



HHS Public Access

Author manuscript

J R Stat Soc Ser C Appl Stat. Author manuscript; available in PMC 2016 November 01.

Published in final edited form as:

J R Stat Soc Ser C Appl Stat. 2015 November ; 64(5): 815–830. doi:10.1111/rssc.12104.

A Marginalized Zero-inflated Poisson Regression Model with Random Effects

D. Leann Long[†],

Department of Biostatistics, West Virginia University, Morgantown, WV USA

John S. Preisser,

Department of Biostatistics, University of North Carolina, Chapel Hill, NC USA

Amy H. Herring, and

Department of Biostatistics, University of North Carolina, Chapel Hill, NC USA; Carolina Population Center, University of North Carolina, Chapel Hill, NC USA

Carol E. Golin

Department of Health Behavior, University of North Carolina, Chapel Hill, NC USA; Department of Medicine, University of North Carolina, Chapel Hill, NC USA

Summary

Public health research often concerns relationships between exposures and correlated count outcomes. When counts exhibit more zeros than expected under Poisson sampling, the zero-inflated Poisson (ZIP) model with random effects may be used. However, the latent class formulation of the ZIP model can make marginal inference on the sampled population challenging. This article presents a marginalized ZIP model with random effects to directly model the mean of the mixture distribution consisting of ‘susceptible’ individuals and excess zeroes, providing straightforward inference for overall exposure effects. Simulations evaluate finite sample properties, and the new methods are applied to a motivational interviewing-based safer sex intervention trial, designed to reduce the number of unprotected sexual acts.

Keywords

Marginalized Models; Repeated Measures; Unprotected Intercourse; Zero-inflation

1. Introduction

Infectious disease researchers are often concerned with reducing risky sexual behavior among HIV-positive individuals. One measure of risky sexual behavior is the Unprotected Anal and Vaginal Intercourse (UAVI) count, the number of unprotected anal or vaginal intercourse acts with any partner over a specified period of time. The SafeTalk program was developed by Golin et al. (2012) to reduce the number of unprotected sexual acts through a multicomponent, motivational interviewing-based, safer sex intervention. Sexual behavior

[†]Address for correspondence: D. Leann Long, Department of Biostatistics, School of Public Health, West Virginia University Health Sciences Center, PO Box 9190, Morgantown, WV 26506-9190, dllong@hsc.wvu.edu.

count data can display a distribution with excess zeros (Heilbron, 1994; Ghosh and Tu, 2009). To examine the efficacy of the SafeTalk program over time, a randomized controlled clinical trial collected risky sexual behavior data at baseline and up to three follow-up visits.

Several methods have been developed for modeling correlated count data with many zeros such as UAVI from the SafeTalk clinical trial. Building upon the zero-inflated Poisson (ZIP) regression model established by Mullahy (1986) and Lambert (1992), Hall (2000) extends the ZIP regression model to include random effects in the Poisson process. In order to account for overdispersion beyond the excess zeros, Yau, Wang and Lee (2003) modify the zero-inflated negative-binomial (ZINB) regression model to include random effects. Instead of using random effects to handle correlated data, Hall and Zhang (2004) employ GEE methodology for zero-inflated models in order to achieve population-averaged interpretations. For each of these zero-inflated methods, two sets of parameter estimates are produced, those associated with the excess zero process that models the probability of being non-susceptible for the disease or condition and those associated with the count process that models the mean count among susceptible individuals. In many applications, the two latent class interpretations are not clinically supported or simply not of interest, and the zero-inflated methodology is used as a convenient modeling technique to account for excess zeros in a population (Mwalili, et al., 2008).

While closely related to the zero-inflated methodology, hurdle models (including zero-altered models) specify a model for the probability of any zero in addition to the model for the mean of the untruncated distribution of the count data process (Mullahy, 1986; Heilbron, 1994). Dobbie and Welsh (2001) use the zero-altered Poisson model, modified to utilize GEE, to account for correlated observations. Min and Agresti (2005) extend the zero-altered model to include random effects.

The choice between the hurdle and zero-inflated model classes has been approached from various angles. Much of the literature pertaining to the analysis of count data with excess zeros focuses on model fit, using fit statistics to provide justification of model class choice. Gilthorpe, et al.(2009) argue that *a priori* knowledge of the data-generating mechanism could be used to identify the class of models from which to choose, supported by statements in Neelon et al.(2010) and Buu et al.(2012). Applications in which all zeros are considered as arising from an identical process indicate a hurdle model, rather than a zero-inflated model, where zeros can occur from the two different processes.

While many health-related fields are implementing zero-inflated techniques, sometimes health researchers wish to make inference upon an entire sampled population rather than the latent classes modeled by ZIP methodology (Preisser, et al., 2012). Albert et al.(2014) contend that interpretations for features of the marginal mixture distribution have been generally overlooked in the zero-inflated literature, such as the overall mean count, owing to the fact that ZIP models and hurdle models do not produce a direct overall estimate of exposure effect for the marginal mean count. In particular, transformation methods, with variance estimation by the delta method or resampling methods, may be used to make inference on overall estimates of a dichotomous exposure effect for ZIP and ZINB models

(Albert, et al., 2014). However, such transformations can be tedious for many analysts, and the treatment of continuous covariates is not necessarily apparent.

Proposing the marginalized model for longitudinal binary data, Heagerty (1999) employs joint models by directly modeling the marginal mean and simultaneously using a linked random effects model to account for correlated responses. Through this joint model, marginalization over random effects achieves population-averaged parameters, while accounting for correlated measures. Extending the marginalized model approach, Lee et al. (2011) focus on the hurdle model formulation for Poisson and negative binomial data with excess zeros while marginalizing over random effects for clustering. Since Lee et al. focus on marginalizing over the random effects, the two sets of parameters from their marginalized hurdle models have the same interpretations as hurdle models for independent responses.

Adapting the marginalized model approach to achieve inference on the marginal mean for independent count responses with excess zeroes, Long et al. (2014) present a new marginalized ZIP model that jointly models the marginal mean and excess zero process to produce estimates for marginal mean inference while accounting for excess zeroes. Where as marginalized models often average over random effects to obtain population-average effect estimates, the marginalized ZIP model averages over the two ZIP model processes to achieve overall effect estimates for expected counts, providing parameter estimates with the same interpretation as Poisson regression. This article builds upon both the marginalized ZIP model and current ZIP methods for correlated data and proposes the marginalized ZIP model with random effects.

Sections 2 and 3 briefly review the ZIP model with random effects from Hall (2000) and the marginalized ZIP model from Long et al. (2014), respectively. Section 4 proposes the marginalized ZIP model with random effects, which has subject-specific parameters, and discusses the situation where those parameters have equivalent population-averaged interpretations. Section 5 presents simulation study results examining the finite sample performance of the new model. In Section 6, we consider data from the SafeTalk randomized controlled clinical trial. A discussion is provided in Section 7.

2. ZIP model with random effects

Extending Lambert's ZIP model to incorporate correlated zero-inflated count data, Hall (2000) developed the ZIP model with random effects. Let $\mathbf{Y} = (\mathbf{Y}'_1, \dots, \mathbf{Y}'_K)$ where K is the number of independent clusters and $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iT_i})'$, where T_i is the number of observations for the i^{th} cluster. Let $s_{ij} = 1$ if Y_{ij} is from the first process (i.e. Y_{ij} is an excess zero) and $s_{ij} = 2$ if Y_{ij} is from the second (Poisson) process; s_{ij} is unobserved when $Y_{ij} = 0$. Then

$$Y_{ij} \sim \begin{cases} 0 & \text{with probability } P(s_{ij}=1) = \psi_{ij} \\ \text{Poisson}(\mu_{ij}^C) & \text{with probability } P(s_{ij}=2) = 1 - P(s_{ij}=1) = 1 - \psi_{ij} \end{cases} \quad (1)$$

Where $\mu_{ij}^C = E(Y_{ij} | s_{ij}=2, b_i)$. The notation μ_{ij}^C indicates that the Poisson mean is conditional on the random effect b_i . The log-linear and logistic regression models are

$$\begin{aligned} \text{logit}(\psi_{ij}) &= \mathbf{Z}'_{ij}\boldsymbol{\gamma} \\ \log(\mu_{ij}^C) &= \mathbf{X}'_{ij}\boldsymbol{\beta} + \sigma b_i, \end{aligned} \quad (2)$$

where $b_1, \dots, b_K \stackrel{i.i.d.}{\sim} N(0, 1)$, and \mathbf{Z}_{ij} and \mathbf{X}_{ij} are the covariate vectors for the logistic and Poisson processes, respectively. Note that $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are latent class parameters, providing separate inference for the excess zero and Poisson processes, respectively. The log-likelihood can be expressed

$$l(\boldsymbol{\Omega}, \mathbf{y}) = \sum_{i=1}^K \log \int_{-\infty}^{\infty} \left[\prod_{j=1}^{T_i} \Pr(Y_{ij} = y_{ij} | b_i, \boldsymbol{\Omega}) \right] \phi(b_i) db_i$$

where $\boldsymbol{\Omega} = (\boldsymbol{\gamma}', \boldsymbol{\beta}', \sigma)$, ϕ is the standard normal probability density and

$$\Pr(Y_{ij} = y_{ij} | b_i, \boldsymbol{\theta}) = \left[\psi_{ij} + (1 - \psi_{ij})e^{-\mu_{ij}^C} \right]^{u_{ij}} \left[\frac{(1 - \psi_{ij})e^{-\mu_{ij}^C} (\mu_{ij}^C)^{y_{ij}}}{y_{ij}!} \right]^{1-u_{ij}} = (1 + e^{\mathbf{Z}'_{ij}\boldsymbol{\gamma}})^{-1} \left\{ u_{ij} \left[e^{\mathbf{Z}'_{ij}\boldsymbol{\gamma} + \exp(-e^{\mathbf{X}'_{ij}\boldsymbol{\beta} + \sigma b_i})} \right] + (1 - u_{ij}) \right\}$$

where $u_{ij} = I(y_{ij} = 0)$. Using the EM algorithm framework that Lambert (1992) proposed, Hall fits this ZIP model with random effects with the EM algorithm with Gaussian quadrature. Generally, the overall conditional mean $E(Y_{ij} | b_i) = (1 - \psi_{ij})\mu_{ij}^C$ will depend on $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and b_i through a complicated function that does not permit easy and direct inference for overall effects, here defined as ratios of such means when a single covariate is allowed to vary. Although Hall (2000) used (2) to account for correlation within the Poisson process only, others have utilized correlated random effects in both processes of the ZIP and hurdle models (Dobbie and Welsh, 2002; Min and Agresti, 2005; Ghosh and Tu, 2009; Neelon et al., 2010).

3. Marginalized ZIP model for independent responses

Rather than jointly modeling the excess zero probability and the latent class Poisson mean μ_i , Long, et al.(2014) instead propose the marginalized ZIP regression model, which directly models the marginal mean of the mixture distribution in addition to the zero-inflation process. For independent outcomes Y_i , the marginalized ZIP model is given by

$$\begin{aligned} \text{logit}(\psi_i) &= \mathbf{Z}'_i\boldsymbol{\gamma} \\ \log(\nu_i) &= \mathbf{X}'_i\boldsymbol{\alpha}. \end{aligned} \quad (4)$$

where v_i is the marginal mean, that is $v_i \equiv E(Y_i)$. The elements of $\boldsymbol{\gamma}$ provide inference on the probability of an excess zero, the same interpretations as ZIP models. However, the modeling of the marginal mean v_i allows log-incidence density rate interpretations of the elements of $\boldsymbol{\alpha}$, providing the same interpretation as in Poisson regression. The marginalized ZIP model utilizes the ZIP likelihood framework and the concept of marginalized models to marginalized over the two processes. Specifically, the Poisson process mean is redefined as a general function of model parameters in (4). Solving $v_i = (1 - \psi_i)\mu_i$, with substitution for (4), provides

$$\mu_i = (1 + e^{\mathbf{Z}'_i \boldsymbol{\gamma}}) e^{\mathbf{X}'_i \boldsymbol{\alpha}}.$$

This definition of μ_i reparameterizes the ZIP model, allowing for inference on the marginal mean. Using this redefined μ_i , $\psi_i = \text{logit}^{-1}(\mathbf{Z}'_i \boldsymbol{\gamma})$ and the ZIP likelihood, the marginalized ZIP likelihood for $(\boldsymbol{\gamma}, \boldsymbol{\alpha})$ is derived to be

$$L(\boldsymbol{\gamma}, \boldsymbol{\alpha} | \mathbf{y}) = \prod_{y_i} (1 + e^{\mathbf{Z}'_i \boldsymbol{\gamma}})^{-1} \prod_{y_i=0} (e^{\mathbf{Z}'_i \boldsymbol{\gamma}} + e^{-(1 + \exp(\mathbf{Z}'_i \boldsymbol{\gamma})) \exp(\mathbf{X}'_i \boldsymbol{\alpha})}) \times \prod_{y_i > 0} [e^{-(1 + \exp(\mathbf{Z}'_i \boldsymbol{\gamma})) \exp(\mathbf{X}'_i \boldsymbol{\alpha})} (1 + e^{\mathbf{Z}'_i \boldsymbol{\gamma}})^{y_i} e^{\mathbf{X}'_i \boldsymbol{\alpha} y_i} / (y_i!)].$$

Long et al.(2014) note that analysts may fit this marginalized ZIP model in SAS NLMIXED, providing sample code as well as details for robust (empirical) standard error estimation. Although derived from a reparameterization of the ZIP model, the marginalized ZIP parameters yield direct inference on the marginal mean rather than the latent classes and gives statistical analysts a new class of models to address marginal exposure effects.

4. Marginalized ZIP model with random effects

4.1. Subject-specific marginalized ZIP model

Building upon both Hall (2000) and Long et al.(2014), we present a marginalized adaptation of the ZIP model with random effects for repeated measures data. Rather than modeling the conditional Poisson process mean μ_{ij}^C as in (2), the marginalized ZIP model for clustered data directly models the overall subject-specific mean $\nu_{ij}^C = E(Y_{ij} | \mathbf{d}_i)$ through

$$\begin{aligned} \text{logit}(\psi_{ij}^C) &= \mathbf{Z}'_{ij} \boldsymbol{\gamma} + \mathbf{w}'_{1ij} \mathbf{c}_i \\ \log(\nu_{ij}^C) &= \mathbf{X}'_{ij} \boldsymbol{\alpha} + \log(N_i) + \mathbf{w}'_{2ij} \mathbf{d}_i, \end{aligned} \quad (5)$$

where $\psi_{ij}^C = P(s_{ij} = 1 | \mathbf{c}_i)$ and $\mathbf{b}_i = (\mathbf{c}_i, \mathbf{d}_i)'$ follows the multivariate normal distribution with

mean zero and covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. Above, N_i represents an off-set variable for situations where the incidence density v_i/N_i is of interest. To account for clustering within each process, we propose correlated random effects \mathbf{c}_i and \mathbf{d}_i and

corresponding column design vectors \mathbf{w}_{1ij} , \mathbf{w}_{2ij} , usually subsets of \mathbf{Z}'_{ij} and \mathbf{X}'_{ij} , respectively. For many applications and focus of our subsequent simulation study and example, random intercepts may adequately model clustering. Note that for independent responses, this marginalized ZIP model with random effects reduces to the Long et al.(2014) marginalized ZIP model.

Because ν_{ij}^C is modeled directly in this marginalized ZIP with random effects model, the k^{th} parameter of $\boldsymbol{\alpha}$, α_k , is interpreted as the subject-specific log-incidence density ratio (IDR) for the k^{th} covariate; that is, for a one-unit increase in corresponding covariate x_k , $\exp(\alpha_k)$ is the amount by which the mean ν_{ij}^C for a particular subject is multiplied, which is the same interpretation as in a Poisson random effects model. The direct modeling of ν_{ij}^C rather than the Poisson process mean μ_{ij}^C in Section 2 provides marginal mean inference often of interest to researchers.

For $\boldsymbol{\theta} = (\boldsymbol{\gamma}', \boldsymbol{\alpha}', \boldsymbol{\Sigma})'$, the log-likelihood for this marginalized ZIP model with random effects can be written

$$l(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^K \log \int_{-\infty}^{+\infty} \left[\prod_{j=1}^{T_i} P(Y_{ij} = y_{ij} | \mathbf{b}_i, \boldsymbol{\theta}) \right] \Phi(\mathbf{b}_i) d\mathbf{b}_i, \quad (6)$$

where Φ is the multivariate normal density $(\mathbf{0}, \boldsymbol{\Sigma})$. Augmenting the ZIP likelihood presented in (3) similar to the Long et al.(2014) reparameterization, the marginalized ZIP likelihood redefines $\mu_{ij}^C = \exp(\delta_{ij}^C)$, where δ_{ij}^C is not necessarily a linear function of covariates. Following from the ZIP likelihood specification in (3),

$$P(Y_{ij} = y_{ij} | \mathbf{b}_i, \boldsymbol{\theta}) = \left[\psi_{ij}^C + (1 - \psi_{ij}^C) e^{-\exp(\delta_{ij}^C)} \right]^{u_{ij}} \left[\frac{(1 - \psi_{ij}^C) e^{-\exp(\delta_{ij}^C)} e^{\delta_{ij}^C y_{ij}}}{y_{ij}!} \right]^{1-u_{ij}}. \quad (7)$$

Using (5) and the knowledge $\nu_{ij}^C = (1 - \psi_{ij}^C) \mu_{ij}^C$, solving for $\delta_{ij}^C = \log(\mu_{ij}^C)$ gives

$$\delta_{ij}^C = \log(N_i) + \log[1 + \exp(\mathbf{Z}'_{ij} \boldsymbol{\gamma} + \mathbf{w}'_{1ij} \mathbf{c}_i)] + \mathbf{X}'_{ij} \boldsymbol{\alpha} + \mathbf{w}'_{2ij} \mathbf{d}_i. \quad (8)$$

Rather than linking a linear function of covariates to the Poisson latent class mean, the form of μ_{ij}^C is derived to express a linear function of covariates on the marginal mean ν_{ij}^C . Through substitution of (8) into (7), this subject-specific marginalized ZIP model with random effects may be fit using SAS NLMIXED (SAS Institute Inc, 2013), which employs an adaptive Gauss-Hermite quadrature to approximate the integral of the likelihood (6) over the random effects. For the simulation study, 25 quadrature points were used, and this was increased to 50 quadrature points for the analysis of the SafeTalk efficacy trial (Lesaffre and Spiessens, 2001). Additionally, SAS NLMIXED can provide robust (empirical) standard error

estimates of the parameters, through the likelihood-based ‘sandwich’ estimator, to address model misspecification (White, 1982).

4.2. Population-averaged marginalized ZIP model for clustered data

The primary objective in the marginalized models literature (e.g. Heagerty, 1999) is to obtain parameters with marginalized (population-averaged) interpretations rather than parameters with subject-specific interpretations. In Section 4.1, we described the marginalized ZIP model with random effects, where the ‘marginalization’ is over the two latent classes of the ZIP model to achieve overall exposure effect estimates. However, because the marginalized ZIP with random effects models $\nu_{ij}^C = E(Y_{ij} | \mathbf{d}_i)$, it yields parameters with subject-specific interpretations.

For data with repeated measures, statistical analysts usually choose between methods employing subject-specific (SS) parameters (mixed models) and methods having population-average (PA) parameters (GEE), though in a few notable cases (e.g. the Gaussian mixed model) parameters have both interpretations. However, Ritz and Spiegelman (2004) and Young et al. (2007) investigate the exact nature of the relationship between SS and PA parameters for Poisson count data, using well-established methods (e.g. McCulloch and Searle, 2001). For models with log links and normally distributed random effects, the mathematical relationships between SS and PA parameters can be quite straightforward.

To explore the connection between SS and PA parameters for the marginalized ZIP model with random effects, we restate model (5) as

$$\begin{aligned}\text{logit}(\psi_{ij}^C) &= \mathbf{Z}'_{ij} \boldsymbol{\gamma}^{SS} + \mathbf{w}'_{1ij} \mathbf{c}_i \\ \log(\nu_{ij}^C) &= \mathbf{X}'_{ij} \boldsymbol{\alpha}^{SS} + \log(N_i) + \mathbf{w}'_{2ij} \mathbf{d}_i,\end{aligned}$$

where the *SS* superscript indicates that subject-specific interpretations are appropriate for these parameters. Then

$$E(Y_{ij} | \mathbf{d}_i) = \exp[\mathbf{X}'_{ij} \boldsymbol{\alpha}^{SS} + \log(N_i) + \mathbf{w}'_{2ij} \mathbf{d}_i]$$

and

$$\begin{aligned}E(Y_{ij}) &= E[E(Y_{ij} | \mathbf{d}_i)] \\ &= N_i \exp(\mathbf{X}'_{ij} \boldsymbol{\alpha}^{SS}) E(\exp(\mathbf{w}'_{2ij} \mathbf{d}_i)) \\ &= N_i \exp(\mathbf{X}'_{ij} \boldsymbol{\alpha}^{SS}) \exp(0.5 \mathbf{w}'_{2ij} \boldsymbol{\Sigma}_{22} \mathbf{w}_{2ij})\end{aligned}\quad (9)$$

where $\mathbf{d}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{22})$. From (9), defining $\nu_{ij}^M = E(Y_{ij})$,

$$\log(\nu_{ij}^M) = \mathbf{X}'_{ij} \boldsymbol{\alpha}^{SS} + \log(N_i) + 0.5 \mathbf{w}'_{2ij} \sum_{22} \mathbf{w}_{2ij}.$$

Now consider the fully marginal model (10), where *PA* denotes population-averaged parameters

$$\log(\nu_{ij}^M) = \mathbf{X}'_{ij} \boldsymbol{\alpha}^{PA} + \log(N_i). \quad (10)$$

The PA parameters in (10) are multiplicatively offset from the SS parameters by the function $\exp(0.5 \mathbf{w}'_{2ij} \sum_{22} \mathbf{w}_{2ij})$ of the (*ij*)-th row of the model matrix for the random effects and respective covariance matrix. Thus, for all fixed effect covariates that do not have corresponding random effects, the respective parameters in $\boldsymbol{\alpha}^{SS}$ are equivalent to corresponding parameters in $\boldsymbol{\alpha}^{PA}$. Consider the model with only a random intercept ($\mathbf{w}'_{2ij}=1$) and $\sum_{22} = \sigma_b^2$; then

$$\log(\nu_{ij}^M) = [\alpha_0^{SS} + (\sigma_b^2/2)] + \tilde{\mathbf{X}}'_{ij} \tilde{\boldsymbol{\alpha}}^{SS} + \log(N_i),$$

where $\tilde{\mathbf{X}}'_{ij}$ and $\tilde{\boldsymbol{\alpha}}^{SS}$ contain all the covariates and corresponding parameters excluding the intercept. In this situation, $\tilde{\boldsymbol{\alpha}}^{SS}$ also have population-averaged interpretations. While analysts may choose to include further normal random effects, such as a random slope over time, all parameters without a corresponding random effect have population-averaged as well as subject-specific interpretations because of the log link and normal random effects.

5. Simulation study

To examine the properties of the marginalized ZIP model with random effects, a simulation study was performed using SAS 9.3 NL MIXED. Let Y_{ij} be a zero-inflated Poisson outcome for the *i*th participant at time *j*, and let g_i be a time-constant exposure variable of interest for each subject. The simulation scenario is motivated by the constant treatment assignment in the SafeTalk clinical trial. In the SafeTalk motivating example, Y_{ij} is the UAVI count outcome and g_i is an indicator of randomization to the SafeTalk intervention group. For this simulation study, three time points were used with $I(j=2)$ and $I(j=3)$ being the indicators of whether an observation occurs at follow-up time 2 or 3. Data were simulated using the marginalized ZIP model with random effects given by

$$\begin{aligned} \text{logit}(\psi_{ij}^C) &= \gamma_0 + \gamma_1 I(j=2) + \gamma_2 I(j=2)g_i + \gamma_3 I(j=3) + \gamma_4 I(j=3)g_i + c_i \\ \log(\nu_{ij}^C) &= \alpha_0 + \alpha_1 I(j=2) + \alpha_2 I(j=2)g_i + \alpha_3 I(j=3) + \alpha_4 I(j=3)g_i + d_i, \end{aligned} \quad (11)$$

where c_i, d_i are bivariate normal random intercepts with variances σ_1^2, σ_2^2 and correlation ρ used to account for correlated outcomes for the *i*th participant. For a fixed sample, g_i was generated from a Bernoulli(0.5) and (c_i, d_i) were generated from a bivariate normal

distribution with $\sigma_1^2 = \sigma_2^2 = 1$ and $\rho = -0.25$. In most scenarios, we expect that the probability of an excess zero will be negatively correlated with the marginal mean as in our motivating example.

The parameters ψ_{ij}^C and ν_{ij}^C are calculated with the specified values of $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$. Using the first model part in equation (11) and $\mu_{ij}^C = \nu_{ij}^C / (1 - \psi_{ij}^C)$, excess zeros and Poisson counts were randomly generated. Define $\boldsymbol{\psi}_i^C = (\psi_{i1}^C, \psi_{i2}^C, \psi_{i3}^C)$ and $\boldsymbol{\nu}_i^C = (\nu_{i1}^C, \nu_{i2}^C, \nu_{i3}^C)$. These simulations were performed for 100, 300, 500 and 1000 participants, respectively, with $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ vectors chosen such that $\boldsymbol{\psi}_i^C = \{0.45, 0.50, 0.50\}$, $\boldsymbol{\nu}_i^C = \{1.75, 1.70, 1.70\}$ for $g_i = 0$ and $\boldsymbol{\psi}_i^C = \{0.45, 0.65, 0.65\}$, $\boldsymbol{\nu}_i^C = \{1.75, 1.275, 1.11\}$ for $g_i = 1$. These marginal mean specifications correspond to IDR values of (0.97, 0.97) in the unexposed group and (0.75, 0.65) in the exposed group across follow-up time 2 and 3. Across the combinations of g_i and time j , the total percent of zero counts ranged from 44% to 69%. For each cluster size, 1,000 simulations were attempted, but the SAS NLMIXED procedure failed to converge for 5% iterations. Others have reported difficulties in convergence of ZIP models with random effects (Min and Agresti, 2005).

Table 1 presents the raw and percent relative median bias, simulation standard deviation and median standard errors (model-based and robust) of each estimate from the marginalized ZIP model. The vectors of parameters to simulate the above values of $\boldsymbol{\psi}_{ij}$ and $\boldsymbol{\nu}_{ij}$ are $\boldsymbol{\gamma} = \{-0.2007, 0.2007, 0.8197, 0.2007, 0.8197\}$ and $\boldsymbol{\alpha} = \{0.5596, -0.0290, -0.2877, -0.0290, -0.4263\}$.

The raw median bias is small for each cluster size K , and both the model-based and robust standard errors are close to the standard deviation of the parameter estimates, indicating adequate estimation of the variability in parameter estimates. The largest percent relative bias in estimating $\boldsymbol{\alpha}$ occur for α_0 , α_1 and α_3 . The parameters α_1 and α_3 are the log-IDR for times 2 and 3 relative to time 1 for the unexposed groups and have true values very close to 0, inflating the relative bias. For $K = 500$, the true α_3 is -0.0290 and the median bias is 0.00638, yielding a percent relative median bias of -22.0%. Despite these inflated relative median biases for true parameters near zero, the marginalized ZIP with random effects model has low bias across the simulation scenarios.

In addition to the marginalized ZIP model with random effects, both a Poisson population-average model with GEE estimation and a Poisson random intercept model were fit in SAS 9.3 GENMOD and NLMIXED, respectively, for comparison in estimating the population-average IDR. The model for the Poisson population-average model is

$$\log(\nu_{ij}^M) = \alpha_0^* + \alpha_1 I(j=2) + \alpha_2 I(j=2)g_i + \alpha_3 I(j=3) + \alpha_4 I(j=3)g_i, \quad (12)$$

with unstructured covariance and model-based standard errors scaled with Pearson's chi-square for potential overdispersion, as well as empirical (robust) standard errors; (11)

expresses the model for the Poisson random intercept model with ν_{ij}^C representing the Poisson mean $E(Y_{ij}|d_i)$. As discussed in Section 4.2, the parameters $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ from

(11) have population-average interpretations (since intercept is the only random effect), so the parameters from the Poisson population-average model with GEE estimation in (12) are estimating the same quantities. For time 2, Table 2 presents the relative median bias in estimating both the log-IDR and IDR corresponding to $\{a_1, a_2, a_3, a_4\}$ for all three models, as well as the 95% Wald-type coverage probabilities and power.

Note that the marginalized ZIP model with random effects has lower percent relative median bias for most scenarios, as well as appropriate coverage. With the model-based standard errors in the Poisson random intercept model, the coverage probabilities are much less than the expected 0.95, indicating these standard errors are underestimating the extra-Poisson variability in the ZIP data due to the excess zeros. The robust standard errors for both Poisson models provide appropriate coverage of the IDR, but the marginalized ZIP model has increased power to detect significance in IDR over both Poisson methods, particularly for a_2, a_4 where parameter estimates deviate further from 0. Using the Pearson-scaled model-based standard errors, the Poisson PA models have very similar absolute bias in many scenarios and only slightly less coverage than the marginalized ZIP model, but there is a marked difference in power with the Poisson PA model having significantly less ability to detect differences in mean IDR.

6. Analysis of the SafeTalk efficacy trial

In safer sex counseling for people living with HIV/AIDS, an outcome of interest is Unprotected Anal or Vaginal Intercourse acts (UAVI), defined as the number of unprotected sexual acts with any partner. Researchers developed the motivational interview-based intervention SafeTalk to reduce the number of unprotected sexual acts (Golin et al., 2007; Golin et al., 2010; Golin et al., 2012). For the clinical trial examining SafeTalk efficacy, participants were randomized to receive either SafeTalk intervention counseling or a control nutritional counseling. These participants completed questionnaires about both nutritional and sexual behavior at baseline as well as at three follow-up visits spaced at four-month intervals. After data cleaning, the sample sizes at each time point are 476, 399, 363 and 301. The overall percentage of zero UAVI counts across both treatment groups and all visits was 83.1%.

While some researchers may choose to focus on the latent class interpretations provided by the ZIP model with random effects, our collaborative researchers are interested in quantifying the effect of the SafeTalk intervention over time among the entire randomized population, leading to a choice of marginal mean inference provided by the marginalized ZIP model with random effects. In order to evaluate the efficacy of the SafeTalk intervention over time, the marginalized ZIP with random effects is fit to the UAVI counts at all four time points. The model of interest is

$$\begin{aligned} \text{logit}(\psi_{ij}^C) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 I(j=2) + \gamma_4 I(j=2)g_i + \gamma_5 I(j=3) + \gamma_6 I(j=3)g_i + \gamma_7 I(j=4) + \gamma_8 I(j=4)g_i + c_i \\ \log(\nu_{ij}^C) &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 I(j=2) + \alpha_4 I(j=2)g_i + \alpha_5 I(j=3) + \alpha_6 I(j=3)g_i + \alpha_7 I(j=4) + \alpha_8 I(j=4)g_i + d_i, \end{aligned}$$

where c_i, d_i are bivariate normal random intercepts with covariance $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$, j is the visit number, g_i is an indicator of randomization to SafeTalk intervention group, and x_{i1} and x_{i2} are fixed effects for study site.

Using SAS NL MIXED (for which the code is presented in the Appendix), the SafeTalk analysis results are presented in Table 3. The contrast testing treatment effect over time $H_0 : (a_4, a_6, a_8)' = (0, 0, 0)'$ is highly significant (Robust-Wald $p = 0.0003$), indicating that the SafeTalk intervention affects UAVI count. At the second follow-up visit, for which the IDR (and 95% Wald-type robust confidence interval) is 0.542 (0.260, 1.128), a participant randomized to SafeTalk has 46% fewer unprotected sexual acts with any partner than he or she would have if randomized to the nutritional intervention. Because the only random effect for the above model is a random intercept, the parameters associated with treatment effect from this analysis additionally have population-averaged interpretations. Thus, at the second follow-up visit, those participants randomized to SafeTalk had on average 46% fewer unprotected sexual acts with any partner than the participants randomized to the nutritional intervention. The SafeTalk intervention appears to have the largest effect on UAVI count at the first follow-up survey, where the estimated IDR (and 95% Wald-type robust confidence interval) of treatment effect is 0.280 (0.145, 0.542). By the third follow-up survey, we observe less reduction in UAVI count due to SafeTalk, with an IDR of 0.769 (0.307, 1.928). Figure 1 displays the predicted mean UAVI over time, as well as the IDR of treatment at each time point. The SafeTalk intervention appears to have a significant effect in reducing UAVI counts at the first follow-up visit, but the difference between the two treatment groups is reduced at each subsequent follow-up visit. From Figure 1 and Table 3, note that the nutritional control arm has a significant reduction in predicted UAVI count at the final visit, numerically represented through a_7 . Additionally, note that the correlation between the random intercepts, estimated to be -0.79, is highly significant, indicating those participants with higher expected UAVI counts have lower odds of excess zero latent class membership. In fact, if independence of the random intercepts is assumed, individual parameter estimates from the marginalized ZIP model differ as much as 40%, leading us to recommend the inclusion of correlated random effects in the two processes.

When the SafeTalk data are examined using a Poisson population-average model with GEE estimation and empirical standard errors, the Wald contrast with 3 degrees of freedom testing treatment effect is non-significant ($p=0.8259$). At the second follow-up, the GEE model estimates the IDR to be 0.768 with 95% Wald-type model-based and empirical confidence intervals (0.391, 1.508) and (0.403, 1.466), respectively. Using the Poisson random intercept model, the treatment efficacy contrast is significant when using the model-based standard errors ($p=0.0303$) but non-significant when robust standard errors are used ($p=0.8446$). At the second follow-up, the random intercept model estimates the IDR to be 0.711 with model-based and robust 95% Wald-type confidence intervals of (0.556, 0.908) and (0.336, 1.502). Because the simulations in Section 5 suggest that the model-based standard errors in the Poisson random intercept model underestimate the variability due to the excess zero process, the conclusions of the robust methods are preferred.

To highlight the differences between the proposed marginalized ZIP model with random effects and the ZIP model with random effects from Section 2, the latter was also fit to the SafeTalk data, given by

$$\begin{aligned} \text{logit}(\psi_{ij}^C) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 I(j=2) + \gamma_4 I(j=2)g_i + \gamma_5 I(j=3) + \gamma_6 I(j=3)g_i + \gamma_7 I(j=4) + \gamma_8 I(j=4)g_i \\ \log(\mu_{ij}^C) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 I(j=2) + \beta_4 I(j=2)g_i + \beta_5 I(j=3) + \beta_6 I(j=3)g_i + \beta_7 I(j=4) + \beta_8 I(j=4)g_i + d_i, \end{aligned}$$

where $d_i \sim N(0, \sigma^2)$. For this model, the contrast of treatment effect is highly significant ($p < 0.0001$) with $\beta_4 = -0.96$, $\beta_6 = -0.89$, and $\beta_8 = -0.42$. In contrast to the marginalized ZIP model with random effects and the Poisson models which model the marginal mean directly, these traditional ZIP parameter estimates are the log-IDR for treatment among the non-excess zero latent class. *Among the non-excess zero latent class*, those participants randomized to SafeTalk had 62%, 59% and 35% fewer UAVI acts than those participants randomized to control at the first, second and third follow-up visits, respectively.

7. Conclusion

Motivated by the aim to estimate overall exposure effects for correlated count observations with excess zeroes, we have proposed a marginalized ZIP model with random effects. Since the overall subject-specific mean is modeled directly, the parameters from this new model allow subject-specific inference rather than inference on the latent class components of the subject-specific ZIP model. Additionally, when the log link is used for the marginal mean and normal random effects are used, those parameters without corresponding random effects have both subject-specific and population-average interpretations.

The new marginalized ZIP model with random effects was applied to repeated measures data from a clinical trial to reduce risky sexual behavior among HIV-positive individuals. We observed that the robust standard errors for intervention effect parameters were notably larger than their model-based counterparts, suggesting the counts are overdispersed. Future research could extend the marginalized ZIP model for random effects to handle overdispersion as well as excess zeroes.

In the SafeTalk data, missing at random (MAR) is assumed, meaning that the probability of attending a visit and having UAVI recorded depends only on observed data. There is evidence that the assumption of missing completely at random (MCAR) is not valid because those participants with any risky baseline behavior have 54.1% retention at the final visit versus 65.6% retention in those with non-risky baseline behavior. Maximum likelihood estimation of the marginalized ZIP with random effects model described in Section 4.1 provides valid inference under MAR when the model is correctly specified (Ibrahim and Molenberghs, 2009).

In the simulation study, we experienced convergence issues similar to ZIP model instability occasionally associated with those effects in the excess zero portion of ZIP models (Min and Agresti, 2005). Future research includes exploring other optimization techniques with more stability for zero-inflated models, such as the Bayesian methods proposed in Neelon et al.

(2010). In addition to other computational strategies, the relatively small number of simulation iterations with failed NLMIXED convergence could possibly be lessened by reducing the complexity of the excess zero model. In marginalized ZIP regression, the excess zero model parameters are considered nuisance parameters, as the primary hypotheses concern the marginalized mean. However, as unintended constraints on the marginal means can be introduced by the omission of covariates in the excess zero model, the reduction of the excess zero model should be carefully considered and rigorously justified.

In contrast to exclusive reliance on fit statistics or conjectures about data-generating mechanisms as a basis for selecting the type of count regression model for handling data with many zeros, we affirm that the choice between marginalized ZIP, ZIP and hurdle model classes should be motivated by the interpretations desired. When inference upon the overall marginal mean is desired, the marginalized ZIP model is preferred. The *a priori* choice of model class for zero-inflation is analogous to the *a priori* choice between PA and SS models for longitudinal data (Heagerty, 1999) where the interpretations of regressions parameters differ in models with non-identity link functions.

Rather than marginalizing over the two processes of the ZIP model, the ZIP model with random effects could be marginalized over the random effects, similar to the marginalized hurdle model in Lee et al. (2011). Additionally, one could marginalize over both the random effects and two ZIP processes to achieve a ‘doubly’ marginalized ZIP model. As shown in Section 4.2, the marginalized ZIP model can be used not only for subject-specific inference on overall conditional effects but also for population-average inference for overall effects in many problems.

Acknowledgments

This work was supported in part by National Institute of Health (NIH) grants T32ES007018 (NIEHS), T32HD007237 (NICHD), R01MH069989 (NIMH), R01ES020619 (NIEHS), U54GM104942 (NIGMS), and University of North Carolina Center for AIDS Research AI50410. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH. This work was conducted as part of the first author's doctoral dissertation in the Department of Biostatistics at the University of North Carolina at Chapel Hill (Long 2013).

8. Appendix

The following SAS NLMIXED code was used for the SafeTalk motivating example.

```
proc nlmixed data=safetalk seed=31415;

  parms b0 0 b1 0 b2 0 b3 0 b4 0 b5 0 b6 0 b7 0 b8 0

        a0 0 a1 0 a2 0 a3 0 a4 0 a5 0 a6 0 a7 0 a8 0

        sigma1 1 sigma12 0 sigma2 1;

  /* linear predictor for the zero-inflation probability */
```

```

logit_psi = a0 + a1*site2 + a2*site3 + a3*v2 + a4*v2*st + a5*v3 + a6*v3*st
          + a7*v4 + a8*v4*st + c1;

*logit(\psi)=Z\gamma + c;

/* useful functions of \psi */

psi1 = exp(logit_psi)/(1+exp(logit_psi));

*\psi = exp(Z\gamma+c)/(1+exp(Z\gamma+c));

psi2 = 1/(1+exp(logit_psi));

*1-\psi = (1+exp(Z\gamma+c))^-1;

/* Overall mean \nu */

log_nu = b0 + b1*site2 + b2*site3 + b3*v2 + b4*v2*st + b5*v3 + b6*v3*st
        + b7*v4 + b8*v4*st + d1;

delta = log(psi2**(-1)) + log_nu;

/* Build the mZIP + RE log likelihood */

if outcome=0 then

    ll = log(psi1 + psi2*(exp(-exp(delta))));

else ll = log(psi2) - exp(delta) + outcome*(delta) - lgamma(outcome + 1);

model outcome ~ general(ll);

random c1 d1~normal([0,0],[sigma1,sigma12,sigma2]) SUBJECT=urn;

contrast "TX" b4, b6, b8;

run;

```

References

- Albert JM, Wang W, Nelson S. Estimating overall exposure effects for zero-inflated regression models with application to dental caries. *Statistical Methods in Medical Research*. 2014; 23(3):257–278. [PubMed: 21908419]
- Buu A, Li R, Tan X, Zucker RA. Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine*. 2012; 31(29):4074–4086. [PubMed: 22826194]

- Dobbie M, Welsh A. Theory & Methods: Modelling correlated zero-inflated count data. *Australian & New Zealand Journal of Statistics*. 2001; 43(4):431–444.
- Ghosh P, Tu W. Assessing sexual attitudes and behaviors of young women: a joint model with nonlinear time effects, time varying covariates, and dropouts. *Journal of the American Statistical Association*. 2009; 104(486):474–485.
- Gilthorpe M, Frydenberg M, Cheng Y, Baelum V. Modelling count data with excessive zeros: The need for class prediction in zero-inflated models and the issue of data generation in choosing between zero-inflated and generic mixture models for dental caries data. *Statistics in Medicine*. 2009; 28(28):3539–3553. [PubMed: 19902494]
- Golin C, Davis R, Przybyla S, Fowler B, Parker S, Earp J, Quinlivan E, Kalichman S, Patel S, Grodensky C. Safetalk, a multicomponent, motivational interviewing-based, safer sex counseling program for people living with HIV/AIDS: A qualitative assessment of patients' views. *AIDS Patient Care and STDs*. 2010; 24(4):237–245. [PubMed: 20377435]
- Golin C, Earp J, Grodensky C, Patel S, Suchindran C, Parikh M, Kalichman S, Patterson K, Swygard H, Quinlivan E, Amola K, Chariyeva Z, Groves J. Longitudinal effects of safetalk, a motivational interviewing-based program to improve safer sex practices among people living with hiv/aids. *AIDS and Behavior*. 2012; 16(5):1182–1191. [PubMed: 21964975]
- Golin C, Patel S, Tiller K, Quinlivan E, Grodensky C, Boland M. Start talking about risks: development of a motivational interviewing-based safer sex program for people living with HIV. *AIDS and Behavior*. 2007; 11:72–83.
- Hall D, Zhang Z. Marginal models for zero inflated clustered data. *Statistical Modelling*. 2004; 4(3): 161–180.
- Hall DB. Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics*. 2000; 56:1030–1039. [PubMed: 11129458]
- Heagerty P. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*. 1999; 55(3):688–698. [PubMed: 11314994]
- Heilbron D. Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*. 1994; 36:531–547.
- Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. *Test*. 2009; 18(1): 1–43. [PubMed: 21218187]
- Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*. 1992; 34:1–14.
- Lee K, Joo Y, Song J, Harper D. Analysis of zero-inflated clustered count data: A marginalized model approach. *Computational Statistics & Data Analysis*. 2011; 55(1):824–837.
- Lesaffre E, Spiessens B. On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2001; 50(3):325–335.
- Long, DL. Ph D thesis. Department of Biostatistics, University of North Carolina; Chapel Hill: 2013. Marginalized Zero-inflated Poisson Regression.
- Long DL, Preisser JS, Herring AH, Golin CE. A marginalized zero-inflated poisson regression model with overall exposure effects. *Statistics in Medicine*. 2014; 33(29):5151–5165. [PubMed: 25220537]
- McCulloch, C.; Searle, S. *Generalized, Linear, and Mixed Models*. Wiley; 2001.
- Min Y, Agresti A. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*. 2005; 5:1–19.
- Mullahy J. Specification and testing of some modified count data models. *Journal of Econometrics*. 1986; 33:341–365.
- Mwalili SM, Lesaffre E, Declerck D. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Statistical Methods in Medical Research*. 2008; 17(2):123–139. [PubMed: 17698937]
- Neelon B, O'Malley A, Normand S. A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling*. 2010; 10(4):421–439. [PubMed: 21339863]

- Preisser JS, Stamm JW, Long DL, Kincade ME. Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. *Caries Research*. 2012; 46(4):413–423. [PubMed: 22710271]
- Ritz J, Spiegelman D. Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research*. 2004; 13(4):309–323.
- SAS Institute Inc. SAS/STAT Software, The NLMIXED Procedure Cary, NC Version 9.3. 2013. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#nlmixed_toc.htm
- White H. Maximum likelihood estimation of misspecified models. *Econometrica*. 1982; 50(1):1–25.
- Yau K, Wang K, Lee A. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal*. 2003; 45(4):437–452.
- Young M, Preisser J, Qaqish B, Wolfson M. Comparison of subject-specific and population averaged models for count data from cluster-unit intervention trials. *Statistical Methods in Medical Research*. 2007; 16(2):167–184. [PubMed: 17484299]

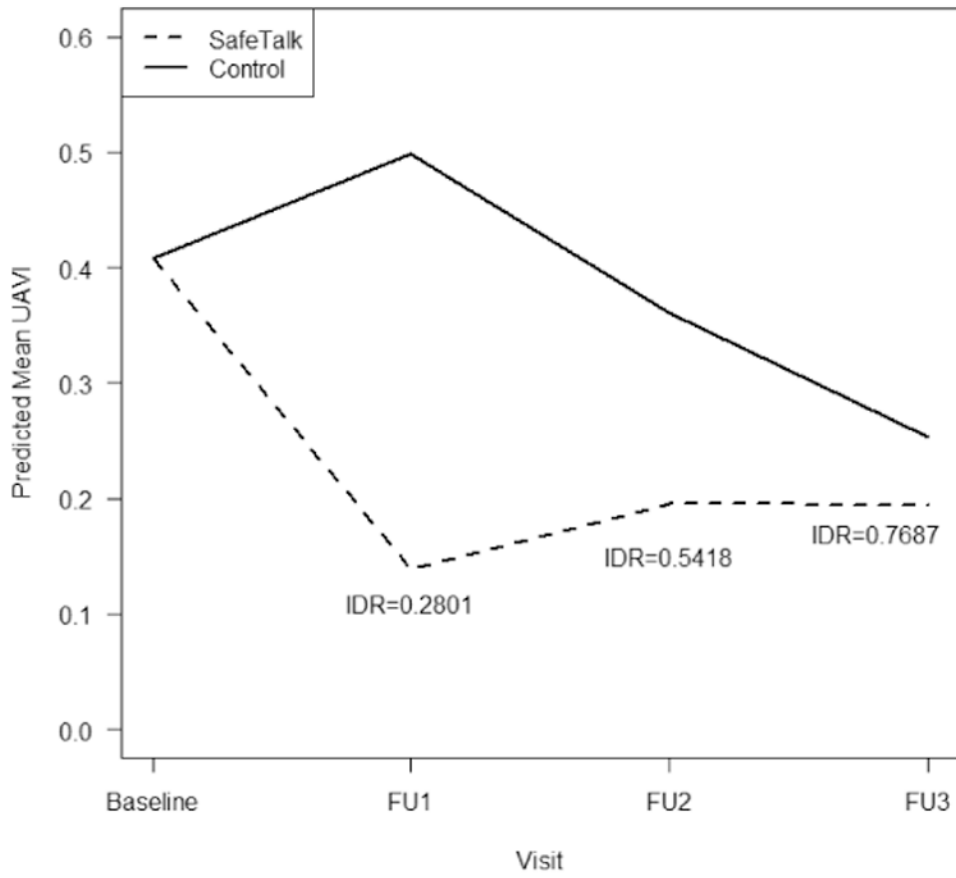


Fig 1. Marginalized ZIP with random effects ($c_i = d_i = 0$) predicted UAVI means over time. Follow-up visits (FU1, FU2, FU3) are at four, eight and twelve months post-randomization.

Table 1
Marginalized ZIP with RE Performance with 1,000 Simulations and Varying Number of Subjects

Parameter	K	Raw Median Bias	Percent Relative Median Bias	Simulation Std Dev	Median Std Error	Median Robust Std Error
γ_0	100	-0.003	1.59	0.2061	0.2518	0.2521
	300	-0.035	17.26	0.1510	0.1554	0.1557
	500	-0.005	2.68	0.1057	0.1159	0.1155
	1000	0.013	-6.34	0.0777	0.0870	0.0870
γ_1	100	0.006	3.19	0.3369	0.3888	0.3656
	300	-0.017	-8.51	0.2391	0.2379	0.2362
	500	-0.011	-5.50	0.1727	0.1798	0.1747
	1000	-0.004	-2.08	0.1340	0.1330	0.1324
γ_2	100	-0.026	-3.18	0.4213	0.4808	0.4744
	300	0.011	1.30	0.2886	0.2924	0.2904
	500	0.000	-0.04	0.2068	0.2202	0.2179
	1000	0.001	0.11	0.1606	0.1628	0.1627
γ_3	100	-0.010	-4.83	0.3604	0.3857	0.3667
	300	0.000	-0.18	0.2514	0.2384	0.2368
	500	-0.012	-5.89	0.1695	0.1792	0.1730
	1000	-0.009	-4.59	0.1333	0.1329	0.1330
γ_4	100	0.003	0.33	0.4135	0.4868	0.4775
	300	0.000	0.02	0.3056	0.2944	0.2941
	500	-0.003	-0.35	0.2069	0.2220	0.2193
	1000	-0.002	-0.20	0.1664	0.1642	0.1640
α_0	100	0.064	11.49	0.1264	0.1685	0.1657
	300	0.124	22.09	0.0976	0.1080	0.1074
	500	0.077	13.71	0.0661	0.0803	0.0782
	1000	0.056	10.08	0.0530	0.0617	0.0613
α_1	100	-0.003	9.19	0.1661	0.1765	0.1582

Parameter	K	Raw Median Bias	Percent Relative Median Bias	Simulation Std Dev	Median Std Error	Median Robust Std Error	
α_2	300	-0.002	5.39	0.1159	0.1207	0.1200	
	500	0.005	-16.51	0.0812	0.0847	0.0820	
	1000	-0.003	10.01	0.0680	0.0655	0.0652	
α_2	100	0.009	-3.03	0.2530	0.2799	0.2632	
	300	0.005	-1.79	0.1826	0.1823	0.1822	
	500	-0.004	1.27	0.1251	0.1302	0.1287	
α_2	1000	0.003	-0.88	0.0997	0.0999	0.0997	
	α_3	100	0.000	0.32	0.1627	0.1726	0.1553
		300	-0.005	16.85	0.1243	0.1201	0.1197
500		0.006	-22.00	0.0847	0.0847	0.0815	
α_3	1000	0.004	-13.75	0.0675	0.0654	0.0652	
	α_4	100	0.007	-1.64	0.2617	0.2823	0.2640
		300	0.001	-0.22	0.1863	0.1848	0.1844
500		-0.007	1.58	0.1244	0.1324	0.1295	
α_4	1000	-0.001	0.30	0.0987	0.1006	0.1005	

True parameter values: $\boldsymbol{\gamma} = \{-0.2007, 0.2007, 0.8197, 0.2007, 0.8197\}$

$\boldsymbol{\alpha} = \{0.5596, -0.0290, -0.2877, -0.0290, -0.4263\}$

Table 2
Percent Relative Median Bias, Coverage & Power for Estimating IDR and log-IDR

		Percent Relative Median Bias (IDR) [†]	Percent Relative Median Bias (Log-IDR)	Model-Based Coverage	Model-Based Power	Robust Coverage	Robust Power		
α_1	100								
		mZIP	-0.27	9.19	0.956	0.054	0.949	0.058	
		Poisson PA	-1.84	64.01	0.944	0.058	0.928	0.074	
		Poisson RI	-0.77	26.72	0.508	0.515	0.936	0.065	
		mZIP	-0.16	1.30	0.964	0.044	0.961	0.047	
		Poisson PA	-0.74	25.69	0.952	0.047	0.940	0.071	
300		Poisson RI	0.99	-34.04	0.508	0.476	0.933	0.071	
		mZIP	0.48	-16.51	0.955	0.057	0.952	0.060	
		Poisson PA	-0.08	2.66	0.961	0.049	0.939	0.071	
		Poisson RI	1.39	-47.56	0.520	0.466	0.946	0.053	
		mZIP	-0.29	10.01	0.935	0.088	0.938	0.088	
		Poisson PA	-0.17	5.83	0.961	0.040	0.949	0.059	
1000		Poisson RI	0.91	-31.18	0.506	0.525	0.943	0.057	
	α_2	100							
			mZIP	0.87	-3.03	0.949	0.426	0.944	0.427
			Poisson PA	-0.37	1.28	0.913	0.294	0.923	0.294
			Poisson RI	-2.09	7.34	0.473	0.718	0.930	0.268
			mZIP	0.52	5.39	0.942	0.333	0.945	0.341
		Poisson PA	-0.02	0.09	0.917	0.237	0.927	0.236	
300		Poisson RI	-2.36	8.29	0.472	0.711	0.933	0.216	
		mZIP	-0.36	1.27	0.952	0.675	0.946	0.682	
		Poisson PA	0.27	-0.92	0.923	0.419	0.935	0.395	
		Poisson RI	-2.33	8.21	0.492	0.846	0.941	0.375	
		mZIP	0.25	-0.88	0.956	0.841	0.954	0.837	
		Poisson PA	-0.36	1.26	0.940	0.525	0.946	0.498	
500		Poisson RI	-2.77	9.76	0.479	0.908	0.946	0.475	
	α_3	100							
			mZIP	-0.01	0.32	0.948	0.058	0.948	0.070

	Percent Relative Median Bias (IDR) [†]	Percent Relative Median Bias (Log-IDR)	Model-Based Coverage	Model-Based Power	Robust Coverage	Robust Power
Poisson PA	-0.77	26.70	0.948	0.043	0.938	0.061
Poisson RI	0.33	-11.22	0.538	0.477	0.952	0.052
mZIP	-0.49	16.85	0.943	0.066	0.943	0.071
Poisson PA	-1.05	36.51	0.959	0.037	0.946	0.062
Poisson RI	0.13	-4.44	0.515	0.483	0.939	0.061
mZIP	0.64	-22.00	0.935	0.068	0.933	0.074
Poisson PA	-0.73	25.12	0.954	0.043	0.933	0.062
Poisson RI	0.85	-29.07	0.497	0.499	0.928	0.059
mZIP	0.40	-13.75	0.948	0.067	0.946	0.073
Poisson PA	0.37	-12.70	0.941	0.067	0.953	0.047
Poisson RI	1.48	-50.64	0.502	0.483	0.942	0.068
α_4						
100 mZIP	0.70	-1.64	0.957	0.572	0.952	0.575
Poisson PA	-0.04	0.08	0.944	0.467	0.942	0.485
Poisson RI	-1.42	3.34	0.488	0.792	0.940	0.442
300 mZIP	0.09	-0.22	0.952	0.651	0.948	0.646
Poisson PA	-1.84	4.36	0.947	0.389	0.937	0.404
Poisson RI	-2.89	6.88	0.485	0.831	0.937	0.378
500 mZIP	-0.67	1.58	0.945	0.925	0.943	0.931
Poisson PA	0.41	-0.96	0.937	0.685	0.928	0.665
Poisson RI	-2.09	4.95	0.488	0.943	0.939	0.641
1000 mZIP	-0.13	0.30	0.946	0.988	0.947	0.990
Poisson PA	0.35	-0.83	0.940	0.811	0.936	0.805
Poisson RI	-2.45	5.81	0.458	0.982	0.941	0.755

* mZIP: Marginalized ZIP model with random effects;

Poisson PA: Poisson population average model with GEE estimation;

Poisson RI: Poisson random intercept model

[†] Simulated scenario: True IDRs $\exp(\alpha_1) = 0.97$, $\exp(\alpha_2) = 0.75$, $\exp(\alpha_3) = 0.97$, $\exp(\alpha_4) = 0.65$

Table 3
Marginalized ZIP Model with Random Effects Results: SafeTalk efficacy trial

	Parameter	Parameter Estimate	Model-Based Std Error	Robust Std Error
Zero-Inflation Model				
Intercept	γ_0	2.1187	0.3581	0.3665
Site 2	γ_1	0.1026	0.4311	0.4184
Site 3	γ_2	0.2445	0.8782	0.9548
Follow-up 1	γ_3	1.2709	0.3287	0.3468
Follow-up 1*Treatment	γ_4	0.8849	0.4144	0.4627
Follow-up 2	γ_5	1.7071	0.3611	0.7011
Follow-up 2*Treatment	γ_6	-0.6021	0.5022	0.9185
Follow-up 3	γ_7	1.0214	0.4577	0.6881
Follow-up 3*Treatment	γ_8	-0.3331	0.6034	1.0968
Marginalized Mean Model				
Intercept	a_0	-0.8966	0.2803	0.2965
Site 2	a_1	0.0362	0.2941	0.2893
Site 3	a_2	-0.0220	0.6191	0.6442
Follow-up 1	a_3	0.2011	0.1471	0.1969
Follow-up 1*Treatment	a_4	-1.2725	0.2197	0.3365
Follow-up 2	a_5	-0.1217	0.1632	0.2264
Follow-up 2*Treatment	a_6	-0.6128	0.2082	0.3742
Follow-up 3	a_7	-0.4762	0.2203	0.3521
Follow-up 3*Treatment	a_8	-0.2630	0.2611	0.4691
Variance Parameters [†]				
	σ_{11}	9.7487	2.1328	2.4313
	σ_{12}	-4.5957	0.8270	0.7345
	σ_{22}	3.4461	0.6929	0.6599

$$^{\dagger} \hat{\rho} = \hat{\sigma}_{12} / \left(\sqrt{\hat{\sigma}_{11} \hat{\sigma}_{22}} \right) = -0.79$$