Matched Case-Control Study with Reporting Bias: An Analysis of Spontaneous Adverse Drug Reaction (ADR) Reports

Technical Report No. ASD/2012/4
Dated: 6 March 2012

Palash Ghosh

And

Anup Dewanji

Applied Statistics Unit
Indian Statistical Institute
Kolkata 700108

palash_r@isical.ac.in, dewanjia@isical.ac.in

Matched Case-Control Study with Reporting Bias: An Analysis of Spontaneous Adverse Drug Reaction (ADR) 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

Matching is widely used to improve the efficiency in case-control studies. To achieve this, it is required that both case and control samples are representative of respective part in the source population. Often this characteristics is violated and we fail to draw valid inference about the parameter of interest. In this work, we will discuss the matched case-control study to analyse spontaneous adverse drug reaction (ADR) reports in the presence of reporting bias. To estimate the parameter of interest unbiasedly we consider a reference sample from the source population. We also discuss the asymptotic properties of the estimates. A real example is considered for the purpose of illustration.

Key words: Adverse drug reaction, Matched Case-control data, Multi-sample model, Reference sample, Source population, Spontaneous Reporting database, Under-reporting.

1 Introduction

Spontaneous reporting (SR) system consists of reports of suspected side effects or adverse drug reactions (ADR). These reports come from clinicians and/or health professionals who are responsible for recognizing and reporting it after the drugs are in the market. Many international and national agencies, such as World Health Organization (WHO), Food and Drug Administration (FDA) and others, maintain such SR database. The SR database is used to monitor the safety of drugs after they are approved by the concerned drug authority and marketed. The primary purpose of maintaining such SR database is to provide early warnings, or suspicions, or signals, of some hazards which have not been recognized prior to marketing because of various limitations of a clinical trial. The findings from a SR database, if positive, are to be followed up with more elaborate and well-designed experiments or epidemiological studies to estimate the related incidence or risks (See Anello and O'Neill, 1998; Bate et al., 1998; and Evans et al., 2001) and establish a potential safety concern identified in SR database. For this purpose, Bate et al. (1998) have proposed a Bayesian approach called Bayesian confidence propagation neural network (BCPNN) and DuMouchel (1999) has also discussed a Bayesian methodology with application to FDA spontaneous reporting system. These are essentially data mining methods, but there is no consensus regarding their applicability and efficiency (Roux et al., 2005). Evans et al. (2004) have introduced a terminology called proportional reporting ratio (PRR) for signal generation from SR database. Often, a particular drug is compared with some reference drugs (or, other drugs) with respect to occurrence of a particular ADR against occurrence of other ADRs.

SR database suffers from severe under-reporting problem (van der Heijden PGM et al., 2002), which makes the estimates from SR database biased for the corresponding quantities in the source population. Existing methods of analyzing a SR database also suffer from high false positive rate largely due to confounding with disease condition. Ghosh and Dewanji (2011) address this problem of confounding by characterizing the source population with having the particular disease for which the concerned drugs are being used; the problem of under-reporting is addressed with the help of supplementary data. In this work, we will extend the idea of Ghosh and Dewanji (2011) to matched case-control study, in which the cases correspond to those experiencing the concerned ADR and controls are those who have not experienced the concerned ADR in the source population. However, the sampling of cases and controls takes place through the SR database which is subject to under-reporting resulting in some selection bias. This selection bias is adjusted by using supplementary data. Besides the particular disease, there may still be some other confounders such as age, gender, or other disease conditions, which are adjusted by means of matching the controls with the cases.

Matching is widely used in epidemiological studies, specially in case-control studies, to avoid biased comparison between cases and controls due to confounding. In a SR database, a particular ADR report has information on, in addition to the status of ADR, presence of different diseases, average daily doses of the drug, whether the report is of first time or a follow-up, start date and end date of medication (duration of drug used), patient's age, gender, weight and residing country, etc., some of which may be confounders. Matching and stratification are the two common ways of adjusting for confounders. The above definition of the source population adjusts for the effect of disease to develop the ADR by means of stratification. Other confounders may need to be adjusted through matching. For example, in the source population of patients suffering from cardiovascular disease and stroke (Ghosh and Dewanji, 2011), one may consider matching with age and obesity to asses the effect of the drug 'diuretic' on the ADR 'congestive heart failure'.

There are many ways to match cases with controls with respect to observed confounders (See Rosenbaum, 2002 p295-331). Depending on the situation, one can use pair matching, or one case with fixed number of controls, one case with variable number of controls, among many others. In this paper, for simplicity, we will consider one case with fixed number of controls; variable number of controls per case can be analyzed in similar fashion. In Section 2, we will discuss how this matching is incorporated into the analysis of SR database. As mentioned before, while analyzing SR database, one has to deal with selection bias due to under-reporting. Lin and Paik (2001) deal with this bias issue by assuming that the cases are free from this bias and selecting controls in two stages in Northern Manhattan Stroke Study (NOMASS), where the selection in the first stage is assumed free of bias. There are, however, real life examples where both case and control samples suffer from reporting (selection) bias. In this work, we propose to deal with this problem using supplementary data from a reference sample, chosen randomly from the source population, in the spirit of Ghosh and Dewanji (2011), as described in Sections 3 and 4. In Section 3, for simplicity, the exposure (drug) is considered binary and there is no other covariate. This is applicable when comparison is made between a specific drug and all other drugs used for the disease in the source population. In Section 4, we consider the exposure to be continuous and there may be other covariates. This model may be applicable when comparison is made between different doses of a particular drug. Nevertheless, this may be considered as a generalization of the modeling of Section 3. Some asymptotic results are derived in Section 5, while the finite sample properties are studied through simulation in Section 6 along with an illustration of a real example. Section 8 ends with some concluding remarks.

2 Preliminaries

A report comes to SR database when an individual in the source population experiences one or more ADRs and are detected by clinicians or health professionals and they report it. Let us define two binary random variables A and R as the ADR status and the reporting status, respectively, where $A = 1$ denotes a case and $A = 0$ denotes a control in the source population and $R = 1$ means a report in SR database and $R = 0$ implies a patient in the source population, not reported in SR database. In the SR database, a case is denoted by $\{A = 1, R = 1\}$ and a control is denoted by $\{A = 0, R = 1\}$. In matched case-control study, information on exposure E and matching variable(s) M are obtained for a case in SR database. Then, a set of c matched controls with $\{A = 0, R = 1\}$ is selected randomly from the SR database. For the i^{th} case and corresponding c matched controls, the likelihood is given by,

$$P(E = E_{i0}, M = M_{i0}|A = 1, R = 1) \times \prod_{h=1}^{c} P(E = E_{ih}|A = 0, R = 1, M = M_{i0}), \tag{1}$$

where E_{i0} and M_{i0} are the observed values of E and M for the case and E_{ih}, $h = 1, \cdots, c$, are the observed values of E for the c matched controls. The i^{th} case along with the corresponding c matched controls is treated as the i^{th} stratum so that n strata corresponding to n cases from the SR database is considered for analysis. Following the approach of Breslow (1996), which considers the conditional probability that exposure E_{i0} is that of the case and $(E_{i1}, \ldots, E_{ic})$ are those of the c controls, as observed, given the unordered set $E^{(i)} = \{E_{i0}, E_{i1}, \ldots, E_{ic}\}$ of $c + 1$ exposures, the conditional likelihood from the i^{th} stratum is

$$L_i = \frac{P(E = E_{i0}, M = M_{i0}|A = 1, R = 1) \times \prod_{h=1}^{c} P(E = E_{ih}|A = 0, R = 1, M = M_{i0})}{\sum_{h=0}^{c} P(E = E_{ih}, M = M_{i0}|A = 1, R = 1) \times \prod_{l \neq h} P(E = E_{il}|A = 0, R = 1, M = M_{i0})}. \tag{2}$$

Assuming that the reporting probabilities do not depend on the matching variables and after some simple probability calculation, the above likelihood L_i becomes

$$\frac{P(R = 1|A = 1, E = E_{i0})P(A = 1|E = E_{i0}, M = M_{i0}) \times \prod_{h=1}^{c} P(R = 1|A = 0, E = E_{ih})P(A = 0|E = E_{ih}, M = M_{i0})}{\sum_{h=0}^{c} P(R = 1|A = 1, E = E_{ih})P(A = 1|E = E_{ih}, M = M_{i0}) \times \prod_{l \neq h} P(R = 1|A = 0, E = E_{il})P(A = 0|E = E_{il}, M = M_{i0})}. \tag{3}$$

When the reporting probability $P(R = 1|A, E)$ depends either on E, or on A, but not on both, this likelihood simplifies further to

$$\frac{P(A = 1|E = E_{i0}, M = M_{i0}) \times \prod_{h=1}^{c} P(A = 0|E = E_{ih}, M = M_{i0})}{\sum_{h=0}^{c} P(A = 1|E = E_{ih}, M = M_{i0}) \times \prod_{l \neq h} P(A = 0|E = E_{il}, M = M_{i0})},$$

which is the conditional likelihood for matched case-control study when both the cases and the controls are random samples from the corresponding source population (Breslow, 1996). As argued in Ghosh and Dewanji (2011), this assumption on the reporting probabilities allows estimation of odds-ratio from the SR database alone, since the 'reporting odds-ratio' becomes equal to the source population odds-ratio.

The SR database carries information only on those reporting ($R = 1$), which does not permit estimation of the reporting probabilities. On the other hand, estimation of model parameters from the likelihood (3) requires information on the reporting probabilities. In order to serve this purpose, as mentioned before, we consider a reference sample randomly chosen from the source population. This combined data of the SR database and the reference sample can be viewed as an outcome-dependent enriched (ODE) sample of Kang (2010). As in Ghosh and Dewanji (2011), we consider varying degree of information from the reference sample supplemented by required assumption, as described in the following two sections.

3 Binary Exposure

In order to analyse spontaneous adverse drug reaction (ADR) reports in the presence of reporting bias, we consider the following reporting probabilities (assumed to be independent of the matching variables):

$$\begin{aligned}
\theta_{11} &= P(R=1|E=1,A=1)\\
\theta_{10} &= P(R=1|E=1,A=0)\\
\theta_{01} &= P(R=1|E=0,A=1)\\
\theta_{00} &= P(R=1|E=0,A=0),
\end{aligned}$$

where the exposure (drug) E is a binary variable with $E=1$ denoting use of the specific drug and $E=0$ denoting use of other drugs. Note that an individual with $A=0$ in the source population does not develop the concerned ADR, but may develop other ADRs and then report the same to the SR database with probability θ_{10} or θ_{00}, depending on the exposure status of the individual. The relationship between the ADR and the exposure along with the matching variable(s) in the i^{th} stratum is modeled as

$$P(A=1|E,M_i) = \frac{exp(\gamma_{M_i}+\beta E)}{1+exp(\gamma_{M_i}+\beta E)}, \tag{4}$$

with M_i being the value of M in the i^{th} stratum, for $i=1,\cdots,n$, where γ_M denotes the effect of the matching variable M which is assumed to be arbitrary depending on the value of M. As a special case, for example, one may consider $\gamma_M=\gamma M$, a linear effect, but the subsequent analysis does not depend on any such modeling assumption. Note that the odds-ratio parameter β is assumed constant over the different strata. Also, simple probability calculation shows, as in Ghosh and Dewanji (2011), that the reporting odds-ratio (ROR), for a given value of the matching variable, in the SR database is given by $ROR=(\theta_{11}\theta_{00}/\theta_{10}\theta_{01})\times OR$, where $OR=exp(\beta)$ is the corresponding population odds-ratio. However, the likelihood L_i from (3), under model (4), becomes

$$\begin{aligned}
&\frac{\theta_{E_{i0}1}\cdot\dfrac{exp(\gamma_{M_{i0}}+\beta E_{i0})}{1+exp(\gamma_{M_{i0}}+\beta E_{i0})}\times\displaystyle\prod_{h=1}^{c}\theta_{E_{ih}0}\cdot\dfrac{1}{1+exp(\gamma_{M_{i0}}+\beta E_{ih})}}{\displaystyle\sum_{h=0}^{c}\theta_{E_{ih}1}\cdot\dfrac{exp(\gamma_{M_{i0}}+\beta E_{ih})}{1+exp(\gamma_{M_{i0}}+\beta E_{ih})}\times\prod_{l\neq h}\theta_{E_{il}0}\cdot\dfrac{1}{1+exp(\gamma_{M_{i0}}+\beta E_{il})}}\\
&=\frac{\theta_{E_{i0}1}\cdot exp(\beta E_{i0})\times\displaystyle\prod_{h=1}^{c}\theta_{E_{ih}0}}{\displaystyle\sum_{h=0}^{c}\theta_{E_{ih}1}\cdot exp(\beta E_{ih})\times\prod_{l\neq h}\theta_{E_{il}0}},
\end{aligned}$$

which is a function of β and also the reporting probabilities θ's. Matched case-control conditional likelihood for n strata from the SR database is, therefore,

$$L_{SR}=\prod_{i=1}^{n}L_i=\prod_{i=1}^{n}\frac{\theta_{11}^{E_{i0}}\theta_{01}^{1-E_{i0}}exp(\beta E_{i0})\prod_{h=1}^{c}\theta_{10}^{E_{ih}}\theta_{00}^{1-E_{ih}}}{\sum_{h=0}^{c}\theta_{11}^{E_{ih}}\theta_{01}^{1-E_{ih}}exp(\beta E_{ih})\prod_{l\neq h}\theta_{10}^{E_{il}}\theta_{00}^{1-E_{il}}}. \tag{5}$$

Note that the SR database does not provide information on θ's, the reporting parameters. Without any assumption or information on the θ's, it is not possible to estimate β, the parameter of interest, from the likelihood (5). Some supplementary information is required from outside (e.g., the reference sample). In practice, getting information from outside SR database is difficult and costly. There is a trade-off between the assumptions we impose on the SR database and the information that is required from the reference sample. Following two sub-sections discuss two scenarios in detail.

3.1 SR database and reference sample with exposure information

Here, we assume that all the cases are reported to the SR database (i.e., $\theta_{11} = \theta_{01} = 1$). As mentioned in Ghosh and Dewanji (2011), this assumption is realistic when the ADR under study is serious and the doctors/health professionals are well-informed about it. We also assume that the source population size N is known. Later, we will show that an approximate value of N will serve the purpose, since the concerned estimate is quite robust against the misspecification of the value of N. With the assumption $\theta_{11} = \theta_{01}$, the likelihood (5) can be written as

$$L_1^* = \prod_{i=1}^{n} \frac{exp(\beta E_{i0})\big(\frac{\theta_{10}}{\theta_{00}}\big)^{\sum_{h=1}^{c} E_{ih}}}{\sum_{h=0}^{c} exp(\beta E_{ih})\big(\frac{\theta_{10}}{\theta_{00}}\big)^{\sum_{l\neq h} E_{il}}}. \tag{6}$$

Note that the two parameters β and $\theta = \theta_{10}/\theta_{00}$ are not estimable by maximization of the likelihood (6) since the two explanatory variables associated with them, E_{i0} and $\sum_{h=1}^{c} E_{ih}$, given $E^{(i)}$, are perfectly correlated. The corresponding two estimating equations turn out to be dependent. To overcome this difficulty, we consider a consistent estimate of θ, and then plug-in that estimate into the likelihood (6), which is then maximized with respect to β to get an estimate of β. The corresponding likelihood equation is given by

$$\sum_{i=1}^{n} \frac{\sum_{h=0}^{c}(E_{i0} - E_{ih})exp(\beta E_{ih})\ \widehat{\theta}^{\sum_{l\neq h} E_{il}}}{\sum_{h=0}^{c} exp(\beta E_{ih})\ \widehat{\theta}^{\sum_{l\neq h} E_{il}}} = 0, \tag{7}$$

where $\widehat{\theta}$ is a consistent estimate of $\theta = \theta_{10}/\theta_{00}$. Note that, from (7), if the exposure status of one case and c controls in a stratum are all same, then this stratum does not contribute anything to the estimation procedure, as expected.

Using the definition of reporting probabilities, the ratio of the reporting probabilities in control sample can be written as

$$\theta = \frac{\theta_{10}}{\theta_{00}} = \frac{P(E=1|A=0,R=1)}{P(E=0|A=0,R=1)} \times \frac{P(E=0) - P(E=0|A=1)P(A=1)}{P(E=1) - P(E=1|A=1)P(A=1)}. \tag{8}$$

All the probabilities, except $P(E)$, are estimable from the SR database with known N and the assumption that $\theta_{11} = \theta_{01} = 1$. Let us consider the SR database as the 2×2 contingency table with the column variable being ADR ($A = 0, 1$) and the row variable being exposure ($E = 0, 1$). Suppose n_{uv}'s are the corresponding cell frequencies, for $u, v = 0, 1$. The two probabilities in the first term of (8) are directly estimable from the controls in the SR database with the ratio estimated by n_{10}/n_{00}. Also, $P(E = 1|A = 1)$ and $P(E = 0|A = 1)$ are estimable from the case population, because of 100% reporting of cases from source population, as $n_{11}/(n_{11} + n_{01})$ and $n_{01}/(n_{11} + n_{01})$, respectively, while $P(A = 1)$ is estimated by $(n_{11} + n_{01})/N$. In order to estimate $P(E = 1)$, we consider a reference sample of size m drawn from the source population and observe the exposure status. Let $m_{1\cdot}$ denote the number of observations in the reference sample with exposure status $E = 1$. Then, we have

$$\widehat{\theta} = \frac{n_{10}}{n_{00}} \times \frac{\frac{m-m_{1\cdot}}{m} - \frac{n_{01}}{n_{11}+n_{01}} \cdot \frac{n_{11}+n_{01}}{N}}{\frac{m_{1\cdot}}{m} - \frac{n_{11}}{n_{11}+n_{01}} \cdot \frac{n_{11}+n_{01}}{N}}. \tag{9}$$

In the expression (9) of $\widehat{\theta}$, individual probabilities are estimated as the binomial proportions in the respective samples. Using Slutsky's theorem, $\widehat{\theta}$, clearly, is a consistent estimate of θ. It is also seen that the estimate is quite robust against misspecification of N. A careful look at (8) and (9) also indicates the robustness of the estimator against the departure from the assumption $\theta_{11} = \theta_{01} = 1$ as long as we have $\theta_{11} = \theta_{01}$ (Ghosh and Dewanji, 2011), since then $P(E = 0|A = 1)/P(E = 1|A = 1) = P(E = 0|A = 1, R = 1)/P(E = 1|A = 1, R = 1)$. In other words, when the case sample is a random sample from the corresponding part of the source population and the control sample may be subject to reporting bias, this methodology still gives satisfactory estimate of the model parameter β. When $c = 2$, it is easy to check that the estimate of β can be obtained as a solution of the quadratic equation

$$2(k_3 + k_4)x^2 + \hat{\theta}(4k_4 + k_3 - 4k_2 - k_1)x - 2\hat{\theta}^2(k_1 + k_2) = 0, \tag{10}$$

where $x = exp(\beta)$, k_1 is the number of strata with case exposure status being $E = 1$ and exactly one control exposure status being $E = 1$, k_2 is the number of strata with case exposure status being $E = 1$ and both the controls having exposure status $E = 0$, k_3 is the number of strata with case exposure status being $E = 0$ and exactly one control status being $E = 1$, and k_4 is the number of strata with case exposure status being $E = 0$ and both the controls having exposure status $E = 1$. The estimate of β is obtained from the positive root of the quadratic equation (10) as

$$\hat{\beta} = log\left(\frac{-\hat{\theta}(4k_4 + k_3 - 4k_2 - k_1) + \sqrt{\hat{\theta}^2(4k_4 + k_3 - 4k_2 - k_1)^2 + 16\hat{\theta}^2(k_1 + k_2)(k_3 + k_4)}}{4(k_3 + k_4)}\right). \tag{11}$$

Writing the i^{th} term of $\partial log L_1^*/\partial\beta$ (See likelihood (6) and equation (7)) as $u_i(\beta, \theta)$, let us assume that the first derivative of u_i with respect to the parameter θ exists and its absolute value in a neighbourhood of the true value of θ is bounded by an integrable function with finite mean. Then the asymptotic distribution of $\hat{\beta}$, obtained from likelihood (6) with the parameter θ replaced by a consistent estimate given by (9), can be shown to be same as that of the estimate of parameter β obtained by using the true value of $\theta = \theta_{10}/\theta_{00}$ in likelihood (6). Under the standard regularity condition, this can be shown to be normal with mean β and a variance that can be estimated from the observed information based on likelihood (6) and using $\hat{\beta}$ and $\widehat{\theta}$.

Note that the estimation procedure using (7) treats $\hat{\theta}$ to be a constant. The corresponding variance estimate using the observed information based on (6) does not account for the variability in $\hat{\theta}$ and, therefore, gives an under-estimate. This has also been found in our simulation study (not reported here). For large m, n and N, when $\hat{\theta}$ is close to the true value of θ with small variation, this under-estimation may be ignored; however, this may require some adjustment for small sample sizes. We propose a sandwich estimate for the variance of $\hat{\beta}$ as given by

$$\left[\left(-\frac{\partial u(\beta, \hat{\theta})}{\partial\beta}\right)^{-1}\right]^T \hat{V}(u(\beta, \hat{\theta})) \left[\left(-\frac{\partial u(\beta, \hat{\theta})}{\partial\beta}\right)^{-1}\right], \tag{12}$$

evaluated at $\beta = \hat{\beta}$, where $u(\beta, \hat{\theta}) = \sum_{i=1}^{n} u_i(\beta, \hat{\theta})$ and $\hat{V}(u(\beta, \hat{\theta}))$ is an estimate of $V(u(\beta, \hat{\theta}))$, as derived in Appendix A, along with the derivation of (12). This seems to perform better than the estimate based on observed information for small sample sizes, as indicated by the simulation study in Section 6.2 (not reported). Alternatively, one can estimate $V(u(\beta, \hat{\theta}))$ by $\frac{1}{n}\sum_{i=1}^{n} u_i(\hat{\beta}, \hat{\theta})u_i^T(\hat{\beta}, \hat{\theta})$ and then use (12) for estimating variance of $\hat{\beta}$. The results turn out to be similar.

3.2 SR database and reference sample with reporting information

In this subsection, we make no assumption on the reporting parameters θ_{uv}'s, but require more information from the reference sample. Unlike the previous subsection, we do not need N to be known. However, along with the information on both ADR and exposure status from the reference sample, information on reporting status is also needed. Suppose there are m_{uv} number of observations in the reference sample with $\{E = u, A = v\}$, for $u, v = 0, 1$, and z_{uv} of them report to the SR database. The SR database conditional likelihood is then given by (5). The reference sample likelihood with information on case-control status along with reporting status can be written as

$$L_R = \prod_{u,v=0}^{1} (\theta_{uv}\pi_{uv})^{z_{uv}}((1-\theta_{uv})\pi_{uv})^{m_{uv}-z_{uv}} = \left\{\prod_{u,v=0}^{1} \theta_{uv}^{z_{uv}}(1-\theta_{uv})^{m_{uv}-z_{uv}}\right\} \prod_{u,v=0}^{1} \pi_{uv}^{m_{uv}}, \qquad (13)$$

where $\pi_{uv} = P(E = u, A = v)$ such that $\theta_{uv}\pi_{uv} = P(R = 1, E = u, A = v)$, for $u, v = 0, 1$. Note that the probability terms π_{uv}'s can be written down, using model (4) and integrating over the distributions of E and M, as functions of β and γ_M's. However, information on β in these π_{uv}'s cannot be extracted in the absence of any knowledge on the γ_M's, the nuisance parameters. We, therefore, consider only the first part of the likelihood (13) involving only the θ_{uv}'s, which can be viewed as a conditional likelihood of observation on reporting status, given the ADR and exposure status (Andersen, 1970; Kalbfleisch and Sprott, 1970, 1973). Combining this with the likelihood (5), we have the total likelihood as

$$L_2^* = \left\{\prod_{i=1}^{n} \frac{\theta_{11}^{E_{i0}}\theta_{01}^{1-E_{i0}} exp(\beta E_{i0}) \prod_{h=1}^{c} \theta_{10}^{E_{ih}}\theta_{00}^{1-E_{ih}}}{\sum_{h=0}^{c} \theta_{11}^{E_{ih}}\theta_{01}^{1-E_{ih}} exp(\beta E_{ih}) \prod_{l\neq h} \theta_{10}^{E_{il}}\theta_{00}^{1-E_{il}}}\right\} \times \left\{\prod_{u,v=0}^{1} \theta_{uv}^{z_{uv}}(1-\theta_{uv})^{m_{uv}-z_{uv}}\right\}. \qquad (14)$$

Numerical maximization can be used to obtain the estimates of the parameters from likelihood (14). The initial values of the reporting parameters θ_{uv}'s, may be based on the reference sample only (i.e., z_{uv}/m_{uv}). The corresponding asymptotic results are discussed in Section 5 and the variance-covariance matrix of the estimated parameters is obtained from the observed information matrix based on likelihood (14).

4 Continuous Exposure

In practice, for a particular drug, there may be higher possibility of developing the concerned ADR at larger doses. The occurrence of this ADR, therefore, can be modeled as a standard dose response relationship. Unlike in the previous sections, instead of comparing occurrence of the concerned ADR in relation with other ADRs for using a particular drug in relation to other drugs, the objective here is to study the dose response relationship for the occurrence of the concerned ADR at different doses of the particular drug. In that case, the exposure can be viewed as a continues variable. Note that, the methodologies developed in the previous section are not able to deal with continuous exposure variable. The dose response model for the presence of ADR is assumed to be the same as that in (4), except that the drug exposure (E) is now considered a continuous variable representing the dose level of the drug. However, the reporting probability, as a function of A and E, is now modeled as

$$P(R=1|A,E) \quad = \quad \frac{exp(\delta_0+\delta_1 A+\delta_2 E)}{1+exp(\delta_0+\delta_1 A+\delta_2 E)}. \tag{15}$$

The conditional likelihood (3) for the i^{th} stratum can be written as

$$\begin{aligned} L_i &= \frac{\frac{exp(\delta_2 E_{i0})}{1+exp(\delta_0+\delta_1+\delta_2 E_{i0})} exp(\beta E_{i0}) \prod_{h=1}^{c} \frac{exp(\delta_2 E_{ih})}{1+exp(\delta_0+\delta_2 E_{ih})}}{\sum_{h=0}^{c} \frac{exp(\delta_2 E_{ih})}{1+exp(\delta_0+\delta_1+\delta_2 E_{ih})} exp(\beta E_{ih}) \prod_{l\neq h} \frac{exp(\delta_2 E_{il})}{1+exp(\delta_0+\delta_2 E_{il})}} \\ &= \frac{g_{i0}(\delta) exp(\beta E_{i0})}{\sum_{h=0}^{c} g_{ih}(\delta) exp(\beta E_{ih})}, \end{aligned} \tag{16}$$

where $\delta' = (\delta_0, \delta_1, \delta_2)$ and $g_{ih}(\delta) = \frac{exp(\delta_2 E_{ih})}{1+exp(\delta_0+\delta_1+\delta_2 E_{ih})} \prod_{l\neq h} \frac{exp(\delta_2 E_{il})}{1+exp(\delta_0+\delta_2 E_{il})}$, for $h = 0, 1, \cdots, c$. When $P(R=1|A,E)$ depends on only A, or only E, then the likelihood (16) is free of δ as, for given $E^{(i)}$, the function $g_{ih}(\delta)$ is constant with respect to h (See Section 2). The matched case-control conditional likelihood from the SR database is, in general,

$$L_{SR} = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} \frac{g_{i0}(\delta) exp(\beta E_{i0})}{\sum_{h=0}^{c} g_{ih}(\delta) exp(\beta E_{ih})}. \tag{17}$$

As in Section 3, here also we have no information on the reporting parameters δ in the SR database. Therefore, in order to estimate β, we need extra information from outside the SR database. As in Section 3.2, we consider a reference sample of size m drawn from the source population and observe (A, E, R) on those in the sample. Writing

$$P(A,E,R) \quad = \quad P(R|A,E) \cdot P(A,E), \tag{18}$$

and, as argued in Section 3.2, the second factor on the right hand side of (18) contains no available information concerning β in the absence of knowledge on the nuisance parameters γ_M's. We, therefore, use only the first part, $P(R|A,E)$, as conditional likelihood of observing R, given A and E, from the reference sample, which can be written as

$$\begin{aligned} L_R &= \prod_{i=1}^{m} P(R_i|A_i, E_i) \\ &= \prod_{i=1}^{m} \left\{ \frac{exp[(\delta_0+\delta_1 A_i+\delta_2 E_i) R_i]}{1+exp[\delta_0+\delta_1 A_i+\delta_2 E_i]} \right\}. \end{aligned} \tag{19}$$

Combining (17) and (19), we have the total likelihood as $L_3^* = L_{SR} \times L_R$ and

$$\begin{aligned} l(\phi) = logL_3^* &= logL_{SR} + logL_R \\ &= \sum_{i=1}^{n} log \left\{ \frac{g_{i0}(\delta) exp(\beta E_{i0})}{\sum_{h=0}^{C} g_{ih}(\delta) exp(\beta E_{ih})} \right\} + \sum_{i=1}^{m} log \left\{ \frac{exp[(\delta_0+\delta_1 A_i+\delta_2 E_i) R_i]}{1+exp[\delta_0+\delta_1 A_i+\delta_2 E_i]} \right\}, \end{aligned} \tag{20}$$

where $\phi' = (\delta', \beta)$. Maximum likelihood estimate of ϕ, denoted by $\hat{\phi}_{n+m}$, is obtained by maximizing $l(\phi)$. The corresponding score vector can be written as

$$\frac{\partial l(\phi)}{\partial \phi} \quad = \quad \sum_{i=1}^{n} U_1(E^{(i)}, M_{i0}; \phi) + \sum_{i=1}^{m} U_2(A_i, E_i, R_i; \phi), \tag{21}$$

where $U_1(E^{(i)}, M_{i0}; \phi) = \left[\frac{\partial}{\partial\delta} log\left\{\frac{g_{i0}(\delta)exp(\beta E_{i0})}{\sum_{h=0}^{C} g_{ih}(\delta)exp(\beta E_{ih})}\right\}, \;\; \frac{\partial}{\partial\beta} log\left\{\frac{g_{i0}(\delta)exp(\beta E_{i0})}{\sum_{h=0}^{C} g_{ih}(\delta)exp(\beta E_{ih})}\right\}\right]$ is the individual (i^{th}) score contribution from the SR database, and

$U_2(A_i, E_i, R_i; \phi) = \left[\frac{\partial}{\partial\delta} log\left\{\frac{exp((\delta_0+\delta_1 A_i+\delta_2 E_i)R_i)}{1+exp(\delta_0+\delta_1 A_i+\delta_2 E_i)}\right\}, \;\; 0\right] = \left[U_2^*(A_i, E_i, R_i; \delta), \;\; 0\right]$ (say) is the individual (i^{th}) score contribution from the reference sample. In the next Section, we consider the asymptotic properties of the estimated parameters using multi-sample structure of the data.

5 Asymptotics

In this section, we derive some asymptotic results concerning the estimates of the model parameters, as described in Section 4. Similar results can be derived for the estimates in Section 3.2 by using the same arguments. We first establish asymptotic normality for the two score vectors arising from the two parts of the likelihood L_{SR} and L_R, and then combine the two using the results of multi-sample central limit theorem (CLT).

Note that, for matched case-control data from the SR database, a case (say, the i^{th} with $A_{i0} = 1$) is first chosen and then the corresponding exposure (E_{i0}) and matching variable (M_{i0}) are recorded; then c controls (with $A_{ij} = 0$, for $j = 1, \cdots, c$) with the same value (M_{i0}) for the matching variable are chosen randomly from the SR database and their exposure ($E_{i1}, \cdots, E_{ic}$) are recorded. This data can be formally written as $\{E_{i0}, M_{i0}|A_{i0} = 1\} \cap \{E_{i1}, \cdots, E_{ic}|M_{i0}, A_{ij} = 0, j = 1, \cdots, c\}$, which, by design, is equivalent to $\{M_{i0}, E_{i0}, E_{i1}, \cdots, E_{ic}\}$. Note that this observation vector, for $i = 1, \cdots, n$, are independent and identically distributed. So, the individual score vectors $U_1(E^{(i)}, M_{i0}; \phi)$, for $i = 1, \cdots, n$, can also be viewed as independent and identically distributed, although these have been derived from a conditional likelihood. Note that, in the i^{th} stratum, the conditional probability that exposure E_{ik} is that of the case with the value of the matching variable M_{i0} and ($E_{ih}, h \neq k$) are those of the c controls, given the unordered set $E^{(i)}$ of $c + 1$ exposures, is

$$\frac{g_{ik}(\delta)exp(\beta E_{ik})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})}, \text{ for } k = 0, 1, \cdots, c. \quad (22)$$

It is shown in Appendix B, that $E[U_1(E^{(i)}, M_{i0}; \phi)] = 0$ and the variance-covariance matrix $V[U_1(E^{(i)}, M_{i0}; \phi)]$ is constant not depending on i, where the expectation is taken with respect to the common distribution of $\{M_0, E_0, E_1, \cdots, E_c\}$ under the probability space as generated by the matched case-control sampling from the SR database. Using central limit theorem, it can be shown that $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_1(E^{(i)}, M_{i0}; \phi)$ follows asymptotically a 4-variate normal distribution with zero mean and variance-covariance matrix given by $V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$, where

$$\begin{aligned} V_{11} &= E_{\mathbb{E}}\left[\frac{\sum_{k=0}^{c} \frac{1}{g_{ik}(\delta)} \left(\frac{\partial g_{ik}(\delta)}{\partial\delta}\right)' \left(\frac{\partial g_{ik}(\delta)}{\partial\delta}\right) exp(\beta E_{ik})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})}\right. \\ &\left. - \frac{\left\{\sum_{h=0}^{c} \frac{\partial g_{ih}(\delta)}{\partial\delta} exp(\beta E_{ih})\right\}' \left\{\sum_{h=0}^{c} \frac{\partial g_{ih}(\delta)}{\partial\delta} exp(\beta E_{ih})\right\}}{\left\{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})\right\}^2}\right], \end{aligned}$$

$$
\begin{aligned}
V_{22} &= E_{\mathbb{E}}\left[\frac{\sum_{k=0}^{c}\left\{\sum_{h=0}^{c}(E_{ik}-E_{ih})g_{ih}(\delta)exp(\beta E_{ih})\right\}^{2} g_{ik}(\delta)exp(\beta E_{ik})}{\left[\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})\right]^{3}}\right], \\
V_{12}' &= V_{21} = E_{\mathbb{E}}\left[\sum_{k=0}^{c}\left[\left\{\frac{\frac{\partial g_{ik}(\delta)}{\partial\delta}}{g_{ik}(\delta))} - \frac{\sum_{h=0}^{c}\frac{\partial g_{ih}(\delta)}{\partial\delta}exp(\beta E_{ih})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})}\right\}'\right.\right. \\
&\times \left.\left.\left\{\frac{\sum_{h=0}^{c}(E_{ik}-E_{ih})g_{ih}(\delta)exp(\beta E_{ih})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})}\right\} \times \frac{g_{ik}(\delta)exp(\beta E_{ik})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})}\right]\right],
\end{aligned}
$$

with the expectation $E_{\mathbb{E}}[\cdot]$ being with respect to the distribution of $E^{(i)}$.

In the reference sample, the observation $\{A_i, E_i, R_i\}$, for $i = 1, \cdots, m$, are independent and identically distributed. So, $U_2^*(A_i, E_i, R_i; \delta)$, for $i = 1, \cdots, m$, are also i.i.d. random variables, where U_2^* denotes the first three non-zero components of U_2. It is again shown in Appendix B, that $E[U_2^*(A, E, R; \delta)] = 0$ and $V[U_2^*(A, E, R; \delta)]$ is constant, where the expectation is taken with respect to the common distribution of $\{A, E, R\}$. Note that, for the i^{th} observation, the conditional probability of R_i, given (A_i, E_i) , is $exp[(\delta_0+\delta_1 A_i+\delta_2 E_i)R_i]/(1+exp[\delta_0+\delta_1 A_i+\delta_2 E_i])$, for $R_i = 0, 1$. Using central limit theorem, $\frac{1}{\sqrt{m}}\sum_{i=1}^{m} U_2^*(A_i, E_i, R_i; \delta)$ follows asymptotically a 3-variate normal distribution with zero mean and variance-covariance matrix given by

$$
W_{11} = E_{(A,E)}\left[\frac{z_i z_i' exp(z_i'\delta)}{[1+exp(z_i'\delta)]^2}\right], \text{ with } z' = (1, A_i, E_i),
$$

where the expectation is taken with respect to the common distribution of (A, E).

Following Lehmann and Casella (1998, p475-476) and Hirose (2005, p68) (See also Bradley and Gart, 1962), we view this as a multi-sample, or s-sample, framework with s=2 corresponding to the matched case-control sample from the SR database and the reference sample from the source population. We assume that $n/(n + m) \to \omega$, $0 < \omega < 1$, as both n and $m \to \infty$. Then, using the results from the s-sample framework, the vector $\sqrt{n + m}(\hat{\phi}_{n+m} - \phi)$ follows asymptotically a multivariate normal distribution with mean zero and variance-covariance matrix given by

$$
\left(\omega\begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22}\end{bmatrix} + (1-\omega)\begin{bmatrix} W_{11} & 0 \\ 0 & 0\end{bmatrix}\right)^{-1}, \tag{23}
$$

which can be estimated from the observed information matrix based on the likelihood L_3^*. As a consequence, the asymptotic distribution of $\sqrt{n + m}(\hat{\beta}_{n+m} - \beta)$, with $\hat{\beta}_{n+m}$ being the estimate of β, which is of primary interest, can be seen to be normal with mean zero and variance given by the $(4, 4)^{th}$ entry of the matrix in (23). The details of these derivations are given in Appendix C.

6 Illustration and Simulation Study

In Section 6.1, we illustrate the estimation procedure of Section 3.1 based on likelihood (6), using an Italian SR database. Section 6.2 presents a simulation study to investigate the properties of the estimates based on (6) of Section 3.1, while the same for the estimates based on (20) of Section 4 is investigated in Section 6.3. All the computation has been made using statistical software R.

6.1 Analysis of an Italian SR database

In this subsection, we perform a matched case-control analysis to investigate safety profile of the drug amoxicillin (AMX representing $E = 1$) and amoxicillin/clavulanic acid (AMC representing $E = 0$) based on the data obtained from spontaneous reporting of suspected adverse drug reaction to an Italian database. This data has been reported from the spontaneous reporting system of six Italian regions from January 1988 to June 2005, as described in Salvo et al. (2007). The drug utilization data is also available for the two drugs. Table 1 presents the organ-specific case-control data for the two drugs, where a case is represented by 'serious' ADR and a control is by 'non-serious' ADR. No significant difference has been found in the two groups (case and control) with respect to distribution of sex, average age, number of patients with concomitant drugs and number of reports in paediatric age (See Salvo et al., 2007). In order to illustrate the methodology described in Section 3.1, we consider organ as the matching variable and each case has been matched with $c = 2$ randomly chosen controls from those with the same value of the matching variable (that is organ). It is clear from Table 1 that, in some of the matched groups, sufficient number of controls is not available for matching. For the sake of simplicity, we consider that many cases in our analysis for which two matched controls per case are available. As a result, the data for the analysis consists of $n = 268$ cases with two randomly chosen matched controls for each case. We assume 100% reporting of the cases in the two exposure groups (i.e., $\theta_{11} = \theta_{01} = 1$), but the controls may be subject to reporting bias with $\theta_{10} \neq \theta_{00}$. The source population size N is required to estimate θ using (9), which is found to be robust for different values of N. For the sake of illustration, we take N as 1,000,000. During the study period, AMX was consumed more often (5.71 DDD/1000 inhabitants/day) than AMC (5.04 DDD/1000 inhabitants/day) in the study area, where DDD stands for defined daily doses. So, the exposure probability for AMX, $P(E = 1)$, in the source population is taken as $5.71/(5.71 + 5.04) = 0.53$. The cell frequencies n_{uv}'s have been obtained by summing up the corresponding cell frequencies over the different organ groups (See Table 1) as $n_{11} = 205, n_{01} = 228, n_{10} = 980$ and $n_{00} = 976$, leading to the estimate of θ as 0.89. The estimate $\hat{\beta}$ from the given data with randomly chosen set of controls is -0.04 with standard error 0.152 based on the adjusted estimate given by (12). The alternative estimate of variance of $\hat{\beta}$, as given at the end of the Section 3.1, also gives a standard error of 0.152. Ignoring the reporting bias in control group and using the SR data only, estimate of β is obtained as 0.07 with standard error 0.151 based on observed information matrix. Although the estimates of β, obtained by both adjusting for and ignoring reporting bias, turn out to be statistically insignificant, they are of opposing sign indicating some difference. Possibly with larger sample size, significant evidence in favour of an association between AMC and serious ADR may be established by adjusting for reporting bias, while ignoring it will give evidence of an opposite nature. By considering the first two organs (Skin and Gastrointestinal) as the only categories for the matching variable, the estimate $\hat{\beta}$ is obtained as 0.1 and 0.217 by considering reporting bias and ignoring it, respectively, with corresponding standard errors 0.181 and 0.180, respectively.

6.2 Simulation with Binary Exposure

We consider a source population of size $N = 20000$ and age as the only matching variable generated from a log-normal distribution with corresponding normal mean and standard deviation as 3.25 and 0.4, respectively. The corresponding mean age is 25.79 with the standard deviation 11.64. Then, for the purpose of matching, we categorize age into 10-year groups 0-9, 10-19, $\cdots$, 90 and above, with the corresponding transformed age-group value 1,2, $\cdots$,10, respectively. For convenience, this transformed age-group value is taken as the matching variable. The binary exposure for each individual in the source population is generated from the the Bernoulli distribution with probability 0.4. We simulate each individual's ADR status by the Bernoulli distribution with probability calculated from the logistic model (4) with $\beta = 0.7$ and $\gamma_{M_i} = \alpha + \gamma M_i$, where $\alpha = -0.9$, $\gamma = -0.6$ and M_i is the value of the transformed age-group of the i^{th} individual. The SR database is generated from the source population by considering the reporting status of all the cases to be $R = 1$ (as in Section 3.1) and that of each control being generated from the Bernoulli distribution with probability θ_{10} or θ_{00}, depending on its exposure status. We consider three sets of values for $(\theta_{10}, \theta_{00})$ as given by $(0.45, 0.75), (0.75, 0.45)$ and $(0.45, 0.45)$, with values of $\theta = \theta_{10}/\theta_{00}$ being less than, more than and equal to 1, respectively, resulting in positive, negative and zero reporting bias in the estimation of β from the SR database alone, as noted in Section 3.1. Now, identifying only those with reporting status $R = 1$ in the source population, we have the SR database. We then randomly select c (here $c = 2$) number of matched controls for each case in the SR data, if available. For the sake of simplicity, we discard the cases for which two matched controls are not available in SR database. For the purpose of illustration, the model parameters have been so chosen such that the resulting SR database is expected to have enough number of controls for each case. On the average, there are about 2000 cases and 8000-10000 controls in the SR database. Binary exposure status of each case and its two matched controls are noted to form the observed data for likelihood (6).

For each simulation, the exposure probability $P(E = 1)$ is estimated by considering a random sample of size m from the source population and observing their exposure status. Using information obtained from the SR database, reference sample and source population size $N = 20000$, the estimate of θ is obtained from (9). The estimate of β is then obtained by solving the quadratic equation (10) using the estimate $\hat{\theta}$ of θ. The variance estimate of $\hat{\beta}$ is obtained from the observed information based on the likelihood (6). This estimated variance is very close to the variance of $\hat{\beta}$ obtained from the simulation study of 5000 simulations, when reference sample size m is large enough. But as discussed in Section 3.1, the observed information based variance estimate (not reported) gives an under-estimate for moderate value of m. For such values of m, the variance estimate is obtained by the expression (12) and also the alternate expression suggested at the end of the Section 3.1, both of which give similar values. Table 2 presents the results of the simulation study for different values of m, which includes estimates of β with corresponding standard errors in parentheses (obtained from the 5000 simulations) along with the same for θ. The results shown in Table 2 suggest that the methodology of Section 3.1 performs well in all the three cases, while the method based on SR data alone gives satisfactory estimate only when $\theta_{10} = \theta_{00}$. Also, as expected, the standard error of the proposed estimate of Section 3.1 decreases with m. However, since the estimate based on SR data alone does not utilize the information in the reference sample, the corresponding standard errors do not change with m. Asymptotic normality of $\hat{\beta}$ has been evident through the Q-Q plot and histogram of the estimates from 5000 simulations (not shown here). The consistency of $\hat{\theta}$ is also evident from the results in Table 2. Interestingly, the relationship between the reporting odd-ratio (ROR) and the population odds-ratio (OR), as mentioned in the beginning of Section 3, can be seen from the two sets of estimates of β in Table 2. Specifically, the estimate of β from the combined data *minus* logarithm of the corresponding $\hat{\theta}$ is expected to be equal to the estimate of β obtained from SR data alone. Consistency of this relationship with increasing m is evident.

6.3 Simulation with Continuous Exposure

The methodology described in Section 4 does not require information on the size N of the source population; however, in order to describe the simulation model, let us consider $N = 20000$. As in subsection 6.2, age is assumed to be the only matching variable having a log-normal distribution with corresponding normal mean and standard deviation being 3.25 and 0.4, respectively. The transformed age-group values are as in subsection 6.2. The exposure value is now generated from a gamma distribution with the shape and scale parameter values as 10 and 0.5, respectively. As in subsection 6.2, the ADR status of each individual is generated by taking a sample from the Bernoulli distribution with success probability given by the logistic model (4) with $\beta = 0.7, \alpha = -0.8$ and $\gamma = -0.1$. The reporting status of an individual is generated by considering the reporting probability model $P(R = 1|A, E)$ given by (15), where the values of the reporting parameters $(\delta_0, \delta_1, \delta_2)$ are taken as $(-2.2, -0.8, 0.5), (-2.2, 0.8, 0.5)$ and $(-2.2, 0, 0)$ to reflect different nature of reporting bias in the analysis of SR data alone for estimating β, the parameter of interest (See Table 3). The SR data and the set of cases with two matched controls are generated following the same procedure of subsection 6.2.

We take a random sample of size m from the source population and observe the ADR status, value of the exposure variable and the reporting status of each individual in the sample. Exposure values for each case and its two matched controls, combined with the information from the reference sample, constitute the observed data to form the log-likelihood (20). The R optimization program has been used to get the parameter estimates, which is repeated for 5000 simulations. The results of the simulation study are presented in Table 3. The estimate of β based on SR data alone is seen to be biased (except in the third case with $\delta_1 = \delta_2 = 0$, as expected), while the same from the combined data seems to perform well. The standard errors obtained from the corresponding observed information matrix (not reported here) are close to the ones obtained from the 5000 simulations which are reported in Table 3. The standard errors decrease with increasing size m of the reference sample, as expected, in the analysis of combined data. As remarked in subsection 6.2, the standard errors in the analysis of SR data alone do not change with m.

7 Concluding remarks

Matched case-control study has been used routinely to adjust for the confounding effects of different variables on the response. But, the situation becomes much more difficult when case or control sample or both suffer from some kind of selection bias. In this work, we have explored different possibilities to adjust for selection bias in matched case-control study in the context of the analysis of SR data, wherein bias arises due to unequal reporting. Two distinct situations have been considered corresponding to the exposure of interest being binary or continuous. To overcome the difficulties of selection/reporting bias, some external informations has been used. Note that there is a trade-off between the assumption made about the available data and the amount of external information required. The use of reference sample in fitting binary regression models has been previously considered by Lee et al. (2006), but without the presence of reporting bias. The design 1 in that work is similar to the situation in Section 3.1, where only the exposure information is used from the reference sample, whereas design 2 is same as the situation in Sections 3.2 and 4.

Although the methodology has been motivated from the analysis of SR data, there are other areas where the similar situation arises. In an interview-based case-control study of breast cancer

(See Rookus, 2000), reporting bias may distort the relation between a history of induced abortion and risk of breast cancer, as induced abortions are markedly under-reported among the controls because of its sensitive nature. The case-control studies where the exposure history is collected through telephonic interview, or in hospital-based studies where controls are selected from the same hospital registry as that of the cases resulting in some selection bias (See Cosslett, 1981), the methodology of this paper can be useful. The standard case-control analysis tend to ignore this problem, whereas the consideration of supplementary data, as in the proposed methodology, can remove the corresponding bias.

In Section 1, we have discussed various types of matching which are commonly used in practice. For the sake of simplicity, we have considered matching of one case with a fixed number of controls. Other types of matching, specially one case matched with variable number controls (See Rosenbaum, 2002) may be more appropriate, depending on the availability of matched controls in application. For the matched case-control study, propensity score can be used for both maintaining the balance and grouping of similar stratum (See Rosenbaum, 2002).

The duration of medication, for a particular drug or drugs may play an important role to develop an ADR. Some ADRs can only occur after a certain period of continuous medication. When this kind of ADR is of concern, the analysis needs to be adjusted for the confounding variable 'medication-duration', which is available in SR data. This can be incorporated by means of an additional covariate in model (4) in the proposed methodology of this paper.

Prentice and Breslow (1978) also considered inclusion probabilities for case and control samples and assumed that those probabilities are independent of explanatory variables. In this paper, this assumption is same as $P(R=1|A,E)$ being independent of E, resulting in the matched conditional likelihood (3) being free of the reporting parameters. The likelihood (3) can also be seen in the context of failure time analysis of retrospective studies, as in Prentice and Breslow (1978), where the product of the likelihood contributions (2), L_i's, is taken over the different failure time points.

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of Royal Statistical Society, Series B* **32**, 283–301.

Anello, C. & O'Neill, R. T. (1998). *Encyclopedia of Biostatistics*, vol. 4, chap. Postmarketing Surveillance of New Drugs and Assessment of Risk. Armitage P, Colton T (eds). Wiley: New York, pp. 3450–3458.

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.

Bradley, R. A. & Gart, J. J. (1962). The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika* **49**, 205–214.

Breslow, N. E. (1996). Statistics in Epidemiology: the case-control study. *Journal of American Statistical Association* , 14–28.

Cosslett, S. R. (1981). Maximum likelihood estimator for choice-based samples. *Econometrica* **49**, 1289–1316.

DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The American Statistician* **53**, 177–190.

Evans, S. J. W. (2004). *Stephens' Detection of New Adverse Drug Reactions*, chap. Statistics: Analysis and Presentation of Safety Data. Talbot J, Waller P (eds). John Wiley and Sons Ltd, 5th ed.

Evans, S. J. W., Waller, P. C. & Davis, S. (2001). Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety* **10**, 483–486.

Ghosh, P. & Dewanji, A. (2011). Analysis of spontaneous adverse drug reaction (ADR) reports using supplementary information. *Statistics in Medicine* **30**, 2040–2055.

Hirose, Y. (2005). *Efficiency of the semi-parametric maximum likelihood estimator in generalized case-control studies.* Ph.D. thesis, Univ. Auckland.

Kalbfleisch, J. D. & Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of Royal Statistical Society, Series B* **32**, 175–208.

Kalbfleisch, J. D. & Sprott, D. A. (1973). Marginal and conditional likelihoo. *Sankhya A* **35**, 311–328.

Kang, Q., Nelson, P. I. & Vahl, C. I. (2010). Parameter estimation from an outcome dependent enriched sample using weighted likelihood method. *Statistica Sinica* **20**, 1529–1550.

Lee, A. J., Scott, A. J. & Wild, C. J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika* **93**, 385–397.

Lehmann, E. L. & Casella, G. (1998). *Theory of Point Estimation.* Springer, 2nd ed.

Lin, I. F. & Paik, M. C. (2001). Matched case-control data analysis with selection bias. *Biometrics* **57**, 1106–1112.

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

Rookus, M. A. (2000). Invited Commentary: Reporting Bias in Case-Control Studies on Induced Abortion and Breast Cancer. *American Journal of Epidemiology* **151**, 1144–1147.

Rosenbaum, P. R. (2002). *Observational Studies.* Springer, 2nd ed.

Roux, E., Thiessard, F., Fourrier, A., Begaud, B. & Tubert-Bitter, P. (2005). Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance. *IEEE Transactions on Information Technology in Biomedicine* **9**, 518–527.

Salvo, F., Polimeni, G., Moretti, U., Conforti, A., Leone, R., Leoni, O., Motola, D., Dusi, G. & Caputi, A. P. (2007). Adverse drug reactions related to amoxicillin alone and in association with clavulanic acid: data from spontaneous reporting in Italy. *Journal of Antimicrobial Chemotherapy* **60**, 121–126.

van der Heijden, P. G. M., van Puijenbroek, E. P., van Buuren, S. & van der Hofstede, J. W. (2002). On the assessment of adverse drug reactions from spontaneous reporting systems: the influence of under-reporting on odds ratios. *Statistics in Medicine* **21**, 2027–2044.

Table 1: Organ Specific serious ADR (as case) and non-serious ADR (as control) due to use of Amoxicillin (AMX) and Amoxicillin/Clavulanic acid (AMC)

Organ	AMX ($E = 1$)		AMC ($E = 0$)	
	Case	Control	Case	Control
Skin	81	820	66	757
Gastrointestinal system	17	65	17	121
Liver and biliary system	4	2	28	12
Haematological	10	0	22	1
Body as whole	41	55	39	44
Urinary and reproductive	30	5	26	6
Respiratory system	15	13	19	14
Cardiovascular system	6	13	10	9
Others	1	7	1	12

Table 2: Simulation results on the estimate of β using SR database with and without exposure information for a source population of size 20,000. The true value of β is 0.7. Corresponding standard errors are in parentheses.

m	$(\theta_{10}, \theta_{00})$	$\theta = \frac{\theta_{10}}{\theta_{00}}$	$\hat{\theta}$ (SE)	$\hat{\beta}$ (SE)	
				SR Data alone	Combined data
100	(0.45, 0.75)	0.60	0.622 (0.144)	1.213 (0.065)	0.713 (0.236)
300			0.607 (0.080)	1.211 (0.065)	0.703 (0.145)
500			0.605 (0.063)	1.212 (0.065)	0.704 (0.120)
1000			0.602 (0.043)	1.212 (0.064)	0.702 (0.094)
5000			0.600 (0.019)	1.212 (0.064)	0.702 (0.069)
100	(0.75, 0.45)	1.67	1.722 (0.405)	0.189 (0.059)	0.706 (0.233)
300			1.687 (0.222)	0.189 (0.059)	0.703 (0.143)
500			1.678 (0.169)	0.188 (0.060)	0.701 (0.115)
1000			1.672 (0.121)	0.189 (0.061)	0.701 (0.092)
5000			1.667 (0.051)	0.191 (0.059)	0.702 (0.065)
100	(0.45, 0.45)	1	1.033 (0.241)	0.701 (0.061)	0.707 (0.233)
300			1.013 (0.134)	0.701 (0.061)	0.704 (0.144)
500			1.006 (0.103)	0.700 (0.061)	0.701 (0.116)
1000			1.004 (0.072)	0.700 (0.061)	0.701 (0.091)
5000			1.000 (0.032)	0.701 (0.062)	0.701 (0.066)

8 Appendix

Appendix A: Adjusted approximate variance formula for variance of $\hat{\beta}$ of Section 3.1
We can write

$$\frac{1}{n}\sum_{i=1}^{n} u_i(\beta, \hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} u_i(\beta, \theta) + (\hat{\theta} - \theta)\frac{1}{n}\sum_{i=1}^{n} \frac{\partial u_i(\beta, \theta)}{\partial \theta} + o_p(1). \tag{24}$$

As $\hat{\theta}$ is consistent for θ and $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial u_i(\beta,\theta)}{\partial \theta} \xrightarrow{P} E(\frac{\partial u(\beta,\theta)}{\partial \theta})$, the second term is also $o_p(1)$, but this convergence may be slow. Therefore, the asymptotic distribution function of $\frac{1}{n}u(\beta, \hat{\theta})$ is same as that of $\frac{1}{n}u(\beta, \theta)$, but there may be bias in variance estimation for small sample because of the

Table 3: Simulation results on the estimate of β using SR database with and without the reference sample for a source population of size 20,000. True value of β is 0.7. Corresponding standard errors are in parentheses.

m	$(\delta_0, \delta_1, \delta_2)$	$\hat{\beta}$ (SE)	
		SR Data alone	Combined data
100	(-2.2, -0.8, 0.5)	0.780 (0.027)	0.687 (0.067)
150		0.779 (0.026)	0.687 (0.058)
200		0.779 (0.027)	0.686 (0.053)
500		0.779 (0.026)	0.688 (0.039)
1000		0.779 (0.027)	0.691 (0.036)
100	(-2.2, 0.8, 0.5)	0.616 (0.023)	0.686 (0.051)
150		0.616 (0.023)	0.689 (0.042)
200		0.616 (0.023)	0.689 (0.038)
500		0.616 (0.022)	0.690 (0.029)
1000		0.616 (0.023)	0.691 (0.027)
100	(-2.2, 0, 0)	0.693 (0.057)	0.686 (0.061)
150		0.694 (0.057)	0.690 (0.059)
200		0.692 (0.056)	0.690 (0.057)
500		0.693 (0.056)	0.692 (0.056)
1000		0.694 (0.056)	0.693 (0.057)

second term. Using higher order asymptotics, and ignoring covariance term due to dependence of $\hat{\theta}$ on SR data as well, this variance may be taken as

$$V\left(\frac{1}{n}u(\beta,\hat{\theta})\right) = V\left(\frac{1}{n}u(\beta,\theta)\right) + \left\{E\left(\frac{\partial u(\beta,\theta)}{\partial\theta}\right)\right\}^2 V(\hat{\theta}). \tag{25}$$

Note that $E\left(\frac{\partial u(\beta,\theta)}{\partial\theta}\right)$ is estimated by $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial u_i(\beta,\theta)}{\partial\theta}$, evaluated at $\beta=\hat{\beta}, \theta=\hat{\theta}$. Variance of $\hat{\theta}$ can be obtained from (9) by applying delta method and using the estimated covariance matrix of $(n_{11}, n_{01}, n_{10}, n_{00}, m_{1\cdot})$. In order to define the covariance matrix of $(n_{11}, n_{01}, n_{10}, n_{00}, m_{1\cdot})$, note that $m_{1\cdot}$ and $(n_{11}, n_{01}, n_{10}, n_{00})$ are independent with

$$\begin{aligned}
m_{1\cdot} &\sim Bin(m, P(E=1)) \\
(n_{11}, n_{01}, n_{10}, n_{00}, N - n_{\cdot\cdot}) &\sim Multinomial\Big(N, P[E=1, A=1], P[E=0, A=1], \\
&\qquad P[E=1, A=0, R=1], P[E=0, A=0, R=1], P[R=0]\Big),
\end{aligned}$$

where $n_{\cdot\cdot} = \sum_{u=0}^{1}\sum_{v=0}^{1} n_{uv}$. From (25), we have

$$\hat{V}\left(u(\beta,\hat{\theta})\right) \approx \hat{V}\left(u(\beta,\theta)\right) + \left\{\sum_{i=1}^{n}\frac{\partial u_i(\beta,\theta)}{\partial\theta}\Big|_{\beta=\hat{\beta},\theta=\hat{\theta}}\right\}^2 \hat{V}(\hat{\theta}), \tag{26}$$

where the first term in RHS is the observed information form L_1^*. We can write

$$\begin{aligned}
u(\hat{\beta},\hat{\theta}) &= u(\beta,\hat{\theta}) + (\hat{\beta}-\beta)\frac{\partial u}{\partial\beta} + o_p(1) = 0 \\
\Rightarrow (\hat{\beta}-\beta) &\approx u(\beta,\hat{\theta})\left(-\frac{\partial u}{\partial\beta}\right)^{-1} \\
\Rightarrow V(\hat{\beta}) &\approx \left[\left(-\frac{\partial u}{\partial\beta}\right)^{-1}\right]^T V(u(\beta,\hat{\theta}))\left[\left(-\frac{\partial u}{\partial\beta}\right)^{-1}\right].
\end{aligned} \tag{27}$$

An estimate of $V(\hat{\beta})$ is obtained from (27) by replacing $\left(-\frac{\partial u}{\partial\beta}\right)$ with $-\sum_{i=1}^{n}\frac{\partial u_i(\beta,\theta)}{\partial\beta}$, evaluated at $\beta=\hat{\beta}, \theta=\hat{\theta}$, and $V(u(\beta,\hat{\theta}))$ with $\hat{V}(u(\beta,\hat{\theta})$ from (26)

Appendix B: Mean and Variance of U_1 and U_2^*

Note that $E[U_1(E^{(i)}, M_{i0}; \phi)] = E_{\mathbb{E}}E[U_1(E^{(i)}, M_{i0}; \phi)|E^{(i)}]$, where the outer expectation is with respect to the distribution of $E^{(i)}$. Using (22),

$$\begin{aligned}
&E[U_1(E^{(i)}, M_{i0}; \phi)|E^{(i)}] \\
= &\sum_{k=0}^{c} \left[\frac{\partial}{\partial\phi}\{log(g_{ik}(\delta)exp(\beta E_{ik})))\} - \frac{\partial}{\partial\phi}\{log\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih}))\} \right] \cdot \frac{g_{ik}(\delta)exp(\beta E_{ik})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})} \qquad (28)
\end{aligned}$$

Differentiating with respect to δ, the first component in (28) is

$$\begin{aligned}
&\sum_{k=0}^{c} \left[\left\{ \frac{\frac{\partial g_{ik}(\delta)}{\partial\delta}}{g_{ik}(\delta)} - \frac{\sum_{h=0}^{c} \frac{\partial g_{ih}(\delta)}{\partial\delta} exp(\beta E_{ih})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})} \right\} \times \frac{g_{ik}(\delta)exp(\beta E_{ik})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})} \right] \\
= &\frac{\sum_{k=0}^{c} \frac{\partial g_{ik}(\delta)}{\partial\delta} exp(\beta E_{ik})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})} - \frac{\left[\sum_{h=0}^{c} \frac{\partial g_{ih}(\delta)}{\partial\delta} exp(\beta E_{ih})\right]\left[\sum_{k=0}^{c} g_{ik}(\delta)exp(\beta E_{ik})\right]}{\left[\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})\right]^2} \\
= &\frac{\sum_{k=0}^{c} \frac{\partial g_{ik}(\delta)}{\partial\delta} exp(\beta E_{ik})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})} - \frac{\sum_{k=0}^{c} \frac{\partial g_{ik}(\delta)}{\partial\delta} exp(\beta E_{ik})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})} \\
= &\ 0
\end{aligned}$$

while, differentiating with respect to β, the second component is

$$\begin{aligned}
&\sum_{k=0} \left[\left\{ E_{ik} - \frac{\sum_{h=0}^{c} E_{ih} g_i(\delta)exp(\beta E_{ih})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})} \right\} \times \frac{g_{ik}(\delta)exp(\beta E_{ik})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})} \right] \\
= &\sum_{k=0}^{c} \left[\frac{\sum_{h=0}^{c} (E_{ik} - E_{ih}) g_{ih}(\delta)exp(\beta E_{ih})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})} \right] \times \frac{g_{ik}(\delta)exp(\beta E_{ik})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})} \\
= &\frac{\sum_{k=0}^{c}\sum_{h=0}^{c} E_{ik} g_{ih}(\delta)exp(\beta E_{ih}) g_{ik}(\delta)exp(\beta E_{ik})}{\left[\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})\right]^2} \\
&\qquad - \frac{\sum_{k=0}^{c}\sum_{h=0}^{c} E_{ih} g_{ih}(\delta)exp(\beta E_{ih}) g_{ik}(\delta)exp(\beta E_{ik})}{\left[\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})\right]^2} \\
= &\ 0.
\end{aligned}$$

Therefore, $E[U_1(E^{(i)}, M_{i0}; \phi)] = 0$. Now,

$$\begin{aligned}
V[U_1(E^{(i)}, M_{i0}; \phi)] &= E_{\mathbb{E}}V[U_1(E^{(i)}, M_{i0}; \phi)|E^{(i)}] + V_{\mathbb{E}}E[U_1(E^{(i)}, M_{i0}; \phi)|E^{(i)}] \\
&= E_{\mathbb{E}}V[U_1(E^{(i)}, M_{i0}; \phi)|E^{(i)}],
\end{aligned}$$

where,

$$\begin{aligned}
&V[U_1(E^{(i)}, M_{i0}; \phi)|E^{(i)}] = E[U_1' U_1|E^{(i)}] \\
= &\sum_{k=0}^{c} \left[\frac{\partial}{\partial\phi}\{log(g_{ik}(\delta)exp(\beta E_{ik})))\} - \frac{\partial}{\partial\phi}\{log\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih}))\} \right]' \left[\frac{\partial}{\partial\phi}\{log(g_{ik}(\delta)exp(\beta E_{ik})))\} \right. \\
- &\left. \frac{\partial}{\partial\phi}\{log\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih}))\} \right] \cdot \frac{g_{ik}(\delta)exp(\beta E_{ik})}{\sum_{h=0}^{c} g_{ih}(\delta)exp(\beta E_{ih})}. \qquad (29)
\end{aligned}$$

The three different entries in the matrix in (29) are

$$\begin{aligned}
&\sum_{k=0}^{c} \frac{\left(\frac{\partial g_{ik}(\delta)}{\partial \delta}\right)'\left(\frac{\partial g_{ik}(\delta)}{\partial \delta}\right)}{g_{ik}(\delta)} \cdot \frac{exp(\beta E_{ik})}{\sum_{h=0}^{c} g_{ih}(\delta) exp(\beta E_{ih})} \\
-&\ \frac{\left[\sum_{h=0}^{c} \frac{\partial g_{ih}(\delta)}{\partial \delta} exp(\beta E_{ih})\right]' \left[\sum_{h=0}^{c} \frac{\partial g_{ih}(\delta)}{\partial \delta} exp(\beta E_{ih})\right]}{\left[\sum_{h=0}^{c} g_{ih}(\delta) exp(\beta E_{ih})\right]^2},
\end{aligned}$$

$$\frac{\sum_{k=0}^{c} \left[\sum_{h=0}^{c} (E_{ik} - E_{ih}) g_{ih}(\delta) exp(\beta E_{ih})\right]^2 \cdot g_{ik}(\delta) exp(\beta E_{ik})}{\left[\sum_{h=0}^{c} g_{ih}(\delta) exp(\beta E_{ih})\right]^3}, \text{ and}$$

$$\begin{aligned}
&\sum_{k=0}^{c} \left[\frac{\frac{\partial g_{ik}(\delta)}{\partial \delta}}{g_{ik}(\delta)} - \frac{\sum_{h=0}^{c} \frac{\partial g_{ih}(\delta)}{\partial \delta} exp(\beta E_{ih})}{\sum_{h=0}^{c} g_{ih}(\delta) exp(\beta E_{ih})}\right]' \left[\frac{\sum_{h=0}^{c} (E_{ik} - E_{ih}) g_{ih}(\delta) exp(\beta E_{ih})}{\sum_{h=0}^{c} g_{ih}(\delta) exp(\beta E_{ih})}\right] \\
\times&\ \frac{g_{ik}(\delta) exp(\beta E_{ik})}{\sum_{h=0}^{c} g_{ih}(\delta) exp(\beta E_{ih})},
\end{aligned}$$

respectively, for the $(1,1)^{th}, (1,2)^{th}$ and $(2,2)^{th}$ partition. Note that, after taking expectation over $\mathbb{E}$, these three terms become independent of i. This implies that this $V[U_1(E^{(i)}, M_{i0}; \phi)]$ is constant and given by $\begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$. Using the same line of arguments, it can be proved that $E[U_2^*(A_i, E_i, R_i; \delta)] = 0$ and $V[U_2^*(A_i, E_i, R_i; \delta)]$ is constant and given by W_{11}.

Appendix C: Proof of the asymptotic results

Assume that both U_1 and U_2^* admit all the second derivatives for almost all of their random arguments and for all parameter values in an open neighbourhood of the true parameter ϕ_0 , which are bounded by integrable functions with finite expectation. We first claim that

$$\omega E\left[\frac{\partial}{\partial \phi} U_1(E^{(i)}, M_{i0}; \phi)\right] + (1-\omega) E\left[\frac{\partial}{\partial \phi} U_2(A_i, E_i, R_i; \phi)\right] = \omega \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} + (1-\omega) \begin{bmatrix} W_{11} & 0 \\ 0 & 0 \end{bmatrix} = \mathit{\Sigma} \text{ (say)}, \quad (30)$$

and assume this to be positive definite. Consider

$$\begin{aligned}
E\left[-\frac{\partial}{\partial \phi} U_1(E^{(i)}, M_{i0}; \phi)\right] &= E_{\mathbb{E}} E\left[-\frac{\partial}{\partial \phi} U_1(E^{(i)}, M_{i0}; \phi) \middle| E^{(i)}\right] \\
&= E_{\mathbb{E}} E\left[U_1(E^{(i)}, M_{i0}; \phi) U_1^T(E^{(i)}, M_{i0}; \phi) \middle| E^{(i)}\right] \\
&= \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}. \qquad (31)
\end{aligned}$$

Similarly, one can prove $E\left[-\frac{\partial}{\partial \phi} U_2^*(A_i, E_i, R_i; \phi)\right] = W_{11}$. This completes the proof of the claim.

Again, as both n and m tend to ∞ , $\hat{\phi}_{n+m}$ tends to the true ϕ in probability. The proof of this consistency follows from the arguments of Lehmann and Casella (1998, p475-476) and Hirose (2005, p68). Also, see Bradley and Gart (1962). Now consider

$$
\begin{aligned}
&\frac{1}{\sqrt{n+m}}\left[\sum_{i=1}^{n} U_1(E^{(i)}, M_{i0}; \hat{\phi}_{n+m}) + \sum_{i=1}^{m} U_2(A_i, E_i, R_i; \hat{\phi}_{n+m})\right] \\
= &\left[\sqrt{\frac{n}{n+m}} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_1(E^{(i)}, M_{i0}; \hat{\phi}_{n+m}) + \sqrt{\frac{m}{n+m}} \cdot \frac{1}{\sqrt{m}} \sum_{i=1}^{m} U_2(A_i, E_i, R_i; \hat{\phi}_{n+m})\right] \\
= &\left[\sqrt{\frac{n}{n+m}} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_1(E^{(i)}, \phi_0) + \sqrt{\frac{m}{n+m}} \cdot \frac{1}{\sqrt{m}} \sum_{i=1}^{m} U_2(A_i, E_i, R_i, \phi_0)\right] \\
+ &\left[\frac{n}{n+m} \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \phi} U_1(E^{(i)}, \phi_0) + \frac{m}{n+m} \cdot \frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial \phi} U_2(A_i, E_i, R_i, \phi_0)\right] \sqrt{n+m}(\hat{\phi}_{n+m} - \phi_0) + o_p(1). \quad (32)
\end{aligned}
$$

Using multi-sample weak law of large numbers (Hirose, 2005, p70), we have

$$
\begin{aligned}
&\left[\frac{n}{n+m} \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \phi} U_1(E^{(i)}, \phi_0) + \frac{m}{n+m} \cdot \frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial \phi} U_2(A_i, E_i, R_i, \phi_0)\right] \xrightarrow{P} \\
&\qquad \left[\omega_1 E\left\{\frac{\partial}{\partial \phi} U_1(E^{(i)}, \phi_0)\right\} + \omega_2 E\left\{\frac{\partial}{\partial \phi} U_2(A_i, E_i, R_i, \phi_0)\right\}\right]. \quad (33)
\end{aligned}
$$

Using (32) and (33), we have

$$
\sqrt{n+m}(\hat{\phi}_{n+m} - \phi_0) = -\Sigma^{-1}\left[\sqrt{\frac{n}{n+m}} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_1(E^{(i)}, \phi_0) + \sqrt{\frac{m}{n+m}} \cdot \frac{1}{\sqrt{m}} \sum_{i=1}^{m} U_2(A_i, E_i, R_i, \phi_0)\right] + o_p(1). \quad (34)
$$

Now, by multi-sample CLT (Hirose, 2005, p70), as both n and $m \to \infty$,

$$
\left[\sqrt{\frac{n}{n+m}} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_1(E^{(i)}, \phi_0) + \sqrt{\frac{m}{n+m}} \cdot \frac{1}{\sqrt{m}} \sum_{i=1}^{m} U_2(A_i, E_i, R_i, \phi_0)\right] \xrightarrow{a} N(0, \Sigma). \quad (35)
$$

Then, from (34) and (35), we have $\sqrt{n+m}(\hat{\phi}_{n+m} - \phi_0) \xrightarrow{a} N(0, \Sigma^{-1})$. See Lin and Paik (2001) for a similar result.

Using the weak law of large numbers, we have

$$
\begin{aligned}
&- \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \phi} U_1(E^{(i)}, M_{i0}; \phi) \xrightarrow{P} E\left[\frac{\partial}{\partial \phi} U_1(E^{(i)}, M_{i0}; \phi)\right] = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \\
\text{and} \quad &- \frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial \phi} U_2(A_i, E_i, R_i; \delta) \xrightarrow{P} E\left[\frac{\partial}{\partial \phi} U_2(A_i, E_i, R_i; \delta)\right] = \begin{bmatrix} W_{11} & 0 \\ 0 & 0 \end{bmatrix},
\end{aligned}
$$

so that the two observed information matrices, evaluated at $\phi = \hat{\phi}_{n+m}$, can be used for estimating the asymptotic variance-covariance matrix of $\hat{\phi}_{n+m}$. In particular, the asymptotic variance of $\hat{\beta}_{n+m}$ can be estimated by the lower right element of the matrix

$$
-\left[\sum_{i=1}^{n} \frac{\partial}{\partial \phi} U_1(E^{(i)}, M_{i0}; \phi) + \sum_{i=1}^{m} \frac{\partial}{\partial \phi} U_2(A_i, E_i, R_i; \delta)\right]^{-1},
$$

evaluated at $\phi = \hat{\phi}_{n+m}$.