# A Comparison of Methods for Correlating Two Variables in the Presence of Non-Detects

**Courtney E. McCracken**
**Dept. of Pediatrics, Emory University, Atlanta, GA 30329**

**Stephen W. Looney**
**Dept. of Biostatistics, Georgia Regents University, Augusta, GA 30912**

## Abstract

This presentation deals with correlating two variables that have values that fall below the known limit of detection (LOD) of the measuring device; these values are known as non-detects (NDs). We use simulation to compare several methods for estimating the association between two such variables. The most commonly used method, simple substitution, consists of replacing each ND with some representative value such as LOD/2. Spearman's correlation, in which all NDs are assumed to be tied at some value just smaller than the LOD, has also been used. We evaluate each method under several scenarios, including small to moderate sample size, moderate to large censoring proportions, extreme imbalance in censoring proportions, and non-bivariate normal (BVN) data. The methods are compared in terms of estimation bias, median absolute deviation, 95% confidence interval width, etc., but our primary focus is on coverage probability. A maximum likelihood approach based on the assumption of BVN data has acceptable performance under most scenarios, even with non-BVN data. Spearman's rho also performs well under many conditions. The methods are illustrated using real data taken from the biomarker literature.

**Keywords:** confidence interval; coverage probability; left censoring; limit of detection; maximum likelihood; Spearman correlation

## 1. Introduction

In biomarker research studies, it is not uncommon for there to be specimens for which the concentration of the biomarker is below the limit of detection (LOD). In other words, any concentration below a certain value (the LOD) cannot be detected with the measuring device used to determine the levels of the analyte in the biological specimens. The only information available for such a specimen is that the analyte is present at some level greater than zero but less than the LOD. Such observations are most commonly referred to as non-detects (NDs), and are usually treated in statistical analyses of the biomarker data as being left-censored.

Applied researchers who work with biomarkers frequently encounter specimens with non-detectable concentrations of the analytes of interest. For example, NDs were a serious problem in the study by Amorim and Alvarez-Leite (1997), who examined the validity of urinary *o*-cresol as a biomarker of exposure to toluene. Using urine samples of individuals exposed to toluene in shoe factories, painting sectors of metal industries, and printing shops, Amorim and Alvarez-Leite correlated the concentrations of urinary *o*-cresol with urinary hippuric acid, which at the time of their study was the most commonly used biomarker for occupational toluene exposure. In 39 of the 54 urine samples in their study (72%), the *o*-cresol concentration was below its LOD (0.2 μg/ml). In 4 of the 39 urine samples with ND *o*-cresol (10%), the hippuric acid concentration was

below its LOD (0.1 mg/ml). In other words, only 15 samples (28%) had "complete data" for both biomarkers.

In another example of a biomarker research study in which NDs were a concern, Atawodi et al. (1998) evaluated hemoglobin adducts as biomarkers of exposure to tobacco smoke. They compared the adduct levels of 18 smokers with those of 52 “never smokers.” In 7 of the 52 samples from the "never smokers" (13%), the hemoglobin adduct levels were below the LOD (9 fmol HPB/g Hb).

From our perspective, the statistical methods that are most commonly used to analyze biomarker data that include NDs are flawed. One of the most frequently used methods is to remove the specimens with NDs from the statistical analysis and analyze only the "complete data." In their evaluation of the urinary concentration of *trans, trans* muconic acid (*t,t*-MA) as a biomarker for low-level benzene exposure, Lagorio et al. (1998) used this approach. Three different pre-analytical procedures (filtration, methanol dilution, ether extraction) were applied to urine samples from 10 Estonian shale oil workers, and Lagorio et al. examined the inter-correlations among the three procedures using only the samples with no NDs for any of the 3 methods. Another commonly used approach for dealing with NDs is "simple substitution." In this approach, one simply replaces the missing biomarker levels with some substitute value and then performs the “usual” statistical analysis on the new sample of data that includes the substituted values in place of the NDs. The values that are most commonly substituted are the LOD (Amorim and Alvarez-Leite, 1997; Atawodi et al., 1998) and LOD/2 (Cook et al., 1993).

Nonparametric methods have also been applied to biomarker data that includes NDs. In this approach, one treats all NDs as if they were tied at some value just below the LOD. For example, if one wished to correlate two biomarkers $X$ and $Y$, at least one of which was undetectable in one or more specimens, one could calculate Spearman’s $r_s$ using the ranks of the $X$ and $Y$ values based on the entire data set. In this approach, all NDs for each variable would be assigned the smallest midrank. Alternatively, if one wished to compare two groups in terms of a biomarker that was subject to NDs, one could use the Mann-Whitney-Wilcoxon test after computing the ranks based on the combined sample of data from the two groups. Again, each of the NDs would be assigned the smallest mid-rank. In their statistical evaluation of hemoglobin adducts as biomarkers of exposure to tobacco smoke, Atawodi et al. (1998) used this approach and found that the HPB-Hb adduct level was significantly higher in smokers than in never smokers ($p = 0.02$).

## 2. Maximum Likelihood Approach

In this presentation, we focus on the problem of correlating two biomarkers that are both subject to left-censoring. Wang (2006) demonstrated via simulation that, for this analysis, none of the "standard" methods described above for dealing with NDs are satisfactory, especially if the two biomarkers are strongly positively correlated ($\rho \geq 0.5$).

When $X$ and $Y$ follow a bivariate normal distribution and both $X$ and $Y$ have known LODs, the preferable approach is to use the maximum likelihood (ML) method proposed by Lyles et al. (2001) to estimate the Pearson correlation between $X$ and $Y$. In this section, we briefly describe the statistical theory behind this method.

Let $X$ and $Y$ denote the two biomarkers to be correlated and let $L_x$ and $L_y$ denote their known detection limits, respectively. Based on the assumption that the non-censored values of $X$ and $Y$ follow a bivariate normal distribution, Lyles et al. developed a method for estimating the population parameter vector $\boldsymbol{\theta} = \left[\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho\right]$ using ML estimation. Let $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$; $i = 1, \ldots, n$ denote the observed sample of $(x, y)$ values (including the NDs). Lyles et al. categorized the observed pairs of $(x, y)$

values as follows: (1) pairs where both $x$ and $y$ are observed (the "complete data"), (2) pairs where $x$ is observed but $y < L_y$, (3) pairs where $y$ is observed but $x < L_x$, and (4) pairs where $x < L_x$ and $y < L_y$. Following the same notation as that used by Lyles et al., the contribution of each Type 1 pair to the likelihood function of the entire sample is:

$$t_{i1} = (2\pi\sigma_x\sigma_{y|x})^{-1}\exp\left\{-0.5\left[\frac{(y_i-\mu_{y|x_i})^2}{\sigma_{y|x}^2}+\frac{(x_i-\mu_x)^2}{\sigma_x^2}\right]\right\}, \quad (1)$$

where $\mu_{y|x_i} = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x_i-\mu_x)$ and $\sigma_{y|x}^2 = \sigma_y^2(1-\rho^2)$. The contribution of each Type 2 pair is:

$$t_{i2} = (2\pi\sigma_x^2)^{-1/2}\exp\left[-0.5\frac{(x_i-\mu_x)^2}{\sigma_x^2}\right]\times\Phi\left(\frac{L_y-\mu_{y|x_i}}{\sigma_{y|x}}\right), \quad (2)$$

where $\Phi(\cdot)$ denotes the standard normal distribution function. Similarly, the contribution of each Type 3 pair is:

$$t_{i3} = (2\pi\sigma_y^2)^{-1/2}\exp\left[-0.5\frac{(y_i-\mu_y)^2}{\sigma_y^2}\right]\times\Phi\left(\frac{L_x-\mu_{x|y_i}}{\sigma_{x|y}}\right), \quad (3)$$

where $\mu_{x|y_i} = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y_i-\mu_y)$ and $\sigma_{x|y}^2 = \sigma_x^2(1-\rho^2)$.

Finally, the contribution of each Type 4 pair is:

$$t_4 = \int_{-\infty}^{L_y}\Phi\left\{\frac{L_x-\left[\mu_x+\frac{\rho\sigma_x(y-\mu_y)}{\sigma_y}\right]}{\sigma_x\sqrt{1-\rho^2}}\right\}\times(2\pi\sigma_y^2)^{-1/2}\exp\left[-0.5\frac{(y-\mu_y)^2}{\sigma_y^2}\right]dy. \quad (4)$$

Without loss of generality, suppose the $(x, y)$ pairs are ordered and indexed by $i$ so that all Type 1 pairs come first, followed by all Type 2, 3, and 4 pairs. In addition, assume that there are $n_j$ terms of type $j$ ($j$ = 1, 2, 3, 4) and define $n_{k\bullet} = \sum_{j=1}^{k} n_j$ for $k$ = 2, 3. Then, the likelihood for the entire sample (including the NDs) is:

$$L(\boldsymbol{\theta}\,|\,\mathbf{x},\mathbf{y}) = \left(\prod_{i=1}^{n_1} t_{i1}\right)\left(\prod_{i=n_1+1}^{n_{2\bullet}} t_{i2}\right)\left(\prod_{i=n_{2\bullet}+1}^{n_{3\bullet}} t_{i3}\right)t_4^{n_4}, \quad (5)$$

where $\mathbf{x}$ is the vector of observed $x$-values, $\mathbf{y}$ is the vector of observed $y$-values, and $t_{i1}$, $t_{i2}$, $t_{i3}$, $t_4$ are given by Equations (1), (2), (3), and (4), respectively.

Once the likelihood function in (5) has been maximized and the ML estimates and their estimated standard errors have been calculated, one can compute an approximate 100(1-$\alpha$)% Wald-type confidence interval (CI) for $\rho$:

$$\hat{\rho}_{ML} \pm z_{\alpha/2}\widehat{SE}(\hat{\rho}_{ML}),$$

where $z_{\alpha/2}$ denotes the upper $\alpha/2$-percentage point of the standard normal. Because Wald-type CIs are can be suspect when $n$ is small, Lyles et al. also considered profile likelihood (PL) CIs. They found that, generally, the PL intervals performed better than the Wald-type intervals. However, PL intervals are more difficult computationally since they usually do not have a closed-form expression. An alternative approach is to use a CI based on an improved Fisher $z$ transformation of $\hat{\rho}_{ML}$, as will be discussed in the following section.

## 3. Correlation Coefficient Confidence Interval Methods When Both Variables Are Subject to Limits of Detection

In her doctoral dissertation, McCracken (2013) compared the following methods for estimating the true correlation between *X* and *Y* when both *X* and *Y* are subject to limits of detection:

(1) Maximum Likelihood [Lyles et al. (2001)]

(2) Simple Substitution: replace each ND by
   (a) *LOD*
   (b) *LOD/2*
   (c) $LOD / \sqrt{2}.$

(3) Complex Substitution (Lynn 2001, McCracken 2013): Replace each ND among the *x*-values by $E(X_i \mid X_i < LOD_x)$ and each ND among the *y*-values by $E(Y_i \mid Y_i < LOD_y).$ That is, replace each ND for each variable by the conditional mean of that variable, given that it is known that the variable is less than its LOD. See Lynn (2001) for computational details.

(4) Random Imputation: Replace each ND among the *x*-values by a value randomly selected from the interval [0, $LOD_x$], and replace each ND among the *y*-values by a value randomly selected from the interval [0, $LOD_y$].

(5) Spearman's correlation: Treat each ND among the *x*-values as being an observation that is tied at some value just smaller than the smallest observed *x*-value, and treat each ND among the *y*-values as being tied at some value just smaller than the smallest observed *y*-value.

For estimation methods (1) - (4) described above, a 2nd-order Fisher *z*-transformation, which provides a more accurate estimate of the variance of $z(\hat{\rho})$, was used to find an approximate 95% CI for the true value of the Pearson correlation between *X* and *Y*. Confidence intervals based on this method have improved coverage probabilities relative to those based on the usual Fisher *z*-transformation, and the improved transformation poses no computational difficulties. Details of this method are provided in Li, Wang, and Chan (2005) and McCracken (2013).

For Spearman's coefficient (method (5)), both the Jackknife and approximate bootstrap confidence interval (ABC) were considered for finding a 95% CI for the true value of the correlation. McCracken examined $r_s$ as a surrogate for *r* (to be used in place of *r* when *X* and *Y* do not follow a BVN distribution), as well as an estimate of the true population rank correlation, $\rho_s$. Defining the true value of the Spearman coefficient is not straightforward (Gibbons and Chakraborti 2003); McCracken followed Newton and Rudel (2007) and used a simulation-based estimate of $\rho_s$ as the true value. In other words, the true value of Spearman's coefficient was taken to be the mean of the 5,000 $r_s$ values calculated from the Monte Carlo samples before the censoring schemes were applied. In the Monte Carlo simulation study described below, McCracken found that the Jackknife method for finding a 95% CI for the true value of the Spearman correlation based on the sample value $r_s$ was generally as good as, if not better than, the ABC method in terms of coverage probability, and was much easier computationally. Thus, the results for Spearman's coefficient provided in the presentation are based on the Jackknife intervals.

Each of the confidence interval methods considered by McCracken (2013) is illustrated in the Example in Section 5, and R code for calculating each of the point estimates and corresponding CIs can be obtained from the second author.

## 4. Simulation Study Comparing the Various Confidence Interval Methods

McCracken (2013) performed extensive Monte Carlo simulations to compare 95% CIs based on the 5 point estimation methods describe above. She examined multiple settings of several simulation parameters: (1) Sample size ($n$ = 20, 30, 50, 75, 100, 200, 500); (2) True correlation between $X$ and $Y$ prior to censoring ($\rho$ = -0.9, -0.6, -0.5, -0.25, 0.0, 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8 and 0.9); and (3) Censoring proportions on $X$ (denoted by $p_1$) and $Y$ (denoted by $p_2$). Fifty-five combinations of censoring proportions were examined, including 12 balanced and 43 unbalanced combinations. The following balanced combinations were considered: ($p_1$, $p_2$) = (0,0), (10, 10), (20, 20), (25, 25), (30, 30), (40, 40), (50, 50), (60, 60), (70, 70), (75, 75), (80, 80), and (90, 90). Examples of unbalanced combinations included: ($p_1$, $p_2$) = (10, 0), (10, 5), (10, 50), (10, 75), (20, 50), (25, 75), (30, 75), (90, 45), and (90, 0).

In addition to the various settings of $n$, $\rho$, and ($p_1$, $p_2$) considered by McCracken, three different distributions were considered for the true bivariate distribution of $X$ and $Y$: bivariate normal (BVN), bivariate gamma (BVG), and bivariate beta (BVB). The non-BVN distributions were chosen to represent varying degrees of departure from the BVN assumption, as measured by Mardia's measures of multivariate skewness and kurtosis (denoted by $\beta_{1,p}$ and $\beta_{2,p}$, respectively). For the BVN, $\beta_{1,2}$ = 0 and $\beta_{2,2}$ = 8. For the BVG distribution used in the simulation study, $\beta_{1,2}$ = 3.5 and $\beta_{2,2}$ = 12, and, for the BVB, $\beta_{1,2}$ = 3 and $\beta_{2,2}$ = 10.

In total, 18,480 different combinations of simulation parameter settings (including the 3 bivariate distributions) were examined. A total of 5,000 Monte Carlo samples were used to evaluate each combination of settings.

In the simulation study, each of the five estimation methods described above were evaluated in terms of: (1) bias (and absolute bias), (2) median absolute deviation, (3) confidence interval width, and (4) confidence interval coverage probability (CP). However, because of space limitations, this presentation is concerned only with the CP of the 95% confidence intervals based on the ML and Spearman methods.

The ML-based confidence intervals had the best overall performance in terms of CP and can generally be recommended even when the BVN assumption is suspect. This is somewhat surprising since the MLEs were derived under the assumption of BVN. For extreme negative values of $\rho$, small sample sizes and/or extremely heavy or imbalanced censoring, the ML method may not produce a valid point estimate (due to failure of the optimization routine to converge) and/or the corresponding ML-based CI may be unreliable. This is especially true when the true joint distribution of $X$ and $Y$ differs substantially from the BVN. Under these circumstances, Spearman's coefficient and the corresponding Jackknife CI would be the best alternative for finding a 95% CI for the true correlation between $X$ and $Y$.

McCracken's simulated CP results are briefly summarized in Tables 1, 2, and 3 for the 95% approximate CIs for $\rho$ based on the ML method (using the improved Fisher $z$ transformation) and the approximate CIs based on Spearman's coefficient (using the Jackknife), considered as an estimate of $\rho_s$. Table 1 examines the effect of the $X$ and $Y$ censoring proportions on the CP of these intervals. Because of space limitations, simulation results for only a subset of the censoring proportions considered by McCracken can be summarized here: ($p_1$, $p_2$) = (0, 0) (0.1, 0.7) (0.25, 0.25) (0.25, 0.75) (0.5, 0.5) (0.75, 0.375) (0.9, 0) (0.9, 0.9). For each of these values of ($p_1$, $p_2$) and each bivariate distribution, the median CP over all settings of the other simulation parameters (i.e., the true value of $\rho$ or $\rho_s$, and $n$) was calculated. In the bottom half of Table 1, the "non-normal" simulated results were obtained by averaging the CPs for the two methods calculated using the simulated BVG and BVB data. For example, for censoring

proportions ($p_1$, $p_2$) = (0.1, 0.7), the median CP over all other simulation parameter settings was 94.8% for the ML method and 91.5% for Spearman's coefficient when the data were generated from the BVN. When data were generated from the non-BVN distributions, the median CPs for the ML and Spearman's methods were 93.9% and 92.9%, respectively.

To determine if the CPs of the CI methods differed in any substantial way from the nominal 95% confidence level, McCracken (2013) used the "liberal" guideline proposed by Bradley (1978) for evaluating the robustness of a statistical test: if the true significance level differs from the nominal value by no more than $\alpha$ /2, the test is robust. If the true significance level differs by more than $\alpha$ /2 from the nominal value (either above or below), the test is not robust. In this presentation, we apply the Bradley criterion as follows: if the CP produced by the CI method differs from the 95% nominal confidence level by no more than 2.5%, the CP is deemed to be acceptable. If the CP differs by more than 2.5% from the 95% level (either above or below), the CP is deemed to be unacceptable. Thus, for this presentation, the CP must be between 92.5% and 97.5% for the CI procedure to be classified as "acceptable" for that combination of simulation parameter settings.

In Table 1, the values in boldface indicate median CPs that were less than the lower acceptability limit of 92.5%. Note that the ML-based CIs yielded acceptable CP for all censoring proportions except (0.9, 0.9) with BVN data and (0, 0) and (0.9, 0.9) for non-BVN data. The Spearman-based CIs performed as well as the ML-based intervals for many combinations of censoring proportions, but failed to maintain the 92.5% level in several instances, especially for BVN data.

Table 2 examines the effect that the true value of the association parameter (either Pearson's correlation or Spearman's coefficient) had on the median CP of the CIs based on the ML and Spearman methods. Values in boldface again indicate median CPs that did not exceed the lower acceptability limit of 92.5%. Note that the ML-based CIs maintained an acceptable CP value for all values of $\rho$ except -0.9 with BVN data; however, they did not perform as well with non-BVN data. The Spearman-based CIs generally performed as well as the ML intervals with non-BVN data, but failed to achieve the 92.5% acceptability limit for several values of $\rho$ with BVN data.

Table 3 examines the effect of $n$ on the median CP of the ML- and Spearman-based CIs. As in Tables 1 and 2, the values in boldface indicate median CPs that did not achieve the lower acceptability limit of 92.5%. Surprisingly, the ML method maintained an acceptable value of CP for all sample sizes except $n$ = 500 when the data were non-BVN. Spearman's performed almost as well as the ML method with non-BVN data, but failed to achieve the 92.5% limit for several sample sizes with BVN data.

To summarize the simulation results presented here, McCracken's study showed that, as expected, with BVN data, the ML method was superior to the Spearman method under all conditions considered and had median CP above 92.5% except when $p_1 = p_2 =$ 0.9. With non-BVN data, the performance of the ML-based intervals was still superior to the Spearman-based intervals except under some scenarios involving moderate to large $n$, small $|\rho|$ and very light censoring combinations. The Jackknife CIs based on Spearman's $r_s$ (as an estimate of $\rho_s$) performed acceptably as long as $|\rho_s|$ was small or moderate, the sample size was not too large (i.e., less than 500), and the censoring proportion for $X$ was small to moderate and there was little or no censoring on $Y$. The CIs based on $r_s$ generally performed better for non-BVN data than for BVN data.

Table 1 Comparison of Median Coverage Probability of 3 CI Methods, by Censoring Proportions

| Distribution | Censoring Proportions ($p_1$, $p_2$) | Method | |
|---|---|---|---|
| | | Maximum Likelihood | Spearman's $r_s$ (Jackknife Interval) |
| Normal | (0, 0) | 94.8 | 93.8 |
| Normal | (0.1, 0.7) | 94.8 | **91.5** |
| Normal | (0.25, 0.25) | 94.8 | 94.1 |
| Normal | (0.25, 0.75) | 94.8 | **89.9** |
| Normal | (0.5, 0.5) | 94.7 | 92.5 |
| Normal | (0.75, 0.375) | 94.9 | **88.4** |
| Normal | (0.9, 0) | 94.9 | **33.8** |
| Normal | (0.9, 0.9) | **90.0** | **61.8** |
| --- | | | |
| Non-Normal | (0, 0) | **91.0** | 93.9 |
| Non-Normal | (0.1, 0.7) | 93.9 | 92.9 |
| Non-Normal | (0.25, 0.25) | 93.4 | 94.1 |
| Non-Normal | (0.25, 0.75) | 94.0 | **91.8** |
| Non-Normal | (0.5, 0.5) | 93.8 | 93.7 |
| Non-Normal | (0.75, 0.375) | 94.0 | 92.5 |
| Non-Normal | (0.9, 0) | 93.2 | **69.7** |
| Non-Normal | (0.9, 0.9) | **85.8** | **74.3** |

Table 2 Comparison of Median Coverage Probability of 3 CI Methods, by True Parameter Value

| Distribution | True Value of $\xi$* | Method | |
|---|---|---|---|
| | | Maximum Likelihood | Spearman's $r_s$ (Jackknife Interval) |
| Normal | -0.9 | **90.8** | **15.8** |
| Normal | -0.5 | 94.6 | **91.6** |
| Normal | -0.25 | 94.9 | 93.3 |
| Normal | 0.0 | 95.0 | 94.4 |
| Normal | 0.25 | 95.0 | 93.0 |
| Normal | 0.5 | 94.9 | **91.9** |
| Normal | 0.75 | 94.7 | **89.0** |
| Normal | 0.9 | 94.4 | **76.7** |
| --- | | | |
| Non-Normal | -0.9 | **87.6** | **15.4** |
| Non-Normal | -0.5 | **91.7** | **92.0** |
| Non-Normal | -0.25 | 92.9 | 93.3 |
| Non-Normal | 0.0 | **89.3** | 94.2 |
| Non-Normal | 0.25 | 94.0 | 94.2 |
| Non-Normal | 0.5 | 93.5 | 93.4 |
| Non-Normal | 0.75 | 93.4 | **91.1** |
| Non-Normal | 0.9 | **91.4** | **79.8** |

*$\xi = \rho$ for Pearson correlation, $\xi = \rho_s$ for Spearman's coefficient

Table 3 Comparison of Median Coverage Probability of 3 CI Methods, by Sample Size

| Distribution | Sample Size | Method | |
|---|---|---|---|
| | | Maximum Likelihood | Spearman's $r_s$ (Jackknife Interval) |
| Normal | 20 | 94.1 | **92.3** |
| Normal | 30 | 94.4 | 93.3 |
| Normal | 50 | 94.5 | 93.0 |
| Normal | 75 | 95.2 | 93.1 |
| Normal | 100 | 95.0 | **92.2** |
| Normal | 200 | 94.8 | **88.0** |
| Normal | 500 | 95.2 | **73.3** |
| --- | | | |
| Non-Normal | 20 | 93.3 | **92.4** |
| Non-Normal | 30 | 93.7 | 93.4 |
| Non-Normal | 50 | 93.8 | 93.7 |
| Non-Normal | 75 | 93.8 | 93.8 |
| Non-Normal | 100 | 93.7 | 93.9 |
| Non-Normal | 200 | 93.2 | 93.7 |
| Non-Normal | 500 | **91.7** | **90.6** |

## 5. Example

In the Introduction, we described the study by Amorim and Alvarez-Leite (1997), who examined the validity of urinary *o*-cresol as a biomarker of exposure to toluene by correlating it with urinary concentrations of hippuric acid in 54 individuals with occupational exposure to toluene. A scatterplot of the data provided in their article is given in Figure 1. In the plot, a zero value is used to represent each ND. The bivariate normality assumption is rejected using the complete cases (Shapiro-Wilk $p < 0.0001$ for both the 15 *o*-cresol values and the 15 corresponding hippuric acid values). The ML method for estimating the Pearson correlation in the presence of NDs yields $\hat{\rho}_{ML} = 0.79,$ with a 95% modified Fisher *z* CI($\rho$) of (0.66, 0.87). Applying the standard approach for Pearson's correlation based on the traditional Fisher *z*-transform to the 15 cases with complete data (the "complete case" analysis) yields $r = 0.76$ with a 95% CI($\rho$) of (0.40, 0.92). Using simple substitution (replacing the NDs by LOD/2, as was done by Amorim and Alvarez-Leite) yields $\hat{\rho}_{LOD/2} = 0.79,$ with a 95% modified Fisher *z* CI($\rho$) of (0.65, 0.87).

A summary of the results obtained using each of the methods examined by McCracken (2013) in her simulation study is provided in Table 4. In this table, we also provide results for two methods that were not included in the simulation study: Kendall's concordance coefficient and simple substitution with 0 in place of each of the NDs. Both of these methods have been employed in published biomarker studies in which NDs were present. The table indicates that the results for the ML-based CI method differ very little from those produced by the various substitution methods, with the exception of random substitution. However, there is quite a discrepancy between the results for the MI-based CI and the CIs obtained using the either the Spearman or Kendall coefficients.

How does one decide which confidence interval method is most appropriate for this set of data? We recommend using the summary results presented in Tables 1-3 to help make this decision. With regard to Table 1, the censoring proportions in this example are $p_1 = 7\%$ (4/54) for hippuric acid and $p_2 = 72\%$ (39/54) for *o-cresol*. The closest censoring proportions in Table 1 are $(p_1, p_2) = (0.1, 0.7)$. For $(X, Y)$ data that do not appear to satisfy the BVN assumption (as in this example), CIs based on either the MLE or Spearman's $r_s$ maintain acceptable CP under $(p_1, p_2) = (0.1, 0.7)$: 93.9% for ML and 92.9% for $r_s$. With regard to Table 2 (non-BVN section), CIs based on the MLE achieve acceptable CP (93.4%) when the true value of $\rho$ is approximately 0.75, which is a reasonable assumption based on the results in Table 4 ( $\hat{\rho}_{ML} = 0.79$ ). Similarly, CIs based on Spearman's $r_s$ (as an estimate of $\rho_s$) achieve acceptable CP (93.4%) when the true value of $\rho_s$ is approximately 0.5, which appears to be a reasonable assumption based on the results in Table 4 ($r_s = 0.58$). Finally, with regard to Table 3 (non-BVN section), we see that CIs based on either the MLE (93.8%) or Spearman's $r_s$ (93.7%) achieve acceptable CP when $n$ is approximately 50, which it is in this example ($n = 54$). Thus, McCracken's simulation results (as summarized in Tables 1 - 3) give no indication that we should doubt the validity of either the ML-based CI or the $r_s$ -based CI ($r_s$ considered as an estimate of $\rho_s$). Given the apparent strong departure from BVN for these data, and that the authors were evaluating the validity of *o*-cresol by examining its association with hippuric acid (not necessarily its linear association), we conclude that Spearman's coefficient is preferable to Pearson's as a measure of the association between $X$ and $Y$ for this biomarker study. The choice of Spearman's as the measure of association for this study is also consistent with Amorim and Alvarez-Leite's use of the nonparametric Kruskal-Wallis test in their comparison of the *o*-cresol concentrations among the three

groups of toluene-exposed individuals included in their study: workers in shoe factories, painting sectors of metal industries, and printing shops. Thus, if we were analyzing these data, we would report the results for Spearman's coefficient: $\hat{\rho}_s = 0.58$, 95% CI (0.34, 0.82). These results suggest that the true association between the urinary concentrations of *o*-cresol and hippuric acid among workers exposed to toluene appears to be quite a bit weaker than that reported by Amorim and Alvarez-Leite ($r = 0.777$).

In general, if the investigators are primarily interested in the linear association between two biomarkers, we would recommend that the ML-based CI be used instead of the one based on $r_s$. Even if there is evidence that the ($X$, $Y$) data do not follow a BVN distribution, McCracken's simulation results show that CIs based on the MLE are still generally preferable to all of the other methods for finding a CI for $\rho$ when NDs are present in both $X$ and $Y$, regardless of the true bivariate distribution. Unless there is reason to doubt the validity of the ML-based CI (as indicated in Tables 1-3), it would generally be acceptable to use the ML-based CI for the Pearson correlation if the primary goal of the analysis is to determine the degree of linear relationship between the two biomarkers.

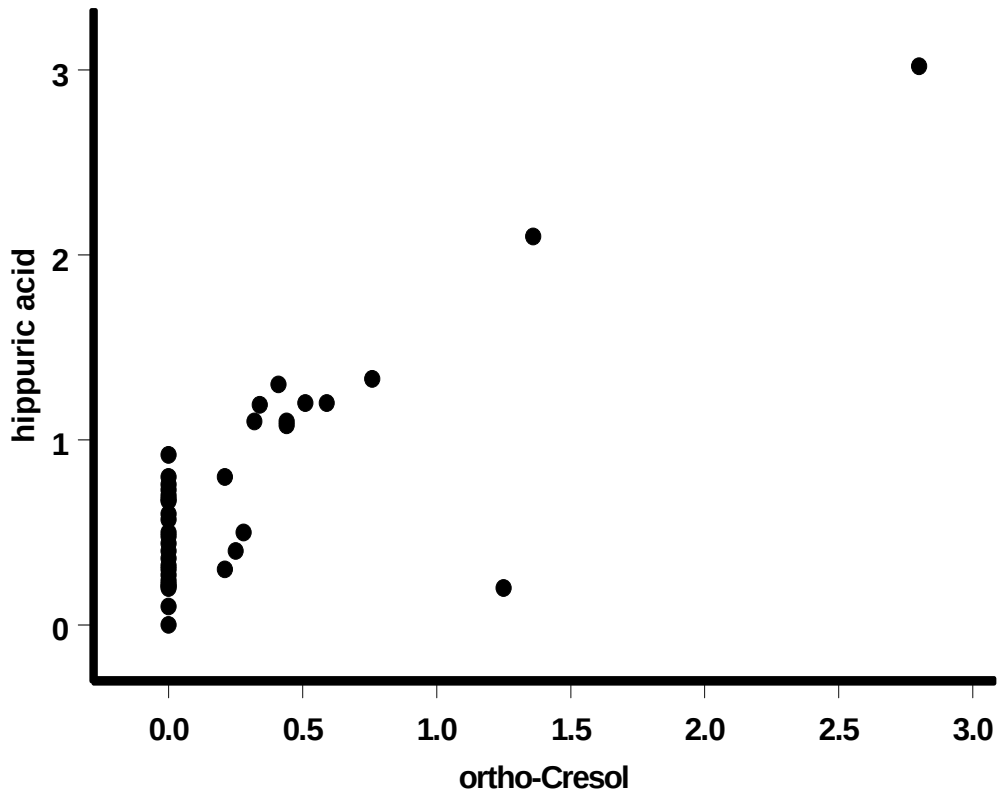


Figure 1. Scatterplot of *o-cresol* vs. hippuric acid concentrations in urine samples of 54 individuals exposed to toluene in shoe factories, painting sectors of metal industries, and printing shops. Observations below the detectable limit of either assay are plotted as zero for purposes of illustration only.

Table 4. Summary of Results for Data from Amorim and Alvarez-Leite (1997)

| Method | *n* | Estimate | *p*-value | 95% CI |
|---|---|---|---|---|
| Complete cases | 15 | 0.76 | < 0.001 | (0.40, 0.92) |
| Substitute zero | 54 | 0.79 | < 0.001 | (0.67,0.88) |
| Substitute *LOD* | 54 | 0.77 | < 0.001 | (0.63, 0.85) |
| Substitute *LOD*/2 | 54 | 0.79 | < 0.001 | (0.65, 0.87) |
| Substitute $LOD/\sqrt{2}$ | 54 | 0.78 | < 0.001 | (0.64, 0.86) |
| Random Substitution | 54 | 0.43 | < 0.001 | (0.17, 0.61) |
| Complex Substitution | 54 | 0.78 | < 0.001 | (0.65, 0.86) |
| Spearman | 54 | 0.58 | < 0.001 | (0.34, 0.82) |
| Kendall | 54 | 0.49 | < 0.001 | (0.12, 00.74) |
| Maximum Likelihood | 54 | 0.79 | < 0.001 | (0.66, 0.87) |

The R code used to produce the above results for all of the methods except "substitute zero" and Kendall's coefficient can be obtained from the second author. These results for "substitute zero" and Kendall's coefficient were obtained using the CORR procedure in SAS.

## References

1. Amorin, L., and Alvarez-Leite, E. (1997), "Determination of *o-cresol* by Gas Chromatography and Comparison with Hippuric Acid Levels in Urine Samples of Individuals Exposed to Toluene," *Journal of Toxicology Environmental Health*, 50, 401-407.
2. Atawodi, S.E., Lea, S., Nyberg, F., Mukeria, A., Constantinescu, V., Ahrens, W., Brueske-Hohlfeld, I., Fortes, C., Boffetta, P., and Friesen, M.D. (1998), "4-Hydroxyl-1-(3-pyridyl)-1-Butanone-Hemoglobin Adducts as Biomarkers of Exposure to Tobacco Smoke: Validation of a Method to be Used in Multicenter Studies," *Cancer Epidemiology Biomarkers and Prevention*, 7, 817–821.
3. Bradley, J.V. (1978), "Robustness?" *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
4. Cook, D.G., Whincup, P.H. , Papacosta, O., Strachan, D.P., Jarvis, M.J., and Bryant, A. (1993), "Relation of Passive Smoking as Assessed by Salivary Cotinine Concentration and Questionnaire to Spirometric Indices in Children," *Thorax*, 48, 14-20.
5. Gibbons, J., and Chakraborti, S. (2003). *Nonparametric Statistical Inference* (4th Edition ed.). New York: Marcel Dekker Inc.
6. Lagorio, S., Crebelli, R., Ricciarello, R., Conti, L., Iavarone, I., Zona, A., Ghittori, S., and Carere, A. (1998), "Methodological Issues in Biomonitoring of Low Level Exposure to Benzene," *Occupational Medicine*, 8, 497-504.
7. Li, L., Wang, W., and Chan, I. (2004), "Correlation Coefficient Inference on Censored Bioassay Data," *Journal of Biopharmaceutical Statistics*, 15, 501-512.
8. Lyles, R.H., Williams, J.K. and Chuachoowong, R. (2001), "Correlating Two Viral Load Assays With Known Detection Limits," *Biometrics*, 57, 1238-1244.
9. Lynn, H. (2001), "Maximum Likelihood Inference for Left-Censored HIV RNA Data," *Statistics in Medicine*, 20, 33-45.
10. McCracken, C.E., (2013), "Correlation Coefficient Inference for Left-Censored Biomarker Data with Known Detection Limits," unpublished Ph.D. dissertation, Georgia Regents University, Dept. of Biostatistics.
11. Newton, E., and Rudel, R. (2007), "Estimating Correlation With Multiply Censored Data Arising From the Adjustment of Singly Censored Data," *Environmental Science and Technology*, 41, 221-228.
12. Wang, H. (2006), "Correlation Analysis for Left-Censored Biomarker Data With Known Detection Limits," unpublished Masters thesis, Louisiana State University Health Sciences Center, School of Public Health, Biostatistics Program.