \right) \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}$$

Hence, using (23), (21) is established.

## Appendix C

### Semiparametric Efficiency

Following Bickel et al. (1993), the asymptotic variance matrix for a regular asymptotically linear (RAL) estimate  $\hat{\boldsymbol{\phi}}$  of  $\boldsymbol{\phi}$  satisfies  $V(\hat{\boldsymbol{\phi}}) \geq B$ , where  $B$  is the semiparametric efficiency bound. Lee & Hirose (2010) have obtained this bound  $B$  for the semiparametric maximum likelihood estimate of parameters in general regression model when data is collected under response-selective sampling scheme. In order to apply their results (Theorem 2) in our context, let us consider the ‘‘population expected likelihood’’ (See also Newey, 1990; Lee et al., 2006) as given by

$$\sum_{j=0}^{J+1} \omega_j E_j [\log f_j(X, \boldsymbol{\phi}, g)], \quad (33)$$

where  $E_j$  is the expectation with respect to the conditional distribution of exposure  $X$ , given  $Y = j$ , having density  $f_j(x, \boldsymbol{\phi}, g) = \mu_j(x, \boldsymbol{\gamma}) p_j(x, \boldsymbol{\beta}) g(x) / \pi_j$  with  $\pi_j = \int \mu_j(x, \boldsymbol{\gamma}) p_j(x, \boldsymbol{\beta}) g(x) dx$ , for  $j = 0, \dots, J$  and  $E_{J+1}$  is the expectation with respect to the unconditional distribution of  $X$  having density  $f_{J+1}(x, \boldsymbol{\phi}, g) = g(x)$ , where  $g(x)$  is the density corresponding to the covariate distribution  $G(x)$  of  $X$ . Then, the efficient scores are given by

$$S_j = \left. \frac{\partial \log f_j(x, \boldsymbol{\phi}, \hat{g}(\boldsymbol{\phi}))}{\partial \boldsymbol{\phi}} \right|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}, \quad j = 0, 1, \dots, J+1, \quad (34)$$

where  $\hat{g}(\boldsymbol{\phi})$  is the maximizer of (33), for fixed  $\boldsymbol{\phi}$ . Then, the corresponding efficiency bound  $B$  is given by

$$B^{-1} = \sum_{j=0}^{J+1} \omega_j E_j (S_j S_j^T). \quad (35)$$

To show that the asymptotic variance matrix of  $\hat{\phi}$  is equal to the semiparametric efficiency bound, we need to show that

$$B^{-1} = I_{\phi\phi} - I_{\phi\rho}I_{\rho\rho}^{-1}I_{\rho\phi}. \quad (36)$$

From (33), the expected loglikelihood

$$\begin{aligned} \sum_{j=0}^J \omega_j \int \{ \log(\mu_j(x, \gamma)p_j(x, \beta)) + \log(g(x)) \} & \frac{\mu_j(x, \gamma)p_j(x, \beta)g_0(x)dx}{\pi_j^0} \\ & + \omega_{J+1} \int \{ \log(g(x)) \} g_0(x)dx - \sum_{j=0}^J \omega_j \log \pi_j, \end{aligned} \quad (37)$$

where  $\pi_j^0 = \int \mu_j(x, \gamma)p_j(x, \beta)g_0(x)dx$ . Considering the terms which involve  $g(x)$ , (37) can be written as

$$\omega_{J+1} \int \log(g(x))\tilde{p}(x)g_0(x)dx - \sum_{j=0}^J \omega_j \log \pi_j, \quad \text{where } \tilde{p}(x) = 1 + \sum_{j=0}^J \frac{\omega_j}{\omega_{J+1}} \frac{\mu_j(x, \gamma)p_j(x, \beta)}{\pi_j^0}. \quad (38)$$

Now, we need to find  $\hat{g}$  which maximizes (38). Consider the class of distribution of  $X$  to be discrete with finite support  $\{x_1, \dots, x_M\}$ . Suppose a general member  $g(\cdot)$  of this class has mass  $g_i$  at  $x_i$ . Note that the true distribution  $g_0(\cdot)$  is a member of this class having mass  $g_0(x_i)$ , say, at  $x_i$ . Then, (38) can be written along with Lagrange multiplier  $\lambda$  to take care of the constraint  $\sum_{i=1}^M g_i = 1$ ,

$$\omega_{J+1} \sum_{i=1}^M \log(g_i)\tilde{p}(x_i)g_0(x_i) - \sum_{j=0}^J \omega_j \log \pi_j(\mathbf{g}) + \lambda \left( \sum_{i=1}^M g_i - 1 \right), \quad (39)$$

where  $\pi_j(\mathbf{g}) = \sum_{i=1}^M \mu_{ji}p_{ji}g_i$ . Differentiating (39) with respect to  $g_i$ , we have

$$\omega_{J+1} \frac{\tilde{p}(x_i)g_0(x_i)}{g_i} - \sum_{j=0}^J \omega_j \frac{\mu_j(x_i, \gamma)p_j(x_i, \beta)}{\pi_j(\mathbf{g})} + \lambda = 0. \quad (40)$$

Multiplying (40) by  $g_i$  and summing over  $i$  gives  $\lambda = -\omega_{J+1}$ . Putting the value of  $\lambda$  in (40), we get the estimate of  $g_i$  as

$$\hat{g}_i = \frac{\tilde{p}(x_i)g_0(x_i)}{1 + \sum_{j=0}^J \frac{\omega_j}{\pi_j(\mathbf{g})\omega_{J+1}} \mu_j(x_i, \gamma)p_j(x_i, \beta)}. \quad (41)$$

In case of general  $g$ , not having finite support, the maximizer of (38) is of the form

$$\hat{g}(x; \phi, \rho) = \frac{\tilde{p}(x)g_0(x)}{1 + \sum_{j=0}^J \frac{\omega_j}{\pi_j(\rho)\omega_{J+1}} \mu_j(x, \gamma)p_j(x, \beta)} \quad (42)$$

(See Lee & Hirose, 2010; Lee et al., 2006), where  $\pi_j(\rho)$  satisfies  $e^\rho = \omega_j/(\pi_j(\rho)\omega_{J+1})$  (See (14)) for  $j = 0, 1, \dots, J$  and  $\rho = \rho(\phi)$  is the solution of the equations

$$\begin{aligned} \int \mu_j(x, \gamma)p_j(x, \beta)\hat{g}(x; \phi, \rho)dx & = \pi_j(\rho) \\ \text{or, equivalently, } \int \left\{ \frac{e^\rho \mu_j(x, \gamma)p_j(x, \beta)}{1 + \sum_{j=0}^J e^\rho \mu_j(x, \gamma)p_j(x, \beta)} \right\} \tilde{p}(x)g_0(x)dx & = \frac{\omega_j}{\omega_{J+1}}, \quad j = 0, 1, \dots, J. \end{aligned} \quad (43)$$

Putting the value of  $\hat{g}$  in the densities  $f_j(x, \phi, g)$ , for  $j = 0, 1, \dots, J$ , we have

$$\log f_j(x, \phi, \hat{g}) = \log \left\{ \frac{e^{\rho(\phi)} \mu_j(x, \gamma) p_j(x, \beta)}{1 + \sum_{j=0}^J e^{\rho(\phi)} \mu_j(x, \gamma) p_j(x, \beta)} \right\} + c_j = \log \{q_j(x, \phi, \rho(\phi))\} + c_j \quad (44)$$

and

$$\begin{aligned} \log f_{J+1}(x, \phi, \hat{g}) &= \log \left\{ \frac{1}{1 + \sum_{j=0}^J e^{\rho(\phi)} \mu_j(x, \gamma) p_j(x, \beta)} \right\} + c_{J+1} \\ &= \log \left\{ 1 - \sum_{j=0}^J q_j(x, \phi, \rho(\phi)) \right\} + c_{J+1}, \end{aligned} \quad (45)$$

where  $c_j$ 's are constants with respect to  $\psi = (\rho, \phi)$ . Using (34), the efficient scores are given as

$$S_j = \frac{\partial}{\partial \phi} \log \{q_j(x, \phi, \rho(\phi))\}, \text{ for } j = 0, 1, \dots, J+1, \quad (46)$$

where  $q_{J+1}(x, \phi, \rho(\phi)) = 1 - \sum_{j=0}^J q_j(x, \phi, \rho(\phi))$  and all the derivatives are evaluated at  $\phi = \phi_0$ . Applying chain rule,

$$S_j = \frac{\partial}{\partial \phi} \log \{q_j(x, \phi, \rho(\phi))\} + \left\{ \frac{\partial \rho(\phi)}{\partial \phi} \right\}^T \frac{\partial}{\partial \rho} \log \{q_j(x, \phi, \rho(\phi))\}. \quad (47)$$

Note that the information matrix (See Lee et al., 2006) is given by

$$I(\psi) = I(\rho, \phi) = \sum_{j=0}^{J+1} \omega_j E_j \left\{ \left( \frac{\partial}{\partial \psi} \log q_j \right) \left( \frac{\partial}{\partial \psi} \log q_j \right)^T \right\}. \quad (48)$$

From (35) and (47), we get

$$B^{-1} = I_{\phi\phi} + \left\{ \frac{\partial \rho(\phi)}{\partial \phi} \right\}^T I_{\rho\phi} + I_{\phi\rho} \left\{ \frac{\partial \rho(\phi)}{\partial \phi} \right\} + \left\{ \frac{\partial \rho(\phi)}{\partial \phi} \right\}^T I_{\rho\rho} \left\{ \frac{\partial \rho(\phi)}{\partial \phi} \right\}. \quad (49)$$

Differentiating (43) under the integral sign

$$\int \frac{\partial q_j}{\partial \phi} \tilde{p} g dx + \left\{ \frac{\partial \rho(\phi)}{\partial \phi} \right\}^T \int \frac{\partial q_j}{\partial \rho} \tilde{p} g dx = 0, j = 0, 1, \dots, J. \quad (50)$$

It can be easily checked that

$$\frac{\partial q_j}{\partial \rho} = q_j \left( 1 - \sum_{j=0}^J q_j \right); -\frac{\partial^2 \log q_j}{\partial \rho^2} = \sum_{j=0}^J \frac{\partial q_j}{\partial \rho} \text{ and } -\frac{\partial}{\partial \phi} \left\{ \frac{\partial \log q_j}{\partial \rho} \right\} = \sum_{j=0}^J \frac{\partial q_j}{\partial \phi}. \quad (51)$$

Now, using (48) and (51),

$$I_{\rho\rho} = \omega_{J+1} \int \sum_{j=0}^J \frac{\partial q_j}{\partial \rho} \tilde{p} g dx \text{ and } I_{\rho\phi} = \omega_{J+1} \int \sum_{j=0}^J \frac{\partial q_j}{\partial \phi} \tilde{p} g dx. \quad (52)$$

Summing over  $j = 0, 1, \dots, J$  in (50) and using (51)

$$\frac{\partial \rho(\phi)}{\partial \phi} = -I_{\rho\rho}^{-1} I_{\rho\phi}. \quad (53)$$

From (35), (49) and (53)

$$B^{-1} = I_{\phi\phi} - I_{\phi\rho} I_{\rho\rho}^{-1} I_{\rho\phi}. \quad (54)$$

This establish the efficiency bound of the semiparametric procedure.



## References

- BATE, A., LINDQUIST, M., EDWARDS, I. R., OLSSON, S., ORRE, R., LANSNER, A. & FREITAS, R. M. D. (1998). A bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology* **54**, 315–321.
- BICKEL, P. J., KLAASSEN, C. A., RITOV, Y. & WELLNER, J. A. (1993). *Efficient and adaptive estimation for semi-parametric models*. Baltimore: Johns Hopkins University Press.
- BRESLOW, N. E. (1996). Statistics in Epidemiology: the case-control study. *Journal of American Statistical Association* **91**, 14–28.
- COSSLETT, S. R. (1981). Maximum likelihood estimator for choice-based samples. *Econometrica* **49**, 1289–1316.
- GHOSH, P. & DEWANJI, A. (2011). Analysis of spontaneous adverse drug reaction (ADR) reports using supplementary information. *Statistics in Medicine* **30**, 2040–2055.
- GILBERT, P. B., LELE, S. R. & VARDI, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika* **86**, 27–43.
- HARTZ, I., SAKSHAUG, S., FURU, K., ENGELAND, A., EGGEN, A. E., NJLSTAD, I. & SKURTVEIT, S. (2007). Aspects of statin prescribing in norwegian counties with high, average and low statin consumption-an individual-level prescription database study. *BMC Clinical Pharmacology* **7**.
- HIROSE, Y. (2005). *Efficiency of the semi-parametric maximum likelihood estimator in generalized case-control studies*. Ph.D. thesis, Univ. Auckland.
- HSIEH, D. A., MANSKI, C. F. & MCFADDEN, D. (1985). Estimation of response probabilities from augmented retrospective observations. *Journal of American Statistical Association* **80**, 651–662.
- LEE, A. & HIROSE, Y. (2010). Semi-parametric efficiency bounds for regression models under response-selective sampling: the profile likelihood approach. *Ann Inst Stat Math* **62**, 1023–1052.
- LEE, A. J., SCOTT, A. J. & WILD, C. J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika* **93**, 385–397.
- MANN, R. D. (1998). Prescription-event monitoring-recent progress and future horizons. *British Journal of Clinical Pharmacology* **46**, 195–201.
- NEUHAUS, J., SCOTT, A. J. & WILD, C. J. (2002). The analysis of retrospective family studies. *Biometrika* **89**, 23–37.
- NEWBY, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5**, 99–135.
- PRENTICE, R. L. & BRESLOW, N. E. (1978). Retrospective studies and failure time models. *Biometrika* **65**, 153–158.
- PRENTICE, R. L. & PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.
- SCOTT, A. J. & WILD, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57–71.

SCOTT, A. J. & WILD, C. J. (2001). Maximum likelihood for generalised case-control studies.  
*Journal of Statistical Planning and Inference* **96**, 3–27.