Bayesian Modeling of Overdispersed Longitudinal Zero-Inflated Binomial Data

Benjamin W. Rogers and Robert E. Weiss

March 25, 2025

Abstract

We develop a class of longitudinal overdispersed zero-inflated binomial (LOZIB) models for use in overdispersed repeated measures count data with a large number of zero observations. These LOZIB models extend previous zero-inflated regression models by incorporating observation-level random effects (OLRE), which allow more flexibility to account for excess variability as well as correlation over time through the specification of within-individual covariance models. We present an analysis of the number of days of heavy drinking in a study of screening, brief intervention, and referral to treatment (SBIRT) using LOZIB models with 10 different covariance models, and show all of them to outperform previously developed subject level random intercept zero-inflated models. These models are implemented in a Bayesian framework using Stan software.

1 Introduction

It is common for count data to have excess zeros relative to standard distributions, known as zero-inflation. This often happens in settings where zeros have a special meaning and can be created by a process separate from that which produces the counts. For example, we consider the number of days of heavy drinking in a population of high

risk individuals. We observe a large number of zeros because not all subjects engage in heavy drinking. Thus, zero counts can come from two different processes – a person may engage in no heavy drinking because this is not an activity they engage in, or a person may occasionally drink heavily, but happen to not do so during the observed time interval. We call these structural and random zeros, respectively.

Our interest is in estimating the effect of a Screening, Brief Intervention and Referral to Treatment (SBIRT) intervention to reduce alcohol and drug use in a population of substance users seeking mental health treatment (Karno et al., 2021). SBIRT has received attention recently as a treatment approach for substance use disorder (SUD) (Saitz et al., 2014; Glass et al., 2015; Barata et al., 2017; Tetrault et al., 2020). We analyze data consisting of 718 patients aged 18 and older randomized into either SBIRT or standard of care treatment groups. To be eligible for the study, subjects had to have a mental health disorder and report use of cannabis, stimulants or one or more days of heavy drinking (≥ 5 drinks for men, ≥ 4 drinks for women). Subjects were brought in for a baseline visit, and followed up with at 3, 6, and 12 months – at each visit, a number of variables were recorded including the number of days out of the previous 90 that each subject engaged in heavy alcohol use. As subjects were eligible for the study for using any of a variety of substances, a large number of zero counts were observed in days of heavy drinking (Figure 1).

A standard approach to modeling such data are a class of mixture models known as zero-inflated models (Cohen Jr, 1966; Johnson et al., 2005; Ridout et al., 1998), which are comprised of two parts – a zero model and a count model. The zero model uses a Bernoulli distribution to model a latent binary variable representing whether or not an observation is a structural zero. In modeling heavy drinking, this binary variable represents whether or not a subject may engage in heavy drinking. Conditional on not being a structural zero, the count model uses a count distribution, such as a binomial, negative binomial or Poisson distribution, to model the observations that are not structural zeros.

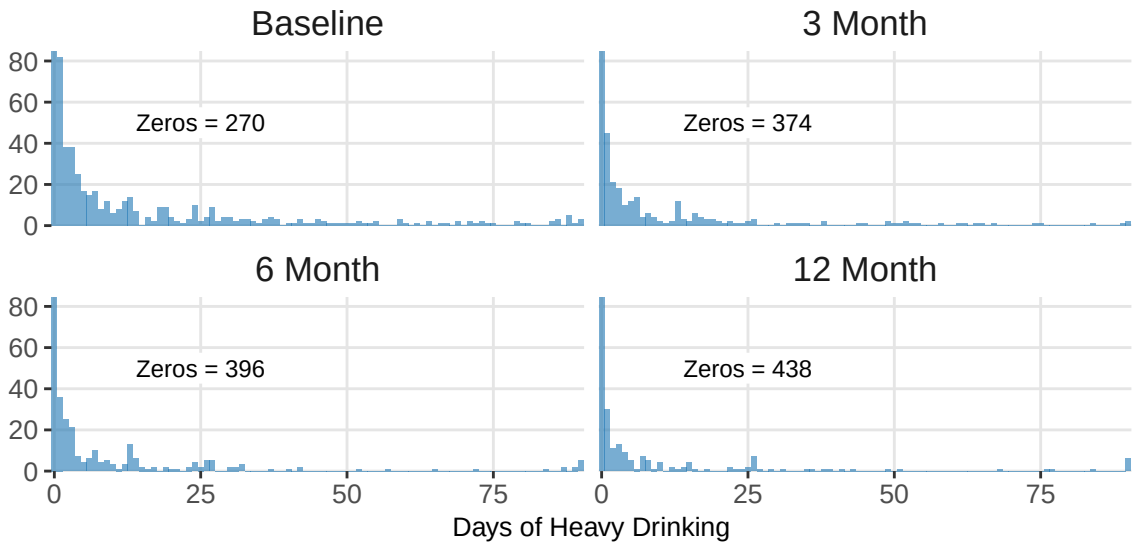

Figure 1: Self-reported days out of the previous 90 in which subjects engaged in heavy drinking at baseline and 3, 6 and 12 month follow-ups. The distribution of responses is both zero-inflated and over-dispersed.

Covariates can be controlled for using generalized linear regression models in both the zero and count models (Mullahy, 1986; Lambert, 1992; Heilbron, 1994). For longitudinal repeated measures data, Hall (2000) developed zero-inflated binomial models using subject-level random intercepts in the count model to account for within-individual correlation over time. Yau and Lee (2001) extend Hall's model by including independent subject-level random intercepts in both the zero and count models. Lee et al. (2006) include include independent cluster and subject-level random effects in both parts of the zero-inflated model. Min and Agresti (2005) make the argument that there is likely correlation between an individual's zero and count model means, which they model using a bivariate normal distribution for the zero and count model random intercepts. In addition, Buu et al. (2012) developed a version of Min and Agresti's model which includes non-linear time trends in both the zero and count regression models. Other authors have extended the zero-inflated Poisson model for multivariate zero-inflated outcomes Wu et al. (2023); Liu and Tian (2015)

We develop models in a Bayesian framework, which allows for numerous advantages including the ability to incorporate prior information and to make inference on possibly

complex functions of model parameters (Gelman et al., 2013). Frequentist methods to produce inference on covariate-adjusted marginal means have been developed for zero-inflated models (Albert et al., 2014; Preisser et al., 2016). Using a Bayesian approach, posterior distributions for such inferences are straightforward to compute. Several authors have proposed Bayesian approaches to zero-inflated models. Rodrigues (2003) and Ghosh et al. (2006) present Bayesian zero-inflated models, using a latent variable data augmentation approach to aid in posterior sampling. Wen et al. (2024) present another data augmentation approach for zero-inflated models using a mixture of Pólya-Gamma distributions. Neelon et al. (2010) give a Bayesian adaptation of Min and Agresti's model, with correlated subject-level random intercepts (Spiegelhalter, 1998; Cooper et al., 2007).

Zero-inflation is one type of overdispersion - when the observed variation is larger than what is expected from a given model. However after accounting for zero-inflation, there may be other sources of overdispersion in the count model. When this occurs one may consider using a different distribution. For example the negative binomial distribution models overdispersion compared to a Poisson (Gardner et al., 1995). Hinde and Demétrio (1998) introduce a score test for deciding between a zero inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) distribution. For overdispersed binomial data, one may use a beta-binomial distribution Hu et al. (2018); Wen et al. (2024). An alternative approach for overdispersion is to include observational level random effects (OLRE) Harrison (2014). Here we explore the use of OLRE in zero-inflated binomial models as they also provide a straightforward framework through which to model longitudinal correlation.

In repeated measures studies, observations on the same individual will generally be positively correlated. Random intercept models allow for this, but do so rather rigidly. For example, it is also reasonable to expect this correlation to differ between different pairs of measurements, decreasing as time between observations increases, a quality which cannot modeled with random intercepts. Additionally, it is common to observe

heterogeneity over time, as variability often increases with time since intervention, which is also not allowed for in random intercept models.

It has been shown that choice of covariance models in longitudinal data impacts inference (Lange and Laird, 1989; Wolfinger, 1996). Models should be complex enough to reflect the data generating process, but modeling unnecessary parameters can result in a loss of power. The random intercept model has been shown to be insufficient in many settings compared to more flexible covariance models (Barr et al., 2013; Kwok et al., 2007). We explore the use of covariance models of varying complexity, ranging from the random intercept (simplest) to the unstructured (most flexible), in zero-inflated binomial models, finding that the more complex covariance structures vastly improve model performance.

We introduce a class of longitudinal overdispersed zero-inflated binomial (LOZIB) regression models. These models extend previous zero-inflated models by using observation-level random effects (OLRE) in the count model to account for additional variation and correlation over time. Within-individual correlation over time is modeled through correlation between an individual's OLRE. We consider various specifications of covariance models and compare against previously used random intercept models, finding the loZIB models to fit the data considerably better.

2 Model

Let Y_{ij} be a zero inflated count response for subject $i = 1, \ldots, N$ at visit $j = 1, \ldots, J$ where N is the total number of subjects, J is the number of visits per subject and y_{ij} is the observed data. Then we consider a two part model

$$P(Y_{ij} = 0|\pi_{ij}, \theta_{ij}) = (1 - \pi_{ij}) + \pi_{ij} f(0|\theta_{ij}), \tag{1}$$

$$P(Y_{ij} = y_{ij}|\pi_{ij}, \theta_{ij}) = \pi_{ij} f(y_{ij}|\theta_{ij}),\ y_{ij} = 1, \ldots, \infty \tag{2}$$

where $1 - \pi_{ij}$ is the probability of being a structural zero for subject i at time j, and $f(k|\theta_{ij})$ is the probability density function for some discrete distribution with parameter θ_{ij}, such as the binomial, negative binomial or Poisson. This is the form of the standard zero-inflated count model. Equation (1) follows from the fact that zeros can either be structural, with probability $1 - \pi_{ij}$, or random, with probability $\pi_{ij} f(0|\theta_{ij})$. Similarly, Equation (2) follows because to observe a positive count, that observation must not be a structural zero, which happens with probability π_{ij}, and is then modeled with the count distribution $f(y_{ij}|\theta_{ij})$.

In the SBIRT application, π_{ij} is the probability that person i is a heavy drinker at visit j. If a subject is a heavy drinker, we model the number of days out of 90 in which they drink heavily with distribution function $f(y_{ij}|\theta_{ij})$ which follows a binomial$(90, \theta_{ij})$ distribution where θ_{ij} is the probability subject i drinks heavily on any individual day during the 90 days observed at visit j.

2.1 Random Intercept Model

The two parameters π_{ij} and θ_{ij} can be modeled using mixed effects regression. For some appropriate link functions $g_1(\cdot)$ and $g_2(\cdot)$

$$g_1(\pi_{ij}) = \boldsymbol{X}_{1ij}\boldsymbol{\beta_1} + \gamma_{1i} \tag{3}$$

$$g_2(\theta_{ij}) = \boldsymbol{X}_{2ij}\boldsymbol{\beta_2} + \gamma_{2i}, \tag{4}$$

where $\boldsymbol{X}_{1ij}$ and $\boldsymbol{X}_{2ij}$ represent fixed effect covariate vectors of lengths K_1 and K_2 for subject i at time j with corresponding unknown coefficient vectors $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$, and each subject i has random intercept γ_{1i} for the zero model and γ_{2i} for the count model. Within-individual correlation over time is modeled through γ_{1i} and γ_{2i}. Let $g_1(\pi_{ij}) = \text{logit}(\pi_{ij}) = \log(\pi_{ij})/(1 - \log(\pi_{ij}))$, the logit link function, as is standard in logistic regression, while $g_2(\theta_{ij})$ is some appropriate link function for the count model. In the SBIRT analysis, we also choose $g_2(\cdot)$ to be the logit function for the binomial

distribution. For other distributions, one may wish to use a different link function, such as the log function for the Poisson distribution.

The random intercepts are assumed to follow a bivariate normal distribution,

$$\gamma_i = \begin{pmatrix} \gamma_{1i} \\ \gamma_{2i} \end{pmatrix} \sim \mathrm{N}_2\left(\mathbf{0}, \mathbf{\Sigma}\right), \tag{5}$$

where the off diagonal element σ_{12} of the 2×2 covariance matrix $\mathbf{\Sigma}$ models the correlation between an individual's zero model and count model random intercepts. Equations (1) - (5) define the zero-inflated model proposed by Min and Agresti (2005). One can set a prior distribution on $\mathbf{\Sigma}$ to take a Bayesian approach to Min and Agresti's model. Neelon et al. (2010) take a Bayesian approach by setting a marginal density for γ_{1i} and a conditional density for $\gamma_{2i}|\gamma_{1i}$ (Spiegelhalter, 1998; Cooper et al., 2007) on the random effects

$$\gamma_{1i} \sim \mathrm{N}(0, \sigma_1^2) \tag{6}$$

$$\gamma_{2i}|\gamma_{1i} \sim \mathrm{N}(\psi\gamma_{1i}, \sigma_2^2), \tag{7}$$

and with some diffuse hyperpriors on σ_1^2, σ_2^2 and ψ.

2.2 Longitudinal Overdispersed Zero-Inflated Binomial Model (LOZIB)

We propose a class of longitudinal overdispersed zero-inflated binomial models (LOZIB) which extend the model of the previous section by introducing observation-level random effects into the count model. Thus, equation (4) become

$$g_2(\theta_{ij}) = \boldsymbol{X}_{2ij}\boldsymbol{\beta_2} + \psi\gamma_{1i} + \gamma_{2ij} \tag{8}$$

where each subject i has random effect γ_{1i} for the zero model and a vector of random effects $\boldsymbol{\gamma_{2i}} = (\gamma_{2i1}, \gamma_{2i2}, \ldots, \gamma_{2iJ})'$ for the count model while equation (3) for the probability of a zero remains the same. We model correlation between the count and zero models through the additive term $\psi\gamma_{1i}$ and center the count model random effects $\boldsymbol{\gamma}_{2i}$ around zero

$$\boldsymbol{\gamma}_2 \sim \mathrm{N}_J(\mathbf{0}, \boldsymbol{\Sigma}_2), \tag{9}$$

where covariance matrix $\boldsymbol{\Sigma}_2$ models within-individual variation and correlation over time and $\mathbf{0}$ denotes a vector of zeros of length J. This class of models has numerous benefits over the random intercept including additional flexiblity for temporal correlation, and heterogeneity. We do not include OLRE in the zero model as this would be overparamaterized given that there is relatively little information for modeling an unobserved binary variable.

2.3 Covariance Models

The LOZIB models provide a convenient framework for modeling within-individual correlation over time through specification of $\boldsymbol{\Sigma}_2$, the count model random effects covariance matrix. Let

$$\boldsymbol{\Sigma}_2 = \mathrm{diag}(\boldsymbol{\sigma}_2)\Omega\mathrm{diag}(\boldsymbol{\sigma}_2), \tag{10}$$

where $\mathrm{diag}(\boldsymbol{\sigma}_2)$ is a diagonal matrix with diagonal entries given by the vector of standard deviations $\boldsymbol{\sigma}_2 = (\sigma_{21}, \ldots, \sigma_{2J})'$, and Ω is a $J \times J$ matrix of correlations. This is a heteroskedastic model, as the standard deviation is allowed to vary over time. One can also consider a homoskedastic, or constant variance, model by assuming the standard deviation to be constant over time such that $\sigma_{2j} \equiv \sigma_2$ for all $j = 1, \ldots, J$. In the homoskedastic case, the scalar σ_2 replaces the $\mathrm{diag}(\boldsymbol{\sigma}_2)$ term in equation (10). Within individual correlation over time is modeled through Ω, which can be parameterized using an appropriate correlation model (Liechty et al., 2004; Weiss, 2005). We present three possible correlation models, each of which may be paired with constant or non-

constant variance over time.

Independence: The independence (IND) model is the simplest of the covariance models presented in this paper. It assumes count model random effects within an individual are independent over time. The correlation matrix $\boldsymbol{\Omega}_{IND} = \boldsymbol{I}_J$ where $\boldsymbol{I}_J$ is the $J \times J$ identity matrix. The IND model assumes no correlation between an individual's random effects across time.

Compound Symmetry: The compound symmetry (CS) model is a one parameter model which assumes equal correlation, ρ, between any two observations from the same individual. Thus, $\text{Corr}(\gamma_{2ij}, \gamma_{2ik}) = \rho$ for all $j \neq k$. While this is an improvement over the IND model, in that it allows for correlation between observations mad on the same individual, one would typically expect observations made closer in time to have higher correlation, which the CS model does not allow for.

Autoregressive: The autoregressive correlation model (AR) is a one parameter model which assumes a constant correlation between any two *adjacent* time points. For the SBIRT data, with $J = 4$ visits, the correlation matrix of $\boldsymbol{\gamma}_{2i}$ is

$$\text{Corr}(\boldsymbol{\gamma_{2i}}) = \boldsymbol{\Omega}_{\text{AR}}(\rho) = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}.$$

As observations get more distant in time, the correlation between the corresponding random effects decays, which one would expect in most longitudinal data. An appeal of using the autoregressive correlation model is that no matter how large J is, only one parameter ρ needs to be estimated.

Antedependent: The antedependence model generalizes the AR model, with $J - 1$ distinct correlations between consecutive observations. It is a good candidate covariance

model for the SBIRT data where the spacing between consecutive time points differs over the course of the study. For the SBIRT setting with $J = 4$ visits,

$$\boldsymbol{\Omega}_{\text{AD}}(\boldsymbol{\rho}) = \begin{pmatrix} 1 & \rho_1 & \rho_1\rho_2 & \rho_1\rho_2\rho_3 \\ \rho_1 & 1 & \rho_2 & \rho_2\rho_3 \\ \rho_1\rho_2 & \rho_2 & 1 & \rho_3 \\ \rho_1\rho_2\rho_3 & \rho_2\rho_3 & \rho_3 & 1 \end{pmatrix}$$

where the lag 1 correlations are given by the elements of $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_3)'$.

Unstructured: The unstructured correlation model (UN) is the most flexible correlation model and makes no assumptions about the correlations between random effects, however the number of correlation parameters to estimate grows as J^2, which can be costly in settings with many repeated measures. Possible priors for the UN correlation matrix include the LKJ prior (Lewandowski et al., 2009) and Wishart-based correlation distribution (Zhang et al., 2006).

2.4 Markove Chain Monte Carlo

The models as parameterized in the previous sections suffered from slow mixing. Here we present alternative parameterizations for each of the models for use in Bayesian sampling, which greatly improved posterior sampling performance.This consisted of sampling standardized random effects from standard normal distributions and then using these to construct random effects with the desired distributions.

For the RI model subject-level random effects, we model Z_{1i}, $Z_{2i} \sim \text{N}(0, 1)$ distributions and then define appropriately scaled random effects $\gamma_{1i} = \sigma_1 Z_{1i}$ and $\gamma_{2i} = \sigma_2 Z_{2i}$.

Similarly, for the LOZIB model observation-level random effects, model $Z_{1i}, Z_{2ij} \sim$N$(0, 1)$. We take two different approaches to parameterizing the LOZIB models, one for the AD, AR, homoskedastic AD (ADcv) and homoskedastic AR models (ARcv), and a slightly different approach for the UN models.

The ADcv, AR and ARcv models are special cases of the more general heteroskedastic AD model. For the AD model, the observation-level random effects are parameterized

$$\gamma_{2i1} = \sigma_{21} Z_{2i1} \tag{11}$$

and for $j > 1$,

$$\gamma_{2ij} = a_{j-1}\gamma_{2ij-1} + s_{2j} Z_{2ij} \tag{12}$$

where $a_{j-1} = \rho_{j-1}\frac{\sigma_{2j}}{\sigma_{2j-1}}$ and $s_{2j} = \sqrt{(1-\rho_{j-1}^2)}\sigma_{2j}$. When $\rho_{j-1} \equiv \rho$ for all $j > 1$, we get the AR model. For both the AR and AD models, letting $\sigma_{2j} \equiv \sigma_2$ for all j, we get the homooskedastic or constant variance models (ADcv and ARcv).

Using the Cholesky factorization $\boldsymbol{\Omega}_{\text{UN}} = LL'$ with lower triangular matrix L, let

$$\boldsymbol{\gamma}_{2i} = \text{diag}(\boldsymbol{\sigma}_2) L \, \boldsymbol{Z}_i \tag{13}$$

where $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iJ})'$, and $\text{diag}(\boldsymbol{\sigma}_2)$ is a diagonal matrix with jth diagonal entry σ_{2j}.

We fit all models in Stan (Gelman et al., 2015) with the `cmdstanr` package (Gabry and Češnovar, 2021) for R (R Core Team, 2023).

2.5 Difference of Differences

Fitting a zero-inflated model results in two sets of parameters, one for the zero model, and one for the count model. Researchers are often primarily interested the overall effect of a covariate on the overall outcome which requires combining the two parts of the model. The Bayesian approach allows straightforward inference on complex functions of parameters. In the SBIRT analysis, can calculate posterior distributions of mean days of heavy drinking for each treatment group at each time point, as well as differences

in treatment effect between groups, combining both model parts rather than making inference on the zero and count models separately.

Our primary goal in the SBIRT analysis is to assess the effectiveness of the SBIRT intervention compared to the standard of care in reducing frequency of heavy drinking. Let μ_{jc} be the mean days of heavy drinking at time j for treatment group $c = 1, 2$ for the control and treatment group respectively. Then the difference of differences (DoD) at visit j is

$$\text{DoD}(\boldsymbol{\mu}_j) = (\mu_{12} - \mu_{02}) - (\mu_{j1} - \mu_{11}), \tag{14}$$

and can be understood as the difference in treatment effect between treatment groups at time j, where treatment effect is measured as the reduction in frequency of heavy drinking.

We quantify this with the difference of differences ($\text{DoD}(\boldsymbol{\mu}_j)$), the difference in change from baseline at visit j in expected number of days of heavy drinking between the SBIRT and control groups. We can calculate the DoDs separately for each of the two parts of the zero-inflated model in a similar fashion.

Calculating the posterior distributions of the various DoDs requires integrating out random effects, which we do using Monte Carlo methods. Let $\zeta^{(s)} = (\boldsymbol{\beta}_1^{(s)}, \boldsymbol{\beta}_2^{(s)}, \sigma_1^{(s)}, \boldsymbol{\Sigma}_{\mathbf{2}}^{(s)})$ be posterior sample $s = 1, \ldots, S$, where S is the total number of posterior draws. For each $\zeta^{(s)}$, we drew 1,000 samples from random effects distributions (6) and (9) and used these samples to integrate out the random effects to compute estimates of the marginal means of the zero and count models at each visit for each treatment group. From the zero and count model means, one can also compute the overall mean of the zero-inflated model, as well as the DoDs described in section 2.5.

2.6 Prior Distributions

Priors were selected to be generally vague to let data dictate the inference. Elements of fixed effect vectors $\beta_{lk} \sim \text{N}(0, 10^2)$ for $l = 1, 2$, and $k = 1, \ldots, K_l$. The standard

deviation parameters σ_1 and σ_2 for homoskedastic models (RI, AR, AD), and σ_{2j} for the heteroskedastic models (AR, AD, UN) were given half-normal $N^+(0,.5^2)$ priors where N^+(a,b) is the Normal distribution with mean a and variance b restricted to the positive domain.

for the zero model random intercepts and count model random effects were given half-normal $N^+(0,.5^2)$ priors where N^+(a,b) is the Normal distribution with mean a and variance b restricted to the positive domain. The regression coefficient ψ modeling the association between the zero and count models was given a $N^+(0,1^2)$ prior, as we expect a positive correlation between an individual's zero model and count model estimates. Correlation parameters ρ for the AR models, and elements of $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$ in the AD models – were assigned flat U(0,1) priors, again reflecting the expected positive within-individual correlation. We set an LKJ correlation prior (Lewandowski et al., 2009) for the correlations of the UN covariance matrix , with shape parameter $\eta = 1$ corresponding to a uniform distribution over all correlation matrices of order J.

3 SBIRT Data Analysis

1. Model Fit

   (a) Make new plot of means and DoDs. Try to do the same with all substances, but maybe remove them.

   (b) Compare the LOO-IC of different models

   (c) INDcv is the best - this model accounts for overdispersion, but has no correlation between random effects at each time point.

   (d) Compare fit between RI and INDcv - how well do they model the data

(e) Next question is, what is the difference between them, why is INDcv doing so much better

(f) Also, why INDcv and not a correlated model. What are correlation posteriors? From AR, AD, models

(g) How does dispersion compare between INDcv and RI?

- What is the variance estimate from each of the models?
- What do they each estimate as the in-sample mean of the data? (Mean use and CrI's, do they cover the observed data?)

- Out of sample means and CrIs - do these cover the observed data?
- mean and variance from data, mean and variance of predictions from RI and INDcv models

We demonstrate the models of the previous section on the SBIRT data set, in which the main outcome of interest is number of days of heavy drinking over the previous 90 days. We fit LOZIB models with 10 different covariance models: independence (IND), independence constant variance (INDcv), compound symmetry (CS), compound symmetry constant variance (CScv), autogregressive (AR), autoregressive constant variance (ARcv), antedependent (AD), antedependent constant variance (ADcv), unstructured (UN), and unstructured constant variance (UNcv). We also include the RI model presented in section 2.1 (2RI) as well as a one random intercept model, which shares a single random intercept differing by a scaling factor between the zero and count models.

Fixed effect regression coefficients for the zero and count models $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, are vectors of length 7, corresponding to indicators for the 4 visits and 3 treatment group differences at the 3, 6, and 12 month follow-ups. We assumed no difference between treatment groups at baseline as groups were randomized before intervention.

Each model was run with 4 Markov chains, each consisting of 20,000 samples after 5,000 burn-in iterations. All models had satisfactory convergence based on trace plots, autocorrelations and $\hat{R}$ statistics Gelman and Rubin (1992).

Table 1: Pareto smoothed approximate leave-one-out cross-validation information criterion (LOO-IC) estimates for LOZIB models fit to the SBIRT data. LOO-IC are given for each of the 12 covariance models for each substance outcome. The models are displayed in ascending order by LOO estimates. All ten observation level random effect LOZIB models fit the SBIRT data considerably better than the random intercept models. Differences between the observation level random effect models were small.

Model	Heavy Drinking	Any Alcohol	Stimulants	Marijuana
INDcv	7,526 (1)	12,397 (10)	4,348 (4)	7,699 (7)
UN	7,529 (2)	12,267 (1)	4,321 (1)	7,659 (6)
AD	7,530 (3)	12,317 (3)	4,418 (8)	7,635 (4)
UNcv	7,549 (4)	12,296 (2)	4,337 (2)	7,591 (3)
IND	7,559 (5)	12,361 (8)	4,350 (5)	7,740 (9)
ADcv	7,577 (6)	12,325 (4)	4,361 (6)	7,531 (1)
ARcv	7,580 (7)	12,327 (5)	4,347 (3)	7,538 (2)
AR	7,584 (8)	12,346 (6)	4,416 (7)	7,649 (5)
CScv	7,591 (9)	12,355 (7)	4,419 (9)	7,729 (8)
CS	7,616 (10)	12,361 (9)	4,433 (10)	7,764 (10)
2RI	16,874 (11)	33,666 (11)	12,182 (12)	38,003 (11)
1RI	16,907 (12)	33,919 (12)	12,172 (11)	38,213 (12)

3.1 Model Fit

We compare model fit between the 12 different models using Bayesian approximate leave-one-out cross-validation (LOO) (Vehtari et al., 2017). Table 1 gives LOO information criterion (LOO-IC) estimates and standard errors coming from the pareto-smoothed approximation used by the loo package in R (Vehtari et al., 2023). Values in the table are -2 times the LOO as defined by Vehtari et al. (2017) to be on the deviance scale, thus smaller values indicate improved fit. The 10 models with observation level random effects fit the SBIRT data substantially better than the random intercept models, however the difference between OLRE models were small. Still, we select the unstructured (UN) model as the best fitting model.

3.2 Mean estimates:

The top plot in Figure 2 gives the overall means of days of heavy drinking out of the previous 90 coming from the entire zero-inflated models. In general, the ARcv

model estimates lower levels of heavy drinking than the RI model, with smaller credible intervals. As in the DoD estimates, the differences between the LOZIB models and the RI model are mainly in the count model. While the LOZIB models do not find much difference between treatment groups, people in both treatment groups reduced rates of heavy drinking – particularly through a decrease in the zero model estimate, that is, a decrease in probability of an observation coming from the count model.

A comparison between the ARcv and RI models is given in Table 2. Again, the ARcv model estimates lower mean days of heavy drinking than the RI model, as well as smaller (less negative) DoD estimates. Given the better fit of the ARcv model, it is likely the RI is overstating the amount of heavy drinking as well as the effectiveness of intervention. **BWR:[REW: Compare to raw data, relating to last 2 sentences]** Table 2 also provides the estimated posterior probability that the difference of differences is less than zero (p(DoD$<$ 0)), which would indicate superiority of SBIRT over the standard of care. There are large differences between the the ARcv and RI models in this estimated probability of intervention effectiveness at the 6 and 12 month visits, with the RI model again overstating the treatment effect compared to the ARcv model.

The bottom portion of Table 2 gives posterior estimates and CrIs for variance and correlation parameters. The ARcv model has wider posterior distributions for the zero and count model standard deviation estimates, as well as for the ψ parameter, which models the within-individual association between the zero and count model estimates. In addition, the within-individual random effects correlation is estimated to be rather high [.732, $95\%CrI$ = (.647, .797)].

3.3 Covariance Inference

Difference of Differences Figure 3 gives difference of differences (DoD) for all seven ZIB models, as well as the DoDs estimated from both the zero and the count parts of the model. The ARcv LOZIB and the RI models are bolded in orange and blue respectively.

The RI model overestimates treatment effects compared to the better fitting LOZIB

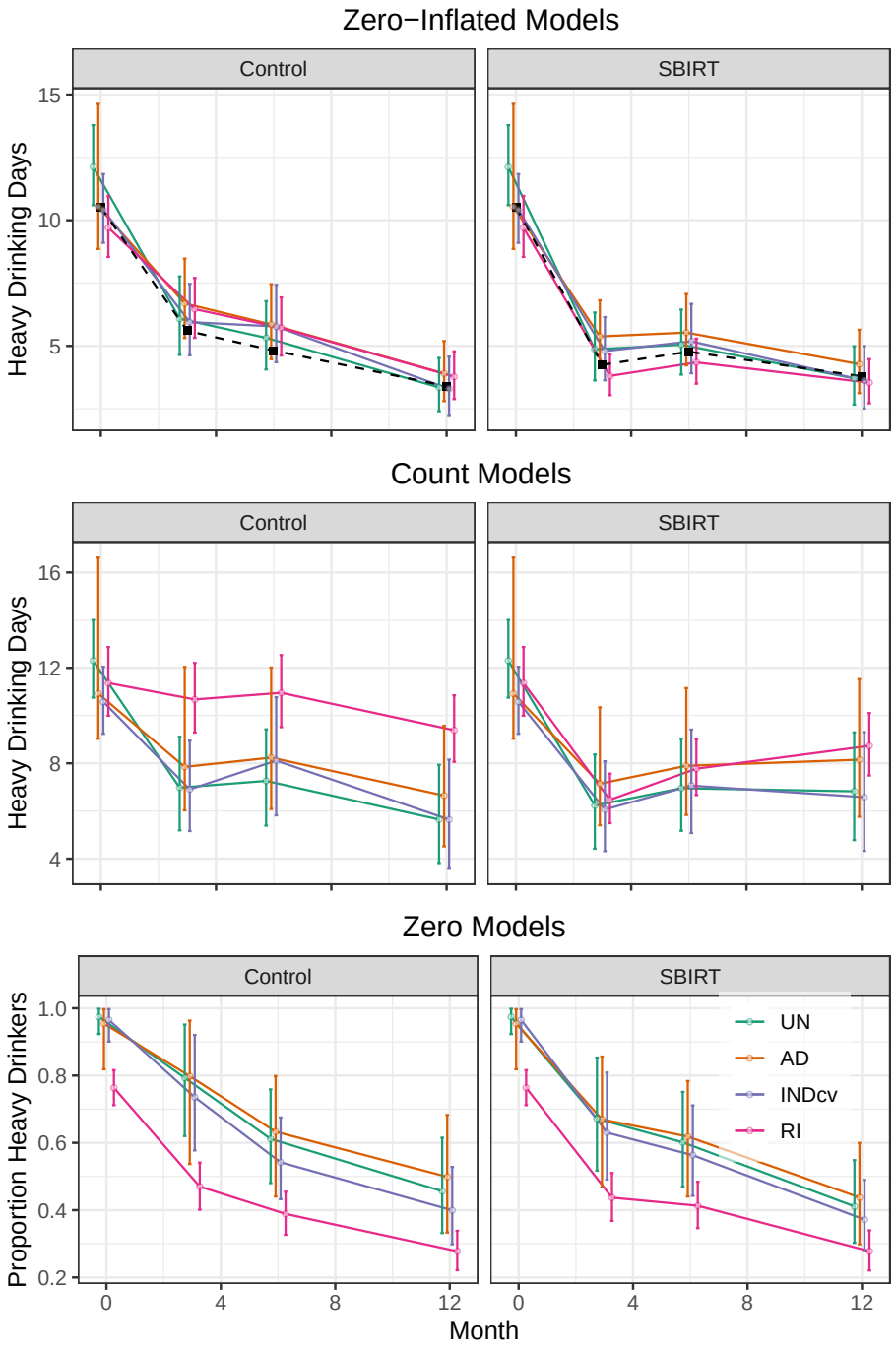

Figure 2: Posterior means and 95% CrIs for the three best fitting LOZIB models (UN, AD and INDcv) as well as the RI model for comparison. Plots on the left show the posterior means and CrIs for the control group and plots on the right show the same for the SBIRT group. The top row of plots show the posterior summaries for the full zero-inflated models, while the middle and bottom rows display the posterior summaries of the count and zero models. The observed sample means for each group and time point are shown in black with a dashed line in top plots.

	3 Month	6 Month	12 Month
AD Model			
Baseline	.11 (.00, .57)		
3 Month		.57 (.33, .78)	
6 Month			.62 (.36, .82)
UN Model			
Baseline	.11 (.00, .57)		
3 Month		.57 (.33, .78)	
6 Month			.62 (.36, .82)

models (Figure 3). At 3 and 6 month follow-up, the RI model 95% Bayesian credible intervals for the DoDs do not include zero, a finding not shared by the six better fitting LOZIB models. In addition, the 3 month posterior means from each of the LOZIB models fall outside of the RI model credible intervals, further illustrating the disagreement between the LOZIB and RI models. Generally, the RI model estimates substantially stronger treatment effects than the LOZIB models .

There is relative consistency between the six LOZIB models. The estimates are not particularly sensitive to choices in covariance models, and thus modeling of overdispersion, rather than temporal correlation, may be the main driver of the improved performance of the LOZIB models. **BWR:[Weiss comment: SHow us e's for AR and AD models first, before drawing this conclusion. I think we will likely find that there is some temporal correlation, it just doesn't have a large influence on LOO-IC.]**

The lower left plot in Figure 3 gives the count model estimates for DoDs in days of heavy drinking out of the previous 90 among heavy drinkers. Again there is consistency among the six LOZIB models, while the RI model estimates greater reductions in drinking with much more narrow credible intervals.

The differences in treatment effects are minimal in the zero models (Figure 3),

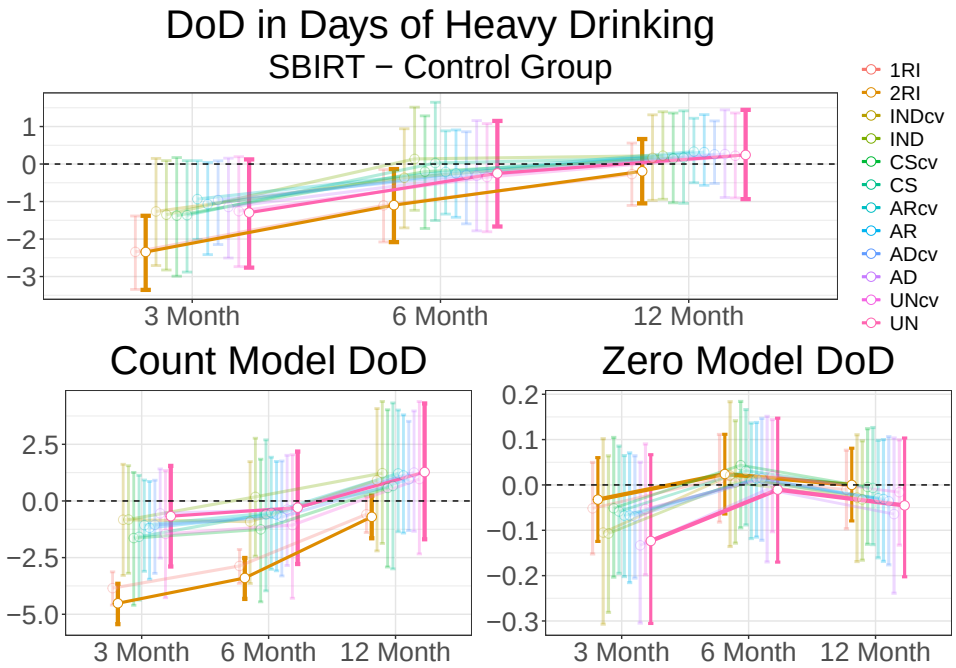

Figure 3: **Top:** Difference of differences (DoD) in mean days of heavy drinking measured as difference in change from baseline at each of the three follow up visits. Negative values indicate SBIRT group exhibited greater reductions in heavy drinking than control group. **Bottom left:** DoD in mean days of heavy drinking coming only from the count model, i.e., only among subjects who were deemed to be heavy drinkers. **Bottom right:** DoD in probability of being a heavy drinker coming from the zero model. Estimates and 95% Bayesian credible intervals (CrI) are given for all seven ZIB models. The random intercept model (RI), given in blue, and the autoregressive constant variance model (ARcv), given in orange, are bolded for emphasis.

which is not surprising as the zero models are parameterized identically across all seven ZIB models, and differences are only introduced through the connection to the count model. Zero model DoD estimates can be understood as the difference in the reduction in proportion of heavy drinkers between the two treatment groups. For example, the ARcv model estimates that at the 3 month follow-up, the reduction in proportion of heavy drinkers was .068 [$95\% CrI =$ (-.065, .202)] greater in the SBIRT group than in the control group.

BWR:[Weiss: Factors of 50 % to 100% different between RI and ARcv. How does raw data compare?]

BWR:[Weiss Notes: Rather large range among models in the zero model estimates.]

		RI		INDcv	
		Control	SBIRT	Control	SBIRT
Mean	Baseline	11.75 (10.42, 13.18)		12.30 (10.87, 13.82)	
	3 Month	8.01 (6.67, 9.45)	5 (4.04, 6.07)	5.57 (4.38, 6.94)	5.57 (4.34, 6.98)
	6 Month	7.14 (5.83, 8.54)	5.61 (4.57, 6.75)	6.21 (4.9, 7.70)	4.84 (3.73, 6.11)
	12 Month	4.92 (3.82, 6.12)	4.64 (3.62, 5.77)	3.65 (2.66, 4.85)	4.51 (3.37, 5.86)
DoD		DoD	p(DoD < 0)	DoD	p(DoD < 0)
	3 Month	-2.35 (-3.36, -1.38)	1	-0.934 (-2.01, .963)	.963
	6 Month	-1.10 (-2.08, -.134)	.987	-.211 (-1.33 , .646)	.646
	12 Month	-.193 (-1.05, .666)	.674	.331 (-.500 , .217)	.217
Zero RE SD	σ_1	1.7 (1.39, 2.02)		1.8 (1.32, 2.31)	
Count RE SD	σ_2	1.98 (1.78, 2.2)		1.82 (1.68, 1.98)	
	ψ	.744 (.308, 1.15)		.826 (.249, 1.41)	
	ρ	-		.732 (.647, .797)	

Table 2: Posterior means and 95% Bayesian CrIs from RI and ARcv models. The top portion of the table gives mean days of heavy drinking out of the previous 90 days. The middle portion of the table gives DoD means and 95% Bayesian CrIs, as well as the posterior probabilities that the DoD was less than zero - indicating a benefit of SBIRT over the control treatment. The bottom portion of the table gives posterior means and CrIs for the zero and count model random effects standard deviations (σ_1 and σ_2), the ψ parameter, which models association between the count and zero models, and ρ, which is the within-individual across-time count model random effects correlation in the ARcv model.

4 Discussion

We developed a class of longitudinal overdispersed zero-inflated binomial (LOZIB) models which extend previous models through the inclusion of observation-level random effects (OLRE) in the count model regression. These OLRE provide increased flexibility to account for both overdispersion and correlation over time. We demonstrated LOZIB models with six different covariance models, which we applied to the SBIRT data to model counts of days of heavy drinking out of the previous 90 days. The LOZIB models showed improved ability to fit the data over a RI ZIB model, however differences between the 6 LOZIB models were small. In addition to improving fit, the LOZIB models yielded different inference than the RI model. These differences could be impactful, for example, if policy makers were to use 95% Bayesian credible intervals to gauge significance of a treatment effect, the SBIRT intervention would be found to be effective at 3 and 6 months under the RI model, but not under the better fitting LOZIB models.

We hypothesize that the superiority of the LOZIB models in the SBIRT analysis

stem mainly from the LOZIB models ability to handle excess variation. Failure to adequately account for overdispersion in count models can lead to bias both mean and variance estimates. In the SBIRT analysis, the general use of a LOZIB model is an important improvement, though the choice of which covariance model to use within the LOZIB model is not particularly influential. In larger data sets with more time points, differences in covariance models may be more noticeable. The random effects covariance matrix grows quickly with the number of time points, as does the difference in number of parameters to estimate between, for example, the ARcv (2 parameters), and the UN models ($\frac{n(n-1)}{2}$ parameters).

We also attempted to fit the SBIRT data with frequentist RI models using the glmmTMB R package (Brooks et al., 2017), but the models failed to converge on reliable model estimates. To our knowledge, it is not possible to fit LOZIB type models using the package. In contrast, under a Bayesian framework, it is straightforward to fit both the RI ZIB and LOZIB models using Stan software (Gelman et al., 2015), and all models converged yielding valid inference. In addition, the SBIRT analysis illustrates the usefulness of Bayesian methods to construct valid credible intervals for functions of model parameters. Using our methods, it was straightforward to produce estimates and intervals for mean days of heavy drinking as well as difference of differences and probability of a treatment effect (Table 2).

There are a number of possible extensions of the LOZIB models presented in this paper. The LOZIB models we developed were designed for nominal time, such as in the SBIRT study where observations were measured at pre-specified times. In other settings one may wish to adapt LOZIB models to handle continuously measured time. This change would be reflected both in the random effects specification as well as the regression parameterization where one may estimate a continuous time effect. Additionally, the LOZIB models may be adapted to model multivariate outcomes where one may specify, for example, a vector autoregressive model to account for correlation both over time and between different outcomes.

Overdispersed longitudinal zero-inflated data can occur in many settings beyond substance use, and for these situations we recommend LOZIB models to researchers due to their ability to account for excess variation and longitudinal correlation. Failure to properly account for these two aspects of data can give rise to misleading inference as we saw in the case of the SBIRT study. Overall, we present LOZIB models as an important improvement over previous methods for overdsipersed longitudinal zero-inflated data.

5 References

References

Albert, J. M., Wang, W., and Nelson, S. (2014). Estimating overall exposure effects for zero-inflated regression models with application to dental caries. *Statistical methods in medical research*, 23(3):257–278.

Barata, I. A., Shandro, J. R., Montgomery, M., Polansky, R., Sachs, C. J., Duber, H. C., Weaver, L. M., Heins, A., Owen, H. S., Josephson, E. B., et al. (2017). Effectiveness of sbirt for alcohol use disorders in the emergency department: a systematic review. *Western journal of emergency medicine*, 18(6):1143.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400.

Buu, A., Li, R., Tan, X., and Zucker, R. A. (2012). Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. *Statistics in medicine*, 31(29):4074–4086.

Cohen Jr, A. C. (1966). A note on certain discrete mixed distributions. *Biometrics*, pages 566–572.

Cooper, N. J., Lambert, P. C., Abrams, K. R., and Sutton, A. J. (2007). Predicting costs over time using bayesian markov chain monte carlo methods: an application to early inflammatory polyarthritis. *Health economics*, 16(1):37–56.

Gabry, J. and Češnovar, R. (2021). cmdstanr: R interface to'cmdstan'. *URL: https://mc-stan. org/cmdstanr, https://discourse. mc-stan. org*.

Gardner, W., Mulvey, E. P., and Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed poisson, and negative binomial models. *Psychological bulletin*, 118(3):392.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gelman, A., Lee, D., and Guo, J. (2015). Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Ghosh, S. K., Mukhopadhyay, P., and Lu, J.-C. J. (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical planning and Inference*, 136(4):1360–1375.

Glass, J. E., Hamilton, A. M., Powell, B. J., Perron, B. E., Brown, R. T., and Ilgen, M. A. (2015). Specialty substance use disorder services following brief alcohol intervention: a meta-analysis of randomized controlled trials. *Addiction*, 110(9):1404–1415.

Hall, D. B. (2000). Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics*, 56(4):1030–1039.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, 2:e616.

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, 36(5):531–547.

Hinde, J. and Demétrio, C. G. (1998). Overdispersion: models and estimation. *Computational statistics & data analysis*, 27(2):151–170.

Hu, T., Gallins, P., and Zhou, Y.-H. (2018). A zero-inflated beta-binomial model for microbiome data analysis. *Stat*, 7(1):e185.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate discrete distributions*, volume 444. John Wiley & Sons.

Karno, M. P., Rawson, R., Rogers, B., Spear, S., Grella, C., Mooney, L. J., Saitz, R., Kagan, B., and Glasner, S. (2021). Effect of screening, brief intervention and referral to treatment for unhealthy alcohol and other drug use in mental health treatment settings: a randomized controlled trial. *Addiction*, 116(1):159–169.

Kwok, O.-m., West, S. G., and Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A monte carlo study. *Multivariate Behavioral Research*, 42(3):557–592.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lange, N. and Laird, N. M. (1989). The effect of covariance structure on variance estimation in balanced growth-curve models with random parameters. *Journal of the American Statistical Association*, 84(405):241–247.

Lee, A. H., Wang, K., Scott, J. A., Yau, K. K., and McLachlan, G. J. (2006). Multilevel zero-inflated poisson regression modelling of correlated count data with excess zeros. *Statistical methods in medical research*, 15(1):47–61.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001.

Liechty, J. C., Liechty, M. W., and Müller, P. (2004). Bayesian correlation estimation. *Biometrika*, 91(1):1–14.

Liu, Y. and Tian, G.-L. (2015). Type i multivariate zero-inflated poisson distribution with applications. *Computational Statistics & Data Analysis*, 83:200–222.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical modelling*, 5(1):1–19.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365.

Neelon, B. H., O'Malley, A. J., and Normand, S.-L. T. (2010). A bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical modelling*, 10(4):421–439.

Preisser, J. S., Das, K., Long, D. L., and Divaris, K. (2016). Marginalized zero-inflated negative binomial regression with application to dental caries. *Statistics in medicine*, 35(10):1722–1735.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ridout, M., Demétrio, C. G., and Hinde, J. (1998). Models for count data with many zeros. In *Proceedings of the XIXth international biometric conference*, volume 19, pages 179–192. International Biometric Society Invited Papers Cape Town, South Africa.

Rodrigues, J. (2003). Bayesian analysis of zero-inflated distributions. *Communications in Statistics-Theory and Methods*, 32(2):281–289.

Saitz, R., Palfai, T. P., Cheng, D. M., Alford, D. P., Bernstein, J. A., Lloyd-Travaglini, C. A., Meli, S. M., Chaisson, C. E., and Samet, J. H. (2014). Screening and brief

intervention for drug use in primary care: the aspire randomized clinical trial. *Jama*, 312(5):502–513.

Spiegelhalter, D. J. (1998). Bayesian graphical modelling: a case-study in monitoring health outcomes. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 47(1):115–133.

Tetrault, J. M., Holt, S. R., Cavallo, D. A., O'Connor, P. G., Gordon, M. A., Corvino, J. K., Nich, C., and Carroll, K. M. (2020). Computerized cognitive behavioral therapy for substance use disorders in a specialized primary care practice: A randomized feasibility trial to address the rt component of sbirt. *Journal of Addiction Medicine*, 14(6):e303–e309.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2023). loo: Efficient leave-one-out cross-validation and waic for bayesian models. R package version 2.6.0.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413–1432.

Weiss, R. E. (2005). *Modeling longitudinal data*, volume 1. Springer.

Wen, C.-C., Baker, N., Paul, R., Hill, E., Hunt, K., Li, H., Gray, K., and Neelon, B. (2024). A bayesian zero-inflated beta-binomial model for longitudinal data with group-specific changepoints. *Statistics in Medicine*, 43(1):125–140.

Wolfinger, R. D. (1996). Heterogeneous variance: covariance structures for repeated measures. *Journal of agricultural, biological, and environmental statistics*, pages 205–230.

Wu, Q., Tian, G.-L., Li, T., Tang, M.-L., and Zhang, C. (2023). The multivariate component zero-inflated poisson model for correlated count data analysis. *Australian & New Zealand Journal of Statistics*, 65(3):234–261.

Yau, K. K. and Lee, A. H. (2001). Zero-inflated poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in medicine*, 20(19):2907–2920.

Zhang, X., Boscardin, W. J., and Belin, T. R. (2006). Sampling correlation matrices in bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics*, 15(4):880–896.