Universal Journal of Public Health 10(6): 555-562, 2022 http://www.hrpub.org
DOI: 10.13189/ujph.2022.100601

Statistical Methods of Handling Ordinal Longitudinal Responses with Intermittent Missing Data

Aluko O.[1,*], Mwambi H.[2]

[1]Department of Biostatistics, Faculty of Health Sciences, School of Biomedical Sciences, University of the Free State, South Africa
[2]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag X01 Scottsville 3209, Pietermaritzburg, South Africa

Received March 18, 2021; Revised October 21, 2022; Accepted November 14, 2022

Cite This Paper in the Following Citation Styles

(a): *[1] Aluko O., Mwambi H. , "Statistical Methods of Handling Ordinal Longitudinal Responses with Intermittent Missing Data," Universal Journal of Public Health, Vol. 10, No. 6, pp. 555 - 562, 2022. DOI: 10.13189/ujph.2022.100601.*

(b): *Aluko O., Mwambi H. (2022). Statistical Methods of Handling Ordinal Longitudinal Responses with Intermittent Missing Data. Universal Journal of Public Health, 10(6), 555 - 562. DOI: 10.13189/ujph.2022.100601.*

Abstract The rate of survival of human immunodeficiency virus (HIV) positive individuals resume to ameliorate with the usage of highly active antiretroviral therapy (HAART), yet pulmonary disease prevalence is growing unstopped amongst some of them. Handling missing data was a difficult challenge in the health data concept. We compared the effect of marginal to the proposed ordinal negative binomial models in handling intermittent missing observations for better numerical performance. The data used was characterised by monotone missing observations due to patients' failure to declare their pulmonary conditions (lung infections and complications) and other vital health information. The use of multiple imputations is one of the latest techniques for handling missing observations, and this technique is relevant to different missing data mechanism assumptions, but occasionally profiles with the complication of the kind of parameters to be imputed and the mechanism underlying the incomplete data. This study focuses on the importance and application of the methods of handling missing health data. The proposed ordinal negative binomial model performs greatly than other models in adjusting for monotone missing data without imputation. In a real application, the proposed ordinal negative binomial model produces low estimates as against direct likelihood, mixed effects proportional odd, and MI-GEE models.

Keywords Ordinal Outcomes, Ordinal Negative Binomial, Multiple Imputations, Monotone Missingness

1. Introduction

As of 2011, approximately 33 million persons have been diagnosed with human immunodeficiency virus (HIV) as reported by the national heart, lung, and blood institute (NHLBI). The rate of survival has improved and turned HIV into a chronic disease based on the development of antiretroviral therapy and other clinical procedures. However, the continuous loss of immunity arising from HIV-infected individuals has resulted in the growing evolution of comorbidity diseases which are remarkably increasing the death rate among the infected individuals. The large number of HIV-infected individuals in the regions with few resources experienced lung complications (pulmonary disease) that are abysmally distinguished. The objective of this work was to provide awareness about the connection between pulmonary disease and HIV infection. The interest is in the description of the axial emphysema distribution outcomes which were classified as ordinal. Axial emphysema was marked by an unusual lasting expansion of the airspaces distal to the terminal bronchioles followed by alveolar wall destruction, with evident fibrosis. Axial emphysema was common among persons living with HIV. HIV-infected patients with axial emphysema burdens were best visualised on the computed tomography (CT) scan because the spirometry test may not be adequate in assessing the extent of distribution as observed in our longitudinal ordinal cohort study. For the classification of patients using diagnosis results, it is necessary to assign individuals to the right levels of the

possibility of having lung complications. For clarity purposes, ordinal data property needs to be applied to the ordinal response categories. Using the methods of analyzing categorical data to ordered categorical data may lead to loss of information. Using models specifically developed for ordinal categorical data has the advantage of taking into account the ordering pattern of the response categories. The models developed for ordered categorical data use limited statistical resources than their unordered counterparts which produce efficient results which facilitate the interpretation of the variables as mentioned in [19]. The two areas of statistical inference that have importance in ordinal data modification are *association and regression* as stated by [1]. The logistic and log-linear regression models were developed in the late 1960s through early 1970 respectively. Around that time ordinal data analysis enjoyed little recognition until [3,26], later research interests were developed by [20] on the logit and [10] on log-linear modeling of cumulative probabilities and odd ratios respectively. In literature, we have three popular methods that defined the logit of ordinal data response. These are the adjacent-category logit, the continuation-ratio logit, and the cumulative logit [20]. Furthermore, we have stereotype models (multiplicative paired-categories logit) used for modeling ordinal data. We applied both cumulative and adjacent-categories logit in our study.

Also [19] stated that other methods used for modeling ordinal data without reference to odd ratios, or logarithms follow a normal distribution. The use of methods from a simple binomial distribution to more flexible distributions that allow for over-dispersion, for example, beta-binomial distribution. However, repeated measurements (longitudinal data) as one of the methods used for analyzing ordered categorical responses has received wide publication in recent times. Repeated measurements and crossover experiments display some sort of clustering. In repeated measurements, data are analyzed using any of these three procedures: *conditional, marginal, and transition models*. Other known methods include Poisson, negative binomial (NB), hurdle, mixed-effect proportional odds, ordinal negative binomial (ONB), and zero-inflated negative binomial models.

However, repeated measurements are often characterized by missing observations either on the covariates or response or both. The missing observations could be monotone or intermittent. In dealing with missing data from a discrete variable, the first method that comes to mind may be to handle the variable as a continuous variable for imputation and then round off the imputed values to the nearest valid discrete value before fitting the substantive model [5]. The intuition behind multiple imputations (MI) is to draw valid and efficient inferences and fit the analysis model to multiply imputed data. The imputed values should be similar to the data structure, but uncertain about the structure and insensitive to the process that led to missing observations [28]. The technique behind the imputation of datasets depends on the missing data pattern. In the case of monotone dropout patterns, the parametric regression method assumes multivariate normality or a nonparametric approach that employs propensity scores may be used [18]. Also, the methods that assume normality had been used [6,7,25] in different studies. Using multiple imputations generates a series of univariate regressions against a single large model (making estimation easier), and drops the normality assumption of the variables. Furthermore, for some reasons, researchers advised against handling ordinal response as a continuous or dichotomized variable. The reasons include efficiency loss of information, reduced statistical power, and decreased generality of conclusions [9]. The use of a continuous model may produce predicted values outside of the range of the ordinal variable and finally, a continuous model may yield correlated residuals and regressors when used for an ordinal response and does not account for the ceiling and floor effects of the ordinal response. This may lead to biased estimates of the regression coefficients [2].

We consider different models in the analysis: direct likelihood (DL), ONB, mixed-effects proportional odds, generalized estimating equations (GEE), and the sub-component of multiple imputations (MI-GEE). GEE and MI-GEE use the marginal model. The marginal model is also termed *population-averaged* effects which means the average of the clusters, this does not rely on random effects but the covariates of interest. Generalized estimating equations (GEE) use quasi-likelihood estimation while Fitting ONB model can be achieved using the standard maximum likelihood estimation by optimizing the respective log-likelihood functions. Mixed-effects proportional odd model for ordinal data accommodate multiple random effects [11].

In our simulation study, ordinal data were simulated from a negative binomial (NB) distribution. After the simulation, we linked the ordinal responses to appropriate count distribution through the known cut-off points within the ordinal outcome. Thereafter, monotone dropout patterns were introduced and later we analyzed the real dataset. Finally, the results and contribution of the proposed ordinal negative binomial model to knowledge were discussed.

In Section 2, the details of the methods used were presented. In Section 3, we describe the simulation study and present empirical data analysis in Section 4. Section 5 concludes the paper with a discussion.

2. Methods

2.1. Direct Likelihood Method

The idea of the direct likelihood (DL) method was introduced by [15] as a better method over imputation in

handling incomplete data for ignorable missing mechanisms such as missing completely at random (MCAR) and missing at random (MAR) [14]. This technique is called likelihood-based ignorable analysis or direct likelihood analysis [17]. In the DL technique, all data are analyzed without deleting or imputation using models that offer a framework from which to analyze clustered data by including both the fixed and random effects in the model. For an extensive description of the DL method as a way of analyzing incomplete data, interested individuals should read [17].

2.2. Ordinal Negative Binomial Model (ONB)

In this Model, an ordinal response is a function of negative binomial (NB) distribution with the assumption that the ordinal response, Y_i for individual $i(i = 1,...,n)$, is an unobserved latent variable, Y_{ij}^*. It is known that the unobserved latent variable may follow any distribution, but our study focuses on the NB. About [16], the probability mass function (PMF) for NB distribution in terms of Y_{ij^*} conditional on covariates, x_i, is

$$f(y_{ij}^*) = P(Y_{ij}^* = y_{ij}^* | x_i) = \frac{\Gamma(y_{ij}^* + a^{-1})}{\Gamma(a^{-1})\Gamma(y_{ij}^* + 1)} (a\mu_1)^{y_{ij}^*} (1 + a\mu_i)^{-(y_{ij}^* + a^{-1})},$$

$$y_{ij}^* = 0, 1, 2, \ldots \quad (1)$$

where $E(Y_{ij}^*) = \mu_i,\ Var(Y_{ij}^*) = \mu_i + \alpha\mu_i^2$, and α is the dispersion parameter. We adopt the log-link function to link the linear predictor to the mean of Y_{ij^*} as

$$log(\mu_i) = x_i'\beta \quad (2)$$

where x_i is a $p \times 1$ vector of covariates (this includes '1' as the first element for the intercept) and β is a $p \times 1$ vector of regression coefficients.

The cumulative distribution function (CDF) for NB distribution is simply the sum of the PMFs such that:

$$F(y_{ij}^*) = \sum_{v=0}^{y_{ij}^*} f(v), \quad (3)$$

where cumulative probability is evaluated at y_{ij}^*.

The next step links the ordinal response, Y_{ij}, to the unobserved variable, Y_{ij}^*. This is achieved using a fixed value of cut-off points to generate the categories of ordinal response where:

$$Y_{ij} = c \ \ if\ k_{c-1} < Y_{ij}^* \leq k_c \quad (4)$$

the threshold k_c defines the upper bound of ordinal response category $c(c = 1,2,...M)$. The probability of observing an outcome in category c can be expressed as the function of the cumulative probabilities of the underlying Y_{ij^*} distribution, that is:

$$P(Y_{ij} = c|x_i) =$$

$$P((k_{c-1}) < Y_{ij}^* \leq k_c)|x_i = F(k_c) - F(k_{c-1}) \quad (5)$$

Specifically, $F(k_c)$ and $F(k_{c-1})$ designate the CDFs measured at the known upper count numbers for categories c and $c - 1$ for a distribution with a mean of μ_i and dispersion of α. The likelihood function over all individuals and categories for ordinal data following underlying count distribution can be expressed as:

$$L_{ONB} = \prod_{i=1}^{n} \left[\prod_{c=1}^{M} \left[F(k_c) - F(k_{c-1}) \right]^{y_{ic}} \right] \quad (6)$$

The ONB technique is not the same as the current practice of proportional odd (PO) because it linked the ordinal responses to the NB CDF as against the logistic CDF. In the proposed ONB, the effect of covariates was studied as if it models the latent responses directly, but (PO) handles the effect of covariates over the cumulative odds across response categories.

2.3. Mixed-effects Proportional Odd Model

Mixed-effects proportional odd model is referred to as the cumulative logit model [11]. It is assumed that the effect of covariates is the same in this model for each cumulative odd ratio. The cumulative probability for the C categories of response Y (with intermittent missing data) is P_{ijc}=Pr($Y_{ij} \leqslant c$), where P_{ijc} represents the (conditional) cumulative probabilities response in C categories. An ordinal proportional odd random-intercept model is written as

$$log\left[\frac{P_{ijc}}{1-P_{ijc}}\right] = \gamma_c - Z_{ij}(c = 1, \ldots, c) \quad (7)$$

The random subject effect follows an assumed standard normal distribution $N(0, 1)$ in the population. The model threshold is γ_c with c-1 increasing order. In c categories, marginal response probabilities are expressed in model intercept and threshold parameters. On the random effect μ_{0i} level, the effects of regression coefficients are conditional.

2.4. Generalized Estimating Equations (GEE)

Likelihood-based and non-likelihood-based methods can be used to estimate the regression coefficients of marginal models as described by [8]. One challenge associated with likelihood models lies in the relationship difficulty between the estimates of the models and the joint probability of the likelihood. GEEs is another option for likelihood-based but it has been explored because it is popular in analyzing non-Gaussian correlated data. With this technique, the joint distribution specified for the repeated outcomes is bypassed but the specification is utilized in marginal distribution. Proportional odd is not a family of a generalized linear model, so transformation is needed before GEE can be applied. As stated by [13], create for each individual at each occasion a (K-1) dimensional expanded vector of binary outcomes, $Y_{ij}^* = (Y_{i1_j}^*, \ldots, Y_{i,(K-1),j})'$, where $Y_{ikj} = 1$ if $Y_{ij} \leqslant K$ and

0, otherwise.

Now,

$$logit\left[Pr\left(Y_{ij} \le k|x_{ij}\right)\right] = logit\left[Pr\left(Y_{ikj}^{*} = 1|X_{ij}\right)\right], \quad k = 1, \ldots \mathrm{K}-1 \tag{8}$$

the logistic regression model is part of the generalized linear model family, then the GEEs technique works, and unbiased estimates of the regression can be obtained by solving the estimating equations

$$\sum_{i=1}^{N} \frac{\partial \pi_i'}{\partial \beta} V_i^{-1}(Y_i^{*} - \pi_i) = 0 \tag{9}$$

where

$Y_i^{*} = (Y_{i_1}^{*}, \ldots, Y_{iT}^{*})'$, $\pi_i = E(Y_i^{*})$, $V_i = A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}}$

with A_i, the diagonal matrix of the variance of the elements of $^{*}{}_i$, and β the expanded vector of intercept and regression coefficients. The working correlation structure (i.e. matrix R_i) is required, this indicates the dependence among the repeated clusters over individuals ranging from independence, unstructured exchangeable, compound symmetry, and autoregressive (1).

2.4.1. Imputation Methods

Multiple imputations (MI) is a unique tool to impute missing data [23]. MI produces several imputed independent datasets. This accounts for uncertainty in the estimation due to missing data. Missing data are predicted from the observed data based on a different set of parameter values in each imputed dataset [24,22]. Therefore, the analysis model is applied to each imputed data set after imputation. Results obtained from the analyses are then pooled to produce a set of final results. In general, three procedures are applied to impute ordinal data with missing data: one handles a normal distribution and uses normal data imputation models (such as linear regression and saturated models), the second produces imputations using a categorical imputation model (such as logistic regression models), and the last produces imputations based on a latent variable model. Knowing how these procedures work, analysts should know the appropriate procedure(s) to use to analyze ordinal missing data after imputation. This study addresses the question in a frequently encountered manner. To explore the technique of fully conditional specification (FCS) in our imputation model, over the commonly used multivariate normal imputation model (MVNI) because the FCS technique imputes variables separately and sequentially at each iteration.

2.4.2. Fully Conditional Specification (FCS)

The fully conditional specification (FCS) procedure serves as an alternative to the multivariate normal imputation method. In the FCS approach, there is a flexible component that specifies the model by a series of univariate conditional models for each of the missing data. The slight difference between the FCS and MVNI is that the former relies not on the normality assumption of MVNI, and the use of univariate regression models for ordered logistic regression for ordinal variables is appropriate if it is well tailored. The Bayesian approach conducts imputations in stepwise order where the least missing observation is filled sequentially until the most missing observation is filled. The imputations involve two stages: fill-in and imputation. In each stage, draws are random from both the posterior distribution of the parameters and the missing observations. In the fill-in stage, the filling of missing observations is in sequential order where the preceding variables serve as covariates. The filled-in values are the starting values for the imputation stage. In the imputation stage, filled-in values replaced the imputed values for each variable at each iteration sequentially.

Let the ordinal response variable Y be characterized by a vector of unknown parameters $\theta = (\mu, \Sigma)$; μ is a mean vector while Σ is a covariance matrix. By Assumption, the complete data is partitioned as, $Y = (Y_o, Y_m)$ where Y_o and Y_m are the observed complete incomplete components of the data respectively. According to [29,30], multiple imputations via FCS proceeds as follows:

Calculate the posterior distribution θ given the observed data, that is, $P(\theta|Y_o)$;

Then θ^{*} is drawn from $P(\theta|Y_o)$;

Then y^{*} from the conditional posterior distribution of y_m given $\theta = \theta^{*}$:

$$y^{*} \sim P(y_m|y_o, \theta = \theta^{*}) \tag{10}$$

Steps two and three are repeated depending on the number of imputations. These simulate approximately independent draws of missing observations for imputed datasets.

2.5. Software Considerations

In the likelihood-based approach where dropout is not modeled valid results are produced especially when MAR is assumed. And in the analysis model, the generalized linear mixed model is used. The SAS procedures are implemented in NLMIXED and GLIMMIX. Imputing missing observations "PROC MI" procedure is used to describe the missing data pattern and perform the multiple imputations. Under this approach, different variables are considered and this offers several methods of imputation whether it is categorical or continuous. In our study, the interest was to compare four methods: the DL method, ONB, mixed-effects proportional odds, and GEE. The FCS statement is implemented in "PROC MI". Markov Chain Monte Carlo (MCMC) is used to draw the imputed values from the multivariate normal distribution. In "PROC MI" procedure, the number of imputations and the imputation model to use are specified. Statistical procedures are implemented to run the analysis model after imputation on each imputed dataset using Imputation statement, and the inferences are located in the output file. In the final process, PROC MIANALYZE is used to combine the estimates from multiple imputed datasets to obtain valid

results [12]. An additional programme may be needed for the analyses of complete categorical data [21]. Combined results assume that the estimation of the statistics follows normal distribution while the estimation of odd ratios, correlation coefficients, and relative risks follows non-normal distribution [22]. These estimates obtain from non-normal distribution need to be normalized before the application of Rubin's combination rules to transform the estimates that are if the focus is on the latter group of estimates. Some transformations are detailed in [27] as a suggestion. The obvious ones for categorical data are log odd ratios and log relative risks.

3. Simulation Study

In the analysis of real data, simulation studies were conducted early to ascertain the effectiveness and evaluate the properties of the selected methods to handle missing data, and the interest was in handling intermittent missing data.

3.1. Data Generation, Simulation Designs, and Analysis of the Simulated Data

In our simulation studies, the specifications and values selected were similar to the real data. Its analyses are expressed in the sub-section below. The simulated ordinal datasets are generated from NB distribution with C categories generated at four visits, j = 1,...4. We simulated 1000 datasets with 500 individuals and covariates with 3 ordinal category response.

$$logit[P(Y_{ij}^{*} \leq c)] = \beta_0 + x'\beta \tag{11}$$

The underlying latent variable (y^*) was assumed to be related to the observed response through the 'threshold concept' as indicated by the ordinal regression model. Three covariates were simulated: x_1 from Bernoulli distribution with a probability of success equal to 0.5, x_2 from a uniform distribution, and x_3 is a four-point assessment time. The following assigned parameters were used for simulation: β_0 = 0.1199, β_1 = 0.3740, β_2 = 0.0214 β_3 = 0.0506. No interaction terms were used in our simulation study. Proportional odd logistic regression with a random intercept for the response can be written as follows:

$$logit\left[P\left(Y_{ij}^{*} \leq c\right)\right] = \beta_0 + 0.374_{sex} + 0.0214_{time} + 0.0506_{age} \tag{12}$$

The inversion of the logit link function leads to the conditional ordinal logistic regression model whose equation can be written as:

$$P(Y_{ij}^{*} \leq c) = \frac{exp(\beta_0 + x'\beta)}{1 + exp(\beta_0 + x'\beta)} \tag{13}$$

Let $\emptyset_{ijc} = P\left(Y_{ij}^{*} \leq c\right)$, then ordinal outcome Y_{ij} (e.g., for C = 3) was obtained by setting the observation rule is defined as:

$$Y = \begin{cases} 0, & \text{if } \phi_{ij} \leq \tau_1, \\ 1, & \text{if } \tau_1 < \phi_{ij} \leq \tau_2, \\ 2, & \text{if } \phi_{ij} > \tau_2. \end{cases} \tag{14}$$

We imposed missing values on the full datasets, after which parameter estimates and standard errors were estimated by a likelihood-based approach. Here, we assume MAR mechanism in the missingness model, where individuals whose outcome was greater than some cut-off point would miss at post-baseline time points 2,3, and 4. Let dropout= $y_{ij} - y_{ij-1}$, j = 2,3,4, producing number ranging from 0 and 2; 0 and 3; 0 and 4 for the ordinal outcome, that is C= 3 categories. We ensured that approximately 45% of the outcome was missing, but ρ = 0.2, and an alpha level of 0.05 was used as the empirical power of proportion. The dropping value probability was independent of the immediate history. In the PROC MI procedure, we indicated missing entries, then applied the FCS method thereafter. Ordinal values imputation was continuous initially and later rounded off to selected categories, then applied the MVNI method. The minimum and maximum values specification were necessary to create imputation values that fall within the response. The "PROC MI" in FCS was used to impute ordinal response and analyzed using the ordinal logistic regression model as the analysis model. For each study, default MCMC and FCS specifications were applied in the simulation studies. The algorithms convergence produced the right posterior distribution and we obtained confidence that the imputed datasets were independent. The use of all covariates was to ensure that the imputation model was strong and satisfied the congeniality requirements under MAR assumption. An increased imputation number may be required to compare methods [32]. We had m = 10 imputations to account for a relatively large fraction of missing data and to reduce the loss of power. However, researchers argued that valid estimates can be obtained when m is $3 \leq m \leq 5$ while some cautioned that pegging it at this range may not be necessary [24]. Furthermore, the values diminish quickly after the first m=2 or m=5 for small and large missing data respectively. The idea of selecting m was suggested by [31] and recommends to be at least equal to the percentage of missing data.

In methods of performance-based evaluation, parameter estimates and standard errors are used for real data; bias and mean squared error (MSE) are used for the simulation studies. Bias is defined as the difference between average parameter estimates and true values.

3.2. Results of the Simulation Study

In Table 1, we have the simulation studies' results. It shows the parameter recovery (i.e.bias) for the NB distribution simulation study with a 3-category response.

The models srecover from random fixed values as they indicate small biases, except β_0 (0.284) which presents the worst-case scenario under DL. In the MSE, the worst performer was mixed-effect proportional odds. For β_1, MI-GEE was the best performer as it produced the smallest bias (-0.5431) and mixed-effect produced the smallest MSE (0.0544) while under β_2 and β_3 Mixed-effect was the best performer as it produced the smallest bias (-0.0073) and (-0.0263) and ONB produced smallest MSE (0.0002) and mixed-effect (0.0007) respectively. Mixed-effects proportional odd is subject-specific, and the results observed were low because of the inclusion of random effects in the model, which are excluded in the MI-GEE and ONB because they are marginal models. Furthermore, a different logit link was applied. The MIGEE used cumulative logit link, while the non-linear systematic (mixed-effects proportional odd and ONB) components used adjacent categories. In most cases, direct likelihood is suitable for analyzing continuous data. All of these are responsible for the difference observed in the results.

4. Axial Emphysema Lung-HIV Data

The dataset for this analysis was from the national heart, lung, and blood institute (NHLBI) longitudinal studies of HIV-associated lung infections and complications (Lung HIV), [4]. The cohort study experimented eight different clinical sites located in South Africa and the United States of America respectively. The response factor is "The best description of the axial emphysema distribution" and expressed in this perspective 0='normal predominantly across individual CT images', 1='peripheral/subpleural predominantly across individual CT images', 2='central/axial predominantly across individual CT images', 3='evenly distributed (central and peripheral) predominantly across individual CT images'.

Before the study was conducted the participants' written consent was sought. The data consist of 325 patients between the ages 19 and 77 who were followed up for 24 months (4 visits), axial emphysema distribution of each participant was graded, with the highest indicating a worst-case scenario. About 47% description of the axial emphysema distribution was missing. The summary of the data is presented in Table 2.

4.1. Results of Axial Emphysema Lung-HIV Data

Next, we illustrate the usefulness of the models using the actual empirical dataset. The data was used to assess the prevalence of pulmonary diseases amongst HIV-infected persons, and the outcome of interest was a three-category ordinal response for five years. Single process models are negative binomial (NB), ONB, linear models, proportional odds, and DL. Thus, dual process models are zero-inflated negative binomial (ZINB), ordinal zero-inflated negative binomial (OZINB), and zero-inflated proportional odd models as stated by [16], but were not part of our study. Testing prevalence among infected participants some of the single process models were applied.

Table 1. Bias and mean squared error of the substantive models after pooling the ordinal outcomes using direct likelihood (D)L, ordinal negative binomial (ONB), mixed-effects proportional odd, and MI-GEE methods

	DL		ONB		Mixed-effects		MI-GEE	
Par	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
β_0	0.2840	0.0880	-0.9562	0.9143	-0.9964	1.0312	0.7480	0.6008
β_1	-0.4491	0.2040	-0.2819	0.0795	-0.2038	0.0544	-0.5431	0.3076
β_2	-0.0153	0.0007	-0.0139	0.0002	-0.0073	0.0029	-0.0352	0.0004
β_3	-0.0398	0.0016	-0.0371	0.0014	-0.0263	0.0007	-0.0747	0.0056

Table 2. Descriptive statistics of the incomplete Lung HIV data

Lung-HIV data	Description	Range	%
Baseline variables Sex	1=Male, 2=Female	1-2	0
Age	Age of patient at enrollment	19-77	0
Time	Number of visits	1-4	0
Response variable Axi emph	Description of axial emphysema distribution	0-3	47

Note: Data missing on the response

Table 3. Parameter estimates(Est),standard errors(SE),confidence limits(CL) obtained from the lung HIV data under the methods of direct likelihood(DL),ordinal negative binomial (ONB),mixed-effects proportional odd and multiple imputation based GEE (MI-GEE) under MAR mechanism

	DL			ONB			Mixed-effects			MI-GEE		
Param	Est	SE	*Pr*> |t|	Est	SE	*Pr*> |t|	Est	SE	*Pr*> |t|	Est	SE	*Pr*> |t|
Intercept	1.2690	0.2089	<.0001	-0.1190	0.1548	0.4445	0.1436	0.2595	0.5834	-3.0124	0.4462	
Intercept	----------			-1.7815	0.3429							
Intercept	----------			-0.7855	0.3220	0.0225						
Intercept	----------			-0.8897	0.3411	0.0166						
Sex	-0.3693	0.1514	0.0263	-0.3740	0.1517	0.0246	0.7529	0.2523	0.0075	-0.7497	0.2497	0.0081
Time	-0.0194	0.0400	0.6316	-0.0214	0.0431	0.6236	0.0758	0.0746	0.3199	0.0760	0.0743	0.3184
Age	-0.0472	0.0460	0.3105	-0.0506	0.0488	0.3050	-0.0581	0.0747	0.4387	-0.0574	0.0738	0.4417

Notes: The missing values on response variable

Table 3 displays the results of DL, ONB, mixed-effects proportional odd, and MI-GEE fitted to the data. We discuss the performance of the models in terms of parameter estimates and standard errors respectively. Comparing the methods suitable for analyzing ordinal data, ONB produces the smallest intercept and for other covariates, except sex under MI-GEE. DL may not be compared directly with ONB or other methods because it treated the ordinal response as a continuous variable, but is suitable for multiple imputations. Furthermore, the smallest standard errors were observed under ONB as compared to other methods. Importantly, the performance of ONB under the simulation study is virtually the same under the empirical data. Consider the estimates for sex, in which DL (-0.37), ONB model (0.37), mixed-effects proportional odd model (0.75), and MI-GEE models (-0.75). The interpretation may be different depending on the model of interest. In the case of ONB, the log of the expected pulmonary disease prevalence is 0.37 units larger in HIV-infected males than in HIV infected females. The incidence rate ratio is another way of interpreting the models. Choose the ONB model, HIV-infected males have increased pulmonary disease occurrence in almost half of the study participants (eg., $e^{.37}$=1.45). Conversely, mixed-effects proportional odd may also be interpreted in terms of cumulative logit or log odd. Thus, the expected log odds of having a pulmonary disease in terms of the age of HIV-infected males is -0.05 units lower than the age of HIV-infected females.

5. Discussion

We introduced a novel framework for the analysis model of multiple imputations for DL, ONB, mixed-effects proportional odds, and MI-GEE methods. But each method followed different computation procedures. In the case of GEE; the marginal model relies on covariates and not on random effects. To capture random effects; the mixed-effects proportional odd model uses the NLMIXED procedure. Thus, ONB uses the NLMIXED procedure but is slightly different from the former because random effects are excluded. We used the cumulative distribution function (PDF) for NB distribution which is the sum of probability mass function (PMF), i.e we expressed the explicit linkage of ordinal responses to the latent variable of interest through cumulative probabilities.

In our simulation study, the disparity between the performance of ONB and other results is small. This is an indication that the latent variable of interest follows the NB distribution and is not skewed. Thus, ONB offers a quantitative and substantive advantage when used to analyze ordinal intermittent missing observations. However, the empirical data results indicate all of the methods perform creditably well like the simulation, but ONB is seen as a strong statistical tool that handles ordinal longitudinal intermittent missing values very well.

However, the study is an extension to the one conducted by [12] where missing data occur on the response variable, but in our situation covariate and response had missing data. Furthermore, in the study conducted in [16], missing data and the use of different cut-off points were not entertained. Missing data were introduced in our simulation studies, and methods to handle them included different cut-off points. Some of the techniques herewith are an extension to research in [16]. This has no limitation on the outcomes of our study, but further studies are advised to be carried investigate how ordinal-count models will perform under misspecification of latent outcome variable distribution, and alternate generating distributions.

REFERENCES

[1] Agresti A, (2010) "Analysis of ordinal categorical data," Second Edition: Wiley Series in Probability and Statistics, 2010, pp. 262-280.

[2] Bauer D. J., Sterba S. K, "Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation," *Psychol Methods*, vol. 16 no. 4, pp.373390, 2011.

[3] Bock R, Jones L., "The measurement and prediction of judgment and choice." San Francisco: Holden-Day, pp. 354-370, 1968.

[4] Bruce T, Zhaoyu L, Arionna S., "Longitudinal studies of HIV-associated lung infections and complications (Lung HIV)," 2013.

[5] Carpenter J. R., Kenward M. G, "Multiple imputation and its application," Wiley, Chichester, 2013.

[6] Choi K. H., Hoff C, Gregorich S. E., Grinstead O, Gomez C, Hussey W, "The efficacy of female condom skills training in HIV risk reduction among women: A randomized controlled trial. *American Journal of Public Health*, vol. 98, no. 10, pp. 1841-1848, 2008. DOI: 10.2105/AJPH.2007.113050

[7] Demirtas H., Hedeker D, "An imputation strategy for incomplete longitudinal ordinal data," *Statistics in Medicine*, vol 27 pp. 4086-4093, 2008. DOI: 10.1002/sim.3239

[8] Donneau A. F., Mauer M, Lambert P, Molenberghs G, Albert A, "Simulation-based study comparing multiple imputation methods for non-monotone missing ordinal data in longitudinal settings," *Journal of Biopharmaucetical Statistical*, vol. 25, no. 3, pp. 560-601, 2015.

[9] Gameroff M. J., "Using the proportional odds model for health-related outcomes: Why, when, and how with various SAS procedures," In SUGI 30 (2005, April), pp. 205-230.

[10] Goodman L. A, "Simple models for the analysis of association in cross-classifications having ordered categories," *Journal of American Statistics Association*, vol. 74, no. 367, pp. 537-552, 1979.

[11] Hedeker D, "A mixed-effects multinomial logistic regression model," *Statistics in Medicine*, vol. 22, pp. 1433-1446, 2003.

[12] Kombo A. Y., Mwambi H., Molenberghs G, "Multiple imputation for ordinal longitudinal data with monotone missing data patterns," *Journal of Applied Statistics*, vol. 44, no. 2, pp. 270-287, 2016. DOI:10.1080/02664763.2016.1168370

[13] Lipsitz S. R., Kim k., and Zhao L, "Analysis of repeated categorical data using generalized estimating equations," *Statistics in Medicine* 13(11):1149-1163, 1994. DOI: 10.1002/sim.4780131106

[14] Little R. J. A., Rubin D. B, "Statistical analysis with missing data" (2nd edition), New York: John Wiley and Sons, 2002.

[15] Mallinckrodt CH, Clark SWS, Carroll RJ, and Molenberghs G "Assessing response profiles from incomplete longitudinal clinical trial data undeer regulatory considerations." *Journal of Biopharmaceutical Statistics*,13, 179-190, 2003.

[16] McGinley J. S., Curran P.J., Hedeker D, "A novel modeling framework for ordinal data defined by collapsed counts," *Statistics in Medicine*, vol. 34, pp. 2312-2324, 2015.

[17] Molenberghs G. and Kenward M. G, "Missing data in clinical studies". Hoboken, NJ: Wiley 2007.

[18] Molenberghs G., Verbeke G, (2005) Models for discrete longitudinal data, Springer, New York, USA, 2005, pp. 35-43.

[19] Moreno U., Mauro G, "A new parsimonious model for ordinal longitudinal data with application to subjective evaluations of a gastrointestinal disease," *Statistical Methods in Medical Research*, 0(0):pp. 1-18 2016. DOI: 10.1177/0962280216661370

[20] McCullagh P, "Regression models for ordinal data (with discussion)," *Journal of Royal Statistical Society*, Series B (Methodological), vol. 42, pp. 109-142, 1980.

[21] Ratitch B., Lipkovich I., O'kelly M, "Combining analysis results from multiply imputed categorical data," pharmaSUG 2013-Paper SP03, pp. 1-10, 2013.

[22] Rubin D. B, (1987) Multiple imputation for nonresponse in surveys, Wiley, New York USA, 1987, pp. 15-22.

[23] Rubin D. B., Schafer J.L, "Efficiently creating multiple imputations for incomplete multivariate normal data in proceedings of the statistical computing section, American Statistical Association, Alexandria, VA, USA. p.83-88, 1990.

[24] Schafer J. L, "Analysis of incomplete multivariate data," Chapman and Hall, New York, USA, 1997

[25] Seitzman R.L., Mahajan V.B., Mangione C.,. Cauley J. A., Ensrud K. E., Stone K. L., Cummings S. R., Hochberg M. C., Hillier T. A., Sinsheimer J. S., Yu F., Coleman A. L. (2008) "Estrogrn receptor alpha and matrix metalloproteinase 2 polymorphisms and age-related maculopathy in older women," *American Journal Epidemiology*, vol. 167, pp. 1217-1225, 2008.

[26] Snell E. A, "Scaling procedure for ordered categorical data," *Biometrics*, vol. 20, pp. 592-607, 1964.

[27] van Buuren S, (2012) Flexible imputation of missing data, Boca Raton, FL, CRC Press.

[28] van Buuren S, "Multiple imputation of discrete and continuous data by fully conditional specification," *Statistical Methods in Medical Research*, vol. 16, pp. 219-242, 2007.

[29] van Buuren S., Brand J. P. L., Groothuis-Oudshoorn K., Rubin D.B, "Fully conditional specification in multivariate imputation," *Journal of Statistical Computation and Simulation*, vol. 77, pp. 1049-1064, 2006.

[30] van Buuren S., Groothuis-Oudshoorn K. "MICE: Multivariate imputation by chained equations in R," *Journal of Statistical Software*,vol. 45 no. 3, 2011.

[31] White I. R., Royston P., Wood A. M, "Multiple imputation using chained equations: Issues and guidance for practice," *Statistics in Medicine*, vol. 30, pp. 377-399, 2011.

[32] Wood A. M., White I.R., Hillsdon M., Carpenter J, "Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes," *International Journal of Epidemiology*, vol. 34, pp. 89-99, 2005.