Effect of reporting bias in the analysis of spontaneous reporting data

Technical Report No. ASD/2012/5
Dated: 10 May 2012

Palash Ghosh

And

Anup Dewanji

Applied Statistics Unit
Indian Statistical Institute
Kolkata 700108

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

Effect of reporting bias in the analysis of spontaneous reporting data

Palash Ghosh and Anup Dewanji
Applied Statistics Unit, Indian Statistical Institute, Kolkata 700108
emails: *palash_r@isical.ac.in*, *dewanjia@isical.ac.in*

April 20, 2012

Abstract

It is well-known that a spontaneous reporting system suffers from significant under-reporting of adverse drug reaction (ADR) from the source population. The existing methods do not adjust for such under-reporting for the calculation of measures of association between a drug and the ADR under study. Often there may be direct or indirect information on the reporting probabilities. This work incorporates the reporting probabilities into existing methodologies, specifically to BCPNN (Bate et al., 1998) and EBGM (Dumouchel, 1999) methods, and shows how the two methods lead to biased results in the presence of reporting bias. Considering all the cases to be reported, the association measure for the source population can be estimated by using only exposure information through a reference sample from the source population.

keywords: Adverse drug reaction, BCPNN, EBGM, Reference sample, Spontaneous Reporting database, Under-reporting.

1 Introduction

In pharmacovigilance, spontaneous reporting (SR) database plays an important role to generate signal regarding the relationship between drug and adverse drug reaction (ADR) based on reported data. These reports are collected from all over the world with the help of clinicians and/or health professionals, who are responsible for recognizing and reporting suspected side effects known as adverse drug reaction (ADR), once the drug is in the market. All ADR reports are stored in database at the national centers and sent to WHO Collaborating Center for International Drug Monitoring (the Uppasala Monitoring Center) as well. The U.S. Food and Drug Administration (FDA) also maintains such SR database (Adverse Event Reporting System) consisting of medical events happening to patients taking different kinds of drugs and so does the UK Yellow Card database [5]. This kind of database are used to provide early warnings or suspicions, which have not been recognized prior to marketing of a drug because of limitations of clinical trials.

It is well-known that a spontaneous reporting database suffers from significant under-reporting of adverse drug reaction (ADR) from the source population. There are several reasons for under-reporting of ADR often depending on the severity of the ADR, the drug associated with the ADR and the disease for which the drug has been prescribed, etc. Presently, the commonly used methodologies to detect signal from SR database are the Bayesian confidence propagation neural network (BCPNN) [2], or DuMouchel's [4] empirical Bayesian (EBGM) methodology, both of which are based solely on reports in SR database. Therefore, these methodologies suffer from reporting bias, as also recognised in these works, and fail to infer about the true drug-ADR association in the source

population. Even if there is some information on the reporting pattern of ADR, these methodologies do not adjust for that. This work is intended to illustrate the effect of ignoring reporting bias. The reporting probabilities are incorporated into the measures used in these methodologies to modify them into measures for the source population. Therefore, while the measures in BCPNN [2] and EBGM [4] relate to the SR database, the modified measures relate to the source population. The reporting probabilities help to find a relationship between these two kinds of measures, as discussed in Section 2. We also show that BCPNN and EBGM are based on the same basic idea of comparison of joint probability of drug and ADR with the corresponding marginal probabilities. In Section 3, we assume that all the cases are reported and obtain the source population association measure by using information on exposure probability. Section 4 ends with some concluding remarks.

2 Reporting bias in BCPNN and EBGM

In this section, we incorporate the reporting probabilities in the expressions of association measures used in BCPNN and EBGM, showing the effect of under-reporting, or the extent of reporting bias. These measures are based on reporting data in the form a 2×2 table as given in Table 1.

Table 1: Summary of the data obtained from SR database

	$A = 1, R = 1$	$A = 0, R = 1$	Total
$D = 1$	n_{11}	n_{10}	$n_{1\cdot}$
$D = 0$	n_{01}	n_{00}	$n_{1\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 0}$	n

Here, A denotes the presence ($A = 1$) or absence ($A = 0$) of the ADR under study and D denotes the use of the specific drug ($D = 1$ means using the drug and $D = 0$ means not using the same). Also, $R = 0, 1$ denotes the reporting status, where $R = 1$ represents the event of reporting to the SR database. In this context, a commonly used association measure, known as odds-ratio (OR), is worth mentioning. As discussed in Ghosh and Dewanji (2011), the observed odds-ratio $n_{11}n_{00}/n_{01}n_{10}$ from the reporting data in Table 1 is a biased estimate for the OR due to reporting bias. This, however, estimates a quantity called reporting odds-ratio (ROR) as given by

$$ROR = \frac{P(D = 1, A = 1|R = 1)P(D = 0, A = 0|R = 1)}{P(D = 0, A = 1|R = 1)P(D = 1, A = 0, |R = 1)}. \tag{1}$$

This ROR and OR are related through the reporting probabilities as

$$OR = ROR \times \frac{P(R = 1|D = 0, A = 1)P(R = 1|D = 1, A = 0)}{P(R = 1|D = 1, A = 1)P(R = 1|D = 0, A = 0)}. \tag{2}$$

The association measures used in BCPNN and EBGM are similarly affected by the reporting probabilities as derived in Section 2.1 and 2.2, respectively. Note that, as discussed in [5], the population here means the set of patients suffering from a particular disease for which the drug under study is prescribed along with some other drugs to be compared.

2.1 BCPNN

Bate et al. [2] have introduced the Bayesian confidence propagation neural network (BCPNN) methodology to generate signal from the SR database. Since 1998, WHO has implemented BCPNN methodology for routine signal detection [1]. The BCPNN method is based on an information component (IC), a measure of association between the two binary variable A and D defined as

$$IC = \log_2 \frac{P(D,A)}{P(D)P(A)}. \tag{3}$$

Note that this IC reflects the observed/expected ratio in logarithm scale (base 2), where the expected quantity is under the null hypothesis of no association. This IC is estimated from the reporting data in Table 1 using a Bayes method. Assuming beta priors for the three probabilities, these are estimated by the corresponding posterior means; hence, IC is estimated by using (3) even for small number of reports in Table 1. As more reports accumulate, the posterior distributions are also updated and the corresponding estimate of IC along with its variance estimate are also improved. A signal is said to be detected if the lower limit of 95% confidence interval of IC is greater than zero.

The IC, as defined, is treated as the population IC. Note that the IC based on SR data only may be defined as

$$IC^{SR} = log_2 \frac{P(D,A|R=1)}{P(D|R=1)P(A|R=1)}. \tag{4}$$

The BCPNN method, in fact, works on this IC^{SR}, where as the population IC is defined differently. The reporting data in Table 1 estimates IC^{SR} as $log_2[n_{11}n/n_{.1}n_{1.}]$. Note that

$$\frac{P(D,A|R=1)}{P(D|R=1)P(A|R=1)} = \frac{P(R=1|D,A)P(R=1)}{P(R=1|D)P(R=1|A)} \cdot \frac{P(D,A)}{P(D)P(A)}, \tag{5}$$

resulting in the following relationship between the two information components:

$$IC = IC^{SR} - log_2 \left\{ \frac{P(R=1|D,A)P(R=1)}{P(R=1|D)P(R=1|A)} \right\}. \tag{6}$$

The objective in this work is to investigate when the IC^{SR} can be used as a representative of the source population IC. Clearly, if reporting status (R) does not depend on any one or both of the ADR status (A) and drug status (D), then the second term in the RHS of (6) is zero, resulting in the equality of the two information components. In practice, this kind of equality rarely holds, as the reporting pattern depends on various factors. There may be some other factors besides ADR-drug status affecting the event of reporting to the SR database [8], but in this work, we assume the influence of those factors to be ignorable.

To illustrate how the reporting probabilities can affect the IC^{SR} to deviate from IC, we consider the data (See Ghosh and Dewanji, 2011) in Table 2. Here, the drug prograph and the ADR liver transplant rejection (LTR), are compared with respect to those patients with liver transplantation in the SR database of US FDA for the time span 2006-2008. The OADR in Table 2 means other ADRs. From Table 2, the estimated reporting odds ratio ($\widehat{ROR}$) is 0.98 and corresponding estimates of IC^{SR}, denoted by $\widehat{IC}^{SR}$ is -0.01. Both the estimates indicate no association between the drug

Table 2: Cross-classified data (2006-08) for Prograph-LTR combination for patients with liver transplantation from the SR database of US FDA

	LTR	OADR	$\widehat{ROR}$	$\widehat{IC}^{SR}$
Prograph	27	152	0.98	-0.01
Other drugs	53	293		

Table 3: Values of IC for different θ for the liver transplant patients of Table 2

θ	IC	θ	IC	θ	IC
0.1	3.31	0.9	0.14	1.7	-0.78
0.2	2.31	1.0	-0.01	1.8	-0.86
0.3	1.73	1.1	-0.15	1.9	-0.94
0.4	1.31	1.2	-0.27	2.0	-1.01
0.5	0.99	1.3	-0.39	2.1	-1.08
0.6	0.73	1.4	-0.50	2.2	-1.15
0.7	0.50	1.5	-0.59	2.3	-1.21
0.8	0.31	1.6	-0.69	2.4	-1.26

prograph and the ADR liver transplant rejection based on the SR data. If there is no reporting anomaly in SR data, we could take the information component measure as the representative of corresponding source population measure. In practice, this may not be the case. Considering different reporting probabilities to see the deviation of the IC from the IC^{SR}, as in Table 3, it is evident that the corresponding source population IC can go either way depending on the quantity

$$\theta = \frac{P(R=1|D=1,A=1)P(R=1)}{P(R=1|D=1)P(R=1|A=1)}. \tag{7}$$

For example, $IC = 3.31$ when $\theta = 0.1$, whereas $IC = -1.26$ when $\theta = 2.4$. Therefore, any decision that has been taken based on IC^{SR}, may turn out to be wrong depending on the reporting probabilities. In other words, there may be false positive or false negative decision based on only SR data.

2.2 EBGM

DuMouchel [4] introduced the signal detection methodology by an empirical Bayes procedure. This methodology assumes that each observed count n_{ij}, corresponding to $D = i$ and $A = j$ (See Table 1), for $i, j = 0, 1$, is a draw from a Poisson distribution with an unknown mean μ_{ij} and interest centers on the ratio $\lambda_{ij} = \mu_{ij}/E_{ij}$, where E_{ij} is the expected count under the null hypothesis of no association between the drug and the ADR. It also assumes that each λ is drawn from a common prior distribution, a mixture of two gamma distributions, instead of treating different λs' as unrelated. The ratio may be interpreted in similar fashion like relative risk with value 1 under the null hypothesis. This method provides Bayesian estimates of the λ_{ij}'s and, in particular, an empirical Bayes measure given by $EBGM_{ij} = 2^{EBlog2_{ij}}$ with $EBlog2_{ij} = E[log_2(\lambda_{ij})|n_{ij}]$. Without getting into the details of this Bayesian methodology, we incorporate the reporting probabilities into the quantities λ_{ij}'s of interest to investigate how these are affected. It can be easily seen that, for $i, j = 0, 1$, and N being the size of the source population,

$$\lambda_{ij} = \frac{\mu_{ij}}{E_{ij}} = \frac{NP(D=i,A=j)}{NP(D=i)P(A=j)} = \frac{P(D=i,A=j)}{P(D=i)P(A=j)}, \tag{8}$$

with the corresponding quantity based on the SR data only written as

$$\lambda_{ij}^{SR} = \frac{P(D=i, A=j|R=1)}{P(D=i|R=1)P(A=j|R=1)}. \tag{9}$$

Similar probability calculation, as in Section 2.1, shows that

$$\lambda_{ij}^{SR} = \lambda_{ij} \times \frac{P(R=1|D=i, A=j)P(R=1)}{P(R=1|D=i)P(R=1|A=j)}. \tag{10}$$

As before, the measure λ_{ij}^{SR} coincides with the corresponding source population measure λ_{ij}, when the reporting status (R) does not depend on any one or both of the ADR status (A) and drug status (D).

From the definition of the IC in Section 2.1 and the quantity λ_{ij}'s in Section 2.2, it is clear that $IC = log_2\lambda_{11}$ and $IC^{SR} = log_2\lambda_{11}^{SR}$. In other words, these two measures are the function of ratio of joint probability and product of marginal probabilities of the ADR and the drug under concern. So, these two methodologies are based on the same principle and we can get one measure from the other.

3 An estimate based on SR database

We have seen in Section 2.1, four different reporting probabilities are required to obtain the source population IC from IC^{SR} using (6). In practice, it is not always easy to get estimate of these reporting probabilities. In this section, we make some assumption on these probabilities and use some external information to estimate IC.

Here, we assume that all the ADRs under study (called 'cases') are reported to the SR database [5] (i.e, 100% reporting of cases with $P[R=1|A=1]=1$). Though, in reality, under-reporting of ADR is a severe problem, steps are being taken to improve the current scenario. For example, the prescription event monitoring (PEM) used by the Drug Safety Research Unit (http://www.dsru.org/) in Southampton, UK, is an observational cohort technique which collects data on all prescriptions for the first 20,000 to 50,000 patients for a new drug leading to 100% reporting of ADR for some drug-ADR combinations in PEM data. Also, this assumption is realistic in case of serious ADR as health professionals are well-informed about the possible adverse effects of the drug. This assumption implies that $P(D=1|A=1, R=1) = P(D=1|A=1)$; in other words, the exposure probability for the cases in SR data is same as that in the source population. Then

$$\begin{aligned}
\frac{P(R=1|D=1, A=1)P(R=1)}{P(R=1|D=1)P(R=1|A=1)} &= \frac{P(R=1|D=1, A=1)P(R=1)}{P(R=1|D=1)} \\
&= \frac{P(R=1, D=1, A=1)}{P(D=1, A=1)} \cdot \frac{P(R=1)P(D=1)}{P(R=1, D=1)} \\
&= \frac{P(D=1|A=1, R=1)P(R=1|A=1)}{P(D=1|A=1)} \cdot \frac{P(D=1)}{P(D=1|R=1)} \\
&= \frac{P(D=1|A=1, R=1)P(D=1)}{P(D=1|A=1)P(D=1|R=1)} \\
&= \frac{P(D=1)}{P(D=1|R=1)}.
\end{aligned} \tag{11}$$

When $P(D=1)$ is known, we can obtain an estimate of the ratio in (11) using the information from the SR data. When exposure probability is unknown, we consider a reference sample of size m from

the source population and observe the exposure status to estimate $P(D=1)$. A reference sample can be drawn from some external sources other than SR database, for example, a prescription database. In UK, virtually all persons are registered with a general practitioner (GP) who provides primary health care and issues prescriptions for the medicines considered medically necessary. The patient takes the prescription to a pharmacist who dispenses the medication and then sends the information to a central Prescription Pricing Authority (PPA) which arranges the reimbursement of the pharmacist [6]. This PPA database, screened for a particular disease, gives the corresponding source population, a random sample from which can be drawn. Exposure information from this sample can be easily obtained. Now, from (6) and (11), we have

$$\begin{aligned} IC &= IC^{SR} - log_2\left\{\frac{P(D=1)}{P(D=1|R=1)}\right\} \\ &= log_2\frac{P(D=1, A=1|R=1)}{P(D=1)P(A=1|R=1)}, \end{aligned} \tag{12}$$

which can be estimated by $\widehat{IC} = log_2[n_{11}m/n_{\cdot 1}m_{1\cdot}]$, where $m_{1\cdot}$ is the number of individuals using the drug under study in the reference sample of size m. Varince of this estimate $\widehat{IC}$ can be obtained using delta method as described in the Appendix.

Table 4: Cross classified data for drug diuretic and ADR CHF from NPCL

Diuretic	Congestive Heart Failure (CHF)	
	Present	Absent
Present	78	1697
Absent	227	7820

The data in Table 4, has been taken from Netherlands Pharmacovigilance Centre Lareb (NPCL) SR database (See van der Heijden et al. 2002) consisting of 9822 reports from health professionals in Netherlands, concerning patients older than 50 years between 1^{st} January 1990 to 1^{st} January 1999. Here, the objective is to find whether the drug diuretic is responsible for the ADR congestive heart failure (CHF). To illustrate the methodology described above, assume that all the drugs were used for cardiovascular disease and stroke, since we have no information about the disease for which the drugs (diuretic and other drugs) have been taken. We also assume 100% reporting of cases to the NPCL SR database.

The probability $P(D=1|R=1)$ can be estimated from Table 4 as $(78+1697)/9822 = 0.18$. However, the exposure probability $P(D=1)$ cannot be estimated from the SR data, since the control sample is suffering from reporting bias. The exposure probability is obtained from the study of Pharmaceutical Use and Expenditure for Cardiovascular Disease and Stroke (PUECDS) [3]. This study reported the average percent of diuretic use (ATC code beginning with C03) from hospital outlets in Netherlands for the time period 1989 to 1999 as 14.8 ([3], p21). In order to estimate the population IC, we use this information to obtain the probability of exposure to diuretic in the source population as 0.148. Since the concerned periods for both NPCL SR-database and PUECDS [3] study are about the same, it is assumed that the information provided in PUECDS study and the events captured in the NPCL database both somewhat represent the source population of interest. Using (12), the estimated source population IC is 0.79, whereas the corresponding estimate of IC^{SR} is 0.5. The standard error of the estimated IC is 0.22, calculated using delta method as

shown in Appendix considering the exposure probability as fixed quantity. This indicates a positive association between the drug and CHF (See [5]).

4 Discussion

The primary purpose of SR database is to generate hypothesis regarding the relationship between the drugs and the ADRs, and not so much about establishing a causal relationship between them [1]. But, the ultimate objective of the pharmacovigilance is to detect whether a particular drug is responsible for a particular ADR. In other words, after detecting a signal from SR database, we have to consider a formal epidemiological study to come up with a decision regarding the association. The main problem of SR-data, which prevents us to get a stronger association measure, is under-reporting of ADRs. This makes all the association measures based on SR data as only 'reporting measures', not as a measure of association in the corresponding source population. This work, intends to bridge this gap. We have shown how source population association measures can be obtained from the measures based on SR data using different reporting probabilities. These reporting probabilities are not easily available in practice. Nevertheless, steps may be taken to gather information on the reporting probabilities from some external sources. In this regard, we have discussed the importance of PPA in Section 3. The example shown at the end of Section 3 indicates how the methodology can be useful in the existing structure. This work also hints at the necessity of linking different databases to make the reference samples readily available. More research in this direction is needed to have a greater understanding of the usefulness of the methodology.

For the the sake of simplicity, we have not gone into the details of the Bayesian approaches in both BCPNN and EBGM. The objective has been to derive the relationship between the source population measures and those in SR database. As a result, when we have information on the reporting probabilities, that can be used to get the corresponding source population measures from those in the SR database. The aim has been to assess the measures in BCPNN and EBGM in view of reporting error and in the direction of detecting the drug-ADR relationship as early as possible keeping in mind the issue of public health and the related cost for delaying the detection.

References

[1] Bate, A., and Evans, S. J. W. Quantitative signal detection using spontaneous adr reporting. *Pharmacoepidemiology and Drug Safety 18* (2009), 427–436.

[2] Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., and Freitas, R. M. D. A bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology 54* (1998), 315–321.

[3] Dickson, M., and Jacobzone, S. Pharmaceutical use and expenditure for cardiovascular disease and stroke: A study of 12 OECD countries. *OECD Health Working Papers, No. 1* (2003).

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

[5] Ghosh, P., and Dewanji, A. Analysis of spontaneous adverse drug reaction (ADR) reports using supplementary information. *Statistics in Medicine 30*, 16 (2011), 2040–2055.

[6] Mann, R. D. Prescription-event monitoring-recent progress and future horizons. *Br J Clin Pharmacol 46* (1998), 195–201.

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

[8] van Puijenbroek, E. P., Bate, A., Leufkens, H. G. M., Lindquist, M., Orre, R., and Egberts, A. C. G. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and Drug Safety 11* (2002), 3–10.

Appendix

Variance of the estimate $\widehat{IC}$ obtained from (12) can be calculated by applying delta method and using the estimated covariance matrix of $(n_{11}, n_{01}, n_{\cdot 0}, m_{1\cdot})$, with $n_{\cdot 0} = n_{10} + n_{00}$, where $n_{ij}, i, j = 0, 1$ are the cell frequencies as in Table 1, and $m_{1\cdot}$ is the number of exposed individuals in the reference sample. In order to define the covariance matrix of $(n_{11}, n_{01}, n_{\cdot 0}, m_{1\cdot})$, note that $m_{1\cdot}$ and $(n_{11}, n_{01}, n_{\cdot 0})$ are independent with

$$\begin{aligned} m_{1\cdot} &\sim Bin(m, P(D=1)) \\ (n_{11}, n_{01}, n_{\cdot 0}) &\sim Multinomial\Big(n, P[D=1, A=1|R=1], P[D=0, A=1|R=1], P[A=0|R=1]\Big). \end{aligned}$$

Variance of $\widehat{IC}$ is given by

$$(g')^T V g', \tag{13}$$

where

$$V = \begin{bmatrix} \frac{p_{11}(1-p_{11})}{n} & -\frac{p_{11}p_{01}}{n} & -\frac{p_{11}p_{\cdot 0}}{n} & 0 \\ -\frac{p_{11}p_{01}}{n} & \frac{p_{01}(1-p_{01})}{n} & -\frac{p_{01}p_{\cdot 0}}{n} & 0 \\ -\frac{p_{11}p_{\cdot 0}}{n} & -\frac{p_{01}p_{\cdot 0}}{n} & \frac{p_{\cdot 0}(1-p_{\cdot 0})}{n} & 0 \\ 0 & 0 & 0 & \frac{p_e(1-p_e)}{n} \end{bmatrix}. \tag{14}$$

Here $p_{11} = P[D=1, A=1|R=1], p_{01} = P[D=0, A=1|R=1], p_{\cdot 0} = P[A=0|R=1]$ and $p_e = P(D=1)$ with their estimates being $n_{11}/n, n_{01}/n, n_{\cdot 0}/n$ and $m_{1\cdot}/m$, respectively. The function g is defined as

$$g = log_2\left[\frac{p_{11}}{(p_{11}+p_{01})p_e}\right], \tag{15}$$

and g' denotes the vector of partial derivatives of g with respect to $(p_{11}, p_{01}, p_{\cdot 0}, p_e)$. The estimate of the variance can be obtained by replacing $(p_{11}, p_{01}, p_{\cdot 0}, p_e)$ in (13) by the corresponding estimates. For the example in Section 3, the variance is estimated by assuming p_e to be known as 0.148. In this case, only the multinomial distribution of $(n_{11}, n_{01}, n_{\cdot 0})$ is used with $(p_{11}, p_{01}, p_{\cdot 0})$ being the parameter vector and the first 3×3 submatrix of V in the calculation of (13).