Fitting logistic regression model when the independent variable present limits of detection

van der Zwaag, Tijmen K.

Award date:
2024

Fitting logistic regression model when the independent variable present limits of detection

T.K. van der Zwaag 1256610
Supervisor: dr. Regis, Marta
Bachelor final project Applied Mathematics
Eindhoven University of Technology

April 29, 2024

Contents

Abstract

In biomedical research it often occurs that biomarker values are not quantifiable for a portion of the observations due to a limit of detection (LOD). Multiple methods exist to deal with data that is missing not at random. In this paper the most commonly used methods from the literature are analyzed by fitting a logistic regression model with the independent variable subject to an upper detection limit. In a simulation study, the performance of these methods is quantified using bias and root-mean-square error, for multiple distributions and LOD values. As an addition to existing literature, this research will also focus on the predictive performance of these methods. Using the results from the simulation we recommend the complete case (CC) method for situations in which a small part of the data is censored ($< 25\%$). For this low censoring percentage all methods scored equivalent, but the CC method has as large benefit its simplicity. For larger censoring proportions the missing data indicator method or a maximum likelihood approach are preferred. Substitution methods were generally found to be biased for moderate and large censoring proportions. The predictive performance on censored values was largely found to be equivalent for all methods across all settings.

1 Introduction

Logistic regression models the log-odds (logit transformation of the probability) of belonging to a certain class as a function of covariates (linear predictor). In medical science, researchers are often interested in the association between variables (such as biomarkers) and outcomes of interest for early detection of deceases. [Yuan and Ghosh, 2008]

However, in practice, it often occurs that these biomarker values are not quantifiable for a portion of the participants because the true value falls below the limit of detection (LOD). [Schisterman et al., 2006] We say that x is subject to left censoring due to a limit of detection if the true value x_t can not be quantified below that limit, i.e.

$$\begin{aligned} x = x_t \quad &\text{if} \quad x \geq LOD \\ x = \text{non-detect (ND)} \ &\text{if} \quad x < LOD \end{aligned}$$

There are cases in the literature in which the level of the biomarker is small enough that the measurement instrumentation is not accurate enough to determine the exact value. An example can be given by a research study performed by McNamara et al. on exposure of ionizing radiation for workers at a nuclear power plant. [McNamara et al., 2018] In this study 27% of all measurements performed by a dosimeter fell below the LOD value of 0.1 mrem. It was assumed that the workers in a power plant had a non-zero exposure to radiation. To deal with these missing values maximum likelihood estimation was used to estimate the distribution of the exposure, then create five data sets and replace missing data with a value sampled from the conditional distribution of the exposure.

It can also occur that the covariate value can not be determined above a certain threshold, then the variable is subject to right censoring. A study performed on treatment of HIV/AIDS by Mwanda et al. had to deal with detection limits during data collection. The measurement equipment for HIV-1 plasma RNA, an important marker for HIV advancement detection, could detect values between 400 copies/mL and 750,000 copies/mL. In this case 12,8% of observations were censored (5,3% left censored and 7,5% right censored). [Mwanda et al., 2009]. A follow up study was performed by Fu et al. to investigate the influence of the censored data on the results. Data analysis with Monte Carlo Expectation-Maximization algorithm was found to be the least biased method. Its estimates were less sensitive to censoring than other methods such as substitution methods and multiple imputation. [Fu et al., 2016]

Various methods have been proposed in the literature for fitting logistic regression models in the presence of censored independent variables. The majority of the studies investigate left censoring for biomarker research or analyte analysis in chemistry, when the true values fall below the measurement sensitivity of the instrument. These include straightforward methods such as a complete

case analysis (CC), that disregards the incomplete data or substitution methods which replace the missing data with fixed values (such as LOD, LOD/2, LOD/$\sqrt{2}$) [Bernhardt et al., 2015]. Other methods include the missing indicator approach(MDI) [Chiou et al., 2019], which accounts for the partial missing data by adding indicator variables to the logistic regression model, maximum likelihood approaches and multiple imputation methods. [Bernhardt et al., 2015] Each of these methods have their own underlying assumptions, advantages and drawbacks. These methods have been compared under multiple settings, such as varying distributional assumptions, multiple censored variables and different LOD values.

In a simulation study performed by Bernhardt et al. a logistic regression model with two covariates was analyzed. [Bernhardt et al., 2015] The two regressor variables were generated according to a multivariate normal distribution and were both subject to left censoring. In this simulation they compared multiple existing methods including CC, substitution (LOD/$\sqrt{2}$, $\mathbb{E}[X|X \leq LOD]$, $\mathbb{E}[X|X \leq LOD, y]$), a maximum likelihood method with their novel improper multiple imputation technique. The performance of the methods was compared at different proportions of censoring (20%, 40%, 60%). It was concluded that the CC-method is a consistent estimator for generalized linear models, and in cases with a limited proportion of missing data values it is robust, valid and easy to apply. However, it is inefficient since partial data is discarded and therefore not suitable for cases in which there is a large proportion of observations subject to the limit of detection. The substitution methods resulted in biased parameter estimations, while CC, MLE and multiple imputation resulted in minimal or no biased estimators. However, the CC method was deemed inferior because the variance in the estimates was larger and the method was less efficient compared to other unbiased methods. The maximum likelihood method was the most efficient. Since most methods required parametric assumptions, it was also tested how well the parametric methods would perform if the distributional assumptions were violated. In this case, the two covariates were actually generated according to a bivariate gamma distribution, while the assumed distribution (for likelihood estimation and expected value substitution) was a normal distribution. Under these wrong assumptions all methods except CC had clear bias, however, standard error estimation for these methods was "still reasonable" even for high censoring rates. The simulation showed that a violation of the distributional assumption for the censored covariate does not have a extreme influence on the estimates.

In a study performed by Ortega-Villa et al. multiple methods were compared to account for the exposure below the detection limit. [Ortega-Villa et al., 2021] This study also focused on investigating the results of the methods under incorrect distributional assumptions. The explanatory variable for this simulation followed a Gaussian distribution and was censored at different LOD values. In this study CC, subtitution (LOD/$\sqrt{2}$), missing data indicator (MDI), maximum likelihood approach, and multiple imputation (not conditioned on the response

variable) were compared. It was noted that for a small portion of data under the LOD($< 20\%$), the CC, MDI and multiple imputation methods provided nearly unbiased results. The LOD/$\sqrt{2}$ substitution produced biased results and this bias only grew with the increase of the proportion of the data below LOD. Generally, it was concluded that the MDI method was best for censored data in practice, despite being less efficient than the maximum likelihood approach under a correctly specified model, since it is less affected by unverifiable modeling assumptions than other methods. These unbiased results were only found under two important assumptions. Firstly, the distribution of the censored covariate is correctly assumed or identified, even below the LOD value. Secondly, the linearity assumptions should hold, even under the LOD. If these assumptions are not met, the regression coefficient can be seriously biased, even with relatively few missing data.

Schisterman et al. evaluated two distributions, bimodal normal and gamma, to simulate biomarkers. [Schisterman et al., 2006] The proportion of values falling below the LOD of 25%, 50% and 75% were investigated. This study examined only substitution-based methods (0, LOD, LOD/2, $\mathbb{E}[X|X < LOD]$). For logistic regression models without intercept the substitution method of value 0 produced minimally biased estimates. Likewise, replacing missing values with $\mathbb{E}[X|X < LOD]$ resulted in minimal bias. Using replacement values of LOD and LOD/2 did result in substantial bias in the estimation of regression parameters.

Chiou et al. studied the performance of MDI relative to other methods such as CC, substitution (LOD/2) and two multiple imputation approaches designed for data missing at random, namely, *mice* and *missForest*. [Chiou et al., 2019] In this simulation study, a model with multiple covariates was considered from which two were left censored. In one simulation they were generated according to a multivariate normal distribution and in the second simulation from a non-normal distribution namely, from a normal copula with exponential and gamma marginal distributions for the two censored variables and a uniform distribution for other covariates. MDI, an expanded version of MDI and CC were among the methods with smallest bias. Generally, CC en MDI had equivalent bias over all settings. However, MDI outperformed the CC approach in terms of mean square error for small and medium sample sizes. In most settings the substitution method had a larger bias than MDI except for the non-normal distribution since LOD/2 would be closer to $\mathbb{E}[X|X \leq LOD]$ for this distribution. The largest bias was obtained by multiple imputation using the *mice* and *missForest* algorithms. This bias was the result of both algorithms producing imputation values inside the range of quantification for the censored covariates. The mean square error of both multiple imputation methods where equal or occasionally smaller compared to MDI. Generally, methods performed better under the non-normal distribution compared to the normal distribution, while maintaining the advantage for MDI over other methods.

The goal of this project is to summarize some of these methods for fitting logistic

regression models in presence of a censored independent variable. An overview will be given listing the assumptions, usage, advantages and disadvantages for all methods. Firstly we will perform a literature research. Afterwards, we will implement these methods in a simulation study with right censored data. This simulation study is inspired by a case study by van Hooff et al. which used a time to recovery variable to diagnose flow limitations in the iliac artery. [van Hooff et al., 2022] In this study the independent time-related variable is right censored due to measurement constrains. The performance for each method will be examined for different distributions and multiple LOD values. In addition to the existing literature, which focuses mainly on bias and variance of the estimates of the coefficients of the regression model, we will compare the methods on their predictive performance by reporting their sensitivity, specificity, AUC and Youden's J statistic.

2 Mathematical Model

2.1 Logistic regression model

Let $Y \in \{0, 1\}$ be a Bernoulli distributed random variable. The logistic regression models the probability of observing an event as function of some independent variable X.

$$p(X) := \mathbf{P}_\beta(Y = 1|X = \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{x}\beta}} \tag{1}$$

in which $\beta = [\beta_0, \ldots, \beta_n]^T$ is a $(n + 1) \times 1$ dimensional vector of regression coefficients and $\mathbf{x} = [1, x_1, \ldots, x_n]$ is a $1 \times (n+1)$ dimensional vector containing covariate values.

$$p(X) = g(\beta X) \tag{2}$$

in which $g(t) = (1 + e^{-t})^{-1}$ is the logistic function.

2.2 Maximum Likelihood Estimator

In order to estimate the parameter of interest β, one can use maximum likelihood estimation. Let $\mathbf{y} = [y_1, \ldots, y_n]$ and $\mathbf{x} = [1, \mathbf{x}_1, \ldots, \mathbf{x}_n]$ be a realization of the data, in which $\mathbf{x}_i = [x_{1,i}, \ldots, x_{k,i}]$ is a vector of covariates for unit i. The maximum likelihood estimator for β is the value $\hat{\beta}$ that probability of the outcome y is maximized under $\hat{\beta}$ given x, i.e.

$$\hat{\beta} = arg \max_{\beta} L(\beta) \tag{3}$$

Assuming Y_i are independent, the likelihood function $L(\beta)$ can be written as,

$$L(\beta|\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} L_i(\beta) \quad , L_i(\beta) = \begin{cases} p_i(\beta) & \textbf{for } y_i = 1 \\ 1 - p_i(\beta) & \textbf{for } y_i = 0 \end{cases}$$

where $p_i(\beta) = (1 + e^{-\mathbf{x}_i\beta})^{-1}$. Now,

$$L(\beta) = \prod_{i=1}^{n} p(\beta)^{y_i} \cdot (1 - p(\beta))^{(1-y_i)} \tag{4}$$

The log-likelihood $l(\beta)$ can be given by

$$\begin{aligned}
l(\beta) = log(L(\beta)) &= log(\prod_{i=1}^{n} p(\beta)^{y_i} \cdot (1-p(\beta))^{(1-y_i)}) \\
&= \sum_{i=1}^{n} [log(p(\beta)^{y_i}) + log((1-p(\beta))^{(1-y_i)})] \\
&= \sum_{i=1}^{n} y_i log(p(\beta)) + (1-y_i) log(1-p(\beta)) \\
&= \sum_{i=1}^{n} y_i (log(p(\beta)) - log(1-p(\beta))) + log(1-p(\beta)) \\
&= \sum_{i=1}^{n} y_i log(\frac{p(\beta)}{1-p(\beta)}) + log(1-p(\beta)) \\
&= \sum_{i=1}^{n} y_i log(\frac{1}{e^{-x_i\beta}}) + log(\frac{e^{-x_i\beta}}{1+e^{-x_i\beta}}) \\
&= \sum_{i=1}^{n} y_i log(e^{x_i\beta}) + log(\frac{1}{1+e^{x_i\beta}}) \\
&= \sum_{i=1}^{n} y_i x_i \beta - log(1+e^{x_i\beta})
\end{aligned}$$

Since the log-function is a strictly increasing function, the value β for which $l(\beta)$ is maximized will be the same value for which the likelihood function $L(\beta)$ is maximized.

The critical values of $l(\beta)$ can be found by setting the gradient to zero, that is,

$$[\frac{\partial l}{\partial \beta_1}, \ldots, \frac{\partial l}{\partial \beta_k}] = 0 \tag{5}$$

The function l is non-linear with respect to β and there is no analytic solution to equation (5). This problem requires the use of numerical approximation methods for which multiple algorithms exist.

3 Methods for dealing with independent variables subjected to detection limits in logistic regression

3.1 Complete case analysis (CC-Analysis)

The most straightforward way to deal with variables subject to a limit of detection is to omit these values in the analysis. Complete case (CC) analysis only performs subsequent analysis on known values and discards incomplete information. This method is easy to use, however, it is not suitable in datasets that contain many missing values since a large part of the data set will not be used and may lead to critically low sample sizes. [Alyabs and Chiou, 2022]

3.2 The substitution method

Another method for dealing with non-detects is to substitute a fixed value for all missing data. For left censoring, values such as LOD, $LOD/2$, $LOD/\sqrt{2}$ and $\mathbf{E}[X|X \leq LOD]$ are often used in practice. [Lubin et al., 2004]. For the simulation data, which is right censored, substitution values of $LOD, 2{\cdot}LOD, \sqrt{2}{\cdot}LOD$ and $\mathbf{E}[X|X > LOD]$ will be used.

3.3 The missing data indicator method (MDI)

In this method, additional parameters get added to the logistic regression model to account for missing values. Let x_i be the variable affected by a limit of detection for observation i. Let z_i denote a vector of uncensored covariates. The logistic regression model can be rewritten as

$$g^{-1}(\mathbf{P}(y_i = 1)) = \beta_0 + x_i\beta_x + z_i\beta_z \tag{6}$$

An indicator variable, denoted by $\mathbb{1}_M$, will be introduced which takes the value $\mathbb{1}_M = 1$ if the value x_i is missing and $\mathbb{1}_M = 0$ otherwise. The new logistic regression model for the MDI method is then given by

$$g^{-1}(\mathbf{P}(Y = 1)) = \beta_0 + x\beta_x(1 - \mathbb{1}_M) + \mathbb{1}_M\beta_M + z\beta_z \tag{7}$$

Note that this model also introduces a correction to the intercept β_M. [Chiou et al., 2019]

3.4 Adjusted Maximum Likelihood Estimation

A way to account for the missing variable in the standard MLE is to introduce an additional term that adds an expected contribution of the censored cases to the likelihood. For left censoring, Ortega-Villa et al. proposed an adjusted likelihood function of the form

$$L = \prod_{i=1}^{n} \left[f(y_i, x_i|z_i)(1 - \mathbb{1}_M) + \int_{-\infty}^{LOD} f(y_i, x_i|z_i)\mathbb{1}_M dx_i \right]. \tag{8}$$

[Ortega-Villa et al., 2021] Bernhardt et al. defined a likelihood that is solely considering the censored values. [Bernhardt et al., 2015]

$$L = \prod_{i=1}^{n} \int_{-\infty}^{LOD} f(y_i, x_i|z_i) dx_i \tag{9}$$

In our simulation study we will adopt the first approach in equation (10) but adapted to fit right censoring.

$$L = \prod_{i=1}^{n} \left[f(y_i, x_i|z_i)(1 - \mathbb{1}_M) + \int_{LOD}^{\infty} f(y_i, x_i|z_i) \mathbb{1}_M dx_i \right]. \tag{10}$$

3.5 Multiple Imputation

Instead of imputing one fixed value, as is the case in the substitution method, multiple data sets can be constructed with probabilistic approaches. For this method the incomplete data set is copied multiple times (recommended 3-5 or 10 times if a great proportion of data is missing [Lubin et al., 2004]), the imputations are filled in by a probabilistic method, such that each copy is now a unique and complete data set. For each copy, estimates for coefficients β are calculated. Lastly, all estimates are pooled together to obtain a single estimate for β.

3.5.1 Sampling from a conditional distribution

A parametric method that will be used in the simulation is a multiple imputation method based on sampling from a conditional distribution. Firstly, an estimate for the distributional parameter set $\hat{\theta}$ is calculated to get an estimate for the assumed distribution $f(x|z, \hat{\theta}, x > LOD)$, from which then all samples will be drawn. [Ortega-Villa et al., 2021]. However, Bernardt et al. showed that ignoring the response y when imputing, generally, leads to biased results and incorrect inference. In order to condition on the response, one can produce an initial estimate $\hat{\beta}$ to obtain the conditional distribution $f(x|z, \hat{\theta}, y, \hat{\beta})$ [Bernhardt et al., 2015]. Since it is known that all missing values are on one side of the limit of detection, one can sample from the probability distribution of the covariate of interest and limit it to the non-detectable ranges. In case of right censoring the new imputed values are sampled from $f(x|z, \hat{\theta}, y, \hat{\beta}, x > LOD)$.

4 Simulation study

To compare all methods discussed in section 3 we perform a simulation study. The simulation is inspired by a research by van Hooff et al. [van Hooff et al., 2022] on sport-related flow limitations in the iliac artery (FLIA). This study focuses on diagnosing FLIA in endurance athletes based on new near-infrared spectroscopy techniques. In the research project, patients were subjected to an increasing physical workload. After maximal performance the participants were asked to rest and multiple measurements were performed to asses the health of the arteries. In the current study we consider their logistic model with a limited number of covariates, of which one subjected to an upper limit of detection (right censored) and investigate the performance of the different methods.

4.1 Simulation variables

This simulation is performed in the programming language R [R Core Team, 2019]. The simulation consist of $m_s = 1000$ studies each containing $n = 300$ participants. Moreover, a test dataset is created containing 1000 participants to assess the out-of-sample performance of the method applied to each of the training sets.

Let $Y \in \{0, 1\}$ denote the binary dependent variable denoting healthy participants ($Y = 0$) and patients suffering from FLIA ($Y = 1$). Two regressor variables are used in the simulation to model Y. Firstly, the mean response time (MRT), which is a time variable in the study. The MRT variable is the only variable subjected to LOD. Secondly, a indicator variable S to indicate the sex of the participant(S=0 indicates a male). One of the interest points of this simulation is to measure the performance of the methods under multiple censoring rates. Three scenarios are created for which a different censoring rate is applied, namely where 25%, 50% and 75% of the data is missing. For multiple imputation 5 new data sets are formed for all censoring proportions.

4.1.1 Distributions

Three different distributions for the variable MRT are analyzed to measure the impact of the distribution.
We generate the MRT measurements according to a normal, a gamma and a mixture distribution. More specifically, $\mathcal{N}(70, 20)$ for the normal distribution, $\Gamma(1, \frac{1}{50})$ for the gamma distribution with shape and rate parameter, and the mixture distribution $\frac{15}{17}\Gamma(1, 1/50) + \frac{2}{17}\mathcal{N}(900, 25)$. This mixture distribution is an approximation for the real distribution of the data measured in the original study. The independent variable for sex is generated according to a Bernoulli distribution $S \sim B(p = 0.5)$.

4.2 Simulation model

$$\mathbf{P}[Y = 1|(\text{MRT}, S)] = (1 + e^{-(\beta_0 + \beta_{\text{MRT}}\text{MRT} + \beta_S S)})^{-1} \tag{11}$$

The intercept β_0, the MRT coefficient β_{MRT} and the female sex coefficient β_S are given by
$\beta_0 = -8.38666$
$\beta_{\text{MRT}} = 0.15549$
$\beta_S = 6.55573$

The values of these coefficients are taken from the original study [van Hooff et al., 2022]. Once the covariates MRT and S are generated for each participant according to the distribution discussed in section 4.1.1, the dependent variable Y_i can be simulated from a Bernoulli distribution with probability

$$p = \mathbf{P}[Y_i = 1|(\text{MRT}_i, S_i)] \tag{12}$$

Now, each study m in the simulation has a complete data set $(\mathbf{y}_m, \mathbf{x}_m)$ where, $\mathbf{y}_m = [\mathbf{y}_{m,1}, \ldots, \mathbf{y}_{m,n}]$ and $\mathbf{y}_{m,i}$ is the dependent variable associated with patient i in study m. Moreover, $\mathbf{x}_m = [\mathbf{x}_{m,1}, \ldots, \mathbf{x}_{m,n}]$ denote a vector of covariates of all patients in study m. Covariates of patient i in study m are given by $\mathbf{x}_{m,i} = [\text{MRT}_i, S_i]$. The censoring of adequate proportion is applied to the MRT variable to create a right censored data set with missing data. Since the censoring is based on quantiles, this will create a different limit of detection LOD_m for each study.

4.3 Applying the methods

For each method the incomplete data set $(\mathbf{y}_m, \mathbf{x}_m^{<LOD_m})$ transformed according to the methods explained in section 3 to create a complete data set $(\mathbf{y}_m, \mathbf{x}_m^*)$ with $\mathbf{x}_m^* = (MRT_m^*, S_m)$ adjusted according to the method. With complete data, the $glm()$ function in R is used to fit a logistic regression model and find estimates $(\hat{\beta}_0, \hat{\beta}_{MRT}, \hat{\beta}_S)$ for the coefficients $(\beta_0, \beta_{MRT}, \beta_S)$. However, the extensive use of numerical approximation methods resulted in a few failed attempts to converge to a solution in the simulation. The adjusted MLE and multiple imputation are the methods affected by this problem. If the numerical estimation algorithm did not converge, then the result for that method in the corresponding study was excluded from the final analysis.

4.4 Evaluating different approaches

In order to compare different methods, we use different measures that assess the accuracy with which coefficients are estimated (bias and RMSE), and the predictive accuracy of the methods (sensitivity, specificity, AUC, and Youden index). In the following, we briefly state the definitions used in the present study. Firstly we investigate the coefficient estimates, since we would like to quantify how "good" the estimates for $\beta_0, \beta_{\text{MRT}}$ and β_S are.

For this we study the bias and RMSE and focus on $\hat{\beta}_{\text{MRT}}$, since that is the estimate of most interest. These quantifications can also be applied to the other two estimators $\hat{\beta}_0$ and $\hat{\beta}_S$.

4.4.1 Bias

The bias of an estimator $\hat{\beta}_{\text{MRT}}$ is the difference between the estimate and the actual value, averaged over all the studies. It can be calculated by

$$bias(\hat{\beta}_{\text{MRT}}) = \frac{1}{m}\sum_{i=1}^{m}(\hat{\beta}_{\text{MRT,i}} - \beta_{\text{MRT}}).$$

4.4.2 Root mean square error

A quantification of the variability of the estimates can be given by the root mean square error (RMSE).

$$RMSE(\hat{\beta}_{\text{MRT}}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{\beta}_{\text{MRT,i}} - \beta_{\text{MRT}})^2}$$

.

4.4.3 Predicting test results

The motivation study that inspired this research aims at diagnosing FLIA from modern measurements, from which (among others) MRT is derived. A participant is either positive, that is, affected by the disease ($y_i = 1$), or negative, not affected ($y_i = 0$). A test is devised to predict the presence of FLIA for a person. This prediction $y_{p,i}$ can come out positive ($y_{p,i} = 1$) or negative ($y_{p,i} = 0$). Let $\mathbf{x}_i, y_i$ denote the covariates and the outcome for person i part of the test set. Based on $\mathbf{x}_i$ a prediction $y_{p,i}$ is made for each test participant i in the test set. The prediction that will be used in the simulation study later is of the form

$$\begin{aligned} y_{p,i} = 0 \quad &\text{if} \quad \mathbf{x}_i\hat{\beta} \leq \ln(\frac{\alpha}{1-\alpha}) \quad \text{i.e.} \quad (\mathbf{P}_{\hat{\beta}}[y_i = 1|\mathbf{x}_i] \leq \alpha) \\ y_{p,i} = 1 \quad &\text{if} \quad \mathbf{x}_i\hat{\beta} > ln(\frac{\alpha}{1-\alpha}) \quad \text{i.e.} \quad (\mathbf{P}_{\hat{\beta}}[y_i = 1|\mathbf{x}_i] > \alpha) \end{aligned} \tag{13}$$

While this works for any value of $\alpha \in (0, 1)$ we will use a value of $\alpha = \frac{1}{2}$ in the simulation. Now, the predicted values $\hat{y}_i$ and the actual test value $y_{p,i}$ can be compared to measure the predictive performance. All four outcomes for an individual prediction are visualized in table 1. Many statistics can be derived from these results. The ones that we will be using in the simulation study evaluation are:

	$y_{p,i} = 0$	$y_{p,i} = 1$
$y_i = 0$	True Negative (TN)	False Positive (FP)
$y_i = 1$	False Negative (FN)	True Positive (TP)

Table 1: Confusion matrix for predictions.

$$\text{sensitivity} = \frac{TP_{total}}{TP_{total} + FN_{total}}$$
$$\text{specificity} = \frac{TN_{total}}{TN_{total} + FP_{total}}$$

Since sensitivity and specificity are influenced by the data set from which it is calculated these values may vary from case to case. Now we will investigate a theoretical value which will be independent of the specific data set. Let $X \in \Omega$ be a vector of continuous covariates that influence the probability of having the disease according to a logistic relation, i.e.

$$\mathbb{P}_\beta[Y = 1|X = x] = \frac{1}{1 + e^{-\beta x}} \tag{14}$$

Moreover, we assume X follows a certain p.d.f. $f(x)$. Let $\Omega_1^\alpha \subset \Omega$ be defined by $\Omega_1^\alpha = \{x \mid \alpha, Y_p = 1\}$, the space of covariates which result in a positive prediction. Now, the sensitivity of the prediction can be calculated as the conditional probability of a positive test result given the person is positive for the disease, i.e.

$$\text{sensitivity} = \mathbb{P}[Y_p = 1|Y = 1] = \frac{\mathbb{P}[Y_p = 1 \cap Y = 1]}{\mathbb{P}[Y = 1]} \tag{15}$$

The probability of a person being positive for the illness is calculated by

$$\mathbb{P}[Y = 1] = \int_\Omega f(x)\mathbb{P}_\beta[Y = 1|X = x]dx \tag{16}$$

The joint probability of having a positive test result and having the disease is given by

$$\mathbb{P}[Y_p = 1 \cap Y = 1] = \int_{\Omega_1^\alpha} f(x)\mathbb{P}_\beta[Y = 1|X = x]dx \tag{17}$$

This results in the sensitivity:

$$\text{sensitivity} = \frac{\int_{\Omega_1^\alpha} f(x)\mathbb{P}_\beta[Y = 1|X = x]dx}{\int_\Omega f(x)\mathbb{P}_\beta[Y = 1|X = x]dx} \tag{18}$$

The same can be done for the specificity of the prediction, namely,

$$\text{specificity} = \mathbb{P}[Y_p = 0|Y = 0] = \frac{\mathbb{P}[Y_p = 0 \cap Y = 0]}{\mathbb{P}[Y = 0]} \tag{19}$$

$$\text{specificity} = \frac{\int_{\Omega \setminus \Omega_1^\alpha} f(x)\mathbb{P}_\beta[Y=0|X=x]dx}{\int_\Omega f(x)\mathbb{P}_\beta[Y=0|X=x]dx} \tag{20}$$

Note that these calculations require the use of value β which, in practice, is unknown. Sensitivity and specificity are therefore a function of the (estimated) parameters beta's.

These are the calculations for theoretical sensitivity and specificity values for an uncensored data set. Since in our simulations we will deal with censored data we will have to adequately adjust these calculations. Let x^* denote the censored variable and z denote the other covariates, then for methods that are not able to predict the response of censored covariates the sensitivity and specificity are calculated as follows:

$$\text{sensitivity} = \frac{\int_{\Omega_1^\alpha} f^*(x)\mathbb{P}_\beta[Y=1|X=x]dx}{\int_\Omega f^*(x)\mathbb{P}_\beta[Y=1|X=x]dx} \tag{21}$$

and

$$\text{specificity} = \frac{\int_{\Omega \setminus \Omega_1^\alpha} f^*(x)\mathbb{P}_\beta[Y=0|X=x]dx}{\int_\Omega f^*(x)\mathbb{P}_\beta[Y=0|X=x]dx} \tag{22}$$

in which $f^*(x)$ is the rescaled version of $f(x)$ with its domain in the uncensored values, given by:

$$f^*(x) = \frac{f(x)\mathbb{1}\{x^* < LOD\}}{\int_\Omega f(x)\mathbb{1}\{x^* < LOD\}dx} \tag{23}$$

For substitution-based methods that impute a single value s we get:

$$\text{sensitivity} = \frac{\int_{\Omega_1^{\alpha *}} f(x)\mathbb{1}\{x^*_{\text{crit}} < s\}\mathbb{P}_\beta[Y=1|X=x]dx}{\int_\Omega f(x)\mathbb{P}_\beta[Y=1|X=x]dx} \tag{24}$$

with $x^*_{\text{crit}} = \frac{1}{\beta_{\text{MRT}}}(\ln(\frac{\alpha}{1-\alpha}) - z\beta_z)$ and
$\Omega_1^{\alpha *} = \{x | x \in \Omega_1^\alpha, x^* > min(x^*_{\text{crit}}, LOD)\}$.

$$\text{specificity} = \frac{\int_{\Omega \setminus \Omega_1^{\alpha *}} f(x)\mathbb{1}\{x^*_{\text{crit}} < s\}\mathbb{P}_\beta[Y=0|X=x]dx}{\int_\Omega f(x)\mathbb{P}_\beta[Y=0|X=x]dx} \tag{25}$$

4.5 ROC and AUC

The value of α has influence on the prediction and therefore the sensitivity and specificity can be seen as functions of α. Plotting the sensitivity and specificity for each $0 \leq \alpha \leq 1$ the ROC curve is created, from which the area under the curve can be calculated. The AUC gives a quantification of the performance of the prediction across all levels of α. It should be noted that the method with the largest AUC does not have to perform the best for a given value of α. Moreover, optimization methods can be performed on the base of different criteria

(optimizing either sensitivity or specificity, or a combination of both depending on the specific use case).

The AUC quantifies overall performance of the methods across all values of α and it can be used to compare methods. On top of that, to find a best α that leads to the maximum sum of sensitivity and specificity we will consider a variation of Youden's J statistic, which is simply;

$$max\{\text{sensitivity+specificity}\} \tag{26}$$

Note that this value lies between 1 and 2, since predictions with Youden statistic smaller than 1 can be changed by inverting all predictions to get a value between 1 and 2. Values near 1 indicate a poor performance while values near 2 indicate great predictions since there are few false positives and false negatives.

5 Results

In this section we will use the the following abreviations for the methods: CC (Complete Case), adjusted MLE, Sub-1, Sub-2, Sub-$\sqrt{2}$ for the substitution method using 1,2 and $\sqrt{2}$ respectively, Sub-Expected for substitution with $\mathbf{E}[x|x > LOD]$. MDI for the Missing Data Indicator method, CS for multiple imputation using conditional sampling, i.e, sampling from the conditional distribution function $f(x|z, \hat{\theta}, x > LOD)$ and MI for multiple imputation sampling from $f(x|y, z, \hat{\theta}, x > LOD)$.

5.1 25% censoring

For the lowest censoring rate, namely 25%, all methods generally scored equivalently for all distributions. Especially for the normal distribution all methods performed similarly with a slight advantage for Sub-$\sqrt{2}$ which has the lowest bias. The adjusted MLE performed slightly worse with the highest RMSE and higher bias. For the gamma distribution, most substitution methods had slightly larger RMSE with Sub-1 having both the largest bias and RMSE of all methods. These results are shown in table 2 and figure 1. For the mixture distribution the only notable result was that the adjusted MLE method had a lower RMSE and slightly larger bias than other methods, which all scored equivalent. Moreover, the adjusted MLE algorithm did not always converge for all 1000 studies. For the mixture distribution there were 11 cases in which the algorithm did not converge and for the gamma distribution there were 13 cases. Since multiple imputation relies on the adjusted MLE method, also the MI results are missing in same proportions. These incomplete results are not used in the analysis of the simulation. For all distributions it holds that CC, adjusted MLE and MI score score lower on AUC and the Youden statistic compared to the other methods. Since these predictions are only applied to the uncensored test set, they do not cover the cases with really high MRT values. Methods

that do predict censored values score better on prediction than the other methods, because only participants with really high MRT are censored which are easier to predict. One should be careful comparing CC, adjusted MLE and MI to the other methods since they are applied to different test sets. Generally these methods should only be compared to one another. Between methods that predict censored values there is no large difference found in performance.

	Method	Relative bias	RMSE	Sensitivity	Specificity	AUC	Youden
1	CC	4.4%	0.0283	0.8170*	0.9129*	0.9528*	1.7589*
2	adjusted MLE	7.8%	0.0216	0.8162*	0.9136*	0.9529*	1.7591*
3	Sub-1	17.4%	0.0381	0.9057	0.9080	0.9714	1.8277
4	Sub-2	-8.0%	0.0376	0.8981	0.9081	0.9717	1.8231
5	Sub-$\sqrt{2}$	-1.4%	0.0276	0.8985	0.9103	0.9725	1.8241
6	Sub-Expected	-6.2%	0.0351	0.8983	0.9085	0.9719	1.8235
7	MDI	4.5%	0.0283	0.8998	0.9103	0.9723	1.8250
8	CS	-1.0%	0.0311	0.9003	0.9087	0.9720	1.8243
9	MI	4.4%	0.0266	0.8510*	0.9029*	0.9529*	1.7595*

Table 2: Results for gamma distribution with 25% censoring. *These values are determined on the complete data set.

5.2 50% censoring

When half of the data is censored more differences are present in the results. For all distributions the methods have a larger RMSE than their 25% censoring counterpart. For the normal distribution, the Sub-1 and Sub-2 methods perform the worst out of all methods with the largest bias and RMSE. In a lesser degree, Sub-$\sqrt{2}$ is also biased with a larger variance. While these methods are biased, the prediction statistics are still equivalent to that of the other methods. All other methods have a relatively small bias and comparable RMSE, which also results in equivalent predictive performance. For the gamma distribution the Sub-1 method also performs poorly with a large bias and RMSE. However, for this distribution the CS approach actually performs the worst in terms of bias, RMSE and AUC. Besides CC and adj.MLE the CS method has a lower sensitivity and specificity compared to the other methods. The prediction performance for one of the studies is visualized in figure 2. The graph shows that the substitution-based methods and MDI behave very similar for predicting test results. It should be noted however, that this specific graph only visualizes the performance for just one study, and while they all differ for each study, multiple ROC graphs have been checked and the general form on the graph remains the same. All other performance statistics are shown in table 3. The results for the mixture distribution are relatively worse compared to the gamma distribution since, generally for all methods, the bias and RMSE increased, while the predictive power decreased. Still, as was the case with the gamma distribution, Sub-1, Sub-$\sqrt{2}$ and CS perform the worst. The CC, adjusted MLE, MDI and

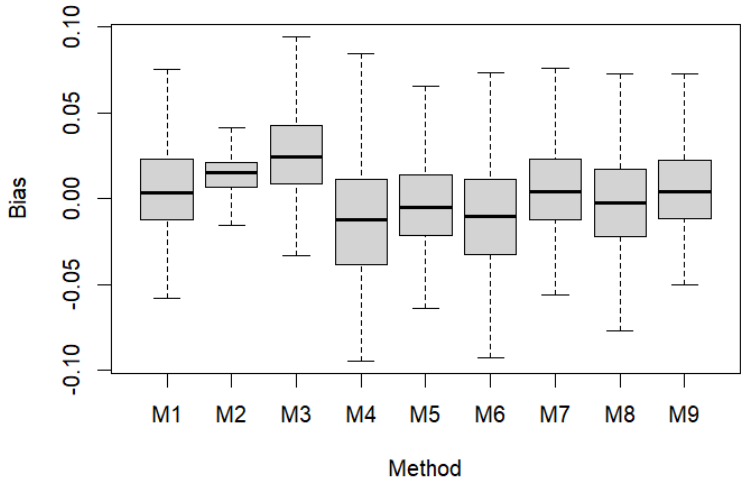

Figure 1: Boxplot for gamma distribution with 25% censoring. M1=CC, M2=MLE, M3=Sub-1, M4=Sub-2, M5=Sub-$\sqrt{2}$, M6=Sub-Expected, M7= MDI, M8= Conditional Sample, M9= Multiple Imputation

MI perform all relatively well with similar bias. The adjusted MLE method had a slight advantage with the lowest RMSE of these methods.

5.3 75% censoring

All methods perform worse with 75% missing data compared to 50% missing data. For the normal distribution, Sub-1 and Sub-2 are both severly biased (relative bias of 127% and 64% respectively) while adjusted MLE and multiple imputation were the least biased. Also, in terms of RMSE Sub-1 scores poorly while adjusted MLE and MI perform the best. Remarkably, Sub-$\sqrt{2}$ performs very well under these circumstances. A possible explanation could be that the imputation value for these circumstances ($\sqrt{2} \cdot LOD$) is very close to the expected value of the censored value ($\mathbb{E}[MRT|MRT > LOD]$), namely 79.84 versus 78.55, respectively on average across all studies. Still, all methods score equivalently on prediction and AUC. Under gamma distribution the bias did not significantly increase for CC, adjusted MLE, MDI, CS compared to the 50% case. The adjusted MLE performs well with the smallest RMSE and low bias. CC en MDI have minimal bias but have large variance in estimation. In around 60 studies it even occurred that both methods resulted in a negative β_{MRT} estimate, which is rather counterintuitive for the case study at hand. This can also be seen in figure 3, visualizing the distributions of β_{MRT}. Sub-2 is comparable to adjusted MLE in terms of RMSE but has a slightly larger bias.

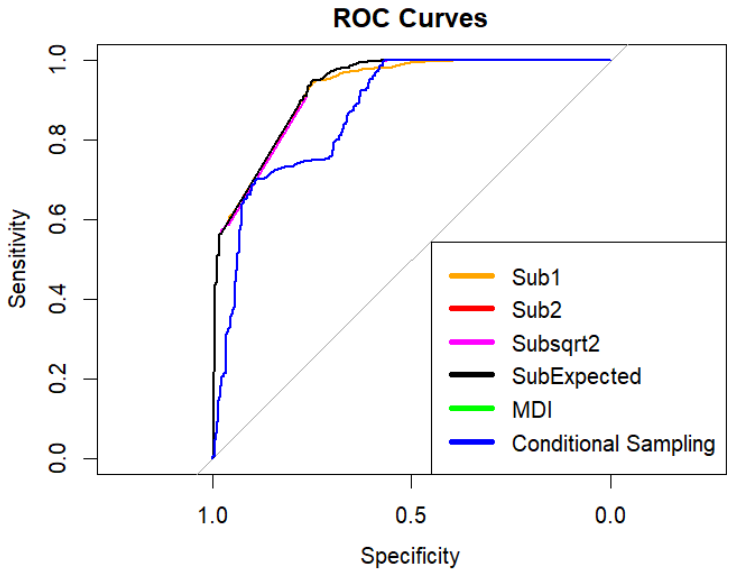

Figure 2: ROC curve for one studie with 50% censoring and a gamma distribution.

Again, Sub-1 and Sub-$\sqrt{2}$ performed the worst (relative bias of 185% and 83% respectively). AUC scores for all methods were comparable outside of CS which scored lower. Results are summarized in table 4. Generally, the same is found when applying the mixture distribution. MDI and MI perform slight better for this distribution.

Lastly, we can verify if our theoretical predicted sensitivity and specificity from section 4.4.3 is in agreement which the simulation results. Note that there was no theoretical sensitivity or specificity for the CS method and is therefore not included in these results. For 25% and 50% censoring the average difference between the theoretical values and simulated values is smaller than 0.03 with most values around 0.01. For 75% censoring and a gamma or mixture distribution the differences become larger for CC, adjusted MLE and MI with values smaller than 0.07. Substitution methods score still around < 0.02 with MDI slightly larger. For normal distribution the difference in values are smaller, this case is visualized in table 5. Since the test set is the same in each study there might be an additional slight bias in the averaged values.

	Method	Relative bias	RMSE	Sensitivity	Specificity	AUC	Youden
1	CC	5.6%	0.0421	0.7766*	0.9290*	0.9526*	1.7705*
2	adjusted MLE	9.1%	0.0272	0.7749*	0.9304*	0.9528*	1.7708*
3	Sub-1	56.1%	0.0962	0.9294	0.7609	0.9304	1.7160
4	Sub-2	-8.0%	0.0386	0.9411	0.7728	0.9340	1.7291
5	Sub-$\sqrt{2}$	19.7%	0.0431	0.9388	0.7765	0.9351	1.7289
6	Sub-Expected	-19.6%	0.0661	0.9468	0.7546	0.9318	1.7291
7	MDI	3.4%	0.0397	0.9397	0.7746	0.9345	1.7291
8	CS	-70.0%	0.1131	0.8338	0.7393	0.8972	1.6574
9	MI	8.3%	0.0324	0.7807*	0.9160*	0.9528*	1.7709*

Table 3: Results for gamma distribution with 50% censoring. *These values are determined on the complete data set.

	Method	Relative bias	RMSE	Sensitivity	Specificity	AUC	Youden
1	CC	5.5%	0.1139	0.3007*	0.9458*	0.8719*	1.6560*
2	adjusted MLE	9.3%	0.0373	0.3198*	0.9492*	0.8815*	1.6601*
3	Sub-1	185%	0.3065	0.6700	0.8636	0.8529	1.5717
4	Sub-2	15.7%	0.0385	0.6824	0.8478	0.8562	1.5687
5	Sub-$\sqrt{2}$	83.0%	0.1394	0.6894	0.8383	0.8559	1.5712
6	Sub-Expected	-35.4%	0.1038	0.6716	0.8554	0.8531	1.5558
7	MDI	-4.7%	0.0973	0.6774	0.8507	0.8546	1.5656
8	CS	-76.8%	0.1257	0.7462	0.6707	0.8209	1.5264
9	MI	-36.7%	0.0749	0.6339*	0.8107*	0.8815*	1.6601*

Table 4: Results for gamma distribution with 75% censoring. *These values are determined on the complete data set.

	Method	Sensitivity	Specificity
1	CC	0.0229*	0.0349*
2	adjusted MLE	0.0244*	0.0340*
3	Sub-1	0.0027	0.0129
4	Sub-2	0.0044	0.0185
5	Sub-2	0.0048	0.0194
6	Sub-Expected	0.0047	0.0169
7	MDI	0.0057	0.0154
8	CS	NA	NA
9	MI	0.0425*	0.0451*

Table 5: Averaged difference in theoretical sensitivity and specificity compared to the values obtained from the simulation. *These values are determined on the complete data set.

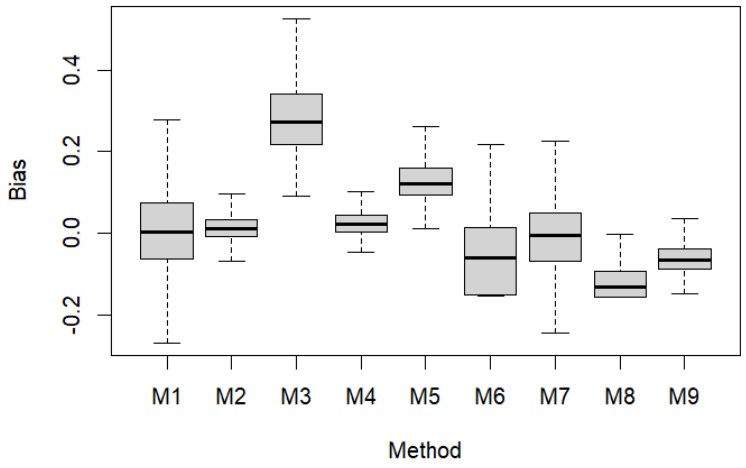

Figure 3: Boxplot for gamma distribution with 75% censoring. M1=CC, M2=MLE, M3=Sub-1, M4=Sub-2, M5=Sub-$\sqrt{2}$, M6=Sub-Expected, M7= MDI, M8= Conditional Sample, M9= Multiple Imputation

6 Conclusion

For the lowest censoring rate (25%) the results showed only minor differences between all methods independent of the applied distribution. Therefore, any of the tested methods can be used to account for censored independent variables in logistic regression given the proportion of missing data is small. However, in practice, the CC method may be preferred because of its simplicity. It does not require any parametric assumptions and is not computationally expensive compared to some of the other methods. One of the drawbacks is its inability to predict the response variable corresponding to censored covariates. For the methods that are able to predict those cases, MDI has the same practical advantages as the CC method and therefore could be recommended as the method of choice when prediction is desired.

For situations in which half of the data is censored the CC method still performs relatively well, but the adjusted MLE has the smallest RMSE across all distributions, with comparable bias to CC. The adjusted MLE outperforms other approaches but comes with multiple weaknesses, since this approach requires parametric assumptions, is more challenging to implement, computationally more expensive and is unable to predict response variables for the censored observations. If distributional assumptions can be made, then this method can be used to obtain the most accurate coefficient estimates. If the distribution is unknown or it is desired to predict censored cases, then the MDI method is suitable since it produces only slightly worse estimations but without all the drawbacks that come with the adjusted MLE.
In the case with heavy censoring (75%) an advantage can be seen for the adjusted MLE and MI approaches which both have a relatively low RMSE but a slight larger bias compared to CC and MDI for all distributions. They are preferred over CC and MDI because of their increased precision in estimation. If predictions need to be made, MDI would be advised for gamma and mixture distributions and Sub-Expected would be recommended for a normal distribution. Based on the results for prediction, the CS method is not recommended for prediction since it performs worse on AUC and Youden's J statistic for moderate and high censoring proportions. Besides CS, most methods perform equivalent and the only big difference is the ability of the method to predict censored cases. A broad overview for all methods is given in table 6.

These results are also partially supported by the literature in which, under low censoring proportions, CC, adjusted MLE, MI produced minimal bias for Bernhardt et al. [Bernhardt et al., 2015] and CC, MDI, were nearly unbiased for Ortega-Villa et al. and Chiou et al. [Ortega-Villa et al., 2021] [Chiou et al., 2019]. Moreover, they found that substitution methods were biased, even for slight censoring. In our case only Sub-1 has an increased bias compared to other methods. This increase in bias becomes more clear with increased censoring as was also observed by Ortega-Villa et al. The equivalence between CC and MDI is also present in our results, since bias and RMSE of the two methods are

Method	Advantages	Disadvantages	Can predict censored cases?
CC	Simple to implement Nonparametric Performs well for low and moderate censoring proportions	Inefficient Needs adequate sample size Less effective when large proportion is censored	No
adjusted MLE	Performs well across all settings especially for much missing data	Computationally intensive Requires good initial parameter estimates for convergence algorithm Complicated to implement Parametric method	No
Sub-1 Sub-2 Sub-$\sqrt{2}$	Simple to implement Nonparametric	Biased Performed poorly for 50% and 75% censoring	Yes
Sub-Expected	Performs moderately across all settings	Parametric method	Yes
MDI	Relatively simple to implement Nonparametric Performs well across all settings	Outperformed by other methods at high proportion censoring	Yes
CS	Performs well for low censoring proportions	Performed poorly for 50% and 75% censoring moderately complicated to implement especially with complex distributions Parametric method	Yes
MI	Performs well across all settings especially for much missing data	Computationally intensive Dependent on response Complicated to implement Time consuming for large sample sizes Parametric method	No

Table 6: Overview of advantages and disadvantages of the tested methods.

mostly equal with a slight advantage for MDI across all settings. Moreover, for large censoring proportions the results agree with Bernhardt et al. that CC is not the most suitable and better methods such as adjusted MLE or multiple imputation exist. The result that non-normal distributions result in better estimates, as found in Chiou et al., can only be partially supported by this data, in which CC, and MDI score better under these circumstances but this does not hold for other methods.

6.1 Discussion

In this paper we focused on a logistic regression model when one independent variable is subjected to a detection limit. The most commonly used methods from the literature have been compared using a simulation study with a simple regression model. This study investigated the influence of an upper detection limit in contrary to existing literature which mainly focuses on left censoring. In a simulation study the methods were tested under different circumstances, namely, varying the distribution of the censored variable and using multiple censoring proportions. Methods were compared by means of bias and RMSE for the regression coefficient of censored variable. Moreover, comparisons were also made on their performance in predicting the response variable. This was done by analyzing sensitivity, specificity, ROC curves, AUC, Youden J statistic on a

test set. Generally, for small censoring proportions ($< 25\%$) the differences between methods are minimal. Because of its simplicity the complete case method is recommended in this situation. For 50% and 75% censoring proportions the adjusted MLE method was best suitable for producing the best estimate for the regression parameters. Since this method is not able to predict censored data, MDI is advised if predictions are desired or if not all modelling assumptions can be met. Methods that are able to predict the response of participants with missing covariates did not show a notable difference on prediction statistics.

However, we suspect that some methods could perform better under circumstances different from this simulation study. Parametric methods that were dependent on only distributional parameter estimates (CS, Sub-Expected) may have been disproportional negatively affected by poor estimates. If the training set had a larger sample size, the quantity of these poor estimates would be reduced which would result in better performance. In addition, the parametric methods (CS, MI, Sub-Expected) which are conditioned on the other covariates may benefit from an extended regression model with more than two covariates to obtain a more reasonable imputation value. In this study a simple logistic regression model was used, but one could examine the performance under a different model with more (censored) variables. Since there is, to our knowledge, no study that also examined prediction performance, additional research should be performed in this area, possibly including different measurement statistics to evaluate their performance. Furthermore, future research could explore more involved distributions other than the normal and gamma distributions.

References

[Alyabs and Chiou, 2022] Alyabs, N. and Chiou, S. H. (2022). The Missing Indicator Approach for Accelerated Failure Time Model with Covariates Subject to Limits of Detection. *Stats*, 5(2):494–506.

[Bernhardt et al., 2015] Bernhardt, P. W., Wang, H. J., and Zhang, D. (2015). Statistical Methods for Generalized Linear Models with Covariates Subject to Detection Limits. *Statistics in Biosciences*, 7(1):68–89.

[Chiou et al., 2019] Chiou, S. H., Betensky, R. A., and Balasubramanian, R. (2019). The missing indicator approach for censored covariates subject to limit of detection in logistic regression models. *Annals of Epidemiology*, 38:57–64.

[Fu et al., 2016] Fu, P., Hughes, J., Zeng, G., Hanook, S., Orem, J., Mwanda, O. W., and Remick, S. C. (2016). A comparative investigation of methods for longitudinal data with limits of detection through a case study. *Statistical Methods in Medical Research*, 25(1):153–166.

[Lubin et al., 2004] Lubin, J. H., Colt, J. S., Camann, D., Davis, S., Cerhan, J. R., Severson, R. K., Bernstein, L., and Hartge, P. (2004). Epidemiologic

evaluation of measurement data in the presence of detection limits. *Environmental Health Perspectives*, 112(17):1691–1696.

[McNamara et al., 2018] McNamara, K., Peters, C., and Burstyn, I. (2018). Forecasting dose from unobserved times: Case study of transient workers at a nuclear power plant. *Annals of Work Exposures and Health*, 62(7):808–817.

[Mwanda et al., 2009] Mwanda, W. O., Orem, J., Fu, P., Banura, C., Kakembo, J., Onyango, C. A., Ness, A., Reynolds, S., Johnson, J. L., Subbiah, V., Bako, J., Wabinga, H., Abdallah, F. K., Meyerson, H. J., Whalen, C. C., Lederman, M. M., Black, J., Ayers, L. W., Katongole-Mbidde, E., and Remick, S. C. (2009). Dose-modified oral chemotherapy in the treatment of AIDS-related non-Hodgkin's lymphoma in East Africa. *Journal of Clinical Oncology*, 27(21):3480–3488.

[Ortega-Villa et al., 2021] Ortega-Villa, A. M., Liu, D., Ward, M. H., and Albert, P. S. (2021). New insights into modeling exposure measurements below the limit of detection. *Environmental Epidemiology*, 5(1).

[R Core Team, 2019] R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Schisterman et al., 2006] Schisterman, E. F., Vexler, A., Whitcomb, B. W., and Liu, A. (2006). The limitations due to exposure detection limits for regression models. *American Journal of Epidemiology*, 163(4):374–383.

[van Hooff et al., 2022] van Hooff, M., Arnold, J., Meijer, E., Schreuder, P., Regis, M., Xu, L., Scheltinga, M., Savelberg, H., and Schep, G. (2022). Diagnosing Sport-Related Flow Limitations in the Iliac Arteries Using Near-Infrared Spectroscopy. *Journal of Clinical Medicine*, 11(24).

[Yuan and Ghosh, 2008] Yuan, Z. and Ghosh, D. (2008). Combining multiple biomarker models in logistic regression. *Biometrics*, 64(2):431–439.

A Appendix

A.1 25% censoring

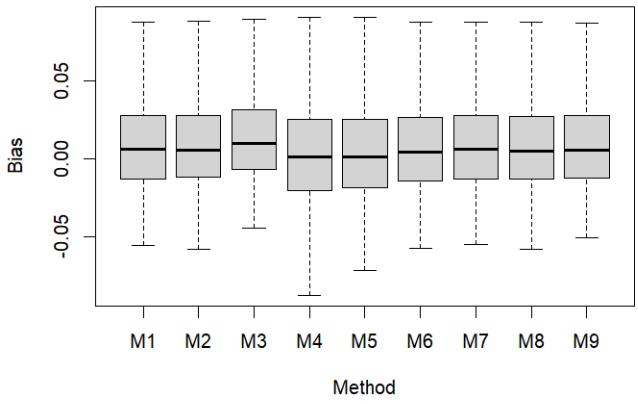

Figure 4: Boxplot for normal distribution with 25% censoring. M1=CC, M2=MLE, M3=Sub-1, M4=Sub-2, M5=Sub-$\sqrt{2}$, M6=Sub-Expected, M7= MDI, M8= Conditional Sample, M9= Multiple Imputation

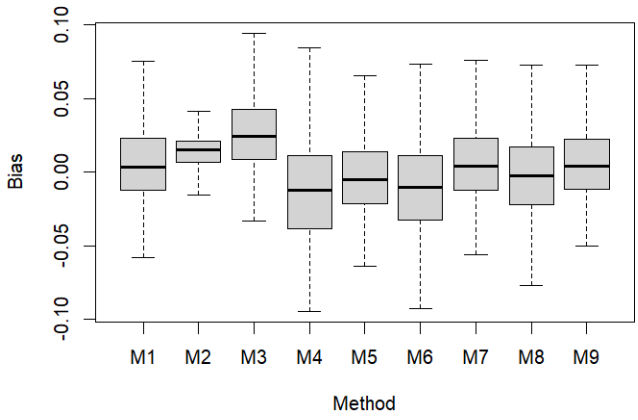

Figure 5: Boxplot for gamma distribution with 25% censoring. M1=CC, M2=MLE, M3=Sub-1, M4=Sub-2, M5=Sub-$\sqrt{2}$, M6=Sub-Expected, M7= MDI, M8= Conditional Sample, M9= Multiple Imputation

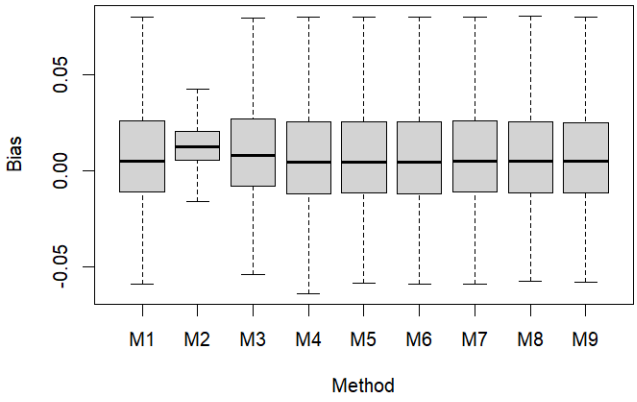

Figure 6: Boxplot for mixture distribution with 25% censoring. M1=CC, M2=MLE, M3=Sub-1, M4=Sub-2, M5=Sub-$\sqrt{2}$, M6=Sub-Expected, M7= MDI, M8= Conditional Sample, M9= Multiple Imputation

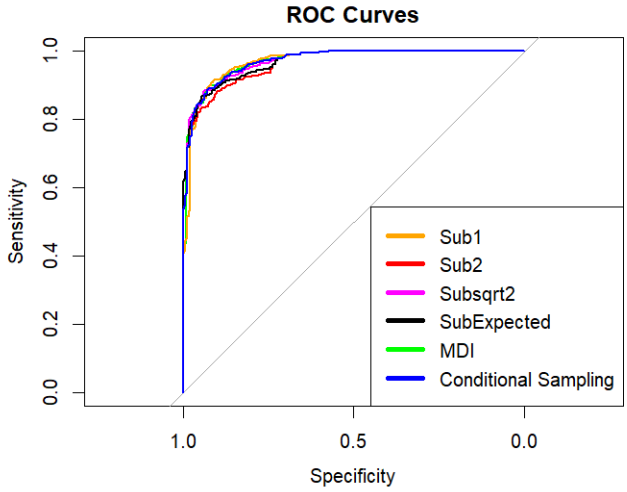

Figure 7: ROC curve for gamma distribution with 25% censoring.

	Method	Bias	RMSE	Sensitivity	Specificity	AUC	Youden
1	CC	0.0106	0.0361	0.9677	0.6552	0.9541	1.7724
2	adjusted MLE	0.0123	0.0432	0.9676	0.6556	0.9540	1.7722
3	Sub-1	0.0150	0.0357	0.9771	0.6550	0.9673	1.8127
4	Sub-2	0.0031	0.0399	0.9777	0.6497	0.9674	1.8134
5	Sub-$\sqrt{2}$	0.0059	0.0367	0.9773	0.6533	0.9674	1.8133
6	Sub-Expected	0.0092	0.0352	0.9771	0.6549	0.9674	1.8131
7	MDI	0.0105	0.0361	0.9770	0.6552	0.9670	1.8130
8	Conditional Sample	0.0102	0.0353	0.9771	0.6548	0.9673	1.8130
9	Multiple Imputation	0.0106	0.0351	0.9630	0.6784	0.9541	1.7724

Table 7: Results for a normal distribution with 25% censoring

	Method	Bias	RMSE	Sensitivity	Specificity	AUC	Youden
1	CC	0.0069	0.0283	0.8170	0.9129	0.9528	1.7589
2	adjusted MLE	0.0122	0.0216	0.8162	0.9136	0.9529	1.7591
3	Sub-1	0.0270	0.0381	0.9057	0.9080	0.9714	1.8277
4	Sub-2	-0.0125	0.0376	0.8981	0.9081	0.9717	1.8231
5	Sub-$\sqrt{2}$	-0.0022	0.0276	0.8985	0.9103	0.9725	1.8241
6	Sub-Expected	-0.0096	0.0351	0.8983	0.9085	0.9719	1.8235
7	MDI	0.0070	0.0283	0.8998	0.9103	0.9723	1.8250
8	Conditional Sample	-0.0015	0.0311	0.9003	0.9087	0.9720	1.8243
9	Multiple Imputation	0.0069	0.0266	0.8510	0.9029	0.9529	1.7595

Table 8: Results for a gamma distribution with 25% censoring

	Method	Bias	RMSE	Sensitivity	Specificity	AUC	Youden
1	CC	0.0091	0.0307	0.8738	0.9292	0.9670	1.8180
2	adjusted MLE	0.0122	0.0193	0.8720	0.9300	0.9669	1.8176
3	Sub-1	0.0117	0.0305	0.9257	0.9290	0.9803	1.8642
4	Sub-2	0.0083	0.0311	0.9251	0.9291	0.9804	1.8638
5	Sub-$\sqrt{2}$	0.0086	0.0307	0.9251	0.9292	0.9804	1.8639
6	Sub-Expected	0.0084	0.0309	0.9251	0.9292	0.9804	1.8639
7	MDI	0.0091	0.0307	0.9252	0.9292	0.9804	1.8640
8	Conditional Sample	0.0090	0.0307	0.9252	0.9292	0.9804	1.8640
9	Multiple Imputation	0.0089	0.0304	0.8873	0.9006	0.9669	1.8180

Table 9: Results for mixture distribution with 25% censoring

A.2 50% censoring

	Method	Bias	RMSE	Sensitivity	Specificity	AUC	Youden
1	CC	0.0135	0.0448	0.9459	0.6901	0.9455	1.7356
2	adjusted MLE	0.0114	0.0377	0.9462	0.6893	0.9456	1.7357
3	Sub-1	0.0462	0.0586	0.9791	0.6424	0.9633	1.8103
4	Sub-2	-0.0598	0.0795	0.9760	0.6483	0.9575	1.8163
5	Sub-$\sqrt{2}$	-0.0276	0.0478	0.9764	0.6572	0.9609	1.8152
6	Sub-Expected	-0.0069	0.0369	0.9767	0.6564	0.9624	1.8139
7	MDI	0.0132	0.0440	0.9769	0.6559	0.9609	1.8124
8	Conditional Sample	0.0004	0.0378	0.9782	0.6474	0.9614	1.8134
9	Multiple Imputation	0.0106	0.0356	0.9389	0.7031	0.9456	1.7358

Table 10: Results for normal distribution with 50% censoring

	Method	Bias	RMSE	Sensitivity	Specificity	AUC	Youden
1	CC	0.0087	0.0421	0.7766	0.9290	0.9526	1.7705
2	adjusted MLE	0.0141	0.0272	0.7749	0.9304	0.9528	1.7708
3	Sub-1	0.0873	0.0962	0.9294	0.7609	0.9304	1.7160
4	Sub-2	-0.0125	0.0386	0.9411	0.7728	0.9340	1.7291
5	Sub-$\sqrt{2}$	0.0306	0.0431	0.9388	0.7765	0.9351	1.7289
6	Sub-Expected	-0.0304	0.0661	0.9468	0.7546	0.9318	1.7291
7	MDI	0.0053	0.0397	0.9397	0.7746	0.9345	1.7291
8	Conditional Sample	-0.1090	0.1131	0.8338	0.7393	0.8972	1.6574
9	Multiple Imputation	0.0129	0.0324	0.7807	0.9160	0.9528	1.7709

Table 11: Results for gamma distribution with 50% censoring

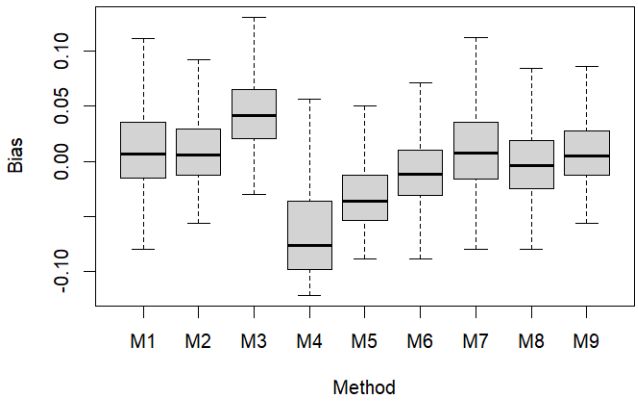

Figure 8: Boxplot for normal distribution with 50% censoring. M1=CC, M2=MLE, M3=Sub-1, M4=Sub-2, M5=Sub-$\sqrt{2}$, M6=Sub-Expected, M7= MDI, M8= Conditional Sample, M9= Multiple Imputation

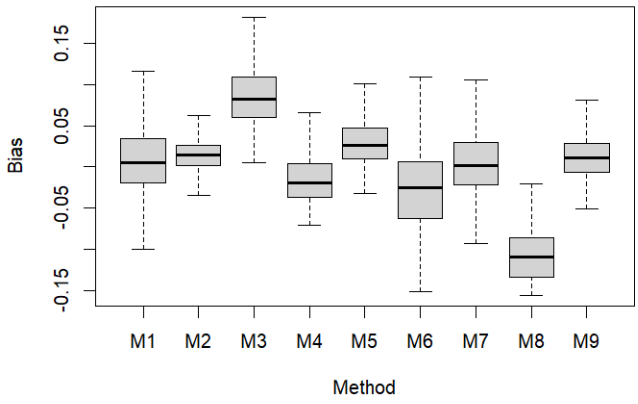

Figure 9: Boxplot for gamma distribution with 50% censoring. M1=CC, M2=MLE, M3=Sub-1, M4=Sub-2, M5=Sub-$\sqrt{2}$, M6=Sub-Expected, M7= MDI, M8= Conditional Sample, M9= Multiple Imputation

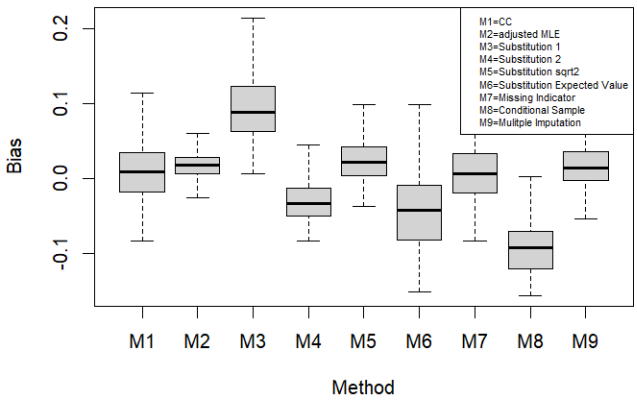

Figure 10: Boxplot for mixture distribution with 50% censoring. M1=CC, M2=MLE, M3=Sub-1, M4=Sub-2, M5=Sub-$\sqrt{2}$, M6=Sub-Expected, M7= MDI, M8= Conditional Sample, M9= Multiple Imputation

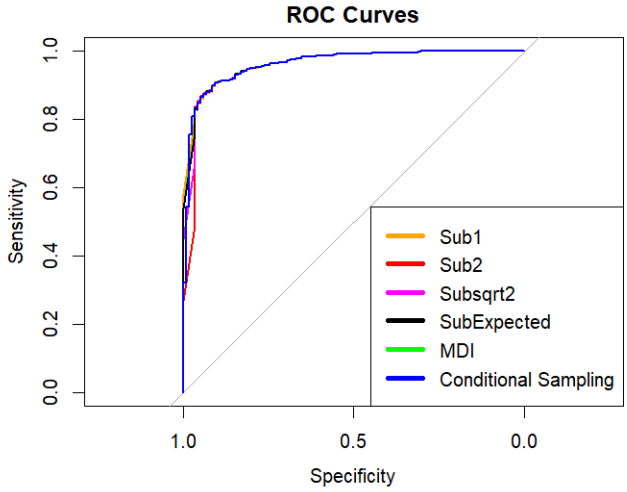

Figure 11: ROC for normal distribution with 50% censoring

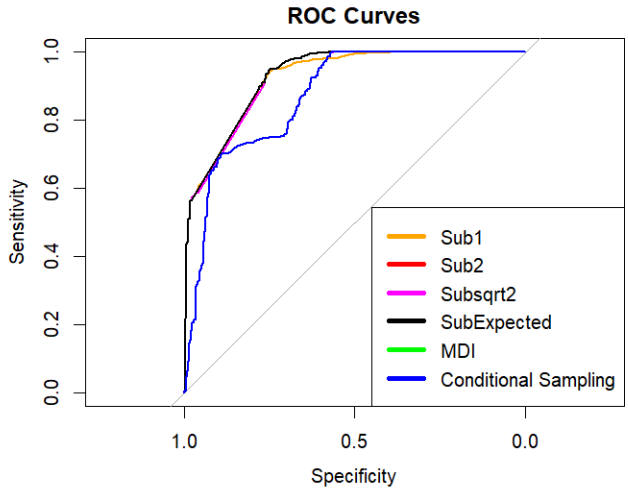

Figure 12: ROC for gamma distribution with 50% censoring

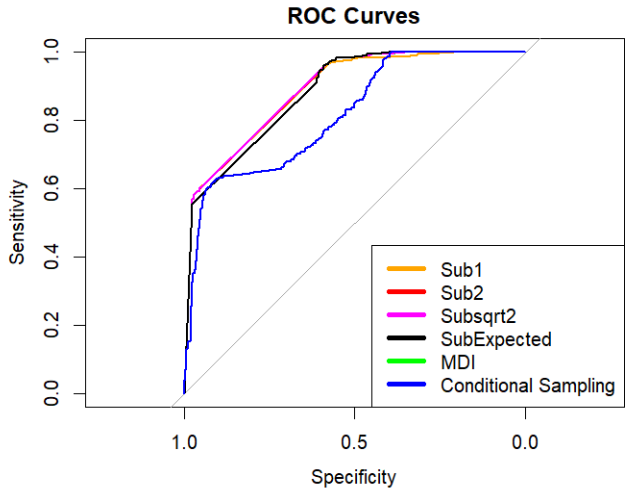

Figure 13: ROC for mixture distribution with 50% censoring

	Method	Bias	RMSE	Sensitivity	Specificity	AUC	Youden
1	CC	0.0133	0.0468	0.8051	0.9437	0.9605	1.8013
2	adjusted MLE	0.0173	0.0268	0.8018	0.9453	0.9603	1.8012
3	Sub-1	0.0983	0.1114	0.9555	0.8324	0.9547	1.8085
4	Sub-2	-0.0259	0.0451	0.9526	0.8518	0.9585	1.8137
5	Sub-$\sqrt{2}$	0.0258	0.0406	0.9494	0.8581	0.9591	1.8137
6	Sub-Expected	-0.0444	0.0725	0.9582	0.8235	0.9569	1.8137
7	MDI	0.0119	0.0455	0.9503	0.8564	0.9591	1.8137
8	Conditional Sample	-0.0946	0.1009	0.8869	0.7891	0.9307	1.7379
9	Multiple Imputation	0.0178	0.0358	0.8089	0.9231	0.9604	1.8012

Table 12: Results for mixture distribution with 50% censoring

A.3 75% censoring

	Method	Bias	RMSE	Sensitivity	Specificity	AUC	Youden
1	CC	0.0466	0.1331	0.8755	0.8961	0.9597	1.8170
2	adjusted MLE	0.0177	0.0504	0.8713	0.9099	0.9601	1.8184
3	Sub-1	0.1986	0.2360	0.9874	0.5518	0.9314	1.6796
4	Sub-2	-0.0994	0.0999	0.9668	0.7100	0.9303	1.6833
5	Sub-$\sqrt{2}$	-0.0424	0.0462	0.9717	0.6854	0.9306	1.6833
6	Sub-Expected	-0.0316	0.0673	0.9736	0.6727	0.9305	1.6830
7	MDI	0.0453	0.1320	0.9785	0.6407	0.9308	1.6821
8	Conditional Sample	-0.0463	0.0727	0.9828	0.5976	0.9275	1.6931
9	Multiple Imputation	0.0177	0.0476	0.8549	0.8841	0.9602	1.8187

Table 13: Results for normal distribution with 75% censoring

	Method	Bias	RMSE	Sensitivity	Specificity	AUC	Youden
1	CC	0.0085	0.1139	0.3007	0.9458	0.8719	1.6560
2	adjusted MLE	0.0145	0.0373	0.3198	0.9492	0.8815	1.6601
3	Sub-1	0.2882	0.3065	0.6700	0.8636	0.8529	1.5717
4	Sub-2	0.0244	0.0385	0.6824	0.8478	0.8562	1.5687
5	Sub-$\sqrt{2}$	0.1295	0.1394	0.6894	0.8383	0.8559	1.5712
6	Sub-Expected	-0.0550	0.1038	0.6716	0.8554	0.8531	1.5558
7	MDI	-0.0073	0.0973	0.6774	0.8507	0.8546	1.5656
8	Conditional Sample	-0.1194	0.1257	0.7462	0.6707	0.8209	1.5264
9	Multiple Imputation	-0.0570	0.0749	0.6339	0.8107	0.8815	1.6601

Table 14: Results for gamma distribution with 75% censoring

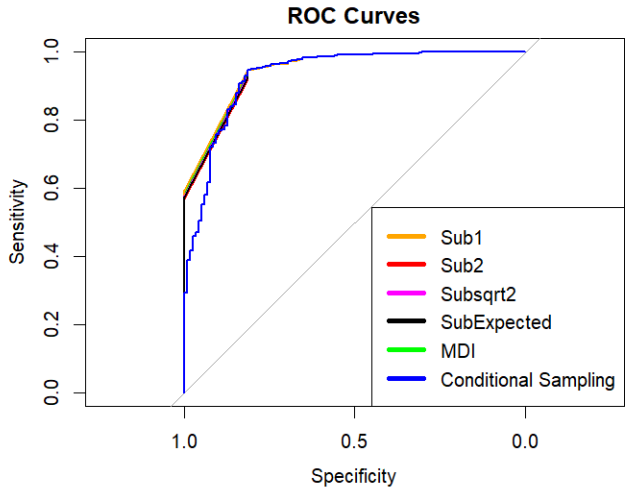

Figure 14: ROC for normal distribution with 75% censoring

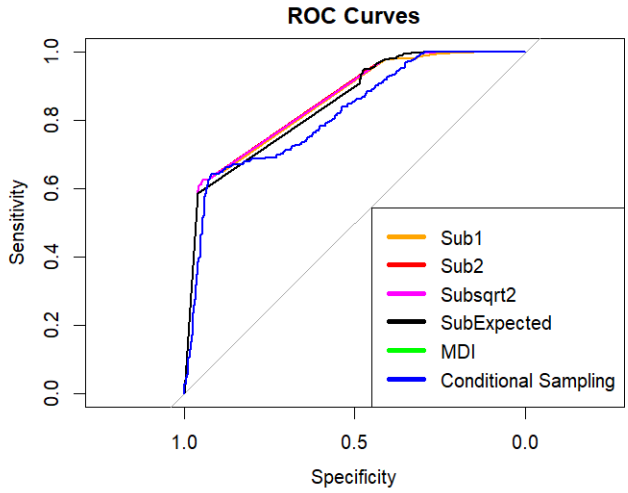

Figure 15: ROC for gamma distribution with 75% censoring

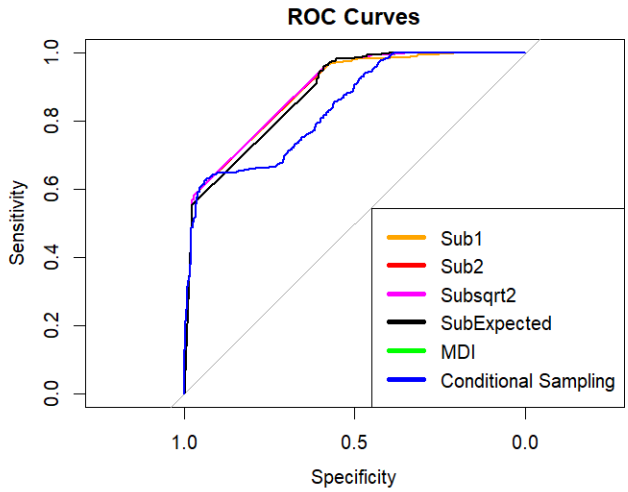

Figure 16: ROC for mixture distribution with 75% censoring

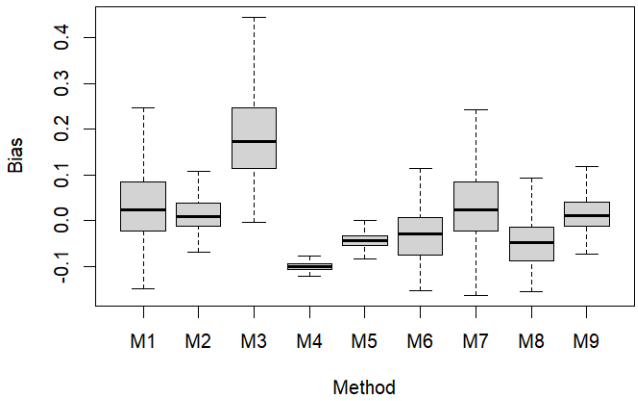

Figure 17: Boxplot for normal distribution with 75% censoring. M1=CC, M2=MLE, M3=Sub-1, M4=Sub-2, M5=Sub-$\sqrt{2}$, M6=Sub-Expected, M7= MDI, M8= Conditional Sample, M9= Multiple Imputation

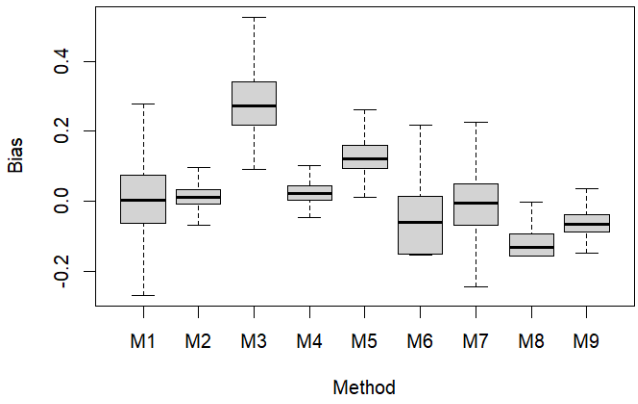

Figure 18: Boxplot for gamma distribution with 75% censoring. M1=CC, M2=MLE, M3=Sub-1, M4=Sub-2, M5=Sub-$\sqrt{2}$, M6=Sub-Expected, M7= MDI, M8= Conditional Sample, M9= Multiple Imputation

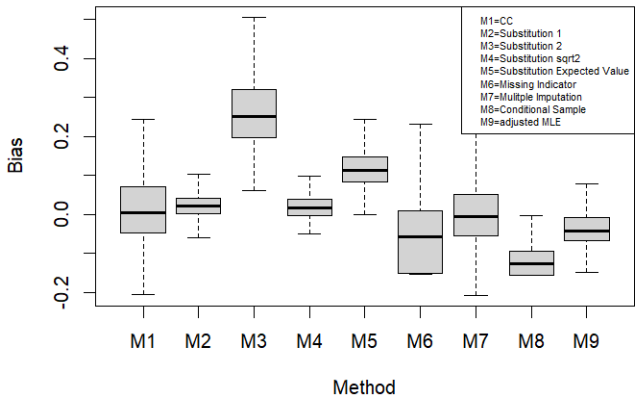

Figure 19: Boxplot for mixture distribution with 75% censoring. M1=CC, M2=MLE, M3=Sub-1, M4=Sub-2, M5=Sub-$\sqrt{2}$, M6=Sub-Expected, M7= MDI, M8= Conditional Sample, M9= Multiple Imputation

	Method	Bias	RMSE	Sensitivity	Specificity	AUC	Youden
1	CC	0.0176	0.0985	0.3453	0.9619	0.8951	1.6431
2	adjusted MLE	0.0237	0.0425	0.3733	0.9613	0.8984	1.6437
3	Sub-1	0.2669	0.2835	0.8828	0.6191	0.8739	1.5584
4	Sub-2	0.0202	0.0378	0.9148	0.5889	0.8764	1.5562
5	Sub-$\sqrt{2}$	0.1185	0.1282	0.9163	0.5882	0.8764	1.5577
6	Sub-Expected	-0.0712	0.1085	0.9050	0.5948	0.8751	1.5531
7	MDI	0.0005	0.0812	0.9115	0.5918	0.8758	1.5551
8	Conditional Sample	-0.1252	0.1303	0.7731	0.6114	0.8196	1.5520
9	Multiple Imputation	-0.0371	0.0598	0.6300	0.8508	0.8984	1.6437

Table 15: Results for mixture distribution with 75% censoring