HHS Public Access

Author manuscript

J Appl Stat. Author manuscript; available in PMC 2016 October 01.

Published in final edited form as:

J Appl Stat. 2015 October 1; 42(10): 2203–2219. doi:10.1080/02664763.2015.1023270.

Distribution-free Inference of Zero-inated Binomial Data for Longitudinal Studies

H. He[a], **W. J. Wang**[a], **J. Hu**[a,b,*], **R. Gallop**[c], **P. Crits-Christoph**[d], and **Y. L. Xia**[a]

[a]Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA

[b]College of Basic Science and Information Engineering, Yunnan Agricultural University, Kunming, Yunnan, China, 650201

[c]Department of Mathematics and Applied Statistics, West Chester University, West Chester, PA 19383, USA

[d]Department of Psychiatry, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

Count reponses with structural zeros are very common in medical and psychosocial research, especially in alcohol and HIV research, and the zero-inflated poisson (ZIP) and zero-inflated negative binomial (ZINB) models are widely used for modeling such outcomes. However, as alcohol drinking outcomes such as days of drinkings are counts within a given period, their distributions are bounded above by an upper limit (total days in the period) and thus inherently follow a binomial or zero-inflated binomial (ZIB) distribution, rather than a Poisson or zero-inflated Poisson (ZIP) distribution, in the presence of structural zeros. In this paper, we develop a new semiparametric approach for modeling zero-inflated binomial (ZIB)-like count responses for cross-sectional as well as longitudinal data. We illustrate this approach with both simulated and real study data.

Keywords

Bounded count response; COMBINE Study; Distribution-free models; Generalized Estimating Equations; Structural zero; Zero-inated binomial (ZIB)

1. Introduction

The issue of structural zeros has drawn a considerable amount of attention during the last decade [7, 8, 11, 20, 23, 26, 29]. Structural zeros refer to zero responses from those subjects whose count responses will always be zero, i.e., constant zeros, in contrast to random (or sampling) zeros that occur to subjects whose count response can be greater than zero, but appear to be zero due to sampling variability. In regression analysis, the zero-inflated

*Corresponding author. hududu@ynau.edu.en.

Appendix

See the Web-based Supplementary Materials.

Poisson (ZIP) has become a popular approach for addressing structural zeros for such count responses [7, 8, 11, 23, 29]. But for some count responses such as the number of days of alcohol drinking within a given period, the range of the variable is bounded above. For example, in the NIH-funded Combined Pharmacotherapies and Behavior Interventions (COMBINE) Study [2], a large randomized trial that combines pharmacotherapies and behavioral interventions, among the primary outcomes are days of drinking (DAD) and days of heavy drinking (DHD) of alcohol during a given period. These drinking outcomes measure the number of days when an individual consumes a certain amount of alcohol within a given period such as a week and thus are bounded from above by an upper limit (total days for the period). Although popular for modelling count responses, Poisson-based approaches are not appropriate for modeling such bounded binomial-based count responses, at least when the range is small. This naturally calls for zero-inflated binomial (ZIB) models.

Most available methods for ZIB follow the parametric approach [8, 26]. However, such an approach has only limited applications because of the strong assumption about the data distribution. Semi-parametric, or distribution-free, models are more robust to misspecification than parametric models. The method of estimating equations, assuming only the conditional mean response, is a popular semi-parametric alternative. For longitudinal studies, the generalized estimating equations (GEE), a generalized version of the method of estimating equations, is commonly used to address correlation among repeated responses. However, since ZIB is a mixture of two distributions, we will not be able to identify the model parameters by simply modeling the mean response [4, 7]. Hall and Zhang [7] developed an approach for zero-inflated Poisson and binomial data by integrating maximum likelihood with GEE to deal with correlated longitudinal responses. This “hybrid” approach by-passes the distribution assumptions about the correlation as in the other methods. However, as it still employs parametric models for the marginal distribution of the response, this approach is sensitive to deviations from the assumed marginal distributions. More importantly, their method does not deal with missing values, a common issue in longitudinal studies such as the COMBINE study. Dobbie and Welsh [5] also developed a GEE approach for zero-inflated count data. However, their approach models the mixture of zeros and truncated Poisson, rather than a mixture of structural zeros and Poisson. As a result, structural zeros are not distinguished from random zeros, failing to provide inference about the likelihood of structural zeros, which is of great interest in practice. Furthermore, like the approach by Hall and Zhang [7], their method does not address missing values either.

In this article, we develop a new semi-parametric approach for modeling zero-inflated count responses bounded above by an upper limit. The new approach not only can handle missing data problem, but also is more robust for deviations from the assumed marginal distributions. Compared to [7, 8], the new approach is less computational intensive and can be used as a benchmark to evaluate the performance of the approaches that either have been used for analyzing such responses in the alcohol studies such as linear regression model (applied to the transformed count response) or existing alternatives such as the zero-inflated Poisson (ZIP) and Hall’s mixed effect and marginal ZIB models [7, 8]. The remainder of this paper is organized as follows. In Section 2, we introduce the notation and ZIB-like models for cross-sectional outcomes. In Section 3, we generalize the development of Section

2 to longitudinal data analysis in the complete data case. The extension of the approach to missing data is discussed in Section 4. In Section 5, we assess performance of the proposed models by simulation studies, while in Section 6, we apply the approach to drinking outcomes from the COMBINE. The paper concludes with a discussion in Section 7.

2. ZIB-like Models for Cross-sectional Data

Let y_i denote a count response and $\mathbf{x}_i$ a set of explanatory variables from the ith subject ($1 \le i \le n$). When $y_i \mid \mathbf{x}_i$ follows a binomial distribution, generalized linear models with a link function for binary responses are often applied. In other words, the probability of success (or fail) for each trial of the binomial, after a transformation, is a linear function of the explanatory variables $\mathbf{x}_i$. For example, if the popular logistic link function is used, we can assume

$$y_i|\mathbf{x}_i \sim \text{Binomial}(m_i, p_i), \text{logit}(p_i) = \beta^\top \mathbf{x}_i, \quad (1)$$

) where m_i is the size of the binomial sample for the ith subject, β is the vector of parameters and the logit transformation of the probability p_i is modeled as a linear function of $\mathbf{x}_i$. In the presence of structural zeros, the model in (1) becomes in-appropriate. First, conceptually binomial is not the correct distribution. Formally, it is straightforward to check that the variance of the response will be more than the variance expected under a binomial distribution. Thus, if a binomial model is applied, one may face overdispersion. More important, the mean of the binomial in (1) modeled is no longer the mean of the mixture distribution in the presence of structural zeros. Thus, although some ad-hoc techniques are available to correct overdispersion, such as sandwich variance estimates, none applies to correct the flaw in the mean when using (1) to model such count responses.

The zero-inflated binomial (ZIB) model acknowledges the mixture distribution in our context defined by the binary response and the indicator of structural zeros. For simplicity of notation, we use the logistic link function for modeling the binary response throughout the paper. However, the same considerations apply to other link functions for modeling binary outcomes such as probit and complementary log-log.

Let ZIB(m_i, p_i, ρ_i) denote a zero-inflated binomial distribution, with ρ_i denoting the probability of structural zero and m_i (p_i) denoting the size (mean) of the binomial sample. Assume $y_i \mid \mathbf{x}_i$ follows a ZIB and generalized logit models are applied to both the structural zero and the binomial components, we may write the ZIB model as:

$$y_i|\mathbf{x}_i \sim \text{i.d.ZIB}(m_i, p_i, \rho_i), \text{logit}(\rho_i|\mathbf{x}_i) = \mathbf{u}_i^\top \beta_\mathbf{u}, \text{logit}(p_i|\mathbf{x}_i) = \mathbf{v}_i^\top \beta_\mathbf{v}, 1 \le i \le n, \quad (2)$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are two subset of $\mathbf{x}_i$ (not necessarily disjoint). The ZIB(m_i, p_i, ρ_i) in (2) has the following distribution function:

$$f_{ZIB}(y_i|m_i, p_i, \rho_i) = \begin{cases} \rho_i + (1-\rho_i)(1-p_i)^{m_i} & \text{if} y_i = 0 \\ (1-\rho_i)\binom{m_i}{y_i} p_i^{y_i}(1-p_i)^{m_i - y_i} & \text{if} y_i > 0 \end{cases}, \quad (3)$$

where the binomial probability at 0, $(1-p_i)^{m_i}$, is modified by $\rho_i+(1-\rho_i)(1-\rho_i)^{m_i}$ to account for the presence of structural zeros. The maximum likelihood estimate may be applied for inference about the model parameters, $\beta=\left(\beta_{\mathbf{u}}^{\top},\beta_{\mathbf{v}}^{\top}\right)^{\top}$. However, parametric approaches are very restricted in the sense that the true distribution may deviate from the assumed models, invaliding the inference. For example, out-comes measuring cumulative incidence over a period of time from the same subject, such as alcohol consumption per day over a week, are generally correlated and as a result may not follow the binomial distribution. Thus, models based on weaker assumptions, such as semiparametric approaches that assume no exact distribution, but rather some aspects of the distribution such as the mean, provide robust for a wider class of data distributions.

The estimating equations (EE) based on conditional mean responses from the generalized linear models (GLM) are commonly used semiparametric alternatives to parametric models. However, existing EEs are mostly for modelling a single mean response and as such are not sufficient to identify the parameters of mixture distributions within the current context. Thus, we model both the probability of structural zero and the mean of a zero-truncated binomial:

$$h_{1i}(\mathbf{x}_i)=\Pr(\mathbf{y}_i=0|\mathbf{x}_i)=\rho_i+(1-\rho_i)(1-p_i)^{m_i}, \quad (4)$$

$$h_{2i}(\mathbf{x}_i)=E(\mathbf{y}_i>0|\mathbf{x}_i)=(1-\rho_i)m_i p_i,$$

$$\rho_i=\frac{\exp(\mathbf{u}_i^{\top}\beta_{\mathbf{u}})}{1+\exp(\mathbf{u}_i^{\mathsf{T}}\beta_{\mathbf{u}})},p_i=\frac{\exp(\mathbf{v}_i^{\top}\beta_{\mathbf{v}})}{1+\exp(\mathbf{v}_i^{\mathsf{T}}\beta_{\mathbf{v}})}.$$

The means, $h_{1i}(\mathbf{x}_i)$ and $h_{2i}(\mathbf{x}_i)$, are derived based on the ZIB model (2), and thus, the ZIB in (2) implies (4). However, since there is no assumption about the exact distribution model in (4), this ZIB-like model yields more robust inference. The mean $h_{1i}(\mathbf{x}_i)$ of a truncated version of (2) at 0 is used in (4) to enable inference based on observed data. Note that the use of the truncated binomial distribution in (4) for $y_i > 0 \mid x_i$ bears some resemblance in theory with the approach of [5], as the latter also used truncated Poisson distributions for the positive response. However, the essential difference is that their approach models the mixture of zeros and truncated (positive) count responses and hence structural and random zeros are not distinguished. In contrast, by modeling the structural zero specifically, (4) has the ability to distinguish structural from random zeros. Note that the membership of the mixture components (zero or positive responses) in the approach of [5] is known, because it models the observed zeros, and thus it is really a hurdle model [24]. However, the membership of the mixture components (structural zeros or not) is unknown and thus (4) truly represents a mixture model for a mixed population consisting the at-risk (defined by random zero and positive responses) and non-risk (defined by structural zeros) subgroups.

The semiparametric, or distribution-free, ZIB-like model in (4) is a model for a 2-dimensional response vector, rather than a single response as in a classic GLM or EE model.

Methods for restricted moment models may be applied for the inference [1, 9, 18, 25]. Let $I(\cdot)$ denotes a set of indicator functions, $s_{1i} = I(\mathbf{y}_i = 0) - E[I(\mathbf{y}_i = 0) \mid \mathbf{x}_i] = I(\mathbf{y}_i = 0) - \rho_i - (1 - \rho_i)(1 - p_i)^{m_i}$, $s_{2i} = I(\mathbf{y}_i > 0)(\mathbf{y}_i - E[\mathbf{y}_i \mid \mathbf{y}_i > 0, \mathbf{x}_i]) = I(\mathbf{y}_i > 0)(\mathbf{y}_i - m_i p_i / [1 - (1 - p_i)^{m_i}])$, and $S_i = (s_{1i}, s_{2i})$. We have

Lemma 2.1

The variance matrix of the response function S_i is given by

$$Var(s_{1i}) = [\rho_i + (1-\rho_i)(1-p_i)^{m_i}][(1-\rho_i)(1-(1-p_i)^{m_i})]$$

$$Var(s_{2i}) = (1-\rho_i)(1-(1-p_i)^{m_i}) \left[\frac{m_i p_i + m_i(m_i-1)p_i^2}{1-(1-p_i)^{m_i}} - \left(\frac{m_i p_i}{1-(1-p_i)^{m_i}} \right)^2 \right]$$

$Cov(s_{1i}, s_{2i}) = 0$.

A proof of the lemma is given in the Web Appedix A.

Within the current context, the ZIB-like model in (4) can be estimated by solving the following EE:

$$\mathbf{W}_n(\beta) = \sum_{i=1}^{n} D_i V_i^{-1} S_i = \mathbf{0}, \quad (5)$$

where $D_i = \frac{\partial}{\partial \beta} S_i = \begin{pmatrix} \frac{\partial}{\partial \beta_{\mathbf{u}}} S_{1i} & \frac{\partial}{\partial \beta_{\mathbf{u}}} S_{2i} \\ \frac{\partial}{\partial \beta_{\mathbf{v}}} S_{1i} & \frac{\partial}{\partial \beta_{\mathbf{v}}} S_{2i} \end{pmatrix}$ and $V_i = Var(S_i)$ given in the above lemma.

It is straightforward to show that the EE in (5) is unbiased and thus provides asymptotically consistent and normally distributed estimates. We summarize the properties of the estimate in the following theorem for ease of reference.

Theorem 2.2

Under the assumption of model (4), estimating equation (5) is un-biased, and hence the estimate $\hat{\beta}$, obtained by solving the estimating equation (5) is consistent. Furthermore, as $n \to \infty$ we have

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \to AN(0, \Sigma), \Sigma = B^{-1} E\left(D_i V_i^{-1} S_i S_i^{\top} V_i^{-1} D_i^{\top}\right) B^{-\top}, B = E\left(D_i V_i^{-1} D_i^{\top}\right).$$

A consistent estimate of the asymptotic variance Σ is readily constructed by substituting consistent estimates of the respective quantities defining Σ, i.e.,

$$\hat{\Sigma} = \hat{B}^{-1}\left(\frac{1}{n-1}\sum_{i=1}^{n} \hat{D}_i \hat{V}_i^{-1} \hat{S}_i \hat{S}_i^{\top} \hat{V}_i^{-1} \hat{D}_i^{\top}\right) \hat{B}^{-\top}, \hat{B} = \frac{1}{n-1}\sum_{i=1}^{n} \hat{D}_i \hat{V}_i^{-1} \hat{D}_i^{\top},$$

where $\hat{V}_i$ *and* $\hat{D}_i$ *are* V_i *and* D_i *with* $\hat{\beta}$ *substituting in place of* β.

A proof of the theorem is given in the Web Appendix B.

Remarks—Note that V_i in (5) can be any invertible matrix function of x_i, although the commonly use of $V_i = V ar\ (S_i)$ provides the most efficient estimates when the data does follow the ZIB distribution. However, regardless of the choice of working correlation for V_i, inference based on (5) is valid as long as the specification in (4), an assumption weaker than the assumption of ZIB, is true.

3. ZIB-like Models for Longitudinal Data

Now consider longitudinal studies. Suppose there are l assessment times. For notational brevity, assume assessment times are fixed a priori. Let y_{it} and $\mathbf{x}_{it}$ be the outcome and the explanatory variables from time t and $\mathbf{y}_i = (y_{i1}, \ldots, y_{il})^\top$ and $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{il})^\top$ be the outcome and the explanatory variables across the time points for the ith subject ($1 \le i \le n$). Both models discussed in the preceding section can be extended to longitudinal data. For the parametric ZIB in (2), the extension can be achieved through random effects, while the semiparametric ZIB-like model in (4) can be extended to longitudinal data by combining the marginal models for each of the assessment points as follows:

$$h_{1it}(\mathbf{x}_i) = E[\mathbf{I}(\mathbf{y}_{it}=0)|\mathbf{x}_{it}] = \rho_{it} + (1-\rho_{it})(1-p_{it})^{m_{it}}, \quad (6)$$

$$h_{2it}(\mathbf{x}_{it}) = E[\mathbf{y}_{it}|\mathbf{y}_{it}>0, \mathbf{x}_{it}] = \frac{m_{it}p_{it}}{1-(1-p_{it})^{m_{it}}},$$

$$\rho_{it} = \frac{\exp(\mathbf{u}_{it}^\top \beta_{\mathbf{u}})}{1+\exp(\mathbf{u}_{it}^T \beta_{\mathbf{u}})}, p_{it} = \frac{\exp(\mathbf{v}_{it}^\top \beta_{\mathbf{v}})}{1+\exp(\mathbf{v}_{it}^T \beta_{\mathbf{v}})}, 1 \le t \le l,$$

where m_{it} is the size of the binomial at time t. Let $S_i = (S_{i1}, \ldots, S_{il})$, and

$$S_{it} = (s_{1it}, s_{2it})^\top = (I(\mathbf{y}_{it}=0) - E[I(\mathbf{y}_{it}=0)|\mathbf{x}_{it}], I(\mathbf{y}_{it}>0)(\mathbf{y}_{it} - E[\mathbf{y}_{it}|\mathbf{y}_{it}>0, \mathbf{x}_{it}]))^\top.$$

The semiparametric ZIB-like model above is also a restricted moment model, with more complicated functional responses than the model in (4), and thus the general theory of restricted moment models applies. An important difference is that the variance-covariance structure of the response functions in the longitudinal study case is more complicated. In fact, it may not be practical to model the true variance structure. Liang and Zeger [12] suggested the working correlation approach which first estimates a correlation for the estimating equation. If the working correlation captures the true variation, then it will provide optimal efficiency. Even though it is misspecified, the estimates are still consistent.

More precisely, we define the generalized estimating equations in similar forms as that in (5), but with S_i defined in (6) and D_i and V_i modified as follows. Let

$$\beta=\left(\beta_{\mathbf{u}}^{\top},\beta_{\mathbf{v}}^{\top}\right)^{\top}, D_{it}=\frac{\partial}{\partial\beta}s_{it}=\begin{pmatrix}\frac{\partial}{\partial\beta_{\mathbf{u}}}s_{1it} & \frac{\partial}{\partial\beta_{\mathbf{u}}}s_{2it}\\ \frac{\partial}{\partial\beta_{\mathbf{v}}}s_{1it} & \frac{\partial}{\partial\beta_{\mathbf{v}}}s_{2it}\end{pmatrix}, \text{and} D_i=(D_{i1},\ldots,D_{il})^{\top}. \quad (7)$$

To define V_i, we assume the following working correlation model:

$$V_i=A_i^{\frac{1}{2}}R(\alpha)A_i^{\frac{1}{2}}, A_i=diag_t(A_{it}), \quad (8)$$

$$A_{it}=\begin{pmatrix}Var(s_{1it}|\mathbf{x}_{it}) & 0\\ 0 & Var(s_{2it}|\mathbf{x}_{it})\end{pmatrix}$$

The working correlation $R(\alpha)$ is often selected as independent or exchangeable for convenience. We may also specify some other forms for a specific study. The exact form selected for $R(\alpha)$ does not affect the consistency of the estimates, but rather the efficiency. For the GEE model, one may need to estimate the parameters in the working correlation $R(\alpha)$ first. Similar to (5), the GEE estimates are consistent as the estimating equations are unbiased. Then we have similar nice asymptotic properties, and we summarize them in the following theorem.

Theorem 3.1

Let $\hat{\beta}$ denote the estimator of β obtained by solving the GEE above and $\hat{\alpha}$ the estimator of α. Under some mild regularity conditions and $\hat{\alpha}$ is $\sqrt{n}-consistent$, the estimator $\hat{\beta}$ is consistent and $\sqrt{n}\left(\hat{\beta}-\beta\right)$ is asymptotically normally distributed with zero mean and covariance matrix:

$$\Sigma_{\beta}=B^{-1}E[(D_iV_i^{-1}S_i)(D_iV_i^{-1}S_i)^{\mathrm{T}}]B^{-\mathrm{T}}, B=E(D_iV_i^{-1}D_i^T)$$

A consistent estimator of Σ_{β} is given by $\hat{\sum}_{\beta}=\hat{B}_0^{-1}\hat{B}_1\hat{B}_0^{-1}$, where

$$\hat{B}_1=\frac{1}{n}\sum_{i=1}^{n}(\hat{D}_i\hat{V}_i^{-1}\hat{S}_i)(\hat{D}_i\hat{V}_i^{-1}\hat{S}_i)^{\mathrm{T}}, \hat{B}_0=\frac{1}{n}\sum_{i=1}^{n}\hat{D}_i\hat{V}_i^{-1}\hat{D}_i^{\mathrm{T}},$$

$\hat{B}_i$, $\hat{D}_i$, $\hat{S}_i$ and $\hat{V}_i$ denote the corresponding quantities with β and α replaced by $\hat{\beta}$ and $\hat{\alpha}$

Remarks—Hall and Zhang [7] considered an extension of ZIB to the longitudinal setting by modeling the marginal response based on (3) and within-subject correlations using the GEE. Specifically, let r_{it} be an indicator with the value 1 if the *i*th subject at time *t* is from the binomial component of ZIB and 0 otherwise, i.e., r_{it} = 1 if y_{it} > 0. Since structure and random zeros cannot be distinguished from each other, r_{it} is unknown if y_{it} = 0. By modeling the (marginal) relationship between y_{it} and x_{it} using (2), Hall and Zhang [7] used a set of estimating equations to account for correlations among the y_{it}'s. Although similar to the

GEE model in (5), their estimating equations cannot be solved directly since r_{it} is unobserved for $y_{it} = 0$. To address the latent issue for variable r_{it} for $y_{it} = 0$, they developed the Expectation-Solution (ES), an expectation-maximization (EM)-type algorithm, with the E-step computing the expected value of r_{it} and the S-step solving the resulted equations. While sound in theory, this approach is quite problematic in practice due to convergence issues with EM algorithms (see our simulation studies below). As in the cross-sectional case, our approach is based on weaker assumptions and thus is more robust than the ES method of [7]; instead of assuming the parametric ZIB for the marginal distribution, we only assume two specific conditional moments. Further we will extend our method to longitudinal data with missing values below.

Hall and Zhang [7] also mentioned an approach to model both the first- and second-moment to address the lack of information for identifying ZIB using GEE, akin to GEE II [19], and [29] applied the approach for modeling zero-inflated count data. Since such an approach imposes additional assumptions about the variance, it involves the 3rd- and 4th-order moments for inference, creating computational complexity and constraints on the variance of the response. Further, it is difficult to accommodate overdispersion in the binomial component of ZIB, a common occurrence in real studies such as the outcomes of DAD and DHD in COMBINE.

4. Inference under Missing Data

Missing data is inevitable for longitudinal studies. Patients often drop out of study in clinical trials, producing missing values in subsequent visits. The GEE approach above may yield biased estimate, if the missing values are simply excluded from analysis, unless the missing data mechanism follows the very restrictive missing completely at random (MCAR) model [22, 24]. The MCAR assumption implies that the missingness is independent of any other variables (observed or otherwise). Unfortunately, missing data mechanisms in clinical studies often depend on some variables and simply ignoring such a dependence structure as in GEE will in general yield biased estimates. As in the literature, we focus on the missing value in the response and assume the missingness follows a monotone missing data patterns (MMDP), i.e., if a missing response occurs at a time point, all subsequent responses after that time point are also missing [9, 14, 21].

In this section, we consider the missing at random (MAR) model, i.e., the missing data mechanism depends on the variables that are always observed. Within the context of longitudinal data discussed in the preceding section, we define a missing (or rather observed) data indicator for each *i*th subject as follows:

$$\mathbf{r}_i=(r_{i1},\ldots,r_{il})^{\mathsf{T}}, r_{it}=\begin{cases} 1 \text{if} y_{it} \text{is observed} \\ 0 \text{if} y_{it} \text{is missing} \end{cases},$$

where $t = 1, 2, \ldots, l$. We assume no missing data at baseline $t = 1$ such that $r_{i1} = 1$ for all $1 \leq i \leq n$. Under MAR, the missingness of y_{it} is independent of y_{it} given the observed history and covariates, i.e.,

$$r_{it} \perp y_{it} | H_{it^-}, H_{it^-} = (y_{i1}, \ldots, y_{i(t-1)}, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{i(t-1)}), 2 \leq t \leq l. \quad (9)$$

The above condition allows one to integrate the inverse probability weighting (IPW) method with the GEE to provide valid inference for our current ZIB-like model within the current context.

Let

$$\pi_{it} = \Pr(r_{it} = 1 | H_{it^-}), \Delta_{it} = \frac{r_{it}}{\pi_{it}}, \mathbf{\Delta}_i = \mathrm{diag}_t(\Delta_{it}), 2 \leq t \leq l, \quad (10)$$

where $\mathrm{diag}_t(\Delta_{it})$ denotes an $l \times l$ diagonal matrix with Δ_{it} on the tth diagonal.

We can estimate β using the following Weighted GEE (WGEE):

$$\mathbf{w}_n(\beta) = \sum_{i=1}^{n} \mathbf{w}_{ni} = \sum_{i=1}^{n} D_i V_i^{-1} \mathbf{\Delta}_i S_i = \mathbf{0}, \quad (11)$$

where S_i, D_i and V_i are defined the same as in (6), (7) and (8). As in the case for GEE, we need to estimate α in the working correlation $R(\alpha)$ and substitute the estimate in place of α before the WGEE can be solved for β. Also, in practice, the weight function π_{it} in (10) is unknown and need to be modeled. In light of the assumption in (9), we can model the conditional probabilities using models for binary responses such as logistic models. Specifically, let $p_{it} = \Pr(r_{it} = 1 \mid r_{i(t-1)} = 1, H_{it^-})$. If using logistic regression, we may model p_{it} as:

$$\mathrm{logit}(p_{it}(\gamma)) = \gamma_{0t} + \sum_{s=1}^{t-1} \gamma_{xs}^{\top} \mathbf{x}_{is} + \sum_{s=1}^{t-1} \gamma_{ys}^{\top} y_{is}, 2 \leq t \leq l, \quad (12)$$

$$\gamma_t = \left(\gamma_{0t}, \gamma_{x1}^{\mathsf{T}}, \ldots, \gamma_{x(t-1)}^{\mathsf{T}}, \gamma_{y1}^{\mathsf{T}}, \ldots, \gamma_{y(t-1)}^{\mathsf{T}}\right)^{\mathsf{T}}, \gamma = \left(\gamma_2^{\mathsf{T}}, \gamma_3^{\mathsf{T}}, \ldots, \gamma_l^{\mathsf{T}}\right)^{\mathsf{T}}.$$

The parameters γ in (12) is estimated based on the subsample of subjects that were observed at least until time t. We estimate π_{it} by the relationship $\pi_{it}(\gamma) = \prod_{s=1}^{t} p_{is}(\gamma)$.

We may estimate γ using the following estimating equations:

$$\mathbf{Q}_n(\gamma) = \sum_{i=1}^{n} \mathbf{Q}_{ni} = \sum_{i=1}^{n} (\mathbf{Q}_{i2}^T, \mathbf{Q}_{i3}^T, \ldots, \mathbf{Q}_{il}^T) = \mathbf{0}, \quad (13)$$

$$\mathbf{Q}_{it} = \frac{\partial}{\partial \gamma_t} \{ r_{i(t-1)} [r_{it} \log p_{it} + (1 - r_{it}) \log(1 - p_{it})] \}, 2 \leq t \leq l.$$

With estimated π_{it}, we can estimate β based on the following generalizing estimating equations:

$$\mathbf{w}_n(\beta)=\sum_{i=1}^{n}\mathbf{w}_{ni}=\sum_{i=1}^{n}\mathbf{D}_i\mathbf{V}_i^{-1}\hat{\mathbf{\Delta}}_i\mathbf{S}_i=\mathbf{0}, \quad (14)$$

where D_i, V_i and S_i are defined the same as in (6), (7) and (8), and $\hat{\mathbf{\Delta}_i}$ denotes Δ_i in (10) with estimated π_{it}. Again, as in the complete data case, V_i may be a function of α if a working correlation model other than the working independence is used. In this case, we must estimate α first and replace the estimated value of α in (14) before we solve (14) for β.

The WGEE estimate $\hat{\beta}$ based on (14) also has nice asymptotic properties summarized in the following theorem.

Theorem 4.1—*Let $\hat{\beta}$ denote the estimate of β obtained by solving the WGEE in* (14) *and $\hat{\alpha}$ denote some estimate of α. Under some mild regularity conditions and $\hat{\alpha}$ is $\sqrt{n}-consistent$, the estimator $\hat{\beta}$ is consistent and $\sqrt{n}\left(\hat{\beta}-\beta\right)$ is asymptotically normally distributed with zero mean and covariance matrix $\Sigma_\beta = B^{-1}E[\Sigma_U + \Phi]B^{-\mathsf{T}}$, where*

$$\Sigma_U=E[(D_iV_i^{-1}\mathbf{\Delta}_iS_i)(D_iV_i^{-1}\mathbf{\Delta}_iS_i)^{\mathsf{T}}], B=E(D_iV_i^{-1}\mathbf{\Delta}_iD_i^T), \Phi=CH^{-T}C^T-G-G^T,$$

$$C=E\left[\frac{\partial}{\partial\gamma_t}(D_iV_i^{-1}\mathbf{\Delta}_iS_i)\right], H=E\left[\frac{\partial}{\partial\gamma_t}\mathbf{Q}_i\right], G=E[D_iV_i^{-1}\mathbf{\Delta}_iS_i\mathbf{Q}_i^TH^{-T}C^T],$$

$$\mathbf{Q}_i=(\mathbf{Q}_{i2},\mathbf{Q}_{i3},\ldots,\mathbf{Q}_{il}).$$

A consistent estimator of Σ_β is given by replacing Σ_U, B and Φ with the following estimates for the corresponding components:

$$\hat{\Sigma}_U=\frac{1}{n}\sum_{i=1}^{n}(\hat{D}_i\hat{V}_i^{-1}\hat{\mathbf{\Delta}}_i\hat{S}_i)(\hat{D}_i\hat{V}_i^{-1}\hat{\mathbf{\Delta}}_i\hat{S}_i)^{\mathsf{T}},$$

$$\hat{B}=\frac{1}{n}\sum_{i=1}^{n}\hat{D}_i\hat{V}_i^{-1}\hat{\mathbf{\Delta}}_i\hat{D}_i^{\top}, \hat{G}=\frac{1}{n}\sum_{i=1}^{n}\left[\hat{D}_i\hat{V}_i^{-1}\hat{\mathbf{\Delta}}_i\hat{S}\hat{\mathbf{Q}}_i^T\hat{H}^{-T}\hat{C}^T\right],$$

$$\hat{H}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{Q}_i|_{\gamma=\hat{\gamma}}, \hat{C}=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\gamma_t}(D_iV_i^{-1}\mathbf{\Delta}_iS_i)|_{\gamma=\hat{\gamma},\beta=\hat{\beta}},$$

$\hat{B_i}$, $\hat{D_i}$, $\hat{S_i}$ and $\hat{V_i}$ denote the corresponding quantities with β replaced by $\hat{\beta}$.

When γ is estimated, the asymptotic variance contains an additional term to account for the variability in estimating γ. A proof of the theorem is sketched in the Web Appendix C.

5. Simulation Study

In this section, we assess the performance of the proposed approach under small to large sample sizes using simulation studies. All simulations were performed with a Monte Carlo size of 1,000. We examine the performance of the new approach for both the cross-sectional and longitudinal data. For space consideration, we only report results for longitudinal data with sample size n = 50, 200 and 1000.

For notational brevity, we considered a relatively simple pre-post longitudinal study design, with only one explanatory variable x_i following a normal N (1, 1). The count response at both the pre and post time points, $\mathbf{y}_i = (y_{i1}, y_{i2})^\top$, is assumed to follow the following marginal ZIB model:

$$y_{it}|x_i \sim \text{ZIB}(m, p_i, \rho_i), \text{logit}(\rho_i|\mathbf{x}_i) = \beta_{u0}, \text{logit}(p_i|\mathbf{x}_i) = \beta_{v0} + \beta_{v1}x_i, 1 \le i \le n. \quad (15)$$

The Copula method was used to simulate correlated data with the fixed marginal model (15) [6, 17, 28].

We first generated $x_i \sim N$ (1, 1), followed by simulating the membership of mixtures by generating an independent Bernoulli random variables a_i with parameters ρ_i. If a_i = 1, set y_{i1} = y_{i2} = 0, i.e., the subject was a structural zero. Otherwise, y_{it} was simulated according to y_{it} ~ Binomial(m, p_i), from the at-risk group, where m was set 7, 15 and 30 corresponding to the cases where data are collected weekly, semimonthly and monthly. We set $\beta_{u0} = -1.2$ to produce about 20% structural zeros.

Copula was used to generate correlated multivariate responses for the at-risk group [17]. Let λ_i be the correlation between the pre- and post- measures among the at-risk group. Then, it is readily checked that for the ZIB in (15), the correlation between y_{i1} and y_{i2} is

$$Corr(y_{i1}, y_{i2}) = \frac{\lambda_i - p_i(\lambda_i - m\rho_i)}{1 - p_i(1 - m\rho_i)}. \quad (16)$$

See Web Appendix D for the derivation.

We examined the performance of the approach for both the complete and missing data case. For the missing data case, we simulated missing values following both the MCAR and MAR to examine the impact of different missing mechanisms on the validity of the GEE inference. We assumed no missing data at baseline (t = 1) for the response and the covariate so that missing values only occurred to y_{i2}. To create missing data under MAR, we simulated the missingness at time t = 2 using the following logistic model:

$$\text{logit}(\pi_{i2}(\gamma)) = \gamma_0 + \gamma_x x_{i1} + \gamma_y y_{i1}, \quad (17)$$

where γ_0, γ_x and γ_y were constants, controlling the amount of missing data as well as the strength of dependence of missingness of y_{i2} on x_{i1} and y_{i1}.

5.1. Complete Data Case

We simulated data using different λ_is to assess the impact of the correlation between the repeated outcomes among the at-risk group. For space issues, we present the cases when λ_i = 0.001 and 0.5. When λ_i = 0.001, the pre and post outcomes are virtually independent for the at-risk group. Thus, the correlation between y_{i1} and y_{i2} for the whole sample, as indicated by (16), is due to the structural zeros. To keep the correlation between y_{i1} and y_{i2} in a reasonable range, we set $\beta_{v0} = -0.1$ and $\beta_{v1} = -0.3$ in our simulation study, where the corresponding correlations between y_{i1} and y_{i2} range from 0.52 to 0.82.

We fitted the proposed ZIB-like model to the simulated data under the working independence model. Shown in Table 1 are the estimates of β and empirical and sandwich standard errors from 1000 MC simulations in the complete data case for the case λ_i = 0.001. For comparison purposes, the simulation results from fitting the simulated data using the ZIB-ES method [7] are also presented. The results suggest that both the proposed and the ZIB-ES methods provide very similar estimates. The estimates of β are both quite close to the their respective true values. Note that Hall and Zhang's ZIB-ES requires much more computing time; the computing time used was on average about 10 times that of our proposed method for these simulation studies. This is because ES is an EM-type algorithm, which is notorious for its slow convergence [10, 13, 15, 16, 27].

The simulation study for the λ_i = 0.5 case suggested similar conclusion and not reported here to save space.

5.2. Missing Data Case

We used model (17) to generate missing values. First we consider the MCAR cases. To generate missing data for MCAR, we set $\gamma_x = \gamma_y = 0$. The value of γ_0 was selected so that there are about 20% missing values in y_{i2}. Under MCAR, missing values can be simply ignored. Thus, we used listwise deletion to deal with missing values and simply applied GEE to the observed data. Shown in Table 2 are the estimates of β and empirical and sandwich standard errors based on 1000 MC simulations for both the ZIB-like WGEE method and ZIB-ES method for the case λ_i = 0.001. The performance of both methods are similar to that of the complete data cases, i.e., they yield similar estimates, but the ZIB-ES approach again requires significantly more time for the computation.

Again, the simulation study for the λ_i = 0.5 case suggested similar conclusion and not reported here.

To simulate missing responses following MAR, we set $\gamma_x = \gamma_y = \frac{1}{2}$ in (17). Again, the value of γ_0 was selected to create about 20% missing responses y_{i2} at post assessment. Unlike the MCAR case, we cannot simply ignore missing values in the MAR cases. We applied our weighted GEE approach to deal with the missing values. We used model (17) for the missing mechanism and the parameters were estimated from the simulated data. As an illustration that ZIB-ES cannot deal with missing values in the MAR cases, we also computed the ZIB-ES estimates ignoring missing values.

Shown in Tables 3 and 4 are the estimated results for both the ZIB-like WGEE method and ZIB-ES method for the cases $\lambda_i = 0.001$ and $\lambda_i = 0.5$, respectively. Our WGEE approach showed similar performance as in the complete and MCAR cases, demonstrating its capability to deal with missing values under MAR assumption. In contrast, the ZIB-ES method fail to address the missing values. The bias in the estimates of the coefficient in the mixture component, β_u, is apparent for both λs. For the count component, the ZIB-ES estimates for β_{v0} and β_{v1} are still good when $\lambda = 0.001$; this is expected because the λ is so small that the pre and post outcomes are almost independent for the at-risk group. However, when $\lambda = 0.5$, obvious bias occur in ZIB-ES estimates for the intercept β_{v0}, although the bias in ZIB-ES estimates for β_{v1} is not obvious. Again, this is expected, and in fact similar phenomenon occurs for binomial regression, i.e., the cases when there are no structural zeros. Thus, the ZIB-ES method does not apply to MAR in general. Further, it continues to suffer from the computation issue, using about 10 times more computing time than our WGEE approach.

6. Real Study Data

To illustrate the proposed approach with real study data, we applied the approach to the COMBINE study. This multi-site randomized clinical trial was conducted from 2001 to 2004 for subjects with alcohol dependence. This study was designed to compare two pharmacological treatments for alcoholism, naltrexone and acamprosate, alone and in combination with an intensive behavioral treatment, combined behavioral intervention (CBI). In the study, 1383 subjects were randomly assigned to one of nine groups [2, 3]. Eight groups (n=1226) received medical management, a 9-session intervention focused on enhancing medication adherence and abstinence. Among the eight groups, four groups (n=619) also received CBI and the remained four (n=607) did not receive CBI. We group the first four group as the group of medical management (MM) and the last four groups as the group of combined MM and CBI treatment (Combined). The ninth group (n=157) received CBI alone (CBI), without pills or medical management, and hence serves as the control group. The subjects were assessed 9 times during the 16 weeks of treatment and at 26, 52, and 68 weeks after randomization, i.e., up to 1 year after treatment ended. We focused on one of the primary outcomes DAD, which measures a subject's days of drinking in the last 30 days.

The DAD outcome did not have any zero at baseline, but had preponderance of zeros after the treatments were initiated. Table 5 shows the average percentage of DAD over 30 days period, i.e., the ratio of the number of days of drinking divided by the 30 days period, and the percentage of zeros of DAD at each assessment for the three treatment groups. A significant percentage of zeros were present during the 16 weeks of treatment (26% – 39%) and at the follow up visits (20% – 29%), providing a strong indication of success of the interventions by increasing the number of alcohol abstainers in this study population.

We first model the outcome DAD at 4-week (y_{i1}), 8-week (y_{i2}), 12-week (y_{i3}) and 16-week (y_{i4}) as a function of treatment conditions (x_{i1}, x_{i2}), the baseline DAD, y_{i0}, and the sites, where x_{i1} is defined as 1 for the treatment condition MM and 0 otherwise, and x_{i2} is 1 for the combined and 0 for others. The sites were controlled in the model because the sites were

significantly different among the treatment conditions [2]. While random effect models are often used for such site effect given the number of the sites, we choose the fixed effect approach here for two reasons. First, it is hard to specify the random effects for mixture outcomes such as ZIB, if no prior information is available. Second, we are focusing semiparametric approaches in the paper, but available random effect models are parametric. Since there were 11 sites, 10 dummy variables (z_{i1}, z_{i2}, …, z_{i10}) were created and included in the model.

We used the ZIB-like model in (6) with

$$\text{logit}(\rho_{it})=\beta_{u0}+\beta_{u1}x_{i1}+\beta_{u2}x_{i2}+\beta_{u3}y_{i0}+\beta_{u4}z_{i1}+\ldots+\beta_{u13}z_{i10}$$

$$\text{logit}(p_{it})=\beta_{v0}+\beta_{v1}x_{i1}+\beta_{v2}x_{i2}+\beta_{v3}y_{i0}+\beta_{v4}z_{i1}+\ldots+\beta_{v13}z_{i10}. \quad (18)$$

We also compared the treatment effect for the whole study period by modeling DAD at additional two assessments time 52-week (y_{i5}) and 68-week (y_{i6}) using the same model as (18).

Among the 1,383 patients, 90 (6.5%) dropped out of the study by the end of treatment period and this number increased to 212 (15.3%) by the end of the study. We examined the missing data mechanism by the logistic regression under MAR and MMDP assumptions. Specifically, we conducted logistic regressions at each assessment time except the baseline with the missing indictor as the dependent variable and the treatment conditions, the DAD from the prior assessment time, baseline demographic information age and gender as the covariates. The weights were estimated and then been used in the ZIB-like WGEE model.

Treating the CBI only as the reference group, the estimated treatment effects are presented in Table 6. The ten estimates for the sites were not included here due to the space limitation. For the treatment period, compared to the CBI group, both the MM and Combined groups had large proportion of subjects abstinent from alcohol drinking (structural zero component), and for those weren't abstinent from alcohol drink, the subjects in these two groups were less likely to drink. Similar patterns were found for the whole study period, though the MM has a borderline p-value 0.075. Our estimates confirmed the findings in [2], but the ability of assessing the treatment effect in turning subjects to be not at-risk as well as reducing the alcohol use among at-risk subjects are more comprehensive, because models in [2] ignores the issue of structural zeros. It can also provide much welcomed information to target subjects who most need the intervention. Although the ZIB-ES method is not appropriate here as illustrated in the simulating studies, we did apply it to the data for comparison purpose. However, the fitting of the ZIB-ES method did not converge.

Discussion

In medical and psychosocial research, we frequently encounter bounded count responses with a preponderance of zeros. Such variables often arise when subjects/patients experience a number of events/activities such as DAD over a period of time. As they are the sum of

finitely many zeros and ones, it is more reasonable to model these outcomes using the zero-inflated binomial, rather than the zero-inflated Poisson. However, since they are sums of dependent Bernoulli variables, which induce overdispersion, they are not exactly (zero-inflated) binomial. By integrating the generalized estimating equation for dependent responses and the inverse probability weighting technique for missing values, we developed a distribution-free approach for modeling such overdispersed ZIB-like outcomes for both cross-sectional and longitudinal studies. Unlike standard log-linear models for count responses in the absence of structural zeros, the proposed ZIB-like model has a more complex bivariate response function, key to identifying the two latent subgroups of a mixed population consisting of the at- and non-risk subgroups. Our approach only models two mean responses, thereby providing more robust inference than parametric alternatives. Further studies are needed to extend the approach to more complex situations such as non-monotone missing data under MAR and non-parametric form of the mean response functions. Addressing these and other limitations will further facilitate building more accurate models for ZIB-like data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank professors Xin Tu and Wan Tang for their constructive comments and suggestions.

Funding

The study was supported in part by National Institute on Drug Abuse grant R33DA027521, National Institute of General Medical Sciences grant R01GM108337, UR CTSI grants 8UL1TR000042-07 and 8UL1TR000042-09.

References

1. Chamberlain G. Asymptotic efficiency in estimation with conditional moment restrictions. Journal of Econometrics. 1987; 34:305–334.
2. COMBINE Study Research Group. Testing combined pharmacotherapies and behavioral interventions in alcohol dependence: rationale and methods. Vol. 27. Alcoholism (NY): 2003. p. 1107-1122.
3. COMBINE Study Research Group. Combined pharmacotherapies and behavioral interventions for alcohol dependence – the COMBINE study: a randomized controlled trial. JAMA: The Journal of the American Medical Association. 2006; 295:2003–2017. [PubMed: 16670409]
4. Crowder M. On linear and quadratic estimating functions. Biometrika. 1987; 74:591–597.
5. Dobbie MJ, Welsh AH. Modelling correlated zero-inflated count data. Aust. N. Z. J. Stat. 2001; 43:431–444.
6. Freesm EW, Valdez E. Understanding relationships using copulas. North American Actuarial Journal. 1998:1–25.
7. Hall D, Zhang Z. Marginal models for zero inflated clustered data. Statistical Modelling. 2004; 4:161–180.
8. Hall DB. Zero-inflated Poisson and binomial regression with random effects: A case study. Biometrics. 2000; 56:1030–1039. [PubMed: 11129458]
9. Kowalski, J.; Tu, XM. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley-Interscience; 2008. Modern applied U-statistics.

10. Kowalski J, et al. On the rate of convergence of the ecme algorithm for multiple regression models with t-distributed errors. Biometrika. 1997; 84:269–281.
11. Lambert D. Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics. 1992; 34:1–14.
12. Liang K, Zeger S. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13–22.
13. Louis T. Finding the observed information matrix when using the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological). 1982:226–233.
14. Ma Y, et al. Inference for kappas for longitudinal study data: Applications to sexual health research. Biometrics. 2008; 64:781–789. [PubMed: 18047535]
15. Meilijson I. A fast improvement to the em algorithm on its own terms. Journal of the Royal Statistical Society. Series B (Methodological). 1989:127–138.
16. Meng X, Rubin D. Using EM to obtain asymptotic variance-covariance matrices: the sem algorithm. Journal of the American Statistical Association. 1991:899–909.
17. Nelsen, RB. An introduction to copulas. Springer: 2006.
18. Newey WK. Adaptive estimation of regression models via moment restrictions. Journal of Econometrics. 1988; 38:301–339.
19. Prentice RL, Zhao LP. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics. 1991:825–839. [PubMed: 1742441]
20. Ridout M, Hinde J, DemeAtrio CG. A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. Biometrics. 2001; 57:219–223. [PubMed: 11252601]
21. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association. 1995; 90:122–129.
22. Rubin DB. Inference and missing data. Biometrika. 1976; 63:581–592.
23. Tang W, He H, Gunzler D. Kernel smoothing density estimation when group membership is subject to missing. Journal of Statistical Planning and Inference. 2012; 142:685–694. [PubMed: 22116738]
24. Tang, W.; He, H.; Tu, X. Applied Categorical and Count Data Analysis. Chapman & Hall/CRC; 2012.
25. Tsiatis, AA. Semiparametric Theory and Missing Data. Springer, New York: Springer Series in Statistics; 2006.
26. Vieira A, Hinde JP, Demétrio CG. Zero-inflated proportion data models applied to a biological control assay. Journal of Applied Statistics. 2000; 27:373–389.
27. Wu C. On the convergence properties of the em algorithm. The Annals of Statistics. 1983:95–103.
28. Yan JR. Package copula on cran, multivariate dependence with copula. 2009
29. Yu Q, et al. Distribution-free models for longitudinal count responses with overdispersion and structural zeros. Statistics in medicine. 2013; 32:2390–2405. [PubMed: 23239019]

Table 1

GEE and ZIB-ES estimates of parameters, empirical standard error (Emp. SE), and asymptotical standard errors (Asym. SE) under complete data from 1000 MC simulations.

			GEE Est.	ZIB-ES Est.
m	**Sample Size**	**Parameter**	**Mean(Emp., Asym.)**	**Mean(Emp., Asym.)**
7	50	β_u	−1.244 (0.354, 0.350)	−1.245 (0.355, 0.366)
		β_{v0}	−0.103 (0.129, 0.131)	−0.104 (0.129, 0.128)
		β_{v1}	−0.302 (0.101, 0.101)	−0.303 (0.102, 0.099)
	200	β_u	−1.210 (0.177, 0.175)	−1.210 (0.175, 0.177)
		β_{v0}	−0.101 (0.065, 0.063)	−0.101 (0.063, 0.064)
		β_{v1}	−0.300 (0.050, 0.050)	−0.301 (0.051, 0.049)
	1000	β_u	−1.200 (0.079, 0.079)	−1.200 (0.079, 0.079)
		β_{v0}	−0.099 (0.027, 0.029)	−0.099 (0.027, 0.029)
		β_{v1}	−0.301 (0.022, 0.022)	−0.301 (0.022, 0.022)
15	50	β_u	−1.235 (0.335, 0.335)	−1.235 (0.335, 0.344)
		β_{v0}	−0.101 (0.082, 0.085)	−0.101 (0.082, 0.083)
		β_{v1}	−0.301 (0.066, 0.063)	−0.301 (0.063, 0.061)
	200	β_u	−1.208 (0.167, 0.168)	−1.208 (0.167, 0.169)
		β_{v0}	−0.100 (0.042, 0.042)	−0.100 (0.042, 0.042)
		β_{v1}	−0.300 (0.031, 0.032)	−0.300 (0.032, 0.031)
	1000	β_u	−1.201 (0.076, 0.075)	−1.201 (0.076, 0.075)
		β_{v0}	−0.099 (0.019, 0.019)	−0.099 (0.019, 0.019)
		β_{v1}	−0.301 (0.014, 0.014)	−0.301 (0.014, 0.014)
30	50	β_u	−1.234 (0.334, 0.334)	−1.234 (0.334, 0.343)
		β_{v0}	−0.101 (0.058, 0.061)	−0.101 (0.058, 0.058)
		β_{v1}	−0.300 (0.044, 0.045)	−0.300 (0.044, 0.43)

m	Sample Size	Parameter	GEE Est. Mean(Emp., Asym.)	ZIB-ES Est. Mean(Emp., Asym.)
	200	β_u	−1.208 (0.167, 0.167)	−1.208 (0.167, 0.169)
		$\beta_{\upsilon 0}$	−0.100 (0.030, 0.030)	−0.100 (0.030, 0.030)
		$\beta_{\upsilon 1}$	−0.300 (0.022, 0.022)	−0.300 (0.022, 0.022)
	1000	β_u	−1.201 (0.076, 0.075)	−1.201 (0.076, 0.075)
		$\beta_{\upsilon 0}$	−0.099 (0.013, 0.013)	−0.099 (0.013, 0.013)
		$\beta_{\upsilon 1}$	−0.301 (0.010, 0.010)	−0.301 (0.010, 0.010)

Simulation summary for GEE and ZIB-ES under complete data $\lambda = 0.001$, $\beta_u = -1.2$, $\beta_{\upsilon 0} = -0.1$, $\beta_{\upsilon 1} = -0.3$

Table 2

WGEE and ZIB-ES estimates of parameters, Emp. SE, and Asym. SE under MCAR from 1000 MC simulations.

m	Sample Size	WGEE Est.		ZIB-ES Est.
		Parameter	Mean(Emp., Asym.)	Mean(Emp., Asym.)
7	50	β_u	−1.254 (0.370, 0.361)	−1.254 (0.367, 0.377)
		$\beta_{\upsilon 0}$	−0.101 (0.140, 0.138)	−0.102 (0.138, 0.134)
		$\beta_{\upsilon 1}$	−0.306 (0.111, 0.106)	−0.305 (0.110, 0.104)
	200	β_u	−1.207 (0.182, 0.181)	−1.208 (0.181, 0.182)
		$\beta_{\upsilon 0}$	−0.101 (0.067, 0.068)	−0.102 (0.066, 0.068)
		$\beta_{\upsilon 1}$	−0.300 (0.053, 0.052)	−0.300 (0.053, 0.052)
	1000	β_u	−1.201 (0.081, 0.081)	−1.201 (0.081, 0.081)
		$\beta_{\upsilon 0}$	−0.099 (0.029, 0.031)	−0.099 (0.029, 0.030)
		$\beta_{\upsilon 1}$	−0.301 (0.023, 0.023)	−0.301 (0.023, 0.023)
15	50	β_u	−1.254 (0.370, 0.361)	−1.241 (0.343, 0.353)
		$\beta_{\upsilon 0}$	−0.101 (0.140, 0.138)	−0.100 (0.089, 0.087)
		$\beta_{\upsilon 1}$	−0.306 (0.111, 0.106)	−0.302 (0.068, 0.065)
	200	β_u	−1.207 (0.182, 0.181)	−1.206 (0.172, 0.173)
		$\beta_{\upsilon 0}$	−0.101 (0.067, 0.068)	−0.100 (0.043, 0.044)
		$\beta_{\upsilon 1}$	−0.300 (0.053, 0.052)	−0.300 (0.033, 0.033)
	1000	β_u	−1.201 (0.081, 0.081)	−1.202 (0.078, 0.077)
		$\beta_{\upsilon 0}$	−0.099 (0.029, 0.031)	−0.099 (0.020, 0.020)
		$\beta_{\upsilon 1}$	−0.301 (0.023, 0.023)	−0.301 (0.015, 0.015)
30	50	β_u	−1.241 (0.344, 0.344)	−1.240 (0.342, 0.352)
		$\beta_{\upsilon 0}$	−0.100 (0.064, 0.064)	−0.100 (0.063, 0.061)
		$\beta_{\upsilon 1}$	−0.301 (0.048, 0.047)	−0.301 (0.047, 0.045)
	200	β_u	−1.205 (0.172, 0.173)	−1.206 (0.171, 0.173)
		$\beta_{\upsilon 0}$	−0.100 (0.031, 0.032)	−0.100 (0.031, 0.031)
		$\beta_{\upsilon 1}$	−0.300 (0.023, 0.023)	−0.300 (0.023, 0.023)
	1000	β_u	−1.201 (0.078, 0.077)	−1.202 (0.078, 0.077)
		$\beta_{\upsilon 0}$	−0.100 (0.014, 0.014)	−0.100 (0.014, 0.014)
		$\beta_{\upsilon 1}$	−0.301 (0.010, 0.010)	−0.301 (0.010, 0.010)

Simulation summary for WGEE and ZIB-ES under MCAR $\lambda = 0.001$, $\beta_u = -1.2$, $\beta_{\upsilon 0} = -0.1$, $\beta_{\upsilon 1} = -0.3$

Table 3

WGEE and ZIB-ES estimates of parameters, Emp. SE, and Asym. SE under MAR from 1000 MC simulations.

m	Sample Size		WGEE Est. Mean(Emp., Asym.)	ZIB-ES Est. Mean(Emp., Asym.)
7	50	β_u	−1.246 (0.366, 0.358)	−1.162 (0.378, 0.374)
		$\beta_{\upsilon 0}$	−0.105 (0.137, 0.139)	−0.099 (0.143, 0.133)
		$\beta_{\upsilon 1}$	−0.302 (0.111, 0.109)	−0.309 (0.113, 0.104)
	200	β_u	−1.209 (0.180, 0.179)	−1.121 (0.182, 0.180)
		$\beta_{\upsilon 0}$	−0.101 (0.068, 0.069)	−0.097 (0.068, 0.067)
		$\beta_{\upsilon 1}$	−0.301 (0.055, 0.054)	−0.303 (0.052, 0.052)
	1000	β_u	−1.201 (0.082, 0.080)	−1.109 (0.080, 0.080)
		$\beta_{\upsilon 0}$	−0.098 (0.029, 0.031)	−0.099 (0.032, 0.030)
		$\beta_{\upsilon 1}$	−0.301 (0.024, 0.024)	−0.302 (0.023, 0.023)
15	50	β_u	−1.236 (0.342, 0.339)	−1.099 (0.347, 0.347)
		$\beta_{\upsilon 0}$	−0.101 (0.094, 0.094)	−0.096 (0.093, 0.089)
		$\beta_{\upsilon 1}$	−0.301 (0.071, 0.070)	−0.305 (0.070, 0.066)
	200	β_u	−1.208 (0.171, 0.170)	−1.078 (0.171, 0.170)
		$\beta_{\upsilon 0}$	−0.100 (0.047, 0.047)	−0.098 (0.045, 0.045)
		$\beta_{\upsilon 1}$	−0.300 (0.036, 0.035)	−0.301 (0.033, 0.033)
	1000	β_u	−1.201 (0.078, 0.076)	−1.071 (0.078, 0.076)
		$\beta_{\upsilon 0}$	−0.099 (0.020, 0.021)	−0.099 (0.022, 0.020)
		$\beta_{\upsilon 1}$	−0.301 (0.015, 0.016)	−0.301 (0.015, 0.015)
30	50	β_u	−1.231 (0.342, 0.341)	−1.084 (0.347, 0.345)
		$\beta_{\upsilon 0}$	−0.102 (0.071, 0.071)	−0.099 (0.068, 0.064)
		$\beta_{\upsilon 1}$	−0.298 (0.052, 0.051)	−0.303 (0.050, 0.046)
	200	β_u	−1.207 (0.172, 0.171)	−1.068 (0.171, 0.170)
		$\beta_{\upsilon 0}$	−0.102 (0.037, 0.036)	−0.098 (0.033, 0.033)
		$\beta_{\upsilon 1}$	−0.299 (0.027, 0.026)	−0.301 (0.024, 0.024)
	1000	β_u	−1.202 (0.079, 0.077)	−1.061 (0.078, 0.076)
		$\beta_{\upsilon 0}$	−0.100 (0.016, 0.017)	−0.100 (0.018, 0.015)
		$\beta_{\upsilon 1}$	−0.300 (0.012, 0.012)	−0.301 (0.011, 0.011)

Simulation summary for WGEE and ZIB-ES under MAR $\lambda = 0.001$, $\beta_u = -1.2$, $\beta_{\upsilon 0} = -0.1$, $\beta_{\upsilon 1} = -0.3$

J Appl Stat. Author manuscript; available in PMC 2016 October 01.

Table 4

WGEE and ZIB-ES estimates of parameters, Emp. SE, and Asym. SE under MAR from 1000 MC simulations.

		WGEE Est.		ZIB-ES Est.
m	**Samp Size**	**Parameter**	**Mean(Emp., Asym.)**	**Mean(Emp., Asym.)**
7	50	β_u	−1.255(0.369, 0.377)	−1.164(0.371, 0.377)
		$\beta_{\upsilon 0}$	−0.106(0.165, 0.159)	−0.125(0.160, 0.157)
		$\beta_{\upsilon 1}$	−0.306(0.130, 0.120)	−0.308(0.125, 0.121)
	200	β_u	−1.211(0.180, 0.181)	−1.121(0.181, 0.181)
		$\beta_{\upsilon 0}$	−0.100(0.080, 0.081)	−0.120(0.078, 0.079)
		$\beta_{\upsilon 1}$	−0.302(0.063, 0.062)	−0.305(0.061, 0.060)
	1000	β_u	−1.201(0.082, 0.080)	−1.111(0.083, 0.080)
		$\beta_{\upsilon 0}$	−0.098(0.035, 0.036)	−0.118(0.034, 0.035)
		$\beta_{\upsilon 1}$	−0.302(0.028, 0.028)	−0.306(0.027, 0.027)
15	50	β_u	−1.236(0.342, 0.349)	−1.107(0.348, 0.347)
		$\beta_{\upsilon 0}$	−0.104(0.110, 0.106)	−0.124(0.104, 0.103)
		$\beta_{\upsilon 1}$	−0.300(0.083, 0.078)	−0.300(0.079, 0.077)
	200	β_u	−1.208(0.171, 0.171)	−1.079(0.175, 0.170)
		$\beta_{\upsilon 0}$	−0.101(0.055, 0.055)	−0.124(0.053, 0.052)
		$\beta_{\upsilon 1}$	−0.300(0.042, 0.041)	−0.301(0.040, 0.039)
	1000	β_u	−1.201(0.078, 0.076)	−1.072(0.079, 0.076)
		$\beta_{\upsilon 0}$	−0.099(0.025, 0.025)	−0.123(0.023, 0.024)
		$\beta_{\upsilon 1}$	−0.301(0.018, 0.018)	−0.301(0.017, 0.017)
30	50	β_u	−1.231(0.342, 0.349)	−1.094(0.348, 0.346)
		$\beta_{\upsilon 0}$	−0.104(0.085, 0.077)	−0.125(0.075, 0.074)
		$\beta_{\upsilon 1}$	−0.298(0.061, 0.055)	−0.296(0.056, 0.054)
	200	β_u	−1.208(0.171, 0.172)	−1.069(0.174, 0.170)

		WGEE Est.		ZIB-ES Est.
m	**Samp Size**	**Parameter**	**Mean(Emp., Asym.)**	**Mean(Emp., Asym.)**
		$\beta_{\upsilon 0}$	−0.102(0.043, 0.041)	−0.125(0.038, 0.037)
		$\beta_{\upsilon 1}$	−0.299(0.031, 0.029)	−0.296(0.028, 0.027)
	1000	β_u	−1.202(0.079, 0.077)	−1.063(0.080, 0.076)
		$\beta_{\upsilon 0}$	−0.100(0.020, 0.020)	−0.125(0.017, 0.017)
		$\beta_{\upsilon 1}$	−0.300(0.014, 0.014)	−0.296(0.012, 0.012)

Simulation summary for WGEE and ZIB-ES under MAR $\lambda = 0.5$, $\beta_u = -1.2$, $\beta_{\upsilon 0} = -0.1$, $\beta_{\upsilon 1} = -0.3$

Table 5

The average percentage of DAD and percentage of zeros for DAD among the three groups at each assessment time for the COMBINE Study

Assessment time	CBI only		MM		Combined	
	Mean (SD)	Zeros (%)	Mean (SD)	Zeros (%)	Mean (SD)	Zeros (%)
Baseline						
	0.76 (0.25)	0 (0.00)	0.74 (0.24)	0 (0.00)	0.75 (0.26)	0 (0.00)
Treatment Period						
Week 4	0.33 (0.35)	41 (26.45)	0.20 (0.26)	213 (35.15)	0.22 (0.29)	225 (36.64)
Week 8	0.36 (0.37)	39 (26.17)	0.25 (0.32)	202 (34.18)	0.24 (0.31)	205 (34.69)
Week 12	0.35 (0.36)	42 (28.38)	0.26 (0.33)	222 (38.47)	0.23 (0.31)	215 (36.69)
Week 16	0.35 (0.37)	45 (30.82)	0.27 (0.33)	201 (35.39)	0.23 (0.32)	225 (38.86)
Follow-up						
Week 52	0.41 (0.37)	28 (20.14)	0.38 (0.37)	132 (24.26)	0.37 (0.37)	149 (26.90)
Week 68	0.42 (0.40)	36 (28.80)	0.38 (0.38)	136 (26.25)	0.37 (0.37)	150 (28.41)

Average percentage of DAD and the number of zeros at each assessment for COMBINE Study

Table 6

WGEE estimates of parameters, standard errors, and p-values for the ZIB-like model under MAR and MMDP for the COMBINE Study

	Treatment period only			Whole study period		
Parameters	**Estimate**	**SE**	**p-value**	**Estimate**	**SE**	**p-value**
Strucural zero part (ρ_{it})						
β_{u0}	−0.784	0.273	0.004	−0.701	0.259	0.007
β_{u1} (MM vs CBI only)	0.383	0.169	0.024	0.274	0.154	0.075
β_{u2} (Combined vs CBI only)	0.427	0.169	0.012	0.345	0.155	0.026
β_{u3} (baseline DAD)	−0.006	0.007	0.327	−0.010	0.006	0.124
Binomial part (p_{it})						
$\beta_{\upsilon 0}$	−1.322	0.208	< 0.0001	−1.264	0.183	< 0.0001
$\beta_{\upsilon 1}$ (MM vs CBI only)	−0.358	0.122	0.003	−0.252	0.106	0.018
$\beta_{\upsilon 2}$ (Combined vs CBI only)	−0.480	0.122	< 0.0001	−0.347	0.107	0.001
$\beta_{\upsilon 3}$ (baseline DAD)	0.059	0.006	< 0.0001	0.060	0.005	< 0.0001

Results of WGEE based on ZIB-like model for COMBINE Study p-value for $H_0 : \beta = 0$