Received: 7 February 2020 | Revised: 26 December 2020 | Accepted: 3 April 2021

DOI: 10.1002/sim.8995

RESEARCH ARTICLE

Statistics in Medicine WILEY

Measuring association among censored antibody titer data

Thao M. P. Tran[1] | **Steven Abrams**[1,2] | **Marc Aerts**[1] | **Kirsten Maertens**[3] | **Niel Hens**[1,2,4]

[1]I-BioStat, Data Science Institute, Hasselt University, Hasselt, Belgium

[2]Global Health Institute, Family Medicine and Population Health, University of Antwerp, Antwerp, Belgium

[3]Centre for Evaluation of Vaccination, Vaccine and Infectious Disease Institute, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium

[4]Centre for Health Economics Research and Modeling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium

Correspondence
Thao M. P. Tran, I-BioStat, Data Science Institute, Hasselt University, Hasselt, Belgium.
Email: maiphuongthao.tran@uhasselt.be

Funding information
Bijzonder Onderzoeksfonds, Grant/Award Number: BOF11NI31; Fonds Wetenschappelijk Onderzoek, Grant/Award Number: 12R5719N; H2020 European Research Council, Grant/Award Number: 68250 - TransMID; Methusalem research grant from the Flemish government: BOF08M01

Censoring due to a limit of detection or limit of quantification happens quite often in many medical studies. Conventional approaches to deal with censoring when analyzing these data include, for example, the substitution method and the complete case (CC) analysis. More recently, maximum likelihood estimation (MLE) has been increasingly used. While the CC analysis and the substitution method usually lead to biased estimates, the MLE approach appears to perform well in many situations. This article proposes an MLE approach to estimate the association between two measurements in the presence of censoring in one or both quantities. The central idea is to use a copula function to join the marginal distributions of the two measurements. In various simulation studies, we show that our approach outperforms existing conventional methods (CC and substitution analyses). In addition, rank-based measures of global association such as Kendall's tau or Spearman's rho can be studied, hence, attention is not only confined to Pearson's product-moment correlation coefficient capturing solely linear association. We have shown in our simulations that our approach is robust to misspecification of the copula function or marginal distributions given a small association. Furthermore, we propose a straightforward MLE method to fit a (multiple) linear regression model in the presence of censoring in a covariate or both the covariate and the response. Given the marginal distribution of the censored covariate, our method outperforms conventional approaches. We also compare and discuss the performance of our method with multiple imputation and missing indicator model approaches.

KEYWORDS
antibody titers, association, geometric mean concentration, left-censored data, maximum likelihood inference

1 | INTRODUCTION

In clinical trials involving vaccination efficacy evaluation or in studies investigating the immunity of a (specific) population to a (particular) disease, antibody titers to specific antigens are of primary interest. It is quite common that there is a limit of quantification (LOQ) or limit of detection (LOD) present in a test sample. The LOD is defined as the "*smallest measured concentration of an analyte from which it is possible to deduce the presence of the analyte in the test sample with*

acceptable certainty",[1] while the LOQ is defined as the "*smallest measured content of an analyte above which the determination can be made with the specified degree of accuracy and precision*".[1] Consequently, antibody titers below a specified threshold (usually provided by the manufacturers) are only known up to the LOQ or LOD. The LOD and LOQ can be similar, but the LOD can also be at a much higher concentration than the LOQ.[2] In addition to left censoring, one might encounter right or interval censoring, which makes the analysis and inference more complicated. An observation is right-censored if its value is not known exactly but it is quantified to be above a certain threshold. Right censoring might occur in cases of extremely high antibody response. Last, interval censoring is present if one cannot determine the exact value of the variable of interest but one only knows that it falls into a certain interval. In this article, we illustrate the proposed method for left-censored data, and we note that the method can be used to deal with right and interval censoring.

While reporting results from vaccine trials, authors have focused on comparing geometric mean concentrations (GMCs) of antibody titers derived from serum samples collected at the same time point between different study groups. The GMC is calculated by taking the n^{th} root of the product of all individual values (n being the number of subjects in the trial). For the GMC results to be interpreted correctly, researchers rely on the assumption that the data on a log-scale are symmetrical.[3-5] The question is whether assuming that antibody titers exhibit a log-normal distribution is always appropriate, or whether other distributions for a nonnegative random variable should be used instead.

When censoring is present due to actual values falling below the LOD (or LOQ), investigators often impute censored observations using the LOD (or LOQ), LOD/2 (or LOQ/2), or LOD/$\sqrt{2}$ (LOQ/$\sqrt{2}$).[6] They then perform analysis on the imputed dataset (substitution method). Another conventional approach is to remove censored observations and analyse data without censoring (complete case analysis, abbreviated as CC analysis). While the CC analysis might lead to biased estimates and loss of efficiency,[7] the performance of a substitution method depends on the setting under consideration. Nonetheless, generally speaking, the imputation method tends to be biased as well.[8-10] Shaarawi and Esterby[8] derived the formulae to calculate the large sample bias for the mean as a function of the true geometric standard deviation (GSD), GMC, the true percentage of censoring, and the sample size. When the sample size increases, the bias asymptotically approaches a fixed value. Furthermore, the substitution approach has no theoretical foundation.[9] In addition to these two conventional approaches, maximum likelihood estimation (MLE) and the nonparametric method (NP) are frequently used. The MLE method is based on maximizing the likelihood function calculated from the probability density of a given distribution, while taking into account the proportions of censored and noncensored data to estimate the sample GMC and GSD. The NP approach recommended by Schmoyeri et al[11] is based on the application of the Kaplan-Meier method to estimate the sample mean when there is censoring present in the data. The MLE method was shown to perform well in many circumstances. On the other hand, the NP approach performed better than the substitution method, but still did poorly when compared with the MLE method, at least in the field of microbial risk assessment.[9]

This article focuses on two aspects: (a) estimating the association between two variables under the presence of censoring in one or both variables and (b) estimating the coefficients of a linear regression model in the presence of censoring in one covariate and/or the response using the MLE approach. Concerning the first aspect, the MLE method based on the assumption of an underlying distribution is commonly used. For example, Lynn,[12] Lyles et al[13,14] proposed their MLE methods on the basis of the bivariate normality assumption. More specifically, Lynn[12] proposed an MLE approach with the assumption of a bivariate normality of the two measurements (HIV RNA concentrations) on a log scale. Similarly, Lyles et al,[13,14] and Benning et al[15] proposed a bivariate normal distribution to quantify the association between two measurements when one or both of them was subject to left censoring. Both approaches relied on the normality assumption, and their performance with nonnormal data has not been investigated. Furthermore, these authors studied only the Pearson product-moment correlation, which is a measure of linear association. To relax the normality assumption, Song et al[16] proposed a generalized estimating equations approach to estimate the Pearson's correlation coefficient. The authors showed the robustness of their method to departures from the normality assumption. This approach, however, is only applicable in cases of censoring in one variable. The extension to data with censoring in both variables is not obvious. In this article, we propose a parametric approach employing a bivariate copula to join the two marginal distributions. The method allows for the simultaneous estimation of both the association and the two marginal distributions while accounting for censoring. Our method aims at estimating the association in a general setting. Depending on the copula used, the association can be described using Pearson's correlation coefficient, Spearman's rho, or Kendall's tau. Additionally, we study the robustness of the proposed method under misspecification of the copula function, or the marginal distribution(s).

The second objective of this article is to make appropriate inferences for a linear regression model in the presence of censoring in one covariate and/or the response. When there is censoring in the response, one can utilize a modified regression model to account for censoring (eg, Tobit regression proposed by Tobin[17]) to provide unbiased estimates. With regard to the case of censoring in one covariate, Tsimikas et al[18] proposed a quasi-score based method that relied

on an estimating function for fitting a generalized linear regression model. The method did not require the specification of a parametric distribution for the response. For linear regression, the method implied the employment of mean imputation of the censored regressor. Recently, Atem et al[19] introduced the idea of combining the imputation and survival regression methods to deal with randomly censored covariates. Multiple imputation (MI) has been used and investigated in many settings, see, for example, Lyles et al,[14] Wei et al,[20,21] Stekhoven and Bühlmann[22] among others. Besides the aforementioned methods, many authors have adapted the missing indicator approach (MID) as proposed by Jones[23] to account for a censored covariate(s). The idea of the MID approach is to introduce a binary variable indicating whether the explanatory variable is observed and include it as a covariate in the model. For example, Chiou et al[24] made use of the MID approach to deal with censored covariates due to the LOD in logistic regression models. Overall, few studies devote attention to methods dealing with censoring in both response and covariate(s). Recently, Jones[25] nominated a pseudo-likelihood approach utilizing plug-in estimators of lower-level parameters (the variance-covariance matrix and mean vector) to estimate higher-level parameters (ie, the regression coefficients) in the context of left censoring that is present in both the covariate and response. We, however, propose a relatively straightforward MLE approach for fitting a regression model with censoring in both the response and covariate variables. In our approach, one assumes that the underlying distribution of the covariate with censoring is known. Under censoring, the likelihood contribution can be obtained by relying on the cumulative distribution function of the censored covariate.

This article is organized as follows: We first introduce our motivating examples in Section 2. In Section 3, the proposed methodology is described in detail. Next, a simulation study is performed to evaluate our proposed method in different scenarios (Section 4). In Section 5, we show the results of applying the new method to the data applications introduced in Section 2. We formulate conclusions and potential avenues for further research in Section 6.

2 | MOTIVATING EXAMPLES

2.1 | Pertussis data from Thailand

Two datasets from two clinical trials inspired the work in this article. The first trial is a prospective, randomized, controlled clinical trial that examined the effect of maternal Tdap (tetanus, diphtheria, acellular pertussis) vaccination on the humoral immune response of infants to acellular pertussis (aP) and whole-cell pertussis (wP) vaccines. The trial was conducted in Bangkok, Thailand. Investigators offered a Tdap vaccine to participating women who were between 26 and 36 weeks of gestation. Study staff collected blood samples from the pregnant women at delivery. In infants, blood samples were taken at birth, at month two (right before the first vaccine dose), and at month seven (1 month after the primary series of three doses of a hexavalent pertussis-containing vaccine). Antibody titers against pertussis (IgG anti-PT, IgG anti-FHA, and IgG anti-PRN) were measured "to assess the influence of vaccine-induced maternal antibodies on the humoral immune response after the administration of different pertussis-containing vaccines to the infant".[26]

More details regarding the study design, study laboratory, and various results of this study can be found in Wanlapakorn et al.[26,27] The GMCs of antibody titers at different time points were calculated to study the differences between both infant groups (aP and wP). Moreover, investigators were also interested in investigating which demographic and clinical factors might influence antibody titer concentrations. Based on the protocol, possible confounders could be birth weight, length at birth, feeding (bottle versus breastmilk), age of the mother, gestational age at the time of vaccination, time from vaccination to delivery, and recent pertussis vaccination of the women. Since there is no recent recommendation to vaccinate adults or pregnant women with a pertussis-containing vaccine in Thailand, we excluded the last confounder from the analysis.

There is an LOD of 5 IU/mL for all three types of antibodies leading to left-censored observations. In the initial analysis, the substitution method with LOD/2 was considered. The analysis was then performed on the imputed data. Here, we will reexamine the association between antibody titers in the cord and at month two, as well as fit a linear regression model to investigate factors affecting the antibody titers in infants at month two using the proposed methods. We compare the proposed methods' results with the CC analysis, substitution method, MI, and MID approaches.

2.2 | Varicella-zoster virus data from Belgium

The second dataset was collected in a prospective multicenter study of pregnant women aged between 18 and 40 years and their offspring. One of the key research objectives was to study the kinetics of maternal antibodies against varicella-zoster

virus (VZV) in infants. VZV is a human alpha-herpesvirus causing varicella (chickenpox) by the first infection, then becoming latent in the human body. Later on, VZV can become reactivated and cause many neurologic diseases such as herpes zoster, postherpetic neuralgia in people with a decline in cell-mediated immunity, or people with a weakened immune system.[28] The trial was conducted in Antwerp (Belgium) from 2006 to 2008. Details on the study design, data collection, and initial analysis are published in Leuridan et al.[29] Briefly, blood samples were taken in both pregnant women and their infants to measure antibody titer concentrations with regard to varicella, among other pathogens. In infants, blood samples were collected at 1, 3, and 12 months of age and randomly at either 6 or 9 months of age. The antibody concentrations were expressed in mIU/mL. There was an LOD of 50 mIU/mL, implying that values below 50 were considered to be left-censored. In this article, we estimated the correlation between antibody titers against VZV at month 9 and month 12 of age in infants since there was a large percentage of censoring in these two measurements.

3 | MATERIALS AND METHODS

We propose to join the two marginal distributions of the variables using a copula function and thereby estimate the association between them. Hence, in this section, we first introduce the bivariate copula function used throughout this article. Next, we briefly discuss the approach for relating two continuous variables while accounting for (left) censoring in one or both variables. Last, we elaborate on a method to fit a (simple) linear regression model when there is (left) censoring in the covariate or both covariate and response.

3.1 | Copula functions

A function $C : [0,1] \times [0,1] \to [0,1]$ is called a copula function when it has the following properties:

(i) For every $u, v \in [0,1]$: $C(u,0) = C(0,v) = 0$ and $C(u,1) = u$ and $C(1,v) = v$.
(ii) For every $u_1, u_2, v_1, v_2 \in [0,1]$ such that $u_1 \le u_2$ and $v_1 \le v_2$: $C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \ge 0$.

From these properties, we note that the bivariate copula function C is the joint (cumulative) distribution function of a two-dimensional random vector (U,V) on the unit cube $[0,1] \times [0,1]$ with uniform margins. Furthermore, from Sklar's theorem,[30] it follows that for any continuous random vector (X,Y) with joint distribution H and respective marginals F and G, there exists a unique copula function C such that for all $x, y \in R$,

$$H(x,y) = C[F(x), G(y)].$$

Several families of copulas have been described in the literature. Among them, Gaussian and Archimedean copulas are quite popular and widely used in real-life applications, see, for example, Shih and Louis,[31] Bárdossy and Li,[32] Danaher and Smith,[33] de Leon and Wu,[34] Song et al,[35] Frahm et al,[36] Shi and Yang.[37] The bivariate Gaussian copula function takes the form:

$$C(u_1, u_2; \rho) = \Phi_2[\Phi^{-1}(u_1), \Phi^{-1}(u_2)].$$

with Pearson product-moment correlation $-1 \le \rho \le 1$ among the marginal random variables. Here, $\Phi_2(\cdot)$ denotes the joint cumulative distribution function of a bivariate normal distribution with correlation ρ and $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal distribution. In addition, the Archimedean copula family is one of the most popular copula families for modeling bivariate survival data. A bivariate Archimedean copula can be represented by:

$$C(u_1, u_2; \vartheta) = \phi_\vartheta[\phi_\vartheta^{-1}(u_1) + \phi_\vartheta^{-1}(u_2)],$$

where $0 \le \phi_\vartheta(\cdot) \le 1$, $\phi_\vartheta(0) = 1$, $\phi'_\vartheta(\cdot) < 0$, and $\phi''_\vartheta(\cdot) > 0$ (see, eg, Reference 30).

Some famous members of the Archimedean copula family are the Clayton, Gumbel, Joe, and Frank copula.[31] Interested readers are referred to Web Appendix A (Supplementary Materials) for more details regarding these copula functions.

3.2 | Estimation of the association and marginal distributions

In the following, we will focus on distributions for nonnegative random variables in the context of our data application, even though our proposed methods apply to any real-valued random variables. Let X and Y denote the antibody titer measurements (for either pregnant women or infants) at a specific time point. In the literature, antibody titers are usually assumed to follow a log-normal distribution and are typically summarized in terms of GMC and its 95% confidence interval. The log-normal distribution is commonly used to model antibody titers in clinical trials that involve the evaluation of vaccine efficacy, see, for example, References 38,39. It has been claimed that generalized linear models assuming a log-normal distribution and a gamma distribution to analyse antibody titers are interchangeable. However, Wiens[40] showed, via a case study, that log-normal and gamma models can lead to different results. Consequently, here, we consider not only a log-normal but also a gamma distribution while modeling antibody titer data. We denote X and Y the two nonnegative variables following two distributions with cumulative distribution functions F and G where special attention is directed towards log-normal and gamma distributions. Without loss of generality, we let $X \sim LN(\mu, \sigma)$ (LN denotes a log-normal distribution with μ and σ being the mean and standard deviation of the data on a log-scale), and $Y \sim \Gamma(\alpha_1, \alpha_2)$ (α_1 and α_2 denote the shape and scale parameters). This parameterization implies that $\mathbf{E}(X) = \exp(\mu + \sigma^2/2), \mathbf{Var}(X) = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]$, and $\mathbf{E}(Y) = \alpha_1\alpha_2, \mathbf{Var}(Y) = \alpha_1\alpha_2^2$.

Let $u = F(X)$ and $v = G(Y)$ denote the cumulative distribution functions for X and Y. Then, $0 \le u, v \le 1$. We assume further that there is a copula joining the two marginal distributions. That is $H(X, Y) = C(F(X), G(Y)) = C(u, v)$. Suppose that one is interested in measuring the association between two variables in the presence of censoring. We denote the total sample size by N. Those X values less than x_{cens} and Y values less than y_{cens} are (left-)censored. The whole dataset is divided into four subsets comprising different censoring patterns. The first subset contains observations with censoring in X only (sample size n_1). The second subset includes observations with censoring in Y only (sample size n_2). The third subset consists of observations with censoring in both X and Y (sample size n_3). Finally, complete observations belong to the fourth subset with the sample size being equal to $N - n_1 - n_2 - n_3$.

We rely on the MLE approach to estimate all parameters of interest. For the set of $(N - n_1 - n_2 - n_3)$ complete observations, the density of the bivariate distribution is given by $f(x, y) = c(F(x), G(y))f(x)g(y)$, where $c(F(x), G(y))$ is the bivariate copula density function. For the set of n_1 (left-)censored observations in X, we have: $f_{X \le x, Y}(x_{\text{cens}}, y) = \int_0^{x_{\text{cens}}} f(x, y)dx$. Similarly, for the set of n_2 (left-)censored observations in Y, we have: $f_{X, Y \le y}(x, y_{\text{cens}}) = \int_0^{y_{\text{cens}}} f(x, y)dy$. And finally, the log-likelihood contribution for the n_3 (left-)censored observations in both X and Y is given by $f_{X \le x, Y \le y}(x_{\text{cens}}, y_{\text{cens}}) = \int_0^{x_{\text{cens}}} \int_0^{y_{\text{cens}}} f(x, y)dxdy$.Hence, we can write down the log-likelihood function as follows:

$$\begin{aligned} LL_a = &\sum_{i=1}^{N-n_1-n_2-n_3} \log[f(x_i, y_i)] + \sum_{j=1}^{n_1} \log\left[\int_0^{x_{\text{cens}}} f(x, y_j)dx\right] \\ &+ \sum_{k=1}^{n_2} \log\left[\int_0^{y_{\text{cens}}} f(x_k, y)dy\right] + n_3 \log\left[\int_0^{x_{\text{cens}}} \int_0^{y_{\text{cens}}} f(x, y)dxdy\right]. \end{aligned} \tag{1}$$

The log-likelihood function in case of censoring in a single variable can be obtained as a special case of Equation (1).

3.3 | Regression analysis

Sometimes, the primary interest lies in fitting a regression model to make inference about the relationship between two variables or to construct a prediction model for individual predictions of the mean outcome conditional on specific covariate information. Here, we demonstrate the method to fit a linear regression model where there is (left) censoring in either covariate (X) or both covariate and response (X and Y) making use of the marginal distribution of the censored covariate X.

Suppose that we want to investigate the linear relationship between Y and X, that is, to estimate the coefficients β_0 and β_1 in which $Y|(X = x) \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$. We group the dataset into four subsets according to (X, Y)'s censoring profiles similar to what has been done in Section 3.2. The log-likelihood function is given by:

$$LL_r = \sum_{i=1}^{N-n_1-n_2-n_3} \log\left\{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right]\right\}$$

$$+\sum_{j=1}^{n_1}\log\left\{\frac{1}{F_X(x_{\text{cens}})}\int_0^{x_{\text{cens}}}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y_j-\beta_0-\beta_1 x)^2}{2\sigma^2}\right]f_X(x)dx\right\}$$
$$+\sum_{k=1}^{n_2}\log[P(y\le y_{\text{cens}}|X=x_k))]$$
$$+\sum_{l=1}^{n_3}\log\left\{\frac{1}{F_X(x_{\text{cens}})}\int_0^{x_{\text{cens}}}P(y\le y_{\text{cens}}|X=x)f_X(x)dx\right\}. \quad (2)$$

Readers are referred to Web Appendix A (Supplementary Materials) for the details of the log-likelihood contributions of different components corresponding to distinctive subset groups of data. When censoring is present in the covariate only, the log-likelihood function can be obtained as a particular case of this setting. Moreover, the proposed method can be extended for use in a multiple linear regression model. Note that, although the specification of the distribution of X is unnecessary in a classical linear regression approach, the distributional choice thereof is required in our method.

4 | SIMULATION STUDY

To gain deeper insights into the proposed methodology's performance, we conducted an extensive simulation study governing different simulation scenarios. We used the methods described in Sections 3.2 and 3.3 to analyse the simulated data. Simulation results with respect to the estimation of the association measure were compared across the existing methods, that is, CC analysis, and substitution by LOD/2 and LOD, and our approach.

For linear regression, we contrasted the proposed method's outputs with the two conventional approaches as well as with the MID and MI methods. Details regarding the MID can be traced back to the work of Jones.[23] Briefly speaking, in the case of a simple linear regression setting with a censored covariate, the linear regression model is specified as $Y|(X=x,\Delta=\delta)\sim\mathcal{N}(\beta_0+\beta_1 x(1-\delta)+c\delta,\sigma^2)$. Here, Δ is the censoring indicator, taking value $\delta=1$ if censored and $\delta=0$ otherwise. Under this model specification, c is a nuisance parameter. When censoring is present in both covariate and the response, we incorporate the Tobit adjustment to the aforementioned model. On the subject of the MI approach, we proposed a method to impute censored covariate based on the assumption that $X\sim F(.)$. In the current simulation setting, X values were generated from the truncated gamma distribution at the LOD such that $x\le$ LOD. However, to assure the relationship we observed between X and Y, we reordered the values of X to preserve the (negative/or positive) association between them. The number of imputed datasets was equal to the percentage of censoring present in the data. The proposed imputation method provides valid inference under censoring at random.

4.1 | Simulation: (left) censoring in one variable

4.1.1 | Estimating the association

In the first simulation setup, we generated bivariate (X,Y) values from a Gaussian copula and two marginal gamma distributions: $X\sim\Gamma(\alpha_1=4,\alpha_2=0.8)$, $Y\sim\Gamma(\beta_1=24,\beta_2=1)$. Here, the gamma distributions were specified in terms of the shape and scale parameters as defined in Section 3.2. A small correlation ($\rho=0.2$) and a large one ($\rho=0.8$) were considered together with various percentages of censoring in X ranging from very small (approximately 10%) to moderate (30%), medium (50%), and large (75%). For each scenario, 500 datasets were simulated with a sample size of $n=240$ observations each. This sample size was motivated by the small to moderate sample sizes in our data applications.

The simulation results are shown in Table 1. In all scenarios, the CC and substitution methods provided biased estimates for α_1,α_2, that is, the distributional parameters associated with the marginal distribution of X (ie, the variable for which censoring is present) as well as for the association parameter. Generally, compared with the LOD substitution, the LOD/2 substitution method gave less biased estimates, at least for moderate to large censoring percentages. The estimates for β_1 and β_2, that is, the parameters associated with the distribution of Y, are essentially unbiased when using the substitution approach but are biased in the CC analysis. Generally, our proposed method gave unbiased estimates for all parameters. In addition, these estimates were very similar to those obtained when we analyzed all data (data without censoring). The MSE values were highest in the CC analysis irrespective of the parameters under study. The substitution

method provided MSE values that were higher than those obtained using our method, at least for parameters associated with the distribution of X. Apparently, the estimation of α_1, α_2 appears to be better than those of β_1, β_2 in terms of bias. However, this observation is likely because the true values for β parameters are larger than those for α parameters. Indeed, the relative bias values (bias divided by the true value) in Web Table 1 (Supplementary Materials) confirmed that the estimation of α parameters was not better than those of β parameters.

To see the consequence of underestimating or overestimating one or both parameters of a specific gamma distribution, we looked at the corresponding means and variances of X and Y given that $E(X) = \alpha_1\alpha_2, E(Y) = \beta_1\beta_2$ and $\text{Var}(X) = \alpha_1\alpha_2^2, \text{Var}(Y) = \beta_1\beta_2^2$ (see Web Table 2 in the Supplementary Materials). As expected, our proposed method provided estimates that were close to the true values (with small biases) for the mean and variance of the two variables. The CC and substitution analyses gave small to large bias depending on the scenarios. The results above hold when we correctly specify the copula function and the two marginal distributions. However, in a real data application, it is challenging to select the right copula function or marginal distributions. Hence, in the subsequent analysis, we investigated how misspecification of the copula function or marginal distribution(s) affected the parameters' estimates.

In order to assess the robustness of our proposed method against the misspecification of the copula function, we reanalyzed the simulated datasets using two correctly specified gamma marginal distributions and a misspecified copula function. More specifically, we used a Frank, Clayton, Gumbel, or Joe copula instead of a Gaussian copula. Web Figure 1 (Supplementary Materials) shows boxplots of empirical estimates for all parameters across 500 simulated datasets. The model with a Frank copula function performed reasonably well compared with the correctly specified model for all censoring percentages under a small association scenario. In approximately 37% of the runs (small association), and 2% of the simulations (large association), the Akaike information criteria (AIC) values obtained using the correctly specified model were higher than those of the model using a Frank copula function. On the other hand, models using a Clayton, Gumbel, or Joe copula function gave good estimates for the parameters of the two marginal distributions, but not for the association measure when the association was small. When the association was large, the misspecified copula model produced biased estimates. The bias was least in the model using the Frank copula model among all.

The subsequent sensitivity analysis was conducted to examine the performance of our approach under the misspecification of one or both marginal distributions. Instead of specifying two gamma marginal distributions, we used a marginal log-normal distribution for either the variable with censoring, the variable without censoring, or both variables (see Web Table 3 in the Supplementary Materials). When one marginal distribution was misspecified, the proposed method still provided estimates that were very close to the true values regarding the other marginal distribution and the association parameter. When both marginal distributions were misspecified, our approach still performed well in estimating the association between the two variables.

4.1.2 | Linear regression

To evaluate the performance of our proposed method in the case of linear regression, a simulation study was performed where $X \sim \Gamma(24, 2)$ and $Y|(X = x) \sim \mathcal{N}(24 + 1.5x, \sigma^2)$ with $\sigma^2 = 1.44$, the variance of the random noise generated from a normal distribution with mean 0. The percentage of censoring was varied from small to large. The results were reported based on 500 simulated datasets in terms of mean estimate (and corresponding empirical standard errors), bias, and MSE for all parameters $(\beta_0, \beta_1, \sigma)$. Moreover, the coverage probabilities for β_0, β_1, and Type II error rates for β_1 (the probabilities of incorrectly not rejecting β_1) were shown as well.

Table 2 shows the simulation results. Regarding the level of bias, our method performed well in all scenarios with different censoring percentages. The two substitution methods provided biased estimates for all parameters, which occurred even with a small percentage of censoring. The bias was more severe when the censoring fraction increased. The performance of the CC analysis was reasonable in all scenarios. More specifically, this approach gave estimates close to the true values, which is not surprising since the CC analysis, though less efficient, provides unbiased estimates when censoring is noninformative (ie, when the censoring does not depend on Y).[41] The MI and MID approaches, in comparison with our method, performed equally well in terms of bias (except for the estimate of σ where the MID approach gave larger bias). With respect to coverage probabilities, the MI method's performance was comparable to our approach while the MID provided higher coverage probabilities for both β_0 and β_1 in all scenarios. However, the SE and MSE values under MI and MID were consistently higher than those of the proposed method. When looking at the widths of confidence intervals produced under MI, MID, and our approach, the proposed method induced the smallest numbers (see Web Figures 3 in the Supplementary Materials). Although the CC analysis provided reasonable estimates for c, the MSE values were larger

TABLE 1 Simulation results

	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE
Method	$\alpha_1 = 4$			$\alpha_2 = 0.8$			$\beta_1 = 24$			$\beta_2 = 1$			$\rho = 0.2$		
Censoring 10%															
All data	0.0535	0.3844	0.1503	−0.0021	0.0794	0.0063	0.3294	2.2006	4.9416	−0.0050	0.0917	0.0084	−0.0013	0.0610	0.0037
Our	0.0389	0.4160	0.1742	0.0020	0.0863	0.0074	0.3294	2.2006	4.9416	−0.0050	0.0917	0.0084	−0.0010	0.0615	0.0038
CC	1.9412	0.5299	4.0485	−0.2146	0.0594	0.0496	0.8007	2.3433	6.1213	−0.0156	0.0942	0.0091	−0.0303	0.0664	0.0053
LOD	0.8452	0.3860	0.8631	−0.1264	0.0619	0.0198	0.3300	2.2006	4.9420	−0.0049	0.0917	0.0084	−0.0041	0.0610	0.0037
LOD/2	−0.5881	0.3342	0.4574	0.1386	0.0910	0.0272	0.3289	2.2007	4.9414	−0.0051	0.0916	0.0084	−0.0043	0.0609	0.0037
Censoring 30%															
Our	0.0453	0.4734	0.2257	0.0030	0.0944	0.0089	0.3242	2.1946	4.9118	−0.0048	0.0914	0.0084	−0.0005	0.0630	0.0040
CC	4.2974	0.8405	19.1724	−0.3403	0.0539	0.1187	1.2180	2.6109	8.2865	−0.0216	0.1027	0.0110	−0.0528	0.0750	0.0084
LOD	2.4752	0.5483	6.4265	−0.2795	0.0529	0.0809	0.3305	2.2006	4.9421	−0.0049	0.0917	0.0084	−0.0109	0.0614	0.0039
LOD/2	−0.9177	0.2405	0.9000	0.2113	0.0794	0.0509	0.3291	2.2006	4.9414	−0.0050	0.0917	0.0084	−0.0077	0.0612	0.0038
Censoring 50%															
Our	0.0632	0.6283	0.3980	0.0060	0.1182	0.0140	0.3294	2.2007	4.9417	−0.0050	0.0917	0.0084	−0.0020	0.0688	0.0047
CC	10.6657	2.2137	118.6490	−0.4861	0.0514	0.2390	1.9075	3.4004	15.1779	−0.0264	0.1295	0.0174	−0.0742	0.0918	0.0139
LOD	9.7835	1.8555	99.1537	−0.5247	0.0402	0.2769	0.3330	2.2011	4.9459	−0.0051	0.0917	0.0084	−0.0301	0.0621	0.0048
LOD/2	−0.7361	0.1734	0.5718	0.1063	0.0690	0.0161	0.3302	2.2006	4.9420	−0.0049	0.0917	0.0084	−0.0235	0.0614	0.0043
Censoring 75%															
Our	−0.2572	0.9859	1.0362	−0.0094	0.1554	0.0242	0.3294	2.2007	4.9418	−0.0050	0.0917	0.0084	−0.0014	0.0783	0.0061
CC	20.5679	5.7034	455.5018	−0.5718	0.0549	0.3299	2.9139	4.8863	32.3192	−0.0347	0.1706	0.0302	−0.0842	0.1234	0.0223
LOD	33.4338	8.697	1193.3200	−0.6777	0.0271	0.4601	0.3410	2.2018	4.9546	−0.0055	0.0917	0.0084	−0.0543	0.0626	0.0069
LOD/2	0.7022	0.4009	0.6535	−0.1796	0.0709	0.0373	0.3315	2.2006	4.943	−0.0049	0.0917	0.0084	−0.0444	0.0618	0.0059
Method	$\alpha_1 = 4$			$\alpha_2 = 0.8$			$\beta_1 = 24$			$\beta_2 = 1$			$\rho = 0.8$		
Censoring 10%															
All data	0.0562	0.3797	0.1471	−0.0026	0.0792	0.0063	0.3360	2.2164	5.0155	−0.0050	0.0923	0.0085	−0.0009	0.0230	0.0005
Our	0.0478	0.4100	0.1701	6.00E-05	0.0851	0.0072	0.3360	2.2162	5.0144	−0.0050	0.0923	0.0085	−0.0009	0.0233	0.0005
CC	1.9418	0.5252	4.0458	−0.2150	0.0588	0.0497	6.4781	2.8121	49.8579	−0.1815	0.0778	0.0390	−0.0540	0.0292	0.0038

(Continues)

TABLE 1 (Continued)

Method	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE
	$\alpha_1 = 4$			$\alpha_2 = 0.8$			$\beta_1 = 24$			$\beta_2 = 1$			$\rho = 0.2$		
LOD	0.8503	0.3829	0.8693	−0.1270	0.0614	0.0199	0.3265	2.2171	5.0125	−0.0041	0.0925	0.0086	−0.0139	0.0230	0.0007
LOD/2	−0.5782	0.3337	0.4455	0.1341	0.0905	0.0262	0.3521	2.2174	5.031	−0.0063	0.0931	0.0085	−0.0126	0.0233	0.0007
Censoring 30%															
Our	0.0474	0.4531	0.2071	0.0016	0.0921	0.0085	0.3367	2.2163	5.0157	−0.0050	0.0923	0.0085	−0.0002	0.0242	0.0006
CC	4.3047	0.8202	19.2015	−0.3410	0.0521	0.1190	11.9894	3.6680	157.1728	−0.2822	0.0754	0.0853	−0.1007	0.0379	0.0116
LOD	2.4778	0.5403	6.4306	−0.2798	0.0521	0.0810	0.3282	2.2167	5.0118	−0.0041	0.0925	0.0085	−0.0409	0.0249	0.0023
LOD/2	−0.9183	0.2350	0.8983	0.2102	0.0794	0.0505	0.3411	2.2183	5.0272	−0.0055	0.0923	0.0085	−0.0301	0.0246	0.0015
Censoring 50%															
Our	0.0722	0.5775	0.3381	0.0019	0.1090	0.3381	0.3367	2.2161	5.0147	−0.0050	0.0923	0.0085	3.90E-05	0.0268	0.0007
CC	10.5578	2.1172	115.9395	−0.4848	0.0499	0.2375	21.9166	6.4716	522.1357	−0.3959	0.0823	0.1635	−0.1804	0.0613	0.0363
LOD	9.7744	1.8225	98.8538	−0.5247	0.0394	5.0156	0.3361	2.2164	5.0156	−0.0046	0.0923	0.0085	−0.1206	0.0319	0.0156
LOD/2	−0.7303	0.1691	0.5619	0.1046	0.0675	0.0155	0.3181	2.2162	5.0031	−0.0036	0.0925	0.0086	−0.0836	0.0294	0.0079
Censoring 75%															
Our	0.1567	0.8274	0.7078	−0.0013	0.1391	0.0192	0.3365	2.2166	5.0166	−0.0050	0.0923	0.0085	0.0015	0.0325	0.0011
CC	20.2281	5.4259	438.5579	−0.5689	0.0554	0.3267	31.5542	11.5165	1128.0320	−0.4582	0.1048	0.2209	−0.2454	0.1010	0.0704
LOD	33.2436	8.4924	1177.114	−0.6774	0.0267	0.4596	0.3438	2.2157	5.0177	−0.0051	0.0921	0.0085	−0.2278	0.0408	0.0536
LOD/2	0.7014	0.3942	0.6470	−0.1806	0.0697	0.0374	0.325	2.2155	5.0042	−0.0036	0.0925	0.0085	−0.1721	0.0372	0.0310

Note: Setup 1: Data simulated from a bivariate Gaussian copula with two gamma marginal distributions. Different scenarios correspond to different proportions of censoring (10%, 30%, 50%, or 75%). The table reports the bias, empirical standard error (SE), and mean squared error (MSE) for 500 simulated data using all data, our approach (correctly specified marginal distributions and copula function), CC analysis, LOD substitution, and LOD/2 substitution.

Abbreviations: CC, complete case; LOD, limit of detection.

TABLE 2 Simulation results for the linear regression (left censoring in the covariate only): Different scenarios corresponding to different proportions of censoring (10%, 30%, 50%, or 75%)

		β_0				σ			β_1				
Censoring	**Method**	**Bias**	**SE**	**MSE**	**CP (%)**	**Bias**	**SE**	**MSE**	**Bias**	**SE**	**MSE**	**CP**	**Type2E**
	All data	0.0123	0.4117	0.1693	93.8000	−0.0034	0.0563	0.0032	−0.0001	0.0086	0.0001	92.6000	0
10%	Our	0.0257	0.4764	0.2272	93.8000	−0.0079	0.0584	0.0035	0.0004	0.0004	0.0001	93.6000	0
	MID	0.0161	0.4818	0.2320	98.8000	0.4475	0.2002	0.2403	−0.0002	0.0098	0.0001	99.0000	0
	MI	0.1442	0.9044	0.8370	93.0000	0.0874	0.0804	0.0141	−0.0037	0.0175	0.0003	93.6000	0
	CC	0.0161	0.4818	0.2319	95.2000	−0.0024	0.0591	0.0035	−0.0002	0.0098	0.0001	93.6000	0
	LOD	−3.3152	1.0260	12.0411	1.4000	0.8603	0.2921	0.8254	0.0606	0.0194	0.0041	2.6000	0
	LOD/2	17.5565	2.4089	314.0203	0.0000	3.1411	0.2980	9.9553	−0.3386	0.0477	0.1169	0.0000	0
30%	Our	0.0216	0.6696	0.4480	94.2000	−0.0109	0.0655	0.0044	−0.0003	0.0128	0.0002	93.2000	0
	MID	0.0131	0.7011	0.4907	100.0000	2.4624	0.7011	6.1856	−0.0001	0.0132	0.0002	100.0000	0
	MI	0.1000	1.9290	3.7237	91.4000	0.0599	0.0893	0.0115	−0.0019	0.0353	0.0012	90.8000	0
	CC	0.0141	0.7000	0.4892	94.2000	−0.0045	0.0669	0.0045	−0.0002	0.0132	0.0002	94.2000	0
	LOD	−17.7191	2.2670	319.0940	0.0000	3.5875	0.4083	13.0365	0.3031	0.0416	0.0936	0.0000	0
	LOD/2	35.5965	1.4626	1269.2450	0.0000	4.4855	0.2465	20.1801	−0.6503	0.0319	0.4239	0.0000	0
50%	Our	0.0835	0.8705	0.7632	94.4000	−0.0151	0.0748	0.0058	−0.0013	0.0156	0.0003	93.6000	0
	MID	0.0576	0.9939	0.9891	100.0000	4.4816	0.4057	20.2486	−0.0009	0.4057	0.0003	100.0000	0
	MI	0.0633	2.2774	5.1801	93.2000	0.0042	0.0969	0.0094	−0.0012	0.0403	0.0016	93.2000	
	CC	0.0532	0.9475	0.8987	94.6000	−0.0066	0.0761	0.0058	−0.0008	0.0169	0.0003	94.8000	0
	LOD	−37.2270	3.8629	1400.7410	0.0000	5.8362	0.4277	34.2443	0.6055	0.0712	0.3717	0.0000	0
	LOD/2	41.2644	1.3800	1704.6490	0.0000	5.4491	0.2465	29.8042	−0.7167	0.0304	0.5145	0.0000	0
80%	Our	−0.0028	1.5759	2.5098	95.8000	−0.0379	0.1279	0.0177	−0.0028	0.0256	0.0007	95.0000	0
	MID	1.0981	1.3583	3.0471	100.0000	8.4208	0.4697	71.1307	−0.0173	0.0217	0.0008	100.0000	0
	MI	0.0772	2.5093	6.2901	93.4000	−0.0331	0.1278	0.0174	−0.0016	0.0425	0.0018	92.8000	0
	CC	0.1681	2.2350	5.0135	94.6000	−0.0179	0.1356	0.0187	−0.0026	0.0354	0.0013	94.2000	0
	LOD	−106.2183	14.9479	11 505.3100	0.0000	9.5980	0.4672	92.3397	1.5684	0.2575	2.5260	0.0000	0
	LOD/2	43.7251	1.5298	1914.2220	0.0000	8.5893	0.4614	73.9893	−0.6824	0.0310	0.4666	0.0000	0

Note: The table reports the empirical SE, bias, MSE (for all parameters), coverage probabilities (for regression coefficient), and type II errors (for β_1) for 500 simulated datasets using all data, the proposed approach, MID, MI, CC analysis, LOD substitution, and LOD/2 substitution.

Abbreviations: CC, complete case; LOD, limit of detection; MI, multiple imputation; MID, missing indicator.

in the CC analysis than those from MI, MID, and our approach due to loss of information. The difference increased when the percentage of censoring increased. Regarding the Type II error rates for β_1, all methods performed well. We observed no cases where the null hypothesis $H_0 : \beta_1 = 0$ was not rejected.

4.2 | Simulation: (left) censoring in both variables

4.2.1 | Estimating the association

In the second simulation setup, we generated bivariate censored (X, Y) values from a Gaussian copula ($\rho = 0.2$ or $\rho = 0.8$) with two marginal gamma distributions and four different censoring percentages as previously mentioned in Section 4.1 (see Table 3). The CC and substitution methods in all scenarios provided biased estimates for all parameters. Similar to what we observed in the case of censoring present in X, the LOD/2 substitution method gave less biased estimates as opposed to the LOD substitution. Again, in all scenarios, our proposed method produced unbiased estimates for all parameters.

Web Table 6 in the Supplementary Materials shows the mean estimate, empirical SE, and MSE for the expected values and variances of the two variables under the different simulation settings. Only our approach produced estimates close to the true values (with small biases) and small MSE values. Depending on the percentage of censoring, other methods yielded moderate to large biases. The results of the sensitivity analysis are shown in Web Figure 2 in the Supplementary Materials. Here, the proposed method was quite robust against the misspecification of the copula function when the association is small. Given a large association and the copula function's choice, there might be moderate bias in the parameter estimates. The higher the censoring percentage in the data, the more biased were the estimates. In the case where we misspecified one of the two marginal distributions or both marginal distributions, our proposed method still performed well in estimating the association parameter.

4.2.2 | Linear regression

A similar simulation procedure as specified in Section 4.1 was considered where we imposed censoring in both X and Y (see Table 4). One can observe the same behavior as in the case of censoring in only the covariate with respect to the level of bias. More specifically, the CC analysis, MI, MID, and our approach provided approximately unbiased estimates for all parameters, while the substitution method gave mildly to severely biased estimates when the percentage of censoring increased. The MSE values using our approaches were the smallest among all methods. The higher coverage probabilities of the CC, MI, and MID analyses were likely due to their larger uncertainties around parameter estimates as visualised by comparing the confidence widths (Web Figure 4 in the Supplementary Materials). While the CC analysis suffered from loss of efficiency due to loss in the number of observations, the MID approach needs to estimate one extra nuisance parameter.

Additional simulation results with a smaller (n=120) or larger sample size (n=1000) were reported in Web Tables 4 and 7 (measuring association scenarios) and Web Tables 8 and 9 (linear regression analysis) in the Web Appendix B (Supplementary Materials). In general, similar conclusions, as obtained from the main simulation study, were drawn.

5 | CASE STUDIES

5.1 | Pertussis data from Thailand

In this section, we reanalyzed the data from the pertussis clinical trial in Thailand using our proposed method and we compared its performance with others. The aim was to analyse antibody titer concentrations in the cord and at 2 months of age in infants. There were 158 infants randomized into the aP group and 157 infants in the wP group. We performed separate analyses for the aP infant group and the wP infant group, although this is theoretically not necessary since before month two, no vaccine was given to infants, and therefore we expected to see no difference between the two groups of infants.

TABLE 3 Simulation results

	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE
Method	$\alpha_1 = 4$			$\alpha_2 = 0.8$			$\beta_1 = 24$			$\beta_2 = 1$			$\rho = 0.2$		
Censoring 10%															
All data	0.0534	0.3844	0.1503	−0.0021	0.0794	0.0063	0.3294	2.2006	4.9415	−0.0050	0.0917	0.0084	−0.0013	0.0610	0.0037
Our	0.0388	0.4251	0.1819	0.0023	0.0879	0.0077	0.3604	2.4021	5.8886	−0.0048	0.0988	0.0098	−0.0003	0.0616	0.0038
CC	2.2763	0.5916	5.5307	−0.2312	0.0601	0.0572	11.0736	3.4468	134.4815	−0.2792	0.0739	0.0834	−0.0546	0.0727	0.0083
LOD	1.0368	0.3996	1.2343	−0.1497	0.0601	0.0260	5.2546	2.4049	33.3824	−0.1671	0.0723	0.0332	−0.0068	0.0615	0.0038
LOD/2	−0.6626	0.3196	0.5410	0.1550	0.0906	0.0322	−13.4426	1.3615	182.5529	1.2375	0.2562	1.5970	−0.0184	0.0604	0.0040
Censoring 30%															
Our	0.0458	0.4735	0.2259	0.0027	0.0945	0.0089	0.4224	2.6825	7.3599	−0.0051	0.1092	0.0119	−0.0007	0.065	0.0042
CC	4.2074	0.8986	18.5081	−0.3264	0.0599	0.1101	20.2223	5.1769	435.6882	−0.4055	0.0728	0.1697	−0.0799	0.0843	0.0135
LOD	2.4752	0.5481	6.4266	−0.2795	0.0528	0.0809	12.4604	3.2641	165.8946	−0.3223	0.065	0.1081	−0.0156	0.0635	0.0043
LOD/2	−0.9182	0.2405	0.9008	0.2111	0.0794	0.0508	−16.6497	0.6344	277.6134	2.0634	0.2235	4.3074	−0.0255	0.0619	0.0045
Censoring 50%															
Our	0.0624	0.6264	0.3954	0.0060	0.1179	0.0139	0.8408	3.9217	16.0563	−0.0098	0.1503	0.0226	−0.0013	0.0774	0.0060
CC	10.1934	2.8774	112.1685	−0.4605	0.0766	0.2179	53.1333	17.0179	3112.1780	−0.6148	0.0834	0.3849	−0.1073	0.1297	0.0283
LOD	9.7852	1.8551	99.1842	−0.5247	0.0402	0.2769	58.7968	12.1659	3604.7740	−0.6797	0.0481	0.4642	−0.0453	0.0713	0.4642
LOD/2	−0.7358	0.1733	0.5714	0.1604	0.0691	0.0161	−18.3799	0.1706	337.8480	2.5017	0.1445	6.2792	−0.0433	0.0658	0.0062
Censoring 75%															
Our	0.2566	0.9861	1.0364	−0.0093	0.1554	0.0242	1.3232	6.0536	38.3239	−0.0023	0.2185	0.0476	−0.0007	0.0934	0.0087
CC	20.9660	10.4493	548.5443	−0.5435	0.0162	0.3066	109.5150	63.1139	15 968.9300	−0.7258	0.115	0.5392	−0.1191	0.2276	0.0659
LOD	33.4893	8.6952	1196.9870	−0.6779	0.0270	0.4603	191.7881	52.6697	39 551.1900	−0.8635	0.0318	0.7467	−0.0775	0.0823	0.0128
LOD/2	0.7028	0.4010	0.6545	−0.1795	0.0710	0.0372	−16.7546	0.4848	280.9530	1.4764	0.2231	2.2295	−0.0702	0.0671	0.0094
Method	$\alpha_1 = 4$			$\alpha_2 = 0.8$			$\beta_1 = 24$			$\beta_2 = 1$			$\rho = 0.8$		
Censoring 10%															
All data	0.0562	0.3798	0.1472	−0.0026	0.0792	0.0063	0.3356	2.2164	5.0154	−0.0050	0.0923	0.0085	−0.0009	0.0230	0.0053
Our	0.0524	0.4179	0.1770	−0.0008	0.0861	0.0074	0.3872	2.3759	5.7835	−0.0061	0.0972	0.0095	−0.0013	0.0235	0.0006
CC	2.5147	0.6027	6.6865	−0.2486	5.79E-02	0.0651	12.9824	3.5790	181.3272	−0.3117	0.0687	0.1018	−0.0730	0.0336	0.0064

(Continues)

TABLE 3 (Continued)

Method	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE
	$\alpha_1 = 4$			$\alpha_2 = 0.8$			$\beta_1 = 24$			$\beta_2 = 1$			$\rho = 0.2$		
LOD	1.0412	0.3963	1.2407	−0.1503	0.0597	0.0261	5.2094	2.3826	32.8037	−0.1656	0.0716	0.0325	−0.0099	0.0243	0.0007
LOD/2	−0.6564	0.3155	0.5302	0.1522	0.0900	0.0312	−13.4503	1.3451	182.7160	1.2382	0.2593	1.6003	−0.0683	0.0307	0.0056
Censoring 30%															
Our	0.0480	0.4596	0.2131	0.0018	0.0930	0.0086	0.4272	2.6683	7.2877	−0.0054	0.1072	0.0115	−0.0009	0.0249	0.0006
CC	4.6085	0.8807	22.0120	−0.3442	0.0532	0.1213	22.8057	5.4226	547.5310	−0.4333	0.0655	0.1921	−0.1170	0.0436	0.0156
LOD	2.4785	0.5402	6.4343	−0.2799	0.0521	0.0810	12.3811	3.2108	163.5796	−0.3207	0.0641	0.1069	−0.0205	0.0272	0.0012
LOD/2	−0.9204	0.2350	0.9022	0.2105	0.0795	0.0506	−16.6432	0.6668	277.4390	2.0618	0.2300	4.3038	−0.0873	0.0334	0.0087
Censoring 50%															
Our	0.0838	0.5959	0.3612	0.0008	0.1111	0.0123	0.6720	3.5890	13.3070	−0.0069	0.1375	0.0189	−0.0010	0.0315	0.0010
CC	10.4896	2.2408	115.0426	−0.4659	0.0569	0.2203	54.2539	13.6314	3128.9280	−0.6225	0.0647	0.3917	−0.2132	0.0813	0.0521
LOD	9.7746	1.8221	98.8554	−0.5247	0.0394	0.2769	58.4774	11.7869	3558.2580	−0.6787	0.0467	0.4628	−0.0578	0.0410	0.0050
LOD/2	−0.7299	0.1692	0.5613	0.1052	0.0676	0.0156	−18.3886	0.1621	338.1680	2.5085	0.1436	6.3133	−0.1153	0.0399	0.0149
Censoring 75%															
Our	0.1825	0.8934	0.8298	−0.0025	0.1460	0.0213	1.1848	5.3518	31.0663	−0.0057	0.1940	0.0376	−0.0004	0.0426	0.0018
CC	18.9094	5.8449	391.6578	−0.5384	0.0706	0.9248	92.6878	33.4874	9710.2030	−0.7135	0.0773	0.5150	−0.2891	0.1388	0.1028
LOD	33.1758	8.4807	1172.4090	−0.6772	0.0267	0.4593	189.8550	50.9150	38 632.0400	−0.8628	0.0306	0.7453	−0.1060	0.0688	0.0157
LOD/2	0.7032	0.3943	0.6497	−0.1799	0.0698	0.0372	−16.7597	0.4902	281.1261	1.4835	0.2293	2.2533	−0.1586	0.0555	0.0282

Note: Setup 1: Data simulated from a bivariate Gaussian copula with two gamma marginal distributions and censoring in both X and Y. Different scenarios correspond to different proportions of censoring (10%, 30%, 50%, or 75%). The table reports bias, empirical standard errors (SE), and MSE values for 500 simulated datasets using all data, our approach (correctly specified marginal distributions and copula function), CC analysis, LOD and LOD/2 substitution.

Abbreviations: CC, complete case; LOD, limit of detection.

TABLE 4 Simulation results of the (simple) linear regression (left censoring in the covariate and the response): Different scenarios corresponding to different proportions of censoring (10%, 30%, 50%, or 75%)

		β_0				σ			β_1				
Censoring	**Method**	**Bias**	**SE**	**MSE**	**CP (%)**	**Bias**	**SE**	**MSE**	**Bias**	**SE**	**MSE**	**CP (%)**	**Type2E**
	All data	−0.1023	0.4117	0.1693	93.8000	0.0034	0.0563	0.0032	0.0001	0.0086	0.0001	92.6000	0
10%	Our	−0.0501	0.4950	0.2470	93.9000	0.0058	0.0599	0.0036	0.0008	0.0100	0.0001	93.2000	0
	MI	0.0059	0.4946	0.2442	94.6000	−0.0098	0.0599	0.0037	−0.0003	0.0100	0.0001	94.6000	0
	MID	0.0075	0.4959	0.2455	93.8000	−0.0056	0.0603	0.0037	−6.0E-5	0.0100	0.0001	93.0000	0
	CC	−0.1305	0.4895	0.2561	94.6000	0.0078	0.0591	0.0036	0.0023	0.0099	0.0001	93.4000	0
	LOD	−0.0879	0.3803	0.1520	95.2000	0.0673	0.0587	0.0080	0.0015	0.0081	0.0001	94.2000	0
	LOD/2	9.8161	2.2113	101.2351	0.6000	−3.6470	0.9073	14.1225	−0.1851	0.0417	0.0360	0.6000	0
30%	Our	−0.0138	0.6808	0.4628	92.7000	0.0085	0.0731	0.0054	0.0002	0.0129	0.0002	92.7000	0
	MI	−0.0409	0.6799	0.4630	94.2000	−0.0240	0.0643	0.0047	0.0008	0.0129	0.0002	93.8000	3
	MID	0.0109	0.7023	0.4923	94.8000	0.0051	0.0669	0.0045	−0.0001	0.0132	0.0002	94.6000	0
	CC	−0.2091	0.6970	0.5286	93.8000	0.0154	0.0669	0.0047	0.0035	0.0131	0.0002	93.8000	0
	LOD	0.9368	0.3323	1.0738	42.4000	0.1957	0.0573	0.0416	−0.0161	0.0094	0.0003	53.0000	0
	LOD/2	14.3759	1.0938	210.1312	0.0000	−5.9934	0.3095	37.6903	−0.2570	0.0298	0.0672	0.0000	0
50%	Our	0.0707	0.9017	0.8169	93.4000	0.0141	0.0758	0.0059	0.0011	0.0162	0.0003	93.4000	0
	MI	−0.0086	0.9292	0.8617	94.8000	−0.0436	0.0742	0.0074	0.0002	0.0166	0.0003	94.4000	0
	MID	0.0445	0.9499	0.9026	97.2000	−0.2610	0.0808	0.0080	−0.2610	0.0808	0.0003	97.6000	0
	CC	−0.2365	0.9414	0.9403	93.0000	0.0177	0.0744	0.0058	0.0038	0.0168	0.0003	93.6000	0
	LOD	3.1422	0.6382	10.2796	0.0000	0.2820	0.0615	0.0833	−0.0512	0.0129	0.0028	0.6000	0
	LOD/2	15.7716	1.6542	251.4716	0.0000	−7.6902	0.3095	61.0762	−0.2692	0.0297	0.0733	0.0000	0
75%	Our	−0.1277	2.0310	4.1332	96.6000	0.0383	0.1498	0.0239	0.0020	0.0326	0.0011	95.8000	0
	MI	0.0435	2.0991	4.3993	95.8000	−0.0869	0.1207	0.0221	−0.0007	0.0334	0.0011	95.0000	0
	MID*	0.1545	2.2433	5.0461	91.3000	−0.1675	0.1643	0.0445	−0.0024	0.0356	0.0013	91.3000	0
	CC	0.3543	2.2173	5.0322	95.2000	0.0296	0.1373	0.0197	0.0054	0.0352	0.0013	94.4000	0
	LOD	8.4155	1.7780	73.9751	0.0000	0.5352	0.0787	0.2926	−0.1241	0.0310	0.0164	0.0000	0
	LOD/2	19.4869	1.3711	381.6160	0.0000	−7.6695	1.7962	62.0442	−0.3004	0.0270	0.0910	29.5295	0

MID*: Convergence achieved only for 23 (out of 500) datasets. Results presented are based on these 23 converged outputs.

Note: The table shows the empirical SE, bias, MSE (for all parameters), coverage probabilities (for regression coefficient parameters), and Type II error (for β_1) for 500 simulated datasets using all data, our approach, MI, MID, CC analysis, LOD substitution, and LOD/2 substitution.

Abbreviations: CC, complete case; LOD, limit of detection; MI, multiple imputation; MID, missing indicator.

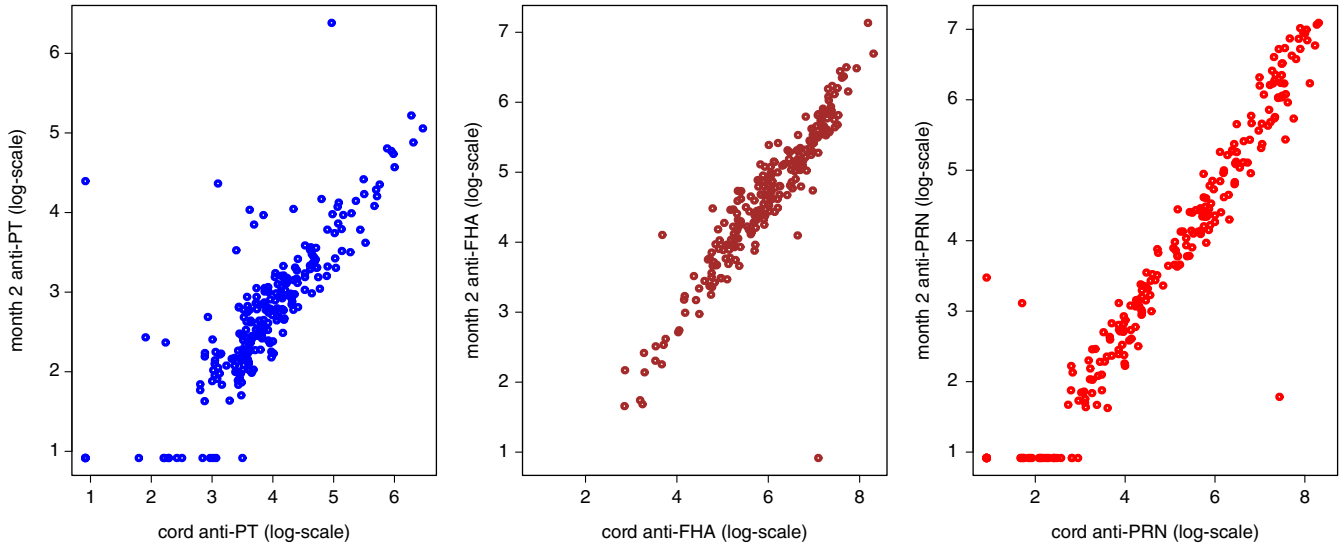

FIGURE 1 Scatter plots between cord antibody titers (*x*-axis) and antibody titers at month two (*y*-axis) on a log-scale: anti-PT (left), anti-FHA (middle), and anti-PRN (right) [Colour figure can be viewed at wileyonlinelibrary.com]

5.1.1 | Association of antibody titers in the cord and at month two in infants

Scatter plots between antibody titers in the cord and at month two (on the natural log-scale) are given in Figure 1. There is a clear linear association between cord antibody titers and antibody titers at month two, especially for anti-FHA and anti-PRN. Generally, if cord antibody titers are high, one expects to see high antibody titers at month two. We presented in this article the results of analyzing anti-PT data. The results of analyzing anti-PRN data are shown in Web Appendix D (Supplementary Materials). Anti-FHA data contain only one censored observation in antibody titers at month two and hence were not considered for analysis. There are 0.4%, 5.1%, and 2.1% of censored observations in cord anti-PT antibodies, antibodies at month two, and both, respectively.

Since antibody titer data are of interest, log-normal and gamma distributions are two sensible starting points for the marginal distributions of the two measurements. All five copula functions mentioned in Section 3 were used. The AIC was utilized to perform model selection (see Table 5). In the aP group, the model with a Frank copula function and two log-normal marginal distributions had the lowest AIC, while in the wP group, the model with a Gaussian copula function and two log-normal marginal distributions gave the smallest AIC. The correlation (expressed by Kendall's tau) between anti-PT IgG titer in the cord and at month two (aP group) using the best model was 0.7002. The Kendall's tau using the substitution method (both LOD and LOD/2) was 0.6736, while the CC analysis gave a Kendall's tau value of 0.6639. In the wP group, Kendall's tau was 0.7364 (for both substitution methods), 0.7032 (CC analysis), and 0.7198 (our proposed method). There was a small difference between various methods due to a limited number of observations below the LOD. In real applications with larger percentages of censoring, it is important to use an appropriate method to obtain a good estimate for the association parameter.

TABLE 5 Akaike information criteria of different fitted models for various choices of copula functions to measure the association of anti-PT IgG antibody titers (data in Thailand) at the cord and month two (using the method for censoring in both *X* and *Y*)

	aP infant group					wP infant group				
	Gaussian copula	Frank copula	Clayton copula	Gumbel copula	Joe copula	Gaussian copula	Frank copula	Clayton copula	Gumbel copula	Joe copula
Gam-Gam	2191.009	2103.344	2116.287	2202.699	2233.166	1718.619	1728.062	1914.64	1723.269	1762.811
LN-LN	2078.675	**2034.153**	2086.067	2065.195	2091.499	**1672.701**	1683.837	1699.557	1845.015	1845.015
LN-Gam	2173.159	2094.762	2132.758	2175.758	2201.149	1703.376	1706.337	1714.848	1716.937	1748.023
Gam-LN	2099.275	2053.909	2083.434	2108.912	2138.833	1715.209	1683.837	1699.557	1845.015	1845.015

The best models indicated that the log-normal assumptions for cord and month two antibody titers were reasonable. More specifically, $X_{aP} \sim LN(4.1918, 0.8268)$, $Y_{aP} \sim LN(2.9810, 0.7726)$ and $X_{wP} \sim LN(4.0151, 0.7233)$, $Y_{wP} \sim LN(2.8435, 0.7495)$ where X and Y denote the random variables representing cord antibody titers and antibodies at month two. There was no difference between the groups, despite the fact that a Gaussian copula was chosen for the wP group, and a Frank copula performed best for the aP group. Therefore, all data were pooled together and reanalyzed with two log-normal marginal distributions. Based on the AIC value, the bivariate Frank copula function performed best. The two marginal distributions were estimated to be $X \sim LN(4.1271, 0.8073)$ and $Y \sim LN(2.9347, 0.7948)$, and this marginal distribution of X was used for the linear regression model in the following section.

5.1.2 | Linear regression model

One of the main objectives of this study was to understand which effects might be important to explain the observed antibody titers in infants at month two (after being born and right before the first dose of the vaccination against pertussis). The primary interest was to see how antibody titers at cord ($antiPT_{Cord}$) might help to predict antibody titers at month two ($antiPT_{M2}$). Other important factors, as specified in the protocol, are the group of infants (group A includes aP infants and group B consists of wP infants), the gestational age at vaccination (*ga_vacc*), the duration between vaccination and delivery (*dur_vd*), infant birth weight (*bw*, recorded in kilograms) and birth length (*bl*, measured in centimeters), age of the mother when giving birth, and feeding manner. Unfortunately, the information regarding feeding (bottle versus breastmilk) was not available. Moreover, since *ga_vacc* and *dur_vd* might give the same information, we decided to include only the *ga_vacc* variables in the analysis. Antibody titer data on a log-scale were used, that is, we define $Y = log(antiPT_{M2})$ and $X = log(antiPT_{Cord})$. First, the random forest method[42,43] was used to investigate which factors were important to predict the level of antibody titers at month two. Web Figure 5 (Supplementary Materials) shows the dotchart of variable importance measured by a random forest analysis. Generally, besides the cord antibody titers, the infant birth weight and birth length were important factors. The results of the random forest approach using data with LOD/2 substitution were reported. However, the output using LOD substitution and CC data pointed in the same direction.

Consequently, in the next step, a linear regression model with three covariates, namely, cord antibody titers (on a log-scale), infant birth weight (*bw*), and birth length (*bl*) standardized around their means was considered. The statistical model is specified as $y = \beta_0 + \beta_1 x + \beta_2 bw + \beta_3 bl + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$. Since there are small percentages of censoring in our dataset, all methods gave quite comparable results (see Table 6). In details, the CC analysis, our approach, and MI analysis provided similar outputs. In conclusion, only cord antibody titers affected antibody levels at month two in infants. The weight and length of infants at birth were not statistically significant.

Sensitivity analysis:

This analysis aims to investigate how the results varied when the percentage of censoring increased. More specifically, a larger LOD value ("new LOD") was imposed to enlarge the number of left-censored observations in the dataset. Three "new LODs" values of 10, 20, and 50 (IU/mL) were chosen leading to higher percentages of censoring in both cord antibody titers and antibody titers at month two (more details in Web Table 10). When the percentage of censoring increased, our proposed method and the MI analysis gave quite consistent results, while the substitution method and the CC analysis produced varying estimates and SE values depending on the scenarios. More specifically, the estimates of the intercept under LOD and LOD/2 substitution switched sign and significance when the new "imputed" LOD was 20

TABLE 6 Parameter estimates with SEs (in brackets) using different approaches

Parms	β_0	β_1	β_2	β_3
CC	−0.465 (0.156)*	0.831 (0.038)*	0.017 (0.036)	0.025 (0.036)
LOD	−0.080 (0.141)	0.739 (0.035)*	0.037 (0.039)	0.033 (0.039)
LOD/2	−0.276 (0.147)	0.780 (0.037)*	0.037 (0.043)	0.048 (0.043)
Our approach	−0.425 (0.162)*	0.818 (0.040)*	0.031 (0.040)	0.035 (0.040)
MI	−0.932 (0.175)*	0.934 (0.043)*	0.024 (0.043)	0.042 (0.043)
MID	−0.446 (0.160)	0.823 (0.039)*	0.030 (0.040)	0.035 (0.040)

Note: *indicates significant results with a *P*-value below .05.

Abbreviations: CC, complete case; LOD, limit of detection; MI, multiple imputation; MID, missing indicator.

TABLE 7 Sensitivity analysis: Parameter estimates with SE(s) (in brackets) using different approaches

Parms	Scenario	β_0	β_1	β_2	β_3
CC	Original data	−0.465 (0.156)*	0.831 (0.038)*	0.017 (0.036)	0.025 (0.036)
	LOD = 10	−0.110 (0.206)	0.765 (0.047)*	−0.007 (0.039)	0.031 (0.038)
	LOD = 20	0.901 (0.360)*	0.583 (0.075)*	−0.073 (0.059)	0.043 (0.052)
	LOD = 50	2.323 (1.350)	0.380 (0.244)	−0.028 (0.224)	0.170 (0.323)
LOD	Original data	−0.080 (0.141)	0.739 (0.035)*	0.037 (0.039)	0.033 (0.039)
	LOD = 10	0.337 (0.136)*	0.655 (0.034)*	0.030 (0.035)	0.024 (0.035)
	LOD = 20	1.408 (0.122)*	0.458 (0.030)*	0.005 (0.029)	0.019 (0.029)
	LOD = 50	2.888 (0.099)*	0.255 (0.023)*	−0.122 (0.016)	0.026 (0.016)
LOD/2	Original data	−0.276 (0.147)	0.780 (0.037)*	0.037 (0.043)	0.048 (0.043)
	LOD = 10	−0.262 (0.167)	0.763 (0.042)*	0.047 (0.048)	0.043 (0.048)
	LOD = 20	0.349 (0.158)*	0.623 (0.039)*	0.061 (0.043)	−0.008 (0.043)
	LOD = 50	2.208 (0.110)*	0.295 (0.028)*	−0.025 (0.028)	0.055 (0.028)
Our approach	Original data	−0.425 (0.162)*	0.818 (0.040)*	0.031 (0.040)	0.035 (0.040)
	LOD = 10	−0.636 (0.201)*	0.863 (0.048)*	0.027 (0.044)	0.038 (0.044)
	LOD = 20	−1.075 (0.343)*	0.945 (0.075)*	0.035 (0.059)	0.004 (0.056)
	LOD = 50	−0.707 (0.748)	0.860 (0.140)*	−0.050 (0.112)	0.150 (0.117)
MI	Original data	−0.446 (0.160)*	0.823 (0.039)*	0.030 (0.040)	0.350 (0.040)
	LOD = 10	−0.680 (0.198)*	0.873 (0.047)*	0.026 (0.044)	0.037 (0.043)
	LOD = 20	−1.111 (0.341)*	0.953 (0.074)*	0.034 (0.058)	0.004 (0.055)
	LOD = 50	−1.240 (0.785)	0.965 (0.146)*	−0.042 (0.106)	0.138 (0.109)
MID	Original data	−0.932 (0.175)*	0.934 (0.043)*	0.024 (0.043)	0.042 (0.043)
	LOD = 10	−2.157 (0.271)*	1.186 (0.064)*	0.022 (0.056)	0.042 (0.056)
	LOD = 20	−3.886 (0.534)*	1.481 (0.116)*	0.097 (0.087)	−0.037 (0.083)
	LOD = 50	−8.007 (1.920)*	2.123 (0.351)*	−0.034 (0.179)	0.226 (0.179)

Note: *indicates significant results with a *P*-value below .05.
Abbreviations: CC, complete case; LOD, limit of detection; MI, multiple imputation; MID, missing indicator.

and 50, respectively. The MID approach produced reasonably consistent results for β_2, β_3 but not for β_1 and the intercept. The antibody in the cord's effect was statistically significant in all scenarios except for the CC analysis when LOD = 50. While the estimates of β_1 in the CC, substitution, and MID analyses varied considerably between different scenarios, our proposed method and MI gave close estimates across all scenarios (Table 7).

5.2 | VZV data from Belgium

Here, we report the correlation estimate between antibody titers against VZV at month 9 and month 12 in infants. There was 88.57%, 88.57%, and 84.29% of censoring in antibody titers at month 9 (marginally), at month 12 (marginally), and both variables, respectively. We assumed that antibody titers followed either a log-normal distribution or a gamma distribution. The AIC values for different model fits with various choices of copula functions are shown in Table 8. All models performed comparably. The model with a log-normal marginal distribution for antibody titers at month 9 and gamma marginal distribution for antibody titers at month 12 joining by a Joe copula function showed the smallest AIC value. The Kendall's tau given by this model was 0.382, while Kendall's tau value using the two substitution methods was 0.553. This result is not unanticipated since the substitution method implies that the correlation given by 84.29% of the observations is one. Hence, the substitution method, in this case, clearly overestimates the association.

TABLE 8 Akaike information criteria values of different fitted models with regard to various choices of copula functions to analyse the association of antibody titers against VZV (data in Belgium) at month 9 and month 12 (using the method for censoring in both X and Y)

	Gaussian	Frank	Clayton	Gumbel	Joe
Gam-Gam	165.257	165.381	165.725	164.553	164.226
LN-LN	166.283	166.205	166.595	165.520	167.348
LN-Gam	165.254	165.253	165.553	164.437	**163.966**
Gam-LN	166.495	166.446	166.89	165.926	167.912

6 | DISCUSSION AND CONCLUSION

In this article, we proposed methods to analyse data for which left censoring is present. The first objective is to account for (left) censoring in one or two variables, while the main interest lies in estimating their marginal distributions and association. To do this, we introduced a copula model approach in which we used a copula function to join the two marginal distributions of the two variables. Within this article, the focus is on the MLE framework, although a Bayesian approach could be employed. Our work has derived the likelihood function in the cases of left censoring for either one measurement or both. However, one might modify the proposed approach to deal with right-censored or interval-censored data. Moreover, the proposed method can be extended to cope with higher-dimensional problems, making use of a copula function to join three or more marginal distributions. Note, however, that when moving away from a two-dimensional problem, the computational burden potentially increases.

The simulation study showed that our approach outperformed the CC analysis, as well as the substitution method for all scenarios with percentages of censoring varying from low (10%) to high (80%) given that the two marginal distributions and the copula function were correctly specified. For a misspecified copula function, our approach still produced reasonable estimates for all parameters of two marginal distributions given a small association. The bias appeared negligible, except for the association parameter. On the other hand, a low level of bias was not guaranteed if the copula function was not correctly specified under a large association scenario. The substitution method should only be used when a small percentage of censoring is present in the data (as in our data application). Otherwise, this method will produce biased estimates. Nevertheless, in our simulation, the substitution method with LOD/2 performed relatively well even when there was more than 50% of censored observations. That result can be partly explained by the fact that, in these simulations, the LOD/2 values were quite close to the values of $\mathbf{E}[X|X<LOD]$ and $\mathbf{E}[Y|Y<LOD]$. It has been shown before by Lynn[12] that the substitution method by $\mathbf{E}[X|X<LOD]$ ranked among the best-performing methods.

Throughout this article, we focused on gamma and log-normal marginal distributions, although other distributions can also be considered. Note that these two distributions are appropriate for describing antibody titers in humans, which is inspired by our motivating examples. As mentioned previously, our method is quite attractive in the sense that it can capture any association between two measurements in terms of Kendall's tau. The method does not limit itself to Pearson's correlation coefficient, which only captures linear associations (see, eg, the work of References 12,14). Moreover, using our method, one does not need to assume normality of the data (on a log scale) or bivariate normality as in Lynn,[12] Lyles et al.[13] We nonetheless assume some parametric forms for the two marginal distributions and a copula function to join them. However, as pointed out in the sensitivity analysis, our method was quite robust against the misspecification of marginal distributions and the choice of copula function.

The second aim of this article is the MLE method to account for censoring within a linear regression context. When the censoring is uninformative, the CC analysis gave valid inference for the linear regression model. By contrast, the substitution method leads to biased estimates (for a percentage of censoring larger than 10%) and hence, should be used only in a limited amount of censoring. Our proposed approach produced unbiased estimates, and the coverage probabilities were close to nominal coverage levels. The method can be adapted to account for right censoring, or interval censoring by changing the lower and upper bounds for the integrals in the likelihood functions. The extension towards the inclusion of interaction effects is straightforward. This extension only implies that the mean structure (linear predictor) is slightly more complicated. Interested readers are invited to the Web Appendix A.2.2 to have a thorough look at the log-likelihood function if we take interaction into account.

Next to our proposed method, the MI and MID approaches also attained good performance. Diverse imputation techniques have been proposed in the literature (see, eg, Atem et al,[44] Wei et al,[20,21] among others). A recent article reviewed comprehensively many imputation frameworks and concluded that the MI method using predictive mean matching

(PMM-MI) ranged in the most optimal approaches.[45] Due to the problem with perfect correlation, we could not perform PMM-IM in our simulation setting. Nevertheless, the suggested MI approach performed reasonably well compared with our proposed method. This result is not unexpected since our proposed imputation scheme is based on the same underlying assumption on the knowledge of the marginal distribution of the censored covariate. Similarly, the MID analysis has demonstrated good performance. This result has been shown before in the logistic regression setting with two censored covariates.[24] In terms of biasedness, the MID and our approaches were comparable, but ours provided a slightly smaller bias. On the other hand, the MID approach exhibited consistently higher coverage probabilities with the cost of wider confidence widths. To the best of our knowledge, only theoretical asymptotic properties for linear regression have been derived under MID.[23] In the future, it is interesting to compare the performance of our proposed method and MID for general link functions, as well as in the case of informative censoring.

A common feature of interest in the MLE framework is model selection. In our data applications, the selection of marginal distributions and the copula function as well as model selection, in general, was performed using AIC. In contrast to standard linear regression in which no distributional assumption with regard to the distribution of covariates is required, such an assumption is required in the proposed methodology in case of censoring present in that covariate. Therefore, we assume the marginal distribution of the censored covariate to be known. Having said that, we propose first to perform model selection to pick up the most "reasonable" distribution for that censored covariate (by fitting either a gamma or a log-normal distribution to the antibody titers at the cord, as in our data application), and then using that information to continue with (multiple) linear regression analysis. In a general setting, one can fit univariate distributions to continuous, censored data using available built-in packages and perform goodness-of-fit check by means of histograms and theoretical densities comparison, empirical and theoretical cumulative distribution functions comparison, Q-Q and P-P plot (see, eg., Delignette-Muller et al[46]). Variable selection is an important feature of any multiple regression analysis. Since our proposed method is parametric under the MLE framework, one can also apply the idea of penalized regression, such as regression shrinkage via the Lasso.[47,48] Following this idea, researchers can extend the specified log-likelihood functions to incorporate a penalty term for nonzero coefficients. This can be done by adding a penalization term consisting of the sum of the absolute values of the coefficients. However, model selection when censoring is present in the data is currently beyond the scope of this article and will be considered in further research.

In our first data application, investigators collected antibody titers in infants from birth, before, and after the primary vaccination series, and before and 1 month after the booster dose at month 18. In the current analysis, we concentrated only on a pairwise analysis. When the interest lies in the kinetics of antibody titers in infants over time, one might consider fitting a nonlinear mixed effect model to pool all information together; see, for example, Maertens et al,[49] Tran et al.[50] Within this framework, the censored observations could be dealt with by applying the idea of Tobit regression given that there is censoring present only in the response. In other data applications, when there is censoring present in both covariate and response in a linear mixed model context, one can extend our proposed method to account for this given that the marginal distribution of the censored covariate is known or can be estimated. However, the computational burden in such analyses might become challenging since we need to integrate over the random effects as well. This research area could be a promising domain for further examination.

The article proposes methods to deal with censoring while estimating the association parameter in the copula model and the regression approach. Indeed, there is an explicit link between the Pearson product-moment correlation coefficients, as a measure of linear association, between covariate X and dependent variable Y, and the ordinary least square estimators for the regression coefficients in a linear regression model. For different copula functions, depending on the association parameter, this link is less straightforward. For instance, in the case of Archimedean copulas such as the Clayton, Frank, or Gumbel copula, the association between the two random variables is more naturally expressed in terms of Kendall's tau, a rank-based association measure. Often, a closed-form expression for the linear Pearson product-moment correlation coefficient in terms of the association parameter is absent. Some work has been done with regard to NP robust inference in a linear regression setting with estimators for the regression parameters based on Kendall's rank correlation tau (see, eg, the Theil-Sen or Kendall-Theil estimator for the regression line in robust regression based on the seminal work by Sen,[51] Theil[52]). However, an extension towards the application of linear regression with censored covariate and response data is considered beyond this article's scope and is an exciting research feature in future work.

Last but not least, our proposed methods are applicable when the limits of detection are known, that is, x_{cens} and y_{cens} have been defined prior to the data analysis since it involves integral calculations with the limits equal to the LOD values (for left censoring). In the two data applications, these limits were known upfront and provided by the

antibody tests' manufacturer. Although the LOD values were equal for all study subjects, the method can be readily extended to accommodate individual-specific limits (or censoring times, in other applications). In case the LOD is unknown, the censoring distribution needs to be modeled as well, and the censoring time (in our current application, the LOD) becomes a random variable. This random variable could follow a degenerate distribution at an unknown value. However, in general, a marginalization over the censoring distribution is necessary. In contrast to the survival context where independent censoring assumption (no relation between the parameters in the censoring and event time distributions) can be relied upon, in our settings, the censoring distribution cannot be ignored. This is because the distribution of the censored covariate information depends clearly on the censoring distribution. Consequently, the extension of the proposed approaches to the situation with unknown limits is not straightforward. Hence, we consider developing methods to handle censored data in the event of unknown detection limits, a promising avenue to explore further in future work.

ACKNOWLEDGEMENTS

We thank Elke Leuridan and Kirsten Maertens for providing us with the two datasets used in our application. This work was supported by the Research Fund of Hasselt University (grant BOF11NI31 to S.A.). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 682540 – TransMID). The authors also gratefully acknowledge financial support from the Methusalem research grant from the Flemish Government awarded to Prof. Dr. Herman Goossens and to Prof. Dr. Geert Molenberghs. K.M is the beneficiary holder of a FWO postdoctoral mandate fellowship from the FWO (Fund for Scientific Research-Flanders; FWO 12R5719N).

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in (repository name, eg, "figshare") at http://doi.org/[doi], reference number (reference number).

ORCID

Thao M. P. Tran https://orcid.org/0000-0003-2336-8302
Steven Abrams https://orcid.org/0000-0001-7353-9304
Marc Aerts https://orcid.org/0000-0002-1803-9072
Kirsten Maertens https://orcid.org/0000-0002-2880-441X
Niel Hens https://orcid.org/0000-0003-1881-0637

REFERENCES

1. EMA. *Guidelines for the Validation of Analytical Methods Used in Residue Depletion Studies*. Brussels, Belgium: VICH; 2009.
2. Armbruster DA, Pry T. Limit of blank, limit of detection and limit of quantitation. *Clin Biochemist Rev*. 2008;29(Suppl 1):S49.
3. Alf EF, Grossberg JM. The geometric mean: confidence limits and significance tests. *Perception Psychophys*. 1979;26(5):419–421.
4. Gad SC. *Clinical Trials Handbook*. Vol 8. Hoboken, NJ: John Wiley & Sons; 2009.
5. Spizman L, Weinstein MA. A note on utilizing the geometric mean: when, why and how the forensic economist should employ the geometric mean. *J Leg Econ*. 2008;15:43.
6. Zeghnoun A, Pascal M, Fréry N, et al. Dealing with the non-detected and non-quantified data. the example of the serum dioxin data in the French dioxin and incinerators study. *Organohalogen Compd*. 2007;69:2288-2291.
7. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. New York, NY: John Wiley & Sons; 2002.
8. El-Shaarawi A, Esterby S. Replacement of censored observations by a constant: an evaluation. *Water Res*. 1992;26(6):835-844.
9. Hewett P, Ganser GH. A comparison of several methods for analyzing censored data. *Ann Occup Hyg*. 2007;51(7):611-632.
10. EFSA. Management of left-censored data in dietary exposure assessment of chemical substances. *EFSA J*. 2010;8(3):1557.
11. Schmoyeri R, Beauchamp J, Brandt C, Hoffman F. Difficulties with the lognormal model in mean estimation and testing. *Environ Ecol Stat*. 1996;3(1):81-97.
12. Lynn HS. Maximum likelihood inference for left-censored HIV RNA data. *Stat Med*. 2001;20(1):33-45.
13. Lyles RH, Fan D, Chuachoowong R. Correlation coefficient estimation involving a left censored laboratory assay variable. *Stat Med*. 2001;20(19):2921-2933.
14. Lyles RH, Williams JK, Chuachoowong R. Correlating two viral load assays with known detection limits. *Biometrics*. 2001;57(4):1238-1244.

15. Benning L, Lyles RH, Gange SJ. Methods for comparing correlations involving left-censored laboratory data. In *ASA Proceedings of Joint Statistical Meetings, Section on Statistics in Epidemiology* (pp. 212-216).
16. Song J, Barnhart HX, Lyles RH. A GEE approach for estimating correlation coefficients involving left-censored variables. *J Data Sci*. 2004;2(3):245-257.
17. Tobin J. Estimation of relationships for limited dependent variables. *Econometrica*. 1958;26(1):24-36.
18. Tsimikas JV, Bantis LE, Georgiou SD. Inference in generalized linear regression models with a censored covariate. *Comput Stat Data Anal*. 2012;56(6):1854-1868.
19. Atem FD, Qian J, Maye JE, Johnson KA, Betensky RA. Linear regression with a randomly censored covariate: application to an Alzheimer's study. *J Royal Stat Soc Ser C (Appl Stat)*. 2017;66(2):313-328.
20. Wei R, Wang J, Jia E, Chen T, Ni Y, Jia W. GSimp: a Gibbs sampler based left-censored missing value imputation approach for metabolomics studies. *PLoS Comput Biol*. 2018;14(1):e1005973.
21. Wei R, Wang J, Su M, et al. Missing value imputation approach for mass spectrometry-based metabolomics data. *Sci Rep*. 2018;8(1):1-10.
22. Stekhoven DJ, Bühlmann P. MissForest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112-118.
23. Jones MP. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *J Am Stat Assoc*. 1996;91(433):222-230.
24. Chiou SH, Betensky RA, Balasubramanian R. The missing indicator approach for censored covariates subject to limit of detection in logistic regression models. *Ann Epidemiol*. 2019;38:57-64.
25. Jones MP. Linear regression with left-censored covariates and outcome using a pseudolikelihood approach. *Environmetrics*. 2018;29(8):e2536.
26. Wanlapakorn N, Maertens K, Chaithongwongwatthana S, et al. Assessing the reactogenicity of Tdap vaccine administered during pregnancy and antibodies to Bordetella pertussis antigens in maternal and cord sera of Thai women. *Vaccine*. 2018;36(11):1453-1459.
27. Wanlapakorn N, Maertens K, Vongpunsawad S, et al. Quantity and quality of antibodies after acellular versus whole-cell pertussis vaccines in infants born to mothers who received tetanus, diphtheria, and acellular pertussis vaccine during pregnancy: a randomized trial. *Clin Infect Dis*. 2019;71(1):72–80.
28. Mueller NH, Gilden DH, Cohrs RJ, Mahalingam R, Nagel MA. Varicella zoster virus infection: clinical features, molecular pathogenesis of disease, and latency. *Neurol Clin*. 2008;26(3):675-697.
29. Leuridan E, Hens N, Hutse V, Aerts M, Van Damme P. Kinetics of maternal antibodies against rubella and varicella in infants. *Vaccine*. 2011;29(11):2222-2226.
30. Nelsen RB. *An Introduction to Copulas*. Berlin, Germany: Springer Science & Business Media; 2007.
31. Shih JH, Louis TA. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*. 1995;51(4):1384-1399.
32. Bárdossy A, Li J. Geostatistical interpolation using copulas. *Water Resour Res*. 2008;44(7).
33. Danaher PJ, Smith MS. Modeling multivariate distributions using copulas: applications in marketing. *Mark Sci*. 2011;30(1):4-21.
34. de Leon AR, Wu B. Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Stat Med*. 2011;30(2):175-185.
35. Song PXK, Li M, Yuan Y. Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*. 2009;65(1):60-68.
36. Frahm G, Junker M, Schmidt R. Estimating the tail-dependence coefficient: properties and pitfalls. *Insur Math Econom*. 2005;37(1):80-100.
37. Shi P, Yang L. Pair copula constructions for insurance experience rating. *J Am Stat Assoc*. 2018;113(521):122-133.
38. Elias J, Findlow J, Borrow R, Tremmel A, Frosch M, Vogel U. Persistence of antibodies in laboratory staff immunized with quadrivalent meningococcal polysaccharide vaccine. *J Occupat Med Toxicol*. 2013;8(1):4.
39. Moulton LH, Halsey NA. A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics*. 1995;51(4):1570-1578.
40. Wiens BL. When log-normal and gamma models give different results: a case study. *Am Stat*. 1999;53(2):89-93.
41. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods*. 2002;7(2):147.
42. Ho T. Random decision forests (PDF). Paper presented at: Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada; 1995.
43. Gareth J. *An Introduction to Statistical Learning: With Applications in R*. New York, NY: Springer Verlag; 2010.
44. Atem FD, Qian J, Maye JE, Johnson KA, Betensky RA. Multiple imputation of a randomly censored covariate improves logistic regression analysis. *J Appl Stat*. 2016;43(15):2886-2896.
45. Do KT, Wahl S, Raffler J, et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics*. 2018;14(10):128.
46. Delignette-Muller ML, Dutang C, et al. fitdistrplus: an R package for fitting distributions. *J Stat Softw*. 2015;64(4):1-34.
47. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc Ser B (Methodol)*. 1996;58(1):267-288.
48. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. 1997;16(4):385-395.
49. Maertens K, Tran TMP, Hens N, Van Damme P, Leuridan E. Effect of prepregnancy pertussis vaccination in young infants. *J Infect Dis*. 2017;215(12):1855-1861.
50. Tran TMP, Maertens K, Hoang HTT, Van Damme P, Leuridan E, Hens N. Elucidating the difference in the kinetics of antibody titres of infants in Belgium and Vietnam. *Vaccine*. 2020;38(45):7079-7086.
51. Sen PK. Estimates of the regression coefficient based on Kendall's tau. *J Am Stat Assoc*. 1968;63(324):1379-1389.
52. Theil H. A rank-invariant method of linear and polynomial regression analysis. *Henri Theil's Contributions to Economics and Econometrics*. New York, NY: Springer; 1992:345-381.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Tran TMP, Abrams S, Aerts M, Maertens K, Hens N. Measuring association among censored antibody titer data. *Statistics in Medicine*. 2021;40:3740–3761. https://doi.org/10.1002/sim.8995