Ann Inst Stat Math (2016) 68:805–825
DOI 10.1007/s10463-015-0511-3

Regression analysis of biased case–control data

Palash Ghosh · Anup Dewanji

Received: 4 September 2013 / Revised: 2 February 2015 / Published online: 17 March 2015

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 this bias is dealt with by introducing the concept of reporting probability. Then, considering a reference sample from the source population, we obtain asymptotically unbiased estimate of the population parameters by fitting a pseudo-likelihood, assuming the exposure distribution in the population to be unknown and arbitrary. The proposed estimates of the model parameters follow asymptotically a normal distribution and are semiparametrically 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.

Keywords Reporting bias · Response-selective sampling · Spontaneous reporting database · Semiparametric estimation · Pseudo-likelihood

1 Introduction

Prentice and 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 carried out in various modifications depending on practical requirements. In this context, Hsieh et al. (1985), Scott and Wild (1997) and Lee et al. (2006) have

P. Ghosh
Centre for Quantitative Medicine, Duke-NUS Graduate Medical School,
Academia-Level 6, 20 College Road, Singapore 169856, Singapore
e-mail: palash.ghosh@duke-nus.edu.sg

A. Dewanji (✉)
Applied Statistics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India
e-mail: dewanjia@isical.ac.in

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 and 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 represent a biased case–control sample of the corresponding counterparts in the source population, defined by the collection of individuals suffering from a particular disease (Ghosh and Dewanji 2011). 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 addresses 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 and Dewanji (2011) explicitly incorporate 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 analyzing ADR data from SR database. The main objective of this paper is to deal with this problem of reporting bias in the regression framework when the exposure can be continuous with only exposure information from the reference sample. In what follows, we also consider multiple response categories, in which the term ‘case–control sampling’ may be replaced by ‘response-selective sampling’ (Lee and Hirose 2010), or choice-based sampling (Cosslett 1981). In Sect. 2, we describe the model along with the selection or reporting probabilities and discuss some existing methods in the similar context. 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 is considered in Sect. 4. Section 5 presents a simulation study to investigate the properties of the estimates, while Sect. 6 illustrates the method through analysis of an example.

2 Modeling and likelihood

Consider a categorical random variable Y, representing the ADR status, having categories $j = 0, \ldots, J$, the distribution of which depends on the vector of covariates X, amount of the drug and others, of dimension $p \geq 1$ through a prospective model of the form

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

for $j = 0, \ldots, J$, 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 in the source population may be reported (selected) to the SR database with certain probability depending even on X as given by the model

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

where R is the binary random variable taking values 1 or 0 representing reporting to the SR database or not and $\boldsymbol{\gamma}$ is the corresponding vector of parameters.

Suppose in the SR database, there are n_j individuals in the sample with $Y = j$ having the observed X values as x_{ji}, for $i = 1, \ldots, n_j$, and $j = 0, \ldots, J$. Note that this observed biased data can be thought of as arising from a random prospective sample of ADRs which are then reported or not with probability $\mu_j(x, \gamma)$. Also, let $g(x)$ denote the marginal density of x. Then, the retrospective likelihood for the SR data is given by

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

If the reporting (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 and Hirose 2010; Scott and Wild 1997). As noted in Ghosh and 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 reported (selected) sample in the SR database 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 and 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 is a random prospective sample from the exposure distribution $G(x)$, say, with density g(.). In this work, we consider $G(x)$ to be arbitrary and unspecified. Note that the likelihood contribution corresponding to the reference sample is free of reporting probability term $\mu_{J+1}(x_{J+1,i}, \boldsymbol{\gamma})$.

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 from the SR database along with the reference sample from $G(x)$ can be written as

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

where $\boldsymbol{\phi} = (\boldsymbol{\gamma}, \boldsymbol{\beta})$ and x_i, for $i = 1, \ldots, n_{J+1}$, 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 $j = 0, \ldots, J$, and $W_{J+1} = 1$, independent of the model parameters. Then, by Theorem 2 of Gilbert et al. (1999), the model (that is, both ϕ and G) 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 $j = 0, \ldots, J$. 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 $\mu_j(., \boldsymbol{\gamma})$'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 $j \in \{0, \ldots, J\}$ 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 μ_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. As remarked in Sect. 1, however, the reporting probabilities $\mu_j(x, \gamma)$'s often depend on both j and x.

Note that the log-likelihood (4) is a function of $\boldsymbol{\phi} = (\boldsymbol{\gamma}, \boldsymbol{\beta})$, which is the parameter vector of interest, and of $G(x)$, the exposure distribution, which is the infinite-dimensional nuisance parameter. In particular, the $\boldsymbol{\beta}$-component of $\boldsymbol{\phi}$ is the quantity of interest. Since estimation of $G(x)$ is not of particular interest, one does not need to carry out the maximum likelihood estimation procedure of Gilbert et al. (1999) for joint estimation of ϕ and $G(x)$. Instead, in order to estimate ϕ alone, a simpler method adjusting for the unknown $G(x)$ may be adopted. Nevertheless, as discussed in Sect. 3, the identifiability condition of Gilbert et al. (1999) still remains valid. In the following section, we develop a semiparametric estimation procedure without making any assumption regarding the functional forms of $\mu_j(., \boldsymbol{\gamma})$ and $p_j(., \boldsymbol{\beta})$, and address the identifiability issue for a specific form in Sect. 4.

Scott and Wild (1997, 2001) and Lee et al. (2006) have considered similar problems with arbitrary $G(x)$ and binary classification (that is, $J = 1$) assuming no selection or reporting bias and developed a pseudo-likelihood-based approach to obtain semiparametric maximum likelihood estimate of the regression parameters for different data configurations (see also Wild 1991). Lee and Hirose (2010) have extended their work for multiple classification and established semiparametric efficiency of the estimates. However, in the presence of selection or reporting bias, these methods may lead to biased estimates. On the other hand, incorporation of reporting probabilities, as given by (2), in these works may lead to identifiability problem. For example, as argued in Sect. 4, the method of Lee et al. (2006) based on case-augmented sample will have identifiability problem if the reporting probabilities are included. Note that the data configuration in the present work is also somewhat different from those of the works mentioned above (see Sect. 3). Our estimation procedure is based on a different pseudo-likelihood approach to alleviate the problem of some lack of information, requiring only one offset parameter unlike that in Lee and Hirose (2010).

3 Semiparametric estimation of model parameters

In order to find the maximum likelihood estimate of $\boldsymbol{\phi}$, we consider the profile log-likelihood $l_p(\boldsymbol{\phi})$ of $\boldsymbol{\phi}$ obtained by maximizing the full log-likelihood (4) with respect to the nuisance parameter $G(x)$ for fixed $\boldsymbol{\phi}$. In this semiparametric framework, when $G(x)$ is completely unspecified, the nonparametric maximum likelihood estimate of $G(x)$ for fixed $\boldsymbol{\phi}$ is discrete with all its mass concentrated on the observed exposure values (Scott and Wild 1997; Gilbert et al. 1999). We, therefore, work with the discrete distribution of X taking values in $\{x_{01}, \ldots, x_{0K}\}$, 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 $j = 0, \ldots, J$, and A_{J+1} as the set of X values in the reference sample. Also, let δ_i denote the probability mass of X at x_{0i}, for $i = 1, \ldots, K$, with $\sum_{i=1}^{K} \delta_i = 1$. Then, the log-likelihood (4) can be written in terms $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)$ as

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

where $\mu_{ji} = \mu_j(x_{0i}, \boldsymbol{\gamma})$, $p_{ji} = p_j(x_{0i}, \boldsymbol{\beta})$ and n_{ji} is the frequency of x_{0i} with $Y = j$, for $j = 0, \ldots, J + 1$, with $Y = J + 1$ denoting the reference sample. The profile log-likelihood $l_p(\boldsymbol{\phi})$ can be obtained from (5) as $l(\boldsymbol{\phi}, \hat{\boldsymbol{\delta}}(\boldsymbol{\phi}))$, where $\hat{\boldsymbol{\delta}}(\boldsymbol{\phi})$ is the maximum likelihood estimate of $\boldsymbol{\delta}$ for given $\boldsymbol{\phi}$. The semiparametric maximum likelihood estimate of $\boldsymbol{\phi}$ that maximizes the log-likelihood $l(\boldsymbol{\phi}, G)$ or $l(\boldsymbol{\phi}, \boldsymbol{\delta})$ can be obtained by maximizing the profile log-likelihood $l_p(\boldsymbol{\phi})$. As noted by Lee and Hirose (2010), this profile log-likelihood depends on a vector of arbitrary parameters $(\rho_1, \ldots, \rho_J)$, where $\log\{\int \mu_j(x, \boldsymbol{\gamma}) p_j(x, \boldsymbol{\beta}) g(x) \mathrm{d}x / \int \mu_0(x, \boldsymbol{\gamma}) p_0(x, \boldsymbol{\beta}) g(x) \mathrm{d}x\} = \rho_j$, for $j = 1, \ldots, J$ (see also Scott and Wild 2001). In this case, estimating all the parameters requires additional information on the population size in each response category. In our context, however, such information is not available. Since $\boldsymbol{\beta}$ is the parameter of interest, we suggest a dimension-reduction approach, where a single offset parameter ρ, independent of j, is required instead of the arbitrary vector $(\rho_1, \ldots, \rho_J)$ (see Appendix A). Note that the argument related to this use of single parameter ρ is asymptotic and justified in view of the underlying prospective mechanism of generating the n_j's, as remarked in the beginning of Sect. 2. The estimator $\hat{\boldsymbol{\phi}}$ can be obtained as the solution of the pseudo-log-likelihood equation $\partial l^*(\boldsymbol{\psi})/\partial \boldsymbol{\psi} = 0$, where $\boldsymbol{\psi} = (\boldsymbol{\phi}, \rho)$ and

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

with $p^*_{ji} = \mu_{ji} p_{ji}$ and ρ being the scalar nuisance parameter. This $l^*(\boldsymbol{\psi})$ is a pseudo-log-likelihood, derived from the log-likelihood (5) using profile log-likelihood approach. It can also be checked that the expression of ρ, given by (12) in appendix, satisfies $\partial l^*(\psi)/\partial \rho = 0$. 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 $\mathrm{e}^{\rho} p^*_{ji}/(1 + \sum_{l=0}^{J} \mathrm{e}^{\rho} p^*_{li})$, $j = 0, \ldots, J$, and $1/(1 + \sum_{l=0}^{J} \mathrm{e}^{\rho} p^*_{li})$, as function of the exposure value x_{0i}. The estimate $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\phi}}, \hat{\rho})$ can be obtained by maximizing (6), which can be carried out using some standard statistical packages. Note that the model parameter $\boldsymbol{\psi}$ is non-identifiable from the likelihood (6), if and only if, for all x, there exist two different values of $\boldsymbol{\psi}$ for which the quantity $\mathrm{e}^{\rho} p^*_{ji}$ evaluated at these two values of $\boldsymbol{\psi}$ are equal, for all $j = 0, 1, \ldots, J$. Therefore, if this equality is violated for at least one j, $\boldsymbol{\psi}$ becomes identifiable from (6).

Note that, from the derivation in Appendix A, the offset parameter ρ 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. For example, information on N can be obtained from a prescription database or some hospital registry. In such case, ρ 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 $\hat{\boldsymbol{\phi}}$, when suitably normalized, follows asymptotically a normal distribution under the standard regularity conditions. The asymptotic variance matrix of $\hat{\boldsymbol{\phi}}$ is estimated by the corresponding partition of the inverse of the observed information matrix $-\partial^2 l^*(\boldsymbol{\phi}, \rho)/\partial\boldsymbol{\phi}\partial\rho$ evaluated at $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}$, $\rho = \hat{\rho}$ (see Appendix B, for a sketch of the proof). The estimate $\hat{\boldsymbol{\phi}}$ is also semiparametrically fully efficient in the sense that the asymptotic variance matrix of $\hat{\boldsymbol{\phi}}$ 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 Sect. 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

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

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

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma)$ and $\boldsymbol{\beta} = (\alpha, \beta)$ are the parameters of interest with $\boldsymbol{\phi} = (\boldsymbol{\gamma}, \boldsymbol{\beta})$. Note that the regression parameter γ in the reporting probability $\mu_j(x, \boldsymbol{\gamma})$'s is the same for $j = 0, 1$, although the intercept parameters γ_0 and γ_1 are different.

As discussed in Sect. 2, the weight functions are given by

$$W_0(x, \boldsymbol{\phi}) = \frac{e^{\gamma_0 + \gamma x}}{1 + e^{\gamma_0 + \gamma x}} \times \frac{1}{1 + e^{\alpha + \beta x}},$$

$$\text{and } W_1(x, \boldsymbol{\phi}) = \frac{e^{\gamma_1 + \gamma x}}{1 + e^{\gamma_1 + \gamma x}} \times \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}},$$

with $W_2(x, \boldsymbol{\phi}) = 1$. 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 Sect. 3, this also ensures identifiability of the model parameters $\boldsymbol{\phi}$ from pseudo-log-likelihood (6).

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

$$l^*(\boldsymbol{\psi}) = \sum_{j=0}^{1} \sum_{i \in A_j} n_{ji}[\rho + \gamma_j + j\alpha + (\gamma + j\beta)x_{0i}]$$

$$
\begin{aligned}
&+\sum_{j=0}^{1}\sum_{i\in A_j} n_{ji}\log[1+\exp(\gamma_{1-j}+\gamma x_{0i})]\\
&+\sum_{i\in A_{J+1}} n_{ji}\log\{[1+\exp(\gamma_0+\gamma x_{0i})][1+\exp(\gamma_1+\gamma x_{0i})]\\
&\times[1+\exp(\alpha+\beta x_{0i})]\}\\
&-\sum_{j=0}^{2}\sum_{i\in A_j} n_{ji}\log[1+\exp(\gamma_1+\gamma x_{0i})+\exp(\alpha+\beta x_{0i})\\
&+\exp(\gamma_0+\alpha+(\gamma+\beta)x_{0i})+(1+\mathrm{e}^{\rho})\{\exp(\gamma_0+\gamma x_{0i})\\
&+\exp(\gamma_0+\gamma_1+2\gamma x_{0i})+\exp(\gamma_1+\alpha+(\gamma+\beta)x_{0i})\\
&+\exp(\gamma_0+\gamma_1+\alpha+(2\gamma+\beta)x_{0i})\}]. \qquad (9)
\end{aligned}
$$

This can be maximized 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 are generated from the model of this section and the pseudo-log-likelihood (9) is maximized.

Note that, while incorporating reporting probabilities in the method of Lee et al. (2006) with case-augmented sample by similar modeling, one will have no observation corresponding to $J = 0$, a biased sample corresponding to $J = 1$ with $W_1(x, \phi)$ as given above and a reference sample corresponding to $J = 2$ with $W_2(x, \phi) = 1$. Then, clearly, the model is not identifiable since the parameter sets $(\gamma_1, \gamma, \alpha, \beta)$ and $(\alpha, \beta, \gamma_1, \gamma)$ give the same pseudo log-likelihood (6).

5 Simulation

For the purpose of the simulation, 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 β, the parameter of interest, based on the combined data using the proposed method of Sect. 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 β, respectively. A reference sample of size $n_{J+1} = n_2$ is drawn from the source population with size N and the corresponding exposure values are recorded. Though for the methodology described in Sect. 3, the source population size is not required, but for the purpose of simulation, we need to specify the value of N. We choose different values of n_2 and N in such a way that the ratio n_2/N remains constant. This, for fixed set of reporting parameters $(\gamma_0, \gamma_1, \gamma)$, ensures that n_j/n tends to the respective constants in probability as $n = n_0 + n_1 + n_2 \to \infty$, for $j = 0, 1, 2$. Here, we consider $n_2 = 200, 400$ and 800 with corresponding N

being 20000, 40000 and 80000. 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 based on 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 β. In Table 1, the average of the estimates of β over the 5000 simulations along with its standard error in parentheses is 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) for the asymptotic 95 % confidence interval of β based on 5000 simulations are also presented.

Depending on the value of $\boldsymbol{\gamma}$ 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 β, respectively, while the method based on combined data seems to produce unbiased estimate in each setting. As expected, the standard error decreases with n_2 and N. The estimated coverage probabilities are also close to 0.95. All this gives 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 ρ can be replaced by $-\log(n_2/N)$ in the pseudo-log-likelihood (9), as remarked in Sect. 3, to reduce the number of parameters to be estimated. For this, we assume $N = 20000, 40000$ and 80000 to be known in respective cases 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 γ in reporting probability (7) is zero making the reporting probability independent of x, as discussed at the end of Sect. 2, identifiability condition is violated. In other words, when data are 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 ρ by $-\log(n_2/N)$ 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 β can be estimated using the standard case–control analysis from only the SR database also. The results based on 5000 simulations are presented in Table 2.

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 the USA. The data contain those reported

Table 1 Simulation results on the estimate of β using SR database alone, SR database with the reference sample and the combined data with known source population size N

$n_{J+1}(=n_2)$	N	$(\gamma_0, \gamma_1, \gamma)$	SR data $\hat{\beta}$ (SSE)	Combined data $\hat{\beta}$ (SSE, ASE, CP)	Combined data with known N $\hat{\beta}$ (SSE, ASE, CP)
200	20000	(−3.5, −4.5, 1.5)	0.633 (0.018)	0.498 (0.037, 0.037, 0.952)	0.500 (0.035, 0.036, 0.954)
400	40000		0.633 (0.013)	0.500 (0.029, 0.026, 0.948)	0.500 (0.025, 0.025, 0.948)
800	80000		0.633 (0.009)	0.500 (0.019, 0.018, 0.944)	0.500 (0.018, 0.018, 0.950)
200	20000	(−3.5, −2.5, 1.0)	0.363 (0.014)	0.512 (0.063, 0.058, 0.944)	0.503 (0.047, 0.047, 0.947)
400	40000		0.363 (0.010)	0.504 (0.038, 0.038, 0.956)	0.503 (0.033, 0.033, 0.953)
800	80000		0.363 (0.007)	0.503 (0.026, 0.027, 0.953)	0.500 (0.023, 0.023, 0.956)

True value of β is 0.5. Corresponding standard errors are in parentheses

Table 2 Simulation results on the estimate of β, when $\gamma = 0$, using SR database alone, SR database combined with the reference sample and known source population size N

$n_{J+1} = n_2$	N	$(\gamma_0, \gamma_1, \gamma)$	SR data alone $\hat{\beta}$ (SSE)	Combined data with known N $\hat{\beta}$ (SSE, ASE, CP)
200	20000		0.500 (0.016)	0.500 (0.016, 0.016, 0.949)
400	40000	(−1, 0.4, 0)	0.500 (0.011)	0.500 (0.011, 0.011, 0.952)
800	80000		0.500 (0.008)	0.500 (0.008, 0.008, 0.949)

True value of β is 0.5. Corresponding standard errors are in parentheses

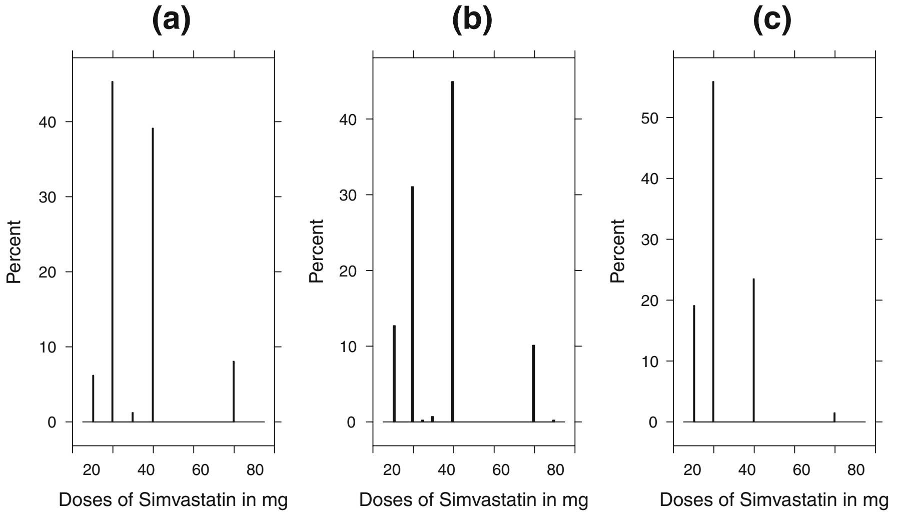

Fig. 1 Dose distribution of the drug simvastatin among **a** 161 cases, **b** 850 controls, **c** reference sample of size 4898

to AERS from all over the world between the time period 2006 to the first quarter of 2011. 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 Sect. 1. Figure 1a, b illustrates 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 1c gives the corresponding dose distribution which seems to be different from those in the case and control samples in Fig. 1a, 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 of the ADR myalgia decreases significantly with higher doses of simvastatin, while ignoring the reporting bias makes this effect insignificant. This analysis indicates the importance of adjusting for reporting bias through a reference sample. The results, however, should be interpreted with some caution keeping in mind that the reference sample is assumed to be a representative of the corresponding source population under study.

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 Sect. 3, the log-likelihood (5) is maximized over $\boldsymbol{\delta}$ for fixed $\boldsymbol{\phi}$. Introducing the Lagrange multiplier λ 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 δ_i to zero, we get

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

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

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

From (11), after setting an offset parameter ρ as

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

we have $\delta_i = (n_{J+1,i} + \sum_{j=0}^{J} n_{ji})/(n_{J+1}(1 + \sum_{j=0}^{J} \mathrm{e}^{\rho} \mu_{ji} p_{ji}))$, which is substituted in (5) to get the pseudo-log-likelihood (6). Note that the ρ in (12) satisfies $\partial l^*(\psi)/\partial \rho = 0$, where $l^*(\psi)$ is given by (6).

To justify this offset parameter ρ 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 and Wild 1997) so that

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

Note that n_j/n tends to ω_j in probability and n_0/n_j tends to ω_0/ω_j in probability as $n \to \infty$ 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

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

the population counterpart of (12). The implicit dependence of ρ on $\boldsymbol{\phi}$, written as $\rho = \rho(\boldsymbol{\phi})$, is clear from the above description.

Appendix B: Asymptotics

The asymptotic properties of the estimator $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\phi}}, \hat{\rho})$ obtained by maximizing the pseudo log-likelihood (6) are established 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 $f_j(x, \boldsymbol{\phi}, g) = \mu_j(x, \boldsymbol{\gamma}) p_j(x, \boldsymbol{\beta}) g(x)/\pi_j$ with $\pi_j = \int \mu_j(x, \boldsymbol{\gamma}) p_j(x, \boldsymbol{\beta}) g(x) \mathrm{d}x$, for $j = 0, \ldots, J$, and E_{J+1} denote the expectation with respect to the unconditional distribution of X having density $f_{J+1}(x, \boldsymbol{\phi}, g) = g(x)$.

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

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

where

$$Z_j(x, \boldsymbol{\psi}) = \left(\frac{\mathrm{e}^{\rho} \mu_j(x, \boldsymbol{\gamma}) p_j(x, \boldsymbol{\beta})}{1 + \sum_{l=0}^{J} \mathrm{e}^{\rho} \mu_l(x, \boldsymbol{\gamma}) p_l(x, \boldsymbol{\beta})} \right), \quad \text{for } j = 0, \ldots, J, \text{ and}$$

$$Z_{J+1}(x, \boldsymbol{\psi}) = \left(\frac{1}{1 + \sum_{l=0}^{J} \mathrm{e}^{\rho} \mu_l(x, \boldsymbol{\gamma}) p_l(x, \boldsymbol{\beta})} \right).$$

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

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

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

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

Hence,

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

Now, we use the results related to estimating function and asymptotic linear estimator (see Hirose 2005, p 72–79). Here, the estimating function is $\partial Z_j(x, \boldsymbol{\psi})/\partial \boldsymbol{\psi}$ with the corresponding asymptotic linear estimator $\hat{\boldsymbol{\psi}}$. Then, the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})$ is multivariate normal (see Hirose 2005, p 67–79) with mean zero and variance–covariance matrix given by

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

where

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

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

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

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

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

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), Σ can be written as

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

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

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

Note that,

$$\begin{aligned}
&E_j \left(\frac{\partial \text{logZ}_j}{\partial \boldsymbol{\psi}} \right) \left(\frac{\partial \text{logZ}_j}{\partial \boldsymbol{\psi}} \right)^T \\
&\quad = \frac{1}{\omega_j} E_X \left[\omega_{J+1} \left(1 + \sum_{l=0}^{J} \mathrm{e}^{\rho} \mu_l(X, \boldsymbol{\gamma}) p_l(X, \boldsymbol{\beta}) \right) \frac{1}{Z_j} \frac{\partial Z_j}{\partial \boldsymbol{\psi}} \frac{\partial Z_j}{\partial \boldsymbol{\psi}^T} \right],
\end{aligned} \tag{25}$$

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

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

Using (25), the information matrix can be written as

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

Now,

$$I_{\boldsymbol{\phi}\boldsymbol{\phi}} = E_X\left[\omega_{J+1}T(X)\sum_{j=0}^{J+1}\frac{1}{Z_j}\frac{\partial Z_j}{\partial\boldsymbol{\phi}}\frac{\partial Z_j}{\partial\boldsymbol{\phi}^T}\right]$$

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

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

using the results,

$$\frac{\partial Z_j}{\partial\rho} = Z_jZ_{J+1}, \quad \text{for } j=0,1,\ldots,J, \quad \text{and} \quad \frac{\partial Z_{J+1}}{\partial\rho} = -Z_{J+1}(1-Z_{J+1}). \tag{28}$$

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

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

Note that, $\sum_{j=0}^{J+1}Z_j = 1$ implies

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

where $T(X) = (1 + \sum_{l=0}^{J} \mathrm{e}^{\rho} \mu_l(X, \boldsymbol{\gamma}) p_l(X, \boldsymbol{\beta}))$. Now, we claim that

$$E_X\left[\omega_{J+1} T(X) \frac{\partial Z_j}{\partial \boldsymbol{\psi}}\right] = -\tau_j(\boldsymbol{\psi}) E_X\left[\omega_{J+1} T(X) \frac{\partial Z_{J+1}}{\partial \boldsymbol{\psi}}\right], \tag{30}$$

where $\sum_{j=0}^{J} \tau_j(\boldsymbol{\psi}) = 1$. Using (29) and (30), we have

$$I(\boldsymbol{\psi})^{-1} E_X\left[\omega_{J+1} T(X) \frac{\partial Z_j}{\partial \boldsymbol{\psi}}\right] = \tau_j(\boldsymbol{\psi}) \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}. \tag{31}$$

From (29) and (31),

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

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 and Hirose (2010) have obtained this bound B for the semiparametric maximum likelihood estimate of parameters in general regression model when data are collected under response-selective sampling scheme. In order to apply their results in our context, let us consider the "population expected likelihood" (see also Newey 1990; Lee et al. 2006) as given by

$$\sum_{j=0}^{J+1} \omega_j E_j[\log \mathrm{f}_j(X, \boldsymbol{\phi}, g)], \tag{32}$$

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 $\pi_j = \int \mu_j(x, \boldsymbol{\gamma}) p_j(x, \boldsymbol{\beta}) g(x) \mathrm{d}x$, for $j = 0, \ldots, J$ 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 exposure distribution $G(x)$ of X. Then, the efficient scores are given by

$$S_j = \left.\frac{\partial \log \mathrm{f}_j(x, \boldsymbol{\phi}, \hat{g}(\boldsymbol{\phi}))}{\partial \boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0}, \quad j = 0, 1, \ldots, J+1, \tag{33}$$

where $\hat{g}(\boldsymbol{\phi})$ is the maximizer of (32), for fixed $\boldsymbol{\phi}$. Then, the corresponding efficiency bound B is given by

$$B^{-1} = \sum_{j=0}^{J+1} \omega_j E_j(S_j S_j^T). \tag{34}$$

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

$$B^{-1} = I_{\boldsymbol{\phi\phi}} - I_{\boldsymbol{\phi}\rho} I_{\rho\rho}^{-1} I_{\rho\boldsymbol{\phi}}. \tag{35}$$

From (32), the expected log-likelihood

$$\sum_{j=0}^{J} \omega_j \int \{\log(\mu_j(x,\boldsymbol{\gamma}) p_j(x,\boldsymbol{\beta})) + \log(g(x))\} \frac{\mu_j(x,\boldsymbol{\gamma_0}) p_j(x,\boldsymbol{\beta_0}) g_0(x)\mathrm{d}x}{\pi_j^0}$$
$$+\omega_{J+1} \int \{\log(g(x))\} g_0(x)\mathrm{d}x - \sum_{j=0}^{J} \omega_j \log \pi_j, \tag{36}$$

where $\pi_j^0 = \int \mu_j(x,\boldsymbol{\gamma_0}) p_j(x,\boldsymbol{\beta_0}) g_0(x)\mathrm{d}x$. Considering the terms which involve $g(x)$, (36) can be written as

$$\omega_{J+1} \int \log(g(x)) \tilde{p}(x) g_0(x)\mathrm{d}x - \sum_{j=0}^{J} \omega_j \log \pi_j,$$
$$\text{where } \tilde{p}(x) = 1 + \sum_{j=0}^{J} \frac{\omega_j}{\omega_{J+1}} \frac{\mu_j(x,\boldsymbol{\gamma_0}) p_j(x,\boldsymbol{\beta_0})}{\pi_j^0}. \tag{37}$$

Now, we need to find $\hat{g}$ which maximizes (37). Consider the class of distribution of X to be discrete with finite support $\{x_1,\ldots,x_M\}$. 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, (37) can be written along with Lagrange multiplier λ to take care of the constraint $\sum_{i=1}^{M} g_i = 1$, as

$$\omega_{J+1} \sum_{i=1}^{M} \log(g_i) \tilde{p}(x_i) g_0(x_i) - \sum_{j=0}^{J} \omega_j \log \pi_j(\mathbf{g}) + \lambda \left(\sum_{i=1}^{M} g_i - 1 \right), \tag{38}$$

where $\pi_j(\mathbf{g}) = \sum_{i=1}^{M} \mu_{ji} p_{ji} g_i$. Differentiating (38) with respect to g_i, we have

$$\omega_{J+1} \frac{\tilde{p}(x_i) g_0(x_i)}{g_i} - \sum_{j=0}^{J} \omega_j \frac{\mu_j(x_i,\boldsymbol{\gamma}) p_j(x_i,\boldsymbol{\beta}))}{\pi_j(\mathbf{g})} + \lambda = 0. \tag{39}$$

Multiplying (39) by g_i and summing over i give $\lambda = -\omega_{J+1}$. Putting the value of λ in (39), we get the estimate of g_i as

$$\hat{g}_i = \frac{\tilde{p}(x_i)g_0(x_i)}{1 + \sum_{j=0}^{J} \frac{\omega_j}{\pi_j(\boldsymbol{g})\omega_{J+1}} \mu_j(x_i, \boldsymbol{\gamma})p_j(x_i, \boldsymbol{\beta}))}. \tag{40}$$

In case of general g, not having finite support, the maximizer of (37) is of the form

$$\hat{g}(x; \boldsymbol{\phi}, \rho) = \frac{\tilde{p}(x)g_0(x)}{1 + \sum_{j=0}^{J} \frac{\omega_j}{\pi_j(\rho)\omega_{J+1}} \mu_j(x, \boldsymbol{\gamma})p_j(x, \boldsymbol{\beta})} \tag{41}$$

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

$$\int \mu_j(x, \boldsymbol{\gamma})p_j(x, \boldsymbol{\beta})\hat{g}(x; \boldsymbol{\phi}, \rho)\mathrm{d}x = \pi_j(\rho), \text{ or, equivalently,}$$
$$\int \left\{ \frac{\mathrm{e}^{\rho}\mu_j(x, \boldsymbol{\gamma})p_j(x, \boldsymbol{\beta})}{1 + \sum_{j=0}^{J} \mathrm{e}^{\rho}\mu_j(x, \boldsymbol{\gamma})p_j(x, \boldsymbol{\beta})} \right\} \tilde{p}(x)g_0(x)\mathrm{d}x = \frac{\omega_j}{\omega_{J+1}}, \tag{42}$$

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

$$\begin{aligned} \log \mathrm{f}_j(x, \boldsymbol{\phi}, \hat{g}) &= \log \left\{ \frac{\mathrm{e}^{\rho(\boldsymbol{\phi})}\mu_j(x, \boldsymbol{\gamma})p_j(x, \boldsymbol{\beta})}{1 + \sum_{j=0}^{J} \mathrm{e}^{\rho(\boldsymbol{\phi})}\mu_j(x, \boldsymbol{\gamma})p_j(x, \boldsymbol{\beta})} \right\} + c_j \\ &= \log\{q_j(x, \boldsymbol{\phi}, \rho(\boldsymbol{\phi}))\} + c_j \end{aligned} \tag{43}$$

and

$$\begin{aligned} \log \mathrm{f}_{J+1}(x, \boldsymbol{\phi}, \hat{g}) &= \log \left\{ \frac{1}{1 + \sum_{j=0}^{J} \mathrm{e}^{\rho(\boldsymbol{\phi})}\mu_j(x, \boldsymbol{\gamma})p_j(x, \boldsymbol{\beta})} \right\} + c_{J+1} \\ &= \log \left\{ 1 - \sum_{j=0}^{J} q_j(x, \boldsymbol{\phi}, \rho(\boldsymbol{\phi})) \right\} + c_{J+1}, \end{aligned} \tag{44}$$

where c_j's are constants with respect to $\boldsymbol{\psi} = (\rho, \boldsymbol{\phi})$. Using (33), the efficient scores are given as

$$S_j = \frac{\partial}{\partial \boldsymbol{\phi}} \log\{q_j(x, \boldsymbol{\phi}, \rho(\boldsymbol{\phi}))\}, \quad \text{for } j = 0, 1, \ldots, J+1, \tag{45}$$

where $q_{J+1}(x, \boldsymbol{\phi}, \rho(\boldsymbol{\phi})) = 1 - \sum_{j=0}^{J} q_j(x, \boldsymbol{\phi}, \rho(\boldsymbol{\phi}))$ and all the derivatives are evaluated at $\boldsymbol{\phi} = \boldsymbol{\phi}_0$. Applying chain rule,

$$S_j = \frac{\partial}{\partial \boldsymbol{\phi}} \log\{q_j(x, \boldsymbol{\phi}, \rho(\boldsymbol{\phi}))\} + \left\{\frac{\partial \rho(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right\}^T \frac{\partial}{\partial \rho} \log\{q_j(x, \boldsymbol{\phi}, \rho(\boldsymbol{\phi}))\}. \tag{46}$$

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

$$I(\boldsymbol{\psi}) = I(\rho, \boldsymbol{\phi}) = \sum_{j=0}^{J+1} \omega_j E_j \left\{\left(\frac{\partial}{\partial \boldsymbol{\psi}} \text{logq}_j\right)\left(\frac{\partial}{\partial \boldsymbol{\psi}} \text{logq}_j\right)^T\right\}. \tag{47}$$

From (34) and (46), we get

$$B^{-1} = I_{\boldsymbol{\phi}\boldsymbol{\phi}} + \left\{\frac{\partial \rho(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right\}^T I_{\rho\boldsymbol{\phi}} + I_{\boldsymbol{\phi}\rho}\left\{\frac{\partial \rho(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right\} + \left\{\frac{\partial \rho(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right\}^T I_{\rho\rho}\left\{\frac{\partial \rho(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right\}. \tag{48}$$

Differentiating (42) under the integral sign

$$\int \frac{\partial q_j}{\partial \boldsymbol{\phi}} \tilde{p} g \, \mathrm{d}x + \left\{\frac{\partial \rho(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right\}^T \int \frac{\partial q_j}{\partial \rho} \tilde{p} g \, \mathrm{d}x = 0, \quad j = 0, 1, \ldots, J. \tag{49}$$

It can be easily checked that

$$\frac{\partial q_j}{\partial \rho} = q_j\left(1 - \sum_{j=0}^{J} q_j\right); -\frac{\partial^2 \log q_j}{\partial \rho^2} = \sum_{j=0}^{J} \frac{\partial q_j}{\partial \rho}$$
$$\text{and} \; -\frac{\partial}{\partial \boldsymbol{\phi}}\left\{\frac{\partial \log q_j}{\partial \rho}\right\} = \sum_{j=0}^{J} \frac{\partial q_j}{\partial \boldsymbol{\phi}}. \tag{50}$$

Now, using (47) and (50),

$$I_{\rho\rho} = \omega_{J+1} \int \sum_{j=0}^{J} \frac{\partial q_j}{\partial \rho} \tilde{p} g \, \mathrm{d}x \text{ and } I_{\rho\boldsymbol{\phi}} = \omega_{J+1} \int \sum_{j=0}^{J} \frac{\partial q_j}{\partial \boldsymbol{\phi}} \tilde{p} g \, \mathrm{d}x. \tag{51}$$

Summing over $j = 0, 1, \ldots, J$ in (49) and using (50)

$$\frac{\partial \rho(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = -I_{\rho\rho}^{-1} I_{\rho\boldsymbol{\phi}}. \tag{52}$$

From (34), (48) and (52)

$$B^{-1} = I_{\boldsymbol{\phi}\boldsymbol{\phi}} - I_{\boldsymbol{\phi}\rho} I_{\rho\rho}^{-1} I_{\rho\boldsymbol{\phi}}. \tag{53}$$

This establish the efficiency bound of the semiparametric procedure.

References

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., et al. (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 semiparametric models*. Baltimore: John Hopkins University Press.

Breslow, N. E. (1996). Statistics in epidemiology: The case–control study. *Journal of American Statistical Association*, *91*(433), 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*(16), 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*(1), 27–43.

Hartz, I., Sakshaug, S., Furu, K., Engeland, A., Eggen, A. E., Njolstad, I., et al. (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*(14), 1–6.

Hirose, Y. (2005). Efficiency of the semi-parametric maximum likelihood estimator in generalized case–control studies. PhD thesis, University of Auckland.

Hsieh, D. A., Manski, C. F., McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations. *Journal of American Statistical Association*, *80*(391), 651–662.

Lee, A., Hirose, Y. (2010). Semi-parametric efficiency bounds for regression models under response-selective sampling: the profile likelihood approach. *Annals of the Institute of Statistical Mathematics*, *62*, 1023–1052.

Lee, A. J., Scott, A. J., Wild, C. J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika*, *93*(2), 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.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, *5*(2), 99–135.

Prentice, R. L., Breslow, N. E. (1978). Retrospective studies and failure time models. *Biometrika*, *65*(1), 153–158.

Prentice, R. L., Pyke, R. (1979). Logistic disease incidence models and case–control studies. *Biometrika*, *66*(3), 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.

Wild, C. J. (1991). Fitting prospective regression models to case–control data. *Biometrika*, *78*(4), 705–717.