New Insights into Modeling Exposure Measurements Below the Limit of Detection

Ana Maria Ortega-Villa[1], Danping Liu[2], Mary H. Ward[3] and Paul S. Albert[2*]

[1] Biostatistics Research Branch, Division of Clinical Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda Maryland, 20892, United States of America.

[2] Biostatistics Branch, Division Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda Maryland, 20892, United States of America.

[3] Occupational and Environmental Epidemiology Branch, Division Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda Maryland, 20892, United States of America.

[*] Corresponding Author. E-mail: albertp@mail.nih.gov, Address: 9609 Medical Center Drive, Rockville, MD 20850 - Room 7E-146, Phone: 240-276-7593

Conflict of Interest: None declared
Type of Manuscript: Original Research Article
Access to Data and Code: Contact corresponding Author

Abstract

Background: In environmental epidemiology it is of interest to assess the health effects of environmental exposures. Some exposure analytes present values that are below the laboratory limit of detection (LOD). There have been many methods proposed for handling this issue in order to incorporate exposures subject to LOD in risk modeling using logistic regression. We present a fresh look at proposed methods to handle exposure analytes that present values that are below the LOD.

Methods: We performed comparisons through an extensive simulation study and a cancer epidemiology example. The methods we considered were a maximum likelihood approach, multiple imputation, Cox regression, complete case analysis, filling in values below the LOD with LOD/$\sqrt{2}$, and a missing indicator method.

Result: We found that the logistic regression coefficient associated with the exposure (subject to LOD) can be severely biased when underlying assumptions are not met, even with a relatively small proportion (under 20%) of measurements below the LOD

Conclusion: We propose the use of a simple method where the relationship between the analyte and disease risk is modeled only above the detection limit with an intercept term at the LOD and a slope parameter, which makes no assumptions about what happens below the LOD. In most practical situations, our results suggest that this simple method may be the best choice for analyzing analytes with detection limits.

Keywords: censored data, environmental epidemiology, limit of detection, logistic regression, missing data, nondetects

What This Study Adds

In many environmental epidemiological applications, exposure analytes present values below the laboratory limit of detection. Many techniques to handle this issue are available in the literature. However, most of these techniques make non-verifiable assumptions about what happens under the LOD. We propose a method in which the relationship between the analyte and the disease risk is modeled only above the detection limit with the aid of a missing indicator, which shows promising results in most practical situations.

Introduction

In epidemiology studies, environmental measurements often have non-detected values below the laboratory limit of detection (LOD), which is the analyte's lowest detectable value that can be differentiated from a blank value (absence of substance).[1]

There have been many methods proposed for incorporating biomarkers subject to lower detection limits as measures of exposure in risk modeling using logistic regression. These include substitution methods that impute a fixed value (LOD, LOD/2, LOD/$\sqrt{2}$, E(X|X ≤LOD), or sample E(X|X >LOD)) for measurements below the LOD,[2-4] multiple imputation,[5] making distributional assumptions on the exposure value, and treating the lower limit of detection (LOD) as being left censored.[6] Recently, a method involving Cox regression and a role reversal between the exposure and the outcome was developed to handle measures subject to LOD.[7] However, for this method, the interpretation of the estimated coefficient is not always comparable to that of a logistic regression but is the log odds ratio of the outcome comparing a subject with $X = x$ versus all subjects with $X > x$.

This paper provides a fresh look at these approaches, particularly with respect to assumptions that are inherently non-verifiable (assuming we do not see measurements below the LOD). Specifically, we examine the robustness of odds ratio estimation in the context of risk using a logistic regression model to the distribution of the exposure as well as to the assumption that the relationship between exposure and risk is the same above and below the LOD. We perform comparisons through an extensive simulation study and in cancer epidemiology. Finally, we provide recommendations regarding the practical choice of different methods to handle LOD issues.

Methods

Study population

Our study population comes from a population-based case-control study of non-Hodgkin lymphoma (NHL) in four National Cancer Institute-Surveillance Epidemiology and End Results Program (NCI-SEER) study sites.[8] Eligible cases were subjects diagnosed with a first primary NHL between July 1998 and June 2000 who were 20-74 years old and free of HIV. Dust was collected at homes at the time of the interview (between February 1999 and May 2001) from vacuum cleaners of participants who gave permission, had used their vacuum cleaner within the past year, and had owned at least half their carpets or rugs for at least 5 years. Dust samples from 682 cases and 513 controls were analyzed between September 1999 and September 2001. The aim of the study was to examine exposure and NHL risk due to a mixture of 27 chemicals in house dust, of which we are focused on the PCB congener 180. The laboratory measurements were subject to missing data, primarily due to concentrations being below the minimum detection level

(20.8 ng/g for PCB 180, 75% under LOD). Further details of the study design can be found in [8]. In this paper, we considered a total of 1,180 subjects, 676 (57%) cases and 511 (43%) controls, where the interest is to estimate the association of NHL incidence with the exposure of PCB 180 in dust.

Statistical Methods

We review each of the analytic approaches in the context of risk using a logistic regression model, discussing important assumptions, advantages, limitations, and relationships between the different methods. Note that these analytes typically follow a log-normal distribution.[9 -11] The methods considered include: utilizing the complete analyte data (i.e. including those values considered to be under the LOD which are known for all simulated datasets), complete case analysis in which one includes only measurements above the LOD,[5] filling the missing values with LOD/$\sqrt{2}$, and the following four approaches:

Maximum-Likelihood approach. This method explicitly accounts for the detection limit by assuming that a transformed exposure variable, t_i, follows a normal distribution left censored at the lower detection limit. Consider the following logistic regression model

$$\operatorname{logit} P(y_i = 1) = \beta_0 + \beta_1 t_i + \alpha^T \boldsymbol{z_i} \qquad (1)$$

where t_i is the log-transformed concentration of the variable subject to LOD, z_i is a vector containing covariates, and y_i represents the health measure. Model parameters are obtained by maximizing the likelihood

$$\mathcal{L} = \prod_{i=1}^{N} \left[f(y_i|t_i, \boldsymbol{z}_i) f(t_i|\boldsymbol{z}_i) \times I_{(t_i \geq LOD)} \right] \times \left[\int_{-\infty}^{LOD} f(y_i|t_i, \boldsymbol{z}_i) f(t_i|\boldsymbol{z}_i) \times I_{(t_i < LOD)} dt_i \right] \qquad (2)$$

where for simplicity we are assuming $f(t_i|\boldsymbol{z_i})$ is a normal distribution with mean $\hat{\mu}(z)$and variance σ^2 and $f(y_i|t_i, \boldsymbol{z_i})$ is a Bernoulli distribution with probability π_i given by

$$\pi_i = \frac{\exp\,(\beta_0 + \beta_1 t_i + \alpha^T \boldsymbol{z_i})}{1 + \exp\,(\beta_0 + \beta_1 t_i + \alpha^T \boldsymbol{z_i})}.$$

The likelihood maximization can be done by using optimization functions in standard software, such as optim in R. However, estimation using this method is computationally intensive due to the required numerical integration in the likelihood. We performed this integration using the trapezoidal rule.

Note that if the model is correct, this method is fully efficient. However, since we do not observe actual exposure measurements that are below the LOD, both the distribution of t_i and the linear association between y_i and t_i below the LOD are non-verifiable assumptions. Treating values below the LOD as censored observations has been referred to as Tobit regression[12,13] in the economics literature.

Multiple Imputation. Rather than accounting for the detection limit by integrating over the unobserved range of the exposure as in the maximum-likelihood method, in this methodology the analyte values under the LOD are imputed based on estimating the analyte value distribution using only values above the LOD. Values below the LOD are then imputed based on the estimated analyte distribution. Lubin et al.[9] proposed the following six-step algorithm for estimation:

Step 1: Let N be the number of subjects in the study. Create a bootstrap sample by sampling subjects with replacement from the original dataset until a bootstrap dataset of size N is obtained.

Step 2: Estimate parameters of the assumed distribution for the analyte. In this case, we considered a log normal distribution with mean μ and variance σ^2, labeled $\tilde{\mu}$ and $\tilde{\sigma}^2$. These parameters can be estimated by maximizing the likelihood:

$$\mathcal{L}_{MI} = \prod_{i=1}^{N} \frac{\phi\left(\frac{\log(t_i) - \mu}{\sigma}\right)}{\sigma\left[1 - \Phi\left(\frac{\log(LOD) - \mu}{\sigma}\right)\right]}$$

where ϕ represents the standard normal probability density function and Φ represents the standard normal cumulative density function.

Step 3: Compute $F(LOD;\, \tilde{\mu}, \tilde{\sigma}^2)$ where $F(\bullet)$ represents the cumulative lognormal distribution. Generate P from a Uniform[0, $F(LOD;\, \tilde{\mu}, \tilde{\sigma}^2)$]. Impute the required observations from $F^{-1}(P; \tilde{\mu}, \tilde{\sigma}^2)$.

Step 4: Fit the logistic regression model using the bootstrap sampled dataset

$$\text{logit}P(y_i = 1) = \beta_0 + \beta_1 t_i + \alpha^T \boldsymbol{z_i}, \qquad i = 1, \ldots, N$$

where t_i represents the concentration of the analyte in the bootstrap dataset.

Step 5: Repeat Steps 1-4 M times; we repeated the process 100 times. It has been stipulated that 3 to 10 times should be enough.[9]

Step 6: Combine the estimates of the M datasets. The sample mean of the bootstrap datasets will be the β coefficient, and the sample standard deviation will be the standard error.[5]

We consider the multiple imputation procedure for cases and controls together,[9] implicitly assuming that $\beta_1 = 0$. Thus, we would expect some attenuation in the estimation of β_1 in this case. This could alternatively be done by doing multiple imputation separately for cases and controls. However, when the study is large and multiple outcomes are present, stratifying the imputation by each outcome may be unfeasible given that the imputation would need to be done for each outcome. Lubin et al.[9] allow the imputation to incorporate individual covariates, so the imputation model in Step 3 can be a linear regression model. We would expect multiple imputation to be less efficient than the maximum likelihood approach.

Cox regression approach. In this methodology the LOD exposure variable is treated as a censored outcome and Cox regression is used to analyze the data.[7] This approach is particularly attractive since it does not explicitly require making any assumptions about the distribution of the analyte under the LOD or about the relationship between exposure and outcome under the LOD. This methodology can be applied as follows:

Step 1: Reverse the scale to change the censoring direction and obtain right censored data. This can be done by selecting M a constant greater than or equal to the maximum of the measure subject to LOD.

Step 2: Use Cox regression to analyze the right censored data. Let $i = 1, \dots, N$ represent the subject, t_i be the concentration of the variable subject to the LOD, c_i represent a concentration variable where $c_i = \max(t_i, LOD)$, δ_i be a censoring indicator which equals one when $t_i > LOD$ and zero otherwise, y_i represent the health measure, $x_i = M - c_i$, and $\boldsymbol{z_i}$ represent covariates. Use standard software to fit the Cox regression model; in our case this was the R function coxph from the survival package,[14] using x_i as the right censored outcome, δ_i as the censoring indicator, and y_i and $\boldsymbol{z_i}$ as covariates for $i = 1, \dots, N$.

Dinse et al.[7] interpret the Cox regression coefficient corresponding to y_i, defined as γ_1, as the log odds ratio relating t_i to the binary variable y_i. Note that the odds ratio estimated in this approach cannot be interpreted in the same way as the odds ratios from all the other approaches. In this case, the interpretation of γ_1 is a log odds ratio of the outcome comparing a subject with $T = t$ versus all subjects with $T > t$, i.e.,

$$\exp(\gamma_1) = \frac{\Pr(y_i = 1|T = t) \,/\, \Pr(y_i = 0|T = t)}{\Pr(y_i = 1|T > t) \,/\, \Pr(y_i = 0|T > t)}.$$

Section A.1 of the Appendix shows the special cases when γ_1 and β_1 have the same interpretation. Although it appears that Cox regression makes few assumptions on the distribution of the exposure, the proportional hazards assumption is indeed very strong and constrains the shape of the dose-response relationship. With simulation studies we found that when equation (1) is the correct model, the proportional hazards assumption will be severely violated for most parameter values (data not shown).

Missing Indicator Method. This method uses only data that can be observed[14]. Let t_i represent a log-transformed concentration of the variable subject to LOD, δ_i be a missing indicator which is equal to one when $t_i > LOD$ and zero otherwise, $\boldsymbol{z_i}$ be a vector containing covariates, and y_i represent the dichotomous disease outcome. The following logistic regression model can be fit using any standard software.

$$\operatorname{logit} P(y_i = 1) = \beta_0 + \beta_1(t_i - LOD) + \beta_2\delta_i + \alpha^T \boldsymbol{z_i} \quad (3)$$

In this model β_2 represents the log odds of disease when at the LOD vs. below the detection limit, and β_1 is the effect of a unit change in the analyte above the detection limit. The main advantage of this approach is that no distributional assumptions are made about information we don't observe, given that we do not model the tail of the exposure distribution.

Simulation Study

We conducted a thorough simulation study to examine the performance of the methodologies described above. We considered the cases where (1) the assumed Gaussian distribution for the log-transformed analyte is correct and the linear effect on the log-odds is true across the whole range of t_i (above and below LOD), (2) the effect of the analyte is different below the LOD than above this value, and (3), the tail distribution below the LOD is non-normal and therefore misspecified. In all cases, log-transformed analyte values t_i were generated from a normal distribution with a mean $\mu = 0$ and variance $\sigma^2 = 2.45$ or 1. For simplicity, no covariates were considered in the simulation study. For all scenarios we set the sample size $N = 2000$ and repeated the simulations 1000 times.

Case 1: Normal distribution of t_i

In this scenario, the log-transformed analyte values follow a Gaussian distribution, and the distribution of t_i below the LOD is correctly specified. Further, we correctly model the dose response as linear on the log odds ratio both above and below the LOD. The disease outcome was generated from a binomial distribution with corresponding logistic regression model

$$\text{logit}P(Y_i = 1) = -0.81 + 0.95t_i. \qquad (4)$$

We investigated the performance of the methodologies at four different LOD cutoff points and two values of σ. For $\sigma^2 = 2.45$ we evaluated cutoffs with 16%, 20%, 30%, and 50% values below the LOD. For $\sigma^2 = 1$ we considered scenarios with 6%, 8%, 20%, and 47% values below the LOD. In this case, the simulated odds ratio of going from the first quantile of the analyte to the third is 8.06.

Table 1 presents the coefficient estimates, standard errors, and Monte Carlo standard errors for each of the methods for a true value of $\beta_1 = 0.95$. We found that with $LOD = 0.2$ and $LOD = 0.25$ (i.e. 16 and 20% below the LOD for $\sigma^2 = 2.45$ and 6 and 8% for $\sigma^2 = 1$), almost all methods (with the exception of Cox regression and fill-in LOD/$\sqrt{2}$) provided nearly unbiased results. When the proportion of values below the LOD increases, the multiple imputation shows increasing attenuation results due to combining cases and controls for imputation. In addition, estimates from Cox regression and fill-in LOD/$\sqrt{2}$ provide increased bias with an increasing proportion of measurements below the LOD. For all approaches, the variance estimation was unbiased, since the mean estimated standard errors were close to the Monte-Carlo standard error.

Further, for this particular case with $\sigma^2 = 2.45$ and 30% values under the LOD, we calculated the empirical type I error rate (simulating $\beta_1 = 0$) and power for three different values of β_1 (0.2, 0.15, and 0.1). Table 3 presents the results of these simulations. Note that all the methods have acceptable empirical type I error rates. In addition, when the model is correctly specified all methods have high power to detect all three values of β_1.

Case 2: Normal distribution of t_i and no effect for values under LOD

In this scenario, the analyte is normally distributed as in Case 1, but there is no effect of the biomarker on outcome below the LOD. For most environmental biomarkers subject to a LOD, it is difficult to verify an assumed relationship between the biomarker and disease outcome for values below the LOD. In some cases, we can use external information such as a more sensitive assay on a smaller number of study participants to assess this relationship, but depending on the study and exposure, this may not be possible. The assumption that this relationship is the same for values below and above the LOD is a very strong assumption. We examine each of the approaches assuming that the relationship has a no-dose response below the LOD. In this case, the disease outcome was generated from a binomial distribution with corresponding logistic regression model given by equation (5), where $I_{(t_i>LOD)}$ is an indicator function with value of one if $t_i > LOD$ and zero otherwise.

$$\text{logit}P(Y_i = 1) = -0.81 + 0.95t_i \times I_{(t_i>LOD)}. \quad (5)$$

In this case, the outcome is affected by the exposure only when the exposure is above the LOD. We considered the same LOD cutoff points as in Case 1, for both $\sigma^2 = 2.45$ and $\sigma^2 = 1$. For this simulation, the odds ratio of going from the first quantile of the analyte to the third is 7.49.

Table 2 presents the results of this simulation. Given that the relationship between the analyte and disease outcome is different for values of t_i above and below the LOD, we find that most methods provide severely biased estimates of β_1 even in cases where the number of values below the LOD is small (i.e. 16% for $\sigma^2 = 2.45$ and 6% for $\sigma^2 = 1$). However, the two methods that focus on observable information, where we either ignore the values below the LOD (complete case analysis) or account for this missingness by using a missing indicator as in equation (3), provide nearly unbiased estimates for all considered missingness evaluations. When there are no additional covariates, these two methods will give identical slope estimates. The difference between the two is that the model (3) provides the estimate β_2, which can be used to assess whether there is an effect for values of the analyte below the LOD.

In addition, we considered $\sigma^2 = 2.45$ and 30% values under the LOD to evaluate the empirical power and type I error rates for β_1 = 0 (type I error rate) and β_1 = 0.2, 0.15, 0.1 (power). We chose the β_1 to be low given we had 100% power for $\beta_1 = 0.95$ and 0.45. Table 3 presents the results of these simulations. We found that all the methods have empirical type I error rates close to the nominal 0.05 rate. In terms of power, we found that all methods had empirical power above 90% to detect β_1 = 0.2, except the Cox regression approach (75.6%). For the case of $\beta_1 = 0.15$, the only methods that had power close to 80% were the maximum likelihood approach (0.807), the complete case analysis (79.8%), missing indicator method (78.1%), the fill in method (84.4%). In this case, the Cox regression case had power less than 60%. For β_1 =0.1, no methods had power above 50%.

Unlike all the other methods, the missing indicator approach requires a two degree of freedom test, for testing the relationship between the analyte subject to LOD and the outcome. For this reason, it could suffer from loss of power. Table 3 demonstrates that this possible loss of power is minimal, and in most cases it results in power comparable to the methods with highest power.

Case 3: Normal distribution of t_i above LOD and uniform distribution below LOD

In this scenario, the effect of the analyte on the outcome remains the same above and below the LOD as in equation (4), but the distribution of log-transformed analyte values is Uniform$[l, u]$ below the LOD and a truncated normal$(0, \sigma^2)$ above the LOD, where the resulting distribution is not discontinuous, as can be seen in Figure 1. Details of the parameters l and u can be found in the Appendix.

Figure 1 presents a visual representation of the distribution from which analyte values t_i were generated for each LOD value. In the figure, the solid line represents the normal distribution, the dashed line represents the uniform distribution, and the shaded area corresponds to the distribution from which the analyte values were generated. Panel (a) corresponds to simulations where $\sigma^2 = 2.45$ and panel (b) to those where $\sigma^2 = 1$.

Table 4 contains the results of this simulation. In this situation, none of the studied methods provide unbiased estimates of β_1. However, for $\sigma^2 = 2.45$, in the case of 16% and 20% under LOD, the multiple imputation approach provides minimum bias. In addition, for 20% under LOD, the maximum likelihood approach ties the multiple imputation for minimum bias, and for larger percentages of missing values, the analytic method provides minimum bias. In the case of $\sigma^2 = 1$, a similar pattern emerges where the multiple imputation and maximum likelihood methods provide the minimum bias, but in addition, the observable-only methods (i.e. complete case analysis and the missing indicator method) match these bias results for 16% and 20% under LOD. For the case of $LOD = 0.95$, the bias in the observable-only methods is due to small sample size bias. For this simulation, the odds ratio of going from the first quantile of the analyte to the third is 7.49.

A separate simulation with a large sample (n = 20,000) was conducted to examine this bias, and we found that for a large sample with 50% values under LOD, the estimates are 0.97, 0.93, 1.11, 0.95, 1.41, and 0.95 for the maximum-likelihood, multiple imputation, Cox regression, complete case analysis, fill-in, and missing indicator approaches, respectively. These results highlight the advantages of approaches that use only the observed data (complete case analysis or the missing indicator approach).

Case 4: Mixture distribution of the analyte

In this scenario the log-transformed analyte values follow the mixture distribution

$$f(t_i) = \theta \times 0 + (1 - \theta)\phi(t_i)$$

where $\phi(t_i)$ is a normal distribution with mean 0 and variance $\sigma^2 = 2.45$. We set the LOD to be 0, and considered two values of $\theta = 0.25$ and 0.5. This corresponds to the situation where LOD values are either censored values of an exposure distribution or true zeros, a situation that occurs quite commonly in environmental epidemiology[16]. We correctly model the dose response as linear on the log odds ratio both above and below the LOD and generated the disease outcome from a binomial distribution as in (4).

Table 5 presents the coefficient estimates, standard errors, and Monte Carlo standard errors for each of the methods for a true value of $\beta_1 = 0.95$. We found that the Cox regression method is highly biased, and the maximum likelihood approach has a small amount of bias in this situation. In addition, we found that for both values of θ, the remaining methods are nearly unbiased.

Data Example: NCI-SEER NHL Study

In this section, we present an illustration of the evaluated methods using one analyte from the NCI-SEER NHL Study, PCB 180. Further, we present parameter estimates and likelihood ratio test results for PCB 180 and γ −Chlordane to illustrate expanding the Missing Indicator approach. The chemical measurements were obtained from carpet dust samples where the quantity of collected material was limited. Given this limited quantity, using a more sensitive assay to characterize the distribution below the LOD was not possible.

The NCI-SEER NHL Study[8] is a population-based case-control study of non-Hodgkin lymphoma (NHL) to determine associations between pesticides found in used vacuum cleaner bags and NHL. Carpet dust samples were collected and analyzed for 30 pesticides in the homes of subjects in across the United States (Detroit, Iowa, Los Angeles, and Seattle).

The laboratory measurements were subject to missing data, primarily due to concentrations being below the minimum detection level (20.8 ng/g for both PCB 180 and γ −Chlordane). In this paper, we considered a total of 1,180 subjects, 676 (57%) cases and 511 (43%) controls; the analytes where chosen due to the number of observations below the LOD, with PCB 180 having 75% values subject to the LOD, and γ −Chlordane 38% values below the LOD. Further details of the study design can be found in Colt et. al.[10]

Table 6 presents estimates of β_1 for PCB 180 obtained through the evaluated methodologies, adjusting for site, gender, education, and age, as in Colt et. al.[10] We find that the complete case analysis method and the missing indicator method ($\beta_2 = 0.51$, non-significant) provide very similar results, whereas the substitution method, the Cox regression, the maximum-likelihood, and the multiple imputation approach provide estimates that are different both in magnitude and in sign (complete case and missing indicator present negative estimates, and the other methods positive estimates). Interestingly, we find that the Cox regression is the only method that provides a significant result. However, we have concerns about interpreting this finding due to empirical evidence that the proportional hazard assumption is not met. Figure A1. in the appendix presents the Kaplan-Meier curve for this scenario.

In addition, we used the missing indicator approach to fit a model with both chemicals (PCB 180 and $\gamma-$Chlordane). We used a likelihood ratio test to simultaneously test the coefficients β_1 and β_2, associated with the observed values of the analytes and the missing indicator. We found that PCB 180 was not statistically significant (β_1= -0.062, $\beta_2 = 0.532, \mathrm{p} = 0.116$), and neither was $\gamma-$Chlordane (β_1=0.124, $\beta_2 = -0.554, \mathrm{p} = 0.229$).

Discussion

In this article, we have compared numerous ways in which exposure data with detection limits have been analyzed in the epidemiology literature. Typically, maximum-likelihood or multiple imputation approaches have been used, which make very strong assumptions about the distribution of the exposure variable and the relationship between this variable and the outcome below the LOD. We note that there is extensive literature on additional methods to model exposure measurements below the limit of detection (examples include Hensel[17] and Gillespie et. al,[18] among others). In this article, we show that the regression coefficient β_1 can be severely biased when assumptions about behavior below the detection limit are not met, even with a relatively small proportion of measurements below the LOD. Unfortunately, unless we actually measure values below the LOD, it is not possible to verify the distribution or the relationship with the outcome below the LOD.[20] The multiple imputation method suffers from even more bias than the maximum-likelihood estimator, resulting in a biased estimator even under the correctly specified parametric modeling assumptions. This happens because the imputation method proposed by Lubin et al.[9] does not impute separately by disease status. For many applications, developing an imputation method that includes disease outcome is difficult since the imputation procedure has to be redone whenever the disease outcome is changed, which can be computationally expensive in large epidemiology studies.

Dinse et al.[7] proposed an interesting approach for analyzing data with a LOD that does not make explicit modeling assumptions about the exposure distribution below the detection limit. Their approach reverses the outcome and covariate by treating the analyte subject to a LOD as a left censored variable and disease outcome as a covariate. The interpretation of their proposed odds ratio is the odds of disease when the analyte is at a value of t divided by the odds when the analyte is less than t, which is not the standard odds ratio used for the other methods. Further, Dinse et al.[7] assumed that the proportional hazards assumption is correct. When we generated data from a logistic regression with left censored analyte values, the resulting hazards for cases and controls were not proportional . For hypothesis testing under the scenarios we considered, the type I error rate was inflated relative to the nominal level of 0.05, and the power was reduced as compared with many of the other approaches considered.

Other, more flexible approaches that non-parametrically model the analyte above the detection limit have been proposed.[20] However, even this methodology makes strong unverifiable modeling assumptions for the distribution and effects below the LOD. This approach uses information about the correlation between other covariates and the analyte to flexibly model the analyte below the LOD. In this case, there is very little information of this type when these other covariates have little correlation with the analyte. Further, the assumption that the correlation structure between covariates and the analyte of interest is the same when the analyte is below the LOD as when it is above is itself a strong assumption.

The missing indicator method[14], where the relationship between the analyte and disease risk is modeled only above the detection limit, does not make any assumptions about what happens below the LOD. Although less efficient than the maximum likelihood approach under a correctly specified model, it is highly robust to unverifiable modeling assumptions. For this method, an intercept term at the LOD and a slope parameter are both needed to assess the relationship between the analyte and disease risk. Thus, a test of association between these two variables would require a joint test of two parameters that are equal to zero.

In most practical situations, our results suggest that the missing indicator method may be the best choice for analyzing analytes with detection limits. These results provide further evidence to the findings of Chiou et. al[14], who recommend the Missing Indicator Method for practical use. This missing indicator method can be adapted for settings where multiple laboratories with different LOD values are used. In this case, a separate missing indicator term could be added to account for the lab-specific LOD. Alternatively, one could fit the model for each laboratory and take a weighted average of the laboratory-specific slope estimates (weighted by the inverse variance of each laboratory-specific slope estimate).

For simplicity, our simulations focused on the simple case where, no confounders were included as covariates, no interactions of covariates with the biomarker subject to an LOD, and only a single biomarker is subject to a LOD. We expect similar results for more complex models that may be used in environmental epidemiology. Namely, the missing indicator model will be more robust to unverifiable modeling assumptions than the competing approaches. Further, for the case of multiple biomarkers subject to LODs, the maximum-likelihood and Cox modeling approaches are very difficult to implement. In contrast, the missing indicator model naturally extends to even high-dimensional biomarkers all subject to LOD.

All methods can be extended to multivariate analyte measurements. The missing indicator method is appealing since it is very easy to extend to this case since it naturally extends to even high-dimensional biomarkers all subject to LOD. All that is required is adding two independent variables corresponding to an intercept and slope effect for each biomarker. The maximum-likelihood approach is computationally challenging to adapt to this case, since it requires multivariate integration to evaluate the likelihood. The reversed method described by Dinse et al.[7] requires multivariate survival analysis techniques to implement in this situation, and this is stated as an area of future research by the authors.

Acknowledgement

We thank the Center for Information Technology, National Institutes of Health, for providing access to the Biowulf cluster computer system and its high-performance computational capabilities. We also thank the associate editor and reviewers for their valuable contributions to this manuscript.

References

1. Browne RW and Whitcomb BW. Procedures for Determination of Detection Limits: Application to High-performance Liquid Chromatography Analysis of Fatsoluble Vitamins in Human Serum. *Epidemiology.* 2010;21(suppl 4):S4-S9.
2. Nie L, Bhu H, Liu C, Cole SR, Vexler A, Schisterman EF. Linear regression with an independent variable subject to a detection limit. *Epidemiology*. 2010;21(suppl4):S17:S24
3. Richardson DB, Ciampi A. Effects of Exposure Measurement Error When an Exposure Variable Is Constrained by a Lower Limit. *American Journal of Epidemiology*. 2003;157(4):355-363
4. Schisterman EF, Vexler A, Whitcomb BW, Liu A. The limitations due to exposure detection limits for regression models. *American journal of epidemiology*. 2006;163(4):374-383
5. Little RJ and Rubin DB. *Statistical Analysis with Missing Data*. New York: John Wiley and Sons.; 1989.
6. Finkelstein MM, Verma DK. Exposure estimation in the presence of nondetectable values: another look. *AIHAJ-American Industrial Hygiene Association*. 2001;62(2):195-198
7. Dinse GE, Jusko TA, Ho LA, et al.; Accommodating measurements below a limit of detection: a novel application of Cox regression. *American Journal of Epidemiology*. 2014;179(8):1018-1024
8. National Cancer Institute-Surveillance Epidemiology and End Results Program (NCI-SEER) study sites. http://seer.cancer.gov/

9. Lubin JH, Colt JS, Camann D, et al.; Epidemiologic evaluation of measurement data in the presence of detection limits. *Environmental health perspectives*. 2004;112(17):1691-1696.

10. Colt JS, Lubin J, Camann D, et al.; Comparison of pesticide levels in carpet dust and self-reported pest treatment practices in four US sites. *Journal of Exposure Science and Environmental Epidemiology*. 2004;14(1):74.

11. Taylor DJ, Kupper LL, Rappaport SM, and Lyles RH. A Mixture Model for Occupational Exposure Mean Testing with a Limit of Detection. *Biometrics*. 2001;57 681-688.

12. Tobin J. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Economietric Society.* 1958;1:24-36.

13. Amemiya T. Tobit models: A survey. *Journal of econometrics. 1984;1:* 24(1-2):3-61

14. Chiou SH, Betensky RA, Balasubramanian R. The missing indicator approach for censored covariates subject to limit of detection in logistic regression models. *Annals of epidemiology*. 2019 ;1(38):57-64.

15. Therneau TM. A package for survival analysis in S. R package version 2.38. 2015. Available at: https://CRAN.R-project.org/package=survival. Accessed spring 2017.

16. Czarnota J, Gennings C, Colt JS, et al.; Analysis of environmental chemical mixtures and non-Hodgkin lymphoma risk in the NCI-SEER NHL study. *Environmental health perspectives*. 2015;123(10):965-970.

17. Helsel DR. *Nondetects and data analysis. Statistics for censored environmental data.* Wiley-Interscience; 2005.

18. Gillespie BW, Chen Q, Reichert H, et al.; Estimating population distributions when some data are below a limit of detection by using a reverse Kaplan-Meier estimator. *Epidemiology*. 1010;S64-70

19. Guo Y, Harel O, Little RJ. How well quantified is the limit of quantification?. *Epidemiology*. 2010;1:S10-6.

20. Kong S, Nan B. Semiparametric approach to regression with a covariate subject to a detection limit. *Biometrika*. 2016;103(1):161-74.

A. Appendix

A.1 Details on the Cox Regression Approach

Logistic regression assumes

$$\text{logit}P(Y = 1) = \beta_0 + \beta_1 X,$$

which is equivalent to the density ratio model:

$$\frac{f(X|T=1)}{f(X|T=0)} = \exp(\beta_0^* + \beta_1 X),$$

with $\beta_0^* = \beta_0 - \log\frac{P(Y=1)}{P(Y=0)}$. On the other hand, a Cox regression model for X assumes that

$$\frac{h(X|T=1)}{h(X|T=0)} = \exp(\gamma_1).$$

As pointed out by Dinse et al. (2014), the interpretation of γ_1 is a log odds ratio of the outcome comparing a subject with $X = x$ versus all subjects with $X > x$, i.e.,

$$\exp(\gamma_1) = \frac{\Pr(y_i = 1|T = t) \,/\, \Pr(y_i = 0|T = t)}{\Pr(y_i = 1|T > t) \,/\, \Pr(y_i = 0|T > t)}.$$

In order for the Cox regression and the logistic regression to be compatible with each other and that $\beta_1 = -\gamma_1$, we need

$$S(X|Y = 1) = S(X|Y = 0)\exp(\beta_0^* + \beta_1 + \beta_1 X)$$

for all values of X. Taking the derivative of X on both sides of the equation yields

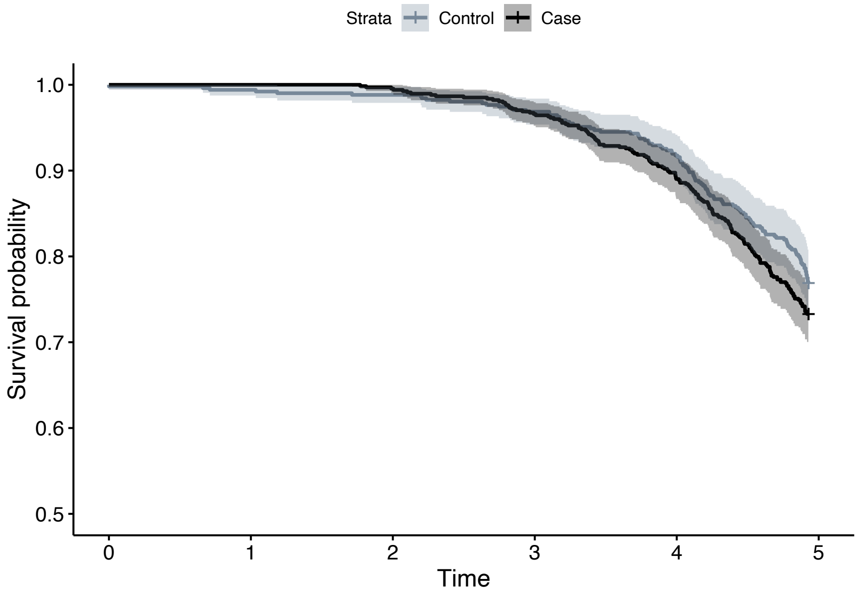
$$-f(X|Y = 1) = \beta_1 S(X|Y = 0)\exp(\beta_0^* + \beta_1 + \beta_1 X) - f(X|Y = 1)\exp(\beta_0^* + \beta_1 + \beta_1 X),$$

or equivalently

$$1 = \exp(\beta_1) - \beta_1 \exp(\beta_1)\frac{s(X|Y = 0)}{f(X|Y = 0)}.$$

The above equation holds if and only if either $\beta_1 = 0$ or $h(X|Y = 0) = \frac{\beta_1 \exp(\beta_1)}{\exp(\beta_1) - 1}$.

Figure A.1 Kaplan-Meier plot of Cases and Controls for the NCI-SEER NHL Study



A.2 Simulation Details and Results : Case 3

In this scenario, the effect of the analyte on the outcome remains the same above and below the LOD, but the distribution of analyte values is Uniform$[l, u]$ below the LOD and normal$(0, \sigma^2)$ above the LOD. The uniform parameters were selected so that the area under the distribution curve is equal to one. Table A presents the uniform distribution parameters used in the simulation.

Table A: Uniform distribution parameters for simulation case 3

	$\sigma^2 = 2.45$		$\sigma^2 = 1$	
	l	u	l	
LOD=0.2	-2.62	4.03	-2.10	7.05
LOD=0.25	-2.48	3.68	-1.93	4.62
LOD=0.45	-2.16	2.31	-1.53	1.92
LOD=0.95	-1.96	1.96	-1.26	1.25

Tables and Figures

Table 1: Simulation results for Case 1: Correct Model Specification. Coefficients represent the average β_1 over the 1000 datasets, the s.e. corresponds to the average s.e over the 1000 datasets, and MC s.e. corresponds to the standard deviation of β_1 over the 1000 datasets.

Variance = 2.45	LOD = 0.2 (16% under)			LOD=0.25 (20% under)			LOD=0.45 (30% under)			LOD=0.95 (50% under)		
	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.
True exposure	0.95	0.06	0.06	0.95	0.06	0.06	0.95	0.06	0.06	0.95	0.06	0.06
Maximum likelihood	0.95	0.05	0.05	0.95	0.05	0.05	0.95	0.05	0.05	0.95	0.06	0.06
Multiple imputation	0.95	0.06	0.06	0.95	0.06	0.06	0.93	0.06	0.06	0.82	0.07	0.06
Cox regression	0.84	0.05	0.05	0.85	0.05	0.05	0.89	0.05	0.06	1.02	0.06	0.07
Complete case analysis	0.95	0.06	0.06	0.95	0.07	0.07	0.95	0.08	0.08	0.96	0.11	0.11
Fill-in LOD/√2	0.99	0.06	0.06	1.01	0.06	0.06	1.106	0.07	0.07	1.33	0.09	0.09
Missing indicator	0.95	0.06	0.06	0.95	0.07	0.07	0.95	0.08	0.08	0.96	0.11	0.11

Variance = 1	LOD = 0.2 (6% under)			LOD= 0.25 (8% under)			LOD=0.45 (20% under)			LOD=0.95 (47% under)		
	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.
True exposure	0.95	0.05	0.05	0.95	0.05	0.05	0.95	0.05	0.05	0.95	0.05	0.05
Maximum likelihood	0.95	0.05	0.05	0.95	0.05	0.05	0.95	0.05	0.05	0.95	0.05	0.05
Multiple imputation	0.95	0.05	0.05	0.94	0.05	0.05	0.91	0.05	0.05	0.81	0.06	0.05
Cox regression	1.25	0.05	0.06	1.26	0.05	0.06	1.31	0.05	0.06	1.45	0.06	0.06
Complete case analysis	0.95	0.05	0.05	0.95	0.06	0.06	0.96	0.07	0.06	0.96	0.09	0.09

Fill-in LOD/√2	1.05	0.05	0.05	1.07	0.05	0.05	1.17	0.06	0.06	1.36	0.07	0.08
Missing Indicator	0.95	0.05	0.05	0.95	0.06	0.06	0.96	0.07	0.06	0.96	0.09	0.09

Table 2: Simulation results for Case 2: Correct Model Specification when $\beta_1 \geq$ LOD and no effect when $\beta_1 <$ LOD. Coefficients represent the average β_1 over the 1000 datasets, the s.e. corresponds to the average s.e over the 1000 datasets, and MC s.e. corresponds to the standard deviation of β_1 over the 1000 datasets.

Variance = 2.45	LOD = 0.2 (16% under)			LOD=0.25 (20% under)			LOD=0.45 (30% under)			LOD=0.95 (50% under)		
	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.
True exposure	0.56	0.04	0.04	0.54	0.04	0.04	0.48	0.03	0.04	0.46	0.03	0.04
Maximum likelihood	0.62	0.04	0.05	0.59	0.04	0.05	0.53	0.04	0.04	0.56	0.04	0.05
Multiple imputation	0.58	0.04	0.04	0.54	0.04	0.04	0.48	0.04	0.03	0.46	0.04	0.03
Cox regression	0.67	0.05	0.06	0.65	0.05	0.06	0.67	0.05	0.06	0.88	0.06	0.06
Complete case analysis	0.95	0.05	0.05	0.95	0.05	0.06	0.96	0.07	0.06	0.96	0.09	0.09
Fill-in LOD/√2	0.79	0.04	0.04	0.78	0.05	0.04	0.80	0.05	0.05	0.94	0.06	0.06
Missing indicator	0.95	0.05	0.05	0.95	0.06	0.06	0.96	0.07	0.06	0.96	0.09	0.09
Variance = 1	LOD = 0.2 (6% under)			LOD= 0.25 (8% under)			LOD=0.45 (20% under)			LOD=0.95 (47% under)		
	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.
True exposure	0.76	0.05	0.06	0.70	0.05	0.06	0.55	0.05	0.05	0.486	0.05	0.05
Maximum likelihood	0.85	0.06	0.06	0.74	0.06	0.07	0.60	0.06	0.06	0.63	0.06	0.07
Multiple imputation	0.76	0.06	0.06	0.70	0.06	0.06	0.55	0.05	0.05	0.488	0.06	0.0495
Cox regression	0.57	0.05	0.06	0.52	0.05	0.06	0.44	0.05	0.05	0.593	0.06	0.06
Complete case analysis	0.95	0.06	0.06	0.95	0.07	0.07	0.95	0.08	0.08	0.949	0.11	0.11
Fill-in LOD/√2	0.86	0.06	0.06	0.84	0.06	0.06	0.80	0.06	0.06	0.93	0.08	0.08
Missing indicator	0.95	0.06	0.06	0.95	0.07	0.07	0.95	0.08	0.08	0.95	0.11	0.11

Table 3: Empirical Type I error rates and power for the case where we have 30% values under LOD, under a correctly specified model (case 1), and a model in which the we have a correct model specification when $t_i \geq$LOD and no effect when $t_i <$LOD (case 2) for several values of β_1. Note the test for the missing indicator approach is a two-degree of freedom test ($H_o: \beta_1 = \beta_2 = 0$).

	Case 1: Correctly Specified Model				Case 2: No effect under LOD			
	Type I rate β_1=0	Power β_1=0.2	Power β_1=0.15	Power β_1=0.1	Type I rate β_1=0	Power β_1=0.2	Power β_1=0.15	Power β_1=0.10
True Exposure	0.052	1.00	0.996	0.875	0.052	0.925	0.724	0.409
Maximum Likelihood	0.046	1.00	0.993	0.879	0.055	0.949	0.807	0.474
Multiple Imputation	0.043	1.00	0.993	0.840	0.043	0.931	0.711	0.372
Cox Regression	0.057	1.00	0.991	0.8223	0.057	0.757	0.529	0.264
Complete Case Analysis	0.046	0.967	0.978	0.982	0.046	0.967	0.798	0.482
Fill in LOD/√2	0.050	1.00	0.991	0.815	0.05	0.982	0.844	0.517
Missing Indicator	0.047	1.00	0.988	0.788	0.047	0.965	0.781	0.439

Table 4: Simulation results for Case 3: Uniform Distribution Under LOD. Correct Model Specification. Coefficients represent the average β_1 over the 1000 datasets, the s.e. corresponds to the average s.e. over the 1000 datasets, and MC s.e. corresponds to the standard deviation of β_1 over the 1000 datasets.

Variance = 2.45	LOD = 0.2 (16% under)			LOD=0.25 (20% under)			LOD=0.45 (30% under)			LOD=0.95 (50% under)		
	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.
True exposure	0.94	0.06	0.06	0.94	0.06	0.06	0.93	0.06	0.05	0.92	0.06	0.05
Maximum likelihood	0.93	0.05	0.04	0.91	0.05	0.05	0.92	0.05	0.05	0.91	0.05	0.06
Multiple Imputation	0.94	0.06	0.06	0.94	0.06	0.06	0.94	0.06	0.06	0.86	0.07	0.06
Cox regression	0.84	0.05	0.05	0.83	0.05	0.05	0.91	0.05	0.05	1.07	0.06	0.06
Complete case analysis	0.94	0.06	0.06	0.94	0.06	0.06	0.92	0.08	0.07	0.89	0.11	0.10
Fill-in LOD/√2	0.98	0.06	0.06	0.98	0.06	0.06	1.10	0.07	0.06	1.34	0.09	0.08
Missing Indicator	0.94	0.06	0.06	0.94	0.06	0.06	0.92	0.08	0.07	0.89	0.11	0.20
Variance = 1	LOD = 0.2 (6% under)			LOD= 0.25 (8% under)			LOD=0.45 (20% under)			LOD=0.95 (47% under)		
	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.	β_1	s.e.	MC s.e.
True exposure	0.94	0.06	0.06	0.94	0.06	0.06	0.93	0.06	0.06	0.91	0.06	0.05
Maximum likelihood	0.99	0.06	0.06	0.98	0.06	0.06	0.93	0.06	0.07	0.94	0.07	0.07
Multiple imputation	0.93	0.06	0.06	0.92	0.06	0.06	0.87	0.06	0.06	0.72	0.07	0.05
Cox regression	0.82	0.05	0.05	0.82	0.05	0.05	0.83	0.05	0.05	0.91	0.06	0.06

Complete case analysis	0.94	0.06	0.06	0.93	0.07	0.06	0.92	0.08	0.07	0.89	0.11	0.10
Fill-in LOD/√2	0.97	0.06	0.06	0.99	0.06	0.06	1.05	0.07	0.06	1.20	0.09	0.08
Missing indicator	0.94	0.06	0.06	0.93	0.07	0.06	0.92	0.08	0.07	0.89	0.11	0.10

Table 5: Simulation results for Case 4: Mixture distribution of the analyte. Coefficients represent the average β_1 over the 1000 datasets, the s.e. corresponds to the average s.e. over the 1000 datasets, and the MC s.e. corresponds to the standard deviation of β_1 over the 1000 datasets.

	θ = 0.25			*θ = 0.5*		
	β_1	*s.e.*	*MC s.e*	β_1	*s.e.*	*MC s.e*
True Exposure	*0.954*	*0.055*	*0.055*	*0.952*	*0.068*	*0.069*
Maximum Likelihood	*0.975*	*0.057*	*0.069*	*0.972*	*0.070*	*0.085*
Multiple Imputation	*0.958*	*0.056*	*0.056*	*0.956*	*0.068*	*0.070*
Cox Regression	*1.825*	*0.060*	*0.059*	*2.442*	*0.075*	*0.073*
Complete Case Analysis	*0.954*	*0.055*	*0.055*	*0.952*	*0.068*	*0.069*
Fill in LOD/2	*0.954*	*0.055*	*0.055*	*0.952*	*0.068*	*0.069*
Missing *Indicator*	*0.954*	*0.055*	*0.055*	*0.952*	*0.068*	*0.069*

Table 6: NCI-SEER NHL Study Data Estimates

	PCB 180	
	β_1	s.e.
Maximum likelihood	0.005	0.039
Multiple imputation	0.036	0.029
Cox regression	0.277	0.120
Complete case analysis	-0.050	0.140
Fill-in LOD/√2	0.112	0.068
Missing indicator	-0.053	0.140
Missing Indicator β_2	0.511	0.576

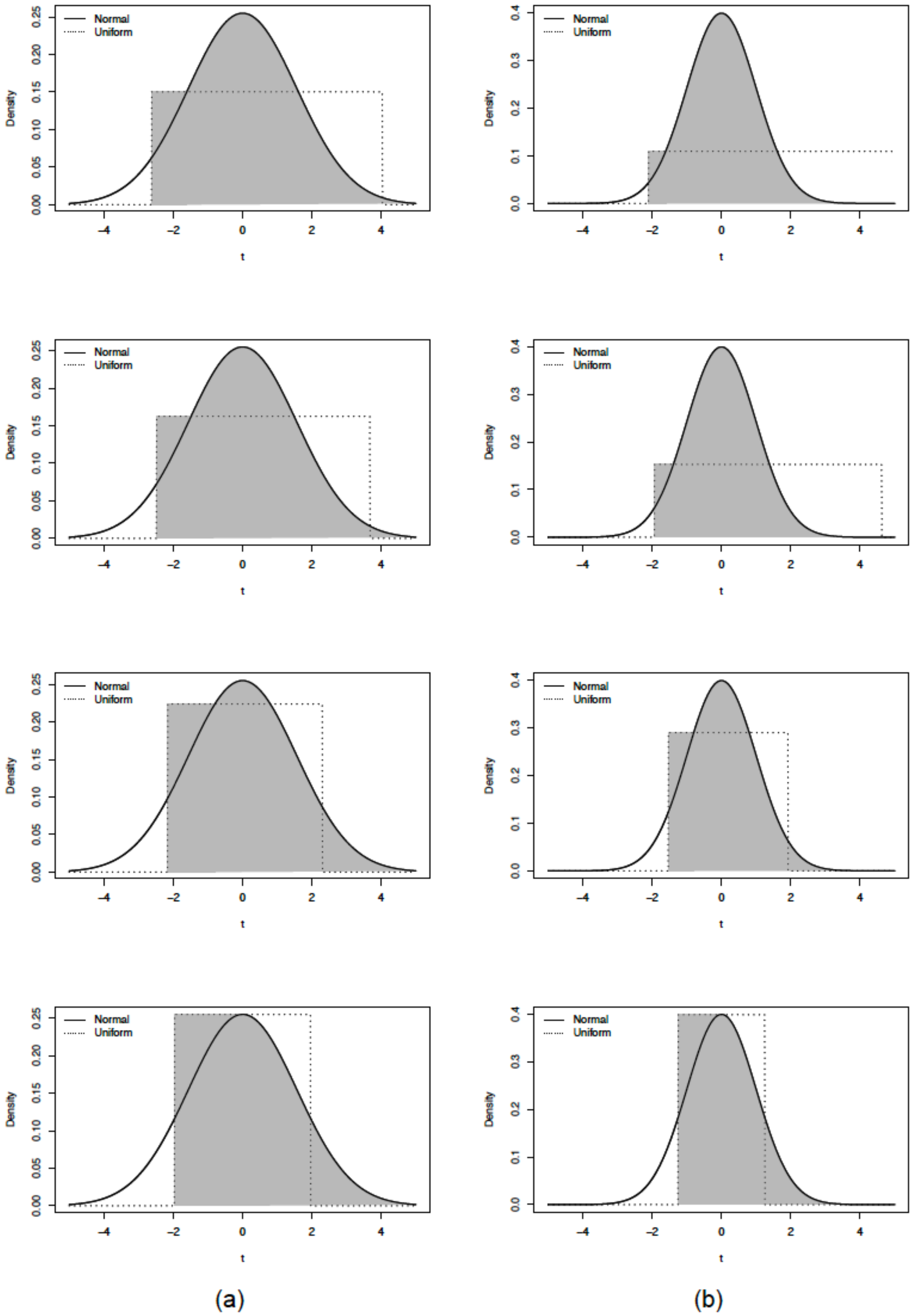


Figure 1: Simulation Case 3: Uniform distribution under LOD. The solid line represents the normal distribution, the dashed line represents the uniform distribution, and the shaded area corresponds to the distribution from which the analyte was generated . (a) and (b) correspond to simulations where $\sigma^2 = 2.45$ and $\sigma^2 = 1$, respectively.