Title
Bayesian Covariance Modeling for Longitudinal Zero-Inflated Count Data

Author
Rogers, Benjamin

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Bayesian Covariance Modeling for Longitudinal Zero-Inflated Count Data

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Benjamin Rogers

2022

ABSTRACT OF THE DISSERTATION

Bayesian Covariance Modeling for Longitudinal
Zero-Inflated Count Data

by

Benjamin Rogers
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2022
Professor Robert E. Weiss, Chair

We develop models for longitudinal count data with a large number of zeros, a feature known as zero-inflation. Familiar distributions for modeling count data (Poisson, binomial, negative binomial) often do not account for the observed frequency of zeros. Further, in longitudinal data, the same subjects are repeatedly measured over time inducing correlation between sets of measurements on the same individual. Modeling of longitudinal data that does not account for this correlation can give rise to misleading inferences. This dissertation develops three classes of models for longitudinal count data: (i) a Bayesian longitudinal hurdle model for data with prespecified measurement times, (ii) a Bayesian longitudinal hurdle model for data with varying measurement times, and (iii) a multivariate longitudinal zero-inflated Poisson model. Approach (i) is an analysis of the number of days of heaving drinking in a study of screening, brief intervention, and referral to treatment (SBIRT), and approaches (ii) and (iii) are motivated by analyses of the Linking Inmates to Care (LINK LA) study. Building on two-part models that predict non-zero versus zero outcomes while incorporating assumptions about the distribution of non-zero outcomes, the newly developed methods use

mixed-effect modeling strategies to account for irregular measurement times and correlated patterns in count data beyond those reflected in random intercept models. The superiority of the proposed methods over random intercept models is established using goodness-of-fit metrics that consider the number of model parameters, and the appeal of modeling multiple count outcomes simultaneously is reflected in Bayesian credible intervals that point to non-zero correlations among the respective count outcomes.

We build upon previous longitudinal zero-inflated and hurdle models by introducing time varying random effects in the count models with random effects distributed a priori as multivariate normal with a parameterized covariance matrix. We propose several covariance models, which improve fit over random intercept models in both the SBIRT and LINK LA data. We introduce latent time varying main and random effects to allow count rates and zero probabilities to change with time since intervention and include exposure offsets to account for varying times over which counts are recorded. Finally, for use with multivariate data, we propose a multivariate longitudinal zero-inflated Poisson model for observations with varying exposure, which we use to simultaneously model three different kinds of doctor visits recorded in the LINK LA study.

The dissertation of Benjamin Rogers is approved.

Nina T. Harawa

Catherine Ann Sugar

Thomas R. Belin

Robert E. Weiss, Committee Chair

University of California, Los Angeles

2022

To Jenny and Josephine,
for your love, encouragement,
smiles and laughs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I must first express my deepest gratitude to my advisor, Dr Robert E. Weiss, for his patience, kindness, and thoroughness in advising me in this work. So much of my growth throughout the program has been due to his guidance and his passion for both statistics and life outside of statistics. I would also like to thank Dr Thomas Belin and Dr Catherine Sugar for their roles as teachers and mentors, and for serving on my dissertation committee. They have both been steady and approachable resources during my studies. Also thank you to Dr Nina Harawa for volunteering for my dissertation committee and for her insightful comments and suggestions. I would also like to thank Dr Mitchell Karno and Dr Suzette Glasner-Edwards for bringing me on to the SBIRT study, which inspired much of this dissertation, and for allowing me to use the SBIRT data. Also thank you to Dr Mary-Lynn Brecht for her mentorship and support during my time at School of Nursing.

Thank you to my parents, Maryanne and David, for their love and encouragement. Without them, none of this would have been possible. Thank you to my brothers, Dan and Tim, for being my closest friends and for helping to shape me into the person I am today.

This research has in part been supported by the AIDS Training Grant T32/AI-07370 and by NIH/NIDA grant R01 DA030781.

VITA

2010 B.A. (Mathematics), Colorado College.

2012–2013 Teaching Assistant, Department of Biostatistics, UCLA.

2016–2020 Graduate Student Researcher, School of Nursing, UCLA.

2019–2021 Teaching Assistant, Department of Biostatistics, UCLA.

PUBLICATIONS

Karno, M. P., Rawson, R., Rogers, B., Spear, S., Grella, C., Mooney, L. J., Saits, R., Kagan, B., & 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.

Hamilton, N., Marques-Garban, D., Rogers, B., Austin, D., Foos, K. Tong, A., Adams, D. Vadgama, J., Brecht, M.-L., & Pietras, R. (2019) Dual therapy with insulin-like growth Factor-I receptor/insulin receptor (IGF1R/IR) and androgen receptor (AR) antagonists inhibits triple-negative breast cancer cell migration in vitro. *SPG BIoMed*, **1** (2).

Wallace, M. S., North, J., Grigsby, E. J., Kapural, L., Sanapati, M. R., Smith, S. G., Willoughby, C., McIntyre, P., Cohen, S., Rosenthal, R., Ahmed, S., Vallejo, R., Ahadian, F., Yearwood, T., Burton, A., Frankoski, E., Shetake, J., Lin, S., Hershey, B., Rogers, B., &

Mekel-Bobrov, N. (2018). An integrated quantitative index for measuring chronic multisite pain: the multiple areas of pain (MAP) study. *Pain Medicine*, **19**(7), 1425-1435.

Li, L., Luo, S., Rogers, B., Lee, S. J., & Tuan, N. A. (2017). HIV disclosure and unprotected sex among Vietnamese men with a history of drug use. *AIDS and Behavior*, **21**(9), 2634-2640.

CHAPTER 1

Introduction

1.1 Overview

In count data, it is common to observe a disproportionately large number of zeros relative to standard count distributions. This quality of excess zeros is called zero-inflation, and often arises in health and behavioral data. Zero-inflation typically occurs as a result of the data being generated by two separate processes, one which determines if an event may happen at all, and another that determines the size of the observed counts. For example, in a study of linkage to HIV care, we model the number of medical visits among a population of previously incarcerated men and transgender women, however not all subjects in the study have reasonable access to care. This leads to a large number of zero observations, and can be considered a separate process from that which determines the number of care visits among those who do have access.

Zero-inflation presents a modeling challenge as standard count distributions cannot account for the excess zeros. A common approach to zero-inflation is to use a two part model consisting of a zero model and a count model. One such class of models, zero-inflated models, allow zeros to come from either the zero model or the count model. Using the linkage to HIV care example, the zero model models the proportion of subjects who have *access* to care, and the count model models the expected number of medical visits among those who do have access to care (Lambert, 1992; Heilbron, 1994). Subjects who have access to care may or may not take advantage of this access, and thus, could still have a zero obser-

vation. An alternative class of models, hurdle models, only allow zeros to come from the zero model, and use a truncated count distribution restricted to positive values to model the counts (Mullahy, 1986; Ridout, Demétrio, & Hinde, 1998). As such, a hurdle model does not attempt to make inference as to who does or does not have access to care, but rather the proportion of subjects who *use* care. The count model in a hurdle model would then measure the number of medical visits among patients that had at least one visit.

We develop zero-inflated and hurdle models for longitudinal studies, where the same subjects are followed over time and repeatedly measured. We expect there to be correlation between the repeated measurements on the same individual. To accurately model variability in the data and draw inferences, this correlation must be accounted for. Traditional longitudinal models for zero-inflated data have used random intercepts either only in the count model (Hall, 2000), or in both the zero and count models (Dagne, 2004), to model within-individual correlation. These approaches assign a single random effect to each individual which is applied to all measurements on that individual. A further benefit of Hall's model is that the random intercepts in the zero and count models can be assumed to be jointly distributed to introduce dependency between both model parts (Min & Agresti, 2005; Neelon, O'Malley, & Normand, 2010). The random intercepts, however, can be restrictive in modeling within-individual correlation and variability. Generally, observations made closer together in time are likely more highly correlated than observations made farther apart in time. This property cannot be accounted for in a random intercept model. We propose zero-inflated and hurdle models which allow for an individual's count model random effect to vary over time, through which we model and compare several possible temporal covariance structures. We find these covariance models greatly improve model fit over random intercept models.

We model data from two separate studies. The first of these comes from a randomized controlled trial designed to test an intervention to reduce rates of heavy drinking in a population of substance users seeking mental health treatment. Number of heavy days of drinking

out of the past 90 was recorded at baseline and at 3 follow-up visits. We model the number of days of heavy drinking to assess the effect of an intervention consisting of screening, brief intervention and referral to treatment (SBIRT), which we compare to a health education control condition. Many of the recruited subjects do not regularly engage in heavy drinking, as it wasn't required for inclusion in the study, resulting in many observations where participants reported zero days. Normally we may model number of days of drinking using a binomial distribution as it is a count with an upper limit, however due to the large number of zeros, the binomial distribution does not fit the data well. In chapter 2 we develop a class of longitudinal binomial hurdle models with multivariate random effects in the count model. We fit several correlation models for the random effects to model within-individual correlation and compare them to traditional random intercept models.

The second study, the Linking Inmates to Care (LINK LA) study, was a longitudinal randomized controlled trial of an intervention to improve linkage to and engagement in care for HIV positive men and transgender women recently released from LA county jail. We model number of primary care visits, including regular visits relating to HIV infection, to assess the effectiveness of the intervention. People living with HIV (PLHIV) with a history of incarceration often experience difficulty accessing necessary care, and as a result, the data is zero-inflated. Observations for each subject were intended to be collected for the one year prior to incarceration, and then at 3 months, 6 months and 12 months following release from jail, however in practice follow-ups occurred at highly irregular intervals. Responses report on all primary care visits since last follow-up, thus different counts are measured over varying amounts of time, making it difficult to compare number of doctor visits both within and between individuals. In chapter 3, we present a longitudinal hurdle model for the LINK LA data, which uses multivariate random effects and accounts for length of time over which visits are counted, called *exposure*.

The LINK LA study recorded data on multiple types of medical visits at each observation. Ultimately, researchers are interested in how well subjects are linked to care and what

the cost of that care is. Therefore, in addition to primary care visits, we are interested in the association between intervention group and counts of each type of doctor visit, such as emergency room visits and specialty care. We present a zero-inflated Poisson model, which is an extension of the models for primary care in chapter 3. Multivariate random effects are used to model within-individual correlation within outcomes and also between outcomes. Modeling these correlations allows us to more accurately model the data and provides behavioral insights through understanding the relationships between usage of different types of healthcare visits.

1.2 Bayesian Longitudinal Hurdle Models for Days of Heavy Drinking

Substance use disorder is estimated to occur in one in five people with mental illness in the United States (Clark, Power, Fauve, & Lopez, 2008). Rates of heavy drinking are disproportionately high in this vulnerable population (Grant et al., 2004; Flynn & Brown, 2008). Screening, brief intervention and referral to treatment (SBIRT) interventions have been developed to help treat substance abuse, but have had mixed results for reducing rates of heavy drinking in a primary care setting (Saitz, 2010). In chapter 2, we consider data from a recent study of the effectiveness of SBIRT interventions for subjects with mental health and substance use disorders, which we refer to as the SBIRT study (Karno et al., 2021). In chapter 2 we develop models for the number of days during which a subject engaged in heavy drinking out of the past 90 days and compare the effect of the SBIRT intervention to a health education standard of care control.

To be eligible for the study, subjects had to have a diagnosis of a mental health disorder and report use of alcohol, cannabis or stimulants within the past 90 days. We model days of heavy alcohol use, where heavy alcohol use is defined as at least 5 drinks for men or 4 drinks for women. Subjects were randomized to SBIRT or standard of care and followed at 3, 6

and 12 months after baseline and self reported substance use was recorded. Many subjects reported no heavy drinking, resulting in a large number of zero observations. We model this data with a two part hurdle model. The zero model uses a Bernoulli distribution to model yes/no whether or not a subject successfully abstained from heavy drinking during the 90 day period. The count model uses a truncated binomial distribution restricted to positive integers to model how many days a subject drank heavily *given that they had at least one day of heavy drinking*. We can think of the two parts of the model as 1) Did the subject abstain from drinking? And 2) If the subject did engage in heavy drinking, to what extent? Thus we jointly model the number of subjects that engage in heavy alcohol use, and the rate of heavy alcohol use among those that do engage.

The SBIRT study was longitudinal, collecting repeated measures on the same subjects over one year. We expect measurements on the same subject to be correlated with each other. Further, we expect that correlation to vary with the spacing of the observation times. Observations closer in time should be more highly correlated, and observations further in time should have lower correlation. Thus, in the SBIRT study where follow-up visits are not equally spaced in time, we expect correlation to differ between different pairs of measurements. To model the SBIRT data, we propose a class of Bayesian longitudinal hurdle regression models that allow researchers to specify a within-individual correlation model.

Previous hurdle models for longitudinal data have used random intercepts to capture within individual correlation over time. Each subject gets one single random effect that applies to all time points. We find this to be too rigid of an assumption for the SBIRT data and that by allowing one separate random effect per observation, with correlation between random effects for an individual, we can fit the data much better. This change to a multivariate random effect distribution allows researchers to specify any of several within-individual correlation models. For example one may use an autoregressive correlation model, which assumes constant correlation between random effects at adjacent time points, and

smaller correlation between random effects at non-adjacent time points. Researchers are then left to strike the right balance between capturing the complexity of the data and over-fitting depending on which correlation model is chosen and how many parameters must be estimated.

Another advantage of the proposed models is that the multivariate random effects allow the models greater ability to model overdispersion, which is a common complication in count data. Overdispersion is when the variation in a data set is greater than what is allowed for by standard distributions. For example, in the SBIRT data, positive counts are assumed to have a zero-truncated binomial distribution, as they are counts with an upper limit. These zero-truncated binomial distributions have one parameter which controls both the mean and the variance, thus the mean and variance do not vary independently. If the data has more variation than what is expected for a given mean, standard distributions will yield misleading inferences. The multivariate random effects provide a mechanism with which to add more variation into the count data model, and help account for possible overdispersion.

1.3 Longitudinal Poisson Hurdle Models for Zero-Inflated Count Data with Variable Follow-up Times

Accessing healthcare can often be difficult for people living with HIV (PLHIV), and is vital for achieving viral suppression. One group that has particular trouble accessing care are HIV positive persons with a history of incarceration. HIV rates among persons in US correctional facilities is estimated to be 3 to 5 times that of the general population (Sabin, Frey, Horsley, & Greby, 2001; Maruschak, 2006). Studies have shown that while in prison, people with HIV have access to care, however upon release, many have difficulty or fail to link to care (Springer et al., 2004).

The Linking Inmates to Care in Los Angeles study (LINK LA) was a randomized controlled trial designed to improve linkage to and retention in care among recently incarcerated

HIV positive men and transgender women (Cunningham et al., 2018). In chapter 3, we model data from the LINK LA study to quantify the effect of a peer navigator intervention to help people recently released from jail access and engage in care. Subjects were recruited upon release from jail and researchers recorded the number of different types of medical visits subjects had attended over the 12 months prior to incarceration. Researchers planned to follow up with subjects at 3 months, 6 months and 12 months after release, and at each follow-up to record the medical visits attended since last follow-up. In many instances, subjects failed to access any care, resulting in a large number of zero observations. In chapter 3 we develop longitudinal Bayesian hurdle models for the LINK LA data to model number of primary care visits from the LINK LA data set.

Similar to chapter 2, we account for the correlation over time inherent to longitudinal studies. This data, however, presents a number of complications due to irregular follow up times. The study was designed to follow subjects over the first year after release from jail, but some follow-ups occurred as late as 3 years after release. Follow-ups were supposed to occur at 3 months, 6 months and 12 months, with each measuring the number of primary care visits since last follow-up. Some subjects' first follow-up, which were supposed to be 3 months after release from jail, did not occur until one year after release. Thus, observations at a given follow-up visit are not directly comparable between subjects. The probability of attending a primary care visit, and the expected number of primary care visits, would both vary with observation time. One subject's first follow-up includes 12 months of primary care visits, while another's includes just 3 months of visits. We call the time frame over which an observation is measured *exposure* (Baetschmann & Winkelmann, 2013) and develop longitudinal hurdle regression models which control for varying exposure for different observations.

In addition we want to know what effect the intervention had on primary care visits, and how this effect changes over time. Thus we allow zero and count rates to vary over time. Doing so, however, is complicated by the fact that we can not directly compare follow-up

reports of doctor visits between subjects, and that we do not know when within an observed time frame primary care visits occurred. To account for this, we introduce partially observed latent parameters coinciding with the time intervals over which primary care visits were intended to be observed – baseline, 0-3 months, 3-6 months, 6-12 months and 12+ months. We include the 12+ month interval as many subjects returned for follow-ups beyond the planned study period, which we wanted to estimate separate zero and count rates for.

We model the zero-inflated counts using weighted averages of the latent parameters corresponding to the time frame over which primary care visits were counted. For example, if a subject first returned at 10 months, that observation is estimated using a weighted average the 0-3 month, 3-6 month and 6-12 month parameters, with weights 3, 3, and 4, corresponding to the number of observed months in each interval. For each time interval, we estimate count model and zero model main effects for both treatment groups and a count random effect. Similar to chapter 2, each subject has a random intercept for the zero model and multivariate random effects for the count model. We fit and compare several covariance models for the count model random effects, showing them to again be an improvement over individual random intercepts.

1.4 Bayesian Zero-Inflated Model for Longitudinal Multivariate Outcomes with Exposure

The LINK LA study also recorded numbers of several specialty care visits and numbers of emergency room visits. To get a better understanding of the effect of intervention, we are interested in how often subjects are attending each kind of visit. This gives us insight into care use patterns and cost of care. For example, specialty medical care visits are estimated to cost nearly twice as much as primary care visits on average, and emergency room visits are estimated to cost over 6 times the price of primary care visits. We are interested not just in the degree to which PLHIV are accessing care, but also which types of medical visits

they are using to access care.

A naïve analysis would be to apply the hurdle models from chapter 3 to each type of medical visit, modeling each separately. However, it is likely that there is correlation between different kinds of visits within an individual. For example, a subject that is likely to go to a a lot of primary care visits may be more or less likely to have a lot of emergency room visits. We account for this correlation by jointly model primary care visits, emergency room visits and specialty care visits. Modeling different types of medical visits jointly as a multivariate outcome allows us to model the correlation between different types of medical visits, sharing information between models and better attributing variation within the data set.

While some multivariate zero-inflated models for cross sectional data have been developed (Li et al., 1999; Liu & Tian, 2015), these models are not designed for longitudinal data or overdispersion. We model within-individual correlation between outcomes using multivariate normal random effects in the count model, allowing each individual to have multiple random effects which differ across both time and outcome. The random effects model within-individual correlation both across time and across different outcomes, while also allowing for modeling of overdispersion. This allows us to make more detailed inference than in the univariate models of chapter 3 about subject linkage to care and the effectiveness of the LINK LA intervention. In chapter 4, we develop an autoregressive multivariate zero-inflated Poisson model for data with varying exposure.

1.5 Outline of Dissertation

We develop three classes of models for longitudinal zero-inflated data and present analyses of data coming from two longitudinal studies, which is organized into three chapters. The models presented in each chapter build upon those in the previous chapters. Each of the chapters is written to stand alone, thus there is some material repeated between chapters.

Chapter 2 develops Bayesian hurdle models for longitudinal data to model the number

of days of heavy drinking out of the previous 90 from the SBIRT study. Chapter 3 develops Bayesian hurdle models for number of primary care visits from the LINK LA data, incorporating the developments of chapter 2, while also extending the models to allow for irregular follow-up times and varying exposure. Chapter 4 develops an autoregressive multivariate extension of the models in chapter 3, which we use to simultaneously model multiple types of medical visits. Finally, chapter 5 includes a discussion of the findings from chapters 2 - 4.

CHAPTER 2

Bayesian Longitudinal Hurdle Models for Days of Heavy Drinking

2.1 Introduction

It is common in count data to encounter an abundance of zeros relative to standard distributions, which is known as zero-inflation. This often happens when zeros have a special significance within a study or can be created by a separate process from that which produces the counts. Researchers often have interest in both of these processes and the interaction between them. A standard class of models for such data are two part mixture models known as hurdle models. In this chapter, we introduce a class of longitudinal hurdle regression models for use with zero-inflated repeated measures count data. To demonstrate these models, we apply them to data coming from a randomized controlled trial by Karno et al. (2021) assessing the effect of a screening, brief intervention and referral to treatment (SBIRT) program to reduce substance use in a population of drug or alcohol users seeking mental health treatment. In particular, our interest is in modeling days of heavy drinking over the past 90 days, which is heavily zero-inflated.

A standard class of models for such data are two part mixture models known as hurdle models. One part of the model is a binary model of whether an observation was zero or positive, sometimes referred to as the *hurdle* part of the model. If an observation crosses this hurdle, and is therefore positive, it is modeled using a truncated count distribution such as a zero truncated Poisson, negative binomial or binomial distribution. For example, in the

SBIRT study, a subject may or may not be a person that engages in heavy alcohol use, this is the hurdle. If a subject does engage in heavy drinking, then we can model the degree of this engagement using a truncated count distribution.

The first hurdle models were introduced by Cragg (1971) using Tobit models (Tobin, 1958) and were further developed for count data by Mullahy (1986). In a hurdle model, the number of zeros are modeled using a binary model predicting the probability of a non-zero observation, which is commonly referred to as the zero model. If that observation is non-zero, that is, it passes the hurdle, the magnitude of the non-zero observation is modeled using a truncated count distribution. This contrasts with the closely related zero-inflated models, which model the counts with an untruncated distribution, allowing zeros to be generated by both the zero and count processes (Lambert, 1992). Thus, zero inflated models have a latent class interpretation in which subjects either belong to an at-risk population or a not at-risk population. In Lambert's model, which she demonstrates on machine manufacturing, machines can either be functioning or imperfect. The imperfect machines produce defects at some rate modeled by a Poisson distribution, which may include zero. The hurdle model version of this analysis would not attempt to classify machines producing zero defects as functioning (not at-risk) or imperfect (at-risk), instead classifying all machines producing zero defects as functioning and all machines producing any positive number of defects as defective, thereby avoiding any attempt at latent classification.

Earlier hurdle models estimated the zero and count models separately from each other, however this independence between model parts seems unlikely in practice, as observations that are more likely to be zero are also likely to have lower counts than other observations. Ridout et al. (1998) present a hurdle model which allows dependence between the two parts of the model by using one regression equation to predict both the probability of a zero and the mean of the count distribution, which is an adaptation of one of Lambert's zero-inflated models. Zero-inflated and hurdle models can also be used in applications other than count data, such as the model by Foundtas and Anastasopoulos (2018), which connects a probit

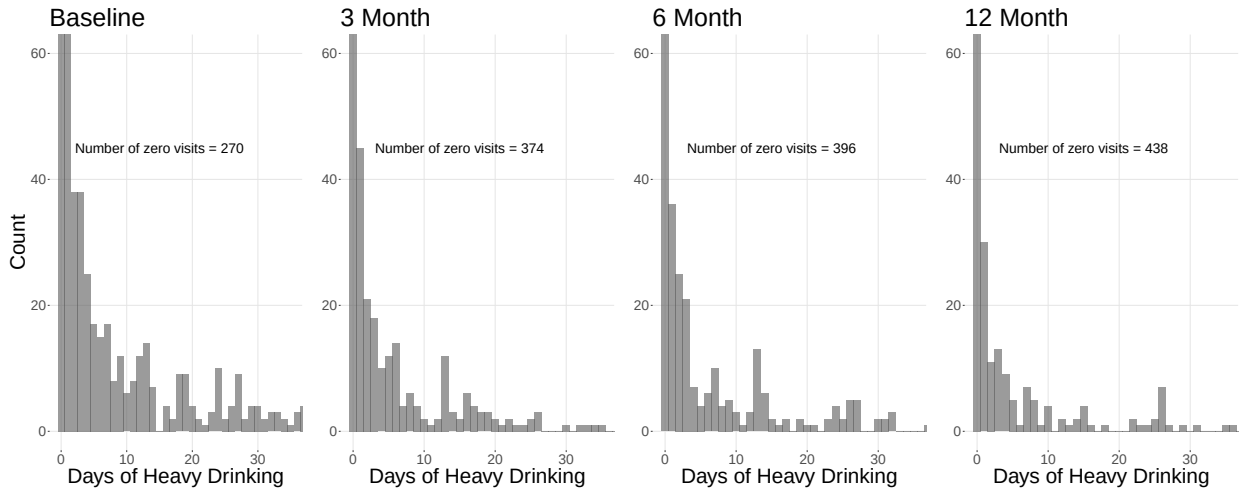


Figure 2.1: Number of days of heavy drinking over the last 90 days recorded at baseline, and each of the 3 follow-up visits. Data is shown to be heavily zero-inflated at all time points.

zero model to an ordered probit model for severity of injuries in car accidents.

Often researchers are interested in estimating the effects of covariates on the entire population, rather than on the subpopulations used for the zero and count models. This can be challenging in two part zero-inflated and hurdle models as it is not straightforward to interpret the coefficients from the zero and count models as effects on the entire population. Lee, Joo, Song, and Harper (2011) develop a marginalized hurdle model with random effects in both the zero and count model which are correlated with each other, and allows for estimation of marginal means using the likelihood. Long, Preisser, Herring, and Golin (2014) take another approach with a marginal zero-inflated Poisson (ZIP) model in which they regress the full population mean on covariates, rather than the mean of only the at-risk population, in addition to the degree of zero-inflation. This allows them to model parameters which describe the effect of covariates on the full population of interest.

Although both zero-inflated and hurdle models are often interchangeable, and should produce comparable inference, we prefer hurdle models for a couple of reasons. First, hurdle models are more flexible, allowing for zero deflation in addition to inflation. Second, in a zero-inflated model, zeros can be generated both by the zero model and by the count model,

whereas in a hurdle model, zeros can only be generated from the zero model. This can make fitting zero-inflated models a bit trickier, requiring use of methods such as the data augmentation approach proposed by Ghosh, Mukhopadhyay, and Lu (2006). In contrast, hurdle models cleanly separate the model into two parts, avoiding some complications of using a mixture model. Still, due to the similarities between hurdle and zero-inflated models, it is straightforward to adapt methodology between the two approaches.

We develop Poisson hurdle models to model longitudinal data where individuals are followed over time. Some researchers have proposed zero-inflated generalized estimating equations to model within unit correlation over time (Dobbie & Welsh, 2002; Hall & Zhang, 2004; Kong, Xu, Levy, & Datta, 2015). Hall (2000) presents zero-inflated Poisson and binomial models with random effects in the count part of the model to capture within unit correlation over repeated measures. Dagne (2004) develop a repeated measures zero-inflated Poisson model which has independent random intercepts for the count and zero models. Min and Agresti (2005) propose using jointly distributed random effects in both the zero and count parts of the models to capture correlation between the two model parts as well as within units over time. These were adapted to a Bayesian framework by Neelon et al. (2010). Both Neelon et al. and Min and Agresti model balanced data with equal follow-up times using random intercepts and neither explore the use of more complicated random effects models. Baetschmann and Winkelmann (2017) propose a dynamic hurdle model, which accounts for zero-inflation by modeling an underlying stochastic process allowing a different Poisson rate for the periods of time before and after the first event. More recently, Burger, Schall, Ferreira, and Chen (2019) developed a zero-inflated model using multivariate normal random effect vectors and a discrete Weibull distribution for robustness in the presence of outliers. Ghosal, Lau, Gaskins, and Kong (2020) present a spatiotemporal negative binomial hurdle model with multivariate normal random effects with an unstructured covariance matrix.

We propose a class of hurdle models which allow researchers the ability to flexibly model within individual correlation over time by allowing random effects to vary over the course of

the study, while also allowing for dependence between the zero and non-zero portions of the model. We present 5 possible covariance models that provide more complexity and flexibility than previous random intercept hurdle models. We then demonstrate these models on the SBIRT study by Karno et al. (2021) and make inference on the effectiveness of the SBIRT intervention on heavy drinking. SBIRT has been recently studied as an approach to help reduce rates of alcohol and substance use (Glass et al., 2015; Barata et al., 2017; Saitz, 2014). The SBIRT data consists of 718 patients aged 18 and older who were randomized into either SBIRT or standard of care treatment groups. Subjects were interviewed at baseline, as well as at 3, 6 and 12 months. The primary outcome of interest collected at each follow up was number of days out of the last 90 that a subject engaged in heavy alcohol use ($\geq$ 5 drinks for men, $\geq$ 4 drinks for women). Since the number of drinking days has an upper limit of 90, we model this outcome using a binomial distribution. In the SBIRT study some subjects were lost to follow-up or missed intermediary visits resulting in approximately 14% missing observations. We assume these observations to be missing at random and thus we do not expect them to bias inference (Little & Rubin, 2002).

2.2 Methods

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

$$P(Y_{ij} = 0|\pi_{ij}) = 1 - \pi_{ij}, \tag{2.1}$$

$$P(Y_{ij} = y_{ij}|\pi_{ij}, \theta_{ij}) = \pi_{ij}\frac{f(y_{ij}|\theta_{ij})}{1 - f(0|\theta_{ij})},\ y_{ij} = 1, \ldots, \infty \tag{2.2}$$

where π_{ij} is the probability of a non-zero response for subject i at time j, and $f(y_{ij}|\theta_{ij})$ is the probability density function of a discrete count distribution with parameter θ_{ij}, such as

the binomial, negative binomial or Poisson. This is the form of a standard hurdle model (Mullahy, 1986). Equation (2.2) is the probability $\mathrm{P}(Y_{ij} > 0|\pi_{ij}) = \pi_{ij}$ multiplied by the probability mass function of a zero truncated distribution with density $f(y_{ij}|\theta_{ij})$. When $\pi_{ij} = f(0|\theta_{ij})$, then this reduces to the distribution $f(y_{ij}|\theta_{ij})$ on the non-negative integers, and can be thought of as the situation in which there is neither zero inflation nor deflation. Often there is interest in both the zero and count parts of the model, as they represent different processes. For example, the process by which one abstains from heavy drinking over a 90 day period may be different than the process which determines the number of days of heavy alcohol use among users.

We can also consider both parts of the model together and calculate the mean of the hurdle distribution as

$$E(Y_{ij}|\pi_{ij}, \theta_{ij}) = \pi_{ij} \frac{\sum_{k=1}^{\infty} k f(k|\theta_{ij})}{1 - f(0|\theta_{ij})} \tag{2.3}$$

$$= \pi_{ij} \frac{E(Y_{ij}|\theta_{ij})}{1 - f(0|\theta_{ij})}. \tag{2.4}$$

For settings in which the counts have a maximum possible value, such as the SBIRT study where the reported number of days of heavy drinking in the past 90 days, a zero-truncated binomial distribution, Binomial(M, θ_{ij}), is a natural choice for the count model, where M is the number of trials and θ_{ij} is the probability of "success". For the SBIRT data, $M = 90$ as that is the number of days on which a subject can engage in heavy drinking, and θ_{ij} is the probability of heavy alcohol use in a day for subject i at visit j. For an outcome with Binomial(M, θ_{ij}) distribution, equation (2.2) becomes

$$P(Y_{ij} = y_{ij}|\pi_{ij}, \theta_{ij}) = \pi_{ij} \binom{M}{y_{ij}} \frac{\theta_{ij}^{y_{ij}} (1 - \theta_{ij})^{M - y_{ij}}}{1 - (1 - \theta_{ij})^{M}} \tag{2.5}$$

with mean

$$E(Y_{ij}|\pi_{ij},\theta_{ij}) = \pi_{ij}\frac{M\theta_{ij}}{1-(1-\theta_{ij})^M}. \tag{2.6}$$

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

$$g_1(\pi_{ij}) = \boldsymbol{X}_{1ij}\boldsymbol{\beta_1} + \gamma_{1i} \tag{2.7}$$

$$g_2(\theta_{ij}) = \boldsymbol{X}_{2ij}\boldsymbol{\beta_2} + \psi\gamma_{1i} + \gamma_{2ij} \tag{2.8}$$

where $\boldsymbol{X}_{1ij}$ and $\boldsymbol{X}_{2ij}$ are fixed effect covariate vectors for subject i at time j with corresponding unknown coefficient vectors $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$, and 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 and $\psi\gamma_{1i}$ models the association between the zero model and count model with ψ being a regression parameter. 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. For the binomial hurdle model $g_2(\theta_{ij}) = \text{logit}(\theta_{ij})$ as well. For other distributions, one may wish to use a different link function, such as log for the Poisson distribution.

In the SBIRT study, we are modeling zero and count rates over time. We treat time discretely and allow the population level zero and count rates to vary over the course of the study and between treatment groups. We use $\boldsymbol{X}_{1ij}$ and $\boldsymbol{X}_{2ij}$ to model time effects and treatment group by time interactions at each of the follow-up visits. We do not allow for a treatment effect at baseline as baseline measures apply to the time before the intervention was administered. Thus both $\boldsymbol{X}_{1ij}$ and $\boldsymbol{X}_{2ij}$ are 7-vectors where the first element is always 1, elements 2, 3 and 4 are indicator variables corresponding to the 3 month, 6 month, or 12 month follow-up, respectively, and 0 otherwise. Elements 5, 6 and 7 are interaction terms for the 3 month, 6 month and 12 month visit by treatment group.

Correlation across time can be accounted for in the prior distributions on γ_{1i} and $\boldsymbol{\gamma_{2i}}$.

We restrict γ_{1i} to be an individual random intercept constant across time as, compared to the count model, there is significantly less information in the binary data used for the zero model, although it would be straightforward to extend to more complex parameterizations. We define $\boldsymbol{\gamma_{2i}}$ as a vector of length J, where each individual's random effect may vary at each follow-up, allowing us to control for correlation between an individual's count measurements over time.

In previous studies it has been assumed that a random intercept model for both parts of the model is sufficient to account for within subject correlation (Min & Agresti, 2005; Neelon et al., 2010), however we test this assumption by specifying autoregressive, antedependent and unstructured correlation models for the count model random effects and fitting these models to the SBIRT data set.

Random Intercept The random intercept is the simplest random effect parameterization, and uses a single fixed individual effect to model the within-individual correlation over time letting $\gamma_{2ij} \equiv \gamma_{2i}$. Each subject i has two random intercepts, γ_{1i} for the zero model, and γ_{2i} for the count model, which are normally distributed

$$\gamma_{1i}|\sigma_1^2 \sim \mathrm{N}(0, \sigma_1^2) \tag{2.9}$$

$$\gamma_{2i}|\sigma_2^2 \sim \mathrm{N}(0, \sigma_2^2). \tag{2.10}$$

In contrast to the parameterization of Min and Agresti (2005) and Neelon et al. (2010), who use bivariate normal distributions, we use a different parameterization of the same model where γ_{1i} is a covariate with coefficient ψ in the count model regression to capture the dependency between hurdle model parts.

2.2.1 Covariance Models

To adapt the count model random effect distribution to more general covariance models, we offer a more flexible formulation of the mixed effects hurdle model. Rather than the model specification (2.9) and (2.10), we allow for the multivariate random effects that we set in (2.8).

$$\gamma_{1i}|\sigma_1^2 \sim N\left(0, \sigma_1^2\right) \tag{2.11}$$

$$\boldsymbol{\gamma}_{2i}|\boldsymbol{\Sigma}_2 \sim N_J(\mathbf{0}, \boldsymbol{\Sigma}_2). \tag{2.12}$$

We decompose the count model random effects covariance matrix as

$$\boldsymbol{\Sigma}_2 = \text{diag}(\boldsymbol{\sigma}_2)\ \boldsymbol{\Omega}\ \text{diag}(\boldsymbol{\sigma}_2), \tag{2.13}$$

where $\text{diag}(\boldsymbol{\sigma}_2)$ is a diagonal matrix with diagonal elements given by $\boldsymbol{\sigma}_2$, a J-vector of standard deviations. Within individual across time correlation is modeled with $J \times J$ correlation matrix $\boldsymbol{\Omega}$, for which we may use a parameterized correlation model such as autoregressive (AR) or antedependent (AD), or take to be unstructured. Matrix $\boldsymbol{\Omega}$ describes the correlations between an individual's count model random effects at each of the J follow-up visits. Let $\boldsymbol{\sigma}_2 = (\sigma_{21}, \ldots, \sigma_{2J})$, allowing for a different variance for each visit random effect for a heteroskedastic variance model. Alternatively, replace $\boldsymbol{\sigma}_2$ with the scalar $\sigma_{2\text{cv}}$, which assumes constant or homoskedastic variance for the random effects across all visits. We consider the autoregressive and antedependent covariance models, each of which we pair with the homoskedastic and heteroskedastic variance parameterizations, and we consider random intercept and unstructured models.

The decomposition of $\boldsymbol{\Sigma}_2$ into variance and correlation components is convenient as variances and correlations are more readily interpretable by researchers than covariance matrices. The decomposition allows one to more readily make modeling decisions regarding variance

and correlation models as well as making prior distribution decisions more straightforward. The following sections present three possible models for $\mathbf{\Omega}$, which we consider for modeling the SBIRT data.

2.2.1.1 Autoregressive Model

The autoregressive (AR) model is a one parameter correlation model assuming a constant correlation, ρ, between any two adjacent time periods. For the SBIRT data, with $J = 4$, the correlation matrix of $\boldsymbol{\gamma_{2i}}$ is

$$\text{Corr}(\boldsymbol{\gamma_{2i}}) = \mathbf{\Omega}_{\text{AR}}(\rho) = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}. \tag{2.14}$$

As time periods get more distant, the correlation between the corresponding random effects decays, which one would expect in most longitudinal data. An appeal of using the AR correlation model is that no matter how large J is, only one parameter, ρ, needs to be estimated. Further, likelihood and posterior calculation under the given models requires one to calculate $\mathbf{\Sigma}_2^{-1} = \text{diag}(\boldsymbol{\sigma}_2)^{-1}\mathbf{\Omega}^{-1}\text{diag}(\boldsymbol{\sigma}_2)^{-1}$, which is straightforward to compute regardless of size, given ρ and $\boldsymbol{\sigma}_2$. For the AR model, the precision matrix $\mathbf{\Omega}^{-1}$ is

$$\mathbf{\Omega}_{\text{AR}}(\rho)^{-1} = \begin{pmatrix} 1 & -\phi & 0 & 0 \\ -\phi & 1+\phi^2 & -\phi & 0 \\ 0 & -\phi & 1+\phi^2 & -\phi \\ 0 & 0 & -\phi & 1 \end{pmatrix}. \tag{2.15}$$

where $\rho = \frac{\phi}{1-\phi^2}$. For the SBIRT data, we consider an AR correlation structure with both homoskedastic variance (ARcv) and heterskedastic variance (AR).

The AR model may be inappropriate, if one believes that the correlation between random effects in adjacent time periods may not be equal across the entire study. For example in the SBIRT data, the AR model assumes that the correlation between random effects at baseline and the 3 month visit is the same as the correlation between the 6 month and 12 month visit. This may be considered unlikely given the difference in elapsed time between the visits.

2.2.1.2 Antedependent Model

Another correlation model, which is more flexible than the AR model, is the antedependent (AD) model, which, for a study with J time points, has $J-1$ distinct correlations between pairs of adjacent time intervals. AD may be a better candidate correlation model as compared to the AR model for the SBIRT study where the spacing between consecutive time points differs over the course of the study. For the SBIRT setting with $J = 4$ visits, the AD correlation matrix for $\boldsymbol{\gamma_{2i}}$ is

$$\text{Corr}(\boldsymbol{\gamma_{2i}}) = \Omega_{AD}(\boldsymbol{rho}) = \begin{pmatrix} 1 & \rho_1 & \rho_1\rho_2 & \rho_1\rho_2\rho_3 \\ \rho_1 & 1 & \rho_2 & \rho_2\rho_3 \\ \rho_1\rho_2 & \rho_2 & 1 & \rho_3 \\ \rho_1\rho_2\rho_3 & \rho_2\rho_3 & \rho_3 & 1 \end{pmatrix} \tag{2.16}$$

where the lag 1 correlations are given by the elements of $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{J-1})$ and higher lag correlations are given by the product of the intermediate lag 1 correlations. Thus the correlation between the random effects for intervals l and m, with $l < m$ is $\prod_{k=l}^{m-1} \rho_k$. The AD model offers more flexibility than the AR model while still restricting the number of parameters to be relatively small. In the AD model, a $J \times J$ correlation matrix is determined from $J-1$ parameters. Another benefit is that there exist closed form solutions for the elements of Ω_{AD}^{-1} given the vector $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{J-1})$, which can provide significant computational savings as

J grows large. We can solve for the elements of Ω_{AD}^{-1} as

$$\Omega_{lm}^{-1}(\boldsymbol{\rho}) = \begin{cases} \dfrac{1}{1-\rho_1^2} & l = m = 1 \\ \dfrac{1}{1-\rho_{L-1}^2} & l = m = J \\ \dfrac{1-\rho_{l-1}^2\rho_l^2}{(1-\rho_{l-1}^2)(1-\rho_l^2)} & l = m \neq 1, J \\ \dfrac{-\rho_l}{1-\rho_l^2} & l = m+1 \\ \Omega_{lm}^{-1} & l = m-1 \\ 0 & |l-m| > 1. \end{cases} \tag{2.17}$$

The last line is due to the conditional independence structure of the AD model, meaning the random effect at a given time interval is assumed to depend only on the random effects at adjacent time intervals, and is independent of the random effects at more distant time intervals conditional on the random effects at adjacent time intervals. As with the AR model, we consider the AD model with both homoskedastic (ADcv) and heteroskedastic (AD) variance.

2.2.1.3 Unstructured Model

We also consider the unstructured covariance model (UN), which makes no assumptions about the correlations between random effects. The cost of using UN is that the number of parameters grows as J^2. However, it offers an advantage in a Bayesian framework. By setting a scaled Inverse-Wishart prior (O'Malley & Zaslavsky, 2008) on the covariance matrix $\boldsymbol{\Sigma}_2$, we can then sample correlations from a conjugate Inverse-Wishart posterior. The scaled Inverse-Wishart distribution is defined by decomposing

$$\boldsymbol{\Sigma}_2 = \text{diag}(\boldsymbol{\omega})\ \boldsymbol{\Sigma}_\omega\ \text{diag}(\boldsymbol{\omega}) \tag{2.18}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_L)$ and $\boldsymbol{\Sigma}_\omega$ is a positive definite matrix with an Inverse-Wishart prior with degrees of freedom ν and $J \times J$ scale matrix $\boldsymbol{S}$. This is slightly different from the variance-correlation decomposition from the previous models, as $\boldsymbol{\Sigma}_\omega$ is not necessarily a correlation matrix. However, the correlation matrix, $\boldsymbol{\Omega}_{\text{UN}}$, is determined directly from $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$, since $\boldsymbol{\Omega}_{\text{UN}} = \text{diag}(\boldsymbol{\Sigma}_{\boldsymbol{\omega}})^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\omega}}\text{diag}(\boldsymbol{\Sigma}_{\boldsymbol{\omega}})^{-1/2}$ where $\text{diag}(\boldsymbol{\Sigma}_{\boldsymbol{\omega}})$ is a diagonal matrix with diagonal entries equal to the diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$. The parameter $\boldsymbol{\omega}$ scales the variances of $\boldsymbol{\Sigma}_2$, allowing us to estimate the random effect variances separately from the correlations through a Metropolis step. The full conditional posterior distribution of $\boldsymbol{\Sigma}_\omega$ is an Inverse-Wishart distribution, which we describe in more detail in the Appendix.

2.2.2 Prior Specification

Priors were selected to be generally non-informative and to let the data guide the inference. Elements of zero model fixed effects $\boldsymbol{\beta}_1$ and count model fixed effects $\boldsymbol{\beta}_2$ were given independent $N(0, 10^2)$ priors. Zero model random effects standard deviation σ_1 and count model standard deviations, $\sigma_{2\text{cv}}$ in the homoskedastic models, RI, ARcv and ADcv, and elements of $\boldsymbol{\sigma}_2$ in the heteroskedastic models, AR, AD and UN, were given half-normal $N^+(0, 1)$ priors where $N^+(a, b)$ is the Normal(a,b) distribution restricted to the positive domain. The regression coefficient ψ modeling the association between the zero and count models was also given a $N^+(0, 1)$, as we expect a positive correlation between probability of at least one day of heavy drinking and the expected number of days of heavy drinking per 90 within an individual. Correlation parameters ρ for the AR model and ρ_j $j = 1, \ldots, J-1$ for the AD model were assigned $N^+(.5, .25^2)$ priors truncated above at 1, as we expect within-individual counts to have positive correlation.

2.2.3 Posterior Sampling

Posterior sampling was performed using Markov Chain Monte Carlo (MCMC) methods methods (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970; Gelfand & Smith, 1990; Casella & George, 1992). Parameters were split into blocks and each block of parameters was sampled from its conditional posterior distribution.

Zero model main effects $\boldsymbol{\beta}_1$, count model main effects $\boldsymbol{\beta}_2$ and count model random effects $\boldsymbol{\gamma}_{2i}$ were sampled separately using Metropolis-Hastings algorithms. In each of these three cases, we used a normal approximation to the conditional posterior distribution using the first two terms of the Taylor series approximation to the log posterior distribution centered at the previous state of the Markov Chain. This method is described in detail by Rue and Held (2005) and requires calculation of the first and second derivatives of the conditional log posterior distributions.

The count model random effects variance for the unstructured model was sampled in two steps. The parameter modeling the random effect correlations $\boldsymbol{\Sigma}_\omega$ was sampled directly from its conditional posterior distribution. The scaling parameter $\boldsymbol{\omega}$ was sampled using a Gaussian random walk Metropolis algorithm scaled to achieve an optimal acceptance rate. All other parameters aside from those described in this section were sampled using Gaussian walk Metropolis algorithms. To perform our posterior sampling we ran 4 chains, each for 40,000 iterations with a thinning of 1, and discarded the first 20,000 iterations.

2.2.4 WAIC

To choose between models, we use the Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010). This is one of several measures of predictive accuracy, known as information criteria, which measure how well the fitted model predicts the data that was used to fit it, the in-sample predictive accuracy. There are two main challenges with measuring model fit. The first is that we do not know the true distribution of our outcome, and using

the distribution of our sample data as a proxy for the true distribution will bias us towards believing the model fit is better than it actually is. The second issue is that adding more parameters to a model guarantees us at least as good of a fit as a smaller model. Different information criteria, such as the Akaike information criterion (AIC), Bayesian information criterion (BIC), Deviance information criterion (DIC) and the WAIC all vary in how they address these issues.

AIC and BIC are both frequentist based methods that estimate the fit of the model based on the maximum likelihood point estimate and then correct for overfitting based on the number of parameters in the model. This is easily done for fixed effects models, however in mixed effect models counting the number of parameters is not as straight forward. Random effects contribute less to overfitting than independent fixed effect parameters, and informative priors further reduce this overfitting. Therefore the Bayesian based DIC and WAIC estimate the number of *effective* parameters of the models. We prefer WAIC to DIC as its estimate for the effective number of parameters tends to be more stable (Gelman, Hwang, & Vehtari, 2014). WAIC also takes a more fully Bayesian approach to estimation by averaging over the entire posterior distribution, rather than calculating fit using only the posterior mean.

WAIC is defined as

$$\text{WAIC} = -2(\text{lppd} - p_{\text{WAIC}}), \tag{2.19}$$

where lppd is the log of the pointwise predictive density and p_{WAIC} is the effective number

of parameters. Defining parameter vector $\boldsymbol{\zeta}_{ij} = (\pi'_{ij}, \theta'_{ij})$, we have

$$\text{lppd} = \log \prod_{i=1}^{N} \prod_{j=1}^{J} p_{\text{post}}(y_{ij}) \tag{2.20}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{J} \log \int p(y_{ij}|\boldsymbol{\zeta}_{ij}) p_{\text{post}}(\boldsymbol{\zeta}_{ij}) d\boldsymbol{\zeta}_{ij}, \tag{2.21}$$

$$p_{\text{WAIC}} = \sum_{i=1}^{N} \sum_{j=1}^{J} \text{var}_{\text{post}}(\log(p(y_{ij}|\boldsymbol{\zeta}_{ij}))), \tag{2.22}$$

where $p_{\text{post}}(y_{ij})$ is the predictive probability of observation y_{ij} averaged over the posterior distribution $p_{\text{post}}(\boldsymbol{\zeta}_{ij})$ as defined in (2.21) and var_{post} is the posterior variance. The integral in (2.20) and the variance in (2.22) are estimated by averaging across the S posterior samples $\boldsymbol{\zeta}_{ij}^s$, $s = 1, \ldots, S$,

$$\widehat{\text{lppd}} = \sum_{i=1}^{N} \sum_{j=1}^{J} \log \left(\frac{1}{S} \sum_{i=1}^{S} p(y_{ij}|\boldsymbol{\zeta}_{ij}^s) \right) \tag{2.23}$$

$$\hat{p}_{\text{WAIC}} = \sum_{i=1}^{N} \sum_{j=1}^{J} V_{s=1}^{S}(\log p(y_{ij}|\boldsymbol{\zeta}_{ij}^s)), \tag{2.24}$$

where $V_{s=1}^{S}$ represents the sample variance, where the sample variance of $a_1, \ldots, a_S$ with sample mean $\bar{a}$ is $V_{s=1}^{S}(a_s) = \frac{1}{S-1} \sum_{s=1}^{S} (a_s - \bar{a})^2$ and S is assumed large enough to capture the posterior distribution. In the hurdle model $p(y_{ij}|\boldsymbol{\zeta}_{ij})$ is given by equations (2.1) and (2.2).

2.2.5 Posterior Inferences

A benefit of fitting these models using Bayesian methods is that it is straightforward to produce inference on a number of different quantities of interest. Let θ_{jc} be the binomial count model parameter at visit j for treatment group $c = 0, 1$ where $c = 0$ for the control group and $c = 1$ for the SBIRT group. Further, $\theta_{10} = \theta_{11} = \theta_1$ since we assume no difference

between treatment groups at baseline. The count model mean, which is the mean number of days of heavy drinking in the last $M = 90$ among subjects that engage in at least one day of heavy drinking at visit j for treatment group c is

$$\mu_{2jc} = \frac{M\theta_{jc}}{1 - (1 - \theta_{jc})^M}. \tag{2.25}$$

Let π_{jc} be the proportion of subjects that engage in heavy drinking at visit $j = 1, 2, 3, 4$ in treatment group $c = 0, 1$, with $\pi_{10} = \pi_{11} = \pi_1$. Then the expected number of days of heavy drinking out of the past 90 for all subjects in treatment group c at visit j is

$$\mu_{jc} = \pi_{jc}\mu_{2jc}, \tag{2.26}$$

which we call the full hurdle mean.

While π_{jc}, μ_{2jc} and μ_{jc} are all of interest to us, our primary goal in this analysis is to assess how effective the SBIRT intervention is compared to the standard of care at getting subjects to reduce heavy drinking. We quantify this with the difference of differences DoD_j, the difference in change from baseline at visit j in expected number of days of heavy drinking between treatment and control groups, where

$$\text{DoD}_j = (\mu_{j1} - \mu_1) - (\mu_{j0} - \mu_1) \tag{2.27}$$

$$= \mu_{j1} - \mu_{j0}. \tag{2.28}$$

We have secondary interest in the difference of differences in proportion of subjects at visit j who use any alcohol, $\text{DoD}_{1j} = \pi_{j1} - \pi_{j0}$, and the expected number of days of heavy drinking among those who drank, $\text{DoD}_{2j} = \mu_{2j1} - \mu_{2j0}$, where μ_{2j1} and μ_{2j0} are given by equation (2.25).

In certain contexts, it is useful to have a single number to summarize the effectiveness of the intervention over the follow-up period. For this we use the difference of differences of

the mean use over the three follow-up visits

$$\text{MDoD} = \frac{1}{3}\sum_{j=2}^{4}\mu_{j1} - \frac{1}{3}\sum_{j=2}^{4}\mu_{j0}. \tag{2.29}$$

We can similarly construct MDoD_1 for the zero model as a single overall measure of how well SBIRT reduced the proportion of people who drank heavily and MDoD_2 for the count model to measure how well it reduced rates of heavy drinking among drinkers.

All of the inferences described in this section first require estimation of π_{jc} and θ_{jc}. To do this, we have to integrate out the random effects, which we do using Monte Carlo integration. Let $\gamma_1^{(r_1)}$ and $\boldsymbol{\gamma}_2^{(r_2)}$ be random samples from the random effects distributions (2.11) and (2.12), with $r_1 = 1, \ldots, R_1$ and $r_2 = 1, \ldots, R_2$ where R_1 and R_2 are large enough to capture the distributions of γ_1 and $\boldsymbol{\gamma}_2$. We calculate

$$\pi_{jc} = \int_{-\infty}^{\infty} \text{logit}^{-1}(\beta_{1jc} + \gamma_1)p(\gamma_1)d\gamma_1 \tag{2.30}$$

$$\approx \frac{1}{R_1}\sum_{r=1}^{R_1} \text{logit}^{-1}(\beta_{1jc} + \gamma_1^{(r_1)}) \tag{2.31}$$

and

$$\theta_{jc} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \text{logit}^{-1}(\beta_{2jc} + \psi\gamma_1)p(\gamma_1)d\gamma_1 d\gamma_2 \tag{2.32}$$

$$\approx \frac{1}{R_2}\sum_{r=1}^{R_2}\frac{1}{R_1}\sum_{r=1}^{R_1} \text{logit}^{-1}(\beta_{1jc} + \psi\gamma_1^{(r_1)} + \gamma_{2j}^{(r_2)}). \tag{2.33}$$

As (2.31) and (2.33) require knowledge of β_{1jc}, β_{2jc} and ψ, we calculate (2.31) and (2.33) for each sample $s = 1, \ldots, S$ from from the posterior. We use the collection of posterior draws to produce mean estimates and 95% Bayesian credible intervals (CrI) for each inferential target of interest. For this analysis we consider a statistically significant result to be found if the 95% Bayesian credible interval does not include zero.

2.3 SBIRT Data Analysis

We apply the models of the previous section to the SBIRT data set, in which the main outcome of interest is number of days of heavy drinking over the previous $M = 90$ days. The sample consists of $N = 718$ patients aged 18 and older whom reported use of alcohol, cannabis or stimulants within the past 90 days. These patients were randomized into either standard of care or a screening, brief intervention and referral to treatment (SBIRT) intervention group. Subjects were interviewed at 3 months, 6 months and 12 months and days of heavy drinking over the past 90 days was recorded. Not all subjects were recruited into the study for alcohol use, thus there was a considerable amount of zero-inflation (Figure 2.1). Our primary interest is to compare the reduction of heavy drinking in the SBIRT group to the control group.

We fit six different binomial hurdle models to the SBIRT data: a random intercept (RI) model, and five multivariate random effects models with autoregressive (AR), autoregressive constant variance (ARcv), antedependent (AD), antedependent constant variance (ADcv), and unstructured (UN) covariance models. We then compare inferences and model fits for these six models. Table 2.1 gives the WAIC from each model in order of model preference. Models using a multivariate normal random effects model (AR, ARcv, AD, ADcv and UN) all show a sizable improvement in model fit over the random intercept model. The random intercept model performs quite poorly as measured by WAIC, suggesting that the random intercept model is not flexible enough to capture the variation in the data. Differences in fit between the multivariate random effect models were modest, though UN had the smallest WAIC and is the preferred model.

2.3.1 SBIRT Treatment Effect

Table 2.2 provides posterior means and credible intervals of SBIRT difference of differences (DoD) treatment effects measured in average number of days of heavy drinking at each of the

	WAIC
UN	7180.7
AD	7181.2
AR	7181.4
ARcv	7183.2
ADcv	7186.2
RI	18975.9

Table 2.1: WAIC for the 6 covariance models: unstructured (UN), autoregressive (AR), antedependent (AD), autoregressive constant variance (ARcv), antedependent constant variance (ADcv) and random intercept (RI) models. A lower WAIC indicates a better fit. UN is the best fitting model.

three follow up points, and the average estimated treatment effect across the three follow-up time points. Only results for the hetereskedastic models are given, as the heteroskedastic models all had lower WAICs than their homoskedastic counterparts. Random intercept model results are also given for comparison. On average, over the one year follow-up period, assuming the UN model, subjects assigned to the SBIRT groups reduced their heavy drinking by .55 days per 90 as compared to those receiving standard of care. The UN, AD and AR models all had no significant treatment effects throughout the study period. Further, the treatment loses effectiveness compared to standard of care over the course of the 12 month follow-up period, with the most negative treatment effect posterior mean occurring at 3 month visit.

The random intercept model did find a significant treatment effect at the 3 month follow-up. This result is not shared by the better fitting multivariate random effect models, UN, AR and AD. In general, the RI model has more narrow credible intervals that the multivariate random effect models, suggesting that the RI model is overconfident in its findings. One important feature of these models is that in both a Bernoulli distribution and a binomial distribution, as we used in the zero and count models, there is only one parameter to control both the mean and the variance, so the mean and variance cannot shift independently of each other. The introduction of random intercepts allows some extra variation, but not as much as the multivariate random effects models. In this case, the RI model is understating the

	UN	AD	AR	RI
3 Month	-1.33 (-2.89, .15)	-1.54 (-3.65, .54)	-1.54 (-3.63, .54)	-1.46 (-2.37, -.59)
6 Month	-.58 (-2.17, .97)	-.12 (-2.30, 2.04)	-.11 (-2.27, 2.05)	-.50 (-1.44, .41)
12 Month	.26 (-1.01, 1.57)	.10 (-1.86, 2.05)	.12 (-1.83, 2.09)	.58 (-.28, 1.47)
All F-U	-.55 (-1.58, .47)	-.52 (-1.93, .88)	-.51 (-1.91, .89)	-.46 (-1.09, .16)

Table 2.2: Posterior mean difference of differences and 95% Bayesian credible intervals (CrI) for number of expected days of heavy drinking for the three best fitting models (AR, AD and Unstructured) and the random intercept model for comparison. Inference is given for the 3 month, 6 month and 12 month follow-up visits and the average of the three follow-up points.

variability of the data, leading to more narrow credible intervals and over-confident results.

Figure 2.2 plots the posterior means and 95% Bayesian credible intervals for the RI and UN models for both SBIRT and control treatment groups. The observed means at each time point from the data are also plotted for comparison. The UN model does a better job estimating the means, particularly at baseline. As the random intercept model shares individual effects across all time points, the dependency between time points may make it more difficult for the RI model to estimate the much higher baseline use rates. The RI model also has more narrow credible intervals suggesting that the model is overconfident in its estimates.

Table 2.3 gives posterior means and 95% Bayesian credible intervals from each of the two parts of the hurdle model using the preferred UN covariance model. There is very little difference between the two groups in the zero model, so the intervention is not more effective than standard of care at getting people to abstain from heavy alcohol use. Posterior means of the number of days of heavy drinking among drinkers from the count model at each follow-up are slightly higher in the SBIRT group than the control group, although the differences are not significant. Among drinkers, we estimate subjects assigned to the SBIRT group reduced their heavy drinking by 2.13 days more than those in the standard of care group at three

		Zero Model		Count Model	
		Ctrl	SBIRT	Ctrl	SBIRT
Mean	Baseline	.61 (.57, .64)		20 (17.8, 22.1)	
	3 Month	.38 (.32, .43)	.33 (.28, .38)	14 (11.1, 17.3)	11.9 (9.2, 14.9)
	6 Month	.32 (.28, .37)	.32 (.27, .37)	13.8 (1.5, 17.5)	12.2 (9.2, 15.7)
	12 Month	.23 (.19, .28)	.22 (.18, .27)	1.4 (7.2, 14.4)	12 (8.4, 16.5)
DoD	3 Month	-.04 (-.11, .02)		-2.13 (-5.56, 1.23)	
	6 Month	-.005 (-.07, .06)		-1.60 (-5.63, 2.36)	
	12 Month	-.01 (-.07, .05)		1.60 (-3.02, 6.36)	
	All F-U	-.02 (-.07, .03)		-.71 (-3.49, 2.10)	
	ψ			.04 (.003, .09)	

Table 2.3: UN model mean estimates and 95% Bayesian CrI from zero and count models for each treatment group at each time point for SBIRT data for both the control and the SBIRT group. Zero model estimates are proportion of subjects that engage in heavy alcohol use, count model estimates are the expected number of days of heavy alcohol use per 90 days among users. The DoD portion of the table gives mean difference of differences (DoD) estimates for each part of the hurdle model at each time point as well as the average across all follow-up time points (All F-U). The parameter ψ models the dependency between zero and count models.

months and 1.6 days at 6 months. The posterior mean at 12 months estimates the control group to drink 1.6 less days per 9.

Estimates and credible intervals for the count model random effect correlation matrices are given in table 2.4. The unstructured model produced substantially different estimates from the 4 structured correlation models with the UN estimating correlations to be much higher than the other models. While the autoregressive and antedependent models assume correlation to decay with spacing between visits, the UN model does not make any such assumptions, however, even the adjacent time point correlations are much higher in the UN model compared to the other multivariate random effect models. The differences in correlation estimates do not result in large differences in ability of the models to fit the data. The disagreement between the models may be due to the relative sparsity of observed positive counts compared to the number of random effects being estimated, which is one per person per visit regardless of whether or not there was a positive observation. Thus, the model

fit may not be very sensitive to changes in the multivariate random effects distributions. In addition, that the WAICs were similar between all 5 multivariate normal random effects models, despite the difference in the UN model correlation estimates suggests that the more complex models mainly improve fit through modeling over dispersion rather than within individual correlation.

Table 2.5 gives the posterior means and credible intervals of the zero and count model random effects standard deviations. There is general agreement between all models. The posterior means from the heterskedastic models, AR, AD and UN, are supportive of the heteroskedastic assumption. The AR and AD models both find a significant difference between 3 month and 12 month random effects variance, and credible intervals in the UN model only narrowly overlap.

Model		3 Month	3-6 months	6-12 months
UN	Baseline	.5 (.38, .61)	.45 (.33, .57)	.51 (.36, .63)
	3 Months		.62 (.5, .72)	.45 (.29, .59)
	6 Months			.63 (.48, .74)
AD	Baseline	.09 (.03, .16)	.01 (.00, .03)	.00 (.00, .00)
	3 Months		.10 (.03, .17)	.01 (.00, .02)
	6 Months			.07 (.01, .14)
ADcv	Baseline	.09 (.02, .15)	.01 (.00, .02)	.00 (.00, .00)
	3 Months		.09 (.03, .16)	.01 (.00, .02)
	6 Months			.08 (.02, .14)
AR	Baseline	.09 (.05, .13)	.01 (.00, .02)	.00 (.00, .00)
	3 Months		.09 (.05, .13)	.01 (.00, .02)
	6 Months			.09 (.05, .13)
ARcv	Baseline	.08 (.04, .12)	.01 (.00, .02)	.00 (.00, .00)
	3 Months		.08 (.04, .12)	.01 (.00, .02)
	6 Months			.08 (.04, .12)

Table 2.4: Count model random effects correlation matrix posterior means and Bayesian 95% credible intervals.

	Zero Model	Count Model			
		Baseline	3 Months	6 Months	12 Months
AR	2.04 (1.78, 2.32)	2.01 (1.87, 2.17)	1.81 (1.63, 2.00)	2.08 (1.87, 2.31)	2.32 (2.02, 2.74)
AD	2.03 (1.78, 2.31)	2.02 (1.88, 2.17)	1.82 (1.64, 2.03)	2.08 (1.88, 2.33)	2.36 (2.03, 2.67)
UN	2.04 (1.78, 2.31)	2.02 (1.88, 2.17)	1.85 (1.66, 2.06)	2.09 (1.87, 2.32)	2.31 (2.02, 2.63)
ADcv	2.04 (1.78, 2.32)	2.02 (1.93, 2.13)	- -	- -	- -
ARcv	2.04 (1.80, 2.32)	2.02 (1.93, 2.13)	- -	- -	- -
RI	2.04 (1.79, 2.31)	2.01 (1.88, 2.14)	- -	- -	- -

Table 2.5: Standard deviation Posterior means and 95% Bayesian Credible intervals for the standard deviations from zero model random effects and count model random effects for each model for the SBIRT data. Constant variance count models (ADcv, ARcv and RI) have one standard deviation estimate shared across the entire study, while the UN, AD, and AR have a separate count model random effects standard deviation for each time interval.

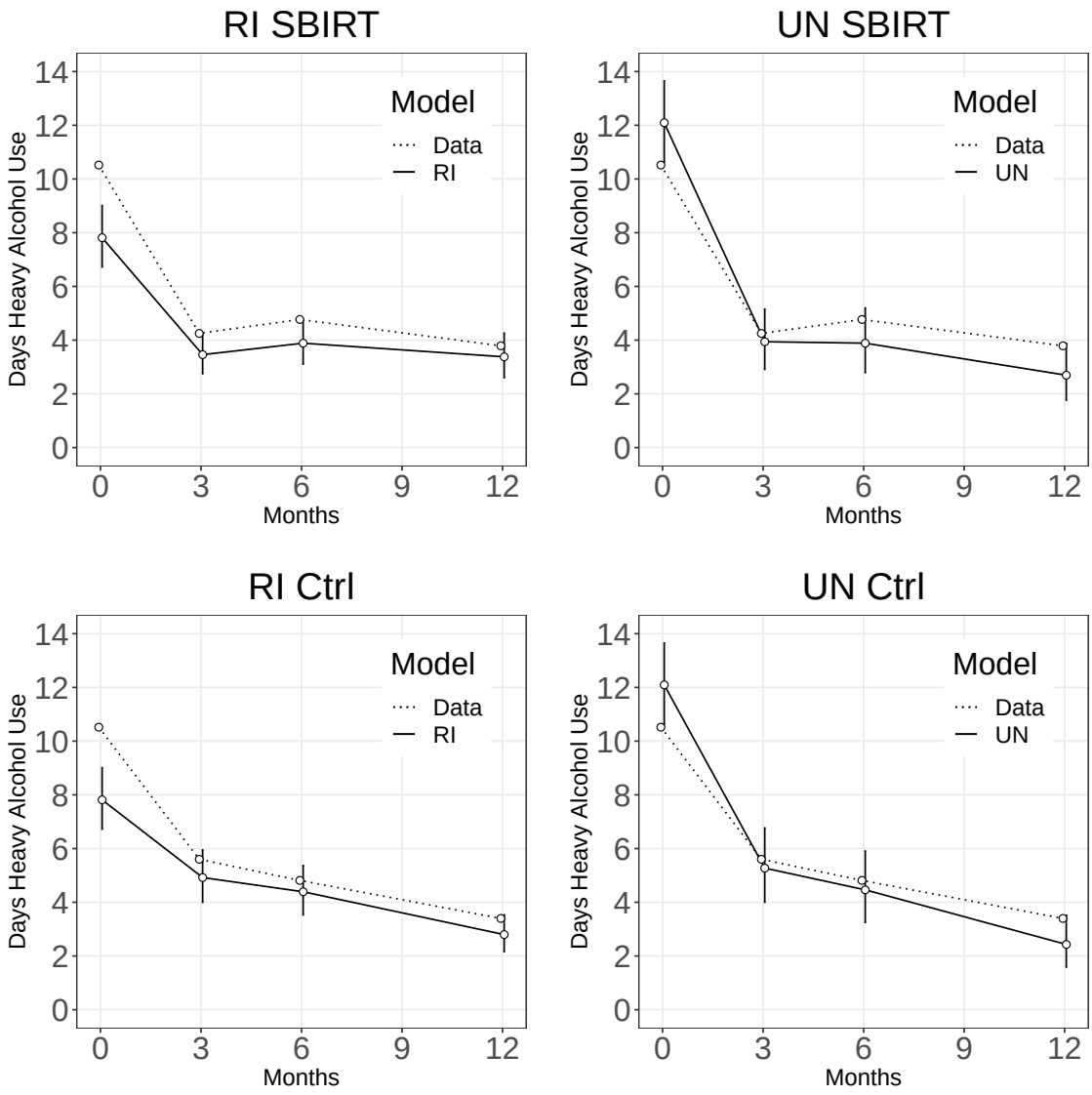


Figure 2.2: Mean number of days of heavy alcohol use at baseline and each follow-up given by the unstructured (UN) and random intercept (RI) models with the observed values from the data plotted for comparison.

2.4 Discussion

In this chapter we developed Bayesian random effect hurdle models for use in zero-inflated data with repeated measures. We demonstrated 5 different covariance models, which we applied to the SBIRT data to model days of heavy drinking in the past 90 days. All models except the random intercept model performed similarly as measured by WAIC, with the unstructured model having the best fit. We introduced multivariate random effects, allowing researchers additional flexibility to model both within individual correlation over time and overdispersion in the data.

We attempted to fit random intercept hurdle models using maximum likelihood methods, using the glmmTMB R package (Brooks et al., 2017) and failed. By comparison, the Bayesian methods presented in this chapter allowed us to fit the random intercept models and 5 other more complex models. In comparing our 5 multivariate random effects models to the random intercept model, we find not only is fit improved, but inference is different. While the random intercept model found a treatment effect in 3 month difference of differences, this result was not shared by the multivariate models.

In hurdle models, the count models generally need to be fit with large amounts of missing data. In the zero model, all observed data contributes to the likelihood. In the count model, only positive observations are used, but unobserved data and zero values are treated as missing data. Count model random effects for missing or zero observations are informed by data for the same subject at other observations through the covariance models, or, if a subject has no positive observations, the posterior of the random effects for that subject are the same as the prior distribution.

None of the multivariate random effects models found a significant difference between treatment groups, however the random intercept model found the SBIRT group to perform better at 3 months. The results presented here cast doubt on that result as none of the better fitting models agreed with this result and generally found the random intercept to be

a poor fit as well as over-confident in its results. This shows the importance of multivariate random effect hurdle models for longitudinal zero-inflated data. In this chapter we proposed 5 possible covariance models, however it is straightforward to use any covariance model that the researchers deem reasonable and to compare model fit to select the best one for the data.

Appendix A

Scaled Inverse-Wishart Posterior

One problem with a standard Inverse-Wishart prior distribution for a covariance matrix is that varainces and correlations cannot vary independently of each other. The scaled Inverse-Wishart distribution proposed by O'Malley and Zaslavsky (2008), decomposes the covariance matrix into two parts, $\boldsymbol{\Sigma}_2 = \text{diag}(\boldsymbol{\omega})\boldsymbol{\Sigma}_\omega\text{diag}(\boldsymbol{\omega})$. In this decomposition, $\boldsymbol{\Sigma}_\omega$ models correlations and some portion of the variances, and $\boldsymbol{\omega}$ scales the covariance matrix up or down such that the total variance may vary independently of the correlations. When used as a prior on the covariance matrix for a multivariate normal distribution, $\boldsymbol{\Sigma}_\omega$ has a conjugate Inverse-Wishart conditional posterior distribution. Letting $\boldsymbol{\Sigma}_\omega$ have prior distribution I-W(ν, S), meaning Inverse-Wishart with degrees of freedom ν and scale parameter S. Then the posterior distribution of $\boldsymbol{\Sigma}_\omega|\boldsymbol{\gamma}_2, a, B \sim$ I-W(a, B) where

$$a = \nu + N \tag{2.34}$$

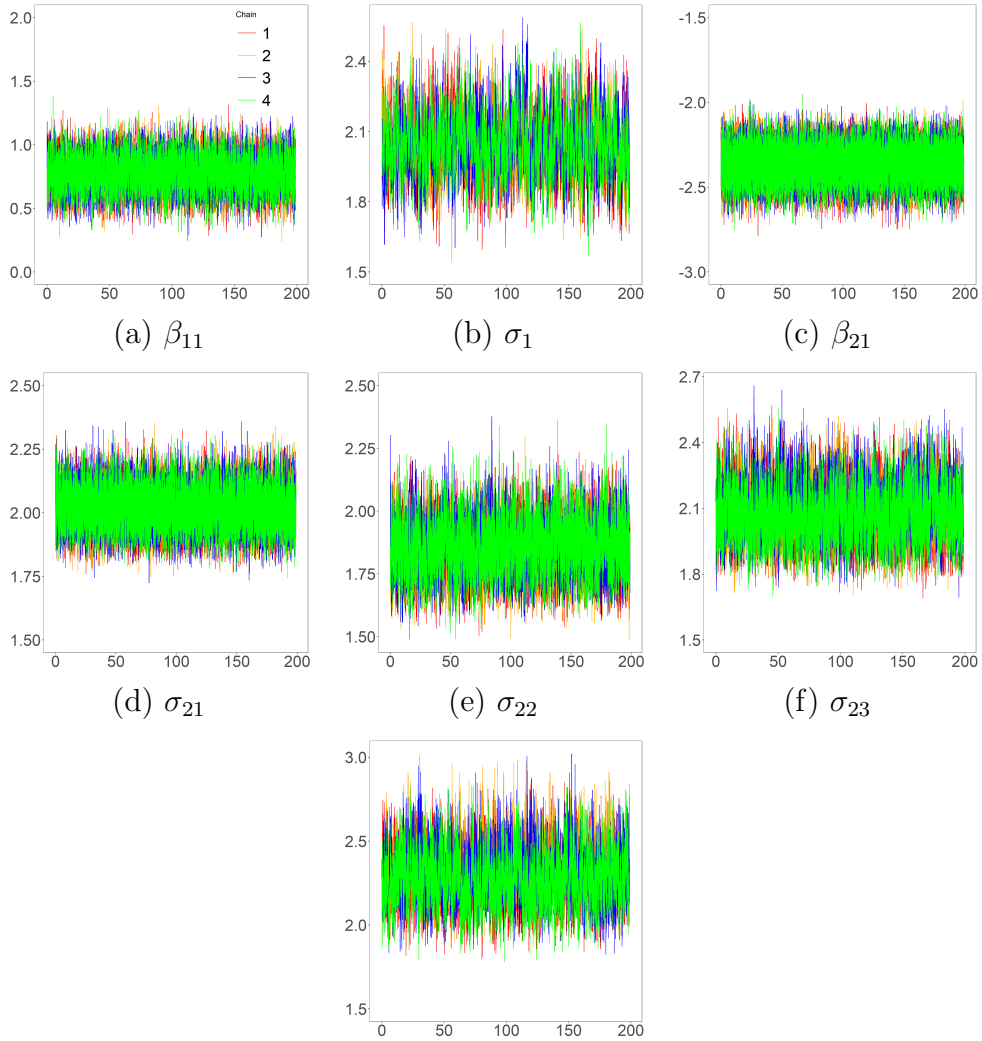
$$B = S + \text{diag}(\boldsymbol{\omega})^{-1}\left(\sum_{i=1}^{N}\boldsymbol{\gamma}_{2i}\boldsymbol{\gamma}_{2i}'\right)\text{diag}(\boldsymbol{\omega})^{-1}, \tag{2.35}$$

allowing us to sample $\boldsymbol{\Sigma}_\omega$ directly from its conditional posterior distribution.

The posterior for $\boldsymbol{\omega}$ does not have a conjugate form. We assign elements of $\boldsymbol{\omega}$ to be a priori independent half-normal $\mathbf{N}^+(0, 2)$ and sample from the posterior using a Gaussian random walk Metropolis algorithm.

Trace Plots

To demonstrate model convergence for the preferred UN model we provide trace plots of 6 different parameters, the baseline main effects β_{11} for the zero model, the baseline main effects β_{21} for the count model, the zero model random effects standard deviation σ_1, and the count model random effects standard deviation at each of the 4 time points, σ_{21}, σ_{22}, σ_{23} and σ_{24}. The MCMC sampler was run with 4 chains of 40,000 samples, discarding the first 20,000 with a thinning of 1. Posterior samples for all parameters were found to have satisfactory mixing and convergence. Trace plots also support a heteroskedastic type model, agreeing with WAIC results.



(a) β_{11} (b) σ_1 (c) β_{21}

(d) σ_{21} (e) σ_{22} (f) σ_{23}

(g) σ_{24}

Figure 2.3: Trace plots for UN model baseline main effects from the zero model β_{11} and count model β_{21}, as well as zero model random effects variance σ_1 and count model random effects variances, σ_{21}, σ_{22}, σ_{23} and σ_{24}.

CHAPTER 3

Bayesian Longitudinal Hurdle Models with Varying Exposure

3.1 Introduction

Successfully addressing the HIV pandemic requires linking people living with HIV (PLWH) to medical care, and then retaining them in that care. Failure to attend regular medical visits has been shown to be associated with higher levels of HIV viremia (Mugavero et al., 2009). We analyze data on primary care visits for PLWH with a history of incarceration, who have been shown to access care much less frequently than the general population of PLWH (Sabin et al., 2001; Maruschak, 2006). We model number of primary care visits in a population of men and transgender women with HIV recently released from LA county jail. The data is heavily zero-inflated as the study cohort faces many barriers to access and utilization of care. Further, the time over which visits are counted varies between observations. To model the data we develop a Bayesian longitudinal hurdle model accounting for exposure.

Hurdle models were first proposed by Cragg (1971), and then refined for count outcomes by Mullahy (1986). Hurdle models are two-part models in which one part models the proportion of zeros and the other part models the positive counts. The motivation for separate modeling of zeros is that standard count distributions cannot account for the excess zeros, which can often be driven by a separate process from that which determines the counts. For example, in the LINK LA data, one may consider the number of doctor visits observed as a realization of two separate processes, the first being whether or not subjects have access to

a doctor, and the second being how often they go to the doctor if they do have access.

While early versions of hurdle models estimated the two parts of the model separately, it is unlikely in practice for the zero probabilities and count rates to be independent of each other. Ridout et al. (1998) address this issue for Poisson hurdle models by developing a model which uses the same regression equation in both the zero and count models. Min and Agresti (2005) use correlated random intercepts in both the zero and count models to model the association between the zero and count models. The random intercepts also model within-individual correlation over time in longitudinal studies where the same subjects are followed and measured repeatedly. Neelon et al. (2010) adapt this model to the Bayesian setting for longitudinal data. Generalized estimating equations have also been used to model longitudinal zero-inflated data to account for within-unit correlation over time (Dobbie & Welsh, 2002; Hall & Zhang, 2004; Kong et al., 2015).

We analyze data from the Linking Inmates to Care study (LINK LA) (Cunningham et al., 2018). LINK LA was a randomized controlled trial of 356 HIV positive men and transgender women recently released from LA County Jail in which subjects were randomized into either receiving intervention or standard of care. The intervention provided peer navigated assistance with linkage, retention and adherence to HIV care, which we compare to a standard transitional case management control. Subjects were followed over time and at each follow-up number of primary care visits since last follow-up was recorded. Data were intended to be collected upon release from jail (baseline) and then at 3, 6 and 12 months following release. Histograms of number of primary care visits since last follow-up are given in figure 3.1 for baseline and each of the 3 follow-ups. We compare the number of doctor visits in the first year following release from jail between the intervention and control groups.

In LINK LA, observed counts of doctor visits accrued over differing amounts of time, which is important to account for as it affects the expected counts and zero probabilities. The greater the elapsed time over which doctor visits accrue, the more opportunity one has to attend doctor visits. The elapsed time over which counts accumulate is called *exposure*. In

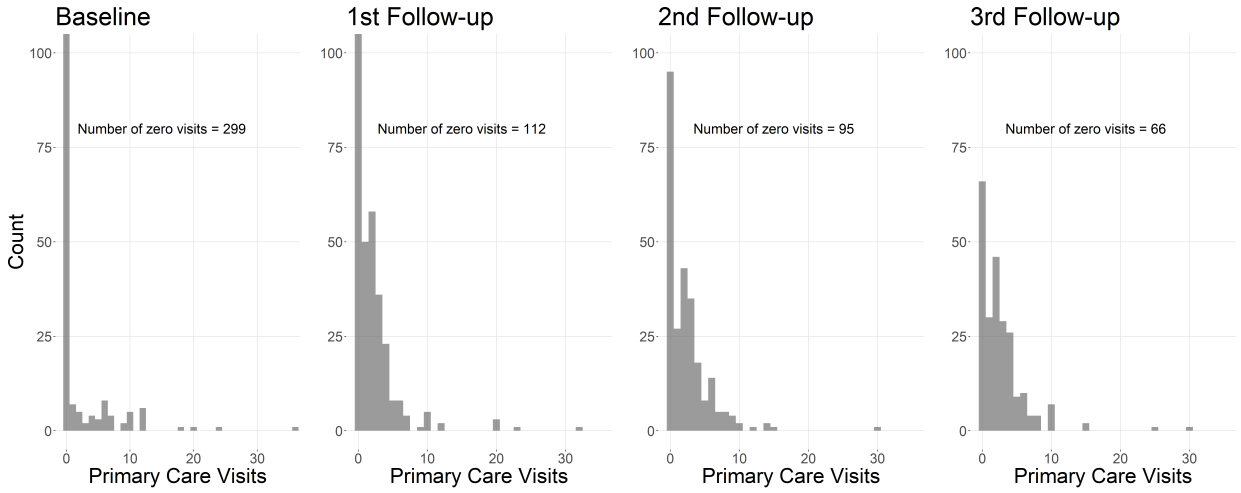


Figure 3.1: Number of primary visits recorded at baseline, and each of the 3 follow-up visits.

the zero model we can account for exposure through the use of a complementary log log link function with a log exposure offset (Baetschmann & Winkelmann, 2013), which is also a form of a Weibull survival model. The survival approach to the zero model follows by considering the zeros as right-censored time to event data, where the event is attending the first primary care visit and the probability of zero visits is the probability that the time is censored. In the Poisson count model, exposure can be accounted for with the combination of a log link function and a log exposure offset in the regression equation, which appropriately scales the expected counts to any given length of time (Cameron & Trivedi, 1998).

An added benefit of using the complementary log log link function for the zero model in Poisson hurdle models is that it makes the models *compatible* (Heilbron, 1994). This means that if the zero model and count model are given the same regression equation with the same covariates and coefficients, the expected number of zeros given by the zero model matches the expected number of zeros given by the Poisson distribution which models the positive counts.

The models developed by Baetschmann and Winkelmann (2013) are for cross sectional data. To model number of primary care visits in the longitudinal LINK LA data, we propose

a mixed effects hurdle model which accounts for exposure. We allow the count rates and zero probabilities for the two treatment groups to vary over the course of the study, as one might expect the effect of the intervention to wear off over time. To do this, we partition the follow-up period into discrete intervals, each interval having its own zero and count rates. For the LINK LA data, we define our intervals based on the planned follow-up time periods: the 12 month period pre-incarceration (baseline) and 0-3 months, 3-6 months, 6-12 months and beyond 12 months post incarceration.

One complication of the LINK LA data set is that follow-up times vary greatly both within and between individuals, and did not resemble the original study design. For example, the first follow-up, which was supposed to occur at 3 months after release from jail, was observed at anywhere from 3 to 12 months after release. Histograms of the observed times since last follow-up at each of three follow-ups are plotted in figure 3.2. Therefore to model this data, exposure must be accounted for. Further, we wish to allow the zero and count rates to vary over the follow up period, as we believe it is unreasonable to assume constant zero and count rates over these time periods. Observations commonly span more than one follow-up time interval. We model observed counts using weighted averages of the parameters corresponding to the time intervals over which the count was observed.

We model within individual correlation over time using random effects in both the zero and count models. We use subject level random intercepts in the zero model and subject level random effect vectors in the count model with entries corresponding to the partitioned intervals of the baseline and follow-up period. The random effects vector in the count model allows us to consider several models for the correlation matrix within an individual. For example, we would expect correlation between observations taken over 0-3 months to be more closely correlated with 3-6 month observations than with the observations taken in the 12 months prior to incarceration, as in many cases incarceration lasted multiple years. We explore different covariance models for the subject specific random effects, in particular autoregressive, antedependence (Zimmerman & Núñez-Antón, 2005) and unstructured

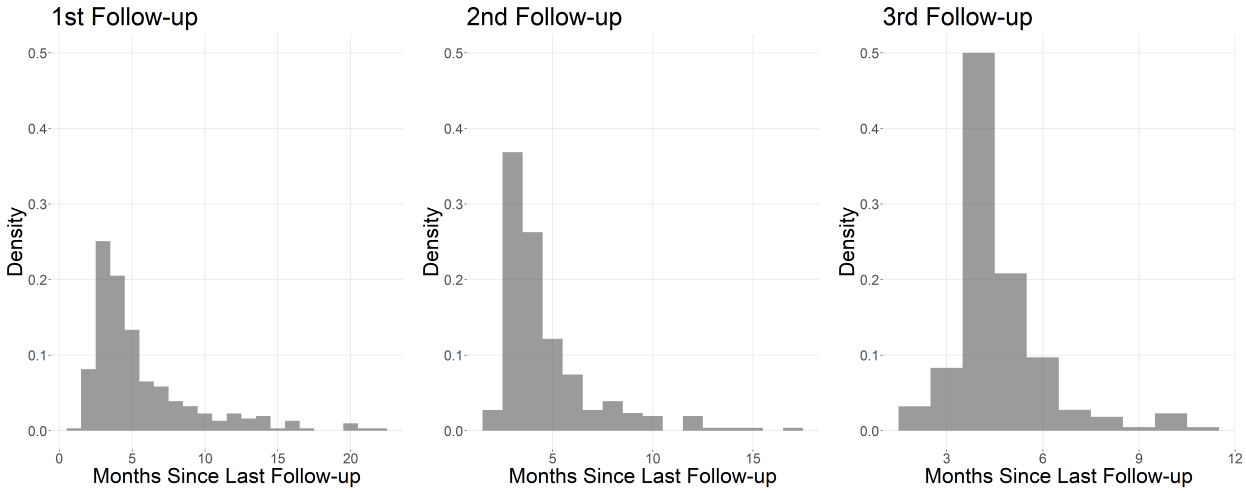


Figure 3.2: Length of time between follow-ups observed in LINK LA data at each visit. The follow-ups were intended to occur at 3 months, 6 months and 12 months after release from jail.

covariance models.

Our model adapts the exposure hurdle models of Baetschmann and Winkelmann (2013) to the repeated measures setting in which follow-up times vary. We decouple the observation times, which we treat on a continuous scale, from the time dependent fixed and random effects, which model time discretely. In addition, our models are Bayesian, allowing researchers to make use of prior information, and allow straightforward inference for a variety of targets. The variability in follow-up times leads us to treat the data as unbalanced and we assume missing observations are missing at random and do not bias inference (Little & Rubin, 2002).

In section 3.2, we present the formulation of longitudinal exposure hurdle models (LEHMs) and how they can be applied to the LINK LA data. We present 6 different covariance models for the LEHM count model random effects, ranging in complexity from random intercept to fully unstructured. We give prior distributions and detail our approach to posterior sampling. In section 3.3 we present the results of fitting the models of section 3.2 to the LINK LA data, and compare model fit and inference across the 6 different models.

3.2 Methods

Let Y_{ij} be a count response for subject $i = 1, \ldots, N$ at visit $j = 1, \ldots, J_i$, where N is the number of subjects and J_i is the number of visits for subject i. We observe that Y_{ij} has excess zeros compared to a Poisson random variable. A Poisson hurdle model is parameterized by two parameters, λ_{ij} and π_{ij} for subject i at visit j, where π_{ij} is the probability of observing a positive count, and λ_{ij} is the mean of an untruncated Poisson which we use to model the non-zero counts, such that

$$P(Y_{ij} = 0|\pi_{ij}) = 1 - \pi_{ij}, \tag{3.1}$$

$$P(Y_{ij} = y_{ij}|\pi_{ij}, \lambda_{ij}) = \pi_{ij} \frac{\lambda_{ij}^{y_{ij}} \exp(-\lambda_{ij})}{y_{ij}!} \frac{1}{1 - \exp(-\lambda_{ij})},\ y_{ij} \geq 1. \tag{3.2}$$

Probability of a zero is modeled only through the zero model, equation (3.1). This allows for either zero inflation or deflation depending on whether $1 - \pi_{ij}$ is greater than (inflation) or less than (deflation) the probability of a zero coming from the Poisson(λ_{ij}) distribution. Probabilities of positive counts in (3.2) are the probability π_{ij} that an observation is greater than 0 from the zero model times the probability of the count for an untruncated Poisson(λ_{ij}) divided by the untruncated Poisson probability $1-\exp(-\lambda_{ij})$ of an observation being greater than .

Define t_{ij} as the months since baseline of visit j for subject i, and Δt_{ij} is the total time over which Y_{ij} was observed, which we call exposure. Then exposure, Δt_{ij}, is given by $\Delta t_{ij} = t_{ij} - t_{ij-1}$ for $j \geq 2$, and $\Delta t_{i1} = 12$. Let λ_{0ij} be the untruncated Poisson rate of doctor visits for one month, and let π_{0ij} be the probability of a subject having at least one doctor visit in a single month over the time interval (t_{ij-1}, t_{ij}). To model the exposure hurdle model parameters λ_{ij} and π_{ij}, we model per month parameters π_{0ij} and λ_{0ij} with mixed effects generalized linear models (GLM) and add a $\log(\Delta t_{ij})$ offset in the linear predictors to appropriately scale to a given exposure, Δt_{ij}. Regression parameters for λ_{0ij} can then be

understood as log per month rate ratios and regression parameters for π_{0ij} can be understood as the log hazard rate ratios where the hazard is the probability of not attending any doctor visit in an individual month.

Link Functions We model λ_{ij} and π_{ij} using mixed effect regression models. For λ_{ij}, we use a log link function. In a hurdle model, the log link is sometimes referred to as a "pseudo link function" as it is used to model the mean of the untruncated Poisson distribution. For the zero observations, we model π_{ij} using the complementary log log link function, $\text{cloglog}(\pi_{ij}) = \log(-\log(1-\pi_{ij}))$, as opposed to the more commonly used logit link function, $\text{logit}(\pi_{ij}) = \log(\pi_{ij}/(1-\pi_{ij}))$. The choice of cloglog and log functions is important as they both allow us to account for exposure.

Poisson Exposure When we observe a Poisson count with mean λ_{0ij} per month over a period of Δt_{ij} months. The total count over the interval is Poisson distributed with mean $\lambda_{ij} = \Delta t_{ij}\lambda_{0ij}$. We model the per month Poisson rate λ_{0ij} as $\log(\lambda_{0ij}) = \eta_{02ij}$, where η_{02ij} is a linear combination of covariates and coefficients. Then the regression equation for λ_{ij} is

$$\log(\lambda_{ij}) = \log(\Delta t_{ij}\lambda_{0ij}) = \log(\Delta t_{ij}) + \log(\lambda_{0ij}) = \log(\Delta t_{ij}) + \eta_{02ij}. \tag{3.3}$$

Zero Model Exposure Time Let the probability of attending at least one doctor visit in a given month be π_{0ij}. We model π_{0ij} with a complementary log log regression model

$$\text{cloglog}(\pi_{0ij}) = \log(-\log(1-\pi_{0ij})) = \eta_{01ij} \tag{3.4}$$

where, as with η_{02ij}, the parameter η_{01ij} is a linear combination of covariates and coefficients. These covariates may or may not be the same as those used in η_{02ij}. For the moment, assume the monthly probability of a positive count, π_{0ij}, to be constant over an interval of length Δt_{ij}. The probability of not attending any doctor visits over the entire interval is $(1-\pi_{0ij})^{\Delta t_{ij}}$,

and its compliment, the probability of attending at least one doctor visit over the interval is

$$\pi_{ij} = 1 - (1 - \pi_{0ij})^{\Delta t_{ij}}. \tag{3.5}$$

Combining equations (3.4) and (3.5), the regression model for π_{ij}, the probability of attending at least one doctor visit over a period of Δt_{ij} months, is a function of η_{01ij} and Δt_{ij},

$$\begin{aligned}
\text{cloglog}(\pi_{ij}) &= \log(-\log(1 - \pi_{ij})) && (3.6)\\
&= \log(-\log((1 - \pi_{0ij})^{\Delta t_{ij}})) && (3.7)\\
&= \log(-\log(1 - \pi_{0ij})\Delta t_{ij}) && (3.8)\\
&= \log(-\log(1 - \pi_{0ij})) + \log(\Delta t_{ij}) && (3.9)\\
&= \eta_{01ij} + \log(\Delta t_{ij}). && (3.10)
\end{aligned}$$

Equation (3.7) is given by the equality (3.5). Equations (3.6) through (3.10) show that, similar to the Poisson regression model, the cloglog regression model can account for exposure through the addition of a $\log(\Delta t_{ij})$ offset. The cloglog regression model with exposure is also known as a Weibull survival model with constant hazard. Indeed, the zero model can be considered a survival model of right censored time to event data over Δt_{ij} months, where the event is a subject attending their first doctors visit.

3.2.1 Regression Model Parameterization

Define zero model fixed time effect coefficient vectors, $\boldsymbol{\alpha}_1^0$ for the control group and $\boldsymbol{\alpha}_1^1$ for the treatment group, and count model fixed time effects $\boldsymbol{\alpha}_2^0$ for the control group and $\boldsymbol{\alpha}_2^1$ for the treatment group. These vectors can be of different lengths, however for simplicity we assume both to be of length L, corresponding to the number of time intervals of the study period having different zero and count means. In the LINK LA analysis, we take $L = 5$ where the first element $l = 1$ corresponds to the 12 month period prior to incarceration

and elements $l = 2, \ldots, 5$ correspond to 0-3 months, 3-6 months, 6-12 months and beyond 12 months partitions after release from jail. Define $J_i \times L$ longitudinal design matrix $\boldsymbol{Z}_i$ with rows $\boldsymbol{Z}_{ij} = (Z_{ij1}, \ldots, Z_{ijL})'$ whose lth element Z_{ijl} is the number of months from lth interval of the partition of the study time included in the observed time period (t_{ij-1}, t_{ij}). For example, $\boldsymbol{Z}_{i1} = (12, 0, 0, 0, 0)$ for $i = 1, \ldots, N$ since the baseline observation always included 12 months prior to incarceration, and the exposure is given by $\Delta t_{i1} = \sum_{l=1}^{L} Z_{i1l} = 12$. Alternatively, if the first follow-up visit for subject i includes the first 4 months post baseline, $\boldsymbol{Z}_{i2} = (0, 3, 1, 0, 0)$ and $\Delta t_{i2} = \sum_{l=1}^{L} Z_{ijl} = 4$. We take weighted averages of the fixed effects corresponding to an individual observation as $\Delta t_{ij}^{-1} \boldsymbol{Z}_{ij}' \boldsymbol{\alpha}_0^{c_i}$ for the zero model and $\Delta t_{ij}^{-1} \boldsymbol{Z}_{ij}' \boldsymbol{\alpha}_1^{c_i}$ for the count model, where $c_i = 0, 1$ (control, intervention) is the treatment group for subject i.

We also define random effects scalar γ_{1i} and L-vector $\boldsymbol{\gamma}_{2i}$ for the zero and count models, respectively, for subject i. We allow each individual to have a different count model random effect associated with each of the L time intervals. This allows flexibility to model the correlation within an individual over the course of the study as well as allowing us to account for possible overdispersion. For the zero model, we restrict γ_{1i} to be a scalar as there is significantly less information contained in the binary data, although it would be straightforward to extend these models to mirror the correlation modeling in the count model. We can account for an individual's correlation over time as well as correlation between the zero and count models through specification of the random effects distributions.

Define the following regression equations for exposure varying longitudinal hurdle model parameters π_{ij} and λ_{ij}

$$\text{cloglog}(\pi_{ij}) = \log(-\log(1 - \pi_{ij})) = \Delta t_{ij}^{-1} \boldsymbol{Z}_{ij}' \boldsymbol{\alpha}_1^{(c_i)} + \gamma_{1i} + \log(\Delta t_{ij}) \tag{3.11}$$

$$\log(\lambda_{ij}) = \Delta t_{ij}^{-1} \boldsymbol{Z}_{ij}' \boldsymbol{\alpha}_2^{(c_i)} + \Delta t_{ij}^{-1} \boldsymbol{Z}_{\boldsymbol{ij}}' \boldsymbol{\gamma}_{2i} + \psi \gamma_{1i} + \log(\Delta t_{ij}). \tag{3.12}$$

where ψ is used to model within subject correlation between the two parts of the hurdle

model. We use a weighted average of count model random effects, $\Delta t_{ij}^{-1} \boldsymbol{Z}_{\boldsymbol{ij}}' \boldsymbol{\gamma}_{2i}$, where the contribution of the random effect for interval l is weighted by the number of months from interval l the count y_{ij} was observed over.

The inclusion of random effects allows three main benefits. The first is that they help in modeling the longitudinal aspect of the study, by allowing us to model correlation within an individual over time, as we discuss in more detail in section 3.2.2. The second benefit is that they provide a convenient means through which to correlate the two parts of the hurdle model, as we would expect some correlation between an individual's zero probabilities and count rates. Lastly, the random effects allow us more flexibility to model overdispersion. This is a common complication in Poisson modeling, as observed variation in count data is often larger than a Poisson model allows for. Failure to account for overdispersion in Poisson models can bias interval estimates and hypothesis tests (Breslow, 1990).

3.2.2 Correlation Models for Random Effects

We develop multivariate random effects for the count model such that each individual has L separate count model random effects, one for each time interval. We model $\boldsymbol{\gamma}_{2i}$ as normal

$$\boldsymbol{\gamma}_{2i} | \boldsymbol{\Sigma}_2 \sim N_L(\mathbf{0}, \boldsymbol{\Sigma}_2), \tag{3.13}$$

with covariance matrix $\boldsymbol{\Sigma}_2$. Covariance matrix $\boldsymbol{\Sigma}_2$ models the within-individual correlation and variance across time in the count model. We consider several models for $\boldsymbol{\Sigma}_2$, which reflect different assumptions about the data and differ in the number of unknown parameters. We present these models in order from simplest, requiring the least number of parameters, to most complex, which is the unstructured model requiring estimation of each entry of the covariance matrix.

It is more straightforward to apply covariance models if one decomposes the count model

random effects covariance matrix $\boldsymbol{\Sigma}_2$ as

$$\boldsymbol{\Sigma}_2 = \text{diag}(\boldsymbol{\sigma}_2)\ \boldsymbol{\Omega}\ \text{diag}(\boldsymbol{\sigma}_2), \tag{3.14}$$

where $\text{diag}(\boldsymbol{\sigma}_2)$ is a diagonal matrix with diagonal elements given by $\boldsymbol{\sigma}_2$, an L-vector of standard deviations. Within individual across time correlation is modeled with $L \times L$ correlation matrix $\boldsymbol{\Omega}$. Thus, modeling choices for $\boldsymbol{\sigma}_2$ reflect assumptions about the variation in the data set, and affect the model's ability to handle excess variation, and modeling choices for $\boldsymbol{\Omega}$ reflect assumptions about within-individual correlation.

We consider two possible models for the count model random effects variance $\boldsymbol{\sigma}_2$, homoskedastic and heteroskedastic. The homoskedastic or equal variance model assumes constant variance in count model random effects throughout the study, replacing L-vector $\boldsymbol{\sigma}_2$ with scalar $\sigma_{2\text{cv}}$. The heteroskedastic or unequal variance model allows the variation to be different at each time interval, requiring estimation of each of the L variance parameters of $\boldsymbol{\sigma}_2$. We consider three different correlation models for the LINK LA data, autoregressive (AR), antedependent (AD), and unstructured.

3.2.2.1 Autoregressive Model

The autoregressive (AR) model is a one parameter correlation model assuming a constant correlation, ρ, between any two adjacent time periods. For the LINK LA data, with $L = 5$, the correlation matrix of $\boldsymbol{\gamma_{2i}}$ is

$$\text{Corr}(\boldsymbol{\gamma_{2i}}) = \boldsymbol{\Omega}_{\text{AR}}(\rho) = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}. \tag{3.15}$$

The autoregressive model requires estimation of only the scalar ρ to estimate the entire correlation matrix, regardless of the number of time intervals. The AR correlation model is best used for longitudinal data where observations are taken equally spaced over time, but can be too strong of an assumption to make in some situations.

For example, in the LINK LA model, the time intervals to which the count model random effects apply are not spaced equally in time, so it may be unreasonable to assume constant correlation between each adjacent pair of random effects. The LINK LA baseline random effects are associated with the year before incarcerations, and in many cases were in jail for years before release and subsequent follow-up measures. One would expect baseline random effects to have a relatively low correlation with follow-up random effects because of this large amount of time between measurements. Meanwhile one might expect relatively high correlation between the 0-3 month random effects and the 3-6 month random effects due to their proximity in time. Thus, we also present another alternative correlation model, which may be better suited for the LINK LA data, the antedependent model.

3.2.2.2 Antedependent Model

The antedependent (AD) model has $L-1$ distinct correlations $\rho_1, \ldots, \rho_{L-1}$ between the l and $l+1$ time intervals. AD may be a better candidate correlation model compared to the AR model for the count model random effects in the LINK LA hurdle models due to the different spacing between time intervals. For the LINK LA setting with $L = 5$ visits, the AD correlation matrix for $\boldsymbol{\gamma_{2i}}$ is

$$\text{Corr}(\boldsymbol{\gamma_{2i}}) = \Omega_{AD}(\boldsymbol{\rho}) = \begin{pmatrix} 1 & \rho_1 & \rho_1\rho_2 & \rho_1\rho_2\rho_3 & \rho_1\rho_2\rho_3\rho_4 \\ \rho_1 & 1 & \rho_2 & \rho_2\rho_3 & \rho_2\rho_3\rho_4 \\ \rho_1\rho_2 & \rho_2 & 1 & \rho_3 & \rho_3\rho_4 \\ \rho_1\rho_2\rho_3 & \rho_2\rho_3 & \rho_3 & 1 & \rho_4 \\ \rho_1\rho_2\rho_3\rho_4 & \rho_2\rho_3\rho_4 & \rho_3\rho_4 & \rho_4 & 1 \end{pmatrix} \tag{3.16}$$

where the lag 1 correlations are given by the elements of $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{L-1})$. Thus the AD model requires estimation of $L-1$ parameters for an $L \times L$ correlation matrix. We consider AR and AD correlation models for the LINK LA data, each with both homoskedastic and heteroskedastic variance.

3.2.2.3 Unstructured Model

If one does not wish to make assumptions about the correlation model or variance of the count model random effects, then one can take the covariance matrix to be unstructured (UN). This is the most general covariance model, requiring estimation of each element of the covariance matrix, which requires estimating $\frac{L(L+1)}{2}$ parameters. However, the unstructured model offers a benefit in Bayesian modeling as one can use an Inverse-Wishart conjugate prior to the multivariate normal distribution, allowing modelers to sample conditional posterior draws directly from an Inverse-Wishart distribution. One concern with this approach, however, is that correlation estimates and variance estimates have an interdependency and thus one cannot use priors to restrict the correlations independent of the variances. To address this problem, we use a scaled Inverse-Wishart prior (O'Malley & Zaslavsky, 2008) on the covariance matrix $\boldsymbol{\Sigma}_2$, which introduces a set of scale parameters which are modeled separately and scale the covariance up or down. The scaled Inverse-Wishart distribution is defined by decomposing $\boldsymbol{\Sigma}_2 = \text{diag}(\boldsymbol{\omega})\boldsymbol{\Sigma}_\omega \text{diag}(\boldsymbol{\omega})$ where $\boldsymbol{\Sigma}_\omega$ is a positive definite matrix with an Inverse-Wishart prior with degrees of freedom ν and $L \times L$ scale matrix $\boldsymbol{S}$ and $\text{diag}(\boldsymbol{\omega})$ is a diagonal matrix with diagonal entries $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_L)$. The full conditional posterior distribution of $\boldsymbol{\Sigma}_\omega$ is an Inverse-Wishart distribution, which we describe in more detail in the appendix.

3.2.3 Hierarchical Mean Centering

Posterior sampling as parameterized in sections 3.2.1 and 3.2.2 mixed poorly. Two alternative parameterizations improved mixing. The first reparameterization was to center the count model random effects around the main effects $\gamma_{2i}|\boldsymbol{\alpha}_2^{c_i}, \boldsymbol{\Sigma}_2 \sim N_L(\boldsymbol{\alpha}_2^{c_i}, \boldsymbol{\Sigma}_2)$. The regression equation for λ_{ij} simplifies to

$$\log(\lambda_{ij}) = \Delta t_{ij}^{-1} \boldsymbol{Z}'_{\boldsymbol{ij}} \boldsymbol{\gamma}_{\boldsymbol{2i}} + \psi\gamma_{1i} + \log(\Delta t_{ij}). \tag{3.17}$$

With this parameterization, with $\boldsymbol{\alpha}_2^c$, $c = 0, 1$, having multivariate normal priors, the full conditional posterior distribution of $\boldsymbol{\alpha}_2^c$ is multivariate normal as well, and we can draw $\boldsymbol{\alpha}_2^0$ and $\boldsymbol{\alpha}_2^1$ directly from the full conditional posterior distributions. More details are given in the appendix.

3.2.3.1 Variance as Coefficient

For the AR and AD models, we employed one further reparameterization. We include the lower triangular Cholesky decomposition of the count model random effects covariance matrix as coefficients in the regression equation and therefore can restrict the prior covariance of the random effects vector to be the $L \times L$ identity matrix $\mathbf{I}_L$. The random effects in these models have long tailed posterior distributions, and parameterizing the variance in this way helped with mixing. Under the new parameterization,

$$\text{cloglog}(\pi_{ij}) = \log(-\log(1 - \pi_{ij})) = \Delta t_{ij}^{-1} \boldsymbol{Z}'_{\boldsymbol{ij}} \boldsymbol{\alpha}_{\mathbf{1}}^{(\mathbf{c_i})} + \sigma_1\gamma_{1i} + \log(\Delta t_{ij}) \tag{3.18}$$

$$\log(\lambda_{ij}) = \Delta t_{ij}^{-1} \boldsymbol{Z}'_{\boldsymbol{ij}} \boldsymbol{\Sigma}_2^{1/2} \boldsymbol{\gamma}_{\boldsymbol{2i}} + \psi\gamma_{1i} + \log(\Delta t_{ij}). \tag{3.19}$$

where $\boldsymbol{\Sigma}_2^{1/2}$ is the lower triangular Cholesky decomposition of $\boldsymbol{\Sigma}_2$ such that $\boldsymbol{\Sigma}_2^{1/2}(\boldsymbol{\Sigma}_2^{1/2})^T = \boldsymbol{\Sigma}_2$. For $i = 1, \ldots, N$

$$\gamma_{1i} \sim \mathrm{N}(0, 1) \tag{3.20}$$

$$\boldsymbol{\gamma}_{2i} \sim N_L(\boldsymbol{\alpha}_2^{(c_i)}, \mathbf{I}_L). \tag{3.21}$$

The fixed effects $\boldsymbol{\alpha}_2^{(c)}$ change in interpretation from the previous parameterizations. The log untruncated per month Poisson rate for each time interval, which was given by the elements of $\boldsymbol{\alpha}_2^{(c)}$ under the previous parameterization are now given by the elements of $\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\alpha}_2^{(c)}$.

3.2.4 Prior Specification

Priors were chosen to be semi-informative to minimize the influence of prior specification on inference, but to have them help keep posterior estimates within reasonable values. For the zero model, each element of $\boldsymbol{\alpha}_1^0$ and $\boldsymbol{\alpha}_1^1$ were given independent N(0,1^2) prior distributions. We assume the first elements of $\boldsymbol{\alpha}_1^0$ and $\boldsymbol{\alpha}_1^1$ to be the same, and separately assume the first elements of $\boldsymbol{\alpha}_2^0$ and $\boldsymbol{\alpha}_2^1$ to be the same as these elements describe the period prior to incarceration, before the intervention was administered. The zero model random effects standard deviation, σ_1, was given a $\mathrm{N}^+(0,.5^2)$ distribution, where $\mathrm{N}^+(\mu, \sigma^2)$ is the normal distribution with mean μ and standard deviation σ, restricted to the positive domain. When $\mu = 0$, the $\mathrm{N}^+(0,\sigma^2)$ distribution is called the half normal distribution.

In the count model, elements of $\boldsymbol{\alpha}_2^0$ and $\boldsymbol{\alpha}_2^1$ were given independent N(-2,$.5^2$) prior distributions. This prior distribution is centered at approximately one doctor visit per year and gives a 95% confidence interval between one visit every 5 years and 3.7 visits per year. The across time correlation parameters in the AD and AR models were given uniform (-1,1) priors, and similarly the matrix, $\boldsymbol{\Sigma}_\omega$, modeling the correlations in the UN model, was given an Inverse-Wishart(6, I_5) prior, corresponding to a non-informative uniform distribution over all correlations. Random effect standard deviation parameters, σ_1, $\boldsymbol{\sigma}_2$ and $\boldsymbol{\omega}$ for the different

models, were also given independent $N^+(0,.5^2)$ distributions.

3.2.5 Posterior Computation

We approach parameter estimation under the Bayesian paradigm, sampling posterior distributions using Markov Chain Monte Carlo (MCMC) methods (Metropolis et al., 1953; Hastings, 1970; Gelfand & Smith, 1990; Casella & George, 1992). We split the parameters into separate blocks and sequentially sampled each block of parameters conditional on all other model parameters. For sampling of $\boldsymbol{\alpha}_1$, $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, a Hamiltonian Monte Carlo step (Betancourt & Girolami, 2015) was used. For the random intercept model, we also sampled $\boldsymbol{\alpha}_2^0$ and $\boldsymbol{\alpha}_2^1$ using an HMC step, however for the multivariate random effects models, we sample $\boldsymbol{\alpha}_2^0$ and $\boldsymbol{\alpha}_2^1$ directly from the multivariate normal conditional posterior distributions. Let α_{21} denote the first element shared by both $\boldsymbol{\alpha}_2^0$ and $\boldsymbol{\alpha}_2^1$, which we first sample from its posterior distribution. We then jointly sample the remaining $L-1$ elements of each vector conditional on the first element. More details are given in the appendix.

Variance parameters σ_1 and $\boldsymbol{\sigma}_2$, as well as ρ in the AR model and $\boldsymbol{\rho}$ in the AD model, were sampled using random walk Metropolis-Hastings steps. In the UN model, $\boldsymbol{\omega}$ was also sampled using a Metropolis-Hastings step, and $\boldsymbol{\Sigma}_\omega$ was drawn directly from its Inverse-Wishart conditional posterior distribution,

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}}|\nu, \boldsymbol{S}, N, \boldsymbol{\eta}, \boldsymbol{D} \sim \text{IW}(\nu + N, \boldsymbol{S} + \text{diag}(\boldsymbol{\eta}^{-1}) \left(\sum_{i=1}^{N} \boldsymbol{D}_i \boldsymbol{D}_i^{'} \right) \text{diag}(\boldsymbol{\eta}^{-1})) \tag{3.22}$$

with $\boldsymbol{D}_i = \boldsymbol{\gamma}_{2i} - \boldsymbol{\alpha}_2^{c_i}$.

In this chapter, we consider the AR, the AR with constant variance (ARcv), AD, and AD with constant variance (ADcv) models, as well as UN and RI models for the LINK LA data. We compare model fit with the Watanabe-Akaike information criterion (WAIC) (Gelman, Hwang, & Vehtari, 2014). WAIC is a fully Bayesian model fit statistic that uses the log posterior predictive density (lppd) to measure how closely a model fits the data, and

penalizes models for overfitting with an estimate p_{WAIC} of the total number of independent parameters estimated. WAIC is defined as

$$\text{WAIC} = -2(\text{lppd} - p_{\text{WAIC}}) \tag{3.23}$$

where a lower WAIC indicates a better model fit.

3.3 Results

In this section, we present results from the longitudinal exposure hurdle models (LEHMs) of section 3.2 applied to the LINK LA data. We consider 6 models: RI, AR, ARcv, AD, ADcv, and UN and compare fit and inference. Table 3.1 gives WAIC, log posterior predictive density (lppd), and estimated number of parameters (p_{WAIC}) for each of the models. Lower WAIC indicates a better model fit, suggesting ARcv as the preferred model. Random intercept and UN fit significantly worse than the parameterized correlation models. The lppd suggests that the random intercept model is not flexible enough to capture the complexity of the LINK LA data and does not fit the data. Conversely the UN model has the highest lppd, but the p_{WAIC} suggests that the UN model is overparameterized and thus, over-fitting the data.

The structured covariance models, AD, ADcv, AR, and ARcv all have very similar fits as measured by lppd, and most of the WAIC difference between these models comes from the p_{WAIC}, hence why the simplest model, ARcv, has the lowest WAIC. Thus, there is not enough improvement in fit from the more complex models to justify additional parameters.

Table 3.2 gives posterior mean estimates for the random effect correlations from each of the covariance models. These can be understood as a measure of the within individual correlation over the course of the study. The unstructured and both autoregressive models estimate the correlations between adjacent time intervals to be quite low, which are given in lag-1 diagonal elements for each model. The 95% credible intervals suggest that the

	WAIC	lppd	p_{WAIC}
ARcv	4546.9	-1634.8	638.7
ADcv	4764.3	-1638.0	744.1
AR	5079.9	-1629.0	911.0
AD	5181.6	-1631.6	959.2
RI	6189.9	-1949.0	1146.0
UN	6997.4	-1427.1	2071.6

Table 3.1: Model fit statistics comparing RI, AR, ARcv, AD, ADcv and UN for LINK LA data. A lower WAIC indicates a better fit. ARcv is shown to be the best fitting model.

correlations are low. The AD and ADcv models also do not find significant correlations but have much wider credible intervals. There is agreement between the ADcv and AD models and between the ARcv and AR models.

Standard deviation posterior estimates for both the count and zero model random effects for each model are given in Table 3.3. In the three heteroskedastic models, AR, AD, and UN, count model standard deviation posterior distributions did not vary much across time intervals, which supports the homoskedastic constant variance models. In addition, the AR, ARcv, AD, and ADcv are mostly in agreement on standard deviations, although in the non-constant variance versions some of the credible intervals associated with later time intervals are a bit wider. Both the UN and RI models, however, estimate much higher random effect standard deviations.

The top row of figure 3.3 gives the posterior mean and credible intervals for the expected number of primary care visits per month at baseline and each follow-up time interval. The second row plots the expected number of primary care visits given that at least one was attended from the count model, and the third row gives the percentage of subjects that attended at list one primary care visit from the zero model. All models agree that subjects in both groups did significantly better at accessing care compared to baseline over the next year, driven mainly by the zero model. Thus, subjects that previously were not linked to care became so after release from jail. None of the models show much difference between the treatment groups, the largest difference between treatment groups is given by the RI model

Model		0-3 months	3-6 months	6-12 months	12+ months
UN	Baseline	.05 (-.18, .28)	.06 (-.18, .29)	.05 (-.18, .27)	.01 (-.20, .22)
	0-3 months		.10 (-.10, .29)	.07 (-.14, .27)	.06 (-.21, .31)
	3-6 months			.10 (-.14, .33)	.08 (-.17, .33)
	6-12 months				.00 (-.21, .21)
AD	Baseline	.35 (-.18, .91)	.09 (-.14, .50)	.01 (-.06, .11)	.00 (-.02, .02)
	0-3 months		.24 (-.33, .9)	.03 (-.15, .27)	.00 (-.06, .04)
	3-6 months			.16 (-.29, .82)	-.01 (-.16, .13)
	6-12 months				-.12 (-.49, .35)
ADcv	Baseline	.36 (-.17, .91)	.05 (-.15, .35)	.00 (-.06, .07)	.00 (-.01, .02)
	0-3 months		.15 (-.32, .77)	.01 (-.13, .16)	.00 (-.03, .04)
	3-6 months			.09 (-.29, .71)	-.01 (-.15, .12)
	6-12 months				-.14 (-.53, .30)
AR	Baseline	.05 (-.20, .37)	.02 (.00, .14)	.00 (-.01, .05)	.00 (.00, .02)
	0-3 months		.05 (-.20, .37)	.02 (.00, .14)	.00 (-.01, .05)
	3-6 months			.05 (-.20, .37)	.02 (.00, .14)
	6-12 months				.05 (-.20, .37)
ARcv	Baseline	.05 (-.21, .33)	.02 (.00, .11)	.00 (-.01, .04)	.00 (.00, .01)
	0-3 months		.05 (-.21, .33)	.02 (.00, .11)	.00 (-.01, .04)
	3-6 months			.05 (-.21, .33)	.02 (.00, .11)
	6-12 months				.05 (-.21, .33)

Table 3.2: Count model random effects correlation matrices from RI, ARcv, AR, ADcv, AD and UN models. Posterior mean and Bayesian 95% credible intervals.

	Zero Model	Count Model				
		Baseline	0-3 Months	3-6 Months	6-12 Months	12+ Months
AR	.42 (.22, .62)	.32 (.21, .47)	.39 (.21, .63)	.31 (.13, .61)	.30 (.17, .47)	.38 (.16, .64)
AD	.41 (.22, .64)	.33 (.21, .49)	.37 (.19, .60)	.29 (.11, .64)	.30 (.16, .49)	.40 (.17, .67)
UN	.78 (.45, 1.07)	1.58 (1.28, 1.94)	1.39 (1.16, 1.66)	1.50 (1.23, 1.81)	1.25 (1.06, 1.48)	1.73 (1.41, 2.1)
ADcv	.4 (.21, .60)	.31 (.21, .44)	- -	- -	- -	- -
ARcv	.41 (.22, .60)	.32 (.21, .45)	- -	- -	- -	- -
RI	1.04 (.89, 1.20)	.92 (.80, 1.07)	- -	- -	- -	- -

Table 3.3: Standard deviation posterior means and 95% Bayesian Credible intervals for the zero model random effects and count model random effects for each model for the LINK LA data. Constant variance count models (ADcv, ARcv and RI) have one standard deviation estimate shared across the entire study, while the UN, AD, and AR have a separate count model random effects standard deviation for each time interval.

over the 0-3 month time interval, however as the RI did not fit the data well compared to the other models. The RI significant result can be attributed to model misspecification and not due to an underlying true effect being identified. The multivariate random effect models, particularly the AD, ADcv, AR and ARcv give very similar inferences.

Posterior summaries for proportion of subjects linked to care, expected number of visits for subjects that had at least one visit, and expected number of visits per person at 12 months post-baseline are given in Table 3.4 for all 6 covariance models. We provide posterior means and 95% credible intervals for both the treatment and control groups, the difference of differences (DoD), measured as the difference between the change from baseline in the intervention group and the change from baseline in the control group in the Posson rate, zero probability and hurdle model means. The zero portion of the table gives the probability of at least one visit in the first year after release from jail.

The RI model had much wider posterior distributions than the other models, with larger posterior means and variances, and it overestimates care usage in both groups. The ARcv model posterior distribution for the expected number of primary care visits has a 95% credible interval of 4.67 to 6.59 visits for the control group and slightly higher at 5.23 to 7.16 visits for the intervention treatment group. For all models, the DoD posterior mean finds the treatment group to improve slightly more in access to care compared to the control group, but none of the credible intervals find statistical significance. The zero model posterior means and intervals are similar across all models, although the ARcv, AR, ADcv and AD models estimate slightly better linkage to care than the RI and UN models, although again the difference between models is neither large nor important. The zero model DoD estimates show very little difference between treatment groups over one year.

		RI	ARcv	AR	ADcv	AD	UN
Hurdle	Ctrl	8.94 (4.63, 12.74)	5.57 (4.67, 6.59)	5.62 (4.74, 6.62)	6.43 (5.02, 8.16)	6.55 (5.03, 8.60)	5.32 (4.19, 6.71)
	Trt	9.60 (5.91, 13.36)	6.15 (5.23, 7.16)	6.18 (5.29, 7.14)	7.15 (5.62, 9.02)	7.22 (5.55, 9.46)	5.73 (4.54, 7.16)
	DoD	.65 (-3.57, 5.87)	.58 (-.36, 1.53)	.56 (-.39, 1.49)	.71 (-.39, 1.88)	.67 (-.44, 1.86)	.40 (-1.25, 2.05)
Zero	Ctrl	.85 (.80, .89)	.91 (.86, .94)	.91 (.86, .94)	.91 (.87, .94)	.91 (.86, .94)	.88 (.82, .92)
	Trt	.85 (.80, .90)	.92 (.88, .95)	.92 (.88, .95)	.92 (.88, .95)	.92 (.88, .95)	.89 (.83, .94)
	DoD	.002 (-.05, .06)	.01 (-.03, .06)	.01 (-.03, .06)	.01 (-.03, .06)	.01 (-.03, .06)	.01 (-.04, .07)
Count	Ctrl	10.50 (5.44, 14.99)	6.15 (5.24, 7.18)	6.20 (5.31, 7.23)	7.10 (5.60, 8.94)	7.23 (5.59, 9.43)	6.07 (4.84, 7.57)
	Trt	11.24 (6.91, 15.65)	6.70 (5.79, 7.70)	6.73 (5.85, 7.69)	7.78 (6.19, 9.76)	7.87 (6.10, 10.24)	6.45 (5.18, 7.97)
	DoD	.74 (-4.14, 6.82)	.55 (-.39, 1.50)	.53 (-.42, 1.45)	.68 (-.44, 1.87)	.64 (-.49, 1.85)	.38 (-1.41, 2.17)

Table 3.4: One year posterior means (95% credible intervals) number of primary care visits for treatment (Trt) and control (Ctrl) groups as well as difference of differences (DoD) at one year. Also given are treatment mean, control mean and DoDs for one year proportion of subjects that attended at least one visit (Zero) and expected number of visits for subjects that had at least one (Count). Posterior summaries are given for random intercept (RI), autoregressive (AR), autoregressive constant variance (ARcv), antedependent (AD), antedependent constant variance (ADcv) and unstructured (UN) models.

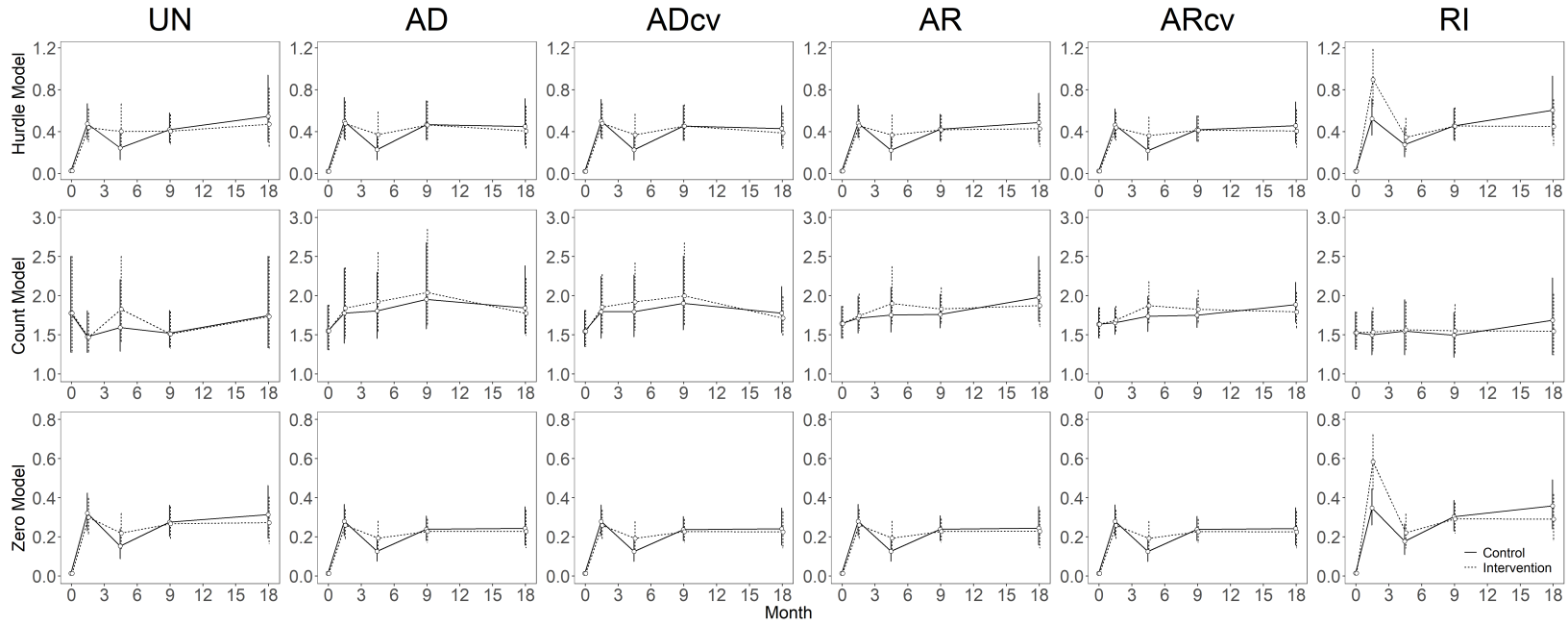


Figure 3.3: Plots of posterior means and 95% credible intervals for monthly rates for each of the planned follow-up periods for the full hurdle model and the count and zero parts of the hurdle model. Baseline is plotted at time 0, and follow-up points are placed at the midpoint of each follow-up period.

3.4 Discussion

The purpose of this chapter is to explore novel Bayesian models for use in zero inflated longitudinal data with varying exposure. These models are designed for zero-inflated outcomes in studies where individuals are followed up over time and follow-up time varies between and within individuals. We demonstrate the use of these models on the LINK LA study to gauge the effect of an intervention on the expected number of doctors visits in one year after release from jail. We also demonstrate the use of covariance models, and find that they significantly improve fit compared to the random intercept and unstructured covariance models. Our analysis shows that choice of random effect model can affect inference, as group mean posterior distributions did not agree between the random intercept model and the more complex structured covariance models.

In zero inflated longitudinal data, the random intercept model is a common choice due to its simplicity and software availability, however we find here that the flexibility of allowing an individual's random effects to vary over the follow-up can improve model fit. In doing so, researchers gain more flexibility in controlling for overdispersion, as well as within subject correlation over time.

If one uses multivariate random effects, the unstructured covariance model is a natural choice, allowing for use of a conjugate Inverse-Wishart variance prior. However, we find in the LINK LA setting that simpler, structured covariance models perform better. The ARcv, for example, which was selected to be the best model among the ones that we tried, allows specification of any size covariance matrix with just two parameters, and was able to fit the data quite well. A benefit of the covariance model framework for these longitudinal hurdle models is that it is straightforward to move between different covariance models and compare WAICs to choose the best option.

The Bayesian approach to these models allows more flexibility than frequentist alternatives. For example, in the LINK LA data set, the Bayesian models in this paper were able

to fit the data and produce inference when frequentist methods struggled to fit even the random intercept model. In addition, Bayesian methods allow researchers to incorporate prior information which can help improve inference, particularly in settings with small sample size (Ghosh et al., 2006). These small sample size properties are particularly important in hurdle models where only a fraction of the observations are used to estimate the count model, as these models are estimated only from non-zero observations.

Overall, we find incorporating Bayesian methods and covariance models into longitudinal hurdle models with exposure help improve fit and inference over alternatives. These models provide a framework for analysis of data from a variety of study designs. In longitudinal count data, zero inflation is common and follow-up times are often unbalanced, requiring use of either zero-inflated models or hurdle models which account for exposure. Further, in longitudinal data, the use of random effects provides a convenient way to account for overdispersion, within-individual correlation both over time and between zero probabilities and positive count probabilities, and, as shown with the LINK LA data, researchers can benefit from allowing these random effects to vary over time with the aid of parameterized covariance models. The models demonstrated in this chapter offer an improvement over previous methods for analysis of the LINK LA data and should be considered by researchers for longitudinal zero-inflated data settings in which exposure varies.

Appendix B

Computing

Posterior sampling for all models was performed using Markov Chain Monte Carlo (MCMC) methods. Here we describe the algorithms in detail. We found random walk Metropolis-Hastings algorithms to mix and converge slowly, and similarly, we could not produce reliable inference using Hamiltonian Monte Carlo (HMC) in Stan (Carpenter et al., 2017). Instead we opted for a hybrid approach, blocking the parameters into groups and running a combination of HMC, random walk Metropolis Hastings and Gibbs sampling steps to sample from the full conditional posterior distribution of each parameter block given the parameters in the other blocks. The zero model was parameterized in the same way across all models presented in this chapter, and consequently posterior sampling for the zero models was also done in the same way regardless of the parameterization of the count model. We took slightly different approaches to sampling the count model parameters for the RI model, for the AR and AD models, and for the UN model, as methods for the RI model had difficulty with the more complex models with more parameters.

Zero Model Sampling

Posterior sampling for the zero model takes the same form across all models as the zero model is always parameterized in the same way across all models in this chapter. The parameter ψ which connects the zero and count models is also always sampled in the same way so we include it in this section. We sample zero model main effects $\boldsymbol{\alpha}_1^0$ and $\boldsymbol{\alpha}_1^1$, and zero model random effects $\boldsymbol{\gamma_1}$ using separate HMC steps and sample zero model random effects variance σ_1 using random walk Metropolis Hastings. We found sampling each of these parameters separately made the algorithm much easier to tune without costing much in computation time. As HMC is a gradient-based algorithm, one can simply use the numerical approximation to the derivative of the log posterior distributions, but computation time can

be greatly improved by using the analytical form of the derivatives which we provide in this section. For a straightforward review of the HMC algorithm with sample code, we refer the reader to Gelman, Carlin, et al. (2014). Here we present, in detail, the steps of our sampling algorithm. For ease of notation, let V_{ij} be the indicator that observation y_{ij} is a non-zero count with $V_{ij} = 1$ when $y_{ij} > 0$ and 0 otherwise and let $\boldsymbol{V}$ denote the vector of all v_{ij}, $i = 1, \ldots, N$ and $j = 1, \ldots, J_i$. We use $f(\alpha_1^0|\cdot)$ to refer to the distribution of α_1^0 conditional on all other parameters and the vector $\boldsymbol{Y}$ of all observed counts, and use a similar notation for other parameters.

1. We sample $\boldsymbol{\alpha_1^c}$, where $c = 0, 1$ from the full conditional posterior distribution

$$\log f(\boldsymbol{\alpha_1^c}|\cdot) \propto \sum_{i=1}^{N}\sum_{j=1}^{J_i} \big\{V_{ij}\log(1-\exp(-\exp(\eta_{1ij}))) - (1-V_{ij})\exp(\eta_{1ij})\big\} - \frac{1}{2}(\boldsymbol{\alpha_1^c} - \boldsymbol{\mu}_{\boldsymbol{\alpha_1^c}})'\boldsymbol{\Sigma}_{\boldsymbol{\alpha_1^c}}^{-1}(\boldsymbol{\alpha_1^c} - \boldsymbol{\mu}_{\boldsymbol{\alpha_1^c}}), \tag{3.24}$$

where $\eta_{1ij} = \Delta t_{ij}^{-1}\boldsymbol{Z'_{ij}}\boldsymbol{\alpha_1^{c_i}} + \sigma_1\gamma_{1i} + \log(\Delta t_{ij})$. The prior mean and variance of $\boldsymbol{\alpha_1^c}$ is given by $\boldsymbol{\mu}_{\boldsymbol{\alpha_1^c}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha_1^c}}$, where in our analysis, $\boldsymbol{\mu}_{\boldsymbol{\alpha_2^0}} = \boldsymbol{\mu}_{\boldsymbol{\alpha_2^1}} = (0,0,0,0,0)'$ and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha_1^0}} = \boldsymbol{\Sigma}_{\boldsymbol{\alpha_1^1}} = \mathbf{I}_5$, the 5×5 identity matrix. To sample from this posterior distribution, we use an HMC algorithm, which requires calculation of the gradient of the log posterior distribution. The gradient of the log posterior distribution with respect to α_{1l}^c, the lth element of $\boldsymbol{\alpha_1^c}$, is given by

$$\frac{\partial}{\partial \boldsymbol{\alpha_{1l}^c}}\log f(\boldsymbol{\alpha_1^c}|\cdot) = \sum_{i=1}^{N}\mathbf{1}_{(c_i=c)}\sum_{j=1}^{J_i}\Delta t_{ij}^{-1}Z_{ijl}\left[V_{ij}\frac{\exp(\eta_{ij})}{\exp(\exp(\eta_{ij}))-1} - (1-V_{ij})\exp(\eta_{ij})\right] - (\Sigma_{\alpha_1^0}^{-1}(\boldsymbol{\alpha}_1^0 - \boldsymbol{\mu}_{\boldsymbol{\alpha_1^0}}))_l \tag{3.25}$$

2. Sampling for $\boldsymbol{\gamma_1}$, the $N \times 1$ vector of random intercepts for each of the N subjects follows a method similar to that of $\boldsymbol{\alpha_1^c}$. The log full conditional posterior distribution is given by

$$\log f(\boldsymbol{\gamma_1}|\cdot) \propto \sum_{i=1}^{N}\sum_{j=1}^{J_i} \big\{ V_{ij} \log(1 - \exp(-\exp(\eta_{1ij}))) \\ - (1 - V_{ij}) \exp(\eta_{1ij}) + y_{ij}\psi\gamma_{1i} - \log\left(\exp(\lambda_{ij}) - 1\right) \big\} - \frac{1}{2}\boldsymbol{\gamma_1'\gamma_1}, \tag{3.26}$$

and the derivative of the log-posterior distribution with respect to γ_{1i} is given by

$$\frac{\partial}{\partial\gamma_{1i}} \log f(\boldsymbol{\gamma_1}|\cdot) = \sum_{j=1}^{J_i}\Bigg[\sigma_1\left(V_{ij}\frac{\exp(\eta_{ij})}{\exp(\exp(\eta_{ij})) - 1} - (1 - V_{ij})\exp(\eta_{ij})\right) \\ + \psi(y_{ij} - \lambda_{ij}(1 - \exp(-\lambda_{ij})^{-1}))\Bigg] - \gamma_{1i}. \tag{3.27}$$

We use 3.26 and 3.27 to construct an HMC algorithm to sample $\boldsymbol{\gamma_1}$ from its full conditional posterior distribution. While it is more computationally efficient to sample $\boldsymbol{\alpha}_1^0$, $\boldsymbol{\alpha}_1^1$ and $\boldsymbol{\gamma_1}$ jointly in an HMC step, we found it much easier to tune the algorithm when we sampled $\boldsymbol{\gamma_1}$ separately, while not costing much in computation time.

3. Sample σ_1 from its full conditional posterior distribution using a Metropolis algorithm with a Gaussian proposal distribution centered around the previous state of the Markov chain with variance chosen to give a desirable acceptance rate.

4. Sample ψ, the parameter modeling the dependency between the zero and count model random effects, using a Gaussian random walk Metropolis step.

Random Intercept Model

In this section, we present the detailed steps of our algorithm for sampling from the posterior distribution of the RI model. For the RI count model, we take a similar approach as in the zero model, sampling $\boldsymbol{\alpha}_2^0$ and $\boldsymbol{\alpha}_2^1$ as a separate block with an HMC step, $\boldsymbol{\gamma_2}$ as a block with a separate HMC step, and then σ_2 individually using a random walk Metropolis-Hastings step.

1. Sample zero model parameters and ψ following the steps described in section 3.4.

2. Jointly sample $\boldsymbol{\alpha}_2^0$ and $\boldsymbol{\alpha}_2^1$ from their full conditional posterior distribution. The full conditional posterior distributions take the same form for both $\boldsymbol{\alpha}_2^0$ and $\boldsymbol{\alpha}_2^1$ so we present the distribution of $\boldsymbol{\alpha}_2^0$,

$$\log f(\boldsymbol{\alpha}_2^0|\cdot) \propto \sum_{i=1}^{N}\sum_{j=1}^{J_i}\left\{y_{ij}\frac{1}{\Delta t_{ij}}\boldsymbol{Z}_{ij}\boldsymbol{\alpha}_2^{c_i} - \log\left(\exp(\lambda_{ij})-1\right)\right\} - \frac{1}{2}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\alpha_2^0}}(\boldsymbol{\alpha}_2^0 - \boldsymbol{\mu}_{\boldsymbol{\alpha_2^0}}), \tag{3.28}$$

where $\lambda_{ij} = \exp(\Delta t_{ij}^{-1}\boldsymbol{Z'_{ij}}\boldsymbol{\alpha_2^{c_i}} + \gamma_{2i} + \psi\gamma_{1i} + \log(\Delta t_{ij}))$, and $\boldsymbol{\mu}_{\boldsymbol{\alpha_2^0}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha_2^0}}$ are the prior mean and variance of $\boldsymbol{\alpha_2^0}$. In our analysis, $\boldsymbol{\mu}_{\boldsymbol{\alpha_2^0}} = \boldsymbol{\mu}_{\boldsymbol{\alpha_2^1}} = c(-2,-2,-2,-2,-2)'$, and and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha_2^0}} = \boldsymbol{\Sigma}_{\boldsymbol{\alpha_2^1}} = \mathbf{I}_5$, the 5×5 identity matrix. This corresponds to an a priori 95% confidence interval of $(1.11, 12.00)$ visits per year with a median of 2 visits per year, which we consider to be reasonable prior bounds for the average over subjects who attend at least one doctor visit in a year. The gradient of the log posterior distribution of $\boldsymbol{\alpha}_2^0$ with respect to the lth element α_{2l}^0 of $\boldsymbol{\alpha}_2^0$ is

$$\frac{\partial}{\partial \boldsymbol{\alpha_{2l}^0}}\log f(\boldsymbol{\alpha}_2^0|\cdot) = \sum_{i=1}^{N}\sum_{j=1}^{J_i} Z_{ijl}\Delta t_{ij}^{-1}(y_{ij} - \lambda_{ij}(1-\exp(-\lambda_{ij})^{-1})) - (\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\alpha_2^0}}(\boldsymbol{\alpha}_2^0 - \boldsymbol{\mu}_{\boldsymbol{\alpha_2^0}}))_l. \tag{3.29}$$

3. As in the zero model, we sample the $N \times 1$ random intercept vector $\boldsymbol{\gamma_2}$ using a separate HMC step. The log posterior distribution is

$$\log f(\boldsymbol{\gamma_2}|\cdot) \propto \sum_{i=1}^{N} \sum_{j=1}^{J_i} \{y_{ij}\sigma_2\gamma_{2i} - \log\left(\exp(\lambda_{ij}) - 1\right)\} - \frac{1}{2}\boldsymbol{\gamma_2'}\boldsymbol{\gamma_2}, \tag{3.30}$$

with derivative,

$$\frac{\partial}{\partial \boldsymbol{\gamma_{2i}}} \log f(\boldsymbol{\gamma_2}|\cdot) = \sum_{j=1}^{J_i} \sigma_2(y_{ij} - \lambda_{ij}(1 - \exp(-\lambda_{ij})^{-1})) - \gamma_{2i}. \tag{3.31}$$

4. Sample σ_2 from full conditional posterior distribution using a random walk Metropolis algorithm with Gaussian proposal, similar to step 3.

Steps 1 through 4 are repeated until satisfactory convergence.

AR/AD Models

Sampling for AR and AD count models takes a different form than for the RI count model, as $\boldsymbol{\gamma_2}$ is now an $N \times L$ matrix of random effects. In the count model parameters, we now use hierarchical mean centering as discussed in section 3.2.3.1, allowing us to sample α_2^c directly from the full conditional posterior distribution.

1. Sample zero model parameters and ψ following the steps described in section 3.4.

2. Letting $\boldsymbol{\gamma}_{2i} \sim N_L(\boldsymbol{\alpha}_2^{(c_i)}, \mathbf{I}_L)$ as described in section 3.2.3.1, we derive the full conditional posterior distributions of $\boldsymbol{\alpha}_2^0$ and $\boldsymbol{\alpha}_2^1$, which we generalize for $\boldsymbol{\alpha}_2^c$, $c = 0, 1$. Given a normal prior with mean $\boldsymbol{\mu}_{\alpha_2}^c$ and variance $\boldsymbol{\Sigma}_{\alpha_2}^c$ the posterior distribution of α_2^c is normal

with mean $\boldsymbol{B}^c_{\alpha_2}$ and variance $\boldsymbol{\Lambda}^c_{\boldsymbol{\alpha}_2}$ given by

$$(\boldsymbol{\Lambda}^c_{\boldsymbol{\alpha}_2})^{-1} = (\boldsymbol{\Sigma}^c_{\alpha_2})^{-1} + N_c\mathbf{I}_L \tag{3.32}$$

$$\boldsymbol{B}^c_{\alpha_2} = \boldsymbol{\Lambda}^c_{\boldsymbol{\alpha}_2}\bigg((\boldsymbol{\Sigma}^c_{\alpha_2})^{-1}_{1\cdot}\boldsymbol{\mu}_{\alpha^c_2} + N_c\boldsymbol{\gamma}_{2i}\bigg), \tag{3.33}$$

where $(\boldsymbol{\Sigma}^c_{\alpha_2})^{-1}_{1\cdot}$ is the first row of $(\boldsymbol{\Sigma}^c_{\alpha_2})^{-1}$. Since the first elements of $\boldsymbol{\alpha}^0_2$ and $\boldsymbol{\alpha}^1_2$ apply to baseline before the intervention was applied, we set these elements to be the same such that $\alpha^0_{21} = \alpha^1_{21} = \alpha_{21}$. This implies $(\mu_{\alpha^0_2})_1 = (\mu_{\alpha^1_2})_1 = \mu_{\alpha_2 1}$ and $(\boldsymbol{\Sigma}_{\alpha^0_2})_{11} = (\boldsymbol{\Sigma}_{\alpha^1_2})_{11} = (\boldsymbol{\Sigma}_{\alpha_2})_{11}$. We can then draw α_{21} directly from the posterior distribution $\mathrm{N}((\boldsymbol{B}_{\alpha_2})_1, (\boldsymbol{\Lambda}_{\boldsymbol{\alpha}_2})_{11})$ where $(\boldsymbol{B}_{\alpha_2})_1$ is the shared first element of $\boldsymbol{B}^0_{\alpha_2}$ and $\boldsymbol{B}^1_{\alpha_2}$ and $(\boldsymbol{\Lambda}_{\boldsymbol{\alpha}_2})_{11}$ is the element in row 1 column 1 shared by $\boldsymbol{\Lambda}^0_{\boldsymbol{\alpha}_2}$ and $\boldsymbol{\Lambda}^1_{\boldsymbol{\alpha}_2}$,

$$(\boldsymbol{\Lambda}_{\boldsymbol{\alpha}_2})^{-1}_{11} = (\boldsymbol{\Sigma}_{\alpha_2})^{-1}_{11} + N \tag{3.34}$$

$$(\boldsymbol{B}_{\alpha_2})_1 = (\boldsymbol{\Lambda}_{\boldsymbol{\alpha}_2})_{11}\bigg((\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_2})^{-1}_{11}\mu_{\alpha_2 1} + \sum_{i=1}^{N}\gamma_{2i1}\bigg), \tag{3.35}$$

where γ_{2i1} is the first element of $\boldsymbol{\gamma_{2i}}$. We then sample the remaining elements of $\boldsymbol{\gamma}_{2i}$, denoted $\boldsymbol{\gamma}_{2i2:L}$, conditional on the first element, γ_{2i1}, using the multivariate normal distributions with precision and mean given in 3.32 and 3.33.

3. In the AR and AD models, similar to the RI model, we update the entire $N \times L$ parameter matrix $\boldsymbol{\gamma}_2$ in a single HMC step. We sample $\boldsymbol{\gamma}_2$ from the log posterior distribution

$$\log f(\boldsymbol{\gamma}_2|\cdot) \propto \sum_{i=1}^{N}\Bigg\{\sum_{j=1}^{J_i}\Big[y_{ij}\boldsymbol{Z}_{ij}\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\gamma_{2i}} - \log\left(\exp(\lambda_{ij}) - 1\right)\Big] - \frac{1}{2}(\boldsymbol{\gamma}_{2i} - \boldsymbol{\alpha}_2^{c_i})'(\boldsymbol{\gamma}_{2i} - \boldsymbol{\alpha}_2^{c_i})\Bigg\}, \tag{3.36}$$

with derivative

$$\frac{\partial}{\partial \boldsymbol{\gamma_{2i}}} \log f(\boldsymbol{\gamma_2}|\cdot) = \left(\sum_{j=1}^{J_i} (y_{ij} - \lambda_{ij}(1 - \exp(-\lambda_{ij})^{-1})) \boldsymbol{Z}_{ij} \boldsymbol{\Sigma}_2^{1/2} \right)' - (\boldsymbol{\gamma}_{2i} - \boldsymbol{\alpha}_2^{c_i}). \tag{3.37}$$

4. Sample correlation parameters using a random walk Metropolis algorithm with Gaussian proposal distribution. In the case of the AR model, the correlation parameter is the scalar ρ_{AR}, and in the AD models correlation is modeled by the $L-1$ parameter vector $\boldsymbol{\rho}_{AD}$. Once a posterior draw of the correlation parameter has been accepted, we construct the correlation matrix Ω using the appropriate model (AR or AD).

5. Variance parameters, which consist of either the scalar σ_2 in the constant variance models, or the L-vector $\boldsymbol{\sigma}_2$ in the non-constant variance models, are sampled using Gaussian random walk Metropolis algorithms.

UN Model

For the UN model, we model the count model random effects covariance through the prior distributions of $\boldsymbol{\gamma}_{2i}$, $i = 1, \ldots, N$ such that $\boldsymbol{\gamma}_{2i} \sim N_L(\boldsymbol{\alpha}_2^{c_i}, \boldsymbol{\Sigma}_2)$. We then set a scaled Inverse-Wishart prior on $\boldsymbol{\Sigma}_2 = \text{diag}(\boldsymbol{\omega})\boldsymbol{\Sigma}_\omega \text{diag}(\boldsymbol{\omega})$ (O'Malley & Zaslavsky, 2008), which allows us to sample $\boldsymbol{\Sigma}_\omega$ directly from its conjugate full conditional posterior distribution.

1. Sample zero model parameters and ψ following the steps described in section 3.4.

2. Sample count model main effects $\boldsymbol{\alpha}_2^0$ and $\boldsymbol{\alpha}_2^1$ directly from their full conditional posterior distributions, which take the same form for $\boldsymbol{\alpha}_2^0$ and $\boldsymbol{\alpha}_2^1$. Here we give the full conditional posterior distribution for $\boldsymbol{\alpha}_2^c$. Given a normal prior with mean $\boldsymbol{\mu}_{\alpha_2}^c$ and variance $\boldsymbol{\Sigma}_{\alpha_2}^c$, such that $\boldsymbol{\alpha}_2^c \sim \text{N}(\boldsymbol{\mu}_{\alpha_2}^c, \boldsymbol{\Sigma}_{\alpha_2}^c)$, the posterior distribution of α_2^c is normal with mean $\boldsymbol{B}_{\alpha_2}^c$

and variance $\boldsymbol{\Lambda}^c_{\boldsymbol{\alpha}_2}$, $\mathrm{N}(\boldsymbol{B}^0_{\alpha_2}, \boldsymbol{\Lambda}^c_{\boldsymbol{\alpha}_2})$, with

$$(\boldsymbol{\Lambda}^c_{\boldsymbol{\alpha}_2})^{-1} = (\boldsymbol{\Sigma}^c_{\alpha_2})^{-1} + N_c \boldsymbol{\Sigma}_2^{-1} \tag{3.38}$$

$$\boldsymbol{B}_{\alpha_2} = \boldsymbol{\Lambda}^c_{\boldsymbol{\alpha}_2}\left((\boldsymbol{\Sigma}^c_{\alpha_2})^{-1}\boldsymbol{\mu}_{\alpha 21} + \boldsymbol{\Sigma}_2^{-1}\sum_{i:\{c_i=c\}}\boldsymbol{\gamma}_{2i}\right). \tag{3.39}$$

where N_c is the number of subjects in treatment group $c = 0, 1$. The distribution of $\alpha^0_{21} = \alpha^1_{21} = \alpha_{21}$ is also normally distributed, $\mathrm{N}((\boldsymbol{B}_{\alpha_2})_1, (\boldsymbol{\Lambda}_{\boldsymbol{\alpha}_2})_{11})$, where

$$(\boldsymbol{\Lambda}_{\boldsymbol{\alpha}_2})^{-1}_{11} = (\Sigma_{\alpha_2})^{-1}_{11} + N(\boldsymbol{\Sigma}_2)_{11} \tag{3.40}$$

$$\boldsymbol{B}_{\alpha_2 1} = (\boldsymbol{\Lambda}_{\boldsymbol{\alpha}_2})_{11}\left((\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_2})^{-1}_{11}\mu_{\alpha_2 1} + (\boldsymbol{\Sigma}_{211})^{-1}\sum_{i=1}^{N}\gamma_{2i1}\right). \tag{3.41}$$

3. Sample γ_2 using an HMC step with log-posterior given by

$$\log f(\boldsymbol{\gamma}_2|\cdot) \propto \sum_{i=1}^{N}\Bigg\{\sum_{j=1}^{J_i}[y_{ij}\boldsymbol{Z}_{ij}\boldsymbol{\gamma_{2i}} - \log\left(\exp(\lambda_{ij}) - 1\right)] - \frac{1}{2}(\boldsymbol{\gamma}_{2i} - \boldsymbol{\alpha}_2^{c_i})'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\gamma}_{2i} - \boldsymbol{\alpha}_2^{c_i})\Bigg\}, \tag{3.42}$$

with derivative

$$\frac{\partial}{\partial \boldsymbol{\gamma_{2i}}}\log f(\boldsymbol{\gamma_2}|\cdot) = \left(\sum_{j=1}^{J_i}(y_{ij} - \lambda_{ij}(1 - \exp(-\lambda_{ij})^{-1}))\boldsymbol{Z}_{ij}\right)' - \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\gamma}_{2i} - \boldsymbol{\alpha}_2^{c_i}). \tag{3.43}$$

4. The scaled Inverse-Wishart distribution separates the covariance matrix into two parts, $\boldsymbol{\Sigma}_2 = \mathrm{diag}(\boldsymbol{\omega})\boldsymbol{\Sigma}_\omega\mathrm{diag}(\boldsymbol{\omega})$. Therefore, rather than cleanly separating the variance and correlation as in the AR and AD models, we sample the matrix $\boldsymbol{\Sigma}_\omega$, which models the

correlations and some amount of the variance, and then separately sample the vector $\boldsymbol{\omega}$ which allows the total variance to scale up or down independent of the correlations. We assume prior distribution $\boldsymbol{\Sigma}_\omega \sim$ I-W(ν, S) with degrees of freedom $\nu = 6$ and scale parameter $S = I_6$. Then the posterior distribution of $\boldsymbol{\Sigma}_\omega|\boldsymbol{\gamma}_2, \alpha_2^0, \alpha_2^1, \nu, S \sim$ I-W$(\hat{\nu}, \hat{S})$ where

$$\hat{\nu} = \nu + N \tag{3.44}$$

$$\hat{S} = \text{diag}(\boldsymbol{\omega})^{-1} \left(\sum_{i=1}^{N} (\boldsymbol{\gamma}_{2i} - \boldsymbol{\alpha}_2^{c_i})(\boldsymbol{\gamma}_{2i} - \boldsymbol{\alpha}_2^{c_i})' \right) \text{diag}(\boldsymbol{\omega})^{-1}. \tag{3.45}$$

We use this result to sample $\boldsymbol{\Sigma}_\omega$ directly from it's full conditional posterior distribution.

5. Sample $\boldsymbol{\omega}$ using a Gaussian random walk Metropolis algorithm.

For each of the models, we ran the MCMC algorithms with 4 separate chains for 50,000 iterations each, discarding the first 10,000 iterations.

Model Convergence

We provide trace plots of the baseline main effects for both the zero and count models as well as random effects variances in both zero and count models for RI, AR, ARcv, AD, ADcv and UN models. For the heterskedastic models, AR, AD and UN, the variance for the baseline random effect was used. Models were run for 4 chains with 50,000 iterations, discarding the first 10,000 iterations and keeping every 5th iteration after that. All models were deemed to have satisfactory convergence and mixing.

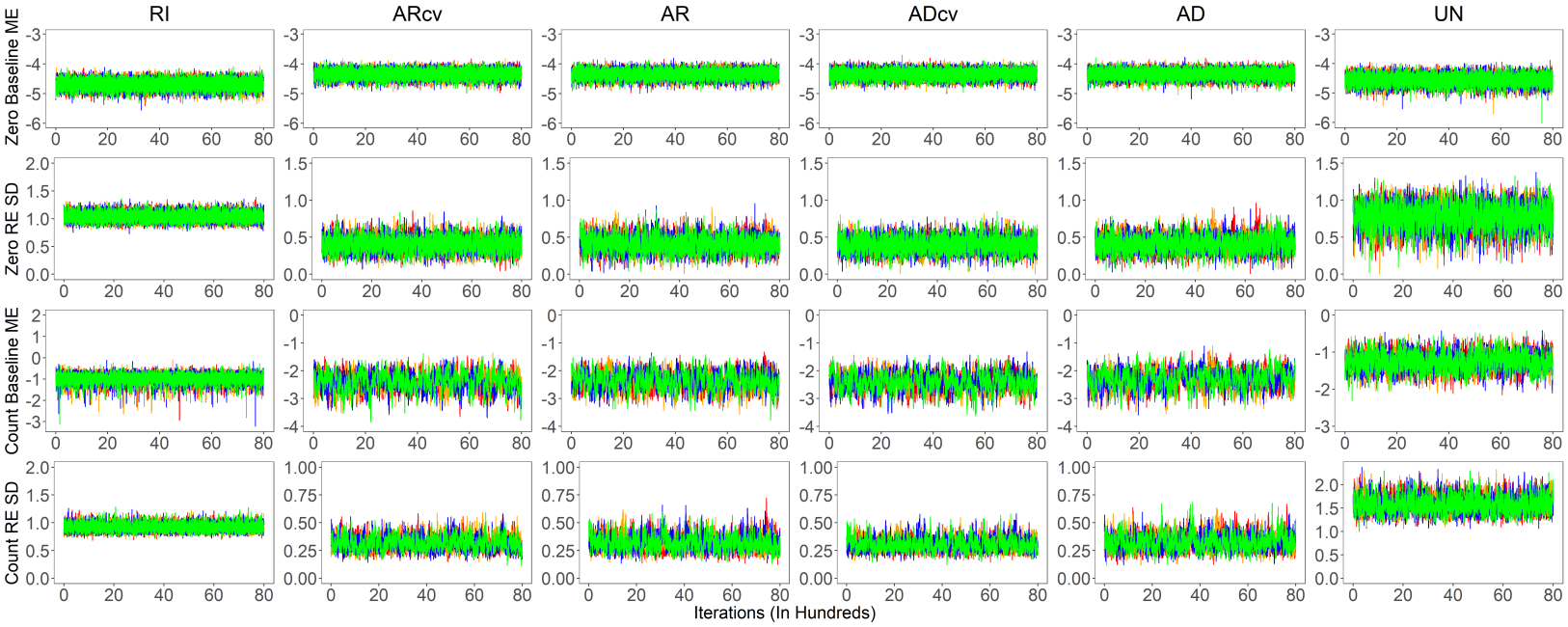


Figure 3.4: Trace plots for the 6 LINK LA models, each with 4 separate chains for 50,000 iterations each, discarding the first 10,000 iterations. The top row plots the zero model main effects associated with baseline, the second row plots the zero model random effects standard deviation, the third row plots the count model main effects associated with baseline and the fourth row plots the count model random effects standard deviation. For the heteroskedastic models, the count model random effects standard deviation is associated with the baseline count model random effects.

CHAPTER 4

A Multivariate Longitudinal Zero-Inflated Poisson Model with Varying Exposure

4.1 Introduction

Care use patterns among people living with HIV (PLHIV) are important determinants of both quality and cost of care. Beyond simply linking PLHIV to care, we want to ensure that they are receiving the right kind of care. Primary care visits have been shown to have better long term care outcomes and cost less than specialty care visits and to result in fewer emergency room visits (Hecht, Wilson, Wu, Cook, & Turner, 1999; Kitahata et al., 1999).

We model data from the Linking Inmates to Care Los Angeles (LINK LA) study, a randomized controlled trial designed to test the effectiveness of an intervention to improve linkage to care in a population of men and transgender women with HIV who have recently been released from Los Angeles county jail (Cunningham et al., 2018). Subjects were recruited to the study upon release from jail, randomized into either intervention or control group, and followed up with repeatedly to record care use since last follow-up. The intervention provided peer navigated assistance with linkage, retention and adherence to HIV care, which we compare to a standard transitional case management control. At each follow-up the number and type of medical care visits subjects had attended since the previous follow-up were recorded. For comparison, researchers also recorded the number and type of medical visits attended in the 12 months prior to incarceration. We model counts of different types of medical care visits to assess the effect of the LINK LA intervention on care use in this

vulnerable population.

The counts of care visits had a large number of zero observations, a quality known as zero-inflation. Generally, zero-inflation is seen when the data is generated by two separate processes, one determines if there is the possibility of a positive count, and if a positive count is possible, a separate process determines the magnitude of the count. For example, in the LINK LA data, subjects may or may not have had access to care during the study. Subjects that did not have access to care would thus have zero medical visits, called structural zeros. Subjects that did have access to care would attend some number of visits, which could also be zero. We call zeros from this counting process random zeros. To model the LINK LA data, we use a zero-inflated Poisson model, which models the excess zeros with a Bernoulli distribution, and the number of medical visits among those with access to care with a Poisson distribution.

Zero-inflated distributions have been well-researched (Cohen, 1966; Johnson & Kotz, 1966), and have been developed for use in regression with covariates by Mullahy (1986), Lambert (1992) and Heilbron (1994). Lambert's models use generalized linear models with a logistic regression to model the probability of an extra zero and a Poisson regression with log link to model the counts of manufacturing defects. To connect the zero and count model, Lambert proposes using a function of the Poisson rate in the logistic regression equation. Hall (2000) adapts Lambert's model to the repeated measures situation by modeling the number of bugs on plants with a zero-inflated Poisson distribution with a random intercept in the count model.

Zero-inflated models have further been developed for use in longitudinal data where the same units are followed over time (Yau & Lee, 2001; Min & Agresti, 2005). Neelon et al. (2010) develop a Bayesian mixed effect zero-inflated Poisson model for psychiatric outpatient service use which uses random intercepts in both the zero and count models to account for within-subject correlation over time. Neelon's model assumes the two random intercepts to have a bivariate normal distribution to model within-individual correlation between the zero

and count parts of the model.

Often count data, particularly in longitudinal studies, is observed over some amount of time. The time over which a count is observed is called *exposure*. For example, in the LINK LA data, subject follow-ups happened irregularly, and at each follow-up, the number of medical visits since last follow-up was measured. Some counts measured the number of medical visits attended over years and some measured the number of medical visits attended in three months. Baetschmann and Winkelmann (2013) develop a zero-inflated Poisson exposure model for cross sectional data, which uses a log exposure term in the Poisson regression and a complementary log log link function with a log exposure term in the zero model regression.

For the non-longitudinal multivariate data setting, Li et al. (1999) present a multivariate zero-inflated Poisson model, which uses a multivariate Poisson distribution for the count model which share a zero model. Liu and Tian (2015) present another possible model which reduces the number of parameters by using independent Poisson random variables which share a zero distribution. Both Li's and Liu's models lack dependency between zero models and count models and neither are developed for longitudinal data. Others such as Chib and Winkelmann (2001) and Tunaru (2002) have modeled multivariate non-zero inflated count data using random intercepts for each outcome and then giving each subject's set of random intercepts a joint multivariate normal distribution to model between outcome correlation.

We develop a Bayesian multivariate longitudinal zero-inflated Poisson exposure model (MLZIPE), which uses multivariate random effects to model within-subject correlation both between outcomes and across time. At each time point a subject has a multivariate set of random effects with one corresponding to each outcome. Correlation over time is modeled through a vector autoregressive (VAR) process (Lütkepohl, 2013). MLZIPE models allow the probability of being in the zero group to be different for different outcomes, reflecting how some subjects may only have access to some types of care. This model advances previous approaches to multivariate longitudinal zero-inflated data by allowing flexible modeling of

the three types of correlation that must be accounted for: between zero and count models, between outcomes, and across time. Section 4.2 defines the MLZIPE model and methods for analyzing the LINK LA data. Section 3 gives the results of the LINK LA data analysis.

4.2 Methods

Let Y_{ijk} be a count response for subject $i = 1, \ldots, N$ at visit $j = 1, \ldots, J_i$, on outcome $k = 1, \ldots, K$, where N is the number of subjects and J_i is the number of visits for subject i, and K is the number of different outcomes observed. For the LINK LA data, $K = 3$ as we are modeling three kinds of medical visits from the LINK LA data. Define π_{ijk} as the probability that subject i belongs to the zero group for outcome k for the time period observed at visit j. Further, define λ_{ijk} as the Poisson rate describing the count for subject i for outcome k over the time period observed at visit j given that subject i has access to medical visit k during the observed time period. Then Y_{ijk} is said to have a zero-inflated Poisson distribution if

$$Y_{ijk} \sim \begin{cases} 0 & \text{with probability } 1 - \pi_{ijk} \\ \text{Poisson}(\lambda_{ijk}) & \text{with probability } \pi_{ijk}. \end{cases} \tag{4.1}$$

The zero-inflated Poisson distribution can be understood as a mixture model with two components. One of these components is a point mass at zero, which an observation belongs to with probability π_{ijk}. The second component is a Poisson(λ_{ijk}) distribution, which an observation belongs to with probability $1 - \pi_{ijk}$. Thus, the zero-inflated Poisson probability density function is given by

$$P(Y_{ijk} = 0|\pi_{ijk}) = (1 - \pi_{ijk}) + \pi_{ijk} \exp(-\lambda_{ijk}), \tag{4.2}$$

$$P(Y_{ijk} = y_{ijk}|\pi_{ijk}, \lambda_{ijk}) = \pi_{ijk} \frac{\lambda_{ijk}^{y_{ijk}} \exp(-\lambda_{ijk})}{y_{ijk}!} \quad \text{for } y_{ij} \geq 1. \tag{4.3}$$

In a zero-inflated Poisson model, zeros can come about through two different processes. One way is through the zero process, which is the Bernoulli distribution modeling the excess zeros with probability π_{ijk}. Zeros generated in this way can be thought of as forced or structural zeros. In the LINK LA data, these would be subjects who do not have access to a particular type of care during the observed time period. The second way zeros can be generated is by chance through the counting process, or random zeros, which we model with a Poisson(λ_{ijk}) distribution. In the LINK LA study, these would be subjects who did have access to care over the observed time period, but did not use it. Equation (4.2) shows the probability of observing these two types of zeros, the structural zeros with probability $1 - \pi_{ijk}$ and the random zeros with probability $\pi_{ijk} \exp(-\lambda_{ijk})$. These models stand in contrast to the hurdle models of the previous two chapters, in which zeros can come about only through the zero process, and the count distribution is truncated at zero.

That zeros can come from either the zero model or the count model presents a difficulty in zero-inflation model estimation, as we do not know which zeros come from which process. Thus zero-inflated models have a latent class interpretation, where we define latent class Bernoulli variable $b_{ijk} \sim \text{Bernoulli}(\pi_{ijk})$ where $b_{ijk} = 1$ if a subject i has access to care type k over the time frame observed at visit j and $b_{ijk} = 0$ if they do not. Then

$$Y_{ijk}|b_{ijk}, \lambda_{ijk} \sim \begin{cases} 0 & \text{if } b_{ijk} = 0 \\ \text{Poisson}(\lambda_{ijk}) & \text{if } b_{ijk} = 1. \end{cases} \tag{4.4}$$

Thus given the latent class variable b_{ijk} the model becomes much more straightforward to fit. We use a data augmentation approach similar to Ghosh et al. (2006) and Tanner and Wong (1987) to fit the zero-inflated model. At each iteration of the MCMC algorithm, we

impute b_{ijk} such that

$$b_{ijk} = 1 \text{ if } y_{ijk} > 0 \tag{4.5}$$

$$b_{ijk} \sim \text{Bernoulli}\left(\frac{\pi_{ijk}\exp(-\lambda_{ijk})}{1-\pi_{ijk}+\pi_{ijk}\exp(-\lambda_{ijk})}\right) \text{ if } y_{ijk} = 0, \tag{4.6}$$

where the probability of success in (4.6) is the probability a zero comes from the Poisson component given a zero was observed. We then sample π_{ijk} and λ_{ijk} conditional on b_{ijk}. We model $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ using mixed effects regression equations accounting for exposure. Each subject i and each visit j has multivariate outcome vector $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijK})'$, augmente with zero model outcome vector $\boldsymbol{b}_{ij} = (b_{ij1}, \ldots, b_{ijK})'$.

4.2.1 Exposure

In LINK LA, subject's visits occurred at varying follow-up times and have varying exposure. Let t_{ij} be the time in months since release from jail for subject i at visit j. Define exposure Δt_{ij} to be total number of months over which the count $\boldsymbol{Y}_{ij}$ was observed. Baseline counts are for the 12 months prior to incarceration, thus $\Delta t_{i1} = 12$ for all subjects and for visits $j > 1$, $\Delta t_{ij} = t_{ij} - t_{ij-1}$ for $j = 2, \ldots, J_i$. We use a log link function for the Poisson model and a complementary log log (cloglog) link function for the zero model, which allows us to account for exposure by including a $\log(\Delta t_{ij})$ term in the regression equations.

Complementary log log link function with exposure Suppose that for each month over the observed interval for subject i at visit j, (t_{ij-1}, t_{ij}), the probability of being linked to care type k in one month is π_{0ijk} which we model with the complementary log log link function

$$\text{cloglog}(\pi_{0ijk}) = \log(-\log(1-\pi_{0ijk})) = \eta_{01ijk} \tag{4.7}$$

where parameter η_{01ijk} is a linear function of covariates and coefficients. The probability of being linked to care at some point over the interval (t_{ij-1}, t_{ij}) is $1-(1-\pi_{0ijk})^{\Delta t_{ij}}$. Then

$$\eta_{01ijk} + \log(\Delta t_{ij}) = \log(-\log(1-\pi_{0ijk})) + \log(\Delta t_{ij}) \tag{4.8}$$

$$= \log(-\log((1-\pi_{0ijk})^{\Delta t_{ijk}})) \tag{4.9}$$

$$= \log(-\log(1-\pi_{ijk})) = \text{cloglog}(\pi_{ijk}). \tag{4.10}$$

Thus the complementary log log regression model for π_{ijk} can be thought of as modeling the per month probability π_{0ijk} and adding a log exposure offset.

Poisson regression with link function We model the per month Poisson rate λ_{0ij} as $\log(\lambda_{0ijk}) = \eta_{02ijk}$, where η_{02ijk} is a linear function of covariates and coefficients. Then the regression equation for the expected Poisson count $\lambda_{ijk} = \Delta t_{ij}\lambda_{0ijk}$ over the entire interval (t_{ij-1}, t_{ij}), is

$$\log(\lambda_{ijk}) = \log(\Delta t_{ijk}\lambda_{0ijk}) = \log(\Delta t_{ij}) + \log(\lambda_{0ijk}) = \log(\Delta t_{ij}) + \eta_{02ijk}. \tag{4.11}$$

Similar to the complementary log log regression for the zero model, we can think of the count model for the Poisson rate over the interval (t_{ij-1}, t_{ij}) as a model for the per month Poisson rate λ_{0ijk} with an added log exposure offset to account for the length of time observed.

4.2.2 Regression Models

We allow the effect of intervention to change over time, but not necessarily linearly. To model this, we partition the study time period into L distinct intervals, each of which we allow to have different count and zero rates. For the LINK LA data, we partition the study into $L = 5$ separate intervals. These intervals are the 12 month period prior to incarceration, and 0-3 months, 3-6 months, 6-12 months and beyond 12 months after release from jail

corresponding to the planned study follow-ups at 3 months, 6 months and 12 months, as well as the baseline measuring the 12 months prior to incarceration.

Define time effects coefficient L-vector $Z_{ij} = (Z_{ij1}, \ldots, Z_{ijL})'$, where Z_{ijl} is the number of months from interval l that are included in observation $\boldsymbol{y}_{ij}$. Thus, $\boldsymbol{Z}_{i1} = (12, 0, 0, 0, 0)'$ for all baseline observations, as they always were observed over the 12 months prior to incarceration, and did not include any doctor visits from the 0-3 month, 3-6 month, 6-12 month or beyond 12 month time intervals. Suppose then subject i returns for follow-ups at $t_{i2} = 4$ months and then again at $t_{i3} = 11$ months. Then $\boldsymbol{Z}_{i2} = (0, 3, 1, 0, 0)'$ as it includes 3 months from the 0-3 month interval and one month from the 3-6 month interval. In addition, $\boldsymbol{Z}_{i3} = (0, 0, 2, 5, 0)'$ as observation $\boldsymbol{y}_{i3}$ includes only the months since last follow up, so 2 months from the 3-6 month interval and 5 months from the 6-12 month interval.

Also define L-vectors $\boldsymbol{\alpha}_{1k}^0$ and $\boldsymbol{\alpha}_{1k}^1$ as the zero model time main effects for the control and intervention groups, respectively, for outcome k. Similarly, define L-vectors $\boldsymbol{\alpha}_{2k}^0$ and $\boldsymbol{\alpha}_{2k}^1$ as the count model time main effects for the control and intervention groups, respectively, for outcome k.

We account for within individual correlation between outcomes and across time through prior distributions on random effects K-vector $\boldsymbol{\gamma}_{1i} = (\gamma_{1i1}, \ldots, \gamma_{1iK})'$ and $K \times L$ matrix $\boldsymbol{\gamma}_{2i} = (\boldsymbol{\gamma}_{2i1}, , \ldots, \boldsymbol{\gamma}_{2iK})'$ where $\boldsymbol{\gamma}_{2ik}$ is an L-vector of count model random effects for subject i and outcome k.

Letting c_i be the treatment group for subject i, where $c_i = 1$ if subject i was in the intervention group and $c_i = 0$ if subject i was in the control group, define regression models

$$\text{cloglog}(\pi_{ijk}) = \log(-\log(1 - \pi_{ijk})) = \Delta t_{ij}^{-1} \boldsymbol{Z}_{ij}' \boldsymbol{\alpha}_{1k}^{(c_i)} + \gamma_{1ik} + \log(\Delta t_{ij}) \tag{4.12}$$

$$\log(\lambda_{ijk}) = \Delta t_{ij}^{-1} \boldsymbol{Z}_{ij}' \boldsymbol{\alpha}_{2k}^{(c_i)} + \Delta t_{ij}^{-1} \boldsymbol{Z}_{ij}' \boldsymbol{\gamma}_{2ik} + \psi_k \gamma_{1ik} + \log(\Delta t_{ij}) \tag{4.13}$$

where $\Delta t_{ij}^{-1} \boldsymbol{Z}_{ij}' \boldsymbol{\alpha}_{1k}^{(c_i)}$ and $\Delta t_{ij}^{-1} \boldsymbol{Z}_{ij}' \boldsymbol{\alpha}_{1k}^{(c_i)}$ are weighted averages of the time main effects contributing to the zero and count rates for observation y_{ijk}. The parameter ψ_k models the

within-individual correlation between the zero and count model means for outcome k.

4.2.3 Random Effects Distribution

In a multivariate longitudinal zero-inflated model, there are three types of within-individual correlation that one must consider: across time, between different outcomes, and between the count and zero models. We model each of these correlations using the random effects in models (4.12) and (4.13). The random effects also help account for overdispersion in the data. In the zero model, we assign each subject i a random intercept γ_{1ik} for each outcome k. We define a joint multivariate normal distribution for subject i

$$\boldsymbol{\gamma}_{1i}|\boldsymbol{\Sigma}_1 \sim N_K(\mathbf{0}, \boldsymbol{\Sigma}_1) \tag{4.14}$$

where $\boldsymbol{\Sigma}_1$ is a $K \times K$ covariance matrix modeling the within-individual correlation between each of the K different outcomes. In the count model, we allow each individual i to have a different random effect for each time interval l for each outcome k. We model the vector of random effects $\boldsymbol{\gamma}_{2il}$ with a stationary vector autoregressive (VAR) process. This assumes the correlation between an individual's random effects at adjacent time intervals is constant over time. We also model the variance as constant across time. At time interval $l = 1$, the random effects K-vector $\boldsymbol{\gamma}_{2i1}$ for subject i is normal

$$\boldsymbol{\gamma}_{2i1}|\boldsymbol{\Sigma}_2^* \sim N_K(\mathbf{0}, \boldsymbol{\Sigma}_2^*) \tag{4.15}$$

where $K \times K$ covariance matrix $\boldsymbol{\Sigma}_2^*$ models the covariance of an individual's K random effects for the K different count outcomes at visit 1. Random effects for time intervals $l = 2, \ldots, L$ are normal conditional on the random effects at the previous time interval $l - 1$

$$\boldsymbol{\gamma}_{2i1}|\boldsymbol{\gamma}_{2il-1}, \boldsymbol{\Sigma}_2 \sim N_K(\mathbf{A}\boldsymbol{\gamma}_{2il-1}, \boldsymbol{\Sigma}_2), \tag{4.16}$$

where $\mathbf{A}$ is a diagonal matrix modeling the correlation between random effects and adjacent time intervals and $\mathbf{\Sigma}_2$ is the innovations covariance matrix. Diagonal elements of $\mathbf{A}$ are restricted to be between 0 and 1, which assumes positive within-individual correlation. Assuming the random effect variance to be constant across time implies

$$\mathbf{\Sigma}_2^* = \mathbf{A}\mathbf{\Sigma}_2^*\mathbf{A} + \mathbf{\Sigma}_2. \tag{4.17}$$

Therefore given $\mathbf{\Sigma}_2$ and $\mathbf{A}$, the elements of $\mathbf{\Sigma}_2^*$ can be solved for since

$$\Sigma_{2st}^* = \frac{\Sigma_{2st}}{1 - \mathrm{A}_{ss}\mathrm{A}_{tt}} \tag{4.18}$$

where Σ_{2st}, Σ_{2st}^*, and A_{st} are the elements in the sth row and tth column of $\mathbf{\Sigma}_2$, $\mathbf{\Sigma}_2^*$ and $\mathbf{A}$.

4.2.4 Prior Specification

Prior distributions were chosen to reflect reasonable bounds on parameters, but to minimally affect inference. Zero model time main effects $\boldsymbol{\alpha}_{1k}^0$ and $\boldsymbol{\alpha}_{1k}^1$ for $k = 1, 2, 3$ were given vague independent $\mathrm{N}(0, 2^2)$ distributions on each element. Count model time main effects $\boldsymbol{\alpha}_{1k}^0$ and $\boldsymbol{\alpha}_{1k}^1$ for $k = 1, 2, 3$ were given independent $\mathrm{N}(-2, 2^2)$ distributions on each element corresponding to a 95% confidence interval from .03 to approximately 87 doctor visits per year.

Covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ were given prior distributions proportional to the product of Inverse-Wishart distributions and half normal distributions on the diagonal elements. The normal distributions are necessary for $\mathbf{\Sigma}_2$ as Inverse-Wishart priors alone results in posterior predictive counts not having finite means due to the exponentiation from the inverse link function (Zhu & Weiss, 2013). The Inverse-Wishart distributions, I-W$(\nu_1, \mathbf{D}_1)$ and I-W$(\nu_2, \mathbf{D}_2)$, for $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ respectively, had degrees of freedom $\nu_1 = \nu_2 = 10$ and scale matrix $\mathbf{D}_1 = \mathbf{D}_2 = \mathbf{I}_3$. The half normal distributions for the diagonal elements were mean

0, variance 1 restricted to the positive real line.

Diagonal elements of $\mathbf{A}$, which model the within individual random effects correlation over time, were uniform(0,1) priors as we expect within-individual correlation over time to be positive, but if they exceed 1 then $\mathbf{\Sigma}_2^*$ will not be positive definite. Similarly, elements of the parameter $\boldsymbol{\psi} = (\psi_1, \psi_2, \psi_3)$, which models the within-individual correlation between the zero and count model, were given independent $\mathrm{N}^+(0, .25^2)$. These last two priors were chosen as we expect there to be some degree of positive correlation within an individual both across time and between model parts.

4.2.5 Posterior Computation

We draw samples from the posterior using Markov Chain Monte Carlo (MCMC) methods. Sampling was performed until satisfactory convergence was reached based on trace plots, autocorrelations, and $\hat{R}$ statistics (Gelman & Rubin, 1992). We ran 8 chains for 40,000 iterations each, keeping every draw after discarding the first 10,000 samples.

Zero model time main effects $\boldsymbol{\alpha}_{1k}^0$ and $\boldsymbol{\alpha}_{1k}^1$, zero model random effects $\boldsymbol{\gamma}_{1i}$, count model time main effects $\boldsymbol{\alpha}_{2k}^0$ and $\boldsymbol{\alpha}_{2k}^1$ and count model random effects $\boldsymbol{\gamma}_{2ij}$ were all sampled using Metropolis-Hastings steps with a Gaussian approximation to the posterior distribution as a proposal distribution. To construct these Gaussian approximations, at each posterior sample, we used a Newton-Raphson algorithm starting at the previous state of the Markov chain to locate the approximate mode of the posterior distribution, and used the second derivative of the posterior distribution at this mode estimate to estimate the covariance. This approach is described in detail by Rue and Held (2005).

Random effect covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ were also sampled using Metropolis algorithms using Inverse-Wishart proposal distributions centered at the previous state of the Markov Chain, with degrees of freedom selected to achieve a desirable acceptance rate.

The autoregressive coefficient matrix $\mathbf{A}$ modeling the across time correlation of an indi-

vidual's count model random effects, and the parameter $\boldsymbol{\psi} = (\psi_1, \psi_2, \psi_3)$ were sampled using random walk Metropolis-Hastings algorithms with Gaussian proposal distributions centered at the previous state of the Markov chain.

4.3 LINK LA Data Analysis

We fit the MLZIPE model to the LINK LA data to model care use patterns over time between primary care, specialty care and emergency care visits. The model finds no discernible difference between treatment groups, although both groups accessed primary and specialty care much more after release from jail than before. Random effect covariance estimates find positive correlation between all three types of medical visits in the count model, and no significant correlation between visit types in the zero model. For this analysis we define significance based on posterior 95% credible intervals. We would consider a treatment effect significant if the credible interval for the difference of differences excludes zero.

4.3.1 Number and Cost of Medical Visits

Table 4.1 gives posterior mean and credible intervals for the expected number of each type of medical care visits in one year. Baseline summaries are for the 12 months prior to incarceration, and control and treatment group summaries are for the first 12 months after release from jail. Difference of differences are also given to quantify treatment effect. The fourth column of the table gives expected cost of care in thousands of dollars.

Cost estimates for each type of medical visit were elicited from the 2009-2012 Medical Expenditure Panel Survey data (*Medical Expenditure Panel Survey*, 2012). Primary care visits were estimated to cost \$139, specialty care were estimated to cost \$271 and emergency care visits were estimated to cost \$852. We assumed these costs as fixed and known and used the MLZIPE model estimates for the expected number of each type of medical visit to estimate expected cost of medical care. This was done by multiplying the cost of each

	Primary Care	Specialty Care	Emergency Care	Cost (thousands)
Baseline	0.50 (0.34, 0.70)	0.45 (0.29, 0.66)	1.79 (1.38, 2.30)	1.71 (1.35, 2.12)
Ctrl	10.71 (8.17, 14.03)	3.66 (2.44, 5.45)	1.88 (1.37, 2.53)	4.08 (3.29, 5.04)
Trt	12.43 (9.41, 16.44)	3.21 (2.12, 4.86)	2.01 (1.48, 2.70)	4.32 (3.48, 5.34)
DoD	1.72 (−1.62, 5.32)	−0.44 (−1.98, 1.01)	0.14 (−0.57, 0.85)	0.24 (−0.77, 1.25)

Table 4.1: Posterior mean and credible intervals for expected number of primary care, specialty care, emergency care visits as well as expected healthcare cost over 12 months. Baseline is for the 12 months prior to incarceration. Ctrl and Trt are for 12 months after release from jail for the control and treatment groups. Difference of differences are also given. The fourth column details estimated cost of treatment in thousands of dollars based on the estimated cost of each visit type.

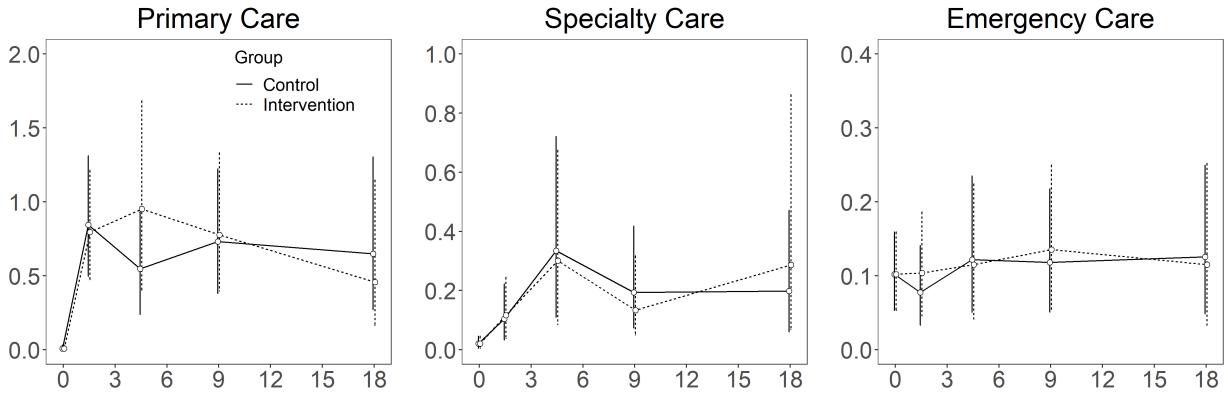


Figure 4.1: Plots of posterior means and 95% credible intervals for monthly expected number of primary care, specialty care and emergency care visits for each of the planned follow-up periods for the full hurdle model and the count and zero parts of the hurdle model. Baseline is plotted at time 0, and follow-up points are placed at the midpoint of each follow-up period.

type of visit by the expected number of visits over one year for each posterior sample to acquire posterior distributions for expected cost of care. The expected one year healthcare costs approximately doubled from baseline to follow-up. In contrast the increases in care use was approximately 20 times higher for primary care visits and 7 to 8 times higher for specialty care over follow-up compared to baseline. The increase large increase in care use compared to the increase in cost of care can be partially explained by the disproportionate cost of emergency care visits which were used with approximately the same frequency at baseline and follow-up. Thus, there is a high floor for the healthcare costs, and changes in number of primary care visits have a small effect by comparison.

Both groups dramatically increased care usage, increasing both in primary care visits and specialty care visits. The posterior mean and credible intervals for expected number of primary care visits jumped from .50 (.34, .70) visits over 12 months at baseline to 10.71 (8.17, 14.03) visits over the fist year of follow-up in the control group and 12.43 (9.41, 16.44) visits in the intervention group. Specialty care visits rose from .45 (.29, .66) expected visits at baseline to 3.66 (2.44, 5.45) visits over the first year after release from jail for the control group and 3.21 (2.12, 4.86) visits for the intervention group. The level of improvement for both groups in both visit types was the same, as the difference of difference credible intervals include zero.

The expected number of visits per month for each type of medical visit are given in figure 4.1. The monthly expected number of visits over the follow-up agree with the one year estimates from table 4.1, as both treatment groups attend visits at roughly the same rate as each other throughout the study. Emergency care use stayed relatively consistent with baseline levels throughout the follow-up, so subjects neither used more nor less emergency care in either treatment group.

4.3.2 Covariance Parameters

Posterior summaries for MLZIPE covariance parameter estimates are given in table 4.2. The zero model random effects do not have significant positive or negative correlation between any of the three types of medical visits, as all off-diagonal elements include zero in their credible intervals.

The count model, however, finds significant positive correlation between each of primary care, specialty care and emergency visits among those with access to care. A question one may be interested in is whether or not increases in primary care and/or specialty care are associated with less emergency room use. We did not find that to be the case. Rather, subjects that used more of one type of medical visit, were more likely to use more of other types of medical visits as well.

The posterior summaries for the across time correlation parameter $\boldsymbol{A}$ suggest a low to moderate within-subject correlation over time for primary care and specialty care visits. There was not strong evidence for a correlation over time for emergency care visits. Posterior summaries for the parameter $\boldsymbol{\psi}$ modeling the association between zero and count models show a moderate positive correlation between the two parts of the MLZIPE model.

4.4 Discussion

The Bayesian MLZIPE models offer an alternative to previous models for use in multivariate longitudinal zero-inflated data where exposure varies. MLZIPE models have several features that are not included in previous models. In multivariate longitudinal zero-inflated data, there are three main types of correlation that should be accounted for: across time, between outcomes, and between zero and count processes. MLZIPE models use random effects in both the zero and count models to model all three of these correlations, while also allowing for exposure time. In the analysis of the LINK LA data, we found significant positive correlations for all three types of correlation, highlighting the importance of including them in the model.

Further, the ability of MLZIPE models to account for varying exposure and observation times is an important feature for longitudinal count data. Counts are often observed over some amount of time, and it is common in health studies for subjects to return for follow-ups at irregular times, particularly in vulnerable populations, as convenience for the study participants determines when researchers are able to collect follow-up observations. In addition, many studies, such as LINK LA, are designed such that spacing between follow-ups varies, with follow-up visits often occurring more frequently closer to the start of studies and becoming less frequent later on. In such study designs, exposure must be accounted for in both parts of the zero-inflated model as exposure affects both probability of acquiring access to care and the expected number of care visits for a given observation.

Another important feature of the MLZIPE model is the ability to estimate count rates over set intervals. For example, comparing LINK LA treatment and control over the first 12 months after release from jail to the 12 months before incarceration would not have been possible using standard models, as it would have required observing exactly 12 months of medical visits. The MLZIPE estimates unobserved time effects which can be used to make inference on any desired time interval.

We also attempted to fit the hurdle versions of these models, in which the zero model models whether or not an observation was greater than zero, and the count model uses a zero-truncated Poisson to model only the positive counts. The hurdle model had difficulty fitting the LINK LA data, possibly due to the relatively large amount of missing data induced by excluding all zeros from the count model. The random effects specification used in the MLZIPE models estimates a large amount of unobserved random effects, which becomes more difficult as the number of observations used for estimation decreases. While a hurdle version of the MLZIPE models may be useful in other data settings, when the hurdle framework better describes the data generating process, or in cases of zero deflation, we found a zero-inflated framework to perform better for the multivariate LINK LA data.

The MLZIPE models also offer more flexibility than previous models to handle overdispersion, which is a common complication in count data modeling. The multivariate random effects for each individual allows for more within-individual variation than standard random intercept zero-inflated models.

Overall, we propose MLZIPE models as an important development in zero-inflated data modeling. The flexible covariance modeling framework allows modelers to better account for the complexities of multivariate longitudinal zero-inflated data.

	Primary Care	Specialty Care	Emergency Care
Zero Model			
Primary Care	.10 (.05, .22)	.003 (-.07, .10)	-.03 (-.27, .22)
Specialty Care	– –	.15 (.05, .37)	.20 (-.11, .62)
Emergency Care	– –	– –	1.22 (.53, 2.27)
Count Model			
$\boldsymbol{\Sigma}_2^*$			
Primary Care	2.68 (2.27, 3.16)	1.96 (1.55, 2.42)	.57 (.28, .87)
Specialty Care	1.96 (1.55, 2.42)	3.93 (3.18, 4.79)	1.10 (.65, 1.57)
Emergency Care	.57 (.28, .87)	1.10 (.65, 1.57)	1.39 (.99, 1.93)
$\boldsymbol{\Sigma}_2$			
Primary Care	2.54 (2.14, 3.01)	1.82 (1.44, 2.25)	.53 (.26, .82)
Specialty Care	– –	3.51 (2.81, 4.35)	1.01 (.61, 1.42)
Emergency Care	– –	– –	1.27 (.94, 1.68)
Across time correlation			
$\boldsymbol{A}$	.22 (.10, .33)	.32 (.18, .44)	.25 (.02, .52)
Zero and count model association			
$\boldsymbol{\psi}$	.73 (.18, 1.24)	.71 (.17, 1.23)	.89 (.63, 1.19)

Table 4.2: Posterior summaries for variance and covariance parameters for the zero model random effects and count model random effects. For the count model, posterior summaries for both the unconditional covariance matrix $\boldsymbol{\Sigma}_2^*$ and the innovations covariance matrix $\boldsymbol{\Sigma}_2$ are given. Also included are posterior sumamries for the autoregressive parameter $\boldsymbol{A}$ and the between model association parameter ψ. Values reported are the posterior mean and 95% Bayesian credible intervals.

Appendix C

Main Effect Posterior Summaries

Posterior means and 95% credible intervals for main effect parameter estimates are given in table 4.3. We parameterized zero model main effects $\boldsymbol{\alpha}_{1k}$ and count model main effects $\boldsymbol{\alpha}_{2k}$ for $k = 1, 2, 3$ as 9-vectors including 5 time effects for each of the 5 time intervals, and treatment by time interaction effects for each of the 4 follow-up time intervals for the treatment groups. Treatment by time parameters in the count model can be understood as log rate ratios of Poisson means. In both models a treatment $\times$ time effect credible interval excluding zero is evidence of a treatment difference over that time interval. No significant treatment effects were found for any of the medical visit types.

	Primary Care	Specialty Care	Emergency Care
Zero Model			
Baseline	−1.98 (−2.61, −0.18)	−0.24 (−1.87, 1.83)	0.29 (−1.02, 2.06)
0–3 Months	1.03 (0.00, 2.48)	0.21 (−1.19, 2.02)	0.36 (−1.12, 2.04)
3–6 Months	0.53 (−0.71, 2.03)	0.42 (−0.97, 2.07)	0.56 (−0.90, 2.23)
6–12 Months	0.40 (−0.57, 1.89)	−0.30 (−1.31, 1.19)	−0.00 (−1.35, 1.81)
12+ Months	−0.15 (−1.14, 1.41)	−0.09 (−1.36, 1.68)	0.18 (−1.26, 1.93)
Trt × 0–3 Months	0.39 (−1.10, 2.06)	0.39 (−1.26, 2.08)	0.35 (−1.30, 2.08)
Trt × 3–6 Months	0.17 (−1.40, 1.96)	0.06 (−1.61, 1.88)	0.24 (−1.51, 2.04)
Trt × 6–12 Months	0.06 (−1.45, 1.88)	−0.25 (−1.63, 1.21)	0.28 (−1.37, 2.04)
Trt × 12+ Months	−0.39 (−1.88, 1.49)	−0.53 (−2.11, 1.49)	−0.07 (−1.80, 1.86)
Count Model			
Baseline.1	−4.31 (−4.64, −4.01)	−5.29 (−5.73, −4.88)	−3.03 (−3.25, −2.81)
0–3 Months	−1.45 (−1.84, −1.07)	−3.92 (−4.55, −3.32)	−3.36 (−3.88, −2.85)
3–6 Months	−1.75 (−2.20, −1.29)	−2.86 (−3.53, −2.21)	−2.97 (−3.57, −2.38)
6–12 Months	−1.40 (−1.75, −1.06)	−3.05 (−3.57, −2.56)	−2.80 (−3.24, −2.39)
12+ Months	−1.28 (−1.78, −0.78)	−3.16 (−3.86, −2.47)	−2.82 (−3.41, −2.26)
Trt × 0–3 Months	−0.10 (−0.62, 0.42)	−0.00 (−0.81, 0.80)	0.22 (−0.45, 0.89)
Trt × 3–6 Months	0.54 (−0.07, 1.14)	−0.09 (−0.95, 0.78)	−0.09 (−0.88, 0.69)
Trt × 6–12 Months	0.08 (−0.41, 0.56)	−0.22 (−0.89, 0.46)	0.06 (−0.50, 0.62)
Trt × 12+ Months	−0.09 (−0.77, 0.59)	0.69 (−0.21, 1.60)	−0.05 (−0.85, 0.75)

Table 4.3: Zero and count model main effect parameter posterior means and 95% credible intervals for the LINK LA data analysis using the MLZIPE model. In the count model, all treatment by time interaction effects include zero in their credible intervals and thus to not find any significant treatment effects.

CHAPTER 5

Conclusions

We developed several advances for modeling longitudinal zero-inflated data. In chapter 2 we presented Bayesian longitudinal hurdle models for the number of days of heavy drinking out of the past 90 days from the SBIRT study. This model uses a multivariate random effects distribution in the count model, allowing researchers flexibility in modeling within-subject correlation over time and in modeling over-dispersion. We showed this model to be an improvement over standard random intercept longitudinal hurdle models. In chapter 3 we developed Bayesian longitudinal exposure hurdle models (LEHM) for longitudinal zero-inflated data in which exposure and observation times vary. We demonstrated the LEHM models on the LINK LA data to model number of primary care visits since last follow-up in a setting where follow-ups occurred highly irregularly. We again used the multivariate count random effects models from chapter 2, again showing them to perform better than random intercept models. In chapter 4 we developed a multivariate outcome zero-inflated model extension of the LEHMs, which we call multivariate longitudinal zero-inflated Poisson exposure models (MLZIPE). We demonstrate an MLZIPE model on number of primary care, specialty care and emergency care visits from the LINK LA study, and used it to make inference on cost of care over a given time period.

In both chapter 2 and chapter 3, we demonstrated several covariance models for the multivariate count random effects, which we compared against random intercept models. In both cases we found the multivariate random effect models to fit substantially better than random intercept models, which had over-confident and sometimes misleading inferences.

That this result was observed in both the SBIRT and LINK LA data highlights the usefulness of the multivariate random effect hurdle models for longitudinal zero-inflated data.

In both chapter 2 and chapter 3, choice of count model random effect covariance models had little impact on fit and inference. The only exception was the UN model in chapter 3, which substantially over-fit the data. Also, correlation estimates were generally very low or not significantly different from 0. Thus, the large improvements of fit seen in both chapter 2 and chapter 3 may be due to the ability of these models to model over-dispersion in the data. While the original motivation for these models was to model correlation over time, it may actually be over-dispersion that drives these models to do so well compared to the random intercept model.

Finally, the MLZIPE model in chapter 4 is an extension of the ARcv model from chapter 3 for multivariate outcomes. An important note is that we switched to a zero-inflated model for this analysis compared to the hurdle models of chapter 3. When originally setting out on the SBIRT analysis, we found hurdle models to perform better, as the data being fit by the zero and count models is fully observed and constant. This contrasts with a zero-inflated model, where we use an unobserved latent variable denoting which zeros are structural and which are random. This latent variable is re-imputed at each posterior MCMC step, and with each imputation changes which zero observations are counted as structural zeros in the zero model, and which are considered random zeros and thus, modeled by the count model. The benefit of this, particularly in settings when observed counts are small, is that a zero-inflated model might use many more observations to fit the count model. Where a hurdle count model only uses positive observations, a zero-inflated count model uses all positive observations as well as some proportion of the zeros. The number of zeros used to fit the count model, and thus the difference in number of observations used between the hurdle and zero-inflated models, is larger when observed counts are small. For example, in the chapter 4 analysis, 64% of the emergency department use observations were zeros and positive counts averaged 2.9 emergency visits per observation. The model had difficulty estimating the count

model random effects in a hurdle framework as a very small portion of the data was used.

One limitation of the proposed models is that their posterior distributions can be difficult to sample from. Posterior distributions had long tails, causing MCMC chains to get stuck often and log posterior calculations were prone to underflow issues. We attempted many sampling approaches and model parameterizations before arriving at those presented in this work. Chains required long run times to explore the posterior distribution well, we ran models for multiple days to yield the inferences we present here. Still, we find the benefits of the developed models to outweigh these drawbacks and recommend them for use in modeling longitudinal zero-inflated data.

References

Baetschmann, G., & Winkelmann, R. (2013). Modeling zero-inflated count data when exposure varies: With an application to tumor counts. *Biometrical Journal*, *55*, 679–686.

Baetschmann, G., & Winkelmann, R. (2017). A dynamic hurdle model for zero-inflation count data. *Communications in Statistics - Theory and Methods*, *46*, 7174–7187.

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., & Macias-Konstantopoulos, W. (2017). Effectiveness of SBIRT for alcohol use disorders in the emergency department: A systematic review. *Western Journal of Emergency Medicine*, *18*, 1143–1152.

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. In *Current Trends in Bayesian Methodology with Applications* (First ed.). Chapman and Hall/CRC.

Breslow, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association*, *85*, 565–571.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & 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. Retrieved from `https://journal.r-project.org/archive/2017/RJ-2017-066/index.html`

Burger, D. A., Schall, R., Ferreira, J. T., & Chen, D.-G. (2019). A robust Bayesian mixed effects approach for zero inflated and highly skewed longitudinal count data emanating from the zero inflated discrete Weibull distribution. *Statistics in Medicine*, *39*, 1275–1291.

Cameron, A. C., & Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge University Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–32.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*, 167–174.

Chib, S., & Winkelmann, R. (2001). Markov chain Monte Carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, *19*(4), 428–435.

Clark, H. W., Power, A. K., Fauve, C. E. L., & Lopez, E. I. (2008). Policy and practice implications of epidemiological surveys on co-occurring mental health and substance use disorders. *Journal of Subtsance Abuse Treatment*, *34*, 3–13.

Cohen, A. C. (1966). A note on certain discrete mixed distributions. *Biometrics*, *22*, 566–572.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, *39*, 829–844.

Cunningham, W. E., Weiss, R. E., Nakazono, T., Malek, M. A., Shoptaw, S. J., Ettner, S. L., & Harawa, N. T. (2018). Effectiveness of a peer navigation intervention to sustain viral suppression among HIV-positive men and transgender women released from jail. *JAMA Internal Medicine*, *178*, 542–553.

Dagne, G. A. (2004). Hierarchical Bayesian analysis of correlated zero-inflated count data. *Biometrical Journal*, *46*, 653–663.

Dobbie, M. J., & Welsh, A. (2002). Modelling correlated zero-inflated count data. *Australia and New Zealand Journal of Statistics*, *43*, 431–444.

Flynn, P. M., & Brown, B. S. (2008). Co-occurring disorders in substance abuse treatment: issues and prospects. *Journal of Substance Abuse Treatment*, *34*, 36–47.

Foundtas, G., & Anastasopoulos, P. C. (2018). Analysis of accident injury-severity outcomes: The zero-inflated hierarchical ordered probit model with correlated disturbances. *Analytic Methods in Accident Research*, *20*, 30–45.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis* (Third ed.). Chapman and Hall/CRC.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*, 997–1016.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 145–472.

Ghosal, S., Lau, T. S., Gaskins, J., & Kong, M. (2020). A hierarchical mixed effect hurdle model for spatiotemporal count data and its application to indentifying factors impacting health professional shortages. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*, 1121–1144.

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

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

Grant, B. F., Stinson, F. S., Dawson, D. A., Chou, S. P., Dufour, M. C., Compton, W., Pickering, R. P., & Kaplan, K. (2004). Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders: results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Arch Gen Psychiatry*, *61*, 807–816.

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

Hall, D. B., & Zhang, Z. (2004). Marginal models for zero inflated clustered data. *Statistical Modeling*, *4*, 161–180.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their

applications. *Biometrika*, *57*, 97–109.

Hecht, F. M., Wilson, I. B., Wu, A. W., Cook, R. L., & Turner, B. J. (1999). Optimizing care for persons with HIV infection. *Annals of Internal Medicine*, *131*, 136–143.

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

Johnson, N. L., & Kotz, S. (1966). *Discrete Distributions: Distributions in Statistics*. Wiley.

Karno, M. P., Rawson, R., Rogers, B., Spear, S., Grella, C., Mooney, L. J., Saitz, R., Kagan, B., & 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.

Kitahata, M. M., Rompaey, S. E. V., Dillingham, P. W., Koepsell, T. D., Deyo, R. A., Dodge, W., & Wagner, E. H. (1999). Primary care delivery is associated with greater physician experience and improved survival among persons with AIDS. *Annals of Internal Medicine*, *131*, 136–143.

Kong, M., Xu, S., Levy, S. M., & Datta, S. (2015). GEE type inference for clustered zero-inflated negative binomial regression with application to dental caries. *Computational Statistics and Data Analysis*, *85*, 54–66.

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

Lee, K., Joo, Y., Song, J. J., & Harper, D. W. (2011). Analysis of zero-inflated clustered count data: A marginalized model approach. *Computational Statistics and Data Analysis*, *55*, 824–837.

Li, C.-S., Lu, J.-C., Park, J., Kim, K., Brinkley, P. A., & Peterson, J. P. (1999). Multivariate zero-inflated Poisson models and their applications. *Technometrics*, *41*(1), 29–38.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley and Sons.

Liu, Y., & Tian, G.-L. (2015). Type I multivariate zero-inflated Poisson distribution with

applications. *Computational Statistics & Data Analysis*, *83*, 200–222.

Long, D. L., Preisser, J. S., Herring, A. H., & Golin, C. E. (2014). A marginalized zero-inflated Poisson regression model with overall exposure effects. *Statistics in Medicine*, *33*, 5151–5165.

Lütkepohl, H. (2013). Vector autoregressive models. In *Handbook of research methods and applications in empirical macroeconomics.* Edward Elgar Publishing.

Maruschak, L. M. (2006). *HIV in Prisons.* Retrieved 2022-02-01, from `https://www.bjs.gov/content/pub/pdf/hivp06.pdf`

Medical expenditure panel survey. (2012). Retrieved from `https://meps.ahrq.gov/mepsweb/index.jsp`

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

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

Mugavero, M. J., Lin, H.-Y., Allison, J. J., Giordano, T. P., Willing, J. H., Raper, J. L., Wray, N. P., Cole, S. R., Schumacher, J. E., Davies, S., & Saag, M. S. (2009). Racial disparities in HIV virologic failure: Do missed visits matter? *Journal of Acquired Immune Deficiciency Syndrome*, *50*, 100–108.

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

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

O'Malley, A. J., & Zaslavsky, A. M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association*, *103*(484), 1405–1418.

Ridout, M., Demétrio, C. G., & Hinde, J. (1998). Models for count data with many zeros. In *Proceedings of the XIXth International Biometric Conference* (Vol. 19, pp. 179–192).

Rue, H., & Held, L. (2005). *Gaussian Markov Random Fields*. Chapman and Hall.

Sabin, K. M., Frey, R., Jr, Horsley, R., & Greby, S. M. (2001). Characteristics and trends of newly identified HIV infections among incarcerated populations: CDC HIV voluntary counseling, testing and referral system, 1992-1998. *Journal of Urban Health*, *78*, 241–255.

Saitz, R. (2010). Co-occurring disorders in substance abuse treatment: issues and prospect-salcohol screening and brief intervention in primary care: absence of evidence for efficacy in people with dependence or very heavy drinking. *Journal of Substance Abuse Treatment*, *29*, 631–640.

Saitz, R. (2014). Screening and brief intervention for unhealthy drug use: little or no efficacy. *Psychiatry*, *5*, 1–5.

Springer, S. A., Pesanti, E., Hodges, J., Macura, T., Doros, G., & Altice, F. L. (2004). Effectiveness of antiretroviral therapy among HIV-infected prisoners: reincarceration and the lack of sustained benefit after release to the community. *Clinical Infectious Diseases*, *38*, 1754–1760.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–550.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *The Econometric Society*, *26*(1), 24–36.

Tunaru, R. (2002). Hierarchical Bayesian models for multiple count data. *Austrian Journal of Statistics*, *31*(2), 221–229.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Yau, K. K. W., & Lee, A. H. (2001). Zero-inflated Poisson regression with random effects

to evaluate an occupational injury prevention programme. *Statistics in Medicine*, *20*, 2907–2920.

Zhu, Y., & Weiss, R. E. (2013). Modeling seroadaptation and sexual behavior among HIV+ study participants with a simultaneously multilevel and multivariate longitudinal count model. *Biometrics*, *69*(1), 214–224.

Zimmerman, D. L., & Núñez-Antón, V. A. (2005). *Antedependence Models for Longitudinal Data*. Chapman and Hall.