8-15-2017

Estimation of Exposure Distribution Adjusting for Association Between Exposure Level and Detection Limit

Yuchen Yang
University of Kentucky, yuchen.y@uky.edu

Brent J. Shelton
University of Kentucky, brent.shelton@uky.edu

Thomas Tucker
University of Kentucky, thomas.tucker@uky.edu

Li Li
Case Western Reserve University

Richard Kryscio
University of Kentucky, kryscio@uky.edu

See next page for additional authors

Authors
Yuchen Yang, Brent J. Shelton, Thomas Tucker, Li Li, Richard Kryscio, and Li Chen

Estimation of Exposure Distribution Adjusting for Association Between Exposure Level and Detection Limit

Notes/Citation Information
Published in *Statistics in Medicine*, v. 36, issue 18, p. 2935-2946.

Digital Object Identifier (DOI)
https://doi.org/10.1002/sim.7335

Published in final edited form as:
Stat Med. 2017 August 15; 36(18): 2935–2946. doi:10.1002/sim.7335.

Estimation of Exposure Distribution Adjusting for Association between Exposure Level and Detection Limit

Yuchen Yang[a], **Brent J. Shelton**[b,c], **Thomas T. Tucker**[c], **Li Li**[d], **Richard Kryscio**[a,b], and **Li Chen**[b,c,*]

[a]Department of Statistics, University of Kentucky, Lexington, KY, USA

[b]Department of Biostatistics, University of Kentucky, Lexington, KY,USA

[c]Markey Cancer Center, University of Kentucky, Lexington, KY, USA

[d]Departments of Family Medicine, Epidemiology, and Biostatistics, Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA

Abstract

In environmental exposure studies, it is common to observe a portion of exposure measurements to fall below experimentally determined detection limits (DLs). The reverse Kaplan-Meier (RKM) estimator, which mimics the well-known Kaplan-Meier estimator for right-censored survival data with the scale reversed, has been recommended for estimating the exposure distribution for the data subject to DLs because it does not require any distributional assumption. However, the RKM estimator requires the independence assumption between the exposure level and DL and can lead to biased results when this assumption is violated. We propose a kernel-smoothed nonparametric estimator for the exposure distribution without imposing any independence assumption between the exposure level and DL. We show the proposed estimator is consistent and asymptotically normal. Simulation studies demonstrate that the proposed estimator performs well in practical situations. A colon cancer study is provided for illustration.

Keywords

detection limits; left-censored data; environmental exposure; kernel smoothing; nonparametric estimator

1. Introduction

In environmental exposure studies, one fundamental question is to estimate distributions of environmental chemicals, such as trace elements and pesticides, in a certain population. However, it is very common to observe a portion of exposure measurements to fall below experimentally determined detection limits (DLs). A detection limit (DL) is "a threshold below which measured values are not considered significantly different from a blank signal, at a specified level of probability" [1]. Therefore, the exposure level of a chemical for a

[*] Correspondence to: Li Chen, Biostatistics and Bioinformatics Shared Resource Facility, Markey Cancer Center and Department of Biostatistics, College of Public Health, University of Kentucky, Lexington, USA. Tel: 859-323-2005.

sample is only reported when its value is not less than the DL and otherwise is reported as a less than value or non-detect. The DL itself can depend on the mass/volume of the analyzed sample and/or on the mass/volume of adjustment factors such as lipid content. The laboratory may report a common DL for all samples or different DLs for different samples. When the latter occurs, it is mostly because the mass/volume of the obtained sample and/or any adjustment factor differs for each individual, and the exposure level and DL may be associated in this case. For example, in the colon cancer study measuring trace element accumulation in toenails [2], we observed a statistically significant association between the exposure level and DL in Appalachian cancer cases for at least 6 trace elements (Table 4). This may be because trace elements can cause adverse effects on metabolism and therefore lead to slow growth rate of toenails [3]. As a result, toenail samples obtained from individuals with high exposure to trace elements tend to have low masses, and thus high DLs. Conversely, a higher toenail mass results in a lower DL (i.e., a better ability to detect low levels of metal accumulation). In this situation, the exposure level and DL are positively associated whereas both are negatively associated with the toenail sample mass.

Ad hoc methods, such as substituting DL, DL/2, or DL/√2 for the value below a DL, are widely used in environmental science literature to estimate the exposure distribution for the data subject to DLs. However, these methods have no theoretical basis and are ill-advised unless relatively few measures fall below DLs [4; 5]. To appropriately account for values below DLs, parametric models for left-censored data, such as the lognormal model [1], can be used since the data subject to DLs can also be treated as left-censored data [1]. But these parametric methods can lead to markedly biased results when the parametric form of the exposure distribution is misspecified [1; 5]. Recently nonparametric methods have received increasing attention because they do not require distributional assumptions, and thus may be a safer choice for data analysis. The reverse Kaplan-Meier (RKM) estimator, which mimics the Kaplan-Meier (KM) estimator for right-censored survival data with the scale reversed, has been recommended [6]. Note that both the RKM estimator and the aforementioned parametric methods require the independence assumption between the exposure level and DL. To our knowledge, there are no appropriate statistical methods available to deal with the case when the exposure level and DL are associated.

In this paper, we utilize a two-step strategy and the kernel smoothing technique to develop a nonparametric consistent estimator for the exposure distribution allowing for the situation that the exposure level and DL are not independent. We first estimate the conditional exposure distribution given the DL by adding kernel weights into the RKM estimator and then obtain the average of the estimated conditional distributions over all DL values in the sample to estimate the marginal exposure distribution. The proposed method does not require any independence assumption between the exposure level and DL and any distributional assumption about the exposure level. In Section 2, we propose the estimator and show that it is consistent and converges weakly to a Gaussian process. In Section 3, the results of several simulation studies are reported to compare the performance of the proposed estimator to both the RKM estimator and a parametric estimator assuming a lognormal exposure distribution. In Section 4, a colon cancer study is provided for illustration. Finally, Section 5 contains discussions and some extensions.

2. Methods

Let $\tilde{T}$ and D be random variables for the exposure level and DL, respectively, and $F(\cdot)$ be the cumulative distribution function (CDF) of the exposure level. Let $T = \mathrm{m}ax(\tilde{T}, D)$ and $\delta = I(\tilde{T} \geq D)$. Here δ indicates whether T is an exposure level value or a DL value. For data subject to DL, only (T, δ, D) are observable for each subject. Suppose the data consist of n replicates $\{(T_i, \delta_i, D_i): i = 1, \cdots, n\}$. Note that the method proposed below requires the DL to be known for each subject in the data.

It is useful to adopt the counting process notation. Analogous to the observed counting process and at-risk process for right censored survival data, we define two counting processes, $N_i(t) = I(T_i \leq t, \delta_i = 1)$ and $Y_i(t) = I(T_i \leq t)$, for the data subject to DLs. Then the RKM estimator can be rewritten as

$$\hat{F}_{\mathrm{RKM}}(t)=\prod_{s>t}\left\{1-\frac{\sum_{j=1}^{n} dN_j(s)}{\sum_{j=1}^{n} Y_j(s)}\right\}, t \geq \tau_n, \qquad (1)$$

where $\tau_n = \min_{i=1,\ldots,n}\{T_i\}$. In addition, when the smallest observation is uncensored, $\hat{F}_{RKM}(t) = 0$ for $t \in (0, \tau_n)$. When the smallest observation is censored, $\hat{F}_{RKM}(t)$ is undefined for $t \in (0, \tau_n)$. This estimator mimics the KM estimator for right-censored survival data with the scale reversed. Similar to the independence assumption between the survival time and censoring time for the KM estimator, the RKM estimator requires the independence assumption between the exposure level and DL and is not a consistent estimator when this assumption is violated.

To develop a consistent estimator for the exposure distribution allowing for the association between the exposure level and DL, we propose a two-step strategy based on the statistical fact that $F(t) = E_D\{F(t; D)\}$, where $F(t; d)$ is the conditional CDF of the exposure level given the DL, i.e. $F(t; d) = \Pr(\tilde{T} \leq t \mid D = d)$, and E_D is the expectation with respect to D. In the first step, we obtain a consistent estimator for the conditional CDF of the exposure level, denoted by $\hat{F}(t; d)$. Specifically, we estimate the conditional CDF by adding kernel weights into the RKM estimator in equation (1) such that subjects whose DL values closest to d receive the largest weights, i.e.

$$\hat{F}(t;d)=\prod_{s>t}\left[1-\frac{\sum_{j=1}^{n} K\{(D_j-d)/h\} dN_j(s)}{\sum_{j=1}^{n} K\{(D_j-d)/h\} Y_j(s)}\right], t \geq \tau_n,$$

where $K(\cdot)$ is a kernel function, and h is a bandwidth such that $nh \to \infty$ and $nh^4 \to 0$ as $n \to \infty$. For each value of d, all subjects contribute to the calculation of $\hat{F}(t; d)$ but with different weights depending on the difference between their DL values and d. In the second step, we estimate $F(t)$ by the average of estimated conditional CDF values over all DL values in the sample, i.e. $\hat{F}(t)=n^{-1}\sum_{i=1}^{n}\hat{F}(t;D_i)$. Similar to the RKM estimator, the proposed estimator $\hat{F}(t)$ is a right-continuous step function with jumps at uncensored observations. When the smallest observation is uncensored, $\hat{F}(t; d) = 0$ and $\hat{F}(t) = 0$ for $t \in (0, \tau_n)$. When the smallest observation is censored, $\hat{F}(t; d)$ and $\hat{F}(t)$ are undefined for $t \in (0, \tau_n)$. The above estimator for the conditional CDF borrows the idea of the kernel conditional KM estimator which added kernel weights into the KM estimator to estimate the conditional survival function for right-censored survival data [7]. In the following, the proposed estimator $\hat{F}(t)$ will be referred to as KRKM estimator. Through the above two-step strategy, in order for $\hat{F}(t)$ to be a consistent estimator for the marginal CDF of the exposure level, we only need the estimator for the conditional CDF given the DL to be a consistent estimator, which only requires the conditional independence between the exposure level and DL given the DL. Since it is true that the exposure level and DL are independent given the DL, the KRKM estimator is consistent without requiring any independence assumption between the exposure level and DL. We show in Appendix A that $\sqrt{n}\{\hat{F}(t) - F(t)\}$ converges weakly to a zero-mean Gaussian process and is asymptotically equivalent to the process $n^{-1/2}\sum_{i=1}^{n}\xi_i(t)$, where

$$\xi_i(t)=F(t;D_i) - F(t) - F(t;D_i)\left\{\frac{\delta_i I(T_i \geq t)}{F(T_i;D_i)}+1 - \frac{1}{F(\max(T_i,t);D_i)}\right\}. \qquad (2)$$

The above theoretic result does not require the kernel function to have any special shape. But numerically, because the kernel function appears in the denominator of the proposed estimator, standard kernel functions, such as Gaussian kernel with fixed standard deviation and Triangular kernel, can produce extremely small kernel weights and thus cause unstable results. Therefore, to ensure computational stability, we suggest using the following modified Silverman kernel [8], which is flatter and less likely to produce extremely small kernel weights,

$$K(u)=\frac{\left|\frac{1}{2}e^{\frac{-|u|}{\sqrt{2}}}\sin\left(\frac{|u|}{\sqrt{2}}+\frac{\pi}{4}\right)\right|}{\int_{-\infty}^{\infty}\left|\frac{1}{2}e^{\frac{-|u|}{\sqrt{2}}}\sin\left(\frac{|u|}{\sqrt{2}}+\frac{\pi}{4}\right)\right|\mathrm{du}}.$$

For the bandwidth, we suggest using $\hat{\sigma}n^{-1/3}$, where $\hat{\sigma}^2$ is the sample variance of the DL. This choice satisfies the conditions that $nh \rightarrow \infty$ and $nh^4 \rightarrow 0$ as $n \rightarrow \infty$. Based on the formula in (2), the variance of the KRKM estimator can be estimated by $n^{-2}\sum_{i=1}^{n}\hat{\xi}_i^2(t)$, where $\hat{\xi}_i(t)$ is obtained by replacing $F(\cdot; D_i)$ and $F(\cdot)$ by $\hat{F}(\cdot; D_i)$ and $\hat{F}(\cdot)$. The log-log transformed 95% confidence intervals for $F(t)$ can then be calculated as that for the survival function in survival analysis. This will be referred to as formula-based variance estimation method. Another approach to estimate the variance is to use the bootstrap method. Similar log-log transformed 95% confidence intervals can be obtained. This approach will be referred to as bootstrap-based variance estimation method. The formula-based variance estimation method is computationally faster than the bootstrap-based method, but may underestimate the variance and thus yield poor coverage probabilities at the points below which there are few observations, as shown in simulation studies of Section 3.

We have implemented the proposed methods in an R package called “KENDL”, which is now available at the Comprehensive R Archive Network. (https://cran.r-project.org/web/packages/KENDL/index.html).

3. Simulation studies

To assess the performance of the proposed KRKM estimator under the situation that the exposure level and DL are associated, we mimicked the cadmium (Cd) and nickel (Ni) data in Appalachian cases from the colon cancer study in Section 4. We generated the DL for each trace element based on their empirical distributions in the data and the exposure level for each trace element from the lognormal regression model: $\log(\tilde{T}) = \mu + \beta \log(D) + \sigma\varepsilon$, where ε follows a standard normal distribution. The parameters μ, β, σ are estimated based on the data for each trace element, which are −3.05, 0.42, and 1.21 for Cd (setting 1) and 0.16, 0.34, and 1.62 for Ni (setting 2). The non-detect rate of the simulated data is 76% and 25% for the above two settings, respectively. We compared the KRKM estimator, with both bootstrap-based and formula-based variance estimation, to the RKM estimator and the parametric estimator assuming a lognormal distribution for the exposure level. The latter two estimators were obtained from NADA R package. Note that the NADA package has an error in reporting the RKM estimator for the CDF, which is described below. Let $t_{(1)} < \cdots < t_{(m)}$ be the ordered unique uncensored exposure values. The RKM estimator at a given uncensored exposure value $t_{(i)}$, i.e. $\hat{F}_{RKM}(t_{(i)})$, should be the NADA-reported CDF estimate at $t_{(i+1)}$. Therefore, we corrected this error by letting $\hat{F}_{RKM}(t_{(i)})$ be the NADA-reported CDF estimate at $t_{(i+1)}$. For any $t \in [t_{(i)}, t_{(i+1)})$, we set $\hat{F}_{RKM}(t)$ equal to $\hat{F}_{RKM}(t_{(i)})$ since $\hat{F}_{RKM}(.)$ is a right-continuous step function. Table 1 summarizes the results for the above three estimators of $F(t)$ at t = 1st, 2nd and 3rd quartiles based on 1000 replicates and 500 bootstraps for both settings. The proposed KRKM estimator with the bootstrap-based variance estimation performs very well except for t = 1st quartile in setting 1: the biases are small and the confidence intervals have proper coverage probabilities. At t = 1st quartile in setting 1, the coverage probability is lower than the nominal value due to the very high non-detect rate of 76%. Compared to the bootstrap-based variance estimation, the formula-based variance estimation for the KRKM estimator is computationally faster. But at the points below which there are few observations, e.g. t = 1st and 2nd quartiles in setting 1, the formula-based variance estimation tends to underestimate the variance and thus yield poor coverage probabilities. In contrast to the KRKM estimator, the RKM estimator has large biases and poor coverage probabilities, especially when the sample size increases, due to its inability to account for the association between the exposure level and DL. Likewise, the lognormal estimator also has large biases and low coverage probabilities, resulting from not accounting for the association between the exposure level and DL and possibly misspecified exposure distribution. To further unravel the impact of not accounting for the association between the exposure level and DL for the lognormal estimator, we considered additional simulations where the DL for each trace element was generated from a lognormal distribution with parameters estimated from the colon cancer data. Under this scenario, the marginal distribution of the exposure level is guaranteed to follow a lognormal distribution so that the parametric distribution is correctly specified for the lognormal estimator. However, as shown in Table 2, the lognormal estimator still yields large biases and poor coverage probabilities.

To compare the performance of the KRKM, RKM and lognormal estimators under the situation that the exposure level and DL are independent, we adopted the above set-up but set $\beta = 0$. The non-detect rate of the simulated data is 78% and 31% for the two settings, respectively. Table 3 summarizes the results for the KRKM, RKM and lognormal estimators of $F(t)$ at t = 1st, 2nd and 3rd quartiles based on 1000 replicates and 500 bootstraps. For all the estimators, the biases are very small, the variance estimators are accurate and the confidence intervals have proper coverage probabilities. The KRKM estimator obtains comparable results as the RKM estimator when the exposure level and DL are independent. The lognormal estimator yields slightly smaller variances than the KRKM and RKM estimators, which is expected since the exposure level and DL are independent and the exposure distribution is lognormal under this set-up.

4. Example

Kentucky has the nation's highest colon cancer incidence rate [10]. Appalachian Kentucky, which has a unique geology that contains high-quality bituminous coal naturally rich in trace elements, has an even higher rate of colon cancer compared to other regions of the state. A case-control study was conducted to explore the association between environmental exposures to trace elements such as arsenic (As), chromium (Cr) and nickel (Ni) and colon cancer and whether exposures to these trace elements contribute to the elevated colon cancer rate in Appalachian Kentucky [11; 2]. For this purpose, 274 colon cancer cases and 253 controls were selected from 23 contiguous rural counties in Kentucky (Appalachian region) and Jefferson County, the largest, most urban county in Kentucky (non-Appalachian region). Among 247 subjects from the Appalachian region, 145 were cases and 102 were controls; among 280 from the non-Appalachian region, 129 were cases and 151 were controls. Toenail samples from these subjects were collected, and the concentrations of 12 trace elements were measured as markers of long-term environmental exposures to these trace elements. The DL varies from one subject to another for these trace element concentrations as a function of the toenail mass. For illustration purposes, we only focus on the Appalachian region. The proportion below the DL is over 20% for most trace elements and is as high as 79% and 83% for Cd in Appalachian cases and controls, respectively (Table 4).

We first examine the independence assumption between the exposure level and DL for each trace element using the following three methods. In the first method, we fitted a lognormal accelerated failure time (AFT) model [12] with the left-censored exposure level as the outcome and the log-transformed DL as a covariate. Under this model, the independence assumption between the exposure level and DL was examined by testing whether the coefficent is equal to 0. The Pearson's correlation coefficient between the exposure level and DL (both log-transformed) was estimated by $\hat{\beta}/\sqrt{\hat{\beta}^2+\hat{\sigma}^2/\hat{\sigma}_1^2}$, where $\hat{\beta}$, $\hat{\sigma}$ are the estimators of the coefficient and scale parameters in the lognormal AFT model and σ_1^2 is the sample variance of $\log(D)$. In the second method, the Pearson's correlation coefficient between the exposure level and DL (both log-transformed) and the corresponding p-value were calculated based on the “clikcorr” R package, which assumes a bivariate normal distribution for the two variables and uses a profile likelihood method [13]. In the third method, the nonparametric Kendall's tau correlation coefficient [14] and the corresponding p-value were calculated based on the “cenken” function in the NADA R package [9]. The results based on the above three methods are reported in Table 4. The results from the first two parametric methods are vey close for all trace elements except for Cd in controls, where the non-detect rate is as high as 83%. For colon cancer cases, there is a statistically significant association between the exposure level and DL for all 12 trace elements based on the two parametric methods. The nonparametric Kendall's tau method, which appears more conservative, identifies 6 trace elements with a significant association between the exposure level and DL. For controls, there is only one trace element showing a significant association between the exposure level and DL based on the three methods.

We then use the trace element Ni to demonstrate our proposed KRKM estimator, comparing to the RKM estimator and the parametric estimator. For cases, the Ni level ranges from 0.02 to 624.4 and the DL ranges from 0.004 to 24.84; for controls, the Ni level ranges from 0.04 to 39.37 and the DL ranges from 0.01 to 38.38. Table 4 shows that for Ni there is a signifcant association between the exposure level and DL for cases but no signficant association for controls. We estimated the exposure distributions of Ni level for cases and controls, respectively. The lognormal distribution was selected for the distributions of Ni for both cases and controls by the Akaike information criterion (AIC) [15] among a number of candidate distributions, including normal, lognormal, Weibull and loglogistic. Figure 1 displays the CDF estimates for colon cancer cases and controls based on the KRKM, RKM and lognormal estimators, and Figure 2 displays the differences in CDF estimates between the KRKM estimator and the latter two estimators along with 95% confidence limits. These figures show that the RKM estimator significantly overestimates the CDF for the Ni level between 0.21 and 5.29 compared to the proposed KRKM estimator for cancer cases. This may be because of the significant association between the exposure level and DL. In contrast, there is no significant difference between the two estimators for controls, which may be because of the insignificant association between the exposure level and DL. As a result, the RKM estimator significantly underestimates the difference between the cases and controls compared to the KRKM estimator. Figures 1 and 2 also show remarkable difference between the lognormal and KRKM estimators for cases, most likely due to a combination of imperfect fit of the lognormal distribution and the significant association between the exposure level and DL. The difference between these two estimators is smaller for controls.

5. Discussion

We have developed a consistent nonparametric estimator for the exposure distribution without requiring any independence assumption between the exposure level and DL. Our proposed estimator outperforms the RKM estimator and the parametric estimator when the exposure level and DL are associated because the latter two estimators are not consistent in that situation. In the case of a common DL, our estimator reduces to the RKM estimator; and in the case of varying DLs but the exposure level and DL are independent, our estimator can obtain comparable results as the RKM estimator. Thus, our estimator provides a unified nonparametric approach for estimating the exposure distribution regardless of whether the exposure level and DL are independent or not and whether the association between the exposure level and DL is linear, curvilinear, or step function, etc. Therefore, the user does not have to test whether the exposure level and DL are associated before using our method, which is an advantage over the RKM method whose validity depends on the test results.

We have utilized a two-step strategy and kernel smoothing technique along with a special feature of data subject to DLs, i.e. the DL is observable for each subject, to completely eliminate the independence assumption between the exposure level and DL. In contrast, the consistent estimators developed based on similar two-step strategies for the marginal survival function for right-censored survival data need to find a set of covariates and require the independence assumption between the censoring time and survival time conditional on those covariates [16; 17]. In our approach, we take advantage of the data characteristic that the DL is observable for each subject and utilize the DL as the conditioning covariate. As a statistical fact, the independence assumption between the DL and exposure level given the DL automatically holds. Therefore, our estimator is free of any independence assumption between the exposure level and DL.

In survival analysis, another approach dealing with dependent censoring for estimating the survival function is the inverse probability of censoring weighting (IPCW) KM estimator [18; 19]. This weighted version of the KM estimator assigns a weight, inversely proportional to an estimate of the conditinal survival function of the censoring time given a set of covariates, to each subject. Under the condition that the censoring time and survival time are independent given that set of covariates, the IPCW KM estimator is consistent. By borrowing this idea, one can construct an IPCW RKM estimator for the exposure distribution by adding subject-specific weights, proportional to each subject's conditional CDF of the DL given a set of covariates, in the RKM estimator. The consistency of this estimator requires that the exposure level and DL are independent given that set of covariates. However, it is not possible to use the IPCW method with DL as the covariate to obtain an estimator free of any independence assumption between the exposure level and DL. The conditional CDF of the DL, given DL, can only take values 0 or 1 and thus cannot be used as an inverse weight.

A key issue in our two-step strategy is how to estimate the conditional CDF of the exposure level given the DL for the data subject to DL. To address this issue, we have added kernel weights into the RKM estimator. The use of the kernel technique assures our estimator is purely nonparametric and free of any distributional assumption. Importantly, our estimator does not suffer the curse of dimensionality of the kernel method because we only need to condition on a one-dimensional variable, i.e. the DL, for estimating the conditional CDF. In addition, our estimator is robust to the choice of bandwidth. Besides the bandwidth of $\hat{\sigma} n^{-1/3}$ presented in the paper, we also conducted simulation studies using several other bandwidths including $\hat{\sigma} n^{-7/24}$, $\hat{\sigma} n^{-2/5}$, and $\hat{\sigma} n^{-1/2}$, which yielded very similar results (data not shown). As an alternative to the kernel method, one can use a parametric AFT model with the left-censored exposure value as the outcome and the DL as the covariate to estimate the conditional CDF. Additional simulation studies reveal that this alternative method performs well and has smaller variance than the proposed estimator when the model is correctly specified but can lead biased results when the model is misspecified (data not shown).

In this paper, we highlight the critical need to account for the association between the exposure and DL and the consequences of ignoring it. This problem of association between the exposure and DL may sometimes be alleviated by improving the design of sample collection. For example, the association between the DL and the exposure level in the colon cancer study could have been reduced if toenail samples had been collected from multiple toes or at multiple time points to obtain larger samples and thus lower the DLs. Having equal DLs for all subjects would eliminate any association, and may be feasible in some settings but logistically difficult in others. In presence of varying DLs, appropriate statistical methods should be used to deal with the possible association between the exposure level and DL so that unbiased analysis results can be obtained.

There are at least two extensions of the proposed method. First, the proposed KRKM estimator requires the data come from a simple random sample of the underlying population. One can extend the proposed estimator to survey data by incorporating sampling weights. Second, our estimator can serve as the building block for a formal test to compare the exposure distributions between two groups by considering the cumulative weighted difference in CDF estimates for the two groups, analogous to the weighted KM statistics for right-censored data [20]. However, it will be more complex than the latter because the proposed KRKM estimator is more complicated than the KM estimator. Of further interest is to incorporate the adjustment of confounding factors in the comparison between two groups. Current literature [21; 22] considered logistic regression models with exposure(s) and confounding factors as covariates and the disease status as the outcome and used the maximum likelihood method to make inferences. However, these methods require the independence assumption between the exposure level and DL. One possible approach to account for the association between the exposure level and DL is to use multiple imputation to impute exposure values below DLs based on our kernel-smoothed conditional CDF given the DL. Since our kernel-smoothed conditional CDF is undefined in $(0, \tau_n)$ when the smallest observation is censored, additional distributional assumptions are needed for that region in order to perform the imputation under this situation.

Acknowledgments

This research was supported by the National Cancer Institute under grants R03CA179661 and P30CA177558. The colon cancer study was supported by the National Cancer Institute under grant R01CA136726. The authors thank two referees and an associate editor for their helpful comments.

Appendix A

Weak convergence of $\sqrt{n}\{\hat{F}(t) - F(t)\}$

In this section, we prove the weak convergence of $\sqrt{n}\{\hat{F}(t) - F(t)\}$ through the modern empirical process theory. Let P_n and P denote the empirical measure and the distribution under the true model, respectively. For a measurable function f and measure Q, the integral $\int fdQ$ is abbreviated as Qf. Specifically, $P_n f(T,\delta,D) = n^{-1}\sum_{i=1}^{n} f(T_i,\delta_i,D_i)$, $P\{f(T, \delta, D)$ is the expectation of $f(T, \delta, D)$, and $P\{f(T, \delta, D)|D\}$ is the conditional expectation of $f(T, \delta, D)$ given D. We express $\sqrt{n}\{\hat{F}(t) - F(t)\}$ as

$$\sqrt{n}(P_n-P)\{F(t;D)\}+\sqrt{n}P\{\hat{F}(t;D)-F(t;D)\}+\sqrt{n}(P_n-P)\{\hat{F}(t;D)-F(t;D)\}. \quad (3)$$

To study the second term in (3), we define

$$R(t;d)=\int_t^{\infty}\frac{dF(u;d)}{F(u;d)}.$$

By some algebras we obtain $R(t; d) = -\log F(t; d)$, which is analogous to the conditional cumulative hazard function in survival analysis but with the conditional survival function replaced by the conditional CDF. We first study

$$\hat{R}(t,d)=\int_t^{\infty}\frac{\sum_{j=1}^{n}K\{(D_j-d)/h\}dN_j(s)}{\sum_{j=1}^{n}K\{(D_j-d)/h\}Y_j(s)}.$$

Let $N(t) = I(T \leq t, \delta = 1)$ and $Y(t) = I(T \leq t)$. We express $\hat{R}(t; d) - R(t; d)$ as

$$\begin{aligned}
&P_n\left[\frac{K\{(D-d)/\}\delta I(T\geq t)}{P_n[K\{(D-d)/h\}Y(u)]|_{u=T}}\right]-P\left[\frac{I(\hat{T}\geq t)}{P\{Y(u)|D=d\}|_{u=\tilde{T}}}\,|D=d\right]=(P_n-P)\left[\frac{K\{(D-d)/h\}\delta I(T\geq t)}{P_n(K\{(D-d)/h\}Y(u)|_{u=T})}\right]\\
&-P\left[\frac{K\{(D-d)/h\}\delta I(T\geq t)(P_n-P)(K\{(D-d)/h\}Y(u)|_{u=T})}{P(K\{(D-d)/h\}Y(u)|_{u=T})P_n(K\{(D-d)/h\}Y(u)|_{u=T})}\right]+\Bigg(P\left[\frac{K\{(D-d)/h\}\delta I(T\geq t)}{P[K\{(D-d)/h\}Y(u)|_{u=T}]}\right]-P\left[\frac{I(\tilde{T}\geq t)}{P[Y(u)|D=d\}|_{u=\tilde{T}}}\middle|D=d\right]\Bigg)\\
&=(P_n-P)\left[\frac{K\{(D-d)/h\}\delta I(T\geq t)}{P(K\{(D-d)/h\}Y(u)|_{u=T})}\right]-P\left(\frac{K\{(D-d)/h\}\delta I(T\geq t)(P_n-P)[K\{(D-d)/h\}Y(u)|_{u=T}]}{P^2[K\{(D-d)/h\}Y(u)|_{u=T}]}\right)+\Bigg(P\left[\frac{K\{(D-d)/h\}\delta I(T\geq t)}{P(K\{(D-d)/h\}Y(u)|_{u=T})}\right]\\
&-P\left[\frac{I(\tilde{T}\geq t)}{P\{Y(u)|D=d\}|_{u=\tilde{T}}}\,|D=d\right]\Bigg)+o_p(n^{-1/2}).]
\end{aligned}$$

(4)

It's straightforward to show that the first term on the right side of (4) is equal to

$$(P_n-P)\int_t^{\infty}\frac{[K\{(D-d)/h\}dN(u)]}{P[K\{(D-d)/h\}Y(u)]}.$$

By Lemma 1 and some algebras, the second term on the right side of (4) is equal to

$$(P_n-P)\int_t^{\infty}\frac{K\{(D-d)/h\}Y(u)dR(u;d)}{P[K\{(D-d)/h\}Y(u)]}+O(h^2).$$

By Lemma 1 and the statistical fact that $\tilde{T}$ and D is independent given D, the third term on the right side of (4) can be shown to be $O(h^2)$. Therefore, we obtain that $\hat{R}(t; d) - R(t; d)$ is equal to

$$(P_n - P)\left(K\{(D-d)/h\}\int_t^\infty \frac{dN(u)+Y(u)dR(u;d)}{P[K\{(D-d)/h\}Y(u)]}\right)+O(h^2)+o_p(n^{-1/2}).$$

By the condition that $\sqrt{n}h^2 = o_p(1)$, the Duhamel equation and Lemma 1, we obtain that the second term on the right side of (3) is asymptotically equivalent to

$$\begin{aligned}&\sqrt{n}(P_n-P)\left(P_{D_*}\left[\right.\right.\\ &\quad -F(t;d)K\{(D-d)/h\}\int_t^\infty \frac{dN(u)+Y(u)dR(u;d)}{P[K\{(D-d)/h\}Y(u)]}\Big]\Big|_{d=D^*}=\sqrt{n}(P_n-P)\Big[\\ &\quad -F(t;d)\int_t^\infty \frac{dN(u)+Y(u)dR(u;D)}{P\{Y(u)|D\}}\Big]+o_p(1),\end{aligned}$$

where D^* is a random variable with the same distribution as D, and P_{D^*} denotes expectation only respective to D^*.

Similarly, we can verify that $P\{\hat{F}(t; D) - F(t; D)\}^2 \rightarrow_p 0$ uniformly for $t \in [0, \infty]$ and that $\hat{F}(t; D)$, $F(t; D)$ belong to a P- Donsker class. It then follows that the third term of (3) converges uniformly to zero in probability by Lemma 19.24 of[23].

Combining the aforementioned results, we conclude that $\sqrt{n}(\hat{F}(t) - F(t))$ is asymptotically equivalent to the process

$$\begin{aligned}&\sqrt{n}(P_n-P)\left\{F(t;D)-F(t;D)\int_t^\infty \frac{dN(u)-Y(u)d\log F(u;D)}{F(u;D)I(D\le u)}\right\}=\\ &n^{-1/2}\sum_{i=1}^n\left[F(t;D_i)-F(t)-F(t;D_i)\left\{\frac{\delta_i I(T_i\ge t)}{F(T_i;D_i)}+1-\frac{1}{F(\max(T_i,t);D_i)}\right\}\right].\end{aligned}$$

Lemma 1. Let $f_D(d)$ be the probability density function of D, then

$$\begin{aligned}P[h^{-1}K\{(D-d)/h\}\delta I(T\ge t)]&=P\{\delta I(T\ge t)|D=d\}f_D(d)+O(h^2)\\ P[h^{-1}K\{(D-d)/h\}Y(u)]&=P\{Y(u)|D=d\}f_D(d)+O(h^2)\end{aligned}$$

Proof: We have

$$P[h^{-1}K\{(D-d)/h\}\delta I(T\ge t)]=\int h^{-1}K\{(x-d)/h\}P[\delta I(T\ge t)|D=x]f_D(x)\mathrm{dx}. \quad (5)$$

Let $g(x) = P[\delta I(T \geq t) \mid D = x] f_D(x)$. Using a simple transformation $s = (x - d)/h$ and the Taylor expansion of $g(d + sh)$ at d, we obtain the right side of (5) is equal to

$$\int K(s)g(d)\mathrm{ds}+\int sK(s)g^{'}(d)\mathrm{ds}+O(h^2). \qquad (6)$$

Because $\int K(s)ds = 1$ and $\int sK(s)ds = 0$, we then obtain the first equation. Similarly, we can obtain the second equation.

References

1. Helsel, DR. Nondetects and Data Analysis: Statistics for Censored Environmental Data. Wiley-Interscience; 2005.
2. Johnson N, Shelton BJ, Hopenhayn C, Tucker TT, Shi X, Unrine JM, Huang B, Christian WJ, Zhang Z, Li L. Concentrations of arsenic, chromium, and nickel in toenail samples from appalachian kentucky residents. Journal of Environmental Pathology, Toxicology and Oncology. 2011; 30(3): 213–223.
3. Razak A, Hafiza N, Praveena SM, Hashim Z. Toenail as a biomarker of heavy metal exposure via drinking water: a systematic review. Reviews on Environmental Health. 2014; 30(3):1–7.
4. Helsel DR. Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. Chemosphere. 2006; 65(11):2434–2439. [PubMed: 16737727]
5. 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. Environmental Health Perspectives. 2004; 112(17):1691–1696. [PubMed: 15579415]
6. Gillespie BW, Chen Q, Reichert H, Franzblau A, Hedgeman E, Lepkowski J, Adriaens P, Demond A, Luksemburg W, Garabrant DH. Estimating population distributions when some data are below a limit of detection by using a reverse Kaplan-Meier estimator. Epidemiology. 2010; 21(4):S64–S70. [PubMed: 20386104]
7. Dabrowska DM. Uniform consistency of the kernel conditional Kaplan-Meier estimate. The Annals of Statistics. 1989; 17(3):1157–1167.
8. Silverman, BW. Density Estimation for Statistics and Data Analysis. CRC press; 1986.
9. Lee, L. NADA: Nondetects And Data Analysis for Environmental Data. 2013. URL https://CRAN.R-project.org/package=NADA, r package version 1.5-6
10. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. CA: A Cancer Journal for Clinicians. 2012; 62(1):10–29. [PubMed: 22237781]
11. Li L, Plummer SJ, Thompson CL, Tucker TC, Casey G. Association between phosphatidylinositol 3-kinase regulatory subunit p85α met326ile genetic polymorphism and colon cancer risk. Clinical Cancer Research. 2008; 14(3):633–637. [PubMed: 18245521]
12. Collett, D. Modelling Survival Data in Medical Research. CRC press; 2015.
13. Li Y, Gillespie BW, Shedden K, Gillespie JA. Calculating profile likelihood estimates of the correlation coefficient in the presence of left, right or interval censoring and missing data. Technical Report. 2015
14. Akritas MG, Murphy SA, LaValley MP. The theil-sen estimator with doubly censored data and applications to astronomy. Journal of the American Statistical Association. 1995; 90(429):170–177.
15. Akaike, H. Prediction and entropy. In: Atkinson, A., Fienberg, SE., editors. A Celebration of Statistics. Springer; 1985. p. 1-24.
16. Murray S, Tsiatis A. A nonparametric approach to incorporating prognostic longitudinal covariate information in survival estimation. Biometrics. 1996; 52(1):137–151. [PubMed: 8934589]
17. Chen L, Lin D, Zeng D. Attributable fraction functions for censored event times. Biometrika. 2010; 97(3):713–726. [PubMed: 23956459]

18. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics. 2000; 56(3):779–788. [PubMed: 10985216]

19. Robins, JM. Proceedings of the Biopharmaceutical Section American Statistical Association. 1993. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers; p. 24-33.

20. Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. Biometrics. 1989; 45(2):497–507. [PubMed: 2765634]

21. Cole SR, Chu H, Nie L, Schisterman EF. Estimating the odds ratio when exposure has a limit of detection. International Journal of Epidemiology. 2009; 38(6):1674–1680. [PubMed: 19667054]

22. May RC, Ibrahim JG, Chu H. Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. Statistics in Medicine. 2011; 30(20):2551–2561. [PubMed: 21710558]

23. Van der Vaart, A., Wellner, J. Weak Convergence and Empirical Processes. Springer; New York: 1998.

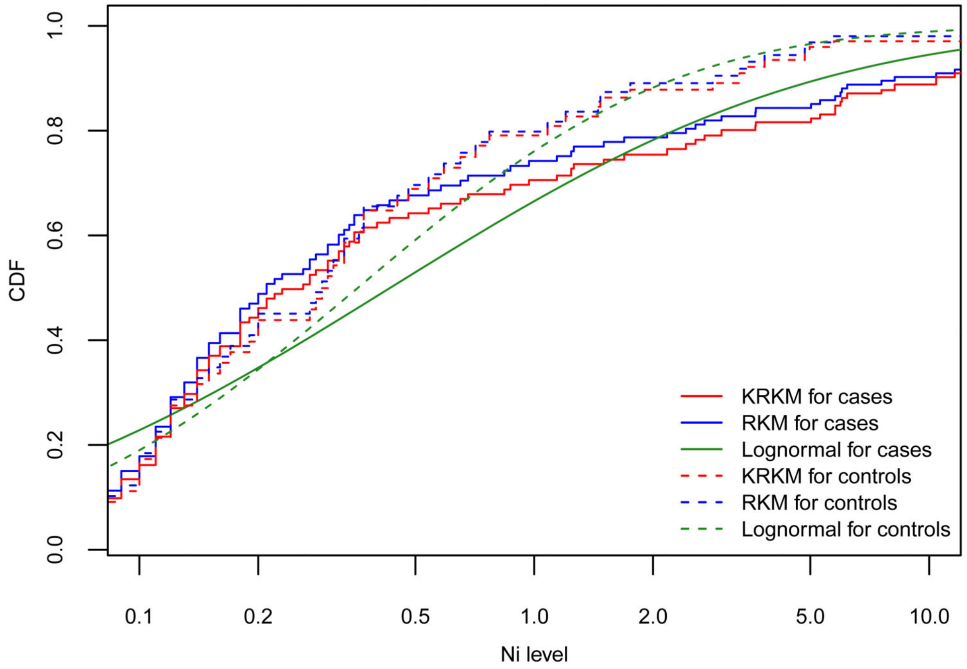

Figure 1.

CDF estimates of Ni exposure distribution for colon cancer cases vs. controls in the Appalachian region based on the KRKM, RKM, and lognormal estimators. The solid curves pertain to the CDF estimates for cases and the dotted curves pertain to those for controls. The red curves are for the KRKM estimator, the blue curves are for the RKM estimator, and the green curves are for the lognormal estimator.

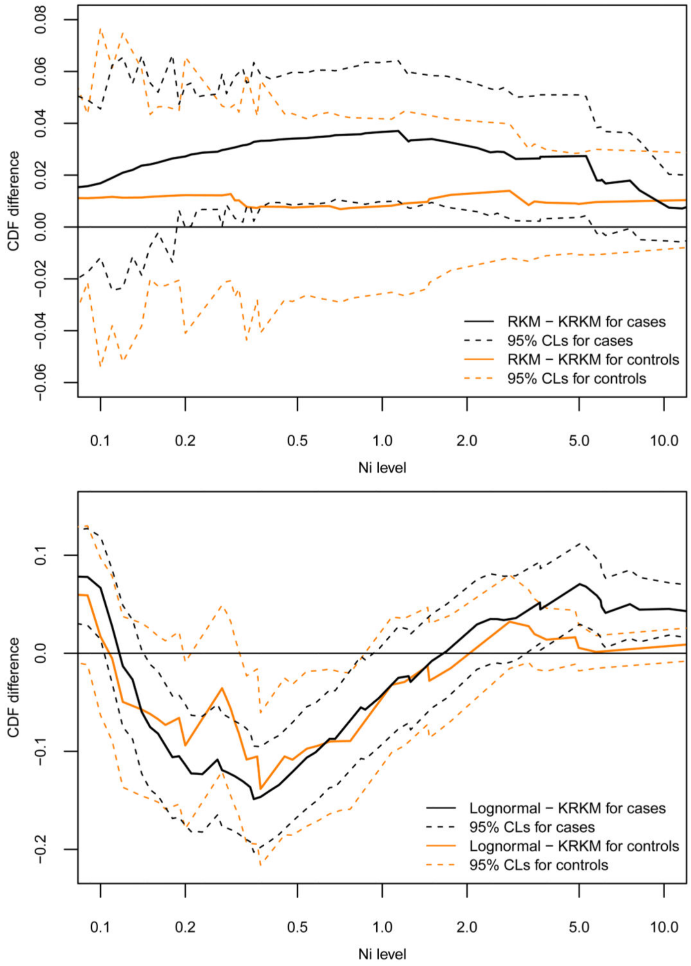

Figure 2.

Differences in CDF estimates of Ni exposure distribution between the RKM and KRKM estimators (upper panel) and between the lognormal and KRKM estimators (lower panel), along with 95% confidence limits. The solid curves are for the point estimates of differences, and the dotted curves are for the corresponding 95% bootstrapped confidence limits (CLs). The black curves pertain to the cases and the orange ones petain to the controls.

Table 1

Comparison of simulation results for the KRKM, RKM and lognormal estimators when the exposure level and DL are associated and the DL was generated based on its empirical distribution in the colon cancer data. True, the true CDF value; Bias, the sampling bias; SSE, the sampling standard error; SEE, the sampling mean of the standard error estimator; CP, the coverage probability of the 95% confidence interval. SEE_B and CP_B pertain to the bootstrap-based variance estimation method using 500 bootstraps, and SEE_F and CP_F pertain to the formula-based variance estimation method. Each entry is based on 1000 replicates.

Setting 1		**KRKM**						**RKM**				**Lognormal**			
n	**True**	**Bias**	**SSE**	**SEE_B**	**CP_B**	**SEE_F**	**CP_F**	**Bias**	**SSE**	**SEE**	**CP**	**Bias**	**SSE**	**SEE**	**CP**
200	.25	.050	.101	.095	.901	.074	.678	.081	.099	.011	.743	.069	.083	.081	.871
	.50	.003	.056	.055	.952	.039	.765	.091	.055	.060	.888	.087	.058	.060	.702
	.75	.002	.038	.036	.946	.030	.938	.048	.034	.036	.843	.054	.036	.036	.700
500	.25	.044	.061	.062	.866	.033	.746	.102	.063	.071	.298	.095	.048	.047	.524
	.50	.003	.036	.035	.948	.025	.801	.096	.035	.037	.752	.088	.035	.036	.346
	.75	.002	.023	.024	.953	.019	.942	.048	.021	.022	.626	.064	.022	.022	.397
1000	.25	.031	.043	.044	.899	.028	.739	.105	.044	.051	.243	.112	.033	.032	.389
	.50	.003	.025	.025	.953	.015	.876	.095	.026	.025	.511	.077	.026	.026	.367
	.75	.001	.019	.018	.953	.015	.946	.049	.016	.015	.607	.056	.016	.015	.327
Setting 2		KRKM						RKM				Lognormal			
n	True	Bias	SSE	SEE_B	CP_B	SEE_F	CP_F	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
200	.25	.012	.041	.040	.922	.036	.942	.025	.041	.045	.902	.023	.034	.033	.925
	.50	.006	.041	.040	.941	.039	.944	.025	.040	.041	.912	.020	.033	.032	.911
	.75	.002	.032	.032	.939	.030	.951	.018	.031	.032	.926	.008	.024	.025	.929
500	.25	.008	.025	.026	.935	.022	.941	.026	.025	.027	.821	.029	.020	.021	.814
	.50	.005	.027	.025	.944	.026	.945	.024	.024	.025	.796	.023	.020	.020	.783
	.75	.002	.021	.021	.949	.021	.944	.014	.019	.020	.842	.009	.016	.015	.859
1000	.25	.009	.018	.018	.940	.016	.940	.024	.018	.019	.808	.028	.015	.015	.529
	.50	.005	.019	.018	.942	.018	.950	.022	.018	.018	.759	.023	.015	.015	.657
	.75	.002	.016	.015	.945	.016	.945	.016	.014	.014	.816	.010	.014	.015	.821

Table 2

Comparison of simulation results for the KRKM, RKM and lognormal estimators when the exposure level and DL are associated and the DL was generated from a lognormal distribution. True, the true CDF value; Bias, the sampling bias; SSE, the sampling standard error; SEE, the sampling mean of the standard error estimator; CP, the coverage probability of the 95% confidence interval. SEE_B and CP_B pertain to the bootstrap-based variance estimation method using 500 bootstraps, and SEE_F and CP_F pertain to the formula-based variance estimation method. Each entry is based on 1000 replicates.

Setting 1		**KRKM**						**RKM**				**Lognormal**			
n	**True**	**Bias**	**SSE**	**SEE_B**	**CP_B**	**SEE_F**	**CP_F**	**Bias**	**SSE**	**SEE**	**CP**	**Bias**	**SSE**	**SEE**	**CP**
200	.25	.032	.100	.101	.902	.074	.692	.165	.104	.110	.561	.117	.075	.074	.683
	.50	.005	.064	.067	.958	.052	.877	.094	.068	.066	.724	.076	.058	.059	.592
	.75	.004	.035	.033	.947	.028	.933	.039	.036	.034	.701	.054	.032	.031	.663
500	.25	.020	.093	.095	.899	.077	.734	.153	.092	.091	.420	.101	.073	.073	.563
	.50	.005	.045	.045	.943	.036	.925	.089	.046	.047	.399	.093	.038	.039	.521
	.75	.004	.027	.025	.931	.025	.952	.038	.026	.025	.467	.046	.024	.022	.429
1000	.25	.017	.076	.074	.901	.059	.781	.121	.078	.078	.663	.094	.057	.056	.334
	.50	.006	.030	.029	.946	.025	.934	.086	.031	.029	.581	.084	.027	.027	.221
	.75	.003	.016	.016	.954	.014	.946	.040	.015	.015	.396	.046	.014	.013	.196
Setting 2		KRKM						RKM				Lognormal			
n	True	Bias	SSE	SEE_B	CP_B	SEE_F	CP_F	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
200	.25	.006	.038	.038	.943	.034	.928	.034	.036	.038	.812	.036	.028	.027	.875
	.50	.005	.031	.032	.940	.029	.937	.029	.030	.033	.793	.029	.029	.028	.858
	.75	.004	.020	.021	.947	.019	.952	.023	.023	.023	.871	.020	.019	.020	.802
500	.25	.006	.023	.024	.951	.022	.932	.037	.022	.020	.662	.035	.018	.018	.671
	.50	.005	.020	.020	.945	.019	.951	.024	.021	.021	.759	.030	.017	.019	.821
	.75	.004	.013	.013	.955	.012	.953	.017	.014	.013	.837	.018	.012	.013	.833
1000	.25	.006	.017	.017	.946	.014	.941	.032	.016	.018	.663	.033	.014	.012	.669
	.50	.005	.014	.014	.945	.013	.945	.026	.014	.014	.589	.030	.011	.011	.571
	.75	.004	.012	.012	.943	.012	.947	.016	.012	.013	.743	.017	.010	.011	.605

Table 3

Comparison of simulation results for the KRKM, RKM and lognormal estimators when the exposure level and DL are independent and the DL was generated based on its empirical distribution in the colon cancer data. True, the true CDF value; Bias, the sampling bias; SSE, the sampling standard error; SEE, the sampling mean of the standard error estimator; CP, the coverage probability of the 95% confidence interval. $\mathrm{SEE_B}$ and $\mathrm{CP_B}$ pertain to the bootstrap-based variance estimation method using 500 bootstraps, and $\mathrm{SEE_F}$ and $\mathrm{CP_F}$ pertain to the formula-based variance estimation method. Each entry is based on 1000 replicates.

Setting 1		**KRKM**						**RKM**				**Lognormal**			
n	**True**	**Bias**	**SSE**	**$\mathrm{SEE_B}$**	**$\mathrm{CP_B}$**	**$\mathrm{SEE_F}$**	**$\mathrm{CP_F}$**	**Bias**	**SSE**	**SEE**	**CP**	**Bias**	**SSE**	**SEE**	**CP**
200	.25	.005	.040	.039	.939	.025	.897	.005	.039	.044	.938	.000	.035	.035	.944
	.50	.000	.041	.039	.947	.027	.918	.000	.039	.041	.943	-.003	.034	.034	.954
	.75	.008	.032	.032	.942	.028	.940	.008	.031	.033	.937	-.006	.026	.026	.937
500	.25	.004	.024	.028	.960	.017	.926	.003	.024	.034	.958	.001	.021	.022	.960
	.50	.002	.026	.025	.941	.023	.942	.002	.025	.026	.946	-.002	.021	.021	.948
	.75	.007	.021	.021	.946	.018	.945	.008	.020	.025	.941	.001	.017	.016	.931
1000	.25	.004	.017	.018	.945	.015	.943	.002	.017	.018	.957	-.003	.015	.015	.953
	.50	.001	.018	.018	.948	.018	.946	.003	.018	.018	.942	.001	.015	.015	.952
	.75	.008	.015	.015	.951	.014	.944	.007	.015	.014	.936	.006	.012	.012	.937
Setting 2		KRKM						RKM				Lognormal			
n	True	Bias	SSE	$\mathrm{SEE_B}$	$\mathrm{CP_B}$	$\mathrm{SEE_F}$	$\mathrm{CP_F}$	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
200	.25	.003	.048	.046	.933	.039	.922	.004	.048	.054	.933	-.003	.054	.049	.931
	.50	-.000	.047	.046	.943	.042	.939	-.000	.045	.048	.944	-.003	.050	.046	.941
	.75	.007	.035	.035	.938	.031	.941	.008	.034	.036	.951	.005	.032	.030	.933
500	.25	.005	.028	.030	.946	.024	.944	.003	.028	.031	.948	.001	.030	.030	.955
	.50	-.002	.028	.030	.957	.023	.944	-.001	.028	.029	.945	-.002	.028	.029	.944
	.75	.006	.023	.023	.942	.021	.939	.006	.022	.022	.948	.005	.020	.020	.940
1000	.25	.003	.020	.021	.954	.016	.943	.001	.018	.020	.951	.002	.021	.021	.942
	.50	-.000	.022	.021	.941	.022	.953	-.004	.018	.020	.949	-.002	.021	.020	.941
	.75	.008	.017	.017	.939	.015	.945	.007	.015	.016	.938	.006	.015	.014	.934

Table 4

The non-detect rate, correlation coefficient between the exposure level and DL and the corresponding p value.

Non-detect rate(%)												
	Ni	**Cd**	**As**	**Cr**	**Pb**	**Co**	**Al**	**Mn**	**Fe**	**Cu**	**Zn**	**Se**
case	23	79	43	13	7	73	7	51	21	0	0	3
control	48	83	45	42	33	79	14	59	37	4	3	8
Correlation coefficient and p value for colon cancer cases												
	Ni	Cd	As	Cr	Pb	Co	Al	Mn	Fe	Cu	Zn	Se
AFT_cor	.513	.369	.290	.638	.436	.538	.562	.332	.446	.320	.202	.174
AFT_p	<.001	.010	.003	<.001	<.001	<.001	<.001	.005	<.001	<.001	.013	.034
Clik_cor	.512	.368	.289	.636	.435	.537	.561	.331	.445	.319	.201	.174
Clik_p	<.001	.035	.007	<.001	<.001	<.001	<.001	.012	<.001	<.001	.014	.036
Ken_cor	.134	.009	.077	.365	.254	.070	.344	.07	.222	.110	.090	-.066
Ken_p	.016	.863	.166	<.001	<.001	.205	<.001	.207	<.001	.050	.108	.241
Correlation coefficient and p value for controls												
	Ni	Cd	As	Cr	Pb	Co	Al	Mn	Fe	Cu	Zn	Se
AFT_cor	.081	-.254	.375	.092	.105	.102	.106	-.199	.141	.097	-.098	-.055
AFT_p	.636	.350	.001	.466	.391	.629	.311	.360	.320	.337	.330	.614
Clik_cor	.068	-.493	.346	.092	.104	.102	.105	-.198	.140	.096	-.097	-.054
Clik_p	.683	.055	.005	.477	.405	.646	.317	.334	.334	.340	.330	.612
Ken_cor	-.032	-.032	.173	.042	.069	-.006	.047	-.030	.039	.046	.003	.024
Ken_p	.633	.629	.009	.533	.302	.932	.481	.650	.561	.491	.968	.717

Note: AFT_cor and Clik_cor are the estimates of the Pearson's correlation coefficient between the exposure level and DL (both log-transformed) based on the lognormal AFT model method and the “clikcorr” R package, respectively. Ken_cor is the Kendall's tau correlation coefficient estimate. AFT_p, Clik_p and Ken_p are the corresponding p-values.