2021

A Semiparametric Modeling Approach for Analyzing Clinical Biomarkers Restricted to Limits of Detection

Sandipan Dutta
Old Dominion University, s1dutta@odu.edu

Susan Halabi

Original Publication Citation

Dutta, S., & Halabi, S. (2021). A semiparametric modeling approach for analyzing clinical biomarkers restricted to limits of detection. *Pharmaceutical Statistics*, *20*(6), 1061-1073. https://doi.org/10.1002/pst.2125

HHS Public Access

Author manuscript

Pharm Stat. Author manuscript; available in PMC 2021 November 21.

Published in final edited form as:

Pharm Stat. 2021 November ; 20(6): 1061–1073. doi:10.1002/pst.2125.

A Semiparametric Modeling Approach for Analyzing Clinical Biomarkers Restricted to Limits of Detection

Sandipan Dutta[1], **Susan Halabi**[2,**]

[1]Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529

[2]Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC 27710

Abstract

Before biomarkers can be used in clinical trials or patients' management, the laboratory assays that measure their levels have to go through development and analytical validation. One of the most critical performance metrics for validation of any assay is related to the minimum amount of values that can be detected and any value below this limit is referred to as below the limit of detection (LOD). Most of the existing approaches that model such biomarkers, restricted by LOD, are parametric in nature. These parametric models, however, heavily depend on the distributional assumptions, and can result in loss of precision under the model or the distributional misspecifications. Using an example from a prostate cancer clinical trial, we show how a critical relationship between serum androgen biomarker and a prognostic factor of overall survival is completely missed by the widely used parametric Tobit model. Motivated by this example, we implement a semiparametric approach, through a pseudo-value technique, that effectively captures the important relationship between the LOD restricted serum androgen and the prognostic factor. Our simulations show that the pseudo-value based semiparametric model outperforms a commonly used parametric model for modeling below LOD biomarkers by having lower mean square errors of estimation.

Keywords

Censored Regression; Limit of Detection; Pseudo-value; Semiparametric Model; Serum Androgen

1. Introduction

Biomarkers are known to play an important role in the disease progression in several diseases, such as oncology and cardiology and as such they have been extensively used in drug development[1,2]. Before biomarkers can be used in patient management or a clinical trial, laboratory assays measuring their levels need to undergo analytical validation. Occasionally values of the markers are so low that they are not detected by the measuring instruments or assays[3]. These observations are known as below the limit of detection (LOD)

**Corresponding Author: susan.halabi@duke.edu.

Data Availability Statement

The clinical trials data used in this article can be obtained through the NCI Data Archive.

or non-detects[4,5]. A large proportion of non-detects for a biomarker in a study would cause serious effects, irrespective if the biomarker is the main outcome or a covariate. Single or multiple detection limits can occur depending on whether the biomarker data contain measurements from one or multiple laboratories. Multiple detection limits can also occur in disease studies when scientists analyze a combination of multiple biomarkers and these biomarkers differ in their detection limits. Deleting observations below the LOD can lead to underpowered inference or biased results and thus it is critical to implement appropriate methods to optimally utilize the values below the LOD.

A common approach is to replace the values below the LOD by a either the LOD or LOD/2.[6,7] This strategy, however, introduces artificial skewness to the data that may lead to a biased inference. Alternative approaches have been suggested[4,8,9] to obtain summary measures of variable having observations below the LOD. However, most of these approaches involve imputations under strong distributional assumptions which has been known to perform poorly in case of mis-specified models with large proportions of non-detects[10]. Moreover, calculating only the summary measures of biomarkers below the LOD is not sufficient to ascertain its importance in clinical studies. It is critical to establish a direct relationship of such a biomarker with some explanatory variables in a disease process. Model- based approaches are widely used in establishing such direct associations. The Tobit model[11,12] is the one of the earliest and most popular model-based approach for observations below LOD as evident from its wide use in various fields of medicine and epidemiology[13,14,15,16]. It uses a likelihood-based approach for estimating model parameters under a normality assumption. There have been recent advancements to such model-based approaches by varying the underlying distributional assumptions, e.g., mixtures of normal and skew-normal distributions[17,18,19] and skewed t-distributions[20]. These parametric approaches are heavily dependent on the knowledge of the underlying distribution, and thus the estimates obtained from them may lead to highly unstable and biased results if the underlying true distribution of the biomarker data deviates from the distributional assumption of the model. This is especially the case when there is a large proportion of non-detects. One such scenario is discussed through a motivating example of a prostate cancer clinical trial, where a widely used parametric model fails to identify an important relationship between a serum androgen and a prognostic factor of prostate cancer survival.

The motivation behind this article is to develop a modeling approach, for biomarkers restricted to LODs, which is robust to the choice of the underlying distribution of the biomarker while producing accurate inference on the biomarker. Moreover, most of the existing parametric models, based on likelihood approaches, were developed mainly to address single lower LOD scenarios without any known optimal way to tackle situations where multiple LODs can occur in the same data. However, such multiple LODs can co-exist if the same biomarker is measured in different laboratories in different batches or if multiple biomarkers, with varying detection limits, are combined to obtain a classifier. In this article, we propose a semiparametric approach of modeling biomarkers subjected to LOD through a pseudo-value approach which is distribution-free and can uniformly handle single as well multiple LODs is the same data. The pseudo-values have been previously implemented in regression of complex quantities in time-to-event analysis[21,22,23,24,25];

however, we extend this approach in modeling biomarkers below LOD for the first time. We implement a semiparametric inference procedure for the biomarkers below the LOD by enabling a pseudo-value based estimation through a generalized estimating equation framework. Then, the resulting inference becomes robust, free from any rigid distributional assumptions, while being efficient as well. This is demonstrated through the several simulations where our proposed method has lower errors of estimation compared to a popular parametric model. Moreover, in our motivating example from a prostate cancer clinical trial, our proposed semiparametric method outperformed the widely used parametric method by identifying a significant association between the LOD restricted serum androgen biomarker and an important prognostic factor of overall survival. Although our method is based on the concept of pseudo-values, its implementation in modeling biomarker data below the LOD is a novel application which leads to a useful semi-parametric alternative to the standard parametric approaches.

The rest of the article is organized as follows. We describe the pseudo-value based semiparametric modeling of observations below the LOD in Section 2. In Section 3, we conduct several simulations to investigate the performance of the proposed approach compared to a standard parametric approach. In Section 4, we revisit our motivating example of a prostate cancer data analysis, along with another real-life example of serum cotinine analysis from the NHANES data, and we discuss the overall results and their implications in Section 5.

2. Method

2.1 Construction of pseudo-values

Our method is based on a two steps approach. In the first step of our semiparametric model is to create the pseudo-values. Suppose X_i is the observed biomarker value (in original or a log-transformed scale) of ith unit in the sample and Z_i is the vector of explanatory variables or covariates corresponding to the ith unit in a set of n i.i.d observations. Let $E(X_i) = \theta$ be the marginal mean and $\theta_i = E(X_i|Z_i)$ be the conditional expectation given the covariates. Suppose $\hat{\theta}$ is an estimator of marginal mean θ and $\hat{\theta}_{-i}$ is the "leave-one-out" estimator of θ obtained from the remaining data after deleting the ith observation. Note that, all the X_is may not be complete, since there are biomarker measurements below the LOD which are not accurately recorded. For example, if the true biomarker value (X_j^*) of the jth unit is below LOD, then the recorded biomarker value for the jth unit is X_j = LOD > X_j^*. This means that observations below a LOD are essentially left-censored. This feature of the data has to be taken into account while computing the marginal estimator $\hat{\theta}$ and the subsequent "leave-one-out" estimators $\hat{\theta}_{-i}$, $i = 1,2,\ldots, n$. Therefore, to account for the left-censored biomarker values below the LOD, we employ a nonparametric Kaplan-Meier[26] based estimator of mean response for obtaining $\hat{\theta}$ and $\hat{\theta}_{-i}$, $i = 1, 2, \ldots, n$. Then, for a data with n observations, the biomarker pseudo-value corresponding to the ith observation is defined as

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}, \quad i = 1, 2, \ldots, n$$

Note that, the pseudo-values are obtained for all the n observations irrespective of whether the biomarker value is below the LOD, i.e. left-censored, or not. Here, the pseudo-value for the unit i measures the contribution of the unit i on the overall mean estimate of the biomarker variable. Therefore, the pseudo-value $\hat{\theta}_i$ can be regarded as an estimate of $\theta_i = E(X_i|Z_i)$, the conditional mean given the covariates.

2.2 Semiparametric modeling using the pseudo-values

In the second step, we use the pseudo-values instead of the observed biomarker values, as the response variables for modeling the biomarkers given the covariates. We formulate the relationship between θ_i and Z_i through a generalized linear model of the form

$$g(\theta_i) = \beta^T Z_i, i = 1, 2, \ldots, n.$$

Here, β is the coefficient vector and $g(.)$ is a suitable link function. For modeling the biomarkers, we choose identity link as a choice for $g(.)$, although other choices, e.g. log-link for count data, are also possible. With the pseudo-values $\hat{\theta}_i$ ($i = 1,2, \ldots, n$) being treated as the observed values of the mean response θ_i and Z_is as explanatory variables, we estimate the coefficient β through a generalized estimating equation (GEE) given below.

$$U(\beta) = \sum_{i=1}^{n} U_i(\beta) = \sum_{i=1}^{n} \frac{\partial}{\partial \beta} g^{-1}(\beta^T Z_i) V_i^{-1} \left(\hat{\theta}_i - g^{-1}(\beta^T Z_i) \right) = 0.$$

Here, V_i is a working variance for $\hat{\theta}_i$. The variance estimates of $\hat{\beta}$ is obtained through the standard sandwich estimator as follows:

$$\hat{S} = I(\hat{\beta})^{-1} \widehat{var}(U(\beta)) I(\hat{\beta})^{-1}$$
$$\text{where } I(\beta) = \sum_{i=1}^{n} \left(\frac{\partial}{\partial \beta} g^{-1}(\beta^T Z_i) \right) V_i^{-1} \left(\frac{\partial}{\partial \beta} g^{-1}(\beta^T Z_i) \right), \widehat{var}(U(\beta)) = \sum_{i=1}^{n} U_i(\hat{\beta}) U_i(\hat{\beta})^T.$$

The use of the generalized estimating equation[27] and the sandwich variance estimator, makes the pseudo-value model robust to model and distributional misspecifications.

The large-sample properties of pseudo-value based regression methods in various settings have been examined in Overgaard et al.[28], Overgaard et al.[29], Graw et al.[30] among others. In particular, Overgaard et al.[28] showed that, under a general framework with independent censoring and certain regularity conditions, the estimated regression coefficients from the pseudo-value method are consistent and asymptotically normally distributed. The reader is referred to in Overgaard et al.[28] for more details on the regularity conditions.

3. Simulation Studies

3.1 Simulation Settings

Simulation setting 1: Single lower limit of detection—Let X_i be the ith observation for the variable of interest and Z_i be the vector of explanatory variables or covariates

corresponding to the ith observation in a set of n i.i.d observations. We use models for $E(\log(X))$ via the identity link. The outcomes are generated using the following model:

$$\log(X_i) = \beta_0 + \beta_1 Z_{Gi} + \beta_2 Z_{Bi} + \epsilon_i\ ,\ i = 1, 2, \ldots, n$$

where $\beta_0 = 0$, Z_G is a continuous covariate with samples drawn from $N(0, 0.5)$, and Z_B is a binary covariate with samples drawn from Bernoulli distribution with success probability of 0.4. Here, $Z_i = (Z_{Gi}, Z_{Bi})$. The errors are generated from one of the two following two distributions:

- **i.** Extreme-valued distribution such that $exp(\epsilon)$ follows a Weibull distribution with shape parameter 0.3 and scale parameter 3, to generate a skewed distribution
- **ii.** Normal distribution with mean 0 and variance 4 to generate a symmetric distribution.

We choose two different sample sizes (n), 60 and 300, to evaluate the performances of the proposed method under small and large sample size scenarios. We consider two choices for the lower limit of detection: the tenth percentile or the twenty-fifth percentile. This means that the lowest 10% or the lowest 25% of the sample biomarker values are set as observations below the LOD.

We apply the semiparametric pseudo-value (PV) model for estimating the regression coefficients. In addition, we use the parametric Tobit regression approach for left-censored outcome to investigate the comparative performance of the PV model. We used the Tobit regression approach because it is one of the most popular parametric models for analyzing LOD restricted responses and has been found out to be the most powerful by Wiegend et al.[14] in certain scenarios among a set of popular parametric models which included imputation based models, Bernoulli-Gaussian mixture models[17,31], and nonparametric Buckley-James estimator based regression models[32]. The Tobit model is based on the assumption that the response variables are normality distributed. For a Tobit Model, if the true biomarker value of the jth unit is X_j^* and the recorded biomarker value for the jth unit is X_j, then $X_j = \begin{cases} X_j^* & if\ X_j^* > LOD \\ LOD & otherwise \end{cases}$ and X_j is modeled using $\log(X_j) = \beta^T Z_j + \epsilon_j$ under the assumption that the error terms ϵ_js are normally distributed. Then, the regression coefficient vector β is estimated through a maximum likelihood estimation approach using the probability density function and the cumulative distribution function of a standard normal distribution. We have used the *censReg* package[33] in the *R* software (version 3.6.3) for estimating the MLE of β from Tobit models in our analyses.

Based on 1,000 Monte-Carlo simulations runs, we compare the two different methods by their average bias, average estimated model-based (theoretical) standard deviation (SD), average mean squared error (MSE), and average estimated 95% coverage probabilities. Moreover, we also compute the empirical SD of the estimators for both the methods based on the Monte-Carlo runs and compare them with the respective theoretical SD of the estimators for investigating the accuracy of the theoretical variance estimators.

Simulation setting 2: Multiple limits of detection—This simulation setting is similar to simulation setting 1, except that the total set of observations (n) is divided into three subsets (subsamples of equal size) and are treated as three independent sets of samples with different limits of detection. In each of the three subsamples, the lowest quartile of observed outcomes is set as the lower limit of detection for that subsample. This simulation setting mimics the scenario where the same biomarker is being evaluated using different measuring instruments (or assays) leading to different limits of detections. Under this simulation setting, we consider three different LOD with an overall percentage of below LOD observations as either of 10% or 25% in the full sample of size n.

We apply the semiparametric PV model for estimating the regression parameters. The application of the Tobit model, however, is limited in this setting since the standard Tobit model has been developed to model data having only one lower LOD. Since this setting has multiple LODs to be handled in the same analysis, we fit separate likelihood based Tobit models for each of the three sets of data which have different lower LODs and pool the estimates from these three models to obtain a final weighted estimate where the weights are based on the sample sizes of the three datasets. The average bias, theoretical SD, empirical SD, MSE and 95% coverage probabilities are calculated based on 1000 Monte-Carlo simulation runs.

3.2 Results of the Simulations

Table 1 presents the results for the extreme-value distributed outcomes having single detection limit. We observe that between the two methods, the PV model has the lowest MSE in estimating β_1 and β_2. Although the bias of estimation for the Tobit model is lower between the two methods, it produces highly unstable estimates with large standard deviations resulting in very high MSE. We also note that the bias of both the methods increase with larger proportion of observations below the LOD. On the other hand, the MSE decrease with the increased sample size. The estimated coverage probabilities for the Tobit model are close to the target coverage probability of 0.95 in almost all cases. Despite the fact the PV method has relatively larger bias, the estimated coverage of the PV model is close to that of the Tobit model as well as the target coverage probability of 0.95 in the majority of scenarios with extreme-value distributed outcomes.

Table 2 displays the results for the normally distributed outcomes having a single lower LOD. The Tobit model, has lower bias than the PV method but high SD leading to the higher MSE. Overall, the PV method has lower MSE than the Tobit model despite being a bit more biased. With the exception in few scenarios, the estimated coverage probabilities of both the PV and the Tobit models are close to the target coverage of 0.95. In few scenarios, the Tobit model slightly overshoots the target coverage, and the PV method has a coverage lower than the target.

We also provide figures in the supplement that describes the results of some extreme scenarios of very high proportion of outcomes below the LOD or a very large sample size. We observe for very large sample size of 1200, the PV model has lower MSE than the Tobit model for extreme-value distributed responses while the Tobit model's MSE tends to be lower in the normally distributed outcomes (Supplementary Figures 1 and 2). The

MSE of both methods, however, tend to decrease with increased sample sizes. In addition, we observe that the MSE of the PV model is lower than the MSE of the Tobit model in the majority of the scenarios under a very high rate of 60% outcomes below LOD (Supplementary Figures 3 and 4). This superiority of the semiparametric PV model as measured by MSE, holds good for such high proportion of below-LOD outcomes even when the underlying normality assumption of the parametric Tobit model in satisfied.

To summarize all the results from simulating setting 1, we note that the PV model outperforms the parametric Tobit model in terms of the MSE, especially for small sample sizes when the distributional assumptions of the parametric models are not met. Even in settings where the distributional assumptions of parametric models are met, the performance of the PV method is at least as good as the parametric model.

Tables 3 and 4 present the results for the bias, SD, MSE and coverage of the methods discussed in simulation setting 2 (multiple detection limits) for extreme-value and normally distributed outcomes, respectively. We note that the bias of the estimators from both the semiparametric PV model and the modified Tobit model, explained in Section 3.1, increase with a larger proportion of observations below LOD, while the MSE values decrease as sample size increases. The PV method has consistently lower MSE than the Tobit model for all the simulation scenarios considered in the setting 2. Moreover, for non-normal responses, the Tobit model can lead to unstable estimation results which is evident from the highly inflated MSE in estimating β_2 (Table 3). This is mainly due to the inaccurate variance estimation of the Tobit model as reflected by the large difference between the theoretical and the empirical SD of the Tobit model estimates for β_2 (Table 3). The estimated coverage probabilities of both models are close to the target coverage probability. While the Tobit model, due to its large variance estimate, has a wider coverage than the PV model in the majority of the cases, there exists some small sample size scenarios where the PV model's estimated coverage exceeds that of the Tobit model (Table 3 and 4). Overall, the semiparametric PV model is much more robust compared to the parametric Tobit model in modeling biomarkers restricted by multiple LODs.

4. Data Analysis

4.1 Serum Androgens Analysis from Prostate Cancer Clinical Trial

We analyze a prostate cancer trial from CALGB 90401 (Alliance) study[34] to demonstrate the usefulness of our proposed semiparametric method in modeling clinical biomarkers subjected to LOD. This study is a randomized phase III clinical trial involving 1050 patients with metastatic castration-resistant prostate cancer where the patients were distributed between two treatment arms (presence or absence of Bevacizumab). Among the laboratory variables, measurements on three types of serum androgens were recorded: testosterone (T), androstenedione (A), and dehydroepiandrosterone (D). Serum androgens are known to play important roles in prostate cancer progression[35]. An important goal here is to study the impact of age on baseline serum biomarker levels in prostate cancer patients. In addition, we also investigate association between race and the biomarker levels. Similarly, association of serum levels post-treatment with age, race, and treatment arm can also be investigated.

We apply the semiparametric PV model for each of the serum androgens. Age and race are considered as covariates for modeling each of the serum androgens separately. Age is modeled as a continuous variable while race (modeled as white or non-white) is a binary variable. In post-treatment analysis treatment group is also included as a binary covariate. The lower limits of detection vary between the three serum androgens depending on their units of measurement, and the resulting proportion of observations below the lower detection limits also differ between the three.

The LOD for T was 1 unit, while that for A and D were 5 units and 20 units respectively. At baseline, the percentage of patients with T, A, and D values that are below the LOD are 39%, 18% and 57%, respectively. The median T at baseline using all the samples was 1 unit, while the median T ignoring the observations below the LOD was 3 units. The baseline median of A using all the samples was 13.5 units while the median ignoring the below LOD samples was 17 units. For T, the baseline median using all the samples was less than 20 units and the median ignoring the below LOD values was 51 units. At 6-weeks post treatment, the percentage of patients with T, A, and D values below the LOD are 78%, 35% and 80%, respectively.

We use the logarithmic values of each serum measurement for modeling. We fit separate models at baseline and 6-weeks. We estimate the regression coefficients, and their sandwich variance estimates in each of the serum androgen PV models, and obtain the z-statistic and the corresponding p-value from each of the estimated coefficients. We also apply the Tobit model for each serum androgen and obtain the estimated coefficients and the corresponding p-values.

Table 5 shows the results from modeling of the three serum androgens at baseline using both the semiparametric PV approach and the Tobit model. We observe that age is negatively associated with the three serums androgens implying that serum levels tend to decrease as age increases. For T and D, the negative association with age was highly significant regardless of the method used. For A regression, the PV model infers marginal significance for association of A with age (p-value=0.08), while the Tobit model does not show evidence of association between A and age (p-value= 0.147). This is an interesting result since the PV model finds an important association between serum androgens and age which the Tobit model could not due a much larger variance estimate.

In Table 6, we present the results for the regression analyses at 6-weeks post-treatment. The proportions of patients with serum observations below the LOD (non-detects) have drastically increased at 6-week time compared to the baseline and this could be explained due to the treatment. Despite the high proportions of observations below the LOD, the PV model is still able to identify significant negative association between age and T (p-value=0.030), while the Tobit model fails to identify any association between T and age (p-value=0.134). These findings of association by the PV model are further supported by other clinical studies that have reported age to be associated with declines in serum androgens[36,37]. This highlights a scenario where the semiparametric PV model can capture a significant association that is missed by the parametric Tobit model in two separate models

(baseline and post-treatment). Statistically significant association is observed between A and treatment arm using both the PV model and the Tobit model.

4.2 Serum Cotinine Analysis from NHANES study

Serum Cotinine is a metabolite that can be used as a marker for both active smoking and passive smoking. In this section, we analyze the NHANES 2003-2004 data to investigate the effect of age and gender on the serum cotinine levels after adjusting for the smoking habits of the respondents. The data set included 1901 individuals who had information on their age, sex, smoking habits, and their serum cotinine levels were measured. From the data, we observed that 16% of the serum cotinine values were below the LOD. In order to assess the relationship of age and sex with serum cotinine, we fit censored regression models with serum cotinine levels as responses, and age, sex, and smoking habit (yes/no) as the covariates. We use our proposed semiparametric model as well as the parametric Tobit model for this analysis (Table 7). It is clear that age has positive significant association with the cotinine levels, while females have lower cotinine levels compared to males (Table 7). These association results are agreed upon by both the PV model and the Tobit model with different, but significant, p-values.

5. Discussion

This article is motivated by the necessity to develop a modeling approach for analyzing biomarkers with observations below the LOD that remains robust to misspecifications in model assumptions. To achieve this goal, we implement a semiparametric model based on pseudo-values which is free from distributional assumptions. To the best of our knowledge, we are the first to implement the pseudo-value (PV) approach for modeling biomarkers below the detection limit. Through our motivating prostate cancer clinical trial example, we emphasize the utility of our proposed semiparametric PV model approach in identifying important association between a LOD-restricted serum androgen biomarker and prognostic factor when the standard parametric Tobit model fails. Moreover, through simulation studies we show that our method produces lower MSE than the standard parametric regression model when the underlying distributional assumptions are violated. Even in simulated data where the distributional assumptions of the standard parametric models are met, the performance of our semiparametric model is competitive to that of the parametric models. In multi-center trials, often biomarkers are measured at institutional laboratories and LODs can differ between these resources. Unlike the standard Tobit model which fits separate likelihoods for different LODs resulting in multiple estimates of the same parameter of biomarker association, the semiparametric PV method can be easily applied to multiple lower detection limits to fit a single model on the same biomarker due to the different sources. Through the prostate cancer trial data, we show that, even in presence of high percentages of non-detects in serum androgens, our semiparametric method can identify signification biomarker associations.

Returning to our motivational example in prostate cancer, it has been reported that serum androgen levels have strong association with age of the patients[36,37]. In the 6-week post-treatment analysis, the serum androgen levels decline to great extents from the baseline

levels, due to the treatment effect, leading to the higher proportions of observations below the LOD than the baseline. Even in such a scenario, the semiparametric PV model is still able to identify significant negative association between age and T (p-value=0.030), while the Tobit model fails to identify any association between T and age (p-value=0.134). This raises the possibility of substantial loss of power of the parametric model due to the departure from the underlying assumptions and highlights the robustness of the PV model.

In summary, the semiparametric model based on pseudo-values is easy to use and implement. Our semiparametric PV model does not require distributional assumptions, it is robust to model misspecification than the standard parametric models. The pseudo-values are based on the leave-one-out jackknife technique. A typical pseudo-value for a unit measures the contribution of the unit on the overall mean of the response variable even if that unit has a censored response, i.e., a value below LOD. Hence, in the presence of multiple LODs in the same data, the PV model incorporates more information on the impact of the below LOD samples on the overall summary statistic, compared to an ad-hoc substitution approach that merely replaces all the observations below LOD by an artificial value, e.g., LOD/2. Investigators are encouraged to use this approach whenever they encounter single or multiple observations below the LOD. The use of biomarkers will continue to be an important area of research not only in diagnosing patients but will be also used in treating patients with disease.

We have focused in this article on the modeling of biomarkers as responses. For the serum androgen data, including baseline serum androgen as a covariate in modeling post-treatment serum level will lead to a more complicated modeling scenario where both response and one of the covariates are restricted to an LOD. As a future research, we plan to examine this complex modeling scenario and aim to develop methods for addressing it. We will also consider modelling the changes of serum levels from baseline to 6-weeks as many patients have serum levels recorded as below LOD at both time points.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

This research was supported in part by National Institutes of Health Grants R21 CA195424-01 and R01 CA256157-01; United States Army Medical Research, Grant/Award Numbers: W81XWH-15-1-0467, W81XWH-18-1-0278; and the Prostate Cancer Foundation

References

1. Henry NL, Hayes DF. Cancer biomarkers. Molecular Oncology 2012;6: 140–146. [PubMed: 22356776]
2. Frank R, Hargreaves R. Clinical biomarkers in drug discovery and development. Nature Reviews Drug Discovery 2003;2: 566–580. [PubMed: 12838269]
3. MacDougall D, Crummett WB.. Guidelines for data acquisition and data quality evaluation in environmental chemistry. Analytical Chemistry 1980;52: 2242–2249.
4. Helsel DR. Less than obvious-statistical treatment of data below the detection limit. Environmental Science & Technology 1990;24:1766–1774.

5. Armbruster DA, Pry T. Limit of blank, limit of detection and limit of quantitation. The Clinical Biochemist Reviews 2008;29:S49. [PubMed: 18852857]
6. Hornung RW, Reed LD. Estimation of average concentration in the presence of nondetectable values. Applied Occupational and Environmental Hygiene 1990;5:46–51.
7. Helsel DR. Nondetects and data analysis. Statistics for censored environmental data. Wiley-Interscience, New York; 2005.
8. Schisterman EF, Vexler A, Whitcomb BW, Liu A. The limitations due to exposure detection limits for regression models. American Journal of Epidemiology 2006;163:374–383. [PubMed: 16394206]
9. Gillespie BW, Chen Q, Reichert H, Franzblau A, et al. Estimating population distributions when some data are below a limit of detection by using a reverse Kaplan-Meier estimator. Epidemiology 2010;21:S64–S70. [PubMed: 20386104]
10. Arunajadai SG, Rauh VA. Handling covariates subject to limits of detection in regression. Environmental and Ecological Statistics 2012;19:369–391.
11. Tobin J Estimation of relationships for limited dependent variables. Econometrica 1958;26:24–36.
12. Uh HW, Hartgers FC, Yazdanbakhsh M, Houwing-Duistermaat JJ. Evaluation of regression methods when immunological measurements are constrained by detection limits. BMC Immunology 2008;9:1–10. [PubMed: 18211710]
13. Tellez-Plaza M, Navas-Acien A, Crainiceanu C, Guallar E. A Tobit Model to Address the Instrumental Limit of Detection in the Study of Blood Cadmium and Peripheral Arterial Disease in US Adults. Epidemiology 2009;20:S187.
14. Wiegand RE, Rose CE, Karon JM. Comparison of models for analyzing two-group, cross-sectional data with a Gaussian outcome subject to a detection limit. Statistical Methods in Medical Research 2016;25: 2733–2749. [PubMed: 24803511]
15. Wang W, Griswold ME. Natural interpretations in Tobit regression models using marginal estimation methods. Statistical Methods in Medical Research 2017;26:2622–2632. [PubMed: 26329751]
16. Soret P, Avalos M, Wittkop L, Commenges D, et al. Lasso regularization for left-censored Gaussian outcome and high-dimensional predictors. BMC Medical Research Methodology 2018;18:159. [PubMed: 30514234]
17. Moulton LH, Halsey NA. A mixture model with detection limits for regression analyses of antibody response to vaccine. Biometrics 1995;51:570–1578.
18. Mattos TDB., Garay AM, Lachos VH. Likelihood-based inference for censored linear regression models with scale mixtures of skew-normal distributions. Journal of Applied Statistics 2018;45: 2039–2066.
19. Zeller CB, Cabral CRB, Lachos VH, Benites L. Finite mixture of regression models for censored databased on scale mixtures of normal distributions. Advances in Data Analysis and Classification 2019;13:89–116.
20. Kim S, Chen Z, Perkins NJ, Schisterman EF, et al. A Model-Based Approach to Detection Limits in Studying Environmental Exposures and Human Fecundity. Statistics in Biosciences 2019;11:524–547. [PubMed: 33072224]
21. Andersen PK, Klein JP, Rosthøj S. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika 2003;90:15–27.
22. Andrei AC, Murray S. Regression Models for the Mean of the Quality-of-Life-Adjusted Restricted Survival Time Using Pseudo-Observations. Biometrics 2007;63:398–404. [PubMed: 17688492]
23. Overgaard M, Andersen PK, Parner ET. Regression analysis of censored data using pseudo-observations: An update. The Stata Journal 2015;15:809–821.
24. Ahn KW, Logan BR. Pseudo-value approach for conditional quantile residual lifetime analysis for clustered survival and competing risks data with applications to bone marrow transplant data. The Annals of Applied Statistics 2016;10:618–637. [PubMed: 29081872]
25. Dutta S, Datta S, Datta S. Temporal prediction of future state occupation in a multistate model from high-dimensional baseline covariates via pseudo-value regression. Journal of Statistical Computation and Simulation 2017;87:1363–1378. [PubMed: 29217870]
26. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958;53:457–481.

27. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73, 13–22.
28. Overgaard M, Parner ET, Pedersen J. Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. The Annals of Statistics 2017;45:1988–2015.
29. Overgaard M, Parner ET, Pedersen J. Pseudo-observations under covariate-dependent censoring. Journal of Statistical Planning and Inference 2019;202:112–122.
30. Graw F, Gerds TA, Schumacher M. On pseudo-values for regression analysis in competing risks models. Lifetime Data Analysis 2009;15:241–255. [PubMed: 19051013]
31. Moulton LH, Curriero FC, Barroso PF. Mixture models for quantitative HIV RNA data. Statistical Methods in Medical Research 2002;11:317–325. [PubMed: 12197299]
32. Buckley J, James I. Linear regression with censored data. Biometrika 1979;66:429–436.
33. Henningsen A censReg: Censored Regression (Tobit) Models. 2020. R package version 0.5-32.
34. Kelly WK, Halabi S, Carducci M, George D, et al. Randomized, double-blind, placebo-controlled phase III trial comparing docetaxel and prednisone with or without bevacizumab in men with metastatic castration-resistant prostate cancer: CALGB 90401. Journal of Clinical Oncology 2012;30:1534–1540. [PubMed: 22454414]
35. Huggins C, Hodges CV. Studies on prostatic cancer. Cancer Research 1941;1:293–297.
36. Vermeulen A. Androgens in the aging male. The Journal of Clinical Endocrinology and Metabolism 1991;73, 221–224. [PubMed: 1856256]
37. Swerdloff RS, Wang C. Androgens and aging in men. Experimental Gerontology 1993;28:435–446. [PubMed: 8224040]

Table 1.

Simulation results based on 1000 Monte-Carlo runs with Weibull error distribution and single lower limit of detection (LOD)

	$n = 60$											
	β_1						β_2					
	10% below LOD			25% below LOD			10% below LOD			25% below LOD		
	Bias	**Theoretical (Empirical) SD**	**MSE (coverage)**	**Bias**	**Theoretial (Empirical) SD**	**MSE (coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (coverage)**
PV model	−0.139	0.872 (0.915)	1.643 (0.926)	−0.271	0.698 (0.746)	1.135 (0.922)	−0.135	0.901 (0.959)	1.763 (0.936)	−0.262	0.727 (0.778)	1.213 (0.912)
Tobit model	−0.064	0.989 (1.006)	2.018 (0.936)	−0.108	0.923 (0.954)	1.793 (0.948)	−0.079	0.995 (1.059)	2.135 (0.942)	−0.098	0.926 (0.986)	1.852 (0.936)
	$n = 300$											
	β_1						β_2					
	10% below LOD			25% below LOD			10% below LOD			25% below LOD		
	Bias	**Theoretical (Empirical) SD**	**MSE (coverage)**	**Bias**	**Theoretial (Empirical) SD**	**MSE (coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (coverage)**
PV model	−0.099	0.398 (0.409)	0.336 (0.944)	−0.241	0.312 (0.324)	0.261 (0.856)	−0.097	0.411 (0.432)	0.366 (0.926)	−0.245	0.325 (0.330)	0.275 (0.890)
Tobit model	−0.031	0.445 (0.450)	0.402 (0.952)	−0.050	0.410 (0.420)	0.348 (0.944)	−0.034	0.452 (0.473)	0.430 (0.940)	−0.068	0.416 (0.419)	0.353 (0.940)

Table 2.

Simulation results based on 1000 Monte-Carlo runs with Normal Error distribution and single (LOD)

	$n = 60$											
	β_1						β_2					
	10% below LOD			25% below LOD			10% below LOD			25% below LOD		
	Bias	**Theoretical (Empirical) SD**	**MSE (coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (coverage)**
PV model	−0.165	0.921 (0.996)	1.894 (0.934)	−0.222	0.797 (0.905)	1.528 (0.904)	−0.092	0.947 (1.095)	1.916 (0.928)	−0.249	0.842 (0.884)	1.564 (0.926)
Tobit model	−0.060	1.044 (1.000)	2.312 (0.940)	0.050	1.067 (1.197)	2.600 (0.922)	0.008	1.051 (1.104)	2.334 (0.936)	−0.012	1.069 (1.121)	2.414 (0.934)
	n=300											
	β_1						β_2					
	10% below LOD			25% below LOD			10% below LOD			25% below LOD		
	Bias	**Theoretical (Empirical) SD**	**MSE (coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (coverage)**
PV model	−0.218	0.383 (0.364)	0.330 (0.892)	−0.293	0.353 (0.336)	0.325 (0.852)	−0.180	0.396 (0.389)	0.342 (0.920)	−0.262	0.367 (0.356)	0.331 (0.896)
Tobit model	−0.017	0.466 (0.425)	0.400 (0.960)	−0.017	0.478 (0.444)	0.426 (0.960)	0.015	0.474 (0.450)	0.427 (0.958)	0.013	0.484 (0.463)	0.449 (0.954)

Table 3.

Simulation results based on 1000 Monte-Carlo runs with Weibull Error distribution with multiple LOD

	$n = 60$											
		β_1					β_2					
	10% below LOD			25% below LOD			10% below LOD			25% below LOD		
	Bias	**Theoretical (Empirical) SD**	**MSE (Coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (Coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (Coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (Coverage)**
PV model	−0.157	0.871 (0.937)	1.688 (0.942)	−0.213	0.788 (0.856)	1.419 (0.916)	−0.174	0.901 (0.960)	1.776 (0.922)	−0.216	0.818 (0.868)	1.480 (0.916)
Tobit model	−0.048	1.012 (1.108)	2.283 (0.926)	−0.086	0.949 (1.034)	2.003 (0.924)	−0.118	1.947 (1.168)	455.2 (0.928)	−0.162	3.647 (1.115)	1257.2 (0.928)
	$n = 300$											
	β_1						β_2					
	10% below LOD			25% below LOD			10% below LOD			25% below LOD		
	Bias	**Theoretical (Empirical) SD**	**MSE (Coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (Coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (Coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (Coverage)**
PV model	−0.093	0.414 (0.425)	0.363 (0.938)	−0.193	0.354 (0.360)	0.293 (0.910)	−0.091	0.428 (0.454)	0.398 (0.930)	−0.189	0.368 (0.391)	0.325 (0.898)
Tobit model	−0.028	0.447 (0.456)	0.409 (0.950)	−0.061	0.415 (0.427)	0.359 (0.942)	−0.033	0.451 (0.480)	0.435 (0.944)	−0.070	0.417 (0.447)	0.380 (0.928)

Table 4.

Simulation results based on 1000 Monte-Carlo runs with Normal Error distribution with multiple LOD

	$n = 60$											
	β_1						β_2					
	10% below LOD			25% below LOD			10% below LOD			25% below LOD		
	Bias	**Theoretical (Empirical) SD**	**MSE (Coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (Coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (Coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (Coverage)**
PV model	−0.140	0.942 (1.015)	1.964 (0.934)	−0.226	0.875 (0.948)	1.745 (0.934)	−0.069	0.967 (1.018)	1.985 (0.936)	−0.159	0.904 (0.969)	1.796 (0.918)
Tobit model	−0.045	1.064 (1.161)	2.507 (0.924)	−0.045	1.093 (1.197)	2.662 (0.928)	0.002	1.042 (1.153)	2.429 (0.926)	−0.030	1.065 (1.178)	2.543 (0.922)
	$n = 300$											
	β_1						β_2					
	10% below LOD			25% below LOD			10% below LOD			25% below LOD		
	Bias	**Theoretical (Empirical) SD**	**MSE (Coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (Coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (Coverage)**	**Bias**	**Theoretical (Empirical) SD**	**MSE (Coverage)**
PV model	−0.207	0.388 (0.372)	0.334 (0.894)	−0.266	0.365 (0.350)	0.328 (0.872)	−0.168	0.401 (0.393)	0.345 (0.922)	−0.232	0.379 (0.366)	0.333 (0.912)
Tobit model	−0.014	0.468 (0.437)	0.411 (0.958)	−0.016	0.480 (0.454)	0.438 (0.956)	0.020	0.472 (0.454)	0.430 (0.956)	0.011	0.483 (0.465)	0.450 (0.948)

Table 5.

Results from the semiparametric PV model and Tobit model analysis of serum androgens at baseline from the CALGB 90401 prostate cancer data

Covariates	Estimated regression coefficients					
	Testosterone model		Androstenedione model		Dehydroepiandrosterone model	
	PV model coefficient (p-value)	Tobit model coefficient (p-value)	PV model coefficient (p-value)	Tobit model coefficient (p-value)	PV model coefficient (p-value)	Tobit model coefficient (p-value)
Age	−0.011 (0.004)	−0.026 (0.004)	−0.006 (0.080)	−0.006 (0.147)	−0.020 (<0.001)	−0.044 (<0.001)
Race (White vs. others)	0.107 (0.280)	0.110 (0.636)	−0.062 (0.488)	−0.080 (0.476)	0.033 (0.636)	0.046 (0.777)

Table 6.

Results from the semiparametric PV model and Tobit model analysis of serum androgens at 6-weeks post-treatment from the CALGB 90401 prostate cancer data

Covariates	Estimated regression coefficients					
	Testosterone model		Androstenedione model		Dehydroepiandrosterone model	
	PV model coefficient (p-value)	Tobit model coefficient (p-value)	PV model coefficient (p-value)	Tobit model coefficient (p-value)	PV model coefficient (p-value)	Tobit model coefficient (p-value)
Age	−0.009 (0.030)	−0.026 (0.134)	−0.002 (0.568)	0.000 (0.999)	−0.008 (0.001)	−0.029 (0.001)
Race (White vs. others)	−0.124 (0.329)	−0.324 (0.469)	−0.058 (0.446)	−0.087 (0.458)	−0.122 (0.065)	−0.635 (0.003)
Treatment Arm	0.080 (0.263)	0.438 (0.139)	0.101 (0.046)	0.177 (0.019)	0.055 (0.116)	0.434 (0.006)

Table 7.

Results from the modeling of serum cotinine from NHANES 2003-04 study to determine the effects of age and gender on cotinine levels obtained after adjusting for smoking habits

Covariates	Regression Coefficient (p-value)	
	PV model	Tobit model
Age	0.134 (<0.0001)	0.147 (<0.0001)
Gender (Female against male)	−0.418 (0.0003)	−0.517 (0.0001)
Smoking status (Smoker against non-smoker)	2.104 (<0.0001)	2.348 (<0.0001)