Published in final edited form as:

J Appl Stat. 2012 ; 39(8): 1733–1747. doi:10.1080/02664763.2012.681362.

A Bayesian Approach for Generalized Linear Models with Explanatory Biomarker Measurement Variables Subject to Detection Limit – an Application to Acute Lung Injury

Huiyun Wu^{a,*}, Qingxia Chen^{a,*}, Lorraine B. Ware^b, and Tatsuki Koyama^a

Huiyun Wu: huiyun.wu@vanderbilt.edu

^aDepartment of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN, USA

^bDepartment of Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA

Abstract

Biomarkers have the potential to improve our understanding of disease diagnosis and prognosis. Biomarker levels that fall below the assay detection limits (DLs), however, compromise the application of biomarkers in research and practice. Most existing methods to handle non-detects focus on a scenario in which the response variable is subject to detection limit; only a few methods consider explanatory variables when dealing with DLs. We propose a Bayesian approach for generalized linear models with explanatory variables subject to lower, upper, or interval DLs. In simulation studies, we compared the proposed Bayesian approach to four commonly used methods in a logistic regression model with explanatory variable measurements subject to DL. We also applied the Bayesian approach and other four methods in a real study, in which a panel of cytokine biomarkers was studied for their association with acute lung injury (ALI). We found that IL8 was associated with a moderate increase in risk for ALI in the model based on the proposed Bayesian approach.

Keywords

detection limit; Bayesian; generalized linear model; biomarker; lung injury

1 Introduction

Biomarkers have drawn growing research interest for their potential to improve disease diagnosis and prognosis based on biomarkers' underlying association with pathologic and physiologic mechanisms [1]. Much effort has been devoted to developing approaches for biomarker discovery and to applying biomarkers to predict both clinical outcomes for patients and relative risk for populations [2]. Many studies to date, however, have been limited in their effectiveness because biomarker levels can only be measured within the thresholds of assays and lab instruments. The actual biomarker levels are unmeasurable outside of these limits, and such a problem is referred to as a detection limit (DL) problem. These unmeasurable values are often called non-detects. Nutrient levels, for example, often suffer from a lower DL [3, 4]. Cytokine biomarkers, many involved in the inflammation processes, also encounter the lower detection constraint. In a study by Kinasewitz *et al.* [5], 4 (IL1 β , TNF- α , IL8, IL10) of the 19 biomarkers under investigation had baseline levels below the DLs in 91%, 47%, 60%, and 59% of the patients, respectively, significantly

Correspondence to: Qingxia Chen.

*These authors contributed equally to this work.

limiting the analysis of the association between each of these 4 biomarkers and disease severity and mortality. In a clinical study by Fremont *et al.* [6] on the association between a panel of biomarkers and the diagnosis and prognosis of acute lung injury (ALI), IL8 suffered from a lower DL, and an artificial value was used in the regression model. Often, the risk of disease or poor clinical outcomes increases as a biomarker level approaches zero, and statistical models involving these non-detects require special care to avoid false negative conclusions.

Conventional methods for approaching datasets with values below a DL threshold include deletion and simple replacement of non-detects with a constant (e.g., zero, DL threshold, or one half DL threshold) [7, 8, 9, 10]. Other sophisticated methods have been developed based on a normal distribution assumption. Helsel [11] has proposed several comprehensive methods including the regression on order statistics (ROS) method. Rubin [12] and Lubin *et al.* [13] have proposed a multiple imputation (MI) approach in which non-detects are regarded as missing data. Two other methods for large DL fractions are Tobit regression and logistic regression [14]. In many of these studies, however, the DL is considered only for the response variable and not for the explanatory variables [15, 16, 17, 18].

Until recently, estimation of the regression coefficient with an explanatory variable subject to DL has been limited. Richardson *et al.* [15] proposed to replace the exposure measurements below the DL with $E(x|x < DL)$, assuming a normal distribution for x . In contrast, Schisterman *et al.* [10] proposed to use the sample mean of $X|X > dl$ for the non-detects so that a distributional assumption is not required. These two methods, however, only apply to linear regression models with a single predictor. Most recently, Nie *et al.* [19] proposed a maximum likelihood approach under a normal distribution for x . This method also considered only a linear regression model with one predictor. In this paper, we propose an unified Bayesian approach, which is able to handle multiple explanatory variables subject to DL in generalized linear models. We also demonstrate that this method can be easily extended to cases where explanatory variables are subject to upper DLs or interval DLs (both lower and upper DLs).

The paper is organized as follows: In Section 2, we provide a detailed description of the motivating study which focused on the predictive capability of a panel of inflammatory cytokine biomarkers for patient mortality in patients with ALI. In Section 3, we describe the proposed Bayesian method for the cases in which single or multiple explanatory variables are subject to lower, upper, and interval DLs. In Section 4, we present an extensive simulation study to examine the performance of the proposed method, comparing it to 4 existing methods. A detailed analysis of the ALI study is given in Section 5, followed by the discussion and extension of the proposed method in Section 6.

2 Motivating study

The study design and protocol were published elsewhere [20]. Briefly, 8 cytokine biomarkers that reflect the complex pathogenesis of acute lung injury/acute respiratory distress syndrome (ALI/ARDS) were measured in baseline plasma from 549 patients enrolled in the National Heart, Lung, and Blood Institute (NHLBI) ARDS Clinical Trials Network clinical trial of two different levels of positive end-expiratory pressure [20, 21]. These biomarkers included markers of inflammation (IL6, IL8, and TNFR1), lung and systemic endothelial activation and injury (VWF), lung epithelial injury (SP-D), adhesion molecules (ICAM-1), and activation of coagulation and inhibition of fibrinolysis (protein C, and PAI-1). Among these biomarkers, IL8 had 35% of values that were below the DL threshold of 2.5 *pg/ml* of the enzyme-linked immunosorbent assay. Except for ICAM-1, the other biomarkers also had missing values. Clinical data including age, cause of ALI/ARDS,

severity of illness scoring (APACHE III), ventilator parameters, hemodynamic data, alveolar-arterial O_2 difference (A-a O_2), and clinical outcomes were collected. Multiple logistic regression was used to model the association between the biomarkers and mortality [22].

3 Methods

Because the biomarker measurement is usually modeled after logarithmic transformation, from now on (unless otherwise indicated) we use x to denote logarithmic transformation of the biomarker values. Suppose that $\{(y_i, x_i, \mathbf{z}_i), i = 1, 2, \dots, n\}$ are independent observations, where y_i is the response variable, $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iq})$ is the $q \times 1$ vector of the logarithmically transformed biomarker or explanatory variable subject to DL, and $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{ip})$ is a $p \times 1$ vector of additional covariates.

A generalized linear model (GLM) is assumed for $[y_i|x_i, \mathbf{z}_i]$, which has the conditional density given by

$$f(y_i|x_i, \mathbf{z}_i, \theta_i, \tau) = \exp\{\tau(y_i\theta_i - b(\theta_i)) + c(y_i, \tau)\}, \quad i = 1, 2, \dots, n, \quad (1)$$

where $\theta_i = \theta(\eta_i)$ is the canonical parameter, $\eta_i = \beta_0 + \beta_1 x_i + \beta_2 \mathbf{z}_i$, β_1 is a $q \times 1$ vector of regression coefficients for biomarkers for which the logarithmic transformation is assumed to improve the distribution, β_2 is a $p \times 1$ vector of regression coefficients for other covariates, and τ is a dispersion parameter. Without loss of generality, we assume $\theta_i = \eta_i$ (canonical link) and $\tau = 1$, as, for example, in logistic regression and Poisson regression models.

3.1 Single Covariate Subject to Detection Limit

We consider the case with one covariate subject to DL ($q = 1$) first and generalize the method to a scenario with multiple covariates subject to DL ($q > 1$) in the next section. With $q = 1$, we assume the lower DL for biomarker measurement x is ld . Because the logarithm transformed biomarker measurement is usually modeled, we describe the observations as $D_{obs} = (y_i, x_i^* = x_i \vee \log(ld), r_i = I(x_i \leq \log(ld)), \mathbf{z}_i, i = 1, 2, \dots, n)$. The likelihood function of the observed data is

$$L = \prod_{i=1}^n [f(y_i|x_i^*, \mathbf{z}_i, \beta) f(x_i^*|\mathbf{z}_i, \alpha, \sigma^2)]^{(1-r_i)} \left[\int_{-\infty}^{\log(ld)} f(y_i|x, \mathbf{z}_i, \beta) f(x|\mathbf{z}_i, \alpha, \sigma^2) dx \right]^{r_i} \quad (2)$$

where $\beta = (\beta_0, \beta_1', \beta_2')'$ and α is a vector of regression coefficients in the biomarker model. The distribution of cytokine biomarker measurements is often positively skewed and fits a log-normal distribution [23, 24]. For illustration purposes, we assume $[x|\mathbf{z}_i]$ has conditional density given by

$$f(x|\mathbf{z}_i, \alpha, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x - \alpha^T \mathbf{z}_i)^2}{2\sigma^2} \right] \quad (3)$$

When $f(y_i|x, \mathbf{z}_i, \beta)$ is a linear regression model, there exists a closed form expression for the integration in (2). In general, however, we have to rely on the numerical approximation for the integration. Let $\tau = 1/\sigma^2$ and $\pi(\beta, \alpha, \tau)$ be the joint prior for (β, α, τ) . Then the joint posterior density of (β, α, τ) based on the observed data is

$$\pi(\beta, \alpha, \tau | D_{obs}) \propto L \times \pi(\beta, \alpha, \tau),$$

where L is given in (2).

In practice, we can specify the independent prior as $\pi(\beta, \alpha, \tau) = \pi(\beta)\pi(\alpha)\pi(\tau)$ and use the non-informative prior for each individual parameter to minimize the influence of the prior on the posterior distribution. In particular, in the simulation study we used a normal prior with mean=0 and precision= 10^{-6} for the regression coefficients including β and α and a gamma prior with mean=1 and variance=100 for τ . As will be demonstrated in the simulation and analysis of the ALI/ARDS example (Sections 4 and 5), the Bayesian method can easily be implemented using WinBUGS or OpenBUGS in R (package R2WinBUGS or R2OpenBUGS) [25, 26].

Extending the proposed method to interval DL (both lower and upper DLs) is straightforward. We assume the upper DL for x is $\log(ud)$ and denote the observations as $D_{obs}=(y_i, x_i^*=(x_i \vee \log(ld) \wedge \log(ud)), r_{il}=I(x_i \leq \log(ld)), r_{iu}=I(x_i \geq \log(ud)), \mathbf{z}_i, i=1, 2, \dots, n)$. The likelihood function of the observed data now becomes

$$L = \prod_{i=1}^n [f(y_i | x_i = x_i^*, \mathbf{z}_i, \beta) f(x_i^* | \mathbf{z}_i, \alpha, \sigma^2)]^{(1-r_{il})(1-r_{iu})} \left[\int_{-\infty}^{\log(ld)} f(y_i | x, \mathbf{z}_i, \beta) f(x | \mathbf{z}_i, \alpha, \sigma^2) dx \right]^{r_{il}(1-r_{iu})} \left[\int_{\log(ud)}^{\infty} f(y_i | x, \mathbf{z}_i, \beta) f(x | \mathbf{z}_i, \alpha, \sigma^2) dx \right]^{(1-r_{il})r_{iu}} \quad (4)$$

The rest of the development remains intact.

3.2 Multiple Covariates Subject to Detection Limit

When there is more than one covariate subject to DL, $f(\mathbf{x}|\mathbf{z})$ in (2) becomes a multivariate density function. An easy extension is to model $f(\mathbf{x}|\mathbf{z})$ through a multivariate Gaussian distribution. The normality assumption may not hold for all biomarkers subject to DL, however. Therefore, following Ibrahim, Lipsitz and Chen (1999) [27] and Chen and Ibrahim (2006) [28], we model the joint distribution of the biomarkers using a series of one-dimensional conditional densities as

$$f(\mathbf{x}|\mathbf{z}, \boldsymbol{\varphi}) = f(x_{i1} | \mathbf{z}_i, x_{i2}, \dots, x_{iq}, \boldsymbol{\varphi}_1) f(x_{i2} | \mathbf{z}_i, x_{i3}, \dots, x_{iq}, \boldsymbol{\varphi}_2) \cdots f(x_{iq} | \mathbf{z}_i, \boldsymbol{\varphi}_q), \quad (5)$$

where $\boldsymbol{\varphi}_k$ is the vector of unknown parameters in the distribution of $f(x_{ik} | \mathbf{z}_i, x_{i;k+1}, \dots, x_{iq}, \boldsymbol{\varphi}_1)$, and $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_q)$. The advantage of (5) is that it allows for a more flexible model specification for the joint distribution of $f(\mathbf{x}|\mathbf{z})$ and is especially useful when the biomarkers follow different distribution functions. However, since the joint model depends on the order of conditioning, it is important to carry out several sensitivity analyses using different orders of conditioning. Ibrahim, Lipsitz, and Chen (1999) [27] and Chen and Ibrahim (2006) [28], among others, have considered this issue in the missing data literature, and their works have shown that estimates are quite robust with respect to the order of conditioning.

When there are q biomarkers subject to lower DL, the observed data are

$D_{obs}=(y_i, x_{i1}^*=(x_{i1} \vee \log(ld_1)), \dots, x_{iq}^*=x_{iq} \vee \log(ld_q), r_{i1}=I(x_{i1} \leq \log(ld_1)), \dots, r_{iq}=I(x_{iq} \leq \log(ld_q)), \mathbf{z}_i, i=1, 2, \dots, n)$, where $\log(ld_k)$ is the lower DL of x_k . Let $\mathbf{x}_{i,-j} = (x_{i1}, \dots, x_{i,j-1}, x_{i,j+1}, \dots, x_{iq})$ denote all the

biomarkers from the i th subject except the j th biomarker. The likelihood function of the i th observed value is

$$L_i = [f(y_i | \mathbf{x}_i = \mathbf{x}_i^*, \mathbf{z}_i, \beta) f(\mathbf{x}_i^* | \mathbf{z}_i, \boldsymbol{\varphi})]^{(1-r_{i1}) \cdots (1-r_{iq})} \prod_{j=1}^q \left[\int_{-\infty}^{\log(ld_j)} f(y_i | \mathbf{x}_{i-j}, x, \mathbf{z}_i, \beta) f(\mathbf{x}_{i-j}, x | \mathbf{z}_i, \boldsymbol{\varphi}) dx \right]^{(1-r_{i1}) \cdots r_{ij} \cdots (1-r_{iq})} \cdots \left[\int_{-\infty}^{\log(ld_q)} \cdots \int_{-\infty}^{\log(ld_1)} f(y_i | \mathbf{x}, \mathbf{z}_i, \beta) f(\mathbf{x} | \mathbf{z}_i, \boldsymbol{\varphi}) dx_1 \cdots dx_q \right]^{r_{i1} \cdots r_{iq}}, \quad (6)$$

where $f(\mathbf{x}_i | \mathbf{z}_i, \boldsymbol{\varphi})$ is defined in (5). The likelihood function of all the observed data is $L = \prod_{i=1}^n L_i$ where L_i is defined in (6). The rest of the development remains the same.

4 Monte Carlo Simulation

To examine the performance of the proposed method, we conducted extensive Monte Carlo simulations for 18 scenarios with 3 distribution functions (normal, t_{25} , and gamma distributions) for biomarker measurements, 3 different percentages of measurements below the DL (10%, 30%, and 50%), and 2 coefficients (a positive true association of $\beta=0.8$ and a null association of $\beta=0$). For each scenario, we considered a sample of 500 subjects and 1000 simulations. In addition, for the simulation of positive association, power comparisons were performed for the nine scenarios with three different distributions and three different DL fractions as well in which β was set at 0.55 and sample size was reduced to 100 so that the full data analysis had 80% power to reject the null hypothesis when the biomarker values were simulated from log-normal distribution. In this study, we compared the following 5 methods for their performance on the estimates of regression coefficients: (1) single replacement of non-detects with one half DL threshold; (2) case-wise deletion of non-detects; (3) regular Multiple Imputation method (MI); (4) extrapolation by ROS; and (5) the proposed Bayesian approach.

Single replacement is the most popular method that assigns a value such as $ld/2$ to undetected values. Deletion is simply excluding (case-wise) all the observations with values below the threshold in the analysis. MI refers to the regular multiple imputation method assuming missing at random [12]. We modified the MI procedure proposed by Lubin *et al.* [13] so that the new procedure takes all aspects of uncertainty in the imputations into account [29, 30, 31, 32]. ROS is a robust method which is often used in environmental science to compute summary statistics [8]. This method involves fitting a regression line on a normal probability plot of the uncensored data (using censored values as place holders) and estimating the model parameters from the line coefficient including mean and standard deviation (SD). We then used the ROS-generated mean and SD to simulate values below the DL according to a truncated normal distribution. Logistic regression was run to estimate the risk with the complete dataset that combined the simulated with the observed data. The simulation of the truncated normal distribution and the following logistic regression were repeated C times to adjust for the uncertainty due to the simulation [33, 34]. The point estimate was the average of the C imputed datasets, and the standard error was calculated using Rubin's rule [35]. In the simulation of the truncated dataset, we used $C=10$. Note that for fair comparison, the ROS algorithm adopted here is a modified version to account for the simulation variability. The original ROS in [8] used $C=1$, as the goal was to estimate summary statistics.

4.1 Simulation I

We simulated datasets of 500 subjects for a biomarker (ICAM-1) and two covariates (age and A-a O_2) using parameters from the motivating study. Specifically, we simulated age from $N(51, 17^2)$, A-a O_2 from $N(315, 147^2)$, and $\log(\text{ICAM-1})$ from $N(3.7, 1.25^2)$. The binary response, dead or alive, was simulated from a logistic regression model with $\text{logit}(P(\text{Died})) = \beta_0 + \beta_1 \log(\text{ICAM-1}) + \beta_2 \text{age} + \beta_3 \text{A-a } O_2$, where $\beta_0 = -5.93$, $\beta_1 = 0.8$ or 0.0 , $\beta_2 = .04$, and $\beta_3 = 0.003$. We then set the threshold value dl for biomarker ICAM-1 so that it had 10%, 30%, and 50% of the values below the threshold. We considered two scenarios in the simulation: (a) The biomarker has a positive true association with mortality risk, i.e., $\beta_1=0.8$, and (b) The biomarker has no association with mortality risk, i.e., $\beta_1=0$.

In this study, all computations were implemented using R 2.10.1 with WinBUGS 14.1 through package R2WinBUGS. We evaluated the methods with bias, empirical standard error (ESE), average standard error (ASE), root mean square error (RMSE), and coverage probability of 95% confidence interval (95%CP), as well as power for scenario (a) and type I error rate for scenario (b). All the statistics summarized in Tables 1–5 were based on the log odds ratio estimates for the logarithmically transformed biomarker values. As shown in Table 1, a small difference was observed at the lower DL fraction across the different methods whereas substantial differences were detected for higher DL fractions. Single replacement with $dl/2$ generated the highest bias and the lowest 95%CP. Deletion had minimal bias, but its estimate was inefficient as evidenced by increased SE. MI had the highest type I error rate with a relatively small bias and larger SE. ROS had a relatively large bias for the alternative but the smallest type I error rate for the null. The Bayesian method had a small bias and RMSE, correct 95% CP, and a nominal type I error rate around 5.0% except for DL fraction at 50% for which the type I error rate was slightly inflated at 6.5% (Table 1). (Table 1 here). In the power analysis with reduced effect size ($\beta_1 = 0.55$) and sample size ($n = 100$), the results showed that the Bayesian method had the highest power and was robust to the increasing DL fraction. The deletion method, on the other hand, had the lowest power among all the methods. The powers of deletion, MI, and ROS dropped significantly when the detection limit fraction increased from 10% to 50%. (Figure 1). (Figure 1 here)

4.2 Simulation II

We studied the robustness of the proposed Bayesian method with respect to the log-normal assumption of the biomarker measurement. We simulated the data in the same manner as in Simulation I, except for the logarithmic transformation of biomarker ICAM-1, which now follows a t_{25} distribution featured with heavier tails. As in Simulation I, we also considered two scenarios, *a* and *b*, one with positive association for ICAM-1 and mortality and one with null association. We adopted the same procedure to evaluate the performance of the different methods.

In this simulation, results similar to those in Simulation I were observed (Table 2 and Figure 1). The Bayesian method had reasonably good performance, outperforming the other methods in terms of RMSE, especially at 50% DL fraction. The Bayesian method also obtained correct 95%CP, type I error rate, and the highest power across the 3 DL fractions. (Table 2 here).

4.3 Simulation III

We further evaluated the proposed Bayesian method for robustness when a gamma distribution was severely misspecified as a normal distribution. We simulated the data as above, except that logarithm transformation of biomarker ICAM-1 follows Gamma(0.5,1) distribution. The results showed that the Bayesian approach did not perform as well as in

normal and t_{25} distributions, especially when evaluated with bias and 95% CP for alternative. As shown in Figure 2, when evaluated by the QQ-plot against a normal distribution for logarithm transformation of the DL biomarker variable in a randomly selected simulation dataset, the normality assumption is obviously violated. Interestingly, single replacement had very small bias, and the bias did not increase across the DL fractions. This is probably because Gamma(0.5, 1) has a highly positive skewness with its 10%, 30% and 50% thresholded values close to zero, which are around 0.008, 0.074 and 0.227, respectively. ROS, another distribution-based method, performed the worst in this setting (Table 3). (Table 3 here, Figure 2 here). All the methods, including the full data analysis, lost their powers under this distribution (Figure 1).

5 ALI/ARDS Example

Before we examined IL8 in this example, we used ICAM-1 to illustrate the properties of different DL methods. We truncated ICAM-1 to generate a dataset with 30% observations below DL. To simplify the modeling, we included only age and A-a O_2 in the multiple logistic regression model since these two covariates showed strong predictive capacity [20]. The distribution of the complete ICAM-1 data after logarithmic transformation shows a bell-shaped pattern (Figure 3), and a normal distribution can be reasonably assumed (Figure 4). Comparing to the full dataset, MI failed to generate any values below DL and had a similar distribution pattern as deletion method. When combined with observed data, the ROS and Bayesian methods had distributions similar to the original complete data, implying that these two methods are likely to yield valid estimates for summary statistics. In terms of the regression coefficient of $\log(\text{ICAM-1})$ in the logistic regression model for mortality, single replacement with $d/2$ resulted in the largest effect attenuation with the smallest standard error followed by MI (Table 4). For this data analysis, estimation by deletion yielded an estimate with larger bias, whereas the ROS and Bayesian methods yielded results better than the others with the Bayesian method having the smallest bias among the 5 methods. (Table 4, and Figure 3 here).

We then extended our model to include both IL8 and ICAM-1 in the following response model:

$$\text{logit} [P(\text{Death}|\log(\text{IL8}), \log(\text{ICAM-1}), \text{age}, A, a)] = \beta_0 + \beta_1 \log(\text{IL8}) + \beta_2 \log(\text{ICAM-1}) + \beta_3 \text{age} + \beta_4 A.a. \quad (7)$$

Though 38% of the IL8 data in this study are subject to DL, a log-normal distribution seemed reasonable based on the QQ-plot of the observed data (Figure 4). When there was more than one covariate subject to DL, we modeled the joint distribution of the DL covariates as in (5). We conducted the sensitivity analysis to the order of the conditioning. In particular, for the Bayesian1 method in Table 5, we modeled

$$f(\log(\text{IL8}), \log(\text{ICAM-1})|\text{age}, A.a) = f(\log(\text{IL8})|\text{age}, A.a, \log(\text{ICAM-1})), \times f(\log(\text{ICAM-1})|\text{age}, A.a) \quad (8)$$

and for the Bayesian2 method, we modeled

$$f(\log(\text{IL8}), \log(\text{ICAM-1})|\text{age}, A.a) = f(\log(\text{ICAM-1})|\text{age}, A.a, \log(\text{IL8}))f(\log(\text{IL8})|\text{age}, A.a). \quad (9)$$

The estimates were not sensitive to the order of sequential conditioning when the two biomarkers were modeled jointly (Table 5). The results showed substantial differences in estimates across the methods. It was also found that ROS and half DL had estimates close to the Bayesian method. Logistic regression from the Bayesian method showed a moderate

increase of mortality risk with IL8 (OR=1.26; 95% CI: 1.12–1.43). (Table 5 and Figure 4 here). The inflammatory marker IL8 was a meaningful predictor of the risk of mortality adjusting for age and A-a O_2 . The result was consistent with previous analyses that indicate that prediction of clinical outcomes using plasma levels of biologic markers provides important information about the pathogenesis of clinical ALI/ARDS [20]. IL8 is a potent neutrophil chemoattractant that plays an important role in the pathophysiology of animal models of ALI [36]. High levels of IL8 have been reported in the pulmonary edema fluid [37] and BAL fluid of patients with ALI/ARDS [38, 39]. In the ARDS Network low tidal volume study, baseline levels of IL8 in the plasma were highly associated with mortality and were reduced by the protective ventilatory strategy [40]. Taken together with these prior studies, the association between plasma IL8 concentrations and mortality in the current study supports the important role of this chemokine in the pathogenesis of clinical ALI/ARDS.

6 Discussion

The proposed Bayesian method performed well in the motivating example and simulation studies I and II with correctly specified or mildly misspecified distribution of the biomarker measurement. However, the proposed Bayesian method was not robust to severe misspecification like Gamma(0.5, 1) as in simulation III, which is an unsymmetric and strongly skewed distribution. The QQ-plots or model selection criteria like deviance information criterion (DIC) can be used to guide the distribution specification in this setting. The proposed Bayesian method was also extended to the scenario with multiple covariates subject to DL. Sensitivity analysis to the order of conditioning should be considered.

Single replacement had the worst performance among all the methods evaluated in simulation study I and II. As we know, single replacement failed to represent the change of exposure level to capture the overall association between the exposure variable and outcome. In a motivating study by Schisterman *et al.* [10], replacement of values below DL with the average of the non-detects, a quantity not available in reality, yielded estimates close to those with all data used. On the other hand, replacement of non-detects with either mean of the data above DL threshold or zero substantially biased the coefficients in linear regression [10]. Nevertheless, when the distribution was severely misspecified as Gamma(0.5,1), single replacement generated reasonable estimates with robustness across all DL fractions to some extent, which is possibly due to the fact that the thresholds are close to zero.

The poor performance of ROS was observed in simulation I and was more substantial in simulation II and III, even at low DL fraction. ROS has been evaluated as one of the most reliable procedures for developing summary statistics of multiple censored data [41]. It has the advantage of generating a normal distribution (Figure 3) and accurate summary statistics (data not shown) for the dataset when combined with observed data. However, when the goal is to estimate regression coefficients, its bias and coverage probability worsened with increased DL fractions. Our study indicates that when DL is present in the explanatory variable, the imputed values may severely distort its relation to the response variable and thus bias the estimate. These problems associated with ROS were also observed in the study by Uh *et al.* [34] in which the response variable suffered from DL and a few explanatory variables were associated with the response. ROS appeared to yield large bias especially at higher DL fractions [34]. Findings from our study and those in Uh *et al.* may be explained by the fact that the relationship between biomarker and outcome is not resolved in the extrapolation algorithm of ROS when there exists a true association. The ignorance of the relationship is not an issue under the null hypothesis, so a reasonable type I rate was observed (Table I, II).

Deletion had small bias but large standard error and RMSE. The result of small bias was expected according to Little and Rubin [42] because the missing biomarker value only depends on the independent variable and not the outcome. On the other hand, deletion always yields larger variability when the DL fraction is high due to dropping a significant amount of information, resulting in a loss of power. Because the deletion method does not require a distribution assumption on the exposure variable with DL, it could be considered as an alternative when the distribution assumption is invalid and the DL fraction is small. When the outcome variable was subject to DL, deletion generated the larger bias than did single replacement, MI and ROS [34] because in this situation the outcome was non-ignorable missing. Poor performance by deletion (or complete case analysis) has been demonstrated in missing data literature [42]. Furthermore, deletion is inappropriate when the goal is to provide the summary statistics for biomarker measurements. MI as a missing data method requires an assumption of missing at random, which DL obviously does not meet. With regard to the imputed values, MI failed to generate substitutions for non-detects.

While various approaches have been proposed to manage non-detects, the application of different methods depends on proper determination of the DL variable distribution function. For the methods considered herein, only the single replacement and deletion method are free of distribution assumption. Others require distribution assumption for the truncated biomarker measurement. A QQ-plot can be used to evaluate this assumption. When the normality assumption is violated, other parametric models, such as GLMs, can be specified for the biomarkers subject to DL and implemented easily in WinBUGS. If the parametric distribution assumption is reasonable, the proposed Bayesian approach should be able to yield valid distribution and best estimates with joint posterior modeling for covariates with non-detects and an outcome variable. Facilitated by the flexibility of Bayesian modeling, our proposed Bayesian approach can handle a single covariate or multiple covariates subject to DL. By properly modifying the attached WinBUGS code, statisticians can easily write the code to fit their specific questions. We would, therefore, recommend the Bayesian approach to DL data analysis when the DL variables are reasonably normally distributed.

Acknowledgments

Contract/grant sponsor: NIH, American Heart Association Established Investigator Award Contract/grant number: HL081332, R21HL097334, UL1 RR024975-01

The authors are grateful to HW Uh for her kind provision of the R code for pooled analysis of multiple simulations with ROS in regression. The authors thank the associate editor and two referees for helpful comments that led to significant improvement of this article. Part of Chen's work was supported by 1R21HL097334 from NHLBI/NIH and UL1 RR024975-01 from NCRR/NIH.

References

1. Pradhan AD, Everett BM, Cook NR, Rifai N, Ridker PM. Effects of initiating insulin and metformin on glycemic control and inflammatory biomarkers among patients with type 2 diabetes. the LANCET randomized trial. *Jama*. 2009; 302:1186–1194. [PubMed: 19755697]
2. Livshits G, Zhai G, Hart DJ, Kato BS, Wang H, Williams FM, Spector TD. Interleukin-6 is a significant predictor of radiographic knee osteoarthritis: The Chingford Study. *Arthritis Rheum*. 2009; 60:2037–2045. [PubMed: 19565477]
3. Browne RW, Whitcomb BW. Procedures for determination of detection limits: Application to high-performance liquid chromatography analysis of fat-soluble vitamins in human serum. *Epidemiology*. 2010; 21:s4–s9.
4. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology*. 1995; 6:356–365. [PubMed: 7548341]
5. Kinasevitz GT, Yan BS, Basson B, Comp P, Russell JA, Cariou A, Um SL, Utterback B, Laterre P, Dhainaut J. Universal changes in biomarkers of coagulation and inflammation occur in patients with

- severe sepsis, regardless of causative micro-organism [ISRCTN74215569]. *Critical Care*. 2004; 8:R82–R90. [PubMed: 15025782]
6. Fremont RD, Koyama T, Calfee CS, Wu W, Dossett LA, Bossert FR, Mitchell D, Wicker-sham N, Bernard GR, Matthay MA, May AK, Ware LB. Acute lung injury in patients with traumatic injuries: utility of a panel of biomarkers for diagnosis and pathogenesis. *J Trauma*. 2010; 68:1121–1127. [PubMed: 20038857]
 7. Finkelstein MM, Verma DK. Exposure estimation in the presence of nondetectable values: another look. *Aihaj*. 2001; 62:195–198. [PubMed: 11331991]
 8. Helsel, D. Nondetects and data analysis: statistics for censored environmental data. John Wiley & Sons, Inc; 2005.
 9. Hornung R, Reed L. Estimation of average concentration in the presence of nondetectable values. *Appl Occup Environ Hyg*. 1990; 5:46–51.
 10. Schisterman EF, Vexler A, Whitcomb BW, Liu A. The limitations due to exposure detection limits for regression models. *Am J Epidemiol*. 2006; 163:374–383. [PubMed: 16394206]
 11. Helsel, DR.; Lee, LR. Analysis of environmental data with nondetects: statistical methods for censored environmental data. Joint Statistical Meetings; 2006; Seattle, Washington: American Statistical Association; 2006.
 12. Rubin DB. Inference and missing data. *Biometrika*. 1976; 63:581–592.
 13. Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, Bernstein L, Hartge P. Epidemiologic evaluation of measurement data in the presence of detection limits. *Environ Health Perspect*. 2004; 112:1691–1696. [PubMed: 15579415]
 14. Takeshi A. Regression analysis when the dependent variable is truncated normal. *Econometrica*. 1973; 41:997–1016.
 15. Richardson DB, Ciampi A. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *Am J Epidemiol*. 2003; 157:355–363. [PubMed: 12578806]
 16. Lyles RH, Williams JK, Chuachoowong R. Correlating two viral load assays with known detection limits. *Biometrics*. 2001; 57:1238–1244. [PubMed: 11764265]
 17. Hughes JP. Mixed effects models with censored data with application to HIV RNA levels. *BIOMETRICS*. 1999; 55:625–629. [PubMed: 11318225]
 18. Chu HT, Moulton LH, Mack WJ, Passaro DJ, Barroso PF, Munoz A. Correlating two continuous variables subject to detection limits in the context of mixture distributions. *J R Stat Soc Ser C Appl Stat*. 2005; 54:831–845.
 19. Nie L, Chu HT, Liu C, Cole SR, Vexler A, Schisterman EF. Linear Regression with an Independent Variable Subject to a Detection Limit. *Epidemiology*. 2010; 21:S17–S24. [PubMed: 21422965]
 20. Ware LB, Koyama T, Billheimer DD, Wu W, Bernard GR, Thompson BT, Brower RG, Standiford TJ, Martin TR, Matthay MA. Prognostic and Pathogenetic Value of Combining Clinical and Biochemical Indices in Patients With Acute Lung Injury. *Chest*. 2010; 137:288–296. [PubMed: 19858233]
 21. Brower RG, Lanken PN, MacIntyre N, Matthay MA, Morris A, Ancukiewicz M, Schoenfeld D, Thompson BT. Higher versus lower positive end-expiratory pressures in patients with the acute respiratory distress syndrome. *N Engl J Med*. 2004; 351:327–336. [PubMed: 15269312]
 22. Schoenfeld D. Survival methods, including those using competing risk analysis, are not appropriate for intensive care unit outcome studies. *Crit Care*. 2006; 10:103. [PubMed: 16420653]
 23. Koch AL. The logarithm in biology. 1. Mechanisms generating the log-normal distribution exactly. *J Theor Biol*. 1966; 12:276–290. [PubMed: 5972197]
 24. Limpert E, Stahel W, Abbt M. Log-normal distributions across the sciences: keys and clues. *BioScience*. 2001; 51:341–352.
 25. WinBUGS. The BUGS Project. [<http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>]
 26. The R Project for Statistical Computing. [<http://www.r-project.org/>]
 27. Ibrahim JG, Lipsitz SR, Chen M-H. Missing Covariates in Generalized Linear Models When the Missing Data Mechanism Is Nonignorable. *Journal of the Royal Statistical Society, Series B*. 1999; 61:173–190.

28. Chen Q, Ibrahim JG. Semiparametric Models for Missing Covariate and Response Data in Regression Models. *Biometrics*. 2006; 62:177–184. [PubMed: 16542244]
29. de Groot JA, Janssen KJ, Zwinderman AH, Moons KG, Reitsma JB. Multiple imputation to correct for partial verification bias revisited. *Stat Med*. 2008; 27:5880–5889. [PubMed: 18752256]
30. Frank, E.; Harrell, J. Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer-Verlag; 2001.
31. Little RA, An H. Robust likelihood-based analysis fo multivariate data with missing values. *Stat Sinica*. 2004; 14:949–968.
32. Siddique J, Belin TR. Multiple imputation using an iterative hot-deck with distance-based donor selection. *Stat Med*. 2008; 27:83–102. [PubMed: 17634973]
33. Hopke PK, Liu C, Rubin DB. Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the Arctic. *Biometrics*. 2001; 57:22–33. [PubMed: 11252602]
34. Uh HW, Hartgers FC, Yazdanbakhsh M, Houwing-Duistermaat JJ. Evaluation of regression methods when immunological measurements are constrained by detection limits. *BMC Immunol*. 2008; 9:59–63. [PubMed: 18928527]
35. Little, RA.; Rubin, DB. Multiple Imputation for Nonresponse in Survey. New York: John Wiley & Sons; 1987.
36. Laffon M, Pittet JF, Modelska K, Matthay MA, Young DM. Interleukin-8 mediates injury from smoke inhalation to both the lung endothelial and the alveolar epithelial barriers in rabbits. *Am J Respir Crit Care Med*. 1999; 160:1443–1449. [PubMed: 10556103]
37. Miller EJ, Cohen AB, Matthay MA. Increased interleukin-8 concentrations in the pulmonary edema fluid of patients with acute respiratory distress syndrome from sepsis. *Crit Care Med*. 1996; 24:1448–1454. [PubMed: 8797614]
38. Donnelly SC, Strieter RM, Kunkel SL, Walz A, Robertson CR, Carter DC, Grant IS, Pollok AJ, Haslett C. Interleukin-8 and development of adult respiratory distress syndrome in at-risk patient groups. *Lancet*. 1993; 341:643–647. [PubMed: 8095568]
39. Wiedermann FJ, Mayr AJ, Kaneider NC, Fuchs D, Mutz NJ, Schobersberger W. Alveolar granulocyte colony-stimulating factor and alpha-chemokines in relation to serum levels, pulmonary neutrophilia, and severity of lung injury in ARDS. *Chest*. 2004; 125:212–219. [PubMed: 14718443]
40. Parsons PE, Eisner MD, Thompson BT, Matthay MA, Ancukiewicz M, Bernard GR, Wheeler AP. Lower tidal volume ventilation and plasma cytokine markers of inflammation in patients with acute lung injury. *Crit Care Med*. 2005; 33:1–6. discussion 230–232. [PubMed: 15644641]
41. Shumway RH, Azari RS, Kayhanian M. Statistical approaches to estimating mean water quality concentrations with detection limits. *Environ Sci Technol*. 2002; 36:3345–3353. [PubMed: 12188364]
42. Little, A.; Rubin, B. Statistical analysis with missing data. 2. New York: John Wiley;

Appendix

WinBUGS code is provided below. Centered values are used in continuous variables.

```
model
{
  for(i in 1: N) {
    mort[i] ~ dbern(p[i])
    logit(p[i]) ← beta0 + beta21*(age[i]-c.age) + beta31*(A.a[i]- c.A.a) +
    beta7*(log.il8[i] - c.il8)+ beta8*(log.icam[i] -c.icam)
    log.il8[i] ~ dnorm(mu.il8[i], tau)I(,logLDL)
    mu.il8[i] ← alpha10 + alpha12 *(age[i]-c.age) + alpha13*(A.a[i] - c.A.a) +
    alpha14*(log.icam[i] - c.icam)
```

```
log.icam[i] ~ dnorm(mu.icam[i], tau2)I(,logLDL2)
mu.icam[i] ← alpha20 + alpha22 *(age[i]-c.age) + alpha23*(A.a[i] - c.A.a)
}
}
```

Figure 1.
Results of power analysis

Figure 2.
Histogram Plot and QQ-plot for a Randomly Selected Dataset from Simulation III

Figure 3.
Distribution of Datasets Completed with Imputed Values

Figure 4.
 QQ-plots for Logarithmic Transformation of Observed ICAM and IL8

Table 1

Estimates of log(OR) in Simulation I

$\mathbf{P}^*$	Method	$\beta_1 = 0.8$				$\beta_1 = 0$	
		Bias [†]	ESE [‡]	ASE [§]	RMSE [¶]	CP(%) ^{**}	Err(%) ^{††}
0.1	Full	0.013	0.097	0.099	0.098	96	4.5
	Half DL	-0.042	0.093	0.094	0.102	92	5.0
	Deletion	0.014	0.116	0.117	0.117	96	4.7
	MI	0.007	0.117	0.116	0.117	96	5.8
	ROS	0.002	0.097	0.097	0.097	95	4.0
	Bayesian	0.024	0.099	0.101	0.102	95	5.1
0.3	Full	0.008	0.104	0.099	0.104	95	4.8
	Half DL	-0.175	0.081	0.077	0.193	39	5.0
	Deletion	-0.000	0.163	0.159	0.163	95	4.8
	MI	-0.026	0.166	0.151	0.168	93	7.2
	ROS	-0.062	0.100	0.100	0.118	89	3.1
	Bayesian	0.019	0.111	0.105	0.113	94	5.4
0.5	Full	0.010	0.102	0.099	0.103	95	4.7
	Half DL	-0.239	0.072	0.072	0.249	11	5.8
	Deletion	0.031	0.239	0.231	0.241	95	4.7
	MI	-0.046	0.243	0.201	0.247	88	9.7
	ROS	-0.166	0.087	0.087	0.188	63	2.4
	Bayesian	0.027	0.118	0.115	0.121	97	6.5

* P is the proportion of DL.

[†] Bias is defined as $\hat{\beta} - \beta$.

[‡] ESE is the empirical standard error.

[§] ASE is the average standard error.

[¶] RMSE is the root mean square error.

^{**} CP is the coverage probability for 95% CI.

⁷⁷Err is the type I error rate under null hypothesis.

Table 2

Estimates of log(OR) in Simulation II

P*	Method	$\beta_1 = 0.8$				$\beta_1 = 0$	
		Bias [†]	ESE [‡]	ASE [§]	RMSE [¶]	CP (%)**	Err (%) ^{††}
0.1	Full	0.007	0.107	0.104	0.107	94	4.4
	Half DL	-0.129	0.128	0.088	0.182	63	6.4
	Deletion	0.011	0.160	0.147	0.160	95	6.6
	MI	-0.012	0.164	0.140	0.164	93	2.8
	ROS	-0.077	0.143	0.143	0.162	75	5.1
	Bayesian	-0.006	0.112	0.106	0.112	93	6.1
0.3	Full	0.006	0.104	0.104	0.104	95	5.4
	Half DL	-0.221	0.075	0.076	0.233	18	5.6
	Deletion	0.017	0.161	0.161	0.162	95	4.6
	MI	-0.012	0.163	0.154	0.164	93	6.0
	ROS	-0.121	0.091	0.091	0.151	77	2.7
	Bayesian	-0.020	0.105	0.105	0.107	94	5.0
0.5	Full	0.005	0.105	0.104	0.106	94	3.2
	Half DL	-0.276	0.069	0.071	0.284	4	3.1
	Deletion	0.019	0.230	0.225	0.231	94	2.5
	MI	-0.059	0.236	0.196	0.243	88	2.5
	ROS	-0.248	0.078	0.078	0.260	22	4.0
	Bayesian	-0.042	0.110	0.110	0.118	92	6.1

See Table 1 for the labels.

Table 3

Estimates of log(OR) in Simulation III

P*	Method	$\beta_1 = 0.8$				$\beta_1 = 0$	
		Bias [†]	ESE [‡]	ASE [§]	RMSE [¶]	CP (%)**	Err (%) ^{††}
0.1	Full	0.026	0.181	0.176	0.183	95	4.9
	Half DL	0.026	0.181	0.176	0.183	95	3.8
	Deletion	0.029	0.187	0.182	0.190	95	3.3
	MI	0.019	0.187	0.182	0.188	95	4.9
	ROS	-0.101	0.151	0.151	0.182	88	3.7
	Bayesian	-0.041	0.166	0.162	0.171	93	3.5
0.3	Full	0.009	0.178	0.175	0.178	95	3.8
	Half DL	0.013	0.179	0.176	0.180	95	4.2
	Deletion	0.019	0.205	0.201	0.206	95	5.3
	MI	-0.017	0.204	0.194	0.205	93	5.3
	ROS	-0.318	0.103	0.104	0.335	26	6.8
	Bayesian	-0.182	0.134	0.136	0.226	71	10.7
0.5	Full	0.012	0.173	0.175	0.173	96	3.9
	Half DL	0.032	0.179	0.181	0.182	95	2.9
	Deletion	0.023	0.237	0.237	0.238	96	0.7
	MI	-0.050	0.235	0.213	0.241	91	6.3
	ROS	-0.453	0.073	0.073	0.459	2	5.0
	Bayesian	-0.265	0.113	0.118	0.288	37	3.3

See Table 1 for the labels.

Table 4Log(OR) estimate of *logICAM-1* in motivating study

Method	Estimate	Bias*	SE [†]
Full	0.775	0.000	0.198
Half DL	0.247	-0.528	0.067
Deletion	0.537	-0.238	0.303
MI	0.446	-0.329	0.288
ROS	0.729	-0.057	0.205
Bayesian	0.782	0.007	0.206

See Table 1 for the labels.

Table 5

Log(OR) estimate of $\log(\text{ICAM-1})$ and $\log(\text{IL8})$ in ALI study on sensitivity analysis

Method	$\log(\text{IL8}) (\beta_1)$		$\log(\text{ICAM-1}) (\beta_2)$	
	Estimate	SE [†]	Estimate	SE [†]
Half DL	0.260	0.059	0.217	0.064
Deletion	0.446	0.122	0.233	0.362
MI	0.483	0.091	0.536	0.269
ROS	0.267	0.068	0.717	0.201
Bayesian1	0.235	0.063	0.632	0.203
Bayesian2	0.240	0.063	0.615	0.194

See Table 1 for the labels. The model is stated in (7). Bayesian1 and Bayesian2 used the models stated in equations (8) and (9), respectively.