21. Ueno S, Tanabe G, Sako K, Hiwaki T, Hokotate H, Fukukura Y, et al. Discrimination value of the new western prognostic system (CLIP score) for hepatocellular carcinoma in 662 Japanese patients. Cancer of the Liver Italian Program. Hepatology 2001;34: 529–34.
22. Marrero JA, Su GL, Wei W, Emick D, Conjeevaram HS, Fontana RJ, et al. Des-gamma carboxyprothrombin can differentiate hepatocellular carcinoma from nonmalignant chronic liver disease in American patients. Hepatology 2003;37:1114–21.
23. Yamashita F, Tanaka M, Satomura S, Tanikawa K. Prognostic significance of *Lens culinaris* agglutinin A-reactive alpha-fetoprotein in small hepatocellular carcinomas. Gastroenterology 1996;111:996–1001.

Japanese Journal of Biometrics Vol. 27, No. 1, 17–33 (2006)

Original Article

# Analysis of Quality of Life Data with Death and Drop-out in Advanced Non-Small-Cell Lung Cancer Patients

Kazutaka Doi, Yutaka Matsuyama and Yasuo Ohashi

Department of Biostatistics/Epidemiology and Preventive Health Sciences,
School of Health Sciences and Nursing, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
e-mail:doi@epistat.m.u-tokyo.ac.jp

In measuring quality of life (QOL), outcome-dependent missing values are inevitable because of longitudinal nature of the study. In particular, in clinical trials of advanced-stage disease, it is desirable to distinguish differences between reasons for missing, death and drop-out, because QOL scores for death cases are not really missing data, but are nonexistent and are simply undefined. We focus on estimating the local average treatment effect among survivors. Standard randomized treatment comparisons cannot be performed because the QOL scores are only defined in the non-randomly selected subgroup of survivors. We propose a new estimation method of the survivor average causal effect (SACE) in the presence of both death and drop-out. The proposed estimator is a weighted average of the standard estimators for survivors where the weight is the probability that the patient would have survived had he/she received the other treatment. For drop-out cases, the multiple imputation method is applied. Two analysis methods (proposed method and analysis based on only observed survivors) were compared by simulation studies. The proposed estimator had smaller biases with smaller MSEs compared with those of the standard estimator. The proposed method was applied to data from a randomized phase III clinical trial for advanced non-small-cell lung cancer patients.

*Key words:* Local average treatment effect, Missing values, Multiple imputation, Potential outcomes, Principal stratification, QOL.

## 1. Introduction

Lung cancer is one of the most common cancers in terms of incidence and mortality in Japan. Non-small-cell lung cancer (NSCLC) accounts for approximately 80% of all lung cancers (Japanese Society of Medical Oncology, 1999). Patients with NSCLC are usually in serious condition and are not expected to be complete cure. Hence, to evaluate therapeutic effectiveness of new treatments for NSCLC, quality of life (QOL) is a matter of great concern and is usually

Received February 2006. Revised April 2006. Accepted May 2006.

investigated as one of major clinical endpoints in addition to traditional variables such as tumor response, survival and toxicity in clinical trials.

In most QOL investigations, QOL data are measured longitudinally, because we are interested in how a disease or an intervention affects an individual's well-being over time. Incompleteness of the data, in particular, loss to follow-up, is a common problem of longitudinal studies (Fairclough, 2002). Death and drop-out are two main reasons why a patient may be lost to follow-up. For example, consider a randomized clinical trial for advanced NSCLC patients and evaluations of QOL at six months after randomization. The patients in the study are often in deteriorated condition, and some will die or drop-out before six months, with the result that the QOL outcome is "missing". Such drop-out is usually attributable to deteriorating health condition (Pauler et al, 2003). Patients are often unable to complete a series of QOL assessments because of deteriorating health and there exist studies where proportion of missing QOL data is approximately 40 to 50 percent (Bernard et al., 1998; Brown et al., 2005; Kelly et al., 2001). A large proportion of missing data not only results in wider confidence interval and reduced power, but causes a severe bias (Donaldson and Moinpour, 2005). Such outcome-dependent missing data cause a statistical difficulty in analyzing QOL data.

Standard analytical approaches for missing data such as multiple imputation method (Rubin, 1978; Schafer, 1999) or inverse weighting method (Robins et al., 1995) attempt to estimate the treatment effect that would have been observed if (contrary to fact) all patients had continued to be observed until the end of the study, which is the global average treatment effect in the entire study population. In such an approach, QOL data is treated as "missing" among patients who have died, and a causal estimand can be defined based on the contrasts of QOL distributions for the standard and new treatment groups. The goal of assessing such an estimand is to compare QOL distribution between the randomized groups had all patients survived during the study. This analysis will be reasonable if the post-treatment variable, that is, survival status is controllable. However, missing QOL data for censored cases due to death cannot exist (Rubin, 2000). Therefore, it is reasonable to restrict the estimation of the causal effects of treatment on QOL to the patients in the living status, which can be affected by the treatment and is a posttreatment variable.

In this paper, we consider the comparison of repeated measures of QOL scores between two groups, standard and new treatment group. Our causal estimand is the local average treatment effect, which is the average effect of treatment among survivors. A standard analysis for estimating the local average treatment effect is to compare the QOL scores among observed survivors. However, such a comparison is problematic if the treatment has any effect on survival, because the subgroup of observed survivors under the standard treatment and that of observed survivors under the new treatment, are not comparable. This problem is known to epidemiologists as a posttreatment selection bias, which implies that a comparison of QOL among observed survivors

does not have a causal interpretation (Rosenbaum, 1984; Robins and Greenland, 1992).

Recently, Frangakis and Rubin (2002) have proposed a principal stratification framework with respect to posttreatment variables. The principal stratification approach is a cross-classification of patients into several potential subgroups, called principal strata, defined by the combinations of all possible values of the posttreatment variable (i.e. death or survival) under each of the treatments. As with two groups comparison studies, patients can be classified into four potential subgroups:

1. the true survivors: those who would survive under either treatment assignment
2. the new treatment only survivors: those who would not survive under the standard but would survive under the new treatment
3. the standard treatment only survivors: those who would survive under the standard but would not survive under the new treatment
4. the non-survivors: those who would not survive under either treatment assignment.

The local average treatment effect is defined as a comparison of outcomes of primary interest within a "principal stratum", which is composed of "the true survivors". In a comparison of QOL, the difference of QOL scores among the true survivors is a matter of concern, because the potential values for QOL scores under two treatments are defined only for this subgroup. The key property of the principal stratification is that it is based on the stratification by the baseline potential characteristics of each patient and is not affected by treatment. Rubin (1998) has called this causal parameter the survivor average causal effect (SACE). However, it is important to note that we cannot directly observe the principal stratum to which a patient belongs, because the indicator of whether a patient would have survived under the other treatment is a counterfactual variable.

Matsuyama and Morita (2006) have proposed an estimation method for the SACE. Their approach is based on the prediction of the survival probability in each treatment group as a function of covariates and the estimation of the treatment difference among patients who would have survived to either treatment. The purpose of this paper is to extend their estimation method to the settings where missing outcomes due to both death and drop-out exist. The next section describes a motivating study, a randomized phase III clinical trial for advanced NSCLC patients to evaluate whether QOL score is different under two different treatments (Kubota et al., 2004). In Section 3, we describe the estimation method for the SACE (Matsuyama and Morita, 2006) and extend their approach. In Section 4, the performances of the proposed method are evaluated through a simulation study. In Section 5, we present the results for the simulations and the application to NSCLC data. In Section 6, we conclude with a discussion.

## 2. NSCLC Randomized Clinical Trial Data

We briefly describe the motivating clinical trial and the data. Full details on the design, conduct, and main clinical results have been reported previously (Kubota et al., 2004). This clinical trial was a two-arm, multi-center, randomized controlled trial of chemotherapies for NSCLC patients: Docetaxel plus Cisplatin (Group A) and Vindesine plus Cisplatin (Group B). The patients had to meet the following inclusion criteria: pathologically diagnosed NSCLC; distant metastases (stage IV); previously untreated; age over 19 and under 75 years; performance status (PS) of 0, 1, 2 (patients with PS of 3 because of pain from bone metastases were admitted to the study); life expectancy greater than three months. Table 1 shows the selected baseline characteristics of patients.

**Table 1.** Baseline characteristics of patients in NSCLC randomized clinical trial.

| Characteristics | | Treatment Group | |
|---|---|---|---|
| | | Group A (Docetaxel plus Cisplatin) | Group B (Vindesine plus Cisplatin) |
| Number of Patients | | 119 | 121 |
| Age (mean (SD)) | | 60.6 (9.6) | 61.4 (9.2) |
| Male/Female | | 77(65%)/42(35%) | 78(64%)/43(36%) |
| Performance Status | 0 | 36(30%) | 34(28%) |
| | 1 | 78(66%) | 86(71%) |
| | 2 | 4(3%) | 1(1%) |
| | 3 | 1(1%) | 0(0%) |
| Baseline QOL Score (mean (SD)) | Functional | 78.5(24.6) | 79.8(24.2) |
| | Physical | 75.6(19.9) | 78.2(16.2) |
| | Mental | 59.6(21.2) | 63.1(19.8) |
| | Psychosocial well-being | 45.2(21.5) | 48.3(19.1) |

In this trial, QOL Questionnaire for Cancer Patients Treated with Anticancer Drugs (QOL-ACD; Kurihara et al., 1999) was measured at baseline, just before each course of chemotherapy, one and three months after the termination of the chemotherapy, and every three months after these measurements. The QOL-ACD is a 22-items, self-reported questionnaire, which consists of four domains (functional, physical, mental, and psychosocial well-being) and a global score evaluated by a 5-point face scale. Here, we only consider the comparisons of QOL mean scores in four domains between two treatment groups. Each domain is composed of five or six questions, in which each question was evaluated by a 5-point scale. The reliability and validity of the questionnaire for advanced NSCLC patients have been evaluated (Matsumoto et al., 2002).

In the analysis of QOL-ACD, each item score is summed up into domain scores, which range from 0 (worst condition) to 100 (best condition). We selected three measurement times: baseline (visit 0), just before the 2nd course of chemotherapy (visit 1) as during the course of chemotherapy, and six months after the termination of chemotherapy (visit 2) as after the course

**Table 2.** The distribution of the observed patterns of QOL scores in each domain.

| | Visit# | | | Domain | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Functional | | Physical | | Mental | | Psychosocial well-being | |
| | 0 | 1 | 2 | Group A | Group B | Group A | Group B | Group A | Group B | Group A | Group B |
| Pattern 1 | *O* | *O* | *O* | 50 (42.0%) | 50 (41.3%) | 51 (42.9%) | 50 (41.3%) | 51 (42.9%) | 49 (40.5%) | 51 (42.9%) | 49 (40.5%) |
| Pattern 2 | *O* | *O* | *M* | 38 (31.9%) | 47 (38.8%) | 38 (31.9%) | 47 (38.8%) | 38 (31.9%) | 46 (38.0%) | 38 (31.9%) | 46 (38.0%) |
| Pattern 3 | *O* | *M* | *M* | 12 (10.1%) | 7 (5.8%) | 12 (10.2%) | 7 (5.8%) | 12 (10.2%) | 8 (6.6%) | 12 (10.2%) | 8 (6.6%) |
| Pattern 4 | *O* | *M* | *O* | 9 (7.6%) | 2 (1.7%) | 8 (6.7%) | 2 (1.7%) | 8 (6.7%) | 3 (2.5%) | 8 (6.7%) | 3 (2.5%) |
| Pattern 5 | *O* | *O* | *D* | 9 (7.6%) | 11 (9.1%) | 9 (7.6%) | 11 (9.1%) | 9 (7.6%) | 11 (9.1%) | 9 (7.6%) | 11 (9.1%) |
| Pattern 6 | *O* | *M* | *D* | 1 (0.8%) | 4 (3.3%) | 1 (0.8%) | 4 (3.3%) | 1 (0.8%) | 4 (3.3%) | 1 (0.8%) | 4 (3.3%) |
| Total | | | | 119 (100%) | 121 (100%) | 119 (100%) | 121 (100%) | 119 (100%) | 121 (100%) | 119 (100%) | 121 (100%) |

#: O means the observed data, M means the missing data and D means the death.

of chemotherapy. In each measurement time, four domain scores per one subject were obtained. Table 2 shows the distribution of the observed patterns of QOL scores in each domain. Pattern 1 is the complete observed case, Pattern 2 and 3 are the drop-outs cases, Pattern 4 is the intermittent missing case, and Pattern 5 and 6 are the death cases during follow-up. In this paper, we refer to pattern 2, 3 and 4 as the drop-out cases without their distinctions. There were a little difference in observed missing patterns among four domains. In both groups, death cases occurred only at visit 3. The larger death proportion was observed in Group B (15/121) compared with Group A (10/119). The risk difference was $8.4\% - 12.4\% = -4.00\%$ (95%CI: $-11.7\%$, $3.7\%$).

## 3. Methods

### 3.1 Estimation of Survivor Average Causal Effect (SACE)

In this subsection, for simplicity, we consider the situation of QOL measurement at only one time point after randomization, although QOL data are usually measured repeatedly. Consider a randomized clinical trial with two drug treatment conditions: a standard treatment and a new treatment, and two outcomes: an indicator of survival ($D = 1$ for survival; $D = 0$ for death) and QOL score ($Y$). We assume that the pre-randomization covariates ($X$) including a baseline QOL score are available. The objective is to draw inferences about the effect of treatment ($Z$) on the QOL score, that is, to compare the mean QOL score between treatment groups. Some patients, however, will die during follow-up, with the result that the outcome $Y$ is not defined. More generally, $D$ is an indicator for a condition that makes $Y$ undefined. In this subsection, we also

assume that there are no missing data due to drop-outs.

Let $\boldsymbol{Z}$ be the vector of treatment assignments for the $N$ randomized patients, with the $i$th element $Z_i$ ($Z_i = 1$ for a new treatment; $Z_i = 0$ for a standard treatment). We define the potential outcomes of the study patients. Let $\boldsymbol{D}(\boldsymbol{Z})$ be the $N$-vector with the $i$th element $D_i(\boldsymbol{Z})$, which is the indicator of whether the $i$th patient would survive given $\boldsymbol{Z}$. For patients with $D_i(\boldsymbol{Z}) = 1$, let $Y_i(\boldsymbol{Z})$ be the QOL score given $\boldsymbol{Z}$. In order to limit the possible potential outcomes for each patient, we adopt Rubin's (1974) stable unit treatment value assumption (SUTVA) throughout. It assumes that $D_i(\boldsymbol{Z}) = D_i(\boldsymbol{Z}')$ whenever $Z_i = Z_i'$, and $Y_i(\boldsymbol{Z}) = Y_i(\boldsymbol{Z}')$ whenever $Z_i = Z_i'$ and $D_i(Z_i) = D_i'(Z_i') = 1$. SUTVA means that potential outcomes for each patient $i$ are unrelated to the assignment $Z_j$ ($j \neq i$) of other patients, and allows $D_i(\boldsymbol{Z})$ and $Y_i(\boldsymbol{Z})$ to be written as $D_i(Z_i)$ and $Y_i(Z_i)$, respectively. Therefore, under SUTVA, each patient has two potential outcomes for survival ($D_i(1)$, $D_i(0)$), and at most two potential outcomes for QOL score ($Y_i(1)$, $Y_i(0)$). For each patient, only one of $D_i(1)$ or $D_i(0)$ is observed. Note that $Y_i(1)$ ($Y_i(0)$) is defined only if $D_i(1) = 1$ ($D_i(0) = 1$). We will also assume the consistency assumption that, for every individual $i$, if the actual value of $Z_i$ turns out to be $z_i$, then the value that $D$ (or $Y$) would take on if $Z_i$ were $z_i$ is equal to the actual value of $D$ (or $Y$). This assumption relates the observed outcome to the potential outcomes.

When randomly assigned treatment is administered correctly and there is no loss to follow-up, the average treatment effect $E\{Y_i(1) - Y_i(0)\}$ can be estimated from the observable data by $E(Y_i \mid Z_i = 1) - E(Y_i \mid Z_i = 0)$ (Holland, 1986). If deaths occur during follow-up, a standard method restricts the estimation procedure to the population with $D_i = 1$ and estimates $E(Y_i \mid Z_i = 1,\ D_i = 1) - E(Y_i \mid Z_i = 0,\ D_i = 1)$. This comparison, however, is not interpreted as a causal effect, because the observed two groups, ($Z_i = 1$, $D_i = 1$) and ($Z_i = 0$, $D_i = 1$) will not be comparable unless the event of death is completely at random (Rosenbaum, 1984).

To overcome this problem, Frangakis and Rubin (2002) have proposed a comparison of potential outcomes under standard versus new treatment within true survivors:

$$E\{Y_i(1) - Y_i(0) \mid D_i(1) = D_i(0) = 1\}. \tag{1}$$

As was suggested in Rubin (1998, 2000), the most meaningful inferences about the causal effects on Y can be drawn only for this subgroup, because both $Y_i(1)$ and $Y_i(0)$ are defined only for this subgroup. This population level causal parameter is the effect of the treatment on the QOL scores ($Y$) for a common set of patients, that is, patients who would survive under both treatments. Therefore, this parameter does not suffer from the complicated interpretations of the standard posttreatment-adjusted one.

The local average treatment effect, which is defined by (1), can be written as follows:

$$\mu = \frac{E[\{Y_i(1) - Y_i(0)\}D_i(0)D_i(1)]}{E[D_i(0)D_i(1)]}. \tag{2}$$

The quantity $D_i(0)D_i(1)$ in both the numerator and denominator of (2) takes the value of one for any patient who would have survived under both treatments and takes the value of zero for all the other patients. It is not possible to estimate (2) without introducing assumptions, because the joint distributions involved in the numerator and denominator of (2) are not observable. For example, when the treatment has no effects on survival, that is, $D_i(z) = D_i(1-z)$, we can estimate (2) consistently from the observed data as

$$\frac{\sum_j Y_j(1)D_j(1)}{\sum_j D_j(1)} - \frac{\sum_k Y_k(0)D_k(0)}{\sum_k D_k(0)}, \tag{3}$$

where $j$ indexes over patients assigned to group $z=1$ and $k$ indexes over patients assigned to group $z=0$. However, if the treatment has any effect on survival, the estimator (3) from the observed survivors will in general be a biased estimate of the causal parameter owing to the posttreatment selection bias.

In order to make (2) to be estimable, we make the following assumption about the potential outcomes:

$$\Pr[D_i(1-z)=1 \mid D_i(z), Y_i(z), X_i] = \Pr[D_i(1-z)=1 \mid X_i], \tag{4}$$

where $X_i$ represents the pre-randomization covariates. This assumption means that the probability that survival would have been observed had the patient received the other treatment can be explained only by measured baseline covariates $X_i$. Let $w_i(z) = E[D_i(z) \mid X_i]$ be the expected value of $D_i(z)$ conditional on $X_i$ for $z=0,1$. Then, under the assumption of (4), we have

$$\begin{aligned}
&E[Y_i(z)D_i(z)D_i(1-z) \mid X_i] \\
&\quad = E[Y_i(z)D_i(z) \mid X_i]E[D_i(1-z) \mid X_i] \\
&\quad = E[Y_i(z)D_i(z) \mid X_i]w_i(1-z) \\
&\quad = E[Y_i(z)D_i(z)w_i(1-z) \mid X_i].
\end{aligned}$$

From this equation, conditional on $X_i$ and with a consistent estimator $\widehat{w}_i(z)$ of $w_i(z)$ for $z=0,1$, a consistent estimator of (2) is given by

$$\widehat{\mu} = \frac{\sum_j Y_j(1)D_j(1)\widehat{w}_j(0)}{\sum_j D_j(1)\widehat{w}_j(0)} - \frac{\sum_k Y_k(0)D_k(0)\widehat{w}_k(1)}{\sum_k D_k(0)\widehat{w}_k(1)}, \tag{5}$$

where $j$ indexes over patients assigned to group $z=1$ and $k$ indexes over patients assigned to group $z=0$ (Matsuyama and Morita, 2006).

Although the probabilities of survival under the other treatment are unknown, we can predict them from the data in each treatment group. Therefore, our proposed estimation procedure for (2) consists of the following three steps:

1. Modeling: A model such as logistic regression is used to predict the probability of survival in each treatment group as a function of baseline covariates.

2. Prediction: Using the estimates of the regression parameters in the other treatment group estimated in step 1, the probability that the patient would have survived had the patient received the other treatment, is predicted for each patient.
3. Weighting: The usual analysis comparing the mean QOL score between treatment groups is conducted among the observed survivors using the individually specific weight, which is the estimated probability in step 2.

The use of the weights induces within-individual correlation, which invalidates the model-based standard error estimates outputted by many standard statistical packages. To construct the confidence intervals for the weighted estimator, we used a robust variance estimator (Huber, 1976; Liang and Zeger, 1986). The robust variance estimator provides conservative confidence intervals for the parameter of interest $\theta$, that is, the 95% Wald confidence intervals calculated as $\theta \pm 1.96 \times$ (robust standard error) is guaranteed to cover the true value $\theta$ at least 95% of the time in large samples (Robins, 1999; Robins et al., 2000).

### 3.2 Analysis Model for the QOL score

Because QOL scores are usually measured repeatedly, a repeated-measures model (Fairclough, 2002; Michiels et al., 2002) is usually used to estimate group-specific means of QOL domain scores at each time. In our analysis, response variables are QOL domain scores at visit 1 and visit 2, and explanatory variables were treatment group, domain, visit, and QOL domain scores at visit 0 (baseline QOL domain score). Among explanatory variables, only QOL domain score at visit 0 is continuous variable and the remainders are categorical variables. We used the following model:

$$Y_{ijkl} = \tau_k x_{ikl} + \alpha_l + \beta_k + \gamma_j + (\alpha\beta)_{kl} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{jkl} + \varepsilon_{ijkl}, \tag{6}$$

where $Y_{ijkl}$ is the QOL score of domain $k$ ($k = 1, \ldots, 4$) at visit $j$ ($j = 1, 2$) for patient $i$ assigned to group $l$ ($l = 0, 1$), $x_{ikl}$ is the baseline QOL score of domain $k$ for patient $i$ assigned to group $l$, $\tau_k$ is the domain-specific regression parameter on $x_{ikl}$, $\alpha_l$ is the main effect of treatment group, $\beta_k$ is the main effect of domain, $\gamma_j$ is the main effect of visit, $(\alpha\beta)_{kl}$ is the interaction effect between domain and group, $(\beta\gamma)_{jk}$ is the interaction effect between visit and domain, $(\alpha\beta\gamma)_{jkl}$ is the interaction effect among visit, domain and group, $\varepsilon_{ijkl}$ is the random error term, which are assumed to be normally distributed with mean zero. For the covariance matrix of $\varepsilon_{il}$, which is a $8 \times 1$ vector, unstructured and compound symmetry structure was assumed among domains and time points, respectively. Because we have no prior information about covariance structure among domains, we assumed unstructured one among domains. In our model, compound symmetry and first-order autoregressive (AR(1)), which are commonly used covariance structure for repeated measurements, are equivalent because analyses times are only two points. In this model, the treatment effect adjusted by baseline QOL score at visit $j$ in domain $k$ is $\alpha_l + (\alpha\beta)_{kl} + (\alpha\beta\gamma)_{jkl}$.

### 3.3 Multiple Imputation Method

We have ignored the presence of missing data due to drop-outs in the above subsections. In this paper, we analyze the missing data due to drop-outs using multiple imputation methods (Rubin, 1978, 1996) under the missing at random (MAR) assumption for the drop-out mechanism (Little and Rubin, 2002).

The idea behind imputation is simple: substitute the values that were not recorded with imputed values. One of the advantages of imputation methods is that, once a filled-in data set has been constructed, standard methods for complete data can be applied. However, methods that rely on just a single imputation, creating only a single filled-in data set, fail to acknowledge the uncertainty inherent in the imputation of the unobserved responses. Multiple imputation circumvents this difficulty. In multiple imputation, the missing values are replaced by a set of $M$ plausible values, thereby taking account of the uncertainty about what values to impute for the missing responses. Typically, a small number of imputations, for instance, $5 \leq M \leq 10$, is sufficient to obtain realistic estimates of the sampling variability (Rubin, 1996; Schafer, 1999).

With multiple imputation, $M$ filled-in data sets are created, producing $M$ different sets of parameter estimates and their standard errors. These are then appropriately combined to provide a single estimate of the parameters of interest, together with standard errors that reflect the uncertainty inherent in the imputation of the unobserved responses. Let $\widehat{Q}_m$ and $\widehat{U}_m$ denote the point estimate of parameter of interest $Q$ and the estimated variance of $\widehat{Q}_m$ from the $m$th filled-in data set ($m = 1, \ldots, M$). A single estimate of $Q$ is given by

$$\widehat{Q} = \frac{1}{M} \sum_{m=1}^{M} \widehat{Q}_m,$$

and the estimated variance of $\widehat{Q}$ is given by

$$T = \widehat{U} + (1 + M^{-1})\widehat{B},$$

where $\widehat{U} = \frac{1}{M} \sum_{m=1}^{M} \widehat{U}_m$ is the within-imputation variance and $\widehat{B} = \frac{1}{M-1} \sum_{m=1}^{M} (\widehat{Q}_m - \widehat{Q})(\widehat{Q}_m - \widehat{Q})^T$ is the between-imputation variance. The statistic $(Q - \widehat{Q})/\sqrt{T}$ is approximately distributed as $t$ distribution with $v = (M-1)(1 + r^{-1})^2$ degrees of freedom, where $r = (1 + M^{-1})\widehat{B}/\widehat{U}$.

Several methods for producing the imputed values for the missing responses have been proposed (Lavori et al., 1995; Schafer, 1999). Two imputation methods are widely used: regression-based and propensity score-based method. We used regression imputation method, because the results from the latter method are severely biased due to the residual confounding within groups of "similar" cases, even when the data are missing completely at random (Allison, 2000). In a regression imputation method, a series of regression models for $Y_{ij}$, given $Y_{i1}, \ldots, Y_{ij-1}$ and covariates $X_i$, are fitted using the observed data. For example, when a response variable at visit $j$ ($Y_{ij}$) is subject to missing, a series of linear regression models

$$E(Y_{ij}) = \phi_1 + \phi_2 Y_{i1} + \cdots + \phi_j Y_{ij-1} + \phi_{j+1} X_i$$

are fitted using the observed data on subjects who have not dropped out by the visit $j$, yielding $\widehat{\phi}$ and $\widehat{\sigma}^2$ (the latter is the residual variance from the linear regression model). Parameters $\phi^*$ (and $\sigma^*$) are then drawn from the distribution of $\widehat{\phi}$ (and $\widehat{\sigma}$), to account for the uncertainty in estimating $\phi$ (and $\sigma$). Missing values for $Y_{ij}$ is then imputed on the basis of the following predictions:

$$\phi_1^* + \phi_2^* Y_{i1} + \cdots + \phi_j^* Y_{ij-1} + \phi_{j+1}^* X_i + \sigma^* e_i,$$

where $e_i$ is simulated from a standard normal distribution. Multiple imputations are obtained by repeating these steps $M$ times. In the regression model, we used age, Eastern Cooperative Group Performance Status, sex, treatment group as covariates $X_i$. We created $M = 5$ imputed values in both of the following two methods.

### 3.4 Two Analytical Approaches

The following two methods were compared in the estimation of mean QOL scores and treatment effects using the repeated-measures model (6). In both methods, the regression imputation method was used for drop-out cases (patterns 2-4 in Table 2).

Method 1: SACE

The weighted repeated-measures model (6) was fitted for filled-in observed survivors (patterns 1-4 in Table 2), in which the weight was the estimated probability that the patient would have survived had he/she received the other treatment. Each weighted result was combined by the multiple imputation technique.

Method 2: Standard

The repeated-measures model (6) was fitted for filled-in observed survivors (patterns 1-4 in Table 2). Each unweighted result was combined by the multiple imputation technique.

## 4. Simulation Studies

In order to evaluate the performance of the methods, simulation studies were carried out under different observed death proportions in each treatment group. We simulated data from two treatment groups, coded as $z = 0$ (standard treatment) or $z = 1$ (new treatment). Equal sample size of 150 for each group was generated. The simulations were based on 1000 replications. For simplicity, repeated QOL scores from only one domain were generated. For each subject $i$, a potential QOL score at visit $j$ ($j = 0, 1, 2$) under the assigned group $z$, $Y_{ij}(z)$, was generated via the linear model:

$$Y_{ij}(z \mid j, x_i, a_{0i}, a_{1i}, \varepsilon_{ij}) = \beta_{1z} + \beta_{2z} j + \beta_{3j} x_i + a_{0i} + a_{1i} j + \varepsilon_{ij}, \tag{7}$$

where group-specific parameters $(\beta_{10}, \beta_{11}) = (80, 80)$, visit specific parameters $(\beta_{20}, \beta_{21}) = (-5, -10)$, and a covariate effect parameters $(\beta_{30}, \beta_{31}, \beta_{32}) = (0, -2, -4)$. The random effects, $a_{0i}$ and $a_{1i}$, were generated from the bivariate normal distribution with mean zero, variance 1, and

correlation 0.8. The random error term, $\varepsilon_{ij}$, was generated from multivariate normal distribution with mean zero and variance

$$Var\begin{pmatrix}\varepsilon_{i0}\\ \varepsilon_{i1}\\ \varepsilon_{i2}\end{pmatrix}=\begin{pmatrix}\sigma^2 & \sigma^2\rho & \sigma^2\rho^2\\ \sigma^2\rho & \sigma^2 & \sigma^2\rho\\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2\end{pmatrix},$$

where $\rho=0.8$ and $\sigma=3$. A covariate $x_i$, whose larger value corresponds to worse health condition, was generated from a standard normal distribution. For each subject $i$, an outcome variable under the other treatment assignment, $Y_{ij}(1-z)$, was also generated from the above model (7).

For each subject $i$, the potential drop-out and death indicators at visit $j$ ($j=0,1,2$) under the assigned group $z$, $M_{ij}(z)$ and $D_{ij}(z)$, were generated via the logistic models:

$$\text{logit}\Pr(M_{ij}(z)=0\mid z,j)=\alpha_1+\alpha_{2j} \tag{8}$$

$$\text{logit}\Pr(D_{ij}(z)=0\mid z,j,x_i)=\gamma_{1z}+\gamma_{2j}+\gamma_3 x_i \tag{9}$$

$M_{ij}(z)=0$ and $D_{ij}(z)=0$ indicate that the QOL score at visit $j$ for subject $i$ assigned group $z$ is missing because of drop-out and death, respectively. In equation (8), the intercept parameter $\alpha_1$ was set to be $\log(5/95)$ and the visit specific drop-out parameters were set at $(\alpha_{20},\alpha_{21},\alpha_{22})=(-20,0,0)$, so that the drop-out dose not occur at visit 0 and the drop-out proportions at visit 2 and 3 are about 5% in both groups.

In equation (9), the visit specific death parameters were set at $(\gamma_{20},\gamma_{21},\gamma_{22})=(-20,-20,0)$, so that death occurs only at visit 2. $\gamma_3$, coefficient for $x_i$, was set to be 1. The intercept parameter $\gamma_{1z}$ was set at several values and one of these values were $(\gamma_{10},\gamma_{11})=(\log(60/40),\ \log(60/40))$, where observed death proportions are expected to be about 60% in both groups. Other values were $(\gamma_{10},\gamma_{11})=(\log(40/60),\ \log(60/40))$, $(\log(20/80),\ \log(60/40))$, and $(\log(20/80,\ \log(40/60))$. At visit 2, both drop-out and death can occur in one case, and it was presumed that the death occurred at visit 2 in such a case.

For each subject $i$, outcome indicator variables under the other treatment assignment, $M_{ij}(1-z)$ and $D_{ij}(1-z)$, were also generated from the above model (8) and (9), respectively.

Analysis model for the simulation study is similar to that of the QOL score (6). The difference is that the number of domain is restricted to one:

$$Y_{ijl}=\tau x_{il}+\alpha_l+\gamma_j+(\alpha\gamma)_{lj}+\varepsilon_{ijl}, \tag{10}$$

where $Y_{ijl}$ is the QOL score at visit $j$ ($j=1,2$) for patient $i$ assigned to group $l$ ($l=0,1$), $\tau$ is the coefficient for $x_{il}$, $\alpha_l$ is the main effect of treatment group, $\gamma_j$ is the main effect of visit, $(\alpha\gamma)_{lj}$ is the interaction effect between group and visit, $\varepsilon_{ijl}$ is the random error term, which are assumed to be normally distributed with mean zero. For the covariance matrix of $\varepsilon_{il}$, which is a $2\times 1$ vector, compound symmetry structure was assumed among time points. In this model, the treatment effect adjusted for baseline QOL score at visit $j$ is $\alpha_l+(\alpha\gamma)_{lj}$.

Two analytical approaches explained in Section 3 were applied to simulation data with the model (10). The individual weights for SACE were estimated by the logistic regression model that included $x_i$ as a covariate. The imputation model for drop-out cases were the linear model that included the previous QOL scores and $x_i$ as covariates. The simulation results were evaluated in terms of the bias from the true value, the mean squared error (MSE), and the 95% coverage probability. The true value was estimated as the treatment effect in the true survivors, that is a subpopulation composed of subjects whose death indicators were $D_i(0) = D_i(1) = 1$.

## 5. Results

### 5.1 Simulation Studies

Table 3 shows the simulation results for the parameters of treatment effect at visit $j$ ($\alpha_l + (\alpha\gamma)_{lj}$ in model (10)), which is the QOL score differences between group A and group B at visit 1 and visit 2. Each row of Table 3 reports the Monte Carlo mean bias, MSE, and coverage probability of the nominal 95% large sample confidence intervals according to the combinations of the death proportion in each group. Examining the upper 6 rows of Table 3, the proposed estimator for SACE was nearly unbiased, while the standard one was largely biased. The proposed estimator had smaller MSE. The coverage probabilities for the proposed estimator based on the robust variance were close to the nominal level of 95%. The bias of the standard estimator was also reflected in the smaller coverage probabilities. Examining the last 2 rows of Table 1, we observe that both methods gave unbiased estimates, as expected, when death proportions were equal between the treatment groups.

**Table 3.** Results of the simulation studies.

| Death Proportion | | | Proposed Method (SACE) | | | Standard Method | | |
|---|---|---|---|---|---|---|---|---|
| Group A | Group B | Visit | Bias | MSE | 95% Coverage Probability | Bias | MSE | 95% Coverage Probability |
| 40% | 60% | 1 | 0.002 | 0.270 | 91.9% | −0.252 | 0.329 | 87.7% |
| | | 2 | −0.003 | 0.481 | 96.6% | −0.560 | 0.798 | 89.8% |
| 20% | 60% | 1 | 0.003 | 0.245 | 91.5% | −0.501 | 0.491 | 74.0% |
| | | 2 | 0.006 | 0.446 | 96.6% | −1.120 | 1.705 | 65.1% |
| 20% | 40% | 1 | 0.007 | 0.214 | 91.5% | −0.244 | 0.273 | 86.8% |
| | | 2 | 0.008 | 0.388 | 95.7% | −0.556 | 0.704 | 87.9% |
| 60% | 60% | 1 | 0.007 | 0.305 | 91.1% | −0.001 | 0.293 | 91.5% |
| | | 2 | −0.004 | 0.541 | 96.2% | −0.021 | 0.540 | 96.0% |

### 5.2 Application to NSCLC Data

The QOL score differences between two groups (Group A − Group B) in each domain averaged over visits, which are interpreted as an overall treatment effect, are presented in Table 4. Mean QOL scores in all domains tended to be improved among patients receiving Docetaxel plus Cisplatin (Group A), but the differences were not statistically significant. The SACE method gave

**Table 4.** Estimates of the overall QOL score differences (Group A – Group B) in each domain.

| Domain | Method | Estimates | 95% Confidence Interval |
|---|---|---|---|
| Functional | SACE | 5.76 | −0.86, 12.38 |
| | Standard | 5.56 | −0.97, 12.09 |
| Physical | SACE | 6.03 | −0.44, 12.50 |
| | Standard | 5.82 | −0.47, 12.11 |
| Mental | SACE | 2.54 | −3.61, 8.69 |
| | Standard | 2.70 | −3.32, 8.72 |
| Psychosocial-well-being | SACE | 2.43 | −3.08, 7.94 |
| | Standard | 2.25 | −3.10, 7.60 |

slightly larger treatment effects than standard one in three domains. The underestimation of QOL score differences by the standard method is probably due to the difference of the drop-out/death proportion between two groups. For example, in the QOL-ACD questionnaire, the physical domain (items 7-11) consists of items which are likely to be affected by the chemotherapy. One of the major side effects of Docetaxel and Vindesine is anorexia, and the appearance of anorexia affects directly item 8 "Did you have a good appetite?" and item 9 "Did you enjoy your meals?" The scores for these items will be high in the observed patients, low in the patients with missing score. Therefore, the observed QOL scores will be high in Group B where missing scores were more observed than Group A. However, in the NSCLC trial data, death cases occurred only at visit 3 and the death proportion was slightly larger in Group B (risk difference $= 8.4\% - 12.4\% = -4\%$). Therefore, the results of the proposed analyses were similar to the standard results. With the increase of death cases, the substantial differences between the results from two methods will be observed as presented in Table 3.

## 6. Discussion

We proposed a method to estimate the local average treatment effect in the presence of missing outcome due to both death and drop-out. The proposed method can be used in a variety of applications where the comparison of treatments adjusted for posttreatment variables are required. As analyzed in this paper, the comparison of QOL data with censoring due to death is one example. Another example is the analysis of responders among randomized patients. Gilbert, Bosch, and Hudgens (2003) considered a randomized study to evaluate the efficacy of a preventive HIV vaccine among infected subjects. They used a similar framework of potential outcomes to formulate the causal estimands, which are defined in terms of the distributions of potential viral loads given assignment to receive vaccine or placebo for subjects in the always-infected principal stratum.

In randomized phase III clinical trials for NSCLC, the evaluation of survival is usually primary concern because of the poor prognosis of the patients. One of the purposes of measuring

QOL is to evaluate the treatment effect that cannot be measured by survival. Then the local average causal effect is appropriate for the interpretation as the difference of QOL scores, which is not affected by survival. It is recommended to use this estimand rather than the standard one for presenting the QOL differences among survivors.

An estimation method for the SACE proposed by Matsuyama and Morita (2006) is simple in that it requires only the prediction of the probability of the event in each treatment as a function of covariates. For any type of outcome variables, their weighted estimator can be easily constructed among patients for whom the event has occurred and can be interpreted as the treatment effect among patients who would have had the event in either treatment group. This weighted analysis can be easily fitted in many standard statistical packages. For the drop-out cases, we proposed that the multiple imputation method was conducted in the weighted analysis under the assumption of MAR.

A closely related estimation method for the principal causal effect has been proposed by Gilbert et al. (2003). Zhang and Rubin (2003) have considered the problem of truncation by death in randomized experiments and derived large sample bounds for the principal causal effect, with or without various identification assumptions. Our approach differs from their approaches in incorporating information from variables related to the posttreatment variable (death) and outcome (QOL score). Furthermore, our approach does not require the assumption which rules out the existence of patients who would survive under control treatment but would not survive under new treatment. This assumption, which is similar to the monotonicity assumption by the Angrist, Imbens, and Rubin (1996), may be reasonable in a placebo controlled study, but is not reasonable in an active controlled study.

The validity of the inferences from our proposed analysis depends on two key assumptions, that is, (i) the value of the counterfactual indicator of survival is independent of the QOL score conditional on covariates, (ii) the drop-out mechanism is MAR. Both these assumptions are non-identifiable assumptions and are not testable from the observed data. For these issues, it is important to collect data on a sufficient number of covariates for the survival and drop-out to ensure that the assumption of no-unmeasured-covariates will be at least approximately true. For the analyses of QOL data, there is the possibility that the different covariates are needed for ensuring these assumptions for different domains. In this situation, it is appropriate to analyze QOL domain scores separately.

In this study, we focused on summarizing one treatment effect averaged over domain. To predict the probability of survival in each group, sex, age, stage, and performance status, which are most clinically important factors for survival, were used as the covariates. We could not use other covariates such as albumin or low-density lipoprotein due to their high proportions of missing values. For the imputation model, baseline QOL scores before the drop-out in addition to the above four baseline covariates were included. We believe that these assumptions seem to be

reasonable to some extent in our example. Otherwise, it will be necessary to extend the proposed method to investigate the sensitivity of our inferences to the fundamental assumptions of no-unmeasured-covariates for death and drop-out separately for each domain (Robins, Rotnitzky, and Scharfstein, 1999; Gilbert et al., 2003). For this issue, further research will be needed.

**Acknowledgements**

We are grateful to the Japanese Taxotere Lung Cancer Study Group for providing us with these valuable data. We would like to thank two referees for their many helpful comments on an earlier version of this paper. This research was supported in part by a Grant-in-Aid for Scientific Research (A) No. 16200022.

## REFERENCES

Allison, P. D. (2000). Multiple imputation for missing data: a cautionary tale. *Sociological Methods and Research* **28**, 301-309.

Angrist, J., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* **91**, 444-472.

Bernhard, F., Cella, D. F., Coates A. S., et al. (1998). Missing quality of life data in cancer clinical trials: Serious problems and challenges. *Statistics in Medicine* **17**, 517-532.

Brown, J., Thorpe, H., Napp, V., et al. (2005). Assessment of quality of life in the supportive care setting of the Big Lung Trial in non-small-cell lung cancer. *Journal of Clinical Oncology* **23**, 7417-7427.

Donaldson, G. W., and Moinpour, C. M. (2005). Learning to live with missing quality-of-life data in advanced-stage disease trials. *Journal of Clinical Oncology* **23**, 7380-7384.

Fairclough, D. L. (2002). *Design and Analysis of Quality of Life Studies in Clinical Trials*. Boca Raton: Chapman & Hall.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21-29.

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59**, 531-541.

Holland, P. W. (1986). Statistics and causal inference. *Journal of American Statistical Association* **81**, 945-960.

Huber, P. J. (1976). The behavior of maximum likelihood estimates under nonstandard conditions. In : *Proceedings of the Fifth Berkley Symposium in Mathematical Statistics and Probability*. 221-233. Berkley: University of California Press.

Japanese Society of Medical Oncology (1999). *Clinical Oncology*, 2nd edition. Tokyo: Cancer and Chemotherapy Publishers. (in Japanese).

Kelly, K., Crowley, J., Bunn, P. A., et al. (2001). Randomized phase III trial of Paclitaxel plus Carboplatin versus Vinorelbine plus Cisplatin in the treatment of patients with advanced non-small-cell lung cancer: A Southwest Oncology Group Trial. *Journal of Clinical Oncology* **19**, 3210-3218.

Kubota, K., Watanabe, K., Kunitoh, H., et al. (2004). Phase III randomized trial of Docetaxel-Cisplatin compared with Vindesine-Cisplatin in patients with stage IV non-small-cell lung cancer. *Journal of Clinical Oncology* **22**, 254-261.

Kurihara, M., Shimizu, H., Tsuboi, K., et al. (1999). Development of quality of life questionnaire in Japan: Quality of life assessment of cancer patients receiving chemotherapy. *Psychooncology* **8**, 355-363.

Lavori, P. W., Dawson, R., and Shera, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine* **14**, 1913-1925.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **31**, 13-22.

Little, R. J.A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. Hoboken: Wiley Interscience.

Matsumoto, T., Ohashi, Y., Morita, S., et al. (2002). The quality of life questionnaire for cancer patients treated with anticancer drugs (QOL-ACD): Validity and reliability in Japanese patients with advanced non-small-cell lung cancer. *Quality of Life Research* **11**, 483-493.

Matsuyama, Y. and Morita, S. (2006). Estimation of the average causal effect among subgroups defined by post-treatment variables. *Clinical Trials* **3**, 1-9.

Michiels, B., Molenberghs, G., Bijnens. L., et al. (2002). Selection models and pattern mixture models to analyse longitudinal quality of life data subject to drop-out. *Statistics in Medicine* **21**, 1023-1041.

Pauler, D. K., McCoy, S., and Moinpour, C. (2003). Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Statistics in Medicine* **22**, 795-809.

Robins, J. M. and Greenland S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, 143-155.

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*. M. E. Halloran and D. Berry (eds.), 95-134. New York: Springer-Verlag.

Robins, J. M., Hernan, M. A. and Brumback, B. (2000). Marginal structual models and causal inference in epidemiology. *Epidemiology* **11**, 95-134.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of American Statistical Association* **90**, 106-121.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (1999). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: *Statistical Models in Epidemiology: The Environment and Clinical Trials*. M. E. Halloran and D. Berry (eds.), 1-94. New York: Springer-Verlag.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A* **147**, 656-666.

Rubin, D. B. (1978). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association* **91**, 473-489.

Rubin, D. B. (1998). More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in Medicine* **17**, 371-385.

Rubin, D. B. (2000). Comment on "Causal inference without counterfactuals," by Dawid AP. *Journal of the American Statistical Association* **95**, 435-437.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research* **8**, 3-15.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics* **28**, 353-368.

Japanese Journal of Biometrics Vol. 27, No. 2, 109–119 (2006)

Original Article

# Sensitivity Analysis of Publication Bias in Meta-analysis: A Bayesian Approach

Kimihiko Sakamoto, Yutaka Matsuyama and Yasuo Ohashi

Department of Biostatistics / Epidemiology and Preventive Health Sciences,
School of Health Sciences and Nursing, University of Tokyo.
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033.
e-mail:sakamoto@epistat.m.u-tokyo.ac.jp

Due to the selection process in academic publication, all meta-analysis of published literature is more or less affected by the so-called publication bias and tends to overestimate the effect of interest. Statistically, publication bias in meta-analysis is a selection bias which results from a non-random sampling from the population of unpublished studies. Several authors proposed methods of modelling publication bias using a selection model approach, which considers a joint modelling of the weight function representing the publication probability of each study and a regression of the outcome of interest. Copas (1999) showed that in this approach some of the model parameters are not estimable and a sensitivity analysis should be conducted. In implementing the Copas's sensitivity analysis of publication bias, a practical difficulty arises in determining the range of sensitivity parameters appropriately. We propose in this article a Bayesian hierarchical model which extends Copas's selectivity model and incorporates the experts' opinions as a prior distribution of sensitivity parameters. We illustrate this approach with an example of the passive smoking and lung cancer meta-analysis.

*Key words*: publication bias; meta-analysis; sensitivity analysis; Gibbs sampling.

## 1. Introduction

Since it is not possible to publish all the results of scientific research ever conducted, selection of papers that is worth publishing is an integral part of academic publication process. As reviewed recently by Rothstein et al. (2005), empirical evidence strongly suggests that a research with positive results is more likely to reach publication than a work with negative or inconclusive findings. Due to this tendency, every research based on published literature, especially systematic review and meta-analysis, is more or less biased toward overestimating the effect of interest. This is called a publication bias, which has been regarded as one of the major difficulties in the statistical aspect of meta-analysis for decades.

Although no standard statistical method for detecting and treating publication bias has

Received June 2006. Revised July 2006. Accepted August 2006.

been established, several methods have been proposed for meta-analysis allowing for publication bias. Hedges (1984) introduced a selection model approach, which assumes a weight function representing the publication probability of each study and considers a joint modelling of the weight function and a regression of the outcome of interest. In this method, a pool of yet unpublished (or even not written) studies is considered, from which each study in a meta-analysis is assumed to be sampled. Publication bias can be understood as a selection bias due to non-random sampling from this hypothetical population.

Copas's model for meta-analysis (Copas, 1999), which we briefly review in Section 2, is a kind of weight function-based method. His contribution to this problem is that some of the parameters in the weight function are not estimable and a sensitivity analysis should be conducted. Copas applied the sensitivity analysis approach to the reanalysis for a meta-analysis of passive smoking and lung cancer risk and showed that the estimated relative risk is lower than that of the usual random effects model (DerSimonian and Laird, 1986).

When performing a sensitivity analysis, it is important to determine the appropriate range of sensitivity parameters so that they cover the plausible range of inference. In the reanalysis of passive smoking meta-analysis, for example, Copas determined the range of parameters by transforming it into the distribution of the number of unpublished studies and assuming its upper limit, somewhat arbitrarily, as ranging from zero to sixty (Copas and Shi, 2000a). Readers without any prior knowledge of this research area may find this assumption as fairly admissible, but experts in this area, including the authors of the original meta-analysis, criticized it as assuming unrealistically many unpublished studies (Hackshaw et al., 2000).

The motivation of this article is to utilize the prior belief expressed by the experts in determining the range of sensitivity parameters. We propose a Bayesian version of Copas's model incorporating the experts' opinion as prior distributions of sensitivity parameters, and perform a sensitivity analysis for various prior beliefs. Thus the problem of determining the range of sensitivity parameters can be reduced to that of sensitivity analysis of various prior distributions in a formal Bayesian inference.

This paper is organized as follows. In Section 2, we first review the Copas's model and accompanying problems of interpreting the values of sensitivity parameters as the number of unpublished studies. Then we introduce a prior distribution for this quantity and construct a method for Bayesian parameter estimation using Gibbs sampling. In Section 3, this method is applied to the passive smoking and lung cancer meta-analysis. A brief discussion and conclusion is given in the last section.

## 2. Sensitivity analysis of publication bias

### 2.1 Copas's selectivity model

Copas (1999) and Copas and Shi (2000a) proposed a sensitivity analysis of publication bias in meta-analysis, following the method for sensitivity analysis of selection bias due to non-random