Open Access Medical Statistics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

METHODOLOGY

# Average effect estimation with dichotomized events when the missing data mechanism is not missing at random

This article was published in the following Dove Press journal:
Open Access Medical Statistics
17 December 2012
Number of times this article has been viewed

Amy M Kwon[1]
Dianxu Ren[2]

[1]Biostatistics and Bioinformatics Core, James Graham Brown Cancer Center, University of Louisville, Louisville, KY, [2]Department of Biostatistics, University of Pittsburgh Center for Research and Evaluation, School of Nursing, University of Pittsburgh, Pittsburgh, PA, USA



**Background:** The purpose of this work was to estimate the average effect of the covariate of interest when the outcome variable is dichotomized from a continuous variable and data are incomplete, with the missing data not missing at random (NMAR). The motivating example is to estimating the effect of vitamin D levels on secondary hyperparathyroidism among patients with chronic kidney disease.

**Methods:** The average effect of the covariate of interest is computed by a two-step procedure. In the first step, we identify the conditional distribution of the original variable given the covariates by obtaining the parameter estimates. In the second step, we draw the predictive values from the identified distribution, and create binary values from the predictive values by dichotomizing them at the threshold.

**Results:** According to the simulation results, the biases of the effects between logistic regression with the complete data and the estimated logistic regression with the converted binary variable are negligible. For the application example, the effect of vitamin D on the occurrence of secondary hyperparathyroidism is highly significant in the complete case analysis, but only a modest effect of vitamin D on secondary hyperparathyroidism is observed under the NMAR assumption.

**Conclusion:** It is impossible to find consistent estimates without knowing the exact nature of the missing data when the missing data mechanism is NMAR. Also, the outcome variable is binary, so we may be faced with an unidentifiability problem when the missing data mechanism is NMAR. To avoid this problem, we estimated the average effect of the covariate of interest in the framework of a generalized linear model from the relationship between a dichotomized outcome and a continuous original outcome, and the estimated effect showed negligible bias according to this simulation.

**Keywords:** average effect, NMAR, not missing at random, dichotomized events, secondary hyperparathyroidism

## Introduction

Biomedical data are noisy and redundant. Because of these characteristics, they are often discretized according to preassigned thresholds to test the inferences of interest. Dichotomization is the simplest form of discretization. Once a certain measurement is discretized, the discretized outcome is treated as a categorical variable, and the original measurement data are generally ignored. Such discretization assigns the same values to subjects who belong to the same interval, so the discretized data look simpler and more convenient to handle, but there are also some drawbacks. Because there is the potential to lose some information about the data, we may lose the power to test for inferences of interest with discretized data. In addition, because the distribution changes, we may

Correspondence: Amy M Kwon
505 S Hancock St, James Graham Brown Cancer Center, Louisville, KY 40218, USA
Tel +1 502 852 1114
Fax +1 502 852 7979
Email amy.kwon@louisville.edu

glimpse on the possible association due to this discretization. Therefore, even though the data are discretized, the original data still have much information which needs to be considered. Data analysis techniques may vary, especially when the data are incomplete, according to the missing data mechanism, which explains the association between the "missingness" and the underlying variables. If the missing values are ignored and the missing data mechanism is not considered, standard statistical techniques may result in severe bias in estimation, depending on the missing data mechanism. For these incomplete data, the original data may still contain more information than discretized data upon which to make assumptions about the missing data mechanism. Also, it may be more natural to make assumptions about the missing data mechanism based on the original data before dichotomization, because the missingness of the discretized data is inherited from that of the original data. Among the categories of the missing data mechanism,[9] when the missing data are not missing at random (NMAR), it is almost impossible to find consistent estimates without knowing the exact nature of the missing data mechanism. A pair-wise form of the pseudo likelihood approach provides asymptotically consistent estimates without specifying the missing data mechanism as long as the missing data mechanism consists of the response variables.[2,13] However, this approach only works when the incomplete data are continuous or multinomial due to the limitation of unidentifiability.[13] Thus, when the data are incomplete, with the NMAR missing data mechanism, and the outcome variable is binary, the choice of statistical techniques to use is limited. There are many biomedical situations in which we are interested in dichotomized outcomes rather than continuous measurements. One of these situations involves the manifestations of secondary hyperparathyroidism in patients with chronic kidney disease. Secondary hyperparathyroidism, which is prevalent in patients with chronic kidney disease, is defined according to whether parathyroid hormone levels exceed a given threshold in a particular disease state. Secondary hyperparathyroidism is of interest because of its association with vitamin D levels, but most researchers use logistic regression to study this association after discarding the missing values.[3,7] This approach is called complete case analysis, but results in severe estimation bias if the missing data mechanism is NMAR. In this paper, we estimate the average effect of vitamin D levels on secondary hyperparathyroidism in patients with chronic kidney disease within a framework of logistic regression assuming the NMAR missing data mechanism. Unlike most other papers, which ignore the original measurements, we first fully identify the underlying distribution of the original data under the NMAR assumption, and stochastically draw the predictive values multiple times and replace the original variable with these predictive values. Next, we convert these values into binary variables by dichotomization according to a given threshold. Those binary variables are treated as complete datasets, and logistic regressions are conducted multiple times to observe the association with the covariate of interest, then the effects of interest are averaged. Before applying this approach, the simulation is conducted with random samples of size 100, 300, 500, and 1000. One hundred replications are performed, and the performance is compared in terms of bias regarding the coefficients.

## Materials and methods

This section describes how to compute the average effect of the covariate of interest when a binary outcome is dichotomized from the original (continuous) variable, and the outcome variable is partially missing under the NMAR missing data mechanism after briefly summarizing notations. The computation is conducted using a two-step procedure. The first step identifies the distribution of the original variable by estimating the consistent parameters of the distribution by pseudo likelihood,[13] and the second step creates multiple binary variables from the predictive values which are drawn from the distribution identified by the first step.

### Notations

Let $X$ and $Y$ be independent and dependent variables. A vector of independent variables for the ith subject is $x_i = \{x_{i1}, \ldots, x_{ip}\}$, and $X$ are always observed. $Y$ is a binary variable, and it is partially missing. Suppose that $Z$ is a continuous variable, and $Y$ is derived from $Z$ by dichotomizing $Z$ according to a given threshold c as (1.1). $Z$ is partially missing and the missing data mechanism is assumed as $m(z)$ where $m(.)$ is an arbitrary function. Therefore, the missing values of $Y$ are inherited from $Z$. Namely, $y_i$ is also missing if $z_i$ is missing for the ith subject, and the number of missing cases between $Y$ and $Z$ are the same.

$$y_i = \begin{cases} 1, & z_i \geq c \\ 0, & z_i < c \end{cases} \tag{1.1}$$

$$r_i = \begin{cases} 1, & z_i \neq MV \\ 0, & z_i = MV \end{cases} \tag{1.2}$$

where $r_i$ is the missing data indicator for the ith subject where $MV$ indicates a missing value. Suppose that we are interested

in the effect of the covariate $x_1$ on $Y$ under the given missing data mechanism. The definition of the missing data mechanism was statistically established by Little and Rubin,[8] and explains the underlying association between the missing data indicator and study variables where the missing data indicator is defined as (1.2). According to Little and Rubin, the missing data mechanism has three categories: one is missing completely at random (MCAR), another is missing at random (MAR), and the other is NMAR. If the missing data mechanism is either MCAR or MAR with regularity conditions, the incomplete data are considered to be ignorable, and we can estimate consistent parameters using the factorization property without specifying the missing data mechanism. Otherwise, the incomplete data are nonignorable, and we cannot find consistent parameter estimates without fully specifying the exact form of the missing data mechanism with nonignorable data. This paper focuses only on nonignorable data, ie, the NMAR missing data mechanism.

## Two-step procedures

### Parameter estimation with original variable

The pair-wise form of the conditionalized pseudo likelihood approach, developed by Tang et al,[13] is applied to incomplete NMAR data to obtain consistent parameters without specifying the missing data mechanism. The pair-wise pseudo likelihood was first introduced by Kalbfleisch[5] and Liang and Qin,[6] and is applied to a nonignorable missing data example later on. They illustrated several nonstandard situations which need specification of the distribution of covariates, and developed a pair-wise form of the pseudo likelihood approach. Subsequently, Chen[2] independently supported the use of the pair-wise approach of Kalbfleisch[5] under the response-biased sampling scheme. Chen also demonstrated the identifiability conditions of the parameters in a regression setting when the probabilities of strata are unknown. He reparameterized the parameter spaces in his proofs, which were similar to the arguments on the same subject in case-control studies reported by Carroll et al.[1] According to Chen's proofs, loss of information is negligible using the pair-wise partial likelihood approach if the probability of being observed can be expressed as a function of responses.[2] Chen's findings are applied to the nonignorable missing data described by Tang et al.[13] Tang et al maintain the pair-wise form, and estimate the regression parameters using a conditionalized pseudo likelihood approach under the assumption that the missing data mechanism consists of response variables with only nonignorable data. This assumption enables us to find asymptotically consistent estimates without specifying the missing data mechanism with nonignorable data, but restricts the identifiability conditions of the parameters to functions of independent variables or of both independent and dependent variables together. Also, if the variable is discrete, the variable needs at least three distinctive values to satisfy the identifiability condition.[13] The identifiability conditions have been described elaborately by Tang et al,[13] Chen,[2] and Carroll et al.[1] Under the given notations, we assume that the conditional distribution of $Z$, given $X$, may be known where $Z$ is continuous. The missing data mechanism is assumed to be $m(z)$ where $m(.)$ is an arbitrary function which is the same as that described by Tang et al.[13] Then, the consistent estimates for parameter estimates of $[Z|x;]$ can be obtained by the pair-wised form of the pseudo likelihood approach[13] as (1.3).

$$\hat{\vec{\gamma}} = \arg\max_{\vec{\gamma}} \prod_{i=1}^{n} \frac{g_z(z_i|x_{i1}, \ldots x_{ip};\vec{\gamma})}{\int\ldots\int g_z(z_i|x_{i1}, \ldots x_{ip};\vec{\gamma}) \cdot f_x(x_{i1}, \ldots x_{ip};\hat{\vec{\alpha}}) dF_1 \ldots dF_p} \tag{1.3}$$

where

$$\hat{\vec{\alpha}} = \arg\max_{\vec{\alpha}} \prod_{i=1}^{n} f_x(x_{i1}, \ldots x_{ip};\vec{\alpha}) \tag{1.4}$$

$F_i$ is the cumulative distribution of $[X]$ for $i = 1, \ldots p$ on (1.3) and $\hat{\vec{\alpha}}$ is a vector of the maximum likelihood estimates of marginal distribution $[X]$ on (1.4). The parameter estimates, $\hat{\vec{\gamma}}$, are asymptotically consistent and normally distributed.[13] The conditional distribution of $[Z|x]$ is identified by plugging $\hat{\vec{\gamma}}$ into $g_z(z|x)$. This method can be also applied as a semi-parametric approach by replacing $f_x(x)$ with its empirical distribution. For more details, refer to Tang et al.[13]

## Binary variable generation

The conditional distribution of $[Z|x;\vec{\gamma}]$ is identified, and this step recomposes the data matrix according to the identified conditional distribution. Assuming that the missing data mechanism is NMAR, we stochastically draw the predictive values of $z_i$ in given $x_i$ for $i = 1, \ldots n$ regardless of the missingness. A binary variable is then created by comparing $Z$ with a preassigned threshold c according to (1.1). With these binary values and $X$, we can estimate the effect of the covariate of interest by logistic regression. This procedure is repeated multiple times, and the average effect can be obtained by averaging the coefficients of the covariate of interest in the logistic regression model. This procedure can be summarized as follows.

1. For subjects $i = 1, \ldots n$, perform (a) and (b):
   a. Stochastically generate the predictive value of $\hat{z}_i$ from $[z|x;\vec{\gamma}]$.
   b. Create a binary value for $\hat{y}_i$ by comparing $\hat{z}_i$ with a given threshold c according to (1.1).
2. Conduct the logistic regression using $\hat{Y}$ and $X$ by treating the dataset as the complete data, and estimate the effect of the covariate of interest, $x_1$.
3. Repeat m times from 1 to 2, and compute the average effect of the covariate of interest by averaging m coefficients obtained from the logistic regression.

Most imputation methods assume that the missing data mechanism is either MCAR or MAR because we can get the consistent estimates from the observed data. If the missing data mechanism is NMAR, we are not able to find the consistent estimates for the imputation techniques only with the observed data. To overcome this hurdle, we obtain consistent estimates for parameters using the likelihood approach, and use those estimates for imputation. The imputation method is easy and convenient to perform, and we can treat the imputed dataset as the complete dataset.[10] In addition, discretized variables contain less information than a continuous variable and they are less easily modeled in general. Instead of using incomplete binary data, this two-step procedure fully identifies the conditional distribution of $[Z|x]$ with its parameter estimates, and draws binary variables from the distribution identified.

## Simulation

Before the analysis, the simulation is conducted using the described two-step procedure, and the performance is observed in terms of bias of the effect of the covariate of interest. The true value is defined as the effect of the covariate of interest obtained by logistic regression before creating the missing values. At first, 100 bivariate random samples ($X$, $Z$) are generated from bivariate normal distribution with different sample sizes of 100, 300, 500, and 1000. Binary samples of $Y$ are correspondingly created by dichotomizing $Z$ according to (1.1) where the threshold c is fixed at 1.2,

$$\begin{aligned} X &\sim N(\mu_x, \sigma_x^2) \\ Z \mid x &\sim N(\gamma_0 + \gamma_1 \cdot x, 1) \end{aligned} \tag{1.5}$$

where $\{\mu_x, \sigma_x^2\} = \{0,1\}$, and $\{\gamma_0,\gamma_1\} = \{1,1\}$. For nonignorable missing data, the missing values are created according to $pr[r_i = 1|x,z] = \Phi(\varphi_0 + \varphi_1 \cdot z_i)$ where $\{\varphi_0,\varphi_1\} = \{1,-1\}$. Each sample set has two types of data, ie, complete data and incomplete. Incomplete data are generated by creating missing cases from the complete data, and the simulation procedure is summarized as follows:

1. Compute the regression coefficient estimates, using logistic regression with $Y$ and $X$ with complete data
2. Create the missing cases of $Z$ and $Y$ according to the described missing data mechanism
3. Define the conditional distribution of $[z|x;\vec{\gamma}]$ by estimating the parameters as described in the first step
4. Conduct the second step.
   a. Stochastically draw values of $\hat{z}_i$ from $[z|x;\vec{\gamma}]$ for $i = 1, \ldots n$.
   b. Create binary values of $\hat{y}_i$ by comparing $\hat{z}_i$ with a given threshold c for $i = 1, \ldots n$
   c. Conduct logistic regression with $\hat{Y}$ and $X$.
5. Repeat step 4 100 times and compute the average of the regression coefficients and standard errors.[8]

Tables 1 and 2 show the simulation results in terms of bias and standard errors. In Table 2, the bias of the slope by pseudo maximum likelihood estimate is slightly bigger when the sample size is 100, and the biases by maximum likelihood estimate with the known missing data mechanism are smaller than by the pseudo maximum likelihood estimate. Tang et al showed that the maximum likelihood estimate with the known missing data mechanism performs better than pseudo maximum likelihood estimates in the simulation.[13] However, overall biases are negligible, and standard errors are also close to the average standard errors by logistic regression with complete data, as long as the conditional distribution of $[Z|x]$ is identified with consistent estimates with the original variable before dichotomization. In Figure 1, the far left plot is a scatter plot of an original random sample of the continuous variable $Z$ on $X$, and the second furthest left plot is a scatter plot of $Y$ on $X$ where $Y$ is created by dichotomizing $Z$ according to (1.1). The right two plots are plots correspondingly estimated by the above two-step procedure for incomplete random samples. In the two tables, we can observe that the average biases are negligible, and that they become smaller as the sample size becomes bigger. Even though we have to use the likelihood approach to obtain

**Table 1** Results with complete data

| Sample size | Intercept | | Slope | |
|---|---|---|---|---|
| | Bias | SE | Bias | SE |
| 100 | –0.0196 | 0.2590 | 0.0197 | 0.3846 |
| 300 | –0.0012 | 0.1462 | –0.0021 | 0.2134 |
| 500 | –0.0024 | 0.1122 | 0.0006 | 0.1624 |
| 1000 | 0 | 0.0792 | 0.0013 | 0.1146 |

**Abbreviation:** SE, standard error.

**Table 2** Results with incomplete data

| Sample size | Using PMLE | | | | Using MLE | | | |
|---|---|---|---|---|---|---|---|---|
| | Intercept | | Slope | | Intercept | | Slope | |
| | Bias | SE | Bias | SE | Bias | SE | Bias | SE |
| 100 | –0.0130 | 0.2684 | 0.1080 | 0.3930 | 0.0045 | 0.2584 | 0.0312 | 0.3781 |
| 300 | 0.0014 | 0.1474 | 0.0291 | 0.2154 | –0.0033 | 0.1455 | 0.0030 | 0.2105 |
| 500 | –0.0064 | 0.1125 | 0.0148 | 0.1633 | –0.0058 | 0.1126 | 0.0098 | 0.1628 |
| 1000 | 0.0093 | 0.0793 | 0.0083 | 0.1142 | –0.0011 | 0.0791 | 0.0020 | 0.1142 |

**Abbreviations:** SE, standard error; PMLE, pseudo maximum likelihood estimation; MLE, maximum likelihood estimation.

consistent estimates, we can avoid unidentifiability by using the original continuous variable, and observe negligible bias in the regression coefficients with incomplete data, for which the missing data mechanism is NMAR as long as the estimates are consistent. For unidentifiability, the reader is referred to the detailed proofs published by Tang et al,[13] Chen,[2] and Carroll et al.[1]

## Application in secondary hyperparathyroidism

The data used in this research are from the medical records of 297 patients with chronic kidney disease attending the Renal Clinic, Grady Memorial Hospital, Atlanta, GA, in 2010. We excluded some patients whose medical records did not match their records for 2011. Biomedical measurements for each subject were not taken exactly on the same date because the data were collected retrospectively from registered medical records, but the measurements were taken at similar time points. So it is assumed that there is no clinical difference in the measurements over time.

The female to male ratio was almost same, 79% of subjects were of African American ethnicity, and the average patient age was 62 years. We chose four variables, ie, glomerular filtration rate, parathyroid hormone, secondary hyperparathyroidism, and vitamin D levels for this analysis. Secondary hyperparathyroidism is the main binary outcome, and is defined as an "event" when the parathyroid hormone level goes beyond the threshold for a particular disease stage, which is defined by the range of glomerular filtration rate, and the threshold may vary according to disease stage. When the parathyroid glands produce too much parathyroid hormone, the patient with chronic kidney disease can progress to chronic renal failure, so the National Kidney Foundation suggests a threshold for each disease stage, and recommends monitoring whether a patient's parathyroid hormone level goes beyond this limit.[5] In addition, if a patient with a disease stage in the range of 3–5 shows higher parathyroid hormone levels, the National Kidney Foundation recommendation is to check the 25-hydroxyvitamin D level. From the medical perspective, secondary hyperparathyroidism is a more important indicator of kidney failure than fluctuations in parathyroid hormone levels, and the association with the vitamin D level is drawing attention. Although some papers have reported that 25-hydroxyvitamin D deficiency might be a cause of secondary hyperparathyroidism,[3,8] this association is still arguable.

The main purpose of our analysis was to evaluate the effect of vitamin D on secondary hyperparathyroidism

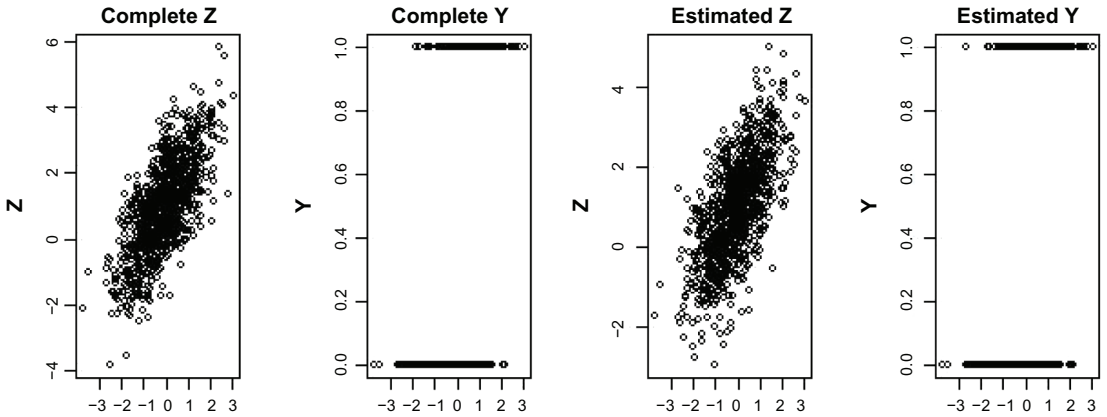


**Figure 1** Comparison of scatter plots: true versus estimated.

controlling for different disease stages. Glomerular filtration rate and vitamin D values are complete in the dataset, but data on parathyroid hormone levels and secondary hyperparathyroidism were missing for 10% of cases. We compared the glomerular filtration rate and parathyroid hormone levels between observed cases and missing cases, and found a significant difference in means of 34.84 versus 43.09 between the two groups. Parathyroid hormone levels tend to have a inverse relationship with glomerular filtration rate, and parathyroid hormone levels may be different between two groups. Also, from the clinical perspective, lower parathyroid hormone levels may be neglected in records because they are not of great concern. Therefore, the NMAR assumption for parathyroid hormone levels may be more plausible for our example. We compared the results using two different methods; one obtained by the two-step procedure described earlier for the NMAR assumption, and the other obtained by complete case analysis under the MCAR assumption, which is commonly used in medical research.

## Results

We denote glomerular filtration rate, vitamin D, parathyroid hormone, and secondary hyperparathyroidism as $X_1$, $X_2$, $Z$, and $Y$, respectively, and assume the following parametric settings to derive pseudo maximum likelihood estimates for these parameters. If the distributions of $[Y|x]$ and $[X|y]$ are not concordant with each other, the efficiency of the model would be lower on (1.3). For example, if we assume the distribution of $[Y|x]$ to be multinomial and that of $[X|y]$ as a normal distribution, we have to discard the term of $x_2$ in the parametric approach, so it is not efficient. Therefore, we assume normal distributions for all, and variables are transformed if necessary. Normality assumptions are checked with Q-Q plots (Figure 3), and $Z$ and $X_2$ are log-transformed parathyroid hormone and vitamin D levels on (1.6).

$$\begin{aligned}
&[Z \mid x_1, x_2] \sim N(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2, \sigma^2) \\
&[X_2 \mid x_1, z] \sim N(\gamma_0 + \gamma_1 \cdot x_1 + \gamma_2 \cdot z, \delta_1^2) \\
&[X_1 \mid z] \sim N(\gamma_3 + \gamma_4 \cdot z, \delta_2{}^2) \\
&[X_2 \mid x_1] \sim N(\mu_{20} + \mu_{21} \cdot x_1, \sigma_2{}^2) \\
&[X_1] \sim N(\mu_1, \sigma_1{}^2)
\end{aligned} \tag{1.6}$$

We can identify $\left\{\frac{\beta_1}{\sigma^2}, \frac{\beta_2}{\sigma^2}, \frac{\beta_1 \cdot \beta_2}{\sigma^2}, \frac{\beta_0 \cdot \beta_2}{\sigma^2}\right\}$ using the above settings,[2] so we are able to estimate the parameters $\vec{\beta} = \{\beta_0, \beta_1, \beta_2, \sigma^2\}$ of $[Z|w,x_2]$ by solving their linear equations, and the solutions are summarized as follows.

**Table 3** Coefficient estimates

| Description | Estimate | SE |
|---|---|---|
| $\hat{\vec{\beta}}$ | | |
| $\hat{\beta}_0$ | 5.7381 | 0.1062 |
| $\hat{\beta}_1$ | –0.0285 | 0.0026 |
| $\hat{\beta}_2$ | –0.1853 | 0.0819 |
| $\hat{\sigma}^2$ | 0.4402 | 0.0427 |

**Abbreviation:** SE, standard error.

$$\begin{aligned}
\hat{\beta}_0 &= \left(\frac{\mu_{20}}{\sigma_2^2} - \frac{\gamma_0}{\delta_1^2}\right) \cdot \left(\frac{\delta_1^2}{\gamma_2}\right) \\
\hat{\beta}_1 &= \left(\frac{1}{\delta_2^2} + \frac{\gamma_1^2}{\delta_1^2} - \frac{\mu_{21}^2}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) \cdot \left(\frac{\gamma_4}{\delta_2^2} - \frac{\gamma_1 \cdot \gamma_2}{\delta_1^2}\right)^{-1} \\
\hat{\beta}_2 &= \left(\frac{1}{\delta_1^2} - \frac{1}{\sigma_2^2}\right) \cdot \left(\frac{\delta_1^2}{\gamma_2}\right) \\
\hat{\sigma}_2 &= \left(\frac{1}{\delta_1^2} - \frac{1}{\sigma_2^2}\right) \cdot \left(\frac{\delta_1^2}{\gamma_2}\right)^2
\end{aligned} \tag{1.7}$$

Accordingly, we can compute the parameter estimates and their standard errors. The standard errors are computed using the bootstrapping technique, and the bootstrapping is conducted with 100 replications. The results are summarized in Table 3. Based on the identified conditional distribution of $[z|x;\vec{\beta}]$, we generate the predictive values of parathyroid hormone, $\hat{z}_i$ and create the predictive values of secondary hyperparathyroidism, $\hat{y}_i$, by dichotomizing $\hat{z}_i$ at given thresholds for $i = 1, \ldots$ n, where n is the number of subjects. We then conduct logistic regression with $\hat{Y}$ on covariates, including vitamin D level. This procedure is repeated 100 times, and the average coefficient estimates with corresponding standard errors are summarized on

**Table 4** Coefficient estimates by logistic regression

| Effects | MCAR | | NMAR | |
|---|---|---|---|---|
| | Estimate | SE | Estimate | SE |
| Intercept | 0.8507 | 0.7575 | –0.4937 | 1.1313 |
| Stage 3 | 1.5185 | 0.4760 | 1.3419 | 0.8998 |
| Stage 4 | 1.4895 | 0.5103 | 1.5803 | 0.9169 |
| Stage 5 | 0.2961 | 0.5668 | –0.1540 | 1.0762 |
| Vitamin D* | –0.5382 | 0.2175 | –0.4871 | 0.2950 |

**Abbreviations:** MCAR, missing completely at random; NMAR, not missing at random; SE, standard error.

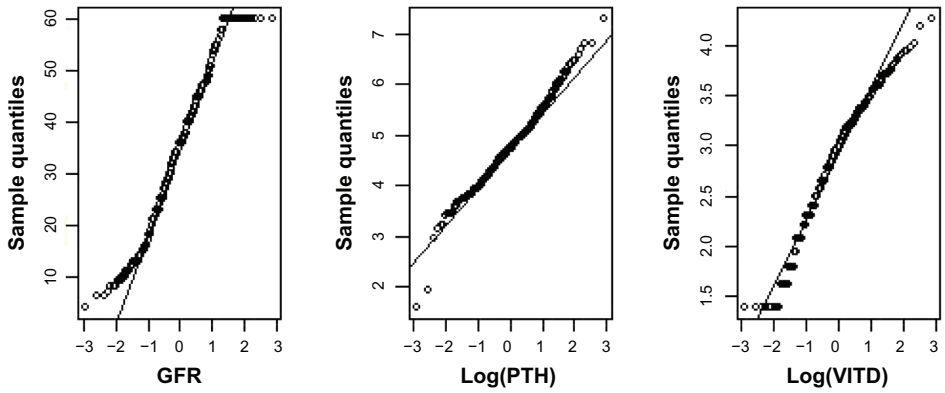


**Figure 2** Q-Q plots for glomerular filtration rate, log-transformed [parathyroid hormone] and log-transformed [Vitamin D].

Table 4. The two left hand columns in Table 4 show the results for the estimates and the standard errors that are obtained by complete case analysis. We can see that the vitamin D level is a significant covariate for secondary hyperparathyroidism, regardless of the missing data mechanism, with $P = 0.014$ under MCAR and $P = 0.049$ under NMAR. However, the estimated odds ratio for the vitamin D level under MCAR is 0.5837 (95% confidence interval 0.3812–0.8941) while that under NMAR is 0.6144 (95% confidence interval 0.3446–1.0954). The effect of vitamin D is highly significant assuming MCAR, but only modest assuming NMAR. We also tested the inference of MCAR using the likelihood ratio test according to the $\chi^2$ distribution,[9] but our dataset does not have enough evidence to show MCAR with d2 = 0.4961. Thus, the NMAR assumption may be more plausible. Figure 2 shows the relationship between estimated parathyroid hormone levels, occurrence of secondary hyperparathyroidism, and log-transformed vitamin D levels using the two-step procedure.

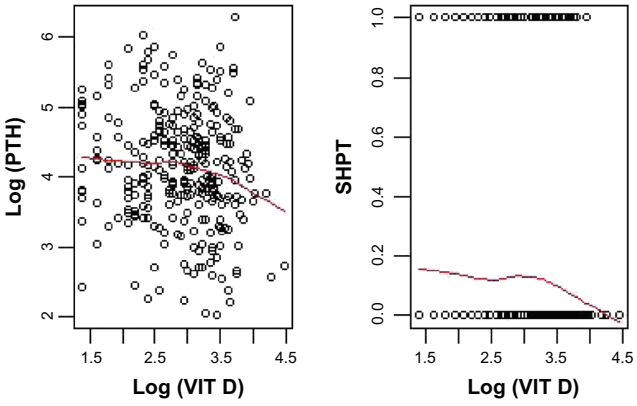


**Figure 3** Predicted values (not missing at random).
**Abbreviations:** PTH, parathyroid hormone; SHPT, secondary hyperparathyroidism.

## Conclusion

If the missing data mechanism is assumed to be composed only of response variables, the pair-wise pseudo likelihood approach enables us to find asymptotically consistent estimates without specifying the missing data mechanism for nonignorable data. This approach holds when the variable is continuous or multinomial, having at least three distinctive points,[13] so when the variable is binary, this method confronts an unidentifiability problem.[13] However, if the original variable is continuous, and the binary variable is driven from the continuous variable by dichotomizing it, such as in our example, we may estimate the average effect of the covariate of interest by regenerating binary variables after fully identifying the conditional distribution of the original data instead. In fact, if the dichotomized variable is incomplete, the missing values are inherited from the original variable, so the original variable has more information than the dichotomized variable about the missing data mechanism. In this research, we estimated the effect of the covariate of interest on a binary outcome with the original data using the two-step procedure, and the simulation shows negligible bias in estimation. However, both the simulation and the application example are conducted with normally distributed datasets, so it needs to be studied further using non-normal data.

## Acknowledgments

We thank Dr K Bashir and his team in Morehouse School of Medicine for providing us with the dataset.

## Disclosure

The authors report no conflict of interest in this work.

## References

1. Carroll RJ, Wang SJ, Wang CY. Prospective analysis of logistic case-control studies. *J Am Stat Assoc*. 1995;90:157–169.

2. Chen K. Parametric models for response-biased sampling. *J R Stat Soc Ser A Stat Soc*. 2001;63(4):775–789.
3. Drueke TB. Treatment of secondary hyperparathyroidism with Vitamin D derivatives and calcimimetics before and after start of dialysis. *Nephrol Dial Transplant*. 2002;17:20–22.
4. Imbens GW. An efficient method of moment estimator for discrete choice models with choice-based sampling. *Econometrica*. 1992;60:1187–1214.
5. Kalbfleisch J. Likelihood method and nonparametric tests. *J Am Stat Assoc*. 1978;78:167–170.
6. Liang KY, Qin J. Regression analysis under non-standard situation: a pairwise pseudo-likelihood approach. *J R Statist Soc B*. 2000;62:773–786.
7. Lips P. Vitamin D deficiency and secondary hyperparathyroidism in the elderly: consequences for bone loss and fractures and therapeutic implication. *Endocr Rev*. 2001;22:477–501.
8. Little RJA, Rubin DB. *Analysis of Missing Data*. 2nd ed. New York, NY: John Wiley & Sons; 2002.
9. Little RJA. A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc*. 1988;83:1198–1202.
10. Nemes S, Jonasson JM, Genell A, et al. Bias in odds ratios by logistic regression modeling and sample size. *BMC Med Res Methodol*. 2009;9:1–5.
11. National Kidney Foundation. KDOQI clinical practice guideline for chronic kidney disease: evaluation, classification, and stratification. *Am J Kidney Dis*. 2002;39:S1–S266.
12. National Kidney Foundation. Test Your Kidney IQ. Glomerular filtration rate (GFR). Available from: http://www.kidney.org//kidneydisease/ckd/knowgfr.cfm. Accessed November 19, 2012.
13. Tang G, Little RJA, Raghunathan TE. Analysis of multivariate missing data with nonignorable nonresponses. *Biometrika*. 2003;90:74–764.