7-6-2016 12:00 AM

Joint Analysis of Zero-heavy Longitudinal Outcomes: Models and Comparison of Study Designs

Erin R. Lundy
The University of Western Ontario

Supervisor
Dr. Charmaine Dean
The University of Western Ontario

Graduate Program in Statistics and Actuarial Sciences
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Abstract

Understanding the patterns and mechanisms of the process of desistance from criminal activity is imperative for the development of effective sanctions and legal policy. Methodological challenges in the analysis of longitudinal criminal behaviour data include the need to develop methods for multivariate longitudinal discrete data, incorporating modulating exposure variables and several possible sources of zero-inflation. We develop new tools for zero-heavy joint outcome analysis which address these challenges and provide novel insights on processes related to offending patterns. Comparisons with existing approaches demonstrate the benefits of utilizing modeling frameworks which incorporate distinct sources of zeros. An additional concern in this context is heaping of self-reported counts where recorded counts are rounded to different levels of precision. Alternatively, more accurate data that is less burdensome on participants to record may be obtained by collecting information on presence/absence of events at periodic assessments. We compare these two study designs in the context of self-reported data related to criminal behaviour and provide insights on choice of design when heaping is expected.

The contributions of this research work include the following: (i) Developing a general framework for joint modeling of multiple longitudinal zero-inflated count outcomes which incorporates a variety of probabilistic structures on the zero counts. (ii) Accommodating a subgroup of subjects who are not at-risk to engage in a particular outcome (iii) Incorporating the effect of a time-dependent exposure variable in settings where some outcomes are prohibited during exposure to a treatment. (iv) Illustrating the extent to which heaping of zero-inflated counts, arising from a variety of heaping mechanisms, can introduce bias, impeding the identification of important risk factors (v) Identifying situations where there is very little loss of efficiency in the analysis of presence/absence data, depending on the partition of the time for the presence/absence records and the underlying rate of events. (vi) Providing recommendations on the design of studies when heaping is a concern. (vii) Modeling of multiple longitudinal binary outcomes where a mixture model approach allows differential rates of recurrence of

events, and where the underlying process generating events may resolve.

Keywords: Zero-inflation, Joint Modeling, Longitudinal Data, Random Effect Model, Discrete Data, Mixture Model, Markov Chain Monte Carlo, Heaped Data

Co-Authorship Statement

Paper title: Joint Analysis of Multivariate Longitudinal Zero-heavy Panel Count Outcomes with Differing Exposures

Publication: In preparation

List of authors: Erin Lundy, Charmaine Dean, Elizabeth Juarez-Colunga, Edward Mulvey, Carol Schubert

The problem of jointly analyzing several longitudinal zero-heavy outcomes in settings with a modulating exposure was formulated by Dr. Dean and Dr. Juarez-Colunga. As well, Dr. Juarez-Colunga initiated the exploratory data analysis. Dr. Dean proposed utilizing a mixture framework in which outcomes are linked by subject-specific random effects. I suggested extending previous work on incorporating the extent of exposure in zero-inflated models to our context. I performed extensive exploratory work to validate the modeling assumptions and conducted the analyses seen in this paper. Dr. Dean and Dr. Juarez-Colunga provided guidance and suggestions on the data analysis and Bayesian methodology. Dr. Mulvey and Ms. Schubert provided insights into the application of these methods to the criminal behaviour setting. All authors contributed to the preparation of the manuscript with regard to the content and relevance of the work.

Paper title: Analyzing Heaped Counts Versus Longitudinal Presence/Absence Data in Joint Zero-inflated Discrete Regression Models

Publication: In preparation

List of authors: Erin Lundy, Charmaine Dean

Dr. Dean proposed an investigation of the two types of data records available in our motivating data set, as well as the use of the zero-inflated Poisson process framework. I performed the data analysis and empirical study in this paper. With guidance from Dr. Dean, I proposed the heaping distributions explored in the simulation study. I formulated the recommendations

on the study design and defined the measure of discrepancy. Both authors contributed to the preparation of the manuscript with regard to the content and relevance of the work.

Paper title: Joint Analysis of Multivariate Longitudinal Presence/Absence Data Subject to Resolution

Publication: In preparation

List of authors: Erin Lundy, Charmaine Dean

I initiated work on the modeling framework utilizing a common latent variable to represent the resolution of the underlying process generating events. Dr. Dean raised the idea of using presence/absence data in this context and, subsequently, I defined the expression for the probability of at least one event for Binomial counts. The fact that the estimation of the probability of permanently quitting utilizes data from the entire observation period was first stated by Dr. Dean. I performed the data analysis and empirical study in this paper. A discussion of the extension for real-time predictions using gap time distributions was presented by Dr. Dean. Both authors contributed to the preparation of the manuscript with regard to the content and relevance of the work.

To my grandparents, Velva and Pekka Roininen

Acknowledgments

I would like express my gratitude to my supervisor, Dr. Charmaine Dean for her support and guidance. I feel fortunate to have had her as my supervisor. I also wish to thank my collaborators Dr. Elizabeth Juarez-Colunga, Dr. Edward Mulvey and Carol Schubert; working with them has been a very rewarding experience. Thanks to Dr. Giovani Silva for his advice and support. I am grateful for the encouragement of Dr. Bethany White to participate in statistics-related projects outside my research; these opportunities have greatly enriched my time at Western. I appreciate the support, encouragement and friendship of my fellow graduate students. Special mention goes to Alisha Albert-Green for the many helpful discussions as well as my fellow lab members.

I would like to acknowledge my high school math teacher, Ron Macdonald, who inspired me to pursue mathematics at a higher level and whose words of encouragement have stayed with me over the years.

Finally, and most importantly, I would like to thank my family and friends for their unwavering support. Especially, Andrew, my perpetual cheerleader, I could not have done this without you.

Contents

List of Figures

List of Tables

List of Appendices

Chapter 1

Introduction

Regression models for zero-inflated count data often need to accommodate within-subject correlation and between-individual heterogeneity; frequently random effects models are utilized for incorporating such complex correlation structures. In cases where several longitudinal zero-heavy count outcomes are jointly considered, zero counts for different outcomes may arise from distinct sources, so that a flexible approach for handling the zero-inflated outcomes jointly becomes imperative. As well, sometimes count outcomes are regulated by an exposure variable, with the length of exposure, for example, being proportional to expected counts. In the case of a joint outcome analysis, it may be that the extent of exposure differs from outcome to outcome. Both of these complications arise in our motivating context; importantly some outcomes are prohibited during a specific treatment leading to some of the zero-heavy nature of the data accounted for in a structural manner based on an exposure variable. In Chapter 2, we develop a flexible mixture modeling approach for handling such joint outcome analyses adopting a conceptual framework similar to a mover-stayer model for handling the excess zeros. We also investigate carry-over effects of time in a secure facility on the outcome in the subsequent panel. Compared with existing methodology, our approach enables a better understanding, offering new insights on processes related to offending patterns.

Self-reported count data are often subject to heaping where reported counts are rounded to

different levels of precision. This arises in settings where exact event times are not available but instead aggregated counts of self-reported events over the observation period are recorded. In situations where counts are aggregated over a long observation period, rounding of the data is not unusual. This yields a distorted distribution of the observed counts and may bias estimation. In Chapter 3, we illustrate the extent to which heaping of zero-inflated counts, arising from a variety of heaping mechanisms, can introduce bias. Alternatively, an accurate recording of presence/absence of events between shorter periodic assessments may provide a competitive approach for self-reported data in terms of high efficiency relative to the analysis of counts. An additional benefit is the reduction of the burden of data collection on respondents. But it is not clear whether there is sufficient benefit of this approach versus an analysis of rounded aggregate counts since certainly some efficiency loss is expected. In our motivating example, the utility of count data aggregated over a year, and rounded, as well as monthly binary data, indicating the presence/absence of events, are contrasted. We compare the analysis of these two types of data records in the context of a joint analysis of two zero-heavy outcomes, where outcomes are linked by a subject-specific random effect. Simulations and empirical studies demonstrate that the analysis of aggregate heaped count data and longitudinal presence/absence data can lead to differing results and, importantly, conflicting conclusions concerning possible risk factors depending on the bias introduced by the heaping. As well, we identify situations where there is very little loss of efficiency in the analysis of presence/absence data. We conclude Chapter 3 by offering recommendations on the design of studies using self-reported data, where heaping may be a concern.

A major aim of studies examining criminal behaviour is understanding the patterns and mechanisms of the process of desistance from criminal activity, as insight so derived is essential for developing effective sanctions and legal policy. In cases where several types of criminal behaviour are considered in a joint outcome analysis, we may conceptualize a latent variable representing the individual susceptibility to engage in criminal activity, which underlies each outcome and hence links outcomes. The analysis of such data is often complicated

by a proportion of subjects who never engage in a particular outcome. Additionally, some subjects eventually desist in engaging in criminal activities leading to what is termed a resolution of the process. As well, longitudinal studies may record only binary data indicating the presence/absence of events between periodic assessments. Finally, incorporating time spent in a secure facility (incarceration, for example) as an exposure variable regulating the occurrence of events is important in these analyses. In Chapter 4, we present a general modeling framework for joint analysis of multiple longitudinal binary outcomes which addresses these challenges. In our novel framework, a mixture model approach accommodates differential rates of recurrence of events, and allows that the underlying process generating events may resolve. Compared with existing approaches, our methodology offers new insights on the processes generating the observed offending patterns. Simulations demonstrate that the proposed methods can accurately differentiate between juvenile offenders who have ceased engaging in criminal behaviour and those who have not.

The methods and models developed in this thesis are motivated by a major study of criminal behaviour patterns. As the application is significant for our developments, in the subsection below we provide contextual background of the study and an in depth description of the data.

1.1 Motivating Study

The juvenile justice system is responsible for keeping communities safe while considering the best interest of the child and rehabilitating young offenders. This requires knowledge and insight concerning the processes related to how and why juveniles desist from committing crime. Unfortunately, the data on either patterns of desistance or escalation or the effects of interventions and sanctions on trajectories of offending during and after adolescence is limited, particularly with regarding serious adolescent offenders.

Sanctions for adolescent offenders are generally determined using commonsense guidelines that have developed through years of practice (Mulvey et al., 2004). As a result, serious offenders are generally given some form of sanction which has strong potential to control crime while

less serious offenders are often enrolled in shorter-term programs. As well, younger serious offenders are more likely to be given an opportunity for rehabilitation.

An important finding from vast literature on risk factors associated with adolescent antisocial behaviour is that relatively few adolescent offenders become serious adult offenders. Consequently, a crucial challenge is reliably distinguishing between juvenile offenders who will continue antisocial behaviour into adulthood and those who will not. The motivating study for this dissertation, the Pathways to Desistance study (Schubert et al., 2004), aims to address this challenge. It is a major study investigating the offending patterns of serious juvenile offenders from adolescence to early adulthood. A total of 1354 adolescents between 14 and 17 years old at the time of their initiating offense were recruited from the juvenile and adult court systems in Philadelphia, Pennsylvania (N=654) and Phoenix, Arizona (N=700) between November 2000 to January 2003. The study sample consists of primarily minority (44% African American and 29% Hispanic) males (86%) with an average of two prior petitions to court. However, 26% of the sample had no prior petitions other than the offense that qualified them for study enrollment. Eligible crimes for enrollment into the study included all felony offenses with the exceptions of less serious property crimes, misdemeanor weapons offenses and misdemeanor sexual assault. As drug law violations represent substantial proportion of offenses for males within this age group, the proportion of male subjects with drug offenses was limited to 15% of the sample at each site.

During the enrollment period, slightly more than one half of the youth determined to be adjudicated on an eligible charge were approached for enrollment. Those not approached were excluded due to operational and design constraints. The participation rate, calculated as the number of participants enrolled divided by the number approached for enrollment, was 67% and the refusal rate, defined as the number of adolescents or guardians who declined to take part in the study divided by the number approached, was 20%. There were several differences between the subjects who were adjudicated, but not enrolled, and the subjects enrolled in the study. The enrolled group was younger at their adjudication hearing, had more prior petitions,

and appeared in the court for the first time at an earlier age. As well, the proportion of girls was higher in the enrolled group. These differences are consistent with the investigators' increased efforts to recruit more serious juvenile offenders and more female subjects. Finally, proportionately more white offenders and fewer African American subjects were enrolled in the study. This discrepancy was likely related to the imposed quota on the proportion of subjects adjudicated on drug charges as there is likely to be an association between adjudications for drug charges and ethnicity.

A baseline interview collecting information about background characteristics and previous offending was conducted at the time of enrollment. Follow up interviews were conducted every 6 months for the first 3 years and every year for an additional 4 years, resulting in a total of ten follow up interviews. A target date for each follow up interview was determined based on the date of baseline interview to ensure approximately equal observation periods for all individuals. Follow up interviews were scheduled in the time period spanning 6 weeks prior to the target interview date and 8 weeks after the target date. The baseline and follow up interviews covered six domains: (a) background characteristics, (b) indicators of individual functioning, (c)psychosocial development and attitudes, (d) family context, (e) personal relationships, and (f) community context. The interviews were conducted electronically, with the computer screen visible to both the interviewer and the participant. Confidentiality was assured through confidentiality protections provided by statute to the U.S. Department of Justice. Each participant was randomly assigned to a single interviewer throughout the course of the study. This consistency in interviewer was important to promote rapport, provide continuity for the participant and hopefully increase disclosure. The self-reported data collected during the interviews was supplemented and validated through interviews with collateral reporters, usually parents, and official record information.

At each interview, two types of data records for illegal and antisocial activity were collected. First, subjects indicated in which months, since the last scheduled interview, they engaged in the antisocial or illegal activity. Secondly, they reported how many times they

engaged in the activity since the last scheduled interview. Therefore, the available data on offending consists of panel count data and repeatedly measured binary data recording presence/absence of events during each month of observation. Table Table A.1 in Appendix A lists the offending outcomes on which the count and binary data were collected. The design of the questionnaire used at the follow up interviews was based on previously developed life calendars. Such methods for constructing life-event calendars have been shown to provide reasonably accurate information about the temporal ordering of events during the period covered by an interview and have been successfully used in studies of criminal offending, antisocial behavior, and mental health service use (Caspi et al., 1996; Horney, Osgood, & Marshall, 1995).

The use of self-reported count data rises questions about recall error, particularly given the long periods between interviews. Previous authors (Monahan and Piquero, 2009) have expressed concerns about the accuracy and reliability of the count data in this data set, especially with respect to recall errors corresponding to frequent and aggressive offenders. Additionally, for three of the illegal or antisocial activities, *carried a gun*, *sold marijuana* and *sold other drugs*, the reported count refers to the number of days the event occurred while for the remaining activities the reported count refers to the number of times the subject engaged in the act. Confusion with regard to what sort of count is requested may have led to outliers as there are a few cases where the reported number of days the event occurs exceeds the maximum possible for the window of observation. For these 0.52% of cases, the number of days was set at the maximum possible.

The participants' high degree of mobility and engagement in illegal activity made tracking and retention of subjects difficult. The Pathways to Desistance study used a wide range of tactics to maintain contact with participants including phone calls during odd hours, unscheduled visits to the participants home, neighborhood, and hangouts, enlisting support and obtaining information from family members and friends mentioned in previous interviews, and conducting address searches with credit databases, community agencies, and criminal justice facilities. Additionally, study participants were paid using a graduated payment schedule.

Overall subject retention was good; at specific follow up interviews the proportion of subjects who completed the interview ranges from 83.5% to 92.5%. As shown in Table 1.1, the proportion of subjects who completed the interview decreased over time. In Table 1.2, we display the percentage of participants with complete data for all follow up interviews as well as for 9,8,...,1,0 interviews. The majority of subjects (79%) completed 9 or 10 interviews. As well, we provide this data, stratified by gender and ethnicity in Table 1.3. Compared to male subjects, a higher proportion of female subjects completed all ten interviews. There are also differences in the number of complete follow up interview across ethnicity with a higher proportion of white and Hispanic subjects completing all ten interviews than Black subjects or subjects of another ethnic origin.

Table 1.1: Status of follow up interview by interview period (IP).

Status	IP 1	IP 2	IP 3	IP 4	IP 5	IP 6	IP 7	IP 8	IP 9	IP 10
Complete (%)	92.54	92.54	89.59	90.32	90.55	90.77	89.66	88.85	86.78	83.53
Missing (%)	6.57	6.79	9.23	9.08	8.86	9.01	10.27	10.76	12.92	16.25
Partial (%)	0.89	0.66	1.18	0.59	0.59	0.22	0.07	0.30	0.30	0.22

Table 1.2: Percentage of participants with complete data for all (10) interviews as well as for 9,8,...,1,0 interviews.

Interviews	0	1	2	3	4	5	6	7	8	9	10
% Participants	1.33	0.74	0.81	0.81	1.33	1.55	2.88	4.43	7.31	17.58	**61.23**

Table 1.3: Percentage of participants with complete data for all (10) interviews as well as for 9,8,...,1,0 interviews, stratified by gender and ethnicity.

	0	1	2	3	4	5	6	7	8	9	10
Male (%)	1.37	0.85	0.94	0.94	1.54	1.45	3.16	4.53	7.69	17.61	**59.91**
Female (%)	1.09	0	0	0	0	2.17	1.09	3.80	4.89	17.39	**69.57**
Black (%)	2.50	0.89	1.25	0.89	1.97	1.07	4.28	4.63	8.73	19.61	**54.19**
Hispanic (%)	0.44	0.88	0.44	0.44	1.10	1.98	1.76	3.52	7.05	16.52	**65.86**
White (%)	0	0.36	0.73	0.73	0.73	2.19	2.19	3.65	5.11	14.23	**70.07**
Other (%)	3.08	0	0	3.08	0	0	1.54	12.31	6.15	21.54	**52.31**

In this dissertation, we view a sanction or intervention as a placement in one of seven different types of facilities without community access: (i) Drug or alcohol treatment units where the

primary focus is providing substance use treatment services. This included both detoxification and longer-term substance use treatment programs, with the vast majority being longer-term treatment facilities. (ii) Psychiatric hospitals or psychiatric units of a general hospital providing inpatient acute care to evaluate and stabilize individuals with mental health problems.(iii) Jails, which are usually locally run and hold youths until trial or for relatively short sentences after trial and prisons; these are typically state-run and hold offenders for a longer sentence after trial. The main goal of these settings is incarceration. (iv) Detention facilities where adolescents await their adjudication hearing or more permanent placement location after adjudication and disposition. (v) State-run, secure juvenile facilities providing secure custody, education, and treatment to committed youth. (vi) Contracted residential treatment (general) facilities providing residential care within a structured environment and that may offer a range of services. (vii) Contracted residential treatment (mental health) facilities where the primary focus is the treatment of the youth's mental health needs. Data on placement in a secure facility including the type of facility and the duration of the placement were recorded monthly. However, only one type of facility can be recorded per month. Therefore, if a subject was in more than one type of facility during a single month, the type of facility with the longest stay was recorded.

1.2 Plan of the Thesis

The motivating data set highlights gaps in the current literature that need to be addressed in order to analyze complex data sets and the complications considered here may arise in a variety of longitudinal data settings. The main methodological challenges include the need to develop methods for multivariate longitudinal discrete data, incorporating modulating exposure variables and several possible sources of zero-inflation. Additionally, we accommodate a subgroup of subjects who eventually desist engaging in criminal activities, utilizing a modeling framework where the simultaneous resolution of several recurrent event processes is possible. As well, we contrast inference based on the analysis of self-reported count data aggregated over the period of observation with that of repeatedly collected binary data indicating the pres-

ence/absence of events between shorter periodic assessments using joint zero-inflated discrete regression models when rounding of the count data is expected. Each chapter addresses different issues related to the joint analysis of zero-heavy longitudinal outcomes and is presented in a style similar to that for publication. As a result, some introductory material is repeated.

This thesis concludes with a discussion of future work emerging from extensions of the methods developed.

References

Caspi, A., Moffitt, T. E., Thornton, A., Freedman, D., Amell, J. W., Harrington, H., et al. (1996). The life history calendar: a research and clinical assessment method for collecting retrospective event-history data. *International Journal of Methods in Psychiatric Research* **6**, 101114.

Horney, J., Osgood, D. W., & Marshall, I. H. (1995). Criminal careers in the short-term: Intra-individual variability in crime and its relation to local life circumstances. *American Sociological Review* **60**, 665673.

Monahan, K. C., & Piquero, A. R. (2009). Investigating the longitudinal relation between offending frequency and offending variety. *Criminal Justice and Behavior* **36**, 653-673.

Mulvey, E. P., Steinberg, L., Fagan, J., Cauffman, E., Piquero, A. R., Chassin, L., et al. (2004). Theory and research on desistance from antisocial activity among serious adolescent offenders. *Youth Violence and Juvenile Justice* **3**, 213-236

Schubert, C. A., Mulvey, E.P., Steinberg, L., Cauffman, E., Losoya, S., Hecker, T., Chassin, L., et al. (2004). Operational lessons from the Pathways to Desistance project. *Youth Violence and Juvenile Justice* **3**, 237-255

Chapter 2

Joint Analysis of Multivariate Longitudinal Zero-heavy Panel Count Outcomes with Differing Exposures

2.1 Introduction

Joint modeling is a generic term used to describe situations where two or more processes are modeled in a way such that the models directly or indirectly influence each other. Outcomes measured on the same subject may be correlated so that conceptualizing a shared latent variable that reflects unobserved individual traits affecting outcomes may be useful for gaining precision for the estimation of parameters. Previous authors have demonstrated that the use of joint models can lead to efficiency gains for the marginal parameters of interest when the association between outcomes is strong (Zeng and Cook, 2007). Furthermore, it has been shown that ignoring the correlation between outcomes can lead to biased estimates (Guo and Carlin, 2004). In this paper, we utilize the general framework proposed by Dunson (2000) in which, conditional on random effects, different members of the exponential family are used to describe the component models in the joint distribution of the set of observed outcomes.

Multi-state stochastic models are useful for the analysis of data from longitudinal studies monitoring individuals moving through various states, when interest centers on the dynamic aspects of the process under investigation. If it is hypothesized that a subgroup of subjects, termed stayers, will remain in the initial state over time, whereas others, termed movers, will make transitions among states, the overall process can be modeled by a finite mixture model called a mover-stayer model. Blumen, Kogan and McCarthy (1955) first introduced discrete time mover-stayer models which consist of a mixture of two independent Markov chains, one degenerate, with a transition matrix equal to the identity matrix for the stayers, and another with an unspecified transition matrix. When several longitudinal count outcomes are jointly considered, and excess zeros may arise from several distinct sources, adopting the basic structure of a mover-stayer model may provide a suitable approach to address a variety of frameworks generating the zero counts.

In settings where the proportion of zero counts is high relative to what is expected based on the distribution of the non-zero counts, standard count distributions such as Poisson, binomial and negative binomial may not provide an adequate fit. Mixture methods for handling zero-inflated counts have received considerable attention in the literature, especially over the last two decades. Two influential foundational papers include Lambert (1992) and Hall (2000). In a manufacturing context, Lambert (1992) introduced zero-inflated Poisson (ZIP) regression models where the probability of a perfect, non-zero defect state and the mean number of defects in the imperfect state are allowed to depend on covariates via canonical link generalized linear models. Motivated by a horticular experiment with a repeated measures design, Hall (2000) adapted Lambert's methodology to the setting with upper bounded counts and proposed the zero-inflated binomial (ZIB) model, including random effects in the mean component of the ZIP and ZIB regression models to accommodate the within-subject correlation and the between-subject heterogeneity typically observed in longitudinal data. Several authors (Boone, Stewart-Koster and Kennard, 2012; Buu et al., 2012; Ghosh and Tu, 2008) have developed zero-inflated count regression models that incorporate correlation structures arising in

longitudinal, clustered or spatial data.

Methods for the joint analysis of several count and zero-inflated count outcomes have been recently developed. Rodrigues-Motta et al. (2013) proposed a joint model for multivariate overdispersed count data where correlation among observations for the same subject is incorporated through the inclusion of correlated outcome- and subject-specific random effects in the mean component. Additionally, they allowed correlated counts to follow different distributions such as Poisson, negative binomial and ZIP. Feng and Dean (2012) discussed joint models for multivariate spatial count data with excess zeros, where outcomes are linked through a shared latent spatial random risk term.

We generalize two existing methodologies: *zero-heavy longitudinal count* models and *joint outcome zero-heavy count* analysis which accommodates longitudinal, multivariate data and which adopts a framework similar to the mover-stayer concept for handling some of the zero counts. Our context for these developments is a major study on criminal behaviour patterns of serious adolescent offenders from adolescence into early adulthood. One goal of this study is to examine the effect of institutional placement on subsequent offending. A specific concern is the carry-over effects of time in a facility with no community access on the offending behavior in the subsequent observation period.

A complicating factor in the analysis of this data set is that the likelihood of some of the criminal activities (e.g. stealing a car) is severely reduced if the individual is in a facility with no community access. Therefore, the length of exposure, defined as the length of time a subject is at-risk to engage in an outcome, varies from outcome to outcome. One possible approach to incorporate the extent of individual exposure in zero-inflated count models is to assume that the mean count is proportional to the exposure time (Lee, Wang and Yau, 2001). Baetschmann and Winkelmann (2013) extended this approach for analysis of a zero-inflated outcome by assuming that structural zeros are generated by a separate process. From this viewpoint, a structural zero occurs if the waiting time until an event exceeds the exposure time. Hence, the probability of a structural zero is equal to the survival function of the waiting time distribution

evaluated at the exposure time. Modeling the probability of a structural zero using a survival function is logical for settings where some outcomes may be prohibited due to a modulating exposure. Specifically, the probability of a structural zero decreases with the length of exposure and if the length of exposure is 0 then the probability of a structural zero is 1. In this application, we incorporate the length of exposure in the structural zero as well as the mean components of the zero-inflated mixture count models.

This article focuses on the development of new tools for zero-heavy joint outcome analysis with a major intent being the illustration of how to build relevant models and what sorts of novel insights they provide in the setting of an analysis of juvenile criminal behaviour. We proceed as follows: In Section 2.2 we describe our general joint modeling framework. In Section 2.3 we introduce our motivating data set and outline the methodological challenges for analysis. We discuss model development in the context of the study of criminal behaviour in Section 2.4. Highlighting innovations and new insights stemming from our mixture modeling approach as well as the framework we adopt for our exposure variable, we present the results of our joint analysis of this study in Section 2.5. A comparison with alternate models, demonstrating the benefits of jointly modeling outcomes in this data set, is provided in Section 2.6. In Section 2.7 we conclude with a discussion of results and limitations, as well as suggestions for future work.

2.2 Joint Analysis of Multivariate Longitudinal Zero-heavy Count Outcomes

Suppose there are N subjects in a study and subject i is observed at T_i follow up interviews, indicating the end of a panel length of time. At each follow up interview, subjects report the number of times they engaged in each of K outcomes during the corresponding panel. We refer to subjects as non-engagers if they are not at-risk to engage in outcome k, i.e. they generate zero values at all panels for outcome k. Note that it is possible for a subject to be a non-engager

for each outcome, resulting in zero values for each outcome. Let y_{itk} be the observed count for subject i at panel t for outcome k and $\mathbf{y}_{ik} = (y_{i1k}, \ldots, y_{iT_ik})'$ be the sequence of counts over $t = 1, \ldots, T_i$ observed for subject i for outcome k, $i = 1, \ldots, N$, $k = 1, \ldots, K$.

We assume each response vector $\mathbf{y}_{ik}$, conditional on random effects, is independently drawn from a mixture model having density

$$f(\mathbf{y}_{ik}|s_{ik}, r_i, u_i, v_i, d_{ik}, b_{ik}) = \begin{cases} \mathbf{I}(\mathbf{y}_{ik} = \mathbf{0}_{T_i\times 1}) & \text{if } s_{ik} = 1 \\ f_{C_k}(\mathbf{y}_{ik}|u_i, v_i, d_{ik}, b_{ik}) & \text{if } s_{ik} = 0 \end{cases} \tag{2.1}$$

where the variables s_{ik} are latent Bernoulli indicators, markers for the outcome-specific non-engagers, with mean function p_{ik}, conditional on a random effect r_i; subject and outcome specific random effects u_i, v_i, d_{ik} and b_{ik} will be discussed later. Specifically, for each outcome, we assume $s_{ik}|p_{ik} \sim Bern(p_{ik})$ with

$$p_{ik} = \{1 + \exp(-\boldsymbol{w}_i'\boldsymbol{\gamma}_k - \lambda_{rk}r_i)\}^{-1} \tag{2.2}$$

where $\boldsymbol{w}_i$ is a $q_1\times 1$ vector of covariates, $\boldsymbol{\gamma}_k$ is a vector of corresponding regression parameters, r_i is a subject-specific random effect and λ_{rk} is a factor loading parameter representing outcome-specific variability related to r_i.

For each outcome, one mixture component places all its mass on the zero vector while the other component distributes mass according to the density, $f_{C_k}(\mathbf{y}_{ik}|u_i, v_i, d_{ik}, b_{ik})$, corresponding to a longitudinal zero-heavy count model. There are several possible choices for the zero-heavy count distribution such as zero-inflated Poisson, zero-inflated negative binomial and zero-inflated binomial. Such different distributions may be required for different outcomes.

Conditional on random effects, we assume the counts for outcome-k-specific engagers follow a zero-inflated count distribution with probability of a structural zero π_{itk} so that $f_{C_k}(\mathbf{y}_{ik}|u_i, v_i, d_{ik}, b_{ik})$ is given by

$$f_{C_k}(\mathbf{y_{ik}}|u_i, v_i, d_{ik}, b_{ik}) = \prod_{t=1}^{T_i} \Big[I(y_{itk} = 0)\{\pi_{itk} + (1 - \pi_{itk}) f_k(0|\mu_{itk})\} + I(y_{itk} > 0)(1 - \pi_{itk}) f_k(y_{itk}|\mu_{itk}) \Big] \tag{2.3}$$

where f_k denotes the probability mass function of the standard (non zero-inflated) count distribution associated with outcome k and μ_{itk} is the corresponding conditional mean. The parameters of the zero-inflated count distributions, π_{itk} and μ_{itk}, $i = 1, \ldots, N$, $t = 1, \ldots, T_i$, $k = 1, \ldots, K$ are modeled as

$$\pi_{itk} = \exp[-\exp\{\boldsymbol{x}_{1it}^{'}(-\boldsymbol{\alpha}_k) + h_1(t, \boldsymbol{\rho}_{1k}) + \delta_k \log(z_{itk}) + \lambda_{uk} u_i + d_{ik}\}], \tag{2.4}$$

the survivor function of a Weibull distribution, and

$$\mu_{itk} = g_k^{-1}\{\boldsymbol{x}_{2it}^{'}\boldsymbol{\beta}_k + h_2(t, \boldsymbol{\rho}_{2k}) + \lambda_{vk} v_i + b_{ik}\} z_{itk} \tag{2.5}$$

where g_k is the canonical link function for the standard count distribution corresponding to the kth outcome; $\boldsymbol{x}_{1it}$ and $\boldsymbol{x}_{2it}$ are $q_2 \times 1$ and $q_3 \times 1$ vectors of covariates for the fixed effects while $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ are vectors of corresponding regression parameters; $h_1(t, \boldsymbol{\rho}_{1k})$ and $h_2(t, \boldsymbol{\rho}_{2k})$ are functions of time describing the temporal trends in π_{itk} and μ_{itk}. We parameterize the model for π_{itk} in terms of $-\boldsymbol{\alpha}_k$ so that a positive covariate effect corresponds to an increased probability of a structural zero. The form of π_{itk}, as well as the term $\delta_k \log(z_{itk})$ reflect the idea proposed by Baetschmann and Winkelmann (2013) to model the probability of a structural zero as the survivor function of a Weibull distribution. Here, the Weibull shape parameter is δ_k and z_{itk}, the waiting time, is the length of exposure in the panel. In the structural zero component, u_i is a subject-specific random effect shared across outcomes and λ_{uk} is the factor loading for this shared effect on outcome k. Correspondingly, in the mean component, v_i is a subject-specific random effect shared across outcomes and λ_{vk} is the factor loading for this shared effect on

outcome k. The outcome- and subject-specific random effect for the structural zero component d_{ik} and the outcome- and subject-specific random effect for the mean component b_{ik} represent additional heterogeneity beyond the shared random effect in the respective model components. We assume the random effects are normally distributed such that $r_i \sim N(0,1)$, $u_i \sim N(0,1)$, $v_i \sim N(0,1)$, $d_{ik} \sim N(0,\sigma^2_{d_k})$ and $b_{ik} \sim N(0,\sigma^2_{b_k})$, $i = 1,\ldots,N$ and $k = 1,\ldots,K$. Thus, the shared frailties allow for the outcomes to be linked in the probability of a non-engager, and for engagers, both in the structural zero and mean components of the model.

Our mixed joint model for multivariate longitudinal zero-heavy count data may be implemented in a Bayesian framework using Markov chain Monte Carlo (MCMC) methods. The joint posterior distribution of the parameters is

$$\begin{aligned} p(\boldsymbol{\Theta},\boldsymbol{r},\boldsymbol{u},\boldsymbol{v},\boldsymbol{d},\boldsymbol{b}|\boldsymbol{Y}) \propto & L(\boldsymbol{Y}|\boldsymbol{\Theta},\boldsymbol{r},\boldsymbol{u},\boldsymbol{v},\boldsymbol{d},\boldsymbol{b})p(\boldsymbol{d}|\boldsymbol{\sigma}^2_d)p(\boldsymbol{b}|\boldsymbol{\sigma}^2_b)\pi(\boldsymbol{\sigma}^2_d)\pi(\boldsymbol{\sigma}^2_b) \\ & p(\boldsymbol{r})p(\boldsymbol{u})p(\boldsymbol{v})\pi(\boldsymbol{\Theta}) \end{aligned} \tag{2.6}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\rho}_1,\boldsymbol{\rho}_2,\boldsymbol{\delta},\boldsymbol{\lambda}_r,\boldsymbol{\lambda}_u,\boldsymbol{\lambda}_v)'$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_K)'$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_K)'$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K)'$, $\boldsymbol{\rho}_1 = (\boldsymbol{\rho}_{11},\ldots,\boldsymbol{\rho}_{1K})'$, $\boldsymbol{\rho}_2 = (\boldsymbol{\rho}_{21},\ldots,\boldsymbol{\rho}_{2K})'$, $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_K)'$, $\boldsymbol{\lambda}_r = (\lambda_{r1},\ldots,\lambda_{rK})'$, $\boldsymbol{\lambda}_u = (\lambda_{u1},\ldots,\lambda_{uK})'$, $\boldsymbol{\lambda}_v = (\lambda_{v1},\ldots,\lambda_{vK})'$, $\boldsymbol{\sigma}^2_d = (\sigma^2_{d_1},\ldots,\sigma^2_{d_K})'$, $\boldsymbol{\sigma}^2_b = (\sigma^2_{b_1},\ldots,\sigma^2_{b_K})'$, $\boldsymbol{r} = (r_1,\ldots,r_N)'$, $\boldsymbol{u} = (u_1,\ldots,u_N)'$, $\boldsymbol{v} = (v_1,\ldots,v_N)'$, $\boldsymbol{d} = (d_{11},\ldots,d_{N1},d_{12},\ldots,d_{NK})'$ and $\boldsymbol{b} = (b_{11},\ldots,b_{N1},b_{12},\ldots,b_{NK})'$. The first term on the right hand side of (2.6) is the likelihood

$$\begin{aligned} L(\boldsymbol{Y}|\boldsymbol{\Theta},\boldsymbol{r},\boldsymbol{u},\boldsymbol{v},\boldsymbol{d},\boldsymbol{b}) \propto \prod_{i=1}^{N}\prod_{k=1}^{K} & [\mathbf{I}(\mathbf{y}_{ik} = \mathbf{0}_{T_i\times 1})\{p_{ik} + (1-p_{ik})f_{C_k}(\mathbf{0}_{T_i\times 1}|u_i,v_i,d_{ik},b_{ik})\} \\ & + \mathbf{I}(\mathbf{y}_{ik} \neq \mathbf{0}_{T_i\times 1})\{(1-p_{ik})f_{C_k}(\mathbf{y}_{ik}|u_i,v_i,d_{ik},b_{ik})\}] \end{aligned} \tag{2.7}$$

The Bayesian model specification is made complete by assigning prior distributions to $\boldsymbol{\Theta}$, $\boldsymbol{\sigma}^2_d$ and $\boldsymbol{\sigma}^2_b$. Inference is then based on the posterior distribution, which can be summarized

using samples drawn from the posterior distribution. This framework for the analysis was implemented through the freely available software JAGS (Plummer, 2003).

2.3 A Study of Antisocial Behaviour Among Serious Juvenile Offenders

The Pathways to Desistance study (Mulvey et al., 2004; Schubert et al., 2004) is a longitudinal study of a group of serious juvenile offenders investigating offending patterns in the period following court adjudication. A total of 1354 youth offenders, aged 14 through 17 years old, who were found guilty of at least one serious offense in metropolitan areas of Phoenix, Arizona or Philadelphia, Pennsylvania were enrolled in the study between 2000 and 2003 and followed for up to 7 years. The primary aim of the study is to identify patterns of desistance or escalation among serious juvenile offenders and evaluate the effects of adolescent development, sanctions and interventions on these offending patterns.

All subjects completed a baseline interview where information about background characteristics and previous offending was collected. Additionally, interviews were conducted over a seven year follow up period. We analyze here panel data recorded approximately annually over the seven year period. At each follow up interview, data pertaining to antisocial and criminal activity in the period since the last scheduled interview were recorded. During the follow up period, subjects may have spent time in a facility with no access to the community, termed a secure facility. Data on placement in a secure facility and, if so, the proportion of the panel spent in a secure facility, are available. Some of the antisocial and criminal activities are highly unlikely to occur in a secure facility and, for this analysis, are considered prohibited in a secure facility. We summarize the eight outcomes considered here: *carried a gun*, *sold marijuana*, *sold other drugs*, *drove drunk*, *aggressive I*, *aggressive II*, *income I*, and *income II*, in Table A.1; we provide a list of the antisocial and criminal activities associated with each outcome, indicate whether the outcome is considered prohibited while a subject is in a secure facility and

the type of response collected. These outcomes may refer to the number of times the subject engaged in an activity and, therefore, represent an unbounded count or they may refer to the number of days the subject engaged in an activity which is bounded by the length of exposure.

A large proportion of the observed counts are zero and there are several distinct patterns in the occurrence of zero counts. In particular, a substantial proportion of subjects, ranging from 81% for *aggressive I* to 21% for *aggressive II*, never report participating in a particular outcome during the follow up period. Furthermore, there are distinct trends in the proportion of zero counts over time across the different outcomes, displayed in Figure A.1. The proportion of zeros substantially increases over time for *aggressive II* and *income II* whereas there is less of a sharp increasing trajectory for the proportion of zeros related to the remaining six outcomes. This motivates consideration of novel zero-inflated models which incorporate a variety of structures on the joint longitudinal zero counts.

2.4 Model Development for Joint Zero-heavy Outcomes Related to Antisocial Behaviour

2.4.1 Model Specification

We restrict our analysis to subjects for whom at least one year of data (N=1170) was available. For each subject, the number of time points included in the analysis, T_i, is defined as the number of consecutive panels of follow up with complete data. We define the length of exposure, z_{itk}, as the number of days in the panel, for outcomes that are not prohibited in a secure facility, and as the number of days spent in the community, for outcomes that are prohibited in a secure facility. Recall that length of each panel is approximately one year. For outcomes not prohibited in a secure facility, we utilize the survivor function of an exponential distribution to model the probability of a structural zero and, hence, set $\delta_k = 1$ as the panel length takes only a few values. On the other hand, for outcomes prohibited in a secure facility, we utilize

the survivor function of a Weibulll distribution to model the probability of a structural zero and estimate δ_k. For outcomes corresponding to bounded counts, we assume, conditional on random effects, the counts for an outcome-specific engager follow a ZIB distribution where the number of trials, z_{itk}, is the number of days the outcome could have occurred; g_k is the logit link function. For outcomes corresponding to unbounded counts, conditional on random effects, we assume the counts for an outcome-specific engager follow a ZIP distribution and g_k is the log link function. We assume piecewise linear temporal trends with a single knot at panel 3 in the structural zero and mean components of the model.

Preliminary results showed a strong positive correlation between the subject-specific random effects in the structural zero and mean components of the model. Thus, a single subject-specific random effect is shared across outcomes in both the structural zero and mean components of the model. This model represents a substantial Watanabe-Akaike information criterion (WAIC, Watanabe 2010) improvement (of approximately 115) over the model with two independent (u_i and v_i in (2.4) and (2.5)) subject-specific random effects. MCMC methods for computing posterior samples from mixed effects models can have convergence issues when the variance of the random effects are near zero. This is the case for the between-subject variability for the probability of being a non-engager for *aggressive II* which is adequately captured by baseline covariates. Relative to the other outcomes, the proportion of subjects who reported never engaging in *aggressive II*, 21%, is low and, therefore, the probability of being a non-engager is low for the majority of subjects. The inclusion of gender in the non-engager component of the model effectively reduces corresponding between-subject variability for *aggressive II* to zero. For all of the outcomes, the shared effect seems to sufficiently characterize the variability in the structural zero component. Additionally, we fit the independence model with independent subject- and outcome-specific random effects (d_{ik} and b_{ik} in (2.4) and (2.5)) and examined the pairwise correlations of the random intercepts. Most of the pairwise estimates of the correlation coefficient corresponding to the mean component for *sold marijuana* and *sold other drugs* were close to zero (all below 0.3), indicating that essentially all of the

variability in the mean component of these drug-related outcomes appears to be absorbed in the term representing additional heterogeneity beyond the shared effect. Therefore, we set the relevant factor loading parameters equal to zero and consider the following model specification

$$\begin{cases} p_{ik} = \{1 + \exp(-\boldsymbol{w}_i'\boldsymbol{\gamma}_k - \lambda_{rk} r_i)\}^{-1} \\ \pi_{itk} = \exp[-\exp\{\boldsymbol{x}_{it}'(-\boldsymbol{\alpha}_k) + \rho_{11k} t + \rho_{12k}(t-3)_+ + \delta_k \log(z_{itk}) + \lambda_{uk} u_i\}] \\ \mu_{itk} = g_k^{-1}\{\boldsymbol{x}_{it}'\boldsymbol{\beta}_k + \rho_{21k} t + \rho_{22k}(t-3)_+ + \lambda_{vk} u_i + b_{ik}\} z_{itk} \end{cases} \tag{2.8}$$

$i = 1, \ldots, 1170$, $t = 1, \ldots, T_i$, $k = 1, \ldots, 8$ where $\lambda_{r6} = \lambda_{v2} = \lambda_{v3} \equiv 0$.

The vector of covariates associated with the probability of an outcome-specific non-engager, $\boldsymbol{w}_i$ consists of a fixed intercept, gender (male/female) and ethnicity (black/Hispanic/other). In the structural zero and mean components, $\boldsymbol{x}_{it}$ consists of an intercept, gender (male/female), ethnicity (black/Hispanic/other), a binary indicator of placement in a secure facility during panel t and a carry-over effect, defined as the proportion of the previous panel spent in a secure facility.

2.4.2 Computational Details

We assign weakly informative prior distributions for the fixed regression effects, $\boldsymbol{\gamma}_k \sim N_{q_1}(\boldsymbol{0}, \boldsymbol{I}_{q_1})$, $\boldsymbol{\alpha}_k \sim N_{q_2}(\boldsymbol{0}, \boldsymbol{I}_{q_2})$, $\boldsymbol{\beta}_k \sim N_{q_3}(\boldsymbol{0}, \boldsymbol{I}_{q_3})$, $\boldsymbol{\rho}_{1k} \sim N_2(\boldsymbol{0}, \boldsymbol{I}_2)$ and $\boldsymbol{\rho}_{2k} \sim N_2(\boldsymbol{0}, \boldsymbol{I}_2)$ $k = 1, \ldots, K = 8$, where $\boldsymbol{I}_n$ denotes an $n \times n$ identity matrix. For the factor loading parameters, λ_{rk}, λ_{uk}, λ_{vk} $k = 1, \ldots, 8$, we adopt moderately informative priors, $\Gamma(1, 1)$, initially, prior to setting some of these to zero in the model development. Feng and Dean (2012) utilized a similar prior specification in the context of a joint analysis of multivariate zero-heavy count outcomes. As well, we specify moderately informative $\Gamma(1, 1)$ priors to, δ_k $k = 1, 4, 5$ and 7, the shape parameter associated with the Weibull survivor function used to model the probability of a structural zero. Finally, we choose Unif$(0, 100)$ priors for the standard deviations of the outcome- and subject-specific random effects in the mean component, σ_{b_k} $k = 1, \ldots, 8$, because of the robust properties of this prior (Gelman, 2006).

The results below reflect two chains, each was run for an initial 10 000 burn-in iterations followed by an additional 40 000 iterations thinned at 40, resulting in a total of 2000 iterations to be used for posterior inference. In order to reduce the number of iterations needed and improve the mixing of the chains, we implement a hierarchical centering reparametrization (Gelfand, Sahu and Carlin, 1996) in the mean component of the model.

2.5 Analysis of Juvenile Offending Behaviour

The focus of this analysis is understanding the processes generating zero counts and assessing the carry-over effects of placement in a secure facility. Within the modeling framework, zero counts may arise from non-engagers, and for engagers, from either the structural zero or mean components of the model.

The posterior medians and 95% equal-tail credible intervals for the baseline covariate effects in the non-engager component are shown in the top row of Figure 2.1. We observed that for all outcomes, compared to male subjects, female subjects have a higher probability of being a non-engager. This effect is significant for all outcomes except *sold other drugs* and *aggressive I*. There are no significant differences in terms of the probability of being a non-engager among ethnicities except that, relative to the baseline group, black subjects have a higher probability of being a non-engager for *drove drunk*. As well, note that there are no significant differences in the probability of being a non-engager for any of the outcomes between black and Hispanic subjects.

The posterior medians for the outcome-specific trajectories for the probability of a structural zero and mean of the standard count distribution are displayed in Figure 2.2. In this figure, the fitted values correspond to a non-black, non-Hispanic male subject who spent no time in a secure facility in the previous or current panel. For illustration purposes, we assume a length of exposure of 365 days. For all the outcomes, the probability of a structural zero is increasing over time. However, the magnitude of this increase varies across outcomes, for example, the probability of a structural zero corresponding to *aggressive II* increases from 0.26 at panel 1

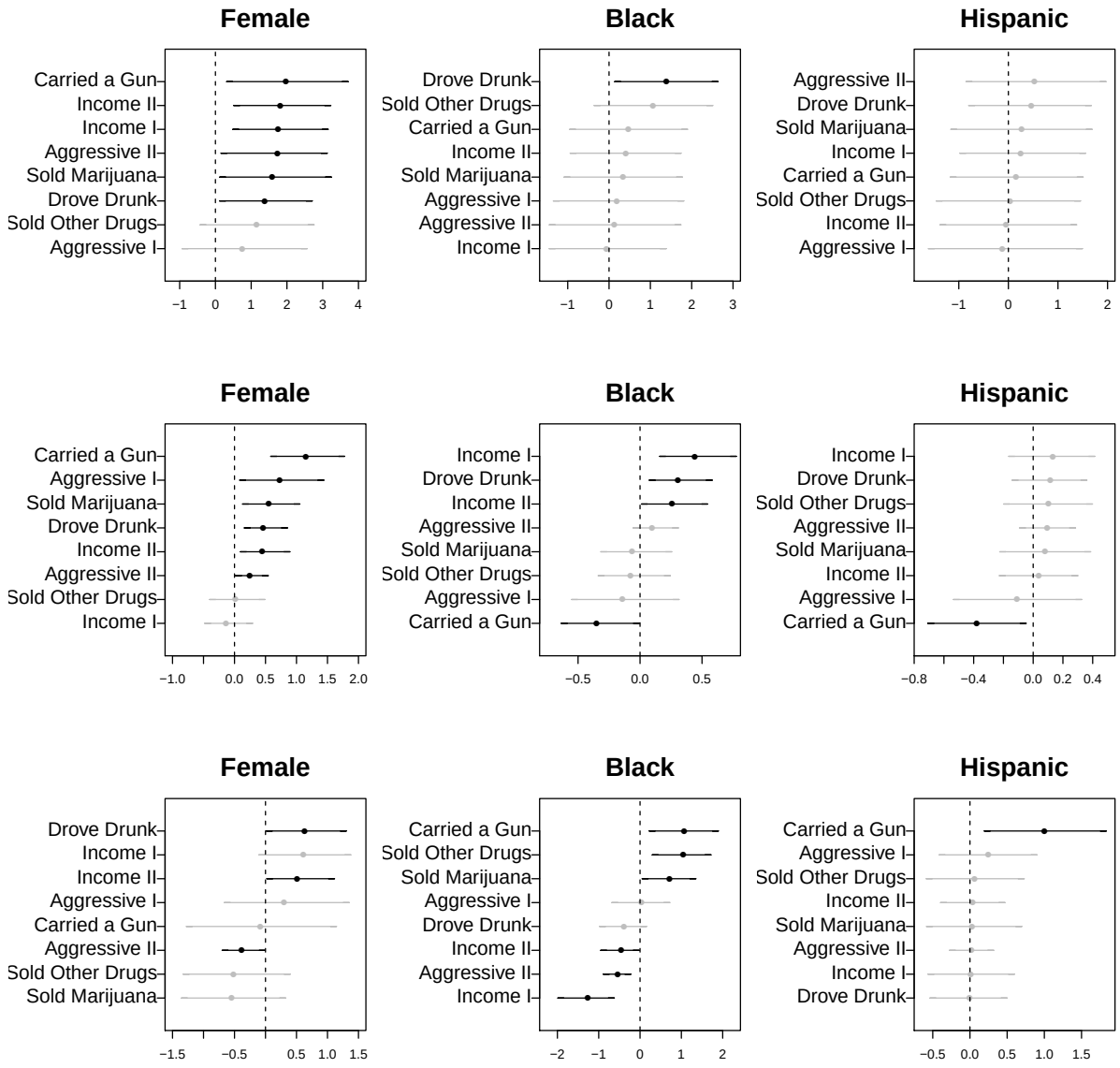

Figure 2.1: Posterior medians and 95% credible intervals for effects of baseline covariates on the probability of a non-engager ($\boldsymbol{\gamma}_k$, top), the probability of a structural zero ($-\boldsymbol{\alpha}_k$, middle), and the mean of the standard count distribution ($\boldsymbol{\beta}_k$, bottom). Credible intervals that exclude the null value of 0 are shaded darker.

to 0.74 at panel 7 while the probability corresponding to *aggressive I* increases from 0.89 to 0.95. There are three distinct types of trajectories associated with the mean of the standard count distribution: increasing over time (*carried a gun*, *sold marijuana*), increasing over panels 1 and 2 followed by a relatively constant mean (*sold other drugs*, *drove drunk*), or relatively constant and low. Overall, the proportion of subjects who are engaging in illegal or antisocial activity is decreasing over time. However, within the subgroup of individuals who continue to engage in illegal or antisocial activity, the frequency of this activity remains relatively constant or increases over time. This suggests that at the end of the seven year follow up period, the majority of subjects have low probability of offending but there exists a small subgroup of subjects whose rate of offending has remained constant or increased over the follow up period. As an example, consider *carried a gun* where the probability of a structural zero increases from 0.85 at panel 1 to 0.95 at panel 7 and, within the at-risk subgroup, the mean of number of days per year a subject carries a gun drastically increases over the follow up period from 3.9 to 38.8.

The posterior medians and 95% equal-tail credible intervals for the baseline covariate effects in the structural zero and mean components are shown in the middle and bottoms rows of Figure 2.1, respectively. Within the outcome-specific engagers, female subjects compared to male subjects have a significantly higher probability of a structural zero for all outcomes except *income I* and *sold other drugs*. Relative to the baseline group, black and Hispanic engagers have a lower probability of a structural zero for *carried a gun*. Additionally, black subjects compared to both the baseline group and Hispanic subjects have a higher probability of a structural zero for *income I*, *income II* and *drove drunk*. Turning to the mean component, relative to male subjects, female subjects who are at-risk to engage in an outcome have a significantly higher mean for *drove drunk* and *income II* and a significantly lower mean for *aggressive II*. Black and Hispanic subjects compared to the baseline group have a higher mean for *carried a gun*. As well, black subjects relative to both the baseline group and Hispanic subjects have a higher mean for *sold marijuana* and *sold other drugs* and a lower mean for *income I*, *income II* and *aggressive II*. Compared to Hispanic subjects, black subjects have a

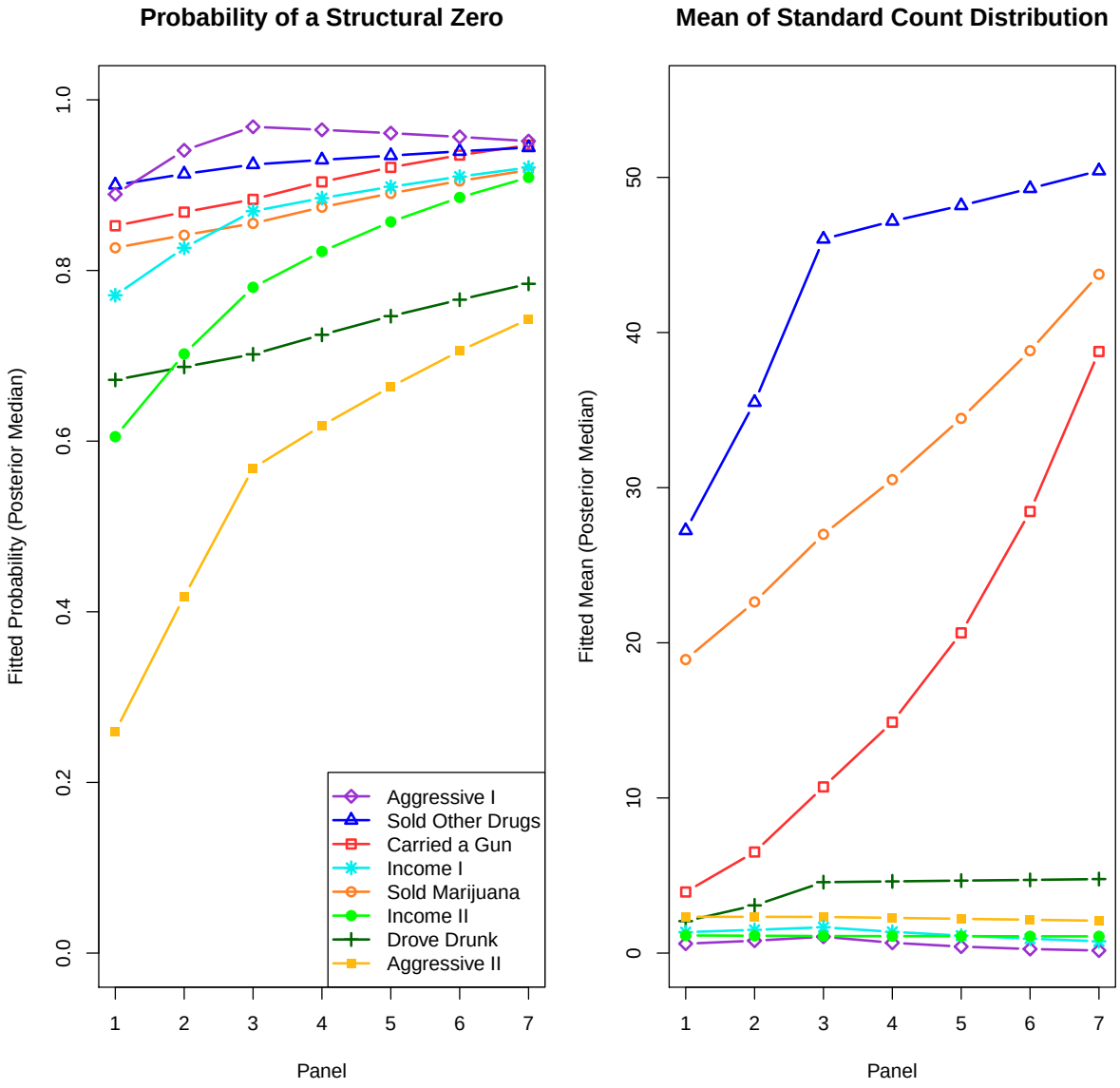

Figure 2.2: Fitted probability of a structural zero and fitted mean of the standard count distribution.

lower mean for *drove drunk*.

There are two effects related to placement in a secure facility. The first is an indicator effect on the probability of a structural zero and the mean count in the current panel. The second is the effect of the proportion of time in a secure facility in the previous panel on the probability of a structural zero and the mean count in the current panel (termed the carry-over effect). The posterior medians and credible intervals for the effect of placement in a secure facility on the probability of a structural zero and the mean of the standard count distribution in the current panel are shown in the top row of Figure 2.3. Corresponding posterior summaries for the carry-over effect are shown in the bottom row of Figure 2.3. For each of the eight outcomes, placement in a secure facility is associated with a lower probability of a structural zero in the current panel. With the exception of the interval corresponding to *income II*, the 95% credible intervals do not contain the null value of 0. Furthermore, placement in a secure facility is associated with higher mean in the current panel for all outcomes except *sold other drugs*. Overall, spending some time in a secure facility is associated with higher rates of offending in the current panel. Note that due to the panel structure of the data, whether placement occurs before or after criminal activity is unknown. It may be that a subject experienced a period of higher offending which led to placement in a secure facility. A higher proportion of the previous panel spent in a secure facility is associated with a higher probability of a structural zero for all of the outcomes in the current panel. Moreover, except for *aggressive I*, a higher proportion of the previous panel spent in a secure facility is associated with a lower mean for the standard count distribution in the current panel. Overall, a higher proportion of the previous panel spent in a secure facility is associated with lower offending in the current panel.

We utilize the survival function of a Weibull distribution to model the probability of a structural zero for *carried a gun*, *drove drunk*, *aggressive I* and *income I*; the corresponding posterior median estimates (95% credible interval) for the shape parameters are 0.29 (0.20, 0.38), 0.19 (0.10, 0.29), 0.01 (0.00, 0.05) and 0.05 (0.00, 0.13). These estimates are less than 1 which indicates that the hazard functions related to the waiting time until an event decreases

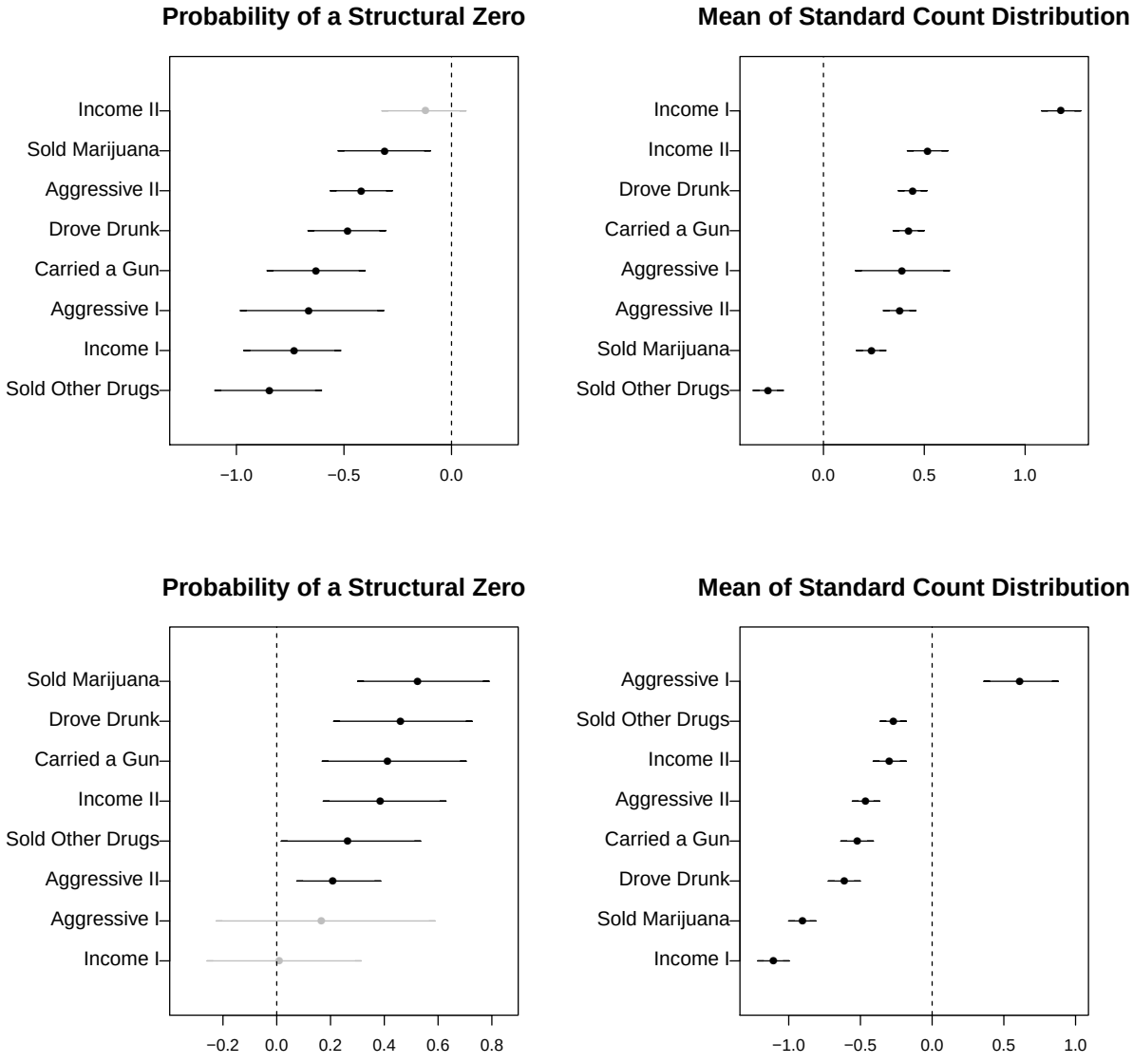

Figure 2.3: Panels in the top row display posterior medians and 95% credible intervals for effects of placement in a secure facility during the current panel on the probability of a structural zero (left) and mean of the standard count distribution (right); panels in the bottom row display posterior medians and 95% credible intervals for the carry-over effect on the probability of a structural zero (left) and mean of the standard count distribution (right). Credible intervals that exclude the null value of 0 are shaded darker.

as time in the community increases; the probability of an event in a fixed time interval in the future decreases as time in the community increases. Furthermore, the shape estimates corresponding to *aggressive I* and *income I* are very close to 0, suggesting that the length of time in the community does not substantially affect the probability of a structural zero for these outcomes.

Considering the subject-specific random effects in the non-engager component of the model, their estimated factor loading parameters vary substantially across the outcomes. This indicates that the between-subject variability for the probability of being a non-engager varies across the outcomes. In particular, this variability is lowest for *income I* and *income II* and highest for *sold marijuana* and *sold other drugs*. In the structural zero component, the estimated factor loading parameters for the subject-specific effects are fairly consistent across the outcomes with the exception of *aggressive II*. The factor loading parameter for *aggressive II* seems to be distinctly smaller; the between-subject variability for the probability of a structural zero is lower for *aggressive II* than the other outcomes. Finally, in the mean component, the factor loading parameter corresponding to the subject-specific effects for *income I* seems to be distinctly larger. The variability of the shared random effect in the mean component of the model is larger for *income I* than the other outcomes. The posterior medians and credible intervals for the factor loading parameters associated with each of the model components are displayed in Figure A.2.

In the top portion of Table 2.1, we display the pairwise estimates of Spearman's rank correlation coefficient for the posterior median estimates of the outcome- and subject-specific random intercepts in the mean component, b_{ik}. Most of the pairwise estimates of the correlation coefficient are close to zero, indicating that shared random effect adequately captures the correlation structure. However, there is evidence of weak positive pairwise correlations of b_{ik} between *carried a gun* and *aggressive I* and *sold marijuana* and *sold other drugs*. This indicates that subjects who report a high count for *carried a gun* tend to report a high count for *aggressive I* which is expected as several of the activities included in *aggressive I* involve

using a gun. An analogous interpretation holds for *sold marijuana* and *sold other drugs*. Also displayed in Table 2.1 is posterior medians for the variance of the random effect representing additional heterogeneity beyond the shared random effect in the mean component, $\sigma^2_{b_k}$. This variance is substantially larger for *carried a gun*, indicating there is large variation in the Binomial mean for *carried a gun* across subjects, distinct from the other outcomes. In the mean component, the variability of all outcomes is decomposed into one common error term that is linked to the structure zero component and, additionally, outcome-specific variability. For each run of the MCMC samples, the empirical variances for the random intercept and common component, $s^2_{b_{ik}+\lambda_{vk}u_i}$ and $s^2_{\lambda_{vk}u_i}$, respectively, are calculated. The fraction of variability explained by the common factor is calculated as the ratio $s^2_{\lambda_{vk}u}/s^2_{b_{ik}+\lambda_{vk}u_i}$. In the final row of Table 2.1, we display the posterior medians for the fraction of variability explained by the common factor for each outcome. The shared random effect accounts for 12% to 48% of the variability in the mean component; some of the variability in the mean component is absorbed by the shared random effect. However, for *carried a gun*, the vast majority of the variability in the mean component is absorbed in the term representing additional heterogeneity beyond the shared random effect, indicating that some latent factors may have distinct effects on this outcome.

Figures 2.4 and 2.5 examine the goodness of fit by comparing the observed counts versus those expected under the model for the structural zero and mean components. The observed trends in the differences between the number of zeros and the number of zeros due to non-engagers and the standard count distribution under the fitted model are overlain on the curves of expected number of structural zeros under the fitted model in Figure 2.4. The trends in the predicted counts of structural zeros follow the observed curves very closely. Figure 2.5 visually compares the trends in the mean of the standard count distribution under the fitted model and the mean of the observed counts, weighted by the inverse probability of the observation arising from the standard count component of the model; these trends are also in general agreement. However, the mean number of counts for *sold other drugs* appears to be consistently overestimated. This may be partially due to a small number of very frequent offenders and the

Table 2.1: Pairwise estimates of Spearman's rank correlation coefficient for posterior median estimates of b_{ik} in the mean component, posterior medians of σ_k^2 and the fraction of variability explained by the shared random effect in the mean component.

	Carried a gun	Sold marijuana	Sold other drugs	Drove drunk	Aggressive I	Aggressive II	Income I	Income II
Carried a gun	1.00	0.14	0.20	0.16	0.29	−0.02	0.02	0.07
Sold marijuana		1.00	0.34	0.10	0.01	0.02	0.04	0.10
Sold other drugs			1.00	0.03	0.03	0.01	0.01	0.01
Drove drunk				1.00	0.11	−0.05	0.06	0.03
Aggressive I					1.00	0.00	0.13	0.09
Aggressive II						1.00	0.01	0.07
Income I							1.00	0.14
Income II								1.00
$\sigma^2_{b_k}$	7.93	4.95	5.17	3.25	2.27	0.92	3.14	1.58
% variability	0.12	− − −	− − −	0.31	0.35	0.45	0.41	0.48

influence of these individuals warrants further investigation.

2.6 Comparison with Alternate Models

We investigate the benefits, above that provided by less complex models, obtained by adopting our mixture model approach for the excess zeros as well as by considering the outcomes jointly rather than modeling each outcome separately. We compared the following models:

Three Component Joint Model: This is the model we used in the analysis

Two Component Joint Model: Three component joint model without non-engager component; i.e., $p_{ik} \equiv 0$.

Separate Model: Three component joint model with $\lambda_{rk} = \lambda_{uk} = \lambda_{vk} \equiv 0$, $k = 1, \ldots, K$ and $\pi_{itk} = \exp[-\exp\{\boldsymbol{x}_{it}^{'}(-\boldsymbol{\alpha}_k) + \rho_{11k}t + \rho_{12k}(t-3)_+ + \delta_k \log(z_{itk}) + d_{ik}\}]$

As measures of comparison, we use the deviance information criterion (DIC, Speighlhalter et al., 2002) and the WAIC. DIC is defined as $\overline{D(\theta)} + \mathrm{p}_D$, where $\overline{D(\theta)}$ is the posterior mean of the deviance. The penalty term p_D is the effective number of model parameters defined by $\mathrm{p}_D = \overline{D(\theta)} - D(\bar{\theta})$ where $\bar{\theta}$ is the posterior mean of θ. WAIC, defined as

$$\mathrm{WAIC} = -2 \sum_{i=1}^{N} \sum_{t=1}^{T_i} \sum_{k=1}^{K} \log \left\{ \frac{1}{R} \sum_{r=1}^{R} L\left(y_{itk} | \theta^{(r)}\right) \right\} + 2 \sum_{i=1}^{N} \sum_{t=1}^{T_i} \sum_{k=1}^{K} V_{r=1}^{R} \{\log L(y_{itk} | \theta^{(r)})\} \qquad (2.9)$$

where $V_{r=1}^{R}$ represents the sample variance and θ denotes the collection of parameters in the model, may be used as a fast and computationally-convenient alternative to cross-validation. Models with lower values of DIC and WAIC are preferred.

The DIC and WAIC values are 129 853 and 57 313 for the three component joint model, 124 009 and 57 058 for the two component joint model and 166 905 and 64 197 for the separate model. The two component joint model seems to provide the best fit according to both measures of fit. Practically, it is hard to distinguish between the fits of the two and three component joint models. The primary difference between these models is at the interpretation level.

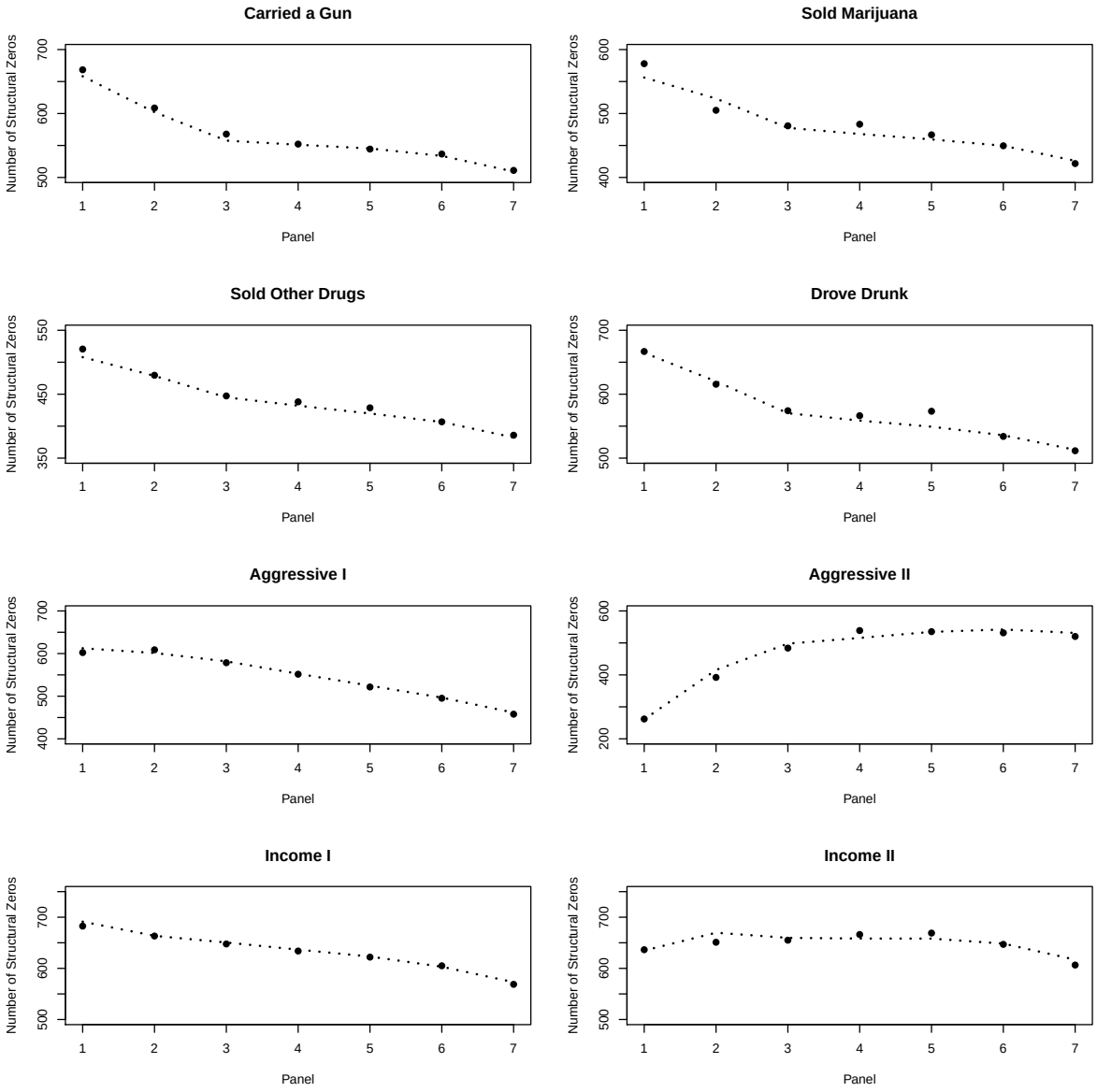

Figure 2.4: The expected number of structural zeros over time (lines) and the observed number of zeros minus the number of expected non-engagers and the expected number of zeros arising from the standard count distribution (points).

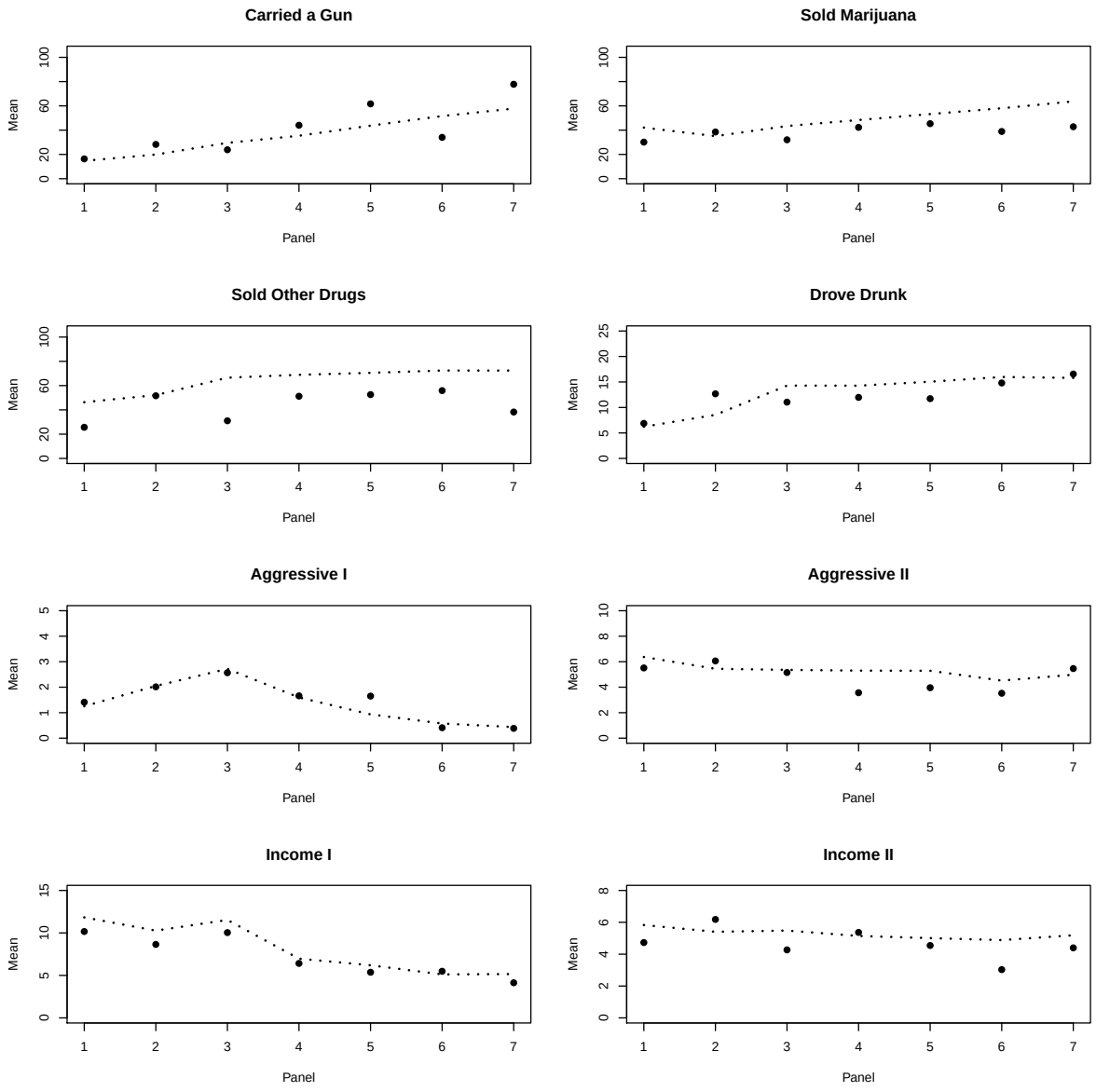

Figure 2.5: The expected mean of the standard count distribution over time (lines) and the mean of the observed counts, weighted by the inverse of the probability of the observation arising from the standard count distribution (points).

Under the two component model we are not able to distinguish between subjects who are not at-risk to engage in a particular outcome and subjects corresponding to a high probability of a structural zero, possibly due to limited exposure. In this application, it is of particular interest to identify subjects who are non-engagers and, hence, not at-risk for offending. Balancing various criteria of model fit and interpretation, the three component model seems most appropriate for modeling this data set. Also, the joint models yield substantially better fits than the separate model. Note that only fixed effects can be included in the model for the probability of a non-engager under the separate model. Linking these probabilities across outcomes allows for the estimation of subject-specific random effects in this component of the model. Therefore, considering the outcomes jointly allows us to account for correlations across outcomes in each of the model components and incorporate additional flexibility in the model for the probability of a non-engager through the inclusion of subject-specific random effects.

Our Bayesian framework provides various measures of subject-specific predictions. In the current context, we may be interested in predicting which subjects will not offend during the next year given that they spend the entire year in the community. We calculate the posterior medians for the probability of not offending during panel T_i+1, assuming the length of exposure is 365 days and no time is spent in a secure facility, across the competing models and display the results in Figure A.3. This probability is expressed as

$$P(y_{it} = \mathbf{0_{K\times 1}}|\boldsymbol{\Theta}, r_i, u_i, b_{ik}) = \prod_{k=1}^{K} [p_{ik} + (1 - p_{ik})\{\pi_{itk} + (1 - \pi_{itk}) f_k(0|\mu_{itk})\}] \tag{2.10}$$

$i = 1, \ldots, 1170$, $t = T_i + 1$. These estimates are very similar for the joint models. Relative to the joint models, the range of these fitted probabilities across all the individuals is substantially narrower under the separate model. Here, considering the outcomes separately affects the detection of individuals with an extreme (very low/very high) probability of not offending in the next year. This may have useful implications from a decision-making perspective, for

example, we may categorize subjects into two groups based on their risk of offending in the next year. A reasonable decision rule would be to select a probability threshold, say 0.8, and classify subjects with a probability of not offending during the next year greater than that threshold as low risk. Decisions concerning the placement of a subject in a secure facility versus enrollment in a community-based treatment would surely differ for low and high risk subjects. Under the two and three component joint models, 372 and 376 subjects, respectively, have an estimated probability of not offending during the next panel that exceeds 0.8. By contrast, only 55 subjects have an estimated probability of not offending in the next panel that exceeds 0.8 under the separate model. However, decision making in this context is complicated and is influenced by factors beyond past behaviour.

Finally, we wish to investigate the accuracy of predictions obtained by modeling these outcomes jointly compared to fitting them separately. We remove last panel of available data and fit joint and separate models using this reduced data set. Then, based on the posterior samples of the rth MCMC iteration, we generate a vector of predicted responses at last panel for each subject, $(y^{(r)}_{iT_i1}, \ldots, y^{(r)}_{iT_iK})'$, $r = 1, \ldots, R$. We calculate the sum of absolute deviation as

$$\sum_{k=1}^{K} \sum_{i=1}^{N} |y_{iT_ik} - \widehat{y}_{iT_ik}| \tag{2.11}$$

where $\widehat{y}_{iT_ik} = \frac{1}{R} \sum_{r=1}^{R} y^{(r)}_{iT_ik}$. The sum of absolute deviation is 60 724.18 for the three component joint model, 60 290.21 for the two component joint model and 66 942.80 for the separate model. We obtain more accurate predictions by modeling the outcomes jointly. We visually compare the distributions of the residuals, $y_{iT_ik} - \widehat{y}_{iT_ik}$, under the three component joint model and the separate model in Figure A.4. Under both models, the distribution of residuals is skewed to the right, indicating that the predicted counts tend to be overestimated. The geometric shape of the residual distributions arises from the fact that the observed responses are counts while the average predicted responses take continuous values. The joint model provides more accurate predictions as the median of the residuals is closer to zero under the joint model than that of the separate model for all of the outcomes. The shift in location of the median is

substantial for *carried a gun*, *drove drunk*, *sold marijuana* and *sold other drugs*.

2.7 Discussion

In this paper, we present a general framework for joint modeling of multiple longitudinal zero-inflated count outcomes which incorporates a variety of probabilistic structures on the zero counts. In particular, we accommodate a subgroup of subjects who are not at-risk to engage in a particular outcome and incorporate the effect of a time-dependent exposure variable in settings where some outcomes are prohibited during exposure to a treatment.

In the context of our motivating example, our three component mixture joint modeling approach enables a clearer understanding of offending patterns than the less complex alternative models considered. Compared to the joint models, considering the outcomes separately impacts the detection of subjects with an extreme probability of offending in a subsequent year and leads to less accurate predictions. On the other hand, it is hard to distinguish between the fits of the two and three component mixture joint models. The primary difference between these models is that under the three component mixture model we are able to identify subjects who are not at-risk for offending. Importantly, the analysis of the three component mixture model identifies differences across gender and ethnicity in terms of the probability of being a non-engager.

In our analysis, the use of the log-log link function in the structural zero component accommodates the presence of high incidence of zeros. In settings where the proportion of zeros exceeds 80 %, traditional ZIP models with symmetric link functions may struggle to explain the high prevalence of zeros, especially to identify important covariates (Ghosh et al., 2012). The three component mixture model may be particularly useful in such settings as under the ZIP (ZIB) models, the non-engager component of the model reduces the proportion of zeros fitted and such large percentages of zeros are hard to accommodate. A comparison with the analysis where the probability of a structural zero is modeled using a logistic link function, and, where the non-engager component is omitted, would be useful here to provide further insights

in this regard.

One potential issue with the proposed approach is that short (possibly zero) lengths of exposure can obscure the distinction between non-engagers and, engagers with a high probability of a structural zero due to limited exposure. However, in our analysis, only three subjects spent no time in the community over their follow up period. Additionally, in the proposed framework, the model for the probability of being an outcome-specific non-engager is linked across outcomes and, therefore, incorporates information from outcomes not prohibited in a secure facility. Caution must be taken when applying the proposed model in studies where the length of exposure may be zero or near zero across all outcomes.

Some alternatives to our modeling in Section 2 should be mentioned. In the Pathways to Desistance study, followup interviews were conducted approximately every 6 months for the first three years and every 12 months for the final four years. For convenience, we considered approximately annual data. However, accommodating irregularly spaced followup times is straightforward mathematically and would require some additional computational algorithmic developments. Incorporating such flexibility is underway.

Our analysis indicates that a higher proportion of a panel spent in a secure facility is associated with lower offending in the subsequent panel. Within the current framework, it is unclear whether this desistance is temporary or permanent. More complex models concerning the longer-term impact of placement in a secure facility on offending patterns could be investigated. Shen and Cook (2014) describe a dynamic mover-stayer model for recurrent event processes in settings where the underlying condition generating the recurrent events may resolve. Adopting this basic model structure by incorporating a time-dependent indicator variable corresponding to non-engager status of a subject which permits a switch from engager to non-engager sometime during the follow up period may be useful for differentiating between temporary and permanent changes in offending. Additionally, an investigation of the carry-over effects associated with placement in a juvenile versus an adult facility has been initiated.

More flexible correlation structures for the random effects could be implemented. In par-

ticular, as an alternative to shared frailties, correlated random effects that follow multivariate normal distributions could be utilized. In our analysis, the moderate pairwise correlations of b_{ik} between some of the outcomes indicates that models with a more flexible correlation structure may be useful. However, computationally efficient estimation of a covariance matrix for correlated random effects is challenging, especially in higher dimensional settings. As well, a copula function could be used to link separate sets of random effects. The shared frality framework utilized here is a special case of the Gaussian copula with a restricted correlation matrix assuming pairwise correlations equal to 1 and Gaussian marginals. Exploring different dependence structures through the use different copula functions warrants further research. Another useful extension would be to allow the random effects in the longitudinal components to evolve through time using an autoregressive structure.

Self-reported counts are often subject to heaping where recorded counts are rounded off to different levels of precision. Indeed, the histograms of non-zero counts corresponding to *carried a gun*, *sold marijuana*, *sold other drugs* and *drove drunk* exhibit heaps at multiples of 5, 10 and 30. Existing models for heaped zero-heavy count data (Wang and Heitjan, 2008) may be adapted to more complex scenarios concerning longitudinal data for multiple outcomes to examine the impact of heaping in Poisson-type analyses. In such settings, specifying the model for the heaping behaviour is complicated, for example, some outcomes may always be reported with the same level of precision and for other outcomes the level of reporting precision may vary by subject. This is an important topic for future investigation.

References

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

Blumen, I., Kogan, M., and McCarthy, P. J. (1955) *The industrial mobility of labor as a probability process*. Cornell Studies of Industrial and Labor Relations, volume 6. Ithaca, New York: Cornell University Press.

Boone, E. L., Stewart-Koster, B., and Kennard, M. J. (2012). A hierarchical zero-inflated Poisson regression model for stream fish distribution and abundance. *Environmetrics* **23**, 207-218.

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

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 355-366

Feng, C. X., and Dean, C. B. (2012). Joint analysis of multivariate spatial count and zero-heavy count outcomes using common spatial factor models. *Environmetrics* **23**, 493-508.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1996). Efficient parametrizations for generalized linear mixed models. *Bayesian Statistics* **5**, 48-74.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515-534.

Ghosh, S., Gelfand, A. E., Zhu, K., & Clark, J. S. (2012). The k-ZIG: flexible modeling for zero-inflated counts. *Biometrics* **68**, 878-885.

Ghosh, P., and Tu, W. (2008). Assessing sexual attitudes and behaviors of young women: a joint model with nonlinear time effects, time varying covariates, and dropouts. *Journal of the American Statistical Association* **103**, 1496-1507.

Guo, X., and Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event

time data using standard computer packages. *The American Statistician* **58**, 16-24.

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

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

Lee, A. H., Wang, K., and Yau, K. K. (2001). Analysis of zero-inflated Poisson data incorporating extent of exposure. *Biometrical Journal* **43**, 963-975.

Mulvey, E. P., Steinberg, L., Fagan, J., Cauffman, E., Piquero, A. R., Chassin, L., et al. (2004). Theory and research on desistance from antisocial activity among serious adolescent offenders. *Youth Violence and Juvenile Justice* **3**, 213-236

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, Austria.

Rodrigues-Motta, M., Pinheiro, H. P., Martins, E. G., Araujo, M. S., and dos Reis, S. F. (2013). Multivariate models for correlated count data. *Journal of Applied Statistics* **40**, 1586-1596.

Schubert, C. A., Mulvey, E.P., Steinberg, L., Cauffman, E., Losoya, S., Hecker, T., Chassin, L., et al. (2004). Operational lessons from the Pathways to Desistance project. *Youth Violence and Juvenile Justice* **3**, 237-255

Shen, H., and Cook, R. J. (2014). A dynamic Mover-Stayer model for recurrent event processes subject to resolution. *Lifetime Data Analysis* **20**, 404-423.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 583-639.

Wang, H., and Heitjan, D. F. (2008). Modeling heaping in self-reported cigarette counts. *Statistics in Medicine* **27**, 3789-3804.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely

applicable information criterion in singular learning theory. *The Journal of Machine Learning Research* **11**, 3571-3594.

Zeng, L., and Cook, R. J. (2007). Transition models for multivariate longitudinal binary data. *Journal of the American Statistical Association* **102**, 211-223.

Chapter 3

Analyzing Heaped Counts Versus Longitudinal Presence/Absence Data in Joint Zero-inflated Discrete Regression Models

3.1 Introduction

Joint outcome recurrent event data arise when events generated by two or more processes may occur repeatedly over a period of observation. In practice, the exact event times may not be readily observed but aggregate responses such as the number of events over the observation period or the presence/absence of events between periodic assessments are recorded. This is the case for the Pathways to Desistance study (Mulvey et al., 2004; Schubert et al., 2004), a major study of criminal behaviour patterns where the available data pertaining to several types of offending consist of both aggregate count data over the period of observation, as well as binary data recording the presence/absence of events repeatedly collected at each month of observation.

A potential issue with the use of the aggregate count data is that self-reported counts are often subject to heaping where recorded counts are rounded to different levels of precision. In the context of a study of smoking behaviour where the responses were zero-inflated counts, Wang and Heitjan (2008) described an example in which heaping attenuated the treatment effect by 20%. Given this concern, the monthly presence/absence data could be used instead to draw inference. However, there is data loss in such an analysis which elicits alternate concerns about loss of precision (Cameron and Trivedi, 2013 Section 3.6). These concerns would be reduced when occurrence rates are low, so presence/absence data provides much of the information available. The utility of both of these types of data records depends on a variety of context-specific factors and, therefore, it is difficult to determine where the use of one type is preferable over the other. This is exemplified in our motivating example where the role of two key features of the data, gender-specific propensities for rounding and occurrence rates that differ substantially from outcome to outcome, warrants investigation.

Heaping is a well-known problem in many applied contexts, particularly those involving retrospective collection of self-reported data. It has been postulated that when Edmond Halley published his Breslau life table in 1693, he grouped some of the reported ages at death due to heaping at multiples of 5 (Bellhouse, 2011). Hence this topic has been of interest to researchers for centuries. Some recent examples include the number of menstrual cycles (Ridout and Morgan, 1991), number of drug partners (Roberts and Brewer, 2001), number of sexual partners (Crawford, Weiss and Suchard, 2015), cigarette use (Wang and Heitjan, 2008) and age at smoking cessation (Bar and Lillard, 2012). Heitijan and Rubin (1991) provided a general framework for heaped data by introducing the concept of data coarsening, in which observations are made only on a subset of the sample space of the response variable. They established conditions under which the stochastic nature of the coarsening mechanism can be ignored and the data can be validly analyzed as group data. In particular, if the data are heaped at random and the parameters of the underlying response and the heaping mechanism are distinct, the coarsening mechanism is ignorable. In the case where heaping depends on the true underly-

ing response, several authors have modeled the latent true response and the heaping behaviour using a mixture model framework. For example, suppose a true count for each subject arises from a distribution with mass function $f(y|\Theta)$ that depends on parameters Θ and that a subject reports a count y^* from a heaping distribution with mass function $f(y^*|y,\rho)$ that depends on the true count y and parameters ρ. The likelihood contribution for an observed value y^* is

$$L(y^*|\Theta,\rho) = \sum_y f(y^*|y,\rho) f(y|\Theta) \tag{3.1}$$

Several authors utlilize this mixing framework in specific applications. Wang and Heitjan (2008) formulated a model for the analysis of heaped cigarette counts in which the probabilities of reporting truthfully and misreporting at different heaping grids is modeled using a proportional odds model. In the longitudinal setting, Wang et al (2012) included a subject-specific random effect in the proportional odds model for heaping behaviour to incorporate between-subject differences in heaping propensity. Crawford, Weiss and Suchard (2014) relaxed the assumption that misreported responses can only take specified grid values. They proposed a novel heaping distribution based on a general birth-death process where specially defined jumping rates ensure that the Markov chain is attracted to heaping grid points. This process accommodates quasi-heaping to values near but not equal to heaping grid points.

Zero-inflated models have been developed for a variety of settings including count data (Lambert, 1992; Hall, 2000; Yu 2008) and continuous data (Olsen and Schafer, 2001; Tooze et al., 2002). These models utilize a mixture model approach to handle the excess zeros, specified as a mixture of a point mass at zero and a specified distribution, e.g., Gaussian, Poisson or binomial. If the support of the specified distribution includes zero, then zero values may arise from either the point mass at zero, termed structural zeros, or as a realization of 0 from the specified distribution, referred to as random zeros.

Similarly, for longitudinal presence/absence data, we may observe a zero response vector for some subjects. In order to account for a high proportion of subjects who never experience an event, Carlin et al. (2001) proposed a mixture model for longitudinal binary data in which

each subject may be either at-risk or not at-risk for an event. Within the at-risk group, the probability of an event is modeled by a mixed logistic regression model.

Methods for the joint analysis of several count and zero-inflated count outcomes have been recently developed. Rodrigues-Motta et al. (2013) proposed a joint model for overdispersed count data where correlation among observations for the same subject is incorporated through the inclusion of correlated outcome- and subject-specific random effects in the mean component. Feng and Dean (2012) discussed joint models for spatial count data with excess zeros, where two outcomes are linked through a shared latent spatial random risk term. In this paper, we utilize the general framework proposed by Dunson (2000) in which, conditional on random effects, different members of the exponential family are used to describe the component models in the joint distribution of the set of observed outcomes.

In order to reduce the burden of data collection on respondents and limit recall error, self-reported data on recurrent events are sometimes recorded as binary responses indicating presence/absence of events or response categories defined by collapsed or grouped count data (0 events, 1-5 events, etc.), leading to partial observation of the underlying counting process. Despite this, little research has focused on developing methods for recurrent event studies with partial observation. Matsui and Miyagishi (1999) discussed the design of clinical trials in which periodic monitoring records whether or not recurrent events occurred. In their analysis, the required number of patients to achieve a specific power for the analysis of presence/absence data recorded every 6 months was not substantially greater than that required to achieve the same power when analyzing exact event times, clearly more so when the baseline event rate was low. McGinley, Curran and Hedeker (2015) considered settings where an underlying count outcome is measured using an ordinal scale and each response category represents a specified range of counts. Through simulations they demonstrated that the analysis of ordinal data defined by grouped counts can accurately recover parameters of the underlying count distribution. Furthermore, in their simulations, there was little loss of precision for the parameter estimates when ordinal data were analyzed instead of aggregate count data, even in the presence of zero-

inflation and overdispersion.

This article focuses on the comparison of the analysis of aggregate heaped count data and longitudinal presence/absence data using joint zero-inflated discrete regression models. Major objectives are the illustration of how heaping can introduce bias, impeding the identification of important risk factors and of determining in which situations the efficiency obtained from the analysis of longitudinal binary data is high, depending on the partition of the time for the presence/absence records and the underlying rate of events. The remainder of this article proceeds as follows. In Section 3.2, we provide a description of the Pathways to Desistance study and identify patterns of heaping observed in this data set. In Section 3.3, we describe joint models for zero-inflated recurrent event data with periodic monitoring and outline relevant heaping distributions which seem prevalent in the context of the study of criminal behaviour. Section 3.4 highlights the differences in inference based on the joint analysis of the aggregate count data and that of the monthly presence/absence data in our motivating data set. A simulation study using the heaping distributions suggested by the criminal behaviour data, contrasts the analysis of heaped count data to the analysis of accurate longitudinal presence/absence data. In Section 3.6, we indicate how one may implement the methodology used in our simulation study to inform decisions concerning the design of recurrent event studies where heaping is a concern. We conclude with a discussion of results and limitations.

3.2 A Study of Antisocial Behaviour Among Serious Juvenile Offenders

We introduce the motivating context as it shapes the model development. Here, we consider the analysis of data on criminal behaviour from a major study of juvenile offenders. The Pathways to Desistance study is a longitudinal study of a group of serious juvenile offenders investigating offending patterns in the period following court adjudication. Our data consist of 1170 youth offenders, aged 14 through 17 years old, who were found guilty of at least one

serious offense in the metropolitan areas of Phoenix, Arizona or Philadelphia, Pennsylvania. Subjects were enrolled in the study between 2000 and 2003 and followed for up to 7 years. We analyze here data corresponding to approximately the first year of follow up. A primary aim of the study is to identify risk factors associated with desistance or escalation of criminal behaviour among serious juvenile offenders.

All subjects completed a baseline interview where information about background characteristics and previous offending was collected. A follow up interview was conducted approximately one year after the baseline interview. At this interview, data pertaining to antisocial and criminal activity in the period since the baseline interview were recorded. Subjects indicated the months in which they engaged in an antisocial or illegal activity and reported how many times they engaged in the activity during this approximately one year period. Therefore, the available data on offending consists of aggregate count data and repeatedly measured binary data recording presence/absence of events during each month of observation. During the observation period, subjects may have spent time in a facility with no access to the community, termed a secure facility, for example, while incarcerated. Data on placement in a secure facility and, if so, the length of time spent in a secure facility, are available monthly. In our analysis, we consider the joint analysis of two outcomes, drunk driving (DD) and aggressive offending (AGG). These two outcomes represent sharply different patterns of occurrence as DD is characterized by a high proportion of zeros counts (81%) and large variability among non-zero counts while AGG is characterized by a moderate proportion of zeros (38%) and relatively few large counts. Importantly, this allows us to contrast the utility of the two types data records under two distinct patterns of occurrence.

A complicating factor in the analysis of the data from the Pathways to Desistance study is that some of the criminal activities are highly unlikely to occur if the individual is in a secure facility. It is is not possible that a subject will engage in DD while in a secure facility and, for this analysis, DD is therefore prohibited in a secure facility. On the other hand, there is no such restriction for AGG. Therefore, the length of exposure, defined as the length of time a subject is

at-risk to engage in an outcome, varies across the two outcomes. Commonly, count outcomes are regulated by such an exposure variable, with the length of exposure being proportional to the expected counts. Baetschmann and Winkelmann (2013) extended the general framework for incorporating an exposure variable in a count analysis to consider how exposures should be handled in the analysis of zero-inflated outcomes. In their approach, a structural zero occurs if the waiting time until an event exceeds the exposure time. This means that the probability of a structural zero is equal to the survivor function of the waiting time distribution evaluated at the exposure time. Using this approach, we incorporate the length of exposure in the structural zero as well as the Poisson components of zero-inflated discrete regression models.

From the distribution of observed non-zero counts for DD, displayed in Figure 3.1, we see evidence of heaping not only at 30, 60 and 90 (representing approximately one, two and three months, respectively) but also at 10, 20, 40, 50 and 80 and to a lesser extent at 5, 15 and 25. Subjects tend to report multiples of five, 10 or 30 and the reported data appear to be coarser as the number of events increases. On the other hand, there is little evidence of heaping for observed counts corresponding to AGG. Furthermore, there is evidence that the proportion of zeros counts that are accurately recorded differs for the two outcomes. For DD, the set of subjects who report a zero annual count is slightly smaller than the set of subjects who report no engagement during each month of observation; whereas the subjects who report a zero annual count for AGG coincide exactly with the subjects who report no engagement during each month of observation. It appears that these two outcomes are recorded with different levels of accuracy.

Another potential issue is that subjects may not be equally likely to under-report and over-report events. In their comparison of self-reported arrest data and official police records, Krohn et al (2013) concluded that adults are much less likely to over-report than under-report the number of arrests. As well, the propensity for rounding counts may differ by gender as a previous analysis of self-reported and official records of arrest data collected as part of the Pathways to Desistance Study (Piquero et al., 2014) observed gender differences in official

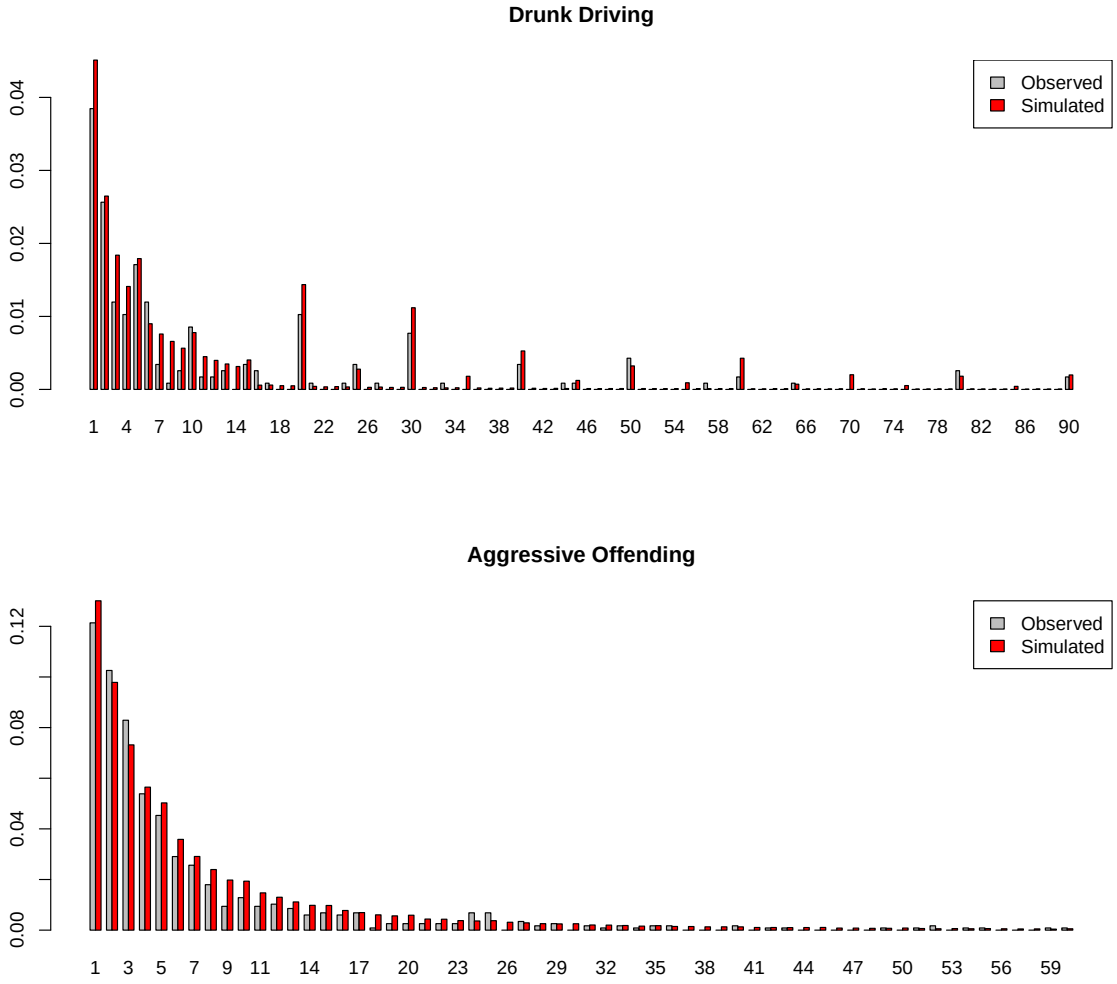

Figure 3.1: Comparison of distributions of non-zero counts between simulated rounded count data corresponding to heaping behaviour H_I, averaged over the 500 replicate data sets, and the Pathways to Desistance data.

arrests that were not accounted for by self-reported arrests. Earlier studies have also found gender differences in the validity of crime measures. Jolliffe et al (2003) compared official court referrals and self-report offending across gender and found higher concurrent validity among male subjects than female subjects. It is hypothesized that the processes related to offending patterns and to rounding differ for male and female subjects and as such gender is a key complicating risk factor in our analysis. Estimation of the effect of gender may be especially problematic in this situation. Hence, it is imperative to assess the impact of heaping which depends on gender in our motivating context.

3.3 Joint Models for Zero-inflated Recurrent Event Data with Periodic Monitoring

We consider a study with N subjects where data related to K outcomes, each corresponding to a recurrent event process, are collected. For each outcome, conditional on random effects, we assume that events arise according to a zero-inflated homogeneous Poisson process. That is, for outcome k, conditional on random effects, events corresponding to subject i arise from a Poisson process with intensity μ_{ik}, with probability $1 - \pi_{ik}$, and a degenerate process with intensity 0, with probability π_{ik}, where

$$\pi_{ik} = \exp\{-\exp(\boldsymbol{x}_{1i}^{'}\boldsymbol{\beta}_{1k} + \delta_k \log(z_{ik}) + \nu_k u_i)\} \tag{3.2}$$

is the survivor function of a Weibull distribution, and

$$\mu_{ik} = \exp(\boldsymbol{x}_{2i}^{'}\boldsymbol{\beta}_{2k} + \lambda_k u_i + b_{ik}) \tag{3.3}$$

Here, $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ are $q_1 \times 1$ and $q_2 \times 1$ vectors of covariates for the fixed effects while $\boldsymbol{\beta}_{1k}$ and $\boldsymbol{\beta}_{2k}$ are vectors of corresponding regression parameters. The form of π_{ik}, as well as the term $\delta_k \log(z_{ik})$ reflect the idea proposed by Baetschmann and Winkelmann (2013) to model

the probability of a structural zero as the survivor function of a Weibull distribution with shape parameter δ_k, evaluated at z_{ik}, the length of exposure during the observation period. Hence π_{ik} represents the probability that the waiting time until an event exceeds the length of exposure. The subject-specific random effect, u_i, is shared across outcomes and model components; ν_k is the factor loading for this shared effect on outcome k in the structural zero component and λ_k is the factor loading for this shared effect on outcome k in the Poisson component. The outcome- and subject-specific random effect for the Poisson component, b_{ik}, represents additional heterogeneity beyond the shared random effect. We assume the random effects are normally distributed such that $u_i \sim N(0, 1)$, without loss of generality since all ν_k and λ_k are not constrained; and $b_{ik} \sim N(0, \sigma^2_{b_k})$, $i = 1, \ldots, N$ and $k = 1, \ldots, K$.

3.3.1 Likelihood for Aggregate Zero-inflated Count Data

Let y^C_{ik} denote the number of events corresponding the the kth outcome and ith subject that occur during the period of observation. Conditional on random effects, u_i and b_{ik}, y^C_{ik} follows a zero-inflated Poisson (ZIP) distribution with probability of a structural zero π_{ik} and Poisson mean $z_{ik}\mu_{ik}$, proportional to the length of exposure. The likelihood can be expressed as

$$L(\boldsymbol{Y}^C|\boldsymbol{\Theta}, \boldsymbol{u}, \boldsymbol{b}) = \prod_{i=1}^{N}\prod_{k=1}^{K}\Bigg[\mathbf{I}(y^C_{ik} = 0)\left\{\pi_{ik} + (1-\pi_{ik})\exp(-z_{ik}\mu_{ik})\right\} + \mathbf{I}(y^C_{ik} \neq 0)\left\{(1-\pi_{ik})\frac{\exp(-z_{ik}\mu_{ik})(z_{ik}\mu_{ik})^{y^C_{ik}}}{y^C_{ik}!}\right\}\Bigg]. \tag{3.4}$$

3.3.2 Modeling Heaping in Zero-inflated Count Data

We assume that the observed number of events for subject i and outcome k, y^{*C}_{ik}, is generated from a heaping distribution with mass function $f(y^{*C}_{ik}|y^C_{ik}, \boldsymbol{\rho})$ that depends on the true count y^C_{ik} and parameters $\boldsymbol{\rho}$. Assuming the true count data arise from the likelihood given in (3.4), the likelihood for the heaped count data can be expressed as

$$L(\boldsymbol{Y}^{*C}|\boldsymbol{\Theta},\boldsymbol{u},\boldsymbol{b},\boldsymbol{\rho}) = \prod_{i=1}^{N}\prod_{k=1}^{K}\sum_{y_{ik}^{C}=0}^{\infty}\Bigg(f(y_{ik}^{*C}|y_{ik}^{C},\boldsymbol{\rho})\Bigg[\mathbf{I}(y_{ik}^{C}=0)\left\{\pi_{ik}+(1-\pi_{ik})\exp(-z_{ik}\mu_{ik})\right\} + \mathbf{I}(y_{ik}^{C}\neq 0)\left\{(1-\pi_{ik})\frac{\exp(-z_{ik}\mu_{ik})(z_{ik}\mu_{ik})^{y_{ik}^{C}}}{y_{ik}^{C}!}\right\}\Bigg]\Bigg). \tag{3.5}$$

The likelihood of the observed heaped count data (3.5) and the extent to which this differs from the likelihood of the true count data (3.4) depends on the heaping distribution. To illustrate the impact of different heaping mechanisms on inference, in our simulation study in Section 3.5, we consider four heaping distributions, denoted H_I to H_{IV}.

The heaping structures aim to reflect various motivations expressed in the literature and seen empirically in our data. As previously noted, gender differences in heaping behaviour as well as differences in under- and over-reporting may be a concern with this data set. Table 3.1 summarizes the four heaping distributions considered here in terms of (i) parameters representing heaping probabilities, (ii) change points for different levels of coarsening as well as (iii) whether rounded counts are the result of symmetrically rounding to the nearest heaping point or rounding down. For each outcome, true counts less or equal to κ are rounded to a multiple of m, $m \in \{5, 10, 30\}$, with probability $\rho_{1_m}^{M}/\rho_{1_m}^{F}$ for male/female subjects. Similarly, true counts greater than κ are rounded to a multiple of m with probabilities $\rho_{2_m}^{M}$ and $\rho_{2_m}^{F}$ for male and female subjects, respectively. The first heaping distribution, H_I, is motivated by insights arising from lengthy analysis of the patterns of heaping observed in the criminal behaviour data without accounting for gender differences. Figure 3.1 visually compares the distribution of non-zeros counts between simulated heaped count data generated according to H_I and the observed data; these distributions are in general agreement. Appendix B provides details on how this heaping distribution was selected. H_{II} links the probability of heaping with gender. H_{III} assumes that rounded counts are the result of subjects under-reporting the true number of events. Finally, H_{IV} incorporates gender differences in both the heaping probabilities and the direction of rounding.

Table 3.1: Summary of four heaping distributions.

	Heaping Parameter	H_I	H_{II}	H_{III}	H_{IV}
Drunk Driving					
	κ	14	14	14	14
	$\rho^M_{1_5}$	0.100	0.300	0.300	0.300
	$\rho^F_{1_5}$	0.100	0.100	0.300	0.100
	$\rho^M_{1_{10}}$	0	0	0	0
	$\rho^F_{1_{10}}$	0	0	0	0
	$\rho^M_{1_{30}}$	0	0	0	0
	$\rho^F_{1_{30}}$	0	0	0	0
	$\rho^M_{2_5}$	0.300	0.100	0.100	0.100
	$\rho^F_{2_5}$	0.300	0.300	0.100	0.300
	$\rho^M_{2_{10}}$	0.400	0.600	0.600	0.600
	$\rho^F_{2_{10}}$	0.400	0.400	0.600	0.400
	$\rho^M_{2_{30}}$	0.100	0.250	0.250	0.250
	$\rho^F_{2_{30}}$	0.100	0.100	0.250	0.100
Aggressive Offending					
	κ	2	–	–	–
	$\rho^M_{1_5}$	0	–	–	–
	$\rho^F_{1_5}$	0	–	–	–
	$\rho^M_{1_{10}}$	0	–	–	–
	$\rho^F_{1_{10}}$	0	–	–	–
	$\rho^M_{1_{30}}$	0	–	–	–
	$\rho^F_{1_{30}}$	0	–	–	–
	$\rho^M_{2_5}$	0.025	0.250	0.250	0.250
	$\rho^F_{2_5}$	0.025	0.025	0.250	0.025
	$\rho^M_{2_{10}}$	0	0	0	0
	$\rho^F_{2_{10}}$	0	0	0	0
	$\rho^M_{2_{30}}$	0	0	0	0
	$\rho^F_{2_{30}}$	0	0	0	0
Rounding rule					
	Male	round nearest	round nearest	round down	round nearest
	Female	round nearest	round nearest	round down	round down

In particular, under H_I, true counts for DD less or equal to $\kappa = 14$ are rounded to the nearest multiple of 5 with probability $\rho^M_{1_5} = \rho^F_{1_5} = 0.100$ and accurately reported otherwise. True counts for DD greater than 14, are rounded the nearest multiples of $\{5, 10, 30\}$ with probabilities $\{0.300, 0.400, 0.100\}$ and accurately reported otherwise. On the other hand, true counts for AGG less than or equal to 2 are accurately reported while true counts greater than 2 are rounded to the nearest multiple of 5 with probability 0.025 and accurately reported with probability 0.975. Relative to H_I, the probability of heaping for male subjects is greater under H_{II} whereas the heaping probabilities for female subjects are the same as those under H_I. As well, we allow true counts of 0,1 and 2 to be misreported for AGG as well as DD. Under H_{III}, the heaping probabilities for all subjects are the same as that specified for male subjects under H_{II}. Under this heaping distribution, misreported counts are the result of rounding down the nearest heaping point as opposed to symmetrically rounding to the nearest heaping point. The gender-specific heaping probabilities for H_{IV} are the same as those under H_{II}. Additionally, under this heaping distribution, there are gender differences in the direction of rounding in that female subjects round down to the nearest heaping point (under-report) while male subjects round to the nearest heaping point.

3.3.3 Joint Mixture Model for Longitudinal Presence/Absence Data

We consider situations where presence/absence of events between several periodic assessments is recorded. For each subject, let $0 = t_0 < t_1 < \cdots < t_{T_i}$ denote successive monitoring times. We assume here, for simplicity in presenting the likelihood, that the monitoring times are common for all subjects and equally spaced. This is true for out motivating data set. For subject i and outcome k, let y^B_{ijk} be the binary response at t_j, so that $y^B_{ijk} = 1$ if one or more events occurred between t_{j-1} and t_j and $y^B_{ijk} = 0$ otherwise; and $\boldsymbol{y}^B_{ik} = (y^B_{i1k}, \ldots, y^B_{iT_ik})'$ be the corresponding sequence of binary responses. Conditional on u_i and b_{ik}, $\boldsymbol{y}^B_{ik}$ can be viewed as arising from a mixture of a zero vector and a vector of independent responses drawn from a Bernoulli distribution. That is, conditional on u_i and b_{ik}, the binary response for subject i at

time t_j for outcome k will correspond to a structural zero with probability π_{ik} (Eq (3.2)) and will follow a Bernoulli(ζ_{ijk}) distribution with probability $1 - \pi_{ik}$ where

$$\zeta_{ijk} = 1 - \exp(-z_{ijk}\mu_{ik}) = 1 - \exp\{-z_{ijk}\exp(\boldsymbol{x}_{2i}^{'}\boldsymbol{\beta}_{2k} + \lambda_k u_i + b_{ik})\} \tag{3.6}$$

Here, z_{ijk} is the length of exposure for subject i between t_{j-1} and t_j for outcome k and $z_{ik} = \sum_{j=1}^{T_i} z_{ijk}$. The corresponding likelihood is given by

$$\begin{aligned} L(\boldsymbol{Y}^B|\boldsymbol{\Theta}, \boldsymbol{u}, \boldsymbol{b}) = \prod_{i=1}^{N}\prod_{k=1}^{K}\Bigg[& \mathbf{I}(\mathbf{y}_{ik}^B = \mathbf{0}_{T_i\times 1})\{\pi_{ik} + (1-\pi_{ik})\exp(-z_{ik}\mu_{ik})\} \\ & + \mathbf{I}(\mathbf{y}_{ik}^B \neq \mathbf{0}_{T_i\times 1})\Bigg\{(1-\pi_{ik})\prod_{j=1}^{T_i}\{1-\exp(-z_{ijk}\mu_{ik})\}^{\mathbf{I}(y_{ijk}^B=1)}\{\exp(-z_{ijk}\mu_{ik})\}^{\mathbf{I}(y_{ijk}^B=0)}\Bigg\}\Bigg] \end{aligned} \tag{3.7}$$

By comparing equations (3.4) and (3.7), we observe that the contribution to the likelihood for a subject with no events during the observation period is the same under the joint zero-inflated Poisson model and the joint mixture model for longitudinal presence/absence data. The magnitude of the loss of precision due to repeatedly recording presence/absence of events instead of aggregate counts depends on occurrence rate and the length of time between monitoring points. As the probability of observing more than one event between monitoring points decreases, due to a low event rate and/or frequent monitoring, the loss of precision will decrease to a possibly negligible level.

3.4 Analysis of Juvenile Offending Behaviour

The focus of this analysis is to examine differences in inference based on the analysis of the heaped aggregate count data and that of the monthly presence/absence data. For DD, we define exposure as number of days spent in the community and utilize the survivor function of a Weibull distribution to model the probability of a structural zero. For AGG, we define exposure

as the number of days under observation; we utilize the survivor function of an exponential distribution to model the probability of a structural zero and, hence, set $\delta_2 = 1$ as the length of the observation period takes only a few values. The fixed effects design matrices, $\boldsymbol{x_{1i}} = \boldsymbol{x_{2i}}$, consist of an intercept, gender (male/female) and a binary indicator of placement in a secure facility during the observation period.

The mixed joint models for aggregate zero-heavy count data and longitudinal presence/absence data may be implemented in a Bayesian framework using Markov chain Monte Carlo (MCMC) methods. The joint posterior distribution of the parameters is

$$p(\boldsymbol{\Theta}, \boldsymbol{u}, \boldsymbol{b}|\boldsymbol{Y}^D) \propto L(\boldsymbol{Y}^D|\boldsymbol{\Theta}, \boldsymbol{u}, \boldsymbol{b}) p(\boldsymbol{b}|\boldsymbol{\sigma}_b^2) \pi(\boldsymbol{\sigma}_b^2) p(\boldsymbol{u}) \pi(\boldsymbol{\Theta}) \tag{3.8}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\delta})'$, $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_{1k}, \ldots, \boldsymbol{\beta}_{1K})'$, $\boldsymbol{\beta}_2 = (\boldsymbol{\beta}_{2k}, \ldots, \boldsymbol{\beta}_{2K})'$,$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)'$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)'$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)'$, $\boldsymbol{\sigma}_b^2 = (\sigma_{b_1}^2, \ldots, \sigma_{b_K}^2)'$, $\boldsymbol{u} = (u_1, \ldots, u_N)'$ and $\boldsymbol{b} = (b_{11}, \ldots, b_{N1}, b_{12}, \ldots, b_{NK})'$. The first term on the right hand side of (3.8) is the likelihood based on either aggregate count data (superscripted by $D = C$) or longitudinal presence/absence data (superscripted by $D = B$).

The Bayesian model specification is made complete by assigning prior distributions to $\boldsymbol{\Theta}$ and $\boldsymbol{\sigma}_b^2$. In preliminary estimation runs using non-informative prior distributions there was some instability in the iterative process, where a chain attempted to switch from the two-component mixture to a reduced one-component model. Carlin et al. (2001) addressed a similar issue by utilizing $N(0, 1)$ priors for the fixed effects regression parameters and a normal prior with a non-zero mean and variance of 1 for the fixed intercept term. Following this, we assign weakly informative prior distributions for the fixed regression effects, $\boldsymbol{\beta}_{1k} \sim N_3(\boldsymbol{\mu}_P, \boldsymbol{I}_3)$, $\boldsymbol{\beta}_{2k} \sim N_3(\boldsymbol{\mu}_P, \boldsymbol{I}_3)$ where $\boldsymbol{\mu}_P = (-6, 0, 0,)'$ and $\boldsymbol{I}_n$ denotes an $n \times n$ identity matrix. Here, the prior mean for the fixed intercepts reflects for fact that our zero-inflated count model includes a term approximately equal to the logarithm of 365 days ($\log(365) = 5.9$) in both model components.

As well, we specify a moderately informative $\Gamma(1,1)$ prior to, δ_1, the shape parameter associated with the Weibull survivor function used to model the probability of a structural zero. For the factor loading parameters, ν_k, λ_k $k = 1, 2$, we adopt moderately informative priors, $\Gamma(1,1)$. Finally, we choose Unif$(0, 100)$ priors for the standard deviations of the outcome- and subject-specific random effects in the mean component, σ_{b_k}, $k = 1, 2$. Other prior distributions (for example, $\sigma^2_{b_k} \sim IG(1,1)$ and $\boldsymbol{\beta}_{2k} \sim N_3(\mathbf{0}, 2 \times \boldsymbol{I}_3)$) were explored in a sensitivity analysis with no substantial change to the results obtained.

Inference is then based on the posterior distribution, which can be summarized using samples drawn from that distribution. This framework for the analysis was implemented through the freely available software JAGS (Plummer, 2003). The posterior estimates and 95% equal-tail credible intervals, displayed in Table 3.2, reflect two chains, each was run for an initial 10 000 burn-in iterations followed by an additional 40 000 iterations thinned at 40, resulting in a total of 2000 iterations to be used for posterior inference.

3.4.1 Key Differences in Inference Between Count and Binary Data Records

Most of the estimates corresponding to AGG are very close for the two types of data records, whereas some of the parameter estimates corresponding to DD are substantially different for the analyses of the aggregate count data and the longitudinal presence/absence data. Using the equal-tail 95% credible intervals, the use of both types of data records identify female subjects, compared to male subjects, as having a lower Poisson mean for AGG. In the count data analysis gender is a to be significant effect in the structural zero component for DD, but this effect is non-significant under the analysis of the presence/absence data. The use of both types of data records identify placement in a secure facility as an important risk factor that contributes to a decreased probability of a structural zero for both outcomes. Placement in a secure facility is also associated with a higher Poisson mean for AGG in the analysis of both types of data. On the other hand, placement in a secure facility is only significant in the

Table 3.2: Posterior median and 95% credible intervals obtained from the analysis of the aggregate count data and longitudinal presence/absence data.

	Aggregate Count Data		Monthly Presence/Absence Data	
Parameter	Estimate	95% CI	Estimate	95% CI
Drunk Driving				
β_{11_0}	-5.08	$(-7.08, -3.03)$	-4.93	$(-6.69, -3.29)$
(female) β_{11_1}	-1.22	$(-2.20, -0.24)$	-0.66	$(-1.68, 0.53)^{ns}$
(secure facility) β_{11_2}	1.56	$(0.56, 2.58)$	2.28	$(1.41, 3.35)$
δ_1	0.51	$(0.19, 0.97)$	0.55	$(0.23, 0.86)$
ν_1	2.26	$(1.21, 4.13)$	2.32	$(1.36, 3.99)$
β_{21_0}	-6.95	$(-8.55, -5.66)$	-6.71	$(-7.77, -5.74)$
(female) β_{21_1}	0.01	$(-1.05, 1.11)^{ns}$	-0.75	$(-1.51, 0.13)^{ns}$
(secure facility) β_{21_2}	1.71	$(0.78, 2.58)$	0.54	$(-0.22, 1.28)^{ns}$
λ_1	2.03	$(0.91, 3.14)$	1.08	$(0.30, 2.11)$
$\sigma^2_{b_1}$	3.21	$(0.11, 6.96)$	2.03	$(0.30, 3.65)$
Aggressive Offending				
β_{12_0}	-5.28	$(-5.79, -4.18)$	-5.13	$(-5.66, -4.08)$
(female) β_{12_1}	-0.59	$(-1.57, 0.35)^{ns}$	-0.58	$(-1.61, 0.64)^{ns}$
(secure facility) β_{12_2}	0.73	$(0.04, 1.78)$	0.83	$(0.09, 1.81)$
δ_2	1.00		1.00	
ν_2	1.33	$(0.25, 3.19)$	1.35	$(0.51, 2.94)$
β_{22_0}	-5.73	$(-6.01, -5.45)$	-5.90	$(-6.12, -5.67)$
(female) β_{22_1}	-0.73	$(-1.13, -0.32)$	-0.67	$(-0.99, -0.32)$
(secure facility) β_{22_2}	0.70	$(0.41, 1.00)$	0.53	$(0.28, 0.79)$
λ_2	0.79	$(0.57, 1.23)$	0.73	$(0.49, 1.02)$
$\sigma^2_{b_2}$	1.28	$(0.47, 1.63)$	0.66	$(0.27, 0.95)$
ns = non-significant				

Poisson component for DD under the analysis of the aggregate count data. Importantly, then some of the covariate effects associated with DD are found to be significant in the analysis of the aggregate count data but non-significant in that of the monthly presence/absence data. We postulate explanations for these differences later and also explore characteristics of the data which lead to such differences in our simulation study.

Considering the subject-specific random effects in the structural zero component of the model, their estimated factor loading parameters are very similar for the two types of data records. The estimated factor loading parameters in the Poisson component are close for AGG under the two types of data records but the corresponding estimates for DD differ substantially, with analysis of the count data indicating much higher outcome-specific variability related to u_i. For both outcomes, the estimated variance of the outcome- and subject-specific random effect representing additional heterogeneity beyond the shared random effect in the Poisson component, $\sigma^2_{b_k}$, is larger but not significantly so, for the analysis of the aggregate count data compared to the longitudinal presence/absence data.

Overall, the differences between the posterior median estimates based on the analysis of the aggregate count data and analysis of the monthly presence/absence data are much smaller for AGG than for DD. This is consistent with the observation that there is strong evidence of heaping in the count data recorded for DD, whereas there is little to no evidence of heaping in the observed counts for AGG. It is unclear the extent to which heaping of aggregate count data may have introduced bias in this analysis.

3.5 Simulation Study

The conflicting conclusions concerning the covariate effects associated with DD for the two types of data records prompts an investigation of the corresponding study designs when heaping is expected. We note that for a homogeneous Poisson process, the analysis of aggregate counts is fully efficient when compared to an analysis of event times. Here, we view an analysis based on true count data, aggregated over the observation period, as the gold stan-

dard and contrast this with analyses based on aggregate heaped count data and longitudinal presence/absence data, through a simulation study. The goals of the study are to determine the sorts of biases which manifest in the analysis of heaped data, and to identify whether accurate presence/absence data along a partitioned longitudinal time scale could provide reasonably efficient estimates. Importantly, the focus of comparison is in the context of the criminal behaviour data. We consider the heaping distributions suggested by the criminal behaviour data and vary the frequency of the monitoring times at which presence/absence data are collected.

We simulate data corresponding to $N = 1000$ subjects from the joint model for zero-inflated aggregate count data with true parameters corresponding to the posterior medians from the analysis of aggregate count data in the criminal behaviour study. The design is specified by the covariate and exposure history of N randomly selected subjects from this study. At the rth replication, we generate

$$\boldsymbol{u}^{(r)} = (u_1^{(r)}, \ldots, u_N^{(r)})' \sim N(\mathbf{0}, \mathbf{I}) \text{ and } \boldsymbol{b_k}^{(r)} = (b_{1k}^{(r)}, \ldots, b_{Nk}^{(r)})' \sim MVN(\mathbf{0}, \sigma_{b_k}^2 \mathbf{I})$$

for $k = 1, 2$. We calculate the probability of a structural zero and Poisson intensity

$$\pi_{ik}^{(r)} = \exp\{-\exp(\boldsymbol{x}_{1i}'\boldsymbol{\beta}_{1k} + \delta_k \log(z_{ik}) + \nu_k u_i^{(r)})\} \text{ and } \mu_{ik}^{(r)} = \exp(\boldsymbol{x}_{2i}'\boldsymbol{\beta}_{2k} + \lambda_k u_i^{(r)} + b_{ik}^{(r)})$$

$i = 1, \ldots, N$, $k = 1, 2$. Then, we generate monthly count data where $y_{ijk}^{C(r)} \sim \text{ZIP}(\pi_{ik}^{(r)}, z_{ijk}\mu_{ik}^{(r)})$. We calculate monthly presence/absence data as $y_{ijk}^{B(r)} = 0$ if $y_{ijk}^{C(r)} = 0$ and $y_{ijk}^{B(r)} = 1$ otherwise and similarly derive bi-monthly and quarterly (every three months) presence/absence data. The true count data, aggregated over the period of observation, are $y_{ik}^{C(r)} = \sum_{j=1}^{T_i} y_{ijk}^{C(r)}$ and heaped aggregate count data, $y_{ik}^{*C(r)}$, generated according to heaping distributions H_I to H_{IV}, detailed in Section 3.3.2, are also summarized. We fit the joint model for aggregate zero-inflated count data using $y_{ik}^{C(r)}$ and $y_{ik}^{*C(r)}$ to obtain the posterior median of the MCMC distribution for each parameter. Similarly, we fit the joint mixture model for longitudinal presence/absence data

using $y_{ijk}^{B(r)}$ to obtain the corresponding posterior estimates. We repeat the above procedure for $R = 500$ replicates.

We compare the use of the two type data records, relative to the analysis of accurately recorded aggregate count data, using average bias (ABIAS) and standard deviation (ASE) computed as

$$\text{ABIAS}(\hat{\theta}) = \sum_{r=1}^{R} \tilde{\theta}^{(r)}/R - \theta$$

$$\text{ASE}(\hat{\theta}) = \left[\sum_{r=1}^{R} \left(\tilde{\theta}^{(r)} - \sum_{r=1}^{R} \tilde{\theta}^{(r)}/R\right)^2 /R\right]^{\frac{1}{2}}$$

where $\tilde{\theta}$ denotes the posterior median for a parameter θ.

Table 3.3 contrasts the ABIAS of the parameters under different heaping distributions and frequency of presence/absence data collection. For the majority of parameters, the ABIAS for count data rounded according to H_I is virtually the same as that of accurately recorded count data.

For the analysis of heaped count data corresponding to heaping distributions H_{II} and H_{IV}, the absolute value of the ABIAS for the majority of the regression parameters as well as the factor loading parameters in the structural zero component increases, relative to the analysis of true count data. We expect this increase in bias for fixed intercepts and gender effects as the heaping probabilities depend on gender. At first glance, the increase in bias for the effects of placement in a secure facility may be surprising but can be explained by the specified covariate structure. Here, male subjects are more likely than female subjects (76% versus 55%) to have spent some time in a secure facility during the observation period. Therefore, under heaping distributions H_{II} and H_{IV}, the extent of heaping differs not only by gender but also for subjects who spent some time in a secure facility versus those who did not. Accordingly, the estimation of the effects of gender and placement in a secure facility are both impacted by gender differences in the heaping probabilities. The poor estimation of the factor loading

Table 3.3: Average bias for parameters across 500 simulated data sets for different heaping distributions and frequency (monthly, bi-monthy and tri-monthly) of presence/absence data collection.

	True Count	Heaped I	Heaped II	Heaped III	Heaped IV	Monthly	Bi-Monthly	Tri-Monthly
Drunk Driving								
$\beta_{11_0} = -5.08$	−0.06	−0.18	−0.14	−0.07	−0.13	−0.08	−0.07	−0.06
(female) $\beta_{11_1} = -1.22$	0.17	0.24	0.29	0.44	0.50	0.18	0.19	0.18
(secure facility) $\beta_{11_2} = 1.56$	−0.05	−0.10	−0.16	−0.26	−0.21	−0.05	−0.04	−0.02
$\delta_1 = 0.51$	0.03	0.03	0.01	−0.02	0.01	0.04	0.04	0.04
$\nu_1 = 2.26$	−0.02	−0.05	−0.21	−0.36	−0.34	−0.06	−0.05	−0.04
$\beta_{21_0} = -6.95$	0.45	0.55	0.74	0.80	0.89	0.55	0.61	0.64
(female) $\beta_{21_1} = 0.01$	0.01	0.01	0.04	0.04	−0.17	0.02	0.01	0.01
(secure facility) $\beta_{21_2} = 1.71$	−0.27	−0.28	−0.30	−0.36	−0.34	−0.32	−0.35	−0.37
$\lambda_1 = 2.03$	−0.31	−0.29	−0.36	−0.49	−0.49	−0.42	−0.49	−0.53
$\sigma^2_{b_1} = 3.21$	0.45	0.20	0.08	0.38	0.14	0.55	0.70	0.82
Aggressive Offending								
$\beta_{12_0} = -5.28$	0.12	0.11	−0.30	−0.71	−0.38	0.13	0.14	0.14
(female) $\beta_{12_1} = -0.59$	0.12	0.12	0.00	0.19	0.52	0.12	0.12	0.12
(secure facility) $\beta_{12_2} = 0.73$	−0.02	−0.02	−0.10	−0.25	−0.12	−0.02	−0.02	−0.01
$\delta_2 = 1.00$								
$\nu_2 = 1.33$	−0.07	−0.08	−0.20	−0.54	−0.26	−0.09	−0.10	−0.13
$\beta_{22_0} = -5.73$	0.01	0.02	0.14	0.18	0.19	0.00	−0.01	−0.01
(female) $\beta_{22_1} = -0.73$	0.01	0.01	0.05	0.03	−0.06	0.01	0.01	0.00
(secure facility) $\beta_{22_2} = 0.70$	−0.03	−0.03	−0.05	−0.06	−0.08	−0.02	−0.02	−0.02
$\lambda_2 = 0.79$	0.07	0.05	0.00	−0.04	−0.04	0.09	0.11	0.12
$\sigma^2_{b_2} = 1.28$	−0.10	−0.06	−0.11	−0.11	−0.18	−0.12	−0.13	−0.15

parameters in the structural zero component, ν_k, under H_{II} and H_{IV} likely reflects the fact that a proportion of low, non-zero counts are inaccurately recorded as zeros under these two heaping distributions. Averaged over the 500 replicate data sets, the proportion of true zeros counts for DD and AGG are 0.73% and 0.29%, respectively. The corresponding observed proportions of zero counts are 0.75% and 0.34% under both H_{II} and H_{IV}.

Relative to the analysis of true count data, the absolute value of the ABIAS for many of the parameters in the structural zero component obtained from the analysis of heaped count data corresponding to heaping distributions H_{III} and H_{IV} drastically increases. As well, there are similar increases for the fixed intercepts in the Poisson component. This increase in bias is a result of a proportion of low, non-zero counts being inaccurately recorded as zeros, leading to an increased number of observed zeros and a decreased observed frequency of low, non-zero counts. The increase in the observed proportion of zero counts, relative to the true count data, is largest under H_{III} with an average of 0.77% and 0.39% of the reported counts being zero for DD and AGG, respectively, under this heaping distribution. Under-reporting can lead to biased estimation in zero-inflated Poisson regression models, particularly in the structural zero component.

We remark that the potential bias in parameter estimation introduced by a particular heaping scheme heavily depends on the underlying event process. In additional simulations (not shown), we considered the situation where both outcomes are recorded with same level of accuracy using the heaping parameters specified for DD with H_I in Table 3.1. In this case, the ABIAS for the parameters corresponding to AGG that are obtained from the analysis of the heaped aggregate count data remains low despite large increases in the heaping probabilities. This is due to the fact that very few counts exceed the change point of 14 and, hence, the proportion of counts that are rounded is far lower for AGG than DD under the same heaping scheme.

We assume presence/absence data are accurately recorded and, hence, should yield unbiased parameter estimates. Indeed, regardless of the frequency of the monitoring times,

the ABIAS for all the parameters in the structural zero component for the analysis of presence/absence data are essentially the same as that for accurately recorded count data, aggregated over the observation period. As well, the ABIAS for the parameters in the Poisson component are comparable for the analyses of monthly presence/absence data and true aggregate count data. For presence/absence data collected under the less frequent monitoring schemes, there is an increase in the resulting ABIAS, relative to true count data, for some parameters in the Poisson component. This is particularly true for DD where the variability of subject-specific event rates is high.

The loss of data arising from the use of longitudinal presence/absence data leads to concerns about loss of precision. In Table 3.4, we contrast the ASE of the parameters obtained from true aggregate count data and that obtained from presence/absence data collected every month, every two months and every three months. Overall, the ASE values are larger for the presence/absence data than the accurately recorded count data with the largest ASE value corresponding to $\sigma^2_{b_1}$, the variance of the outcome- and subject-specific random effect in Poisson component for DD, under monitoring every three months. As expected, the ASE of the parameters decreases as the length of time between monitoring points decreases. For all of the parameters, the ASE corresponding to analyzing true count data and analyzing monthly presence/absence data are similar. Using longitudinal presence/absence data instead of accurately recorded count data, aggregated over the observation period, results in loss of precision but this loss is minimal if monitoring is frequent enough in our context. Determining an appropriate monitoring scheme for presence/absence data depends on the occurrence rate of the process under observation which is primarily driven by the baseline event rate and the between-subject variability in the Poisson component. Under the current parameter values, the ASE when presence/absence data collected every three months are analyzed instead of accurately recorded aggregate count data increased by at most 32% for DD and 23% for AGG, reflecting the differences in variability for these two outcomes. By contrast, in simulations where $\sigma^2_{b_2}$ is increased from 1.28 to 3.06 (not shown), the corresponding increases in ASE are at most 52% for AGG.

Table 3.4: Average standard deviation for parameters across 500 simulated data sets for the aggregate true count data and presence/absence data collected monthly, bi-monthy and tri-monthly.

	True Count	Monthly	Bi-Monthly	Tri-Monthly
Drunk Driving				
$\beta_{11_0} = -5.08$	0.30	0.31	0.31	0.30
(female) $\beta_{11_1} = -1.22$	0.34	0.34	0.34	0.35
(secure facility) $\beta_{11_2} = 1.56$	0.28	0.28	0.29	0.29
$\delta_1 = 0.51$	0.11	0.08	0.08	0.10
$\nu_1 = 2.26$	0.52	0.51	0.53	0.53
$\beta_{21_0} = -6.95$	0.45	0.47	0.47	0.49
(female) $\beta_{21_1} = 0.01$	0.41	0.42	0.42	0.44
(secure facility) $\beta_{21_2} = 1.71$	0.31	0.31	0.33	0.34
$\lambda_1 = 2.03$	0.48	0.51	0.54	0.56
$\sigma^2_{b_1} = 3.21$	1.30	1.43	1.58	1.72
Aggressive Offending				
$\beta_{12_0} = -5.28$	0.28	0.28	0.28	0.29
(female) $\beta_{12_1} = -0.59$	0.36	0.37	0.37	0.37
(secure facility) $\beta_{12_2} = 0.73$	0.32	0.32	0.32	0.32
$\delta_2 = 1.00$				
$\nu_2 = 1.33$	0.31	0.32	0.33	0.34
$\beta_{22_0} = -5.73$	0.15	0.15	0.16	0.18
(female) $\beta_{22_1} = -0.73$	0.18	0.19	0.19	0.20
(secure facility) $\beta_{22_2} = 0.70$	0.13	0.13	0.14	0.15
$\lambda_2 = 0.79$	0.13	0.13	0.14	0.16
$\sigma^2_{b_2} = 1.28$	0.18	0.19	0.21	0.22

The heaping observed in the Pathways to Desistance data set reflects elements of both heaping distributions H_I and H_{II}. Specifically, the distribution of aggregate counts under heaping distribution H_I appears to capture the observed patterns of heaping. However, there are gender differences in the propensity for rounding incorporated in H_{II} that need to be considered in this context. In our simulations, the biases obtained under H_I were comparable to those obtained from the analysis of accurately reported count data. On the other hand, under H_{II} there was an increase in bias, relative to H_I, with a more substantial increase for parameters corresponding DD than those corresponding to AGG. This may explain the differences in the effects for gender and placement in a secure facility for DD between the two types of data records obtained in our analysis of the data.

Overall, in the motivating context, it appears that the analysis of either of the available data records, count data aggregated over a year and rounded, or binary data recording presence/absence of events repeatedly collected each month of observation, may accurately and efficiently recover the true parameter values. In general, caution should be taken when analyzing count data with suspected heaping. In situations where the propensity for rounding is linked to one or more of the covariates or where rounded counts are the result of subjects underreporting the number of events, estimation can be substantially biased. In such situations, the use of longitudinal presence/absence data is preferable. Furthermore, our simulations show that the precision of the estimates obtained from longitudinal presence/absence data under modestly frequent monitoring can be comparable to that obtained from accurately recorded aggregate count data.

3.6 Discussion

We contrast inference based on the analysis of self-reported count data aggregated over the observation period with that of longitudinal presence/absence data using joint zero-inflated discrete regression models when heaping is expected. Taken together, the simulation and empirical studies demonstrate that the analysis of aggregate heaped count data and longitudinal

presence/absence data can lead to different results and, importantly, mismatched sets of significant risk factors.

In our motivating context, it seems that the use of both aggregate count, with evidence of heaping, and monthly presence/absence data may yield accurate parameter estimates. However, the utility of the two types of data records depends on the underlying processes generating events and the heaping behaviour. In our simulations, we observe that heaping may lead to substantial bias in parameter estimation; the magnitude of this bias depends on the heaping behaviour, the occurrence rate and the interplay of the two.

Here, relative to the use of accurately recorded aggregate count data, the loss of precision for parameter estimates obtained from the analysis of presence/absence data collected monthly may be minimal but this loss of precision increases as the length of time between monitoring points increases. Additionally, the advantage of accurate recording of presence/absence data is also more likely under shorter time intervals between monitoring points. Presence/absence data can provide much of the available information if the monitoring is frequent enough. Determining optimal frequency of monitoring times depends on the occurrence rate.

We note that the context we study may yield errors in both counts and presence/absence data since both types of data records are obtained in a retrospective manner. In particular, the presence/absence data may be subject to recall errors where the number of months with at least one event is underreported. Nevertheless, consideration of these two types of data records is useful for the broader context of design of recurrent event studies. As well, in our context, the heaping observed in the count data indicates that it is more likely that errors are observed there than in the presence/absence data. Additionally, a few very large observed counts for DD may have artificially inflated the variance estimates in the mean component. An investigation of the leverage of such outliers is underway.

The majority of studies collect a single type of data record and, therefore, it is usually not possible to directly contrast different study designs in terms of the resulting inference. Through the analysis of the Pathways to Desistance study data, we are able to compare the two study

designs in the context of self-reported data related to criminal behaviour and provide insights on choice of design when heaping is expected. It is clear that the benefits of one design over the other will heavily depend on a particular application. In light of this, in Appendix B, we detail how the methodology of our simulation study can be used in conjunction with pilot study data to inform decisions concerning the design of studies using self-reported data where heaping may be an issue.

In our simulations, we assume a homogeneous Poisson process to understand the design issues under a common modeling scenario. However, other underlying processes could be similarly considered. For example, using a piecewise constant Poisson process to accommodate non-homogeneous event processes is not unusual and results from utilizing such a modeling framework could add substantially to our understanding of these issues.

In this work, we assume that peaks in the observed count data reflect misreporting and investigate the implications of analyzing a distorted version of the true data. In some situations these peaks may be a feature of the underlying data generating process, for example, in studies of smoking cessation subjects may consume exactly one pack per day, corresponding to a heap at 20. In this case, conclusions presented here are not applicable and a model that accounts for inflation not only at 0 but at other peaks believed to represent likely 'true heaping' should be employed. We note that, from discussions with subject experts, no such arguments for distortions are hypothesized in the criminal behaviour context.

In the Pathways to Desistance study, data are also available on several other outcomes related to different types of antisocial and criminal behaviour. These outcomes may refer to the number of times the subject engaged in an activity, such as the two outcomes considered here, or they may refer to the number of days the subject engaged in an activity which is bounded by the length of exposure. In the latter case, we expect peaks corresponding to daily activity. Indeed, histograms of counts for drug-related outcomes exhibit a heap at 365 days. In the analysis, we link the conditional intensity of the Poisson process with the conditional mean of a Bernoulli response using the complementary log-log link function. In the case of

bounded counts, if the conditional Binomial probability is modeled using a complementary log-log link function then an analogous longitudinal presence/absence model with expected counts proportional to the number of Binomial trials can be utilized for zero-inflated Binomial outcomes. The impact of heaping at the upper bound of the sample space of the response variable is unclear and generalizing our findings to other count distributions warrants further research.

References

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

Bar, H. Y., & Lillard, D. R. (2012). Accounting for heaping in retrospectively reported event data-a mixture-model approach. *Statistics in Medicine* **31**, 3347-3365.

Bellhouse, D. R. (2011). A new look at Halley's life table. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174**, 823-832.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). New York, NY: Cambridge University Press.

Carlin, J. B., Wolfe, R., Brown, C. H., & Gelman, A. (2001). A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics* **2**, 397-416.

Crawfoard, F.W., Weiss, R.E., & Suchard, M.A. (2015). Sex, lies and self-reported counts: Bayesian mixture models for heaping in longitudinal count data via birth-death processes. *Annals of Applied Statistics* **9**, 572-596.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 355-366

Feng, C. X., and Dean, C. B. (2012). Joint analysis of multivariate spatial count and zero-heavy count outcomes using common spatial factor models. *Environmetrics* **23**, 493-508.

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

Heitjan, D. F., & Rubin, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association* **85**, 304-314.

Heitjan, D. F., & Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics* **19**, 2244-2253.

Jolliffe, D., Farrington, D. P., Hawkins, J. D., Catalano, R. F., Hill, K. G., & Kosterman, R. (2003). Predictive, concurrent, prospective and retrospective validity of self-reported delinquency. *Criminal Behaviour and Mental Health* **13**, 179-197.

Krohn, M. D., Lizotte, A. J., Phillips, M. D., Thornberry, T. P., & Bell, K. A. (2013). Explaining systematic bias in self-reported measures: Factors that affect the under-and over-reporting of self-reported arrests. *Justice Quarterly* **30**, 501-528.

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

Matsui, S., & Miyagishi, H. (1999). Design of clinical trials for recurrent events with periodic monitoring. *Statistics in Medicine* **18**, 3005-3020.

McGinley, J. S., Curran, P. J., & Hedeker, D. (2015). A novel modeling framework for ordinal data defined by collapsed counts. *Statistics in Medicine* **34**, 2312-2324.

Mulvey, E. P., Steinberg, L., Fagan, J., Cauffman, E., Piquero, A. R., Chassin, L., et al. (2004). Theory and research on desistance from antisocial activity among serious adolescent offenders. *Youth Violence and Juvenile Justice* **3**, 213-236

Olsen, M. K., & Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* **96**, 730-745.

Piquero, A. R., Schubert, C. A., & Brame, R. (2014). Comparing official and self-report records of offending across gender and race/ethnicity in a longitudinal study of serious youthful offenders. *Journal of Research in Crime and Delinquency* **51**, 526-555.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, Austria.

Ridout, M. S., & Morgan, B. J. (1991). Modelling digit preference in fecundability studies. *Biometrics* **47**, 1423-1433.

Roberts, J. M., & Brewer, D. D. (2001). Measures and tests of heaping in discrete quantitative distributions. *Journal of Applied Statistics* **28**, 887-896.

Rodrigues-Motta, M., Pinheiro, H. P., Martins, E. G., Araujo, M. S., and dos Reis, S. F. (2013). Multivariate models for correlated count data. *Journal of Applied Statistics* **40**, 1586-1596.

Schubert, C. A., Mulvey, E.P., Steinberg, L., Cauffman, E., Losoya, S., Hecker, T., Chassin, L., et al. (2004). Operational lessons from the Pathways to Desistance project. *Youth Violence and Juvenile Justice* **3**, 237-255

Tooze, J. A., Grunwald, G. K., & Jones, R. H. (2002). Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research* **11**, 341-355.

Wang, H., & Heitjan, D. F. (2008). Modeling heaping in self-reported cigarette counts. *Statistics in Medicine* **27**, 3789-3804.

Wang, H., Shiffman, S., Griffith, S. D., & Heitjan, D. F. (2012). Truth and memory: Linking instantaneous and retrospective self-reported cigarette consumption. *The Annals of Applied Statistics* **6**, 1689-1706.

Yu, B. (2008). A frailty mixture cure model with application to hospital readmission data. *Biometrical Journal* **50**, 386-394.

Chapter 4

Joint Analysis of Multivariate Longitudinal Presence/Absence Data Subject to Resolution

4.1 Introduction

Joint outcome recurrent event data arise when events generated by two or more processes may occur repeatedly over a period of observation. In some settings, the condition generating the events may resolve over time; following the point of resolution the subject no longer experiences events corresponding to any of the processes. In the context of criminology, research has repeatedly documented that a substantial proportion of serious adolescent offenders likely will not continue their criminal career into adulthood (Mulvey et al., 2010). Based on empirical and theoretical grounds, some researchers have posited groups with distinct criminal behaviour trajectories. Blumstein, Farrington, and Moitra (1985) hypothesized groups called *desisters*, *persisters* and *nonoffenders*; Moffitt (1993) differentiates between *persistent chronic offenders* and *offenders who not not persist beyond adolescence* in criminal activities, offering distinct sociological explanations for these two types of offenders. The motivation for the de-

velopments in this paper is a major study on criminal behaviour patterns of serious adolescent offenders from adolescence into early adulthood. One goal of this study is to reliably distinguish between juvenile offenders who will continue antisocial and illegal behaviour beyond adolescence and those who will not. Here, decisions concerning legal sanctions and interventions are made at specific evaluation points. Such decision-making may be improved when an individual's likelihood of future offending, given their offending history over a long window of time, is considered. These sorts of tools as developed here to accommodate resolution of events are important in other scenarios. In the medical context, for some chronic diseases, the disease process may resolve naturally, as for example rheumatological conditions where the disease goes into remission (Shen and Cook, 2015), but it can be difficult to determine if and when such changes take place.

Often the exact event times are not readily available, particularly for self-reported data. This is the case for our motivating study where data pertaining to several types of offending consist of binary data recording the presence/absence of events repeatedly collected at each month of observation. For each outcome, this represents partial observation of a counting process. In our modeling framework, the use of the complementary log-log link function allows us to explicitly link binary responses to a suitable underlying count distribution as will be described in more detail later.

When several longitudinal binary outcomes are jointly considered, responses for a specific subject are likely correlated both over time and across outcomes. Previous authors (for example, Agresti, 1997; Ribaudo and Thompson, 2002) have jointly analyzed several longitudinal binary outcomes using random effect models to incorporate complex correlation structures. We utilize the general framework proposed by Dunson (2000) in which, conditional on random effects, different members of the exponential family are used to describe the component models in the joint distribution of the set of observed outcomes.

For longitudinal presence/absence data, we may observe a zero response vector for some subjects. In order to account for a high proportion of subjects who never experience an event,

Carlin et al. (2001) proposed a mixture model for longitudinal binary data in which each subject may be either at-risk or not at-risk for an event. Within the at-risk group, the probability of an event is modeled by a mixed logistic regression model. Here, we consider the joint analysis of several longitudinal binary outcomes using a similar mixture model approach where outcomes are linked by subject-specific random effects.

Methods for the analysis of a single recurrent event process subject to resolution have been recently developed. Rondeau et al. (2013) discussed cure frailty models that account for the possibility of a cure after each event. Shen and Cook (2014) developed a dynamic Mover-Stayer model for recurrent event processes in which a latent variable associated with each event indicates whether the underlying disease has resolved. Given that an individual's disease process has not resolved, events follow a standard point process model governed by a latent intensity. This framework has been extended (Shen and Cook, 2015) to accommodate the analysis of interval-censored recurrent event data where the exact event times are not available but the cumulative event count is recorded at periodic assessment times. These models enable a clearer understanding of occurrence patterns when the possibility of resolution of an underlying process can be justified. Omitting the possibility that the underlying process generating events may resolve may lead to underestimating the event rate among subjects for whom the underlying process has not resolved. We utilize here some ideas from Shen and Cook (2014, 2015) whereby a latent variable is associated with resolution of the underlying process, extending this to multiple outcome analysis. Furthermore, our extensions allows us to evaluate the effects of time-dependent interventions on the event rate among subjects for whom the underlying process has not resolved and on the probability of this resolution.

We note also that other types of transitional models have been utilized in different contexts. Motivated by a smoking cessation study, Luo et al (2008) proposed a discrete time model with three behavioural states; smoking, transient quitting and permanent quitting. When a subject is in the smoking state, a quit attempt may be made at the beginning of each assessment period. Once a quit attempt is successfully made, the subject may enter the transient quitting state or

permanent quitting state. The model for the resolution of the process generating events considered here in fact follows the same basic structure; the underlying process may only resolve following a period of offending. Importantly, the general framework proposed in Section 4.3 accommodates the use of more flexible models for the latent variable indicating whether the underlying process has resolved, which may be warranted in different applications.

Models for multivariate longitudinal outcomes using a shared underlying latent variable which focus on describing transitions through distinct states, have been previously considered. Scott et al (2005) proposed a hidden Markov model for data collected in a clinical trial of schizophrenia patients where the conditional distribution of multivariate outcomes given a latent health state follows a multivariate t-distribution. For medical utilization data, Wall and Li (2009) introduced a hidden Markov model which assumes a common unobserved health state governs the counts of several types of medical encounters. This approach takes advantage of conceptualizing a latent variable underlying the multivariate longitudinal data to provide a succinct way of summarizing the process.

This article develops methods for the joint analysis of several longitudinal binary outcomes denoting the presence/absence of events between periodic assessments in settings where an underlying process generating events can resolve. In Section 4.2, we introduce our motivating data set and outline the methodological challenges that motivated this work. In Section 4.3, we describe our general modeling framework. The novelty is that we model the simultaneous resolution of several recurrent event processes and utilize a mixture approach which accommodates the possibility that a subject may not be at-risk to engage in one or more of the outcomes. Accommodating such subjects is imperative in settings, such as our motivating example, where there is no initiating event triggering the start of the observation period as the point of resolution may occur prior to start of the study. We consider mixture models which assume the underlying process may only resolve following the occurrence of at least one event for any of the outcomes, accommodating effects due to time spent in a secure facility. Highlighting novel insights arising from the conceptualization of a latent variable underlying the multivari-

ate longitudinal data, we present the results of our joint analysis of this study in Section 4.4. Focusing on the model component associated with the resolution of the underlying process, we investigate the performance of our methodology through a simulation study in Section 4.5. We conclude with a discussion of results and limitations, as well as suggestions for further research.

4.2 A Study of Antisocial Behaviour Among Serious Juvenile Offenders

The Pathways to Desistance study (Mulvey et al., 2004; Schubert et al., 2004) is a longitudinal study of a group of serious juvenile offenders investigating offending patterns in the period following court adjudication. A total of 1354 youth offenders, aged 14 through 17 years old, who were found guilty of at least one serious offense in the metropolitan areas of Phoenix, Arizona or Philadelphia, Pennsylvania were enrolled in the study between 2000 and 2003 and followed for up to 7 years. The primary aim of that study was to identify patterns of desistance or escalation among serious juvenile offenders and evaluate the effects of adolescent development, sanctions and interventions on these offending patterns.

All subjects completed a baseline interview where information about background characteristics and previous offending was collected. Additionally, interviews were conducted over a seven year follow up period. At each interview, data pertaining to antisocial and criminal activity in the period since the previous interview were recorded. Specifically, subjects indicated the months in which they engaged in an antisocial or illegal activity and, therefore, the available data consists of repeatedly measured binary data, indicators of presence/absence of events during each month of observation.

During the follow up period, subjects may have spent time in a facility with no access to the community, termed a secure facility. Data on placement in a secure facility and, if so, the proportion of the month spent in a secure facility, are available. Some of the antisocial

and criminal activities are highly unlikely to occur in a secure facility and, for this analysis, are considered prohibited in a secure facility. We assume the expected number of events is regulated by an exposure variable corresponding to the number of days a subject is able to engage in an outcome. This exposure variable varies from outcome to outcome. An additional concern is the effect of institutional placement on current and subsequent offending. To address this issue, we consider two time-varying covariates related to placement in a secure facility. The first is a binary indicator of placement in a secure facility during the current month. The second is the cumulative time, in years, spent in a secure facility during prior months from the start of the study.

We summarize the eight outcomes considered here: *carried a gun*, *sold marijuana*, *sold other drugs*, *drove drunk*, *aggressive I*, *aggressive II*, *income I*, and *income II*, in Table A.1; we provide a list of the antisocial and criminal activities associated with each outcome, indicate whether the outcome is considered prohibited while a subject is in a secure facility and the type of response collected. These outcomes may refer to the number of times the subject engaged in an activity and, therefore, represent an unbounded count, or they may refer to the number of days the subject engaged in an activity, which is bounded by the length of exposure. Our analysis will therefore need to provide flexibility with regard to different discrete distributions, for example, both Poisson-type for unbounded and conditional Binomial for bounded.

If the process generating events can resolve at some point, we would expect to observe uncommonly long periods of time without the occurrence of any criminal behaviour at the end of the follow up period for some subjects. As a rough way of understanding the patterns observed in the data, for each of the eight outcomes considered, we examined the distribution of the time between successive positive responses over all subjects. That is, for each positive response, we considered the waiting time before the next positive responses as a single, possibly censored, observation. As well, we examined the distribution of time between successive occurrences of any type of criminal behaviour. Figure 4.1 depicts the associated Kaplan-Meier curves. We observed that all curves exhibit a similar shape characterized by an initial rapid decline followed

by a plateau, suggesting that some subjects may permanently cease offending over the duration of the observation period. Note that curve for the occurrence of at least one outcome closely follows the curve corresponding to *aggressive II* which reflects that much of the information in the data is carried in this outcome.

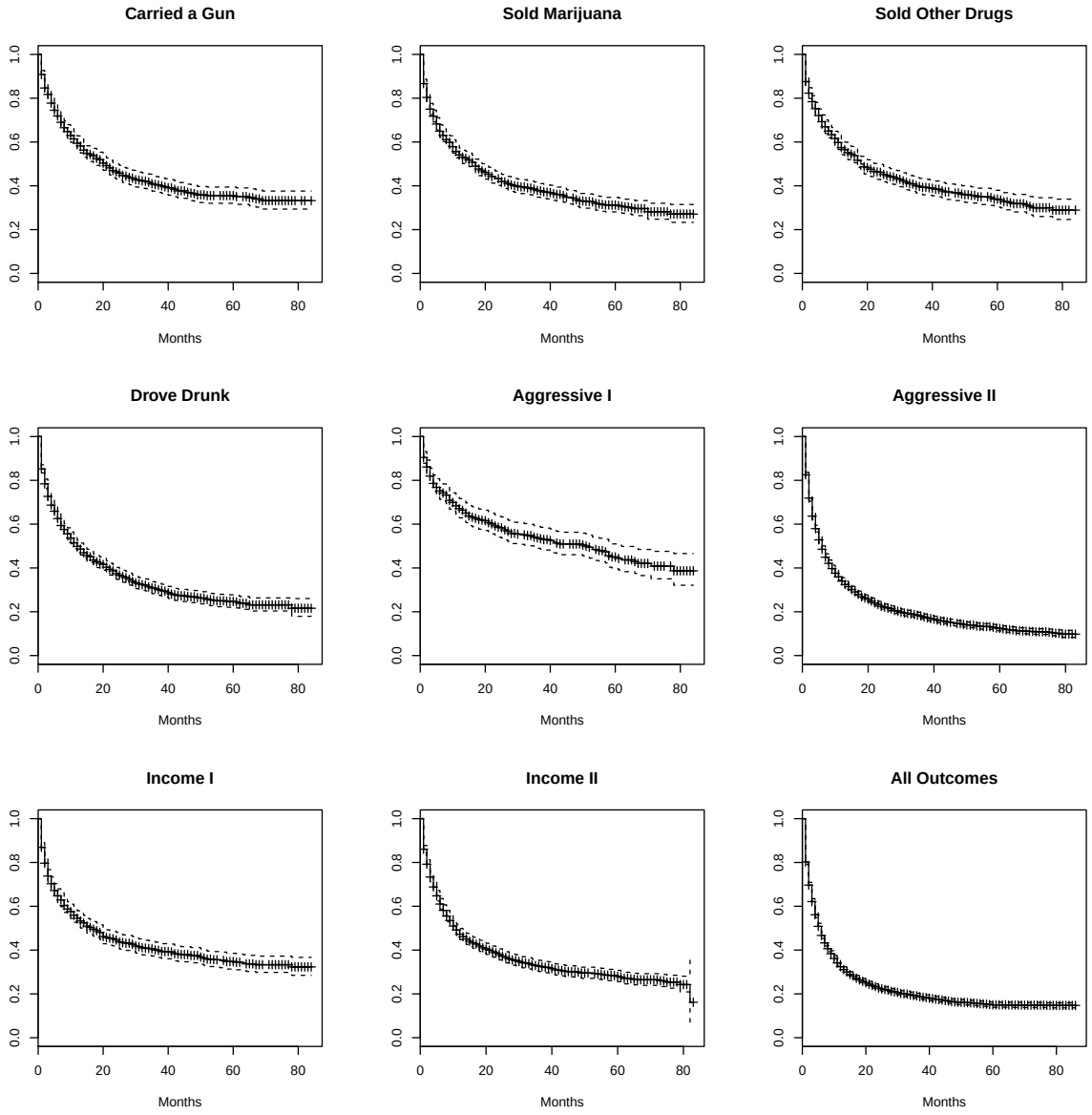

Figure 4.1: Kaplan-Meier curves for the waiting time between successive positive responses over all subjects.

4.3 Joint Model for Multivariate Longitudinal Presence/Absence Data Subject to Resolution

An important focus of our model development is enabling a clearer understanding of the processes generating zeros. Within the proposed modeling framework, zeros may arise from three distinct sources. Firstly, some subjects, termed non-engagers, are not at-risk to engage in a particular outcome at any point during the observation period, i.e. they generate zero values at all time points for a particular outcome. Secondly, the process generating events may resolve at some point, resulting in unusually long runs of zeros for each outcome at the end of the observation period. Finally, for outcome-specific engagers for whom the underlying process generating events has not resolved, at each assessment point, there is the possibility of observing a zero corresponding to the realization of a zero count from the underlying count distribution. Commonly, such count outcomes are regulated by an exposure variable, with the length of exposure being proportional to the expected count. Here, some of the outcomes are prohibited during a specific treatment leading to some of the zero counts being accounted for in a structural manner based on an exposure variable. We discuss the model components associated with each of the distinct sources of zeros in turn.

Suppose there are N subjects in a study and for each subject, let $0 = t_0 < t_1 < \cdots < t_{T_i}$ denote successive monitoring times. For simplicity in presenting the likelihood, we assume here, that the monitoring times are common for all subjects and equally spaced. This is true for our motivating data set. We assume data are collected on K related outcomes at each monitoring point. Let y_{ijk} be a binary response for subject i at t_j, so that $y_{ijk} = 1$ if one or more events corresponding to outcome k occurred between t_{j-1} and t_j and $y_{ijk} = 0$ otherwise; and $\mathbf{y}_{ik} = (y_{i1k}, \ldots, y_{iT_ik})'$ be the sequence of binary responses over $j = 1, \ldots, T_i$ observed for subject i and outcome k, $i = 1, \ldots, N$, $k = 1, \ldots, K$.

We assume each response vector $\mathbf{y}_{ik}$, conditional on random effects, is independently drawn from a mixture model having density

$$f(\mathbf{y_{ik}}|s_{ik}, r_i, v_i, b_{ik}) = \begin{cases} \mathbf{I}(\mathbf{y}_{ik} = \mathbf{0}_{T_i \times 1}) & \text{if } s_{ik} = 1 \\ f_{B_k}(\mathbf{y}_{ik}|\mathbf{q}_i, v_i, b_{ik}) & \text{if } s_{ik} = 0 \end{cases} \tag{4.1}$$

where the variables s_{ik} are latent Bernoulli indicators, markers for the outcome-specific non-engagers, with mean function p_{ik}, conditional on a random effect r_i; subject and outcome specific random effects v_i and b_{ik} will be discussed later. Specifically, for each outcome, we assume $s_{ik}|p_{ik} \sim Bern(p_{ik})$ with

$$p_{ik} = \{1 + \exp(-\boldsymbol{x}_{p_i}^{'}\boldsymbol{\beta}_{p_k} - \lambda_{r_k} r_i)\}^{-1} \tag{4.2}$$

where $\boldsymbol{x}_{p_i}$ is a $l_p \times 1$ vector of covariates, $\boldsymbol{\beta}_{p_k}$ is a vector of corresponding regression parameters, r_i is a subject-specific random effect and λ_{r_k} is a factor loading parameter representing outcome-specific variability related to r_i.

For each outcome, one mixture component places all its mass on the zero vector while the other component distributes mass according to the density, $f_{B_k}(\mathbf{y}_{ik}|\mathbf{q}_i, v_i, b_{ik})$, corresponding to a longitudinal binary model. The longitudinal binary model accommodates the possibility that the process generating the events may resolve at some point. In particular, the resolution of this underlying process means that a subject will no longer experience events related to any of the K outcomes, perhaps resulting in unusually long event-free periods of time at the end of the observation period. All subjects who are engagers for at least one of the outcomes may experience a resolution of the underlying process generating events. We let q_{ij} denote a time-dependent latent indicator such that $q_{ij} = 1$ if underlying process generating events for subject i has resolved by t_{j-1} and $q_{ij} = 0$ otherwise, $j = 1, \ldots, T_i$. We may view $q_{ij} = 1$ as reflecting the absorbing state of resolution in that once $q_{ij} = 1$, the value of this latent variable will be 1 at all subsequent time points. The set of time-dependent indicators of resolution for subject i is denoted $\mathbf{q}_i = \{q_{il}, \ldots, q_{iT_i}\}$. Note that $\boldsymbol{q}_i$ is partially observed as we know q_{ij}= 0 for any assessment period up to and including the one corresponding to the last observed positive

response. Conditional on random effects, we assume the binary responses for the outcome-k-specific engagers for whom the underlying process generating events has not resolved follows a Bernoulli distribution with probability of success ζ_{ijk} so that $f_{B_k}(\mathbf{y}_{ik}|\mathbf{q}_i, v_i, b_{ik})$ is given by

$$f_{B_k}(\mathbf{y}_{ik}|\mathbf{q}_i, u_i, v_i, b_{ik}) = \prod_{j=1}^{T_i}\left[(1-q_{ij})\{\zeta_{ijk}^{y_{ijk}}(1-\zeta_{ijk})^{1-y_{ijk}}\} + q_{ij}\right] \tag{4.3}$$

$i = 1, \ldots, N$, $j = 1, \ldots, T_i$, $k = 1, \ldots, K$.

For each outcome, this Bernoulli random variable represents the partial observation of a counting process where the probability of success, ζ_{ijk}, corresponds to the probability of observing at least one event. For outcome-k-specific engagers for whom the underlying process generating events has not resolved, we assume the number of events that occur between t_{j-1} and t_j, conditional on random effects, follows a count distribution with mean

$$\mu_{ijk} = g_k^{-1}\{\boldsymbol{x}_{\mu_{ij}}^{'}\boldsymbol{\beta}_{\mu_k} + h(j, \boldsymbol{\rho}_k) + \lambda_{v_k} v_i + b_{ik}\} z_{ijk} \tag{4.4}$$

Here g_k is a link function; $\boldsymbol{x}_{\mu_{ij}}$ is a $l_\mu \times 1$ vector of covariates for the fixed effects and $\boldsymbol{\beta}_{\mu_k}$ is vector of corresponding regression parameters; $h(t, \boldsymbol{\rho}_k)$ is a function of time describing the temporal trends in μ_{ijk}. The expected count is proportional to the length of exposure, z_{ijk}, for subject i and outcome k between t_{j-1} and t_j. The subject-specific random effect, v_i, is shared across outcomes and λ_{v_k} is the factor loading for this shared effect on outcome k. The outcome- and subject-specific random effect b_{ik} represent additional heterogeneity beyond the shared random effect. We assume the random effects are normally distributed such that $r_i \sim N(0, 1)$, $v_i \sim N(0, 1)$ without loss of generality since all factor loadings are not constrained; and $b_{ik} \sim N(0, \sigma_{b_k}^2)$, $i = 1, \ldots, N$ and $k = 1, \ldots, K$.

If conditional on random effects, we assume the underlying counts follows a Poisson distribution with mean μ_{ijk} with g_k being the log link function then

$$\zeta_{ijk} = 1 - \exp\{-\exp(\boldsymbol{x}_{\mu_{ij}}^{'}\boldsymbol{\beta}_{\mu_k} + h(j, \boldsymbol{\rho}_k) + \log(z_{ijk}) + \lambda_{v_k} v_i + b_{ik})\} \tag{4.5}$$

Alternatively, for outcome-k-specific engagers for whom the underlying process generating events has not resolved, the number of events that occur between t_{j-1} and t_j may be bounded and follow, conditional on random effects, a Binomial($z_{ijk}, \frac{\mu_{ijk}}{z_{ijk}}$) distribution. If we assume g_k is the complementary log-log link function then the probability of observing at least one event can be expressed as

$$\begin{aligned}\zeta_{ijk} &= 1 - \left(1 - \frac{\mu_{ijk}}{z_{ijk}}\right)^{z_{ijk}} \\ &= 1 - [\exp\{-\exp(\boldsymbol{x}'_{\mu_{ij}}\boldsymbol{\beta}_{\mu_k} + h(j, \boldsymbol{\rho}_k) + \lambda_{v_k} v_i + b_{ik})\}]^{z_{ijk}} \\ &= 1 - \exp\{-\exp(\boldsymbol{x}'_{\mu_{ij}}\boldsymbol{\beta}_{\mu_k} + h(j, \boldsymbol{\rho}_k) + \log(z_{ijk}) + \lambda_{v_k} v_i + b_{ik})\} \qquad (4.6)\end{aligned}$$

The longitudinal model in (4.3) provides a flexible modeling framework and requires specifying a model for the latent indicator variables q_{ij}. In the case where $q_{ij} \equiv 0 \ \forall i, j$, the model reduces to an extension of the mixture model for longitudinal binary data proposed by Carlin et al. (2001) in which several outcomes are linked by subject-specific random effects. We contrast the proposed full model, detailed below, and this reduced model in the analysis of our motivating example and through simulations.

As the focus of our analysis is identifying factors related to desistance among serious juvenile offenders, we model the probability of permanently quitting, defined as the probability that the underlying process generating events resolves following an assessment period with the occurrence of at least one event for any of the outcomes. That is, we assume the underlying process generating events may only resolve at a time point t_j if one or more events corresponding to any of the K outcomes occurred between t_{j-1} and t_j. The probability of permanently quitting is modeled as

$$\mathrm{P}\left(q_{ij} = 1 \middle| \left\{q_{i(j-1)} = 0, \sum_{k=1}^{K} y_{i(j-1)k} > 0, \boldsymbol{x}_{\phi_{ij}}\right\}\right) = \phi_{ij} = \exp\{-\exp(\boldsymbol{x}'_{\phi_{ij}}\boldsymbol{\beta}_\phi)\} \qquad (4.7)$$

$j = 2, \ldots, T_i$; $\boldsymbol{x}_{\phi_{ij}}$ is a $l_\phi \times 1$ vector of covariates for the fixed effects and $\boldsymbol{\beta}_\phi$ is the vector

of corresponding regression parameters. A critical issue involved in regression models for binary response data is the choice of an appropriate link function. This involves choosing between a symmetric link and a skewed link and, if applicable, the direction of the skewed link. Preliminary results indicate that the probability of a subject permanently quitting offending is very low and, hence, we utilize the negatively skewed log-log link function. Under the proposed model, inference for ϕ_{ij} is based on all available data, including data collected after t_j.

The mixed joint models for multivariate longitudinal presence/absence data subject to resolution may be implemented in a Bayesian framework using Markov chain Monte Carlo (MCMC) methods. The joint posterior distribution of the parameters is

$$p(\boldsymbol{\Theta}, \boldsymbol{r}, \boldsymbol{v}, \boldsymbol{b}|\boldsymbol{Y}) \propto L(\boldsymbol{Y}|\boldsymbol{\Theta}, \boldsymbol{r}, \boldsymbol{v}, \boldsymbol{b}) p(\boldsymbol{b}|\boldsymbol{\sigma}_b^2) \pi(\boldsymbol{\sigma}_b^2) p(\boldsymbol{r}) p(\boldsymbol{v}) \pi(\boldsymbol{\Theta}) \tag{4.8}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\beta}_p, \boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\phi, \boldsymbol{\rho}, \boldsymbol{\lambda}_r, \boldsymbol{\lambda}_v)'$, $\boldsymbol{\beta}_p = (\boldsymbol{\beta}_{p_1}, \ldots, \boldsymbol{\beta}_{p_K})'$, $\boldsymbol{\beta}_\mu = (\boldsymbol{\beta}_{\mu_1}, \ldots, \boldsymbol{\beta}_{\mu_K})'$, $\boldsymbol{\rho} = (\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_K)'$, $\boldsymbol{\lambda}_r = (\lambda_{r_1}, \ldots, \lambda_{r_K})'$, $\boldsymbol{\lambda}_v = (\lambda_{v_1}, \ldots, \lambda_{v_K})'$, $\boldsymbol{\sigma}_b^2 = (\sigma_{b_1}^2, \ldots, \sigma_{b_K}^2)'$, $\boldsymbol{r} = (r_1, \ldots, r_N)'$, $\boldsymbol{v} = (v_1, \ldots, v_N)'$ and $\boldsymbol{b} = (b_{11}, \ldots, b_{N1}, b_{12}, \ldots, b_{NK})'$. The first term on the right hand side of (4.8) is the likelihood

$$\begin{aligned} L(\boldsymbol{Y}|\boldsymbol{\Theta}, \boldsymbol{r}, \boldsymbol{v}, \boldsymbol{b}) \propto \prod_{i=1}^{N} \prod_{k=1}^{K} [&\mathbf{I}(\mathbf{y}_{ik} = \mathbf{0}_{T_i \times 1})\{p_{ik} + (1 - p_{ik}) f_{B_k}(\mathbf{0}_{T_i \times 1}|\boldsymbol{q_i}, v_i, b_{ik})\} \\ &+ \mathbf{I}(\mathbf{y}_{ik} \neq \mathbf{0}_{T_i \times 1})\{(1 - p_{ik}) f_{B_k}(\mathbf{y}_{ik}|\boldsymbol{q_i}, v_i, b_{ik})\}] \end{aligned} \tag{4.9}$$

The Bayesian model specification is made complete by assigning prior distributions to $\boldsymbol{\Theta}$ and $\boldsymbol{\sigma}_b^2$. Inference is then based on the posterior distribution, which can be summarized using samples drawn from the posterior distribution. This framework for the analysis was implemented through the freely available software JAGS (Plummer, 2003).

4.4 Application to the Pathways to Desistance Study

We restrict our analysis to subjects who completed at least the first follow up interview (N=1259). For each subject, the number of time points included in the analysis, T_i, is defined as the number of consecutive months of follow up with complete data. We define the length of exposure, z_{itk}, as the number of days under observation for subject i and outcome k between t_{j-1} and t_j, for outcomes that are not prohibited in a secure facility, and as the number of days spent in the community, for outcomes that are prohibited in a secure facility. We assume piecewise linear temporal trends with a single knot at 36 months in the mean component of the model. Covariates considered include gender (male/female), ethnicity (black/Hispanic/other), a binary indicator of placement in a secure facility between t_{j-1} and t_j (in $\boldsymbol{x}_{\mu_{ij}}$) and the cumulative time, in years, spent in a secure facility between t_0 and t_{j-1} (in $\boldsymbol{x}_{\mu_{ij}}$ and $\boldsymbol{x}_{\phi_{ij}}$).

4.4.1 Computational Details

We assign weakly informative prior distributions for the fixed regression effects, $\boldsymbol{\beta}_{p_k} \sim N_{l_p}(\mathbf{0}, \boldsymbol{I}_{l_p})$, $\boldsymbol{\beta}_{\mu_k} \sim N_{l_\mu}(\mathbf{0}, \boldsymbol{I}_{l_\mu})$, $\boldsymbol{\rho}_k \sim N_2(\mathbf{0}, \boldsymbol{I}_2)$ $k = 1, \ldots, K = 8$, and $\boldsymbol{\beta}_\phi \sim N_{l_\phi}(\mathbf{0}, \boldsymbol{I}_{l_\phi})$ where $\boldsymbol{I}_n$ denotes an $n \times n$ identity matrix. For the factor loading parameters, λ_{r_k} and λ_{v_k} $k = 1, \ldots, 8$, we adopt moderately informative priors, $\Gamma(1, 1)$. Finally, we choose Unif(0, 100) priors for the standard deviations of the outcome- and subject-specific random effects in the mean component, σ_{b_k} $k = 1, \ldots, 8$.

The results below arise from two chains, each was run for an initial 2000 burn-in iterations followed by an additional 10 000 iterations thinned at 10, resulting in a total of 2000 iterations to be used for posterior inference. In order to reduce the number of iterations needed and improve the mixing of the chains, we implement a hierarchical centering reparametrization (Gelfand, Sahu and Carlin, 1996) in the mean component of the model.

4.4.2 Results

We investigate the insights, above that provided by less complex models, obtained by accounting for the possibility of the underlying process resolving following the occurrence of at least one event. We compare our proposed "full" model and a "reduced" model without the permanent quit component i.e., $q_{ij} \equiv 0$.

Standard implementations of random effects models assume a known correlation structure. In some settings commonly used assumptions concerning this structure may not be appropriate. For example, the assumption that the random effects covariance matrix has the same structure across outcomes may not be realistic. Here, the outcome- and subject-specific variability in the mean component for *aggressive I* appears to be adequately captured by the shared random effect, v_i. Only 2.5% of the positive responses for *aggressive I* do not coincide with a positive response for at least one other outcome. The inclusion of the shared random effect effectively reduces the variance of the outcome- and subject-specific random effect for *aggressive I* to zero. Additionally, it appears that *aggressive II* is distinct from the remaining outcomes in terms of the probability of being an non-engager. In particular, 16% of subjects report only engaging in *aggressive II* while the corresponding proportion is approximately 1% for the remaining outcomes. Therefore, we set the relevant parameters (λ_{r_k} for *aggressive II* ($k = 6$) and $\sigma^2_{b_k}$ for *aggressive I* ($k = 5$)) equal to zero. This approach utilizes ideas from Chen and Dunson (2003) whereby random effects may have zero variance and effectively drop out of the model.

Probability of Non-engager: The posterior medians and 95% equal-tail credible intervals for the baseline covariate effects in the non-engager component are shown in the top row of Figure 4.2. For all outcomes, under the full and reduced models, female subjects have a higher probability than males of being a non-engager. This effect is significant for all outcomes. Relative to the baseline group, for both the full and reduced models, black subjects have a significantly higher probability of being a non-engager for *drove drunk*, *income I* and *income II*.

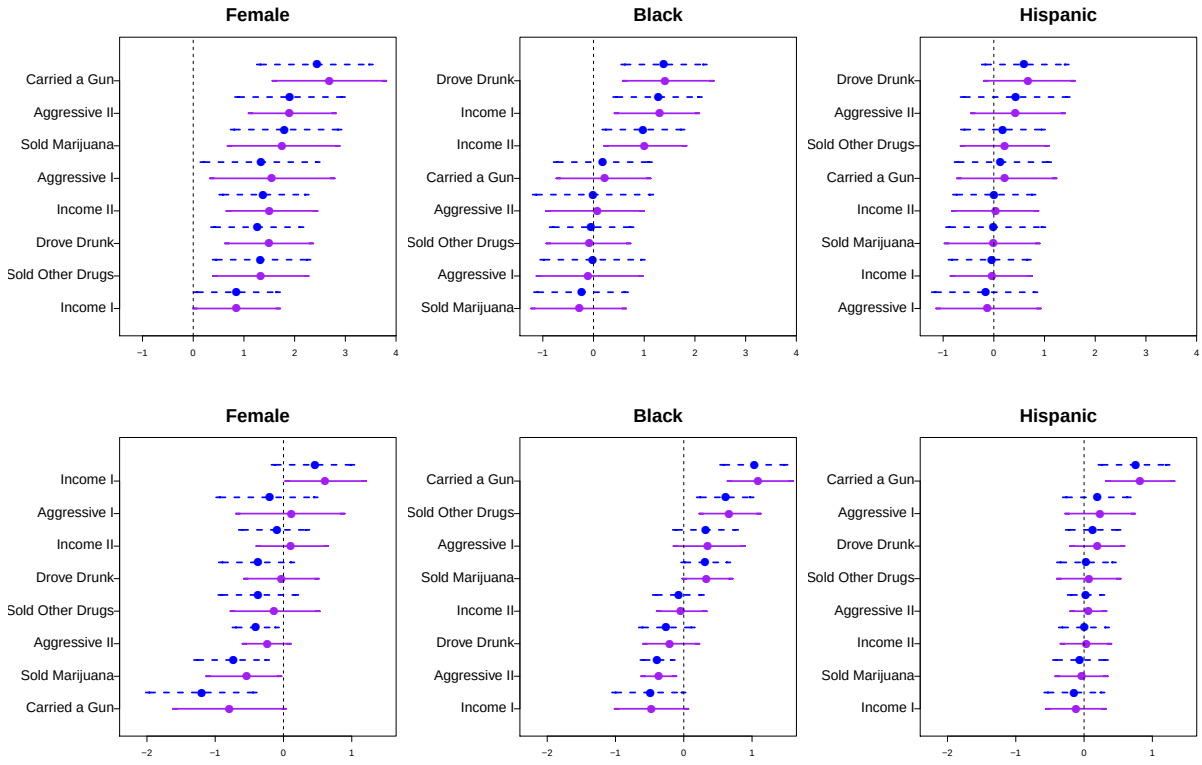

Figure 4.2: Comparison of effects of baseline covariates on the probability of a non-engager ($\boldsymbol{\beta}_{p_k}$, top) and the mean count ($\boldsymbol{\beta}_{\mu_k}$, bottom). Credible intervals corresponding to full/reduced model are shaded in solid purple/dashed blue.

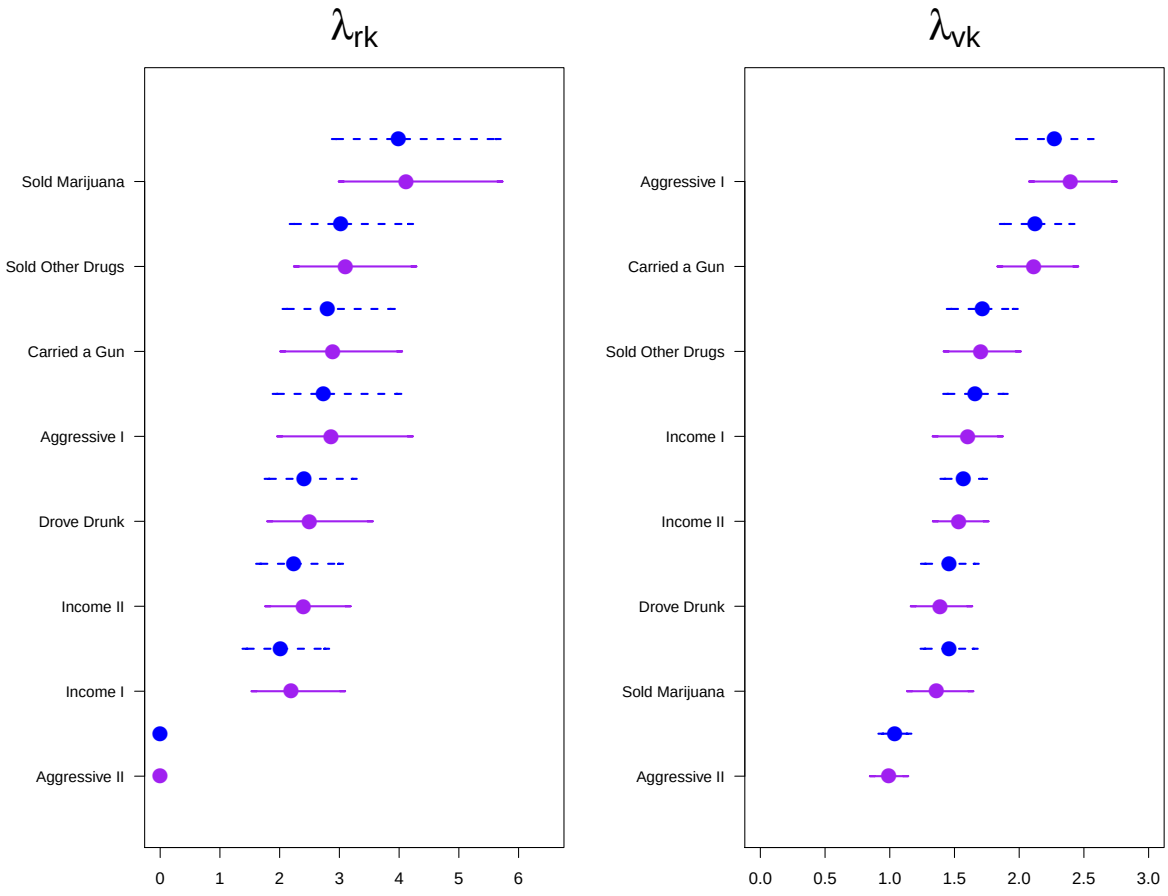

Figure 4.3: Posterior medians and 95% credible intervals for the factor loading parameters for the probability of a non-engager (λ_{r_k}, left) and the mean (λ_{v_k}, right) model components. Credible intervals corresponding to full/reduced model are shaded in solid purple/dashed blue.

The posterior medians and credible intervals for the factor loading parameters associated with permanent quit model component are displayed in the left column of Figure 4.3. Considering the subject-specific random effects in the non-engager component of the model, the factor loading parameter obtained from the full (and reduced) model for *sold marijuana* seems to be distinctly larger, indicating larger variability for this outcome.

Mean of Partially Observed Count Distribution: The posterior medians for the outcome-specific trajectories for the mean of the count distribution are displayed in Figure 4.4. Relative to the full model, under the reduced model, the mean is consistently underestimated over time for all of the outcomes. Furthermore, the shape of the time trend differs for *carried a gun*, *sold*

marijuana, *sold other drugs* and *drove drunk* under the two models. Under the reduced model, the time trends for these four outcomes remains relatively flat while under the full model the mean is increasing over time. Here, the reduced model is essentially averaging over the increasing event rate within a shrinking group for whom the underlying process has not resolved, and long periods without recurrence at the end of the observation period corresponding to a growing subgroup of subjects who have permanently quit offending.

The posterior medians and 95% equal-tail credible intervals for the baseline covariate effects in the mean component are shown in the bottom row of Figure 4.2. Under the full model, within the outcome-specific engagers, female subjects compared to male subjects have a significantly higher mean for *income I* and a significantly lower mean for *sold marijuana*. For all outcomes, the effect of gender in the mean component is lower under the reduced model, compared to the full model. This arises from the fact that female subjects compared to male subjects are more likely to permanently quit offending. Removing the permanent quit component yields more female subjects with long periods of non-offending in the mean component of the model and, hence, the frequency of events for female subjects across all outcomes appears lower. This change is most apparent for *aggressive II* where the gender effect is not significant under the full model but significant under the reduced model. This is expected as *aggressive II* is the most frequently reported outcome and appears to primarily determine the resolution process.

As well, under both the full and reduced models, there are several differences in the mean component among ethnicities. Relative to the baseline group, black and Hispanic engagers have a higher mean for *carried a gun*. Additionally, black subjects compared to both the baseline group and Hispanic subjects have a higher mean for *sold marijuana* and *sold other drugs* and a lower mean for *aggressive II* and *income I*. Finally, compared to Hispanic subjects, black subjects have a lower mean for *drove drunk*.

There are two effects related to placement in a secure facility in the mean component of the model. The first is an indicator effect on the mean count in the current month. The posterior

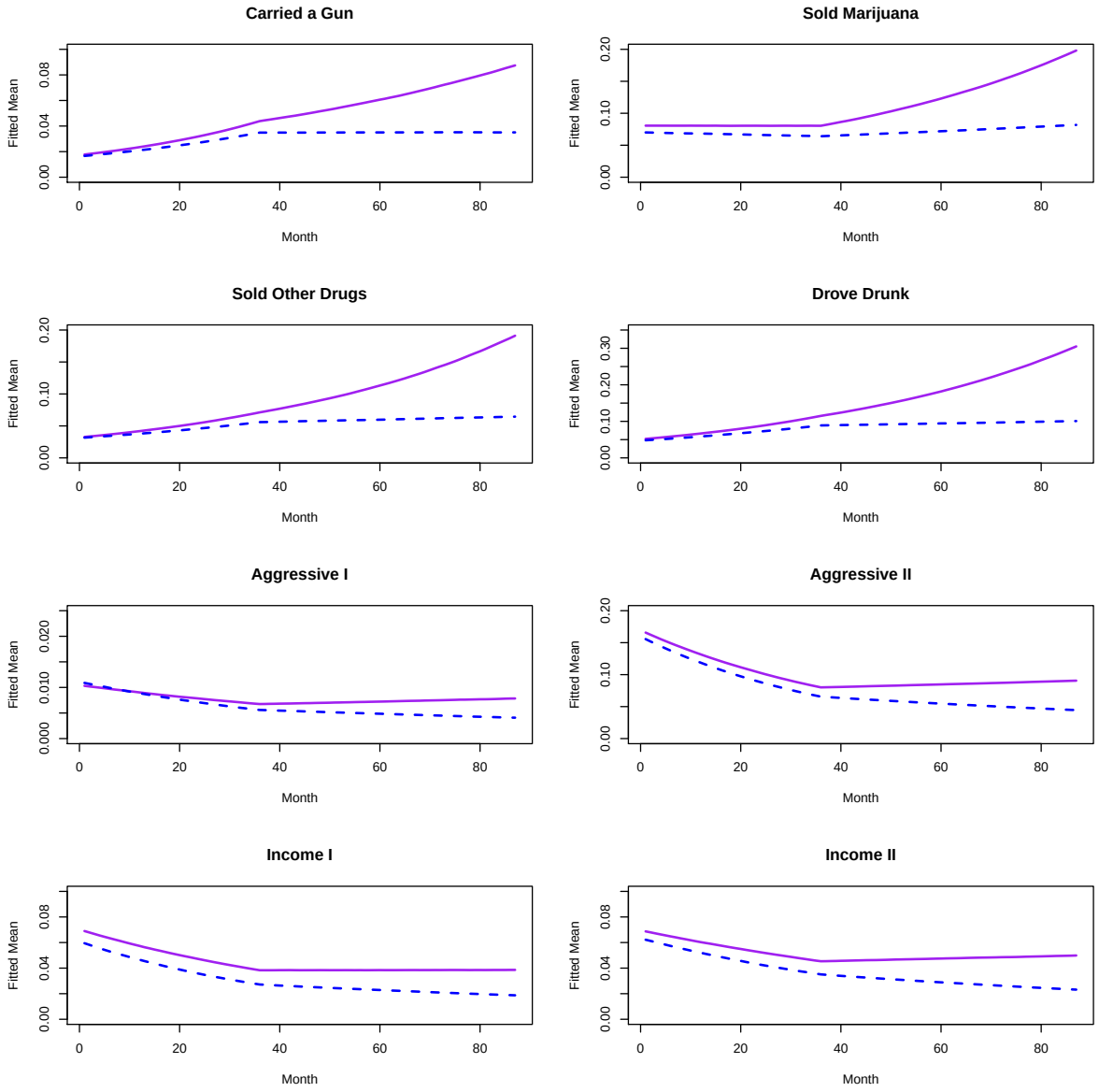

Figure 4.4: Comparison of time trends in the mean component. Fitted values correspond to a non-black, non-Hispanic male subject who spent no time in a secure facility between t_0 and t_j and with length of exposure of 31 days. Posterior medians corresponding to full/reduced model are shaded in solid purple/dashed blue.

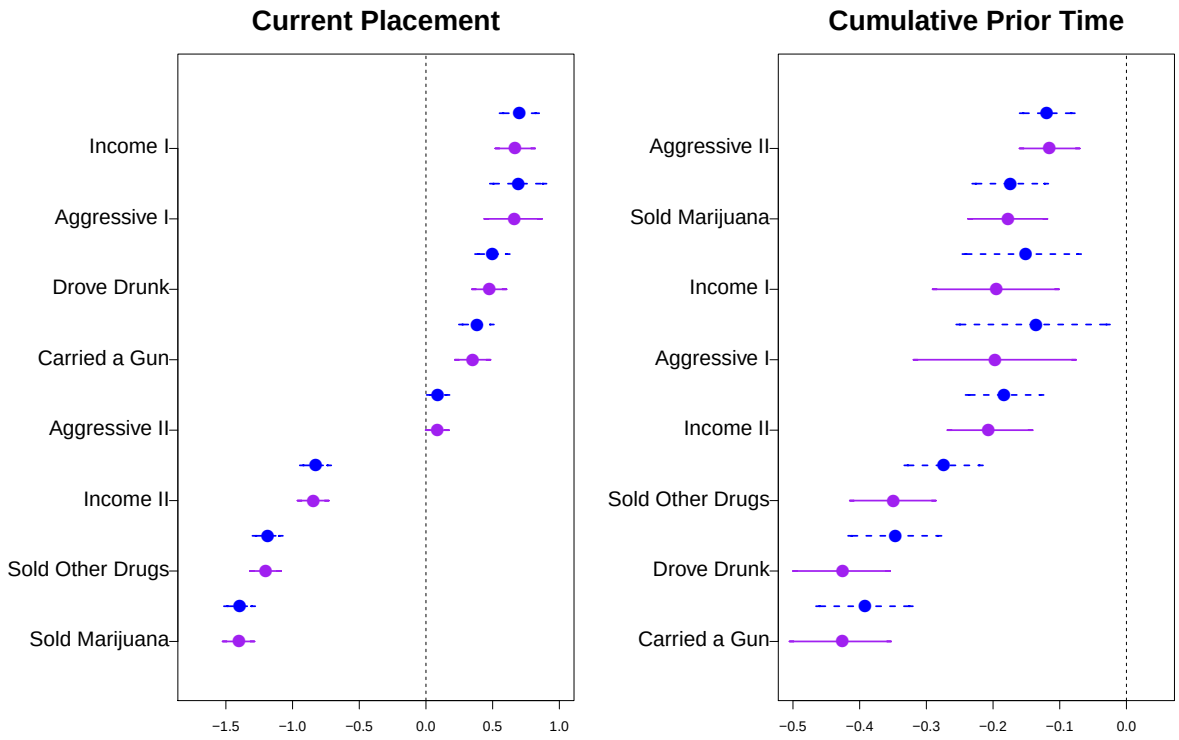

Figure 4.5: Comparison of effect of placement in a secure facility during the current panel (left) and the effect of prior cumulative time spent in a secure facility on the probability of at least one event (right). Credible intervals corresponding to full/reduced model are shaded in solid purple/dashed blue.

medians and credible intervals for this effect are shown in the left column of Figure 4.5. For three of the outcomes, *sold marijuana*, *sold other drugs* and *income I*, under both the full and reduced models, placement in a secure facility is associated with a lower mean in the current panel. On the other hand, again for both models, for the remaining outcomes, placement in a secure facility is associated with a higher mean in the current panel. Note that whether placement occurs before or after criminal activity is unknown. It may be that, for example, a subject experienced a period with a high event rate for *carried a gun* which led to placement in a secure facility.

The second is the effect of the cumulative time in a secure facility in the previous months on the mean count in the current month; posterior summaries for this effect are shown in the right column of Figure 4.5. For each of the eight outcomes, under the full and reduced

Table 4.1: Posterior medians obtained from the full model for σ_k^2 and the fraction of variability explained by the shared random effect in the mean component.

	Carried a gun	Sold marijuana	Sold other drugs	Drove drunk	Aggressive I	Aggressive II	Income I	Income II
$\sigma^2_{b_k}$	0.99	0.69	0.89	1.25	− − −	0.49	1.10	0.46
% variability	0.77	0.70	0.74	0.60	1.00	0.64	0.69	0.84

models, a longer cumulative time in a secure facility in the previous months is associated with a significantly lower mean in the current panel. For all of the outcomes, the effect of cumulative time spent in a secure facility during prior months in the mean component is approximately the same or slightly attenuated under the reduced model.

The posterior medians and credible intervals for the factor loading parameters associated mean model component are displayed in the right column of Figure 4.3. These estimated factor loading parameters vary substantially across the outcomes; this variability is lowest for *aggressive II* and highest for *aggressive I*.

Pairwise estimates (not shown) obtained from the full model of Spearman's rank correlation coefficient for the posterior median estimates of the outcome- and subject-specific random intercepts in the mean component, b_{ik}, are all fairly close to zero, indicating that the shared random effect adequately captures the correlation structure. Table 4.1 provides posterior medians for the variance of the random effect representing additional heterogeneity beyond the shared random effect in the mean component, $\sigma^2_{b_k}$. This variance is largest for *drove drunk*, indicating there may be additional variation in the Poisson mean for *drove drunk* across subjects, distinct from the other outcomes. For each run of the MCMC samples, the empirical variances for the random intercept and common component, $s^2_{b_{ik}+\lambda_{v_k}V_i}$ and $s^2_{\lambda_{v_k}v_i}$, respectively, are calculated. The fraction of variability explained by the common factor is calculated as the ratio $s^2_{\lambda_{v_k}v}/s^2_{b_{ik}+\lambda_{v_k}v_i}$. Table 4.1 also displays the posterior medians for the fraction of variability explained by the common factor for each outcome. The shared random effect accounts for 60% to 77% (excluding *aggressive I*) of the variability in the mean component; hence the majority of the variability in the mean component is absorbed by the shared random effect.

Probability of Resolution: Finally, we examine the probability that the process generating

events resolves following an assessment period with one or more events. Compared to male subjects, female subjects are more likely to permanently quit. There are no significant differences in terms of the probability of permanently quitting offending among ethnicities. The posterior median estimates (95% credible interval) of the regression parameters in the permanent quit component corresponding to female, Black and Hispanic subjects are -0.22 (-0.32, -0.12), -0.07 (-0.14, 0.00) and -0.02 (-0.09, 0.06), respectively. More cumulative time spent in a secure facility since the start of observation is associated with a higher probability of permanently quitting offending; the posterior median estimate for this effect is -0.07 (-0.09, -0.04). Although this effect is significant, it is not practically meaningful. For example, the probability of permanently quitting for a black, male subject increases from 0.021 (0.017, 0.026) with no prior time in a secure facility to 0.027 (0.022, 0.032) with one year spent in a secure facility.

We investigate how an individual's pattern of offending affects their estimated probability of permanently quitting, computed as $\hat{\alpha}_i = \frac{1}{B}\sum_{b=1}^{B} q_{iL_i}^{(b)}$ where B is the number of MCMC iterations, B=2000 here, and L_i denotes the month following the occurrence of the last observed event for any of the outcomes. Figure 4.6 displays the estimated probability of permanently quitting versus the number of event-free months for subject i following L_i, stratified by the proportion of prior months with at least one event for any of the outcomes. The probability of permanently quitting increases with the number of event-free months following the occurrence of the last observed event. This curve becomes steeper as the proportion of prior months with at least one event increases. For subjects with a very low rate of offending in the prior months, $\hat{\alpha}_i$ does not exceed approximately 0.6. In contrast, for subjects with a high rate of offending in the prior months, the probability of permanently quitting increases sharply with the number event-free months following L_i.

These sorts of insights offer a striking advantage of the full versus the reduced model. Estimates of the probabilities of permanently quitting (after L_i) may be useful for distinguishing between juvenile offenders who will continue engaging in criminal behaviour beyond adolescence and those who will not. A possible approach would be to classify individuals into two

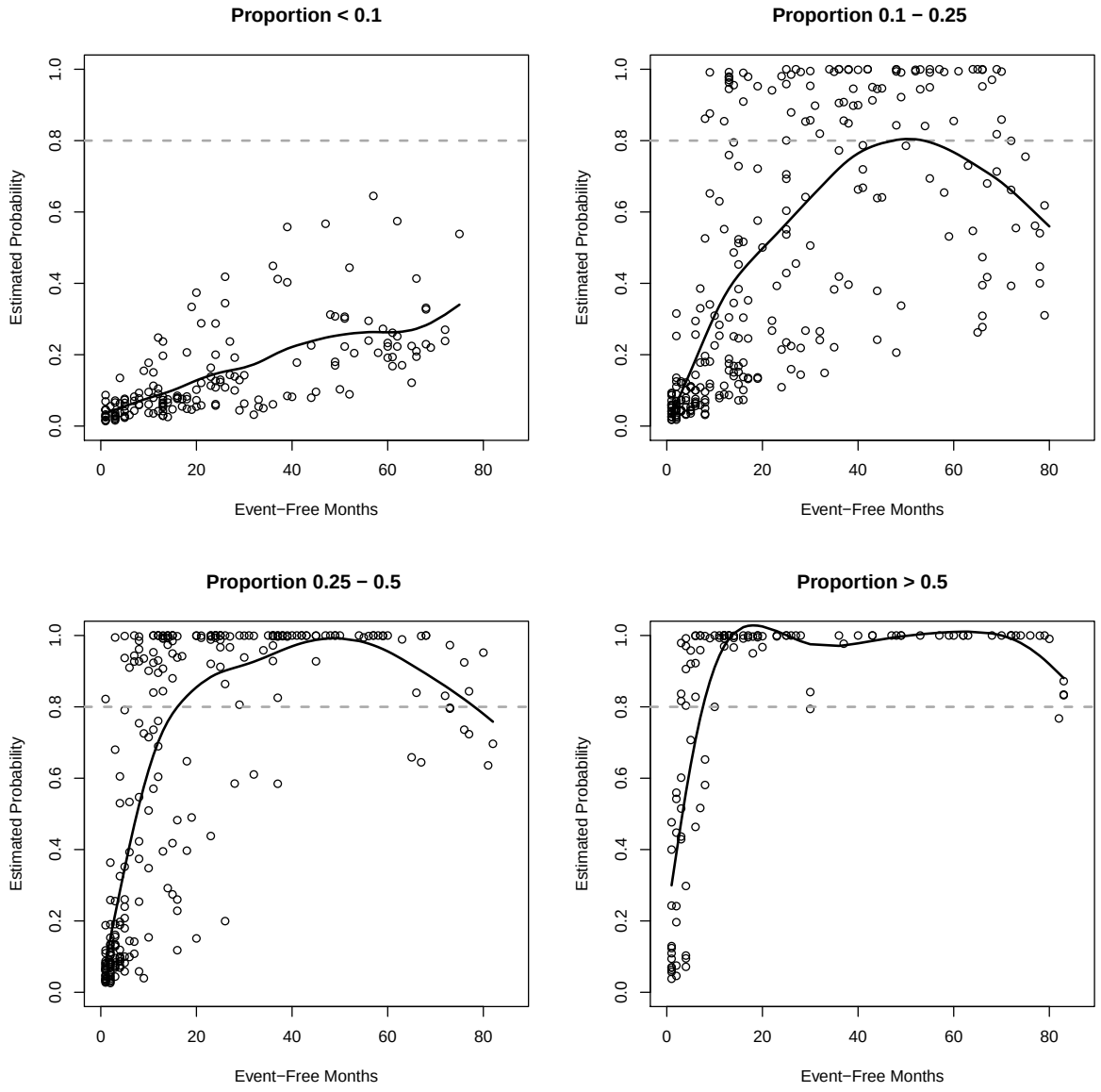

Figure 4.6: Estimated probability of permanently quitting offending, $\hat{\alpha}_i$, versus the number of event-free months following L_i, stratified by the proportion of months prior to L_i with at least one event.

groups, e.g. permanent quitters and nonpermanent quitters. One decision rule would be to classify subjects as permanent quitters if $\hat{\alpha}_i > p_0$, a selected threshold, and nonpermanent quitters otherwise. Using this approach and a threshold of $p_0 = 0.8$, 35% of the subjects were classified as permanent quitters.

Figure 4.7 displays, for each outcome, the expected versus observed number of individuals with a presence of event by month showing no striking evidence of lack of fit. Similar comparisons (not shown here) by gender and ethnicity also show reasonable agreement.

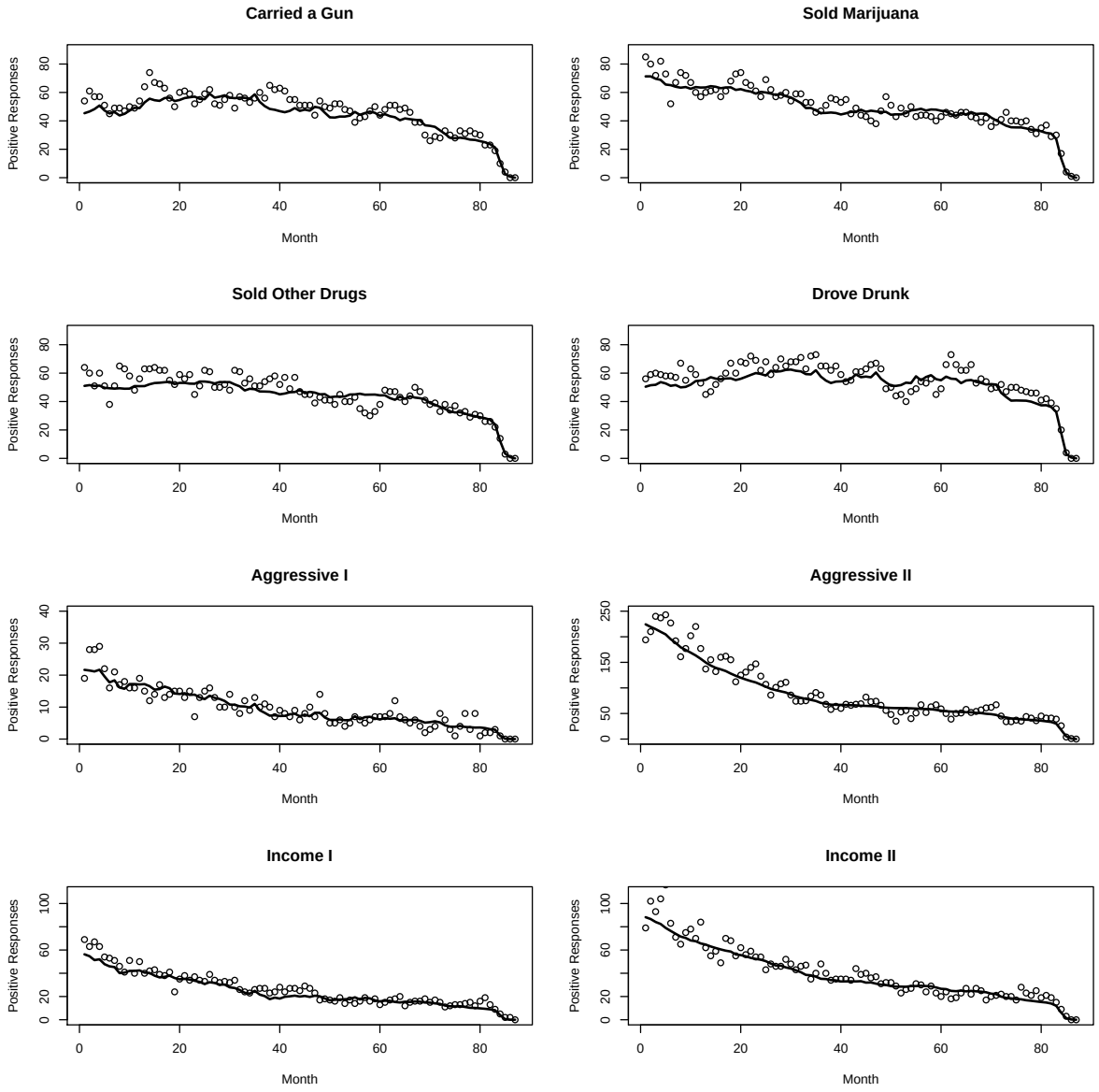

Figure 4.7: The expected number of positive responses over time (lines) and the number of observed positive responses (points).

4.5 Simulation Study

We investigate the added value of incorporating the permanent quitting component to our model. The differences in estimates obtained under the full and reduced models in the analysis of the Pathways to Desistance data prompts an investigation of the potential bias in the mean component of the model when the underlying process generating events can indeed resolve but the reduced model is implemented for analysis. We simulated data corresponding to $N = 1000$ subjects from a simpler joint model for multivariate longitudinal presence/absence data subject to resolution and incorporating non-engagers as well as an exposure variable, as seen in the Pathways to Desistance study. Specifically, we consider $K = 5$ outcomes with rates similar to that observed for *carried a gun*, *sold marijuana*, *aggressive I*, *aggressive II* and *income I*. The true values of the fixed intercepts represent approximately the fitted probabilities corresponding to a male, non-Black, non-Hispanic subject at the first month of the observation period who spent no time in a secure facility during the current month. All model components depend only on one binary covariate x_i denoting gender, simulated independently from a Bernoulli distribution with probability 0.14 which is the about the prevalence of female subjects in Pathways to Desistance study, for $i = 1, \ldots, N$. The length of exposure by month for each outcome reflects approximate average values for these outcomes with the Pathways to Desistance study. The algorithm for data generation is described in the Appendix C.

Using the Bayesian methodology described in Section 4.3, we obtain the joint posterior distribution for all parameters under the full model. For each of 250 simulated data sets, we run two chains, each for an initial 2000 burn-in iterations followed by an additional 5000 iterations used for inference. As well, we fit the reduced model without the permanent quit component using the simulated data.

The bias for the gender effect in the mean component under the full and reduced models is reported in Table 4.2. We observe that the bias is smaller for outcomes where the proportion of positive responses is higher. It may be difficult to obtain accurate parameter estimates in

settings with very sparse data as seen here for *sold marijuana*, *carried a gun* and *aggressive I* (approx. 1% positive responses). As seen in our analysis of the Pathways to Desistance data, the estimate of gender in the mean component is lower under the reduced model, relative to the full model, resulting in an increase in bias. The increase in bias is most apparent for *aggressive II* which is the outcome that corresponds to the highest proportion of positive responses.

Table 4.2: Average bias for gender effect in mean component under the full and reduced models and in the permanent quit component under the full model across 250 simulated data sets. The first column displays the true parameter value, the second and third columns display the average posterior median, the average bias obtained under the full model and the fourth and fifth columns reports the the average posterior median and the average bias under the reduced model. Outcomes are listed in ascending order according to the proportion of positive responses.

	True	Full	Bias Full	Reduced	Bias Reduced
Sold Marijuana	−0.500	−0.746	−0.246	−0.996	−0.496
Carried a Gun	−0.800	−1.131	−0.331	−1.257	−0.457
Aggressive I	0.150	−0.176	−0.326	−0.443	−0.593
Income I	0.650	0.465	−0.185	−0.197	−0.847
Aggressive II	−0.200	−0.269	−0.069	−0.594	−0.394
ϕ	−0.250	−0.249	0.001		

Table 4.2 also displays the average posterior median and the average bias for the gender effect in ϕ (the permanent quit model component). The bias for this effect is substantially smaller than the corresponding bias for any of the outcomes in the mean component. This illustrates another aspect of the conceptualization of the latent variable corresponding to this effect underlying the multivariate longitudinal data. The model for the permanent quit component borrows information across different outcomes effectively increasing the sample size used for parameter estimation.

We study here the use of $\hat{\alpha}_i$ for classifying juvenile offenders as permanent quitters as discussed in Section 4.2. We vary the length of the observation period as well as the threshold p_0 to determine the effects on the sensitivity and specificity of this classification process. Sensitivity is calculated as $\frac{1}{|Q|}\sum_{i\in Q} I\{\hat{\alpha}_i > p_0\}$, where Q is the set of true permanent quitters and $|\bullet|$ is the cardinality of the set, while, specificity is computed as $\frac{1}{|S\setminus Q|}\sum_{i\in S\setminus Q} I\{\hat{\alpha}_i \leq p_0\}$, where S is the set of all subjects. The length of the observation period takes values 36, 60 and 84 months, while p_0 takes values from 0.3 to 0.95 in increments of 0.05.

Figure 4.8 displays the mean sensitivity and specificity at each threshold p_0 by the length of the observation period, T. Sensitivity increases as the length of the observation period increases with the sensitivity remaining above approximately 0.8 for thresholds $p_0 \in [0.3, 0.95]$ when the observation period is 84 months long. Regardless of the length of observation, the specificity increases from approximately 0.9 for $p_0 = 0.3$ to one for $p_0 = 0.95$. Overall, these results indicate that our methodology may be an useful tool to accurately identify permanent and nonpermanent quitters.

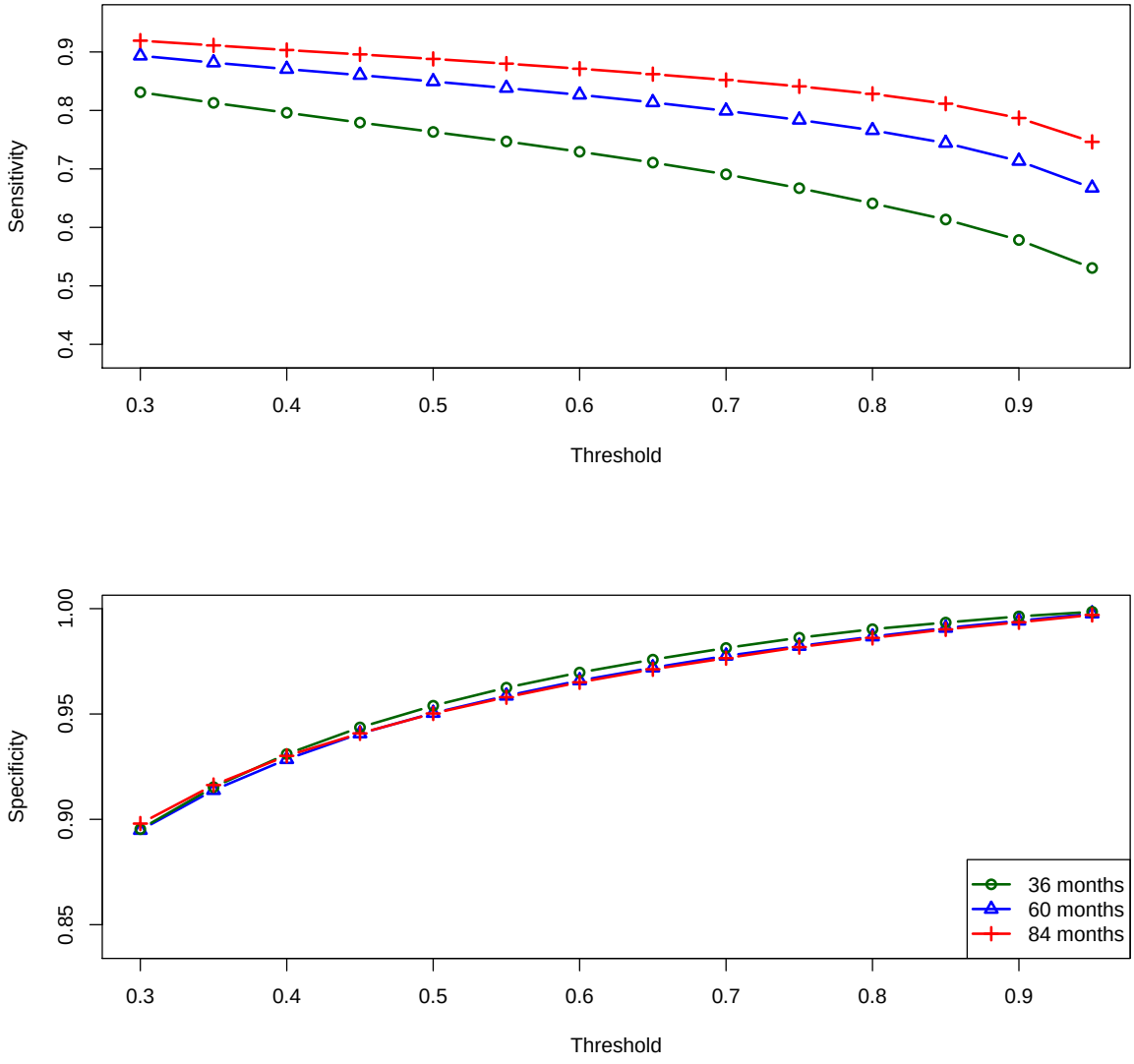

Figure 4.8: The mean sensitivity (top) and specificity (bottom) from 250 simulated data sets with observation periods covering 36, 60 and 84 months.

4.6 Discussion

The approach of utilizing a modeling framework where the process generating events may resolve offers new insights on processes related to offending patterns. The observed decrease in offending over time, across all outcomes, is primarily due to long periods without recurrence at the end of the observation period corresponding to a growing subgroup of subjects who seem to have permanently quit offending. On the other hand, for those subjects who continue to offend, the frequency of offending remains constant or increases over time. Importantly, in our analysis both gender and cumulative time spent in a secure facility since the start of the study were found to significantly affect the probability of permanently quitting. However, the magnitude of the increase associated with increased time in a secure facility was not seen to be meaningful in this application.

Under less complex models, omitting the possibility of the underlying process generating the events resolving, we are unable to distinguish between subjects who are no longer at-risk to offend and subjects with low event rates. The simulation and empirical studies demonstrate that omitting the permanent quit component of the model when the possibility of resolution of an underlying process is justified can yield biased estimates of parameters in the mean component. As well, simulations indicate that our methodology which utilizes data pertaining to individual's engagement in antisocial or illegal activity during the entire observation period is able to accurately identify permanent and nonpermanent quitters.

Our model considers settings where resolution of the event generating process occurs following the presence of an event for any of the outcomes. A natural extension is accommodating the resolution of the underlying process at some specific intervention points in the observation period, for example, following a placement in a secure facility or perhaps some treatment. This may also lead to different interpretations reqarding q_{ij} and ϕ_{ij}.

The resolution process may also differ from outcome to outcome whereby an underlying process generating events for each outcome may resolve at some point during the observation

period. In this work, by assuming a single underlying resolution process, the model is driven by the outcome that is least likely to resolve. The rationale for utilizing a common underlying resolution process in our motivating context was two-fold. First, a major goal of the Pathways to Desistance study is to reliably distinguish between juvenile offenders who will continue criminal behaviour beyond adolescence and those who will not. The conceptualization of a single resolution process directly addresses this objective. Second, due to the limited number of positive responses for all of the outcomes, with possible exception of *aggressive II*, there is insufficient data to permit such an extension which allows resolution for each outcome. However, in other studies such an extension might be important. As well, the development of methods to identify which outcomes convey the most information for the parameters of a single resolution process may add substantially to our understanding of the desistance process.

Under the proposed approach, the probability of permanently quitting is calculated using an individual's data collected over a specified window of time. In our motivating context, decisions concerning, for example, the placement of a subject in a secure facility versus enrollment in a community-based treatment would be made at some specific evaluation points and, hence, such an approach is useful. An alternative approach would be to consider real-time predictions of the probability of the underlying process generating events resolving. Recently, there has been considerable interest in the development of methods for real-time individual predictions, particularly in medical settings. These methods utilize joint models to dynamically predict a subject's risk of occurrence of an event using information pertaining to their medical history. For example, Mauguen et al. (2013) established dynamic predictions of the risk of death using history of cancer recurrences where predictions can be updated following the occurrence of a new event. In some contexts, it may be useful to update the probability of the underlying process resolving following each event-free assessment period. Specifically, we could compare the current censored time since the last occurrence to a gap time distribution based an individual's previous pattern of recurrence of events. The development of such methods is underway.

Our analysis indicates that a higher cumulative time spent in a secure facility during prior

months is associated with a lower rate of offending and a marginally higher probability of permanently quitting offending. It is unclear how these effects cumulate over longer time periods and the impact of placement in a secure facility at different periods within adolescence. Adapting the approach for modeling the cumulative effects of time-dependent exposure proposed by Sylvestre and Abrahamowicz (2009) may provide insights in this regard. Under this approach, cumulative effects of exposure, weighted by recency, are estimated using cubic regression splines. In this work, placement in a secure facility refers to several distinct types of institutional settings beyond incarceration including, for example, substance abuse treatment units and facilities which target mentally ill adolescents. We also note that exposure considered here does not account for placement in a secure facility prior to the start of the study. Future work will incorporate such information, as well as age effects and information on history of offending prior to study enrollment.

More flexible correlation structures for the random effects could be implemented. Correlated random effects that follow a stationary multivariate autoregressive process could be utilized to incorporate correlation between observations with the same subjects across time. However, computationally efficient estimation of a covariance matrix for correlated random effects would need to be developed, especially for higher dimensional settings as seen here.

References

Agresti, A. (1997). A model for repeated measurements of a multivariate binary response. *Journal of the American Statistical Association* **92**, 315-321.

Blumstein, A., Farrington, D. P., & Moitra, S. (1985). Delinquency careers: innocents, desisters, and persisters. *Crime and Justice* **6**, 187-219.

Carlin, J. B., Wolfe, R., Brown, C. H., & Gelman, A. (2001). A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics* **2**, 397-416.

Chen, Z., & Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59**, 762-769.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 355-366

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1996). Efficient parametrizations for generalized linear mixed models. *Bayesian Statistics* **5**, 48-74.

Luo, S., Crainiceanu, C. M., Louis, T. A., & Chatterjee, N. (2008). Analysis of smoking cessation patterns using a stochastic mixed-effects model with a latent cured state. *Journal of the American Statistical Association* **103**, 1002-1013.

Mauguen, A., Rachet, B., Mathoulin-Pélissier, S., MacGrogan, G., Laurent, A., & Rondeau, V. (2013). Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models. *Statistics in Medicine* **32**, 5366-5380.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. *Psychological Review* **100**, 674701.

Mulvey, E. P., Steinberg, L., Fagan, J., Cauffman, E., Piquero, A. R., Chassin, L., et al. (2004). Theory and research on desistance from antisocial activity among serious adolescent offenders. *Youth Violence and Juvenile Justice* **3**, 213-236

Mulvey, E. P., Steinberg, L., Piquero, A. R., Besana, M., Fagan, J., Schubert, C., et al.

(2010). Trajectories of desistance and continuity in antisocial behavior following court adjudication among serious adolescent offenders. *Development and Psychopathology* **22**, 453-475.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, Austria.

Ribaudo, H. J., & Thompson, S. G. (2002). The analysis of repeated multivariate binary quality of life data: a hierarchical model approach. *Statistical Methods in Medical Research* **11**, 69-83.

Rondeau, V., Schaffner, E., Corbire, F., Gonzalez, J. R., & Mathoulin-Plissier, S. (2013). Cure frailty models for survival data: Application to recurrences for breast cancer and to hospital readmissions for colorectal cancer. *Statistical Methods in Medical Research*, **22**, 243-260.

Schubert, C. A., Mulvey, E.P., Steinberg, L., Cauffman, E., Losoya, S., Hecker, T., Chassin, L., et al. (2004). Operational Lessons from the Pathways to Desistance Project. *Youth Violence and Juvenile Justice* **3**, 237-255

Scott, S. L., James, G. M., & Sugar, C. A. (2005). Hidden Markov models for longitudinal comparisons. *Journal of the American Statistical Association* **100**, 359-369.

Shen, H., & Cook, R. J. (2014). A dynamic Mover-Stayer model for recurrent event processes subject to resolution. *Lifetime Data Analysis* **20**, 404-423.

Shen, H., & Cook, R. J. (2015). Analysis of intervalcensored recurrent event processes subject to resolution. *Biometrical Journal* **57**, 725-742.

Sylvestre, M. P., & Abrahamowicz, M. (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in Medicine* **28**, 3437-3453.

Wall, M. M., & Li, R. (2009). Multiple indicator hidden Markov model with an application to medical utilization data. *Statistics in Medicine* **28**, 293-310.

Chapter 5

Future Work

Directions for future work will consider extensions of the modeling framework where the process generating events may resolve as well as Bayesian methods for handling missing data.

5.1 Outcome-specific Process Resolution

In Chapter 4 of this thesis, we develop a modeling framework which utilizes a common latent variable representing the resolution of the process generating events, borrowing information across different outcomes. This assumption of a single underlying resolution process results in the probability of permanently quitting being driven by the outcome that is least likely to resolve. While this suited the criminal behaviour context considered, allowing the resolution process to differ from outcome to outcome whereby different underlying processes generating events for each outcome lead to resolution at some point during the observation period may be important in other situations.

Here, we may consider latent indicators, q_{ijk}, denoting whether or not the underlying process generating events for subject i and outcome k has resolved by time t_{j-1} and model the probability of permanently quitting as follows

$$\mathrm{P}\left(q_{ijk}=1\middle|\left\{q_{i(j-1)k}=0, y_{i(j-1)k}=1, \boldsymbol{x}_{\phi_{ij}}\right\}\right)=\phi_{ijk}=\exp\{-\exp(\boldsymbol{x}_{\phi_{ij}}^{'}\boldsymbol{\beta}_{\phi_k}+\lambda_{u_k}u_i)\} \tag{5.1}$$

where $\boldsymbol{x}_{\phi_{ij}}$ is a $l_\phi \times 1$ vector of covariates, $\boldsymbol{\beta}_{\phi_k}$ is a vector of corresponding regression parameters, u_i is a subject-specific random effect and λ_{u_k} is a factor loading parameter representing outcome-specific variability related to u_i.

This model specifies independent subject-specific random effects for each model component. Forms which include a single subject-specific random effect, shared across outcomes in multiple components of the model may be considered.

Such an extension raises concerns about model selection as we would need to determine if it is more appropriate to assume outcome-specific resolution processes or a common resolution process in a given application. Wall and Li (2009) discussed testing the hypothesis of a shared common state underlying several longitudinal outcomes in a hidden Markov model where an unobserved health state governs the counts of several types of medical encounters. In their model, if the behaviour observed for one outcome is inconsistent with the remaining outcomes, it could be detected by examining the estimated regression parameter associated with the underlying state in the mean model for the outcome-specific counts. However, in our modeling framework such an approach is not available. In this case, model selection would rely on measures of fit and model assessment. Gelman, Hwang and Vehtari (2014) recommended the use of the Watanabe-Akaike information criterion (Watanabe, 2010) over the more commonly utilized deviance information criterion (DIC, Speighlhalter et al. 2002), particularly for mixture models as considered in this thesis. The Bayesian framework utilized in this thesis facilitates the use of posterior predictive assessments based on any parameter-dependent function, or so-called discrepancy (Gelman, Meng and Stern, 1996). This approach allows us to tease apart the impact of modeling choices on the ability of our model to capture key aspects of the data. For example, in a joint analysis of several count outcomes related to sexual

behaviour, Zhu and Weiss (2012) examined the ability of their model to accurately model high activity portions of the study population using posterior predictive distributions. In the context of the criminal behaviour study, it may be of interest to examine the distribution of the number of event-free months following the occurrence of the last observed event.

5.2 Dynamic Prediction of the Probability of Permanently Quitting

The classification of juvenile offenders as permanent or nonpermanent quitters discussed in Chapter 4 is based an individual's data collected over a window of time. In the criminal behaviour context, decisions concerning, for example, the placement of a subject in a secure facility versus enrollment in a community-based treatment, would be made at some specific evaluation points and, hence, such an approach is useful. In other settings, it may be beneficial to consider real-time predictions of the probability of the underlying process generating events resolving.

Recently, there has been considerable interest in the development of methods for real-time individual predictions, particularly in medical settings. These methods utilize joint models to dynamically predict a subject's risk of occurrence of an event using information pertaining to their medical history. Rizopoulos (2011) discussed the prediction of survival probabilities for patients infected with the human immunodeficiency virus based on their longitudinal CD4 cell count measurements. Additionally, the capability of the longitudinal marker to differentiate between subjects who experience an event within a specified time frame, and those who do not, was assessed. Mauguen et al. (2013) established dynamic predictions of the risk of death using history of cancer recurrences where predictions can be updated following the occurrence of a new event.

One approach may be to update the probability of the underlying process resolving following each event-free assessment period. Specifically, we could compare the current censored

time since the last occurrence to a gap time distribution based an individual's previous pattern of recurrence of events. This requires the estimation of subject-specific recurrent event gap time distributions. Peña et al. (2001) proposed Nelson-Aalen and Kaplan-Meier-type estimators for distribution functions governing the time to occurrence of a recurrent event in the presence of censoring. The parallel to the survival setting yields a natural framework for extensions involving covariates, Cox-type regression and frailty models. Recently, Lee et al. (2015), relaxed the commonly used assumptions that individuals are enrolled in a study due the occurrence of an event of interest, and subsequently experience recurrent events of the same type. They developed a nonparametric estimator of the joint distribution of the time from the start of the study to the first event and the gap times between consecutive events.

A two-stage prediction algorithm could be considered that first, following each event-free assessment period, compares an individual's current censored time since the last occurrence to the distribution of gap times between previous occurrences, then calculates the probability of permanently quitting as an increasing function of the time since the last observed occurrence. Previous authors (Li, Wileyto and Heitjan, 2011) have used two-stage algorithms in the context of prediction using frailty models.

5.3 Bayesian Methods for Handling Nonigorable Missing Observations

In our analysis of the data from the Pathways to Desistance study, we consider only data obtained from consecutive follow up interviews with complete data, leading to dropouts. This is not uncommon in large observational studies. Dropouts can be ignored if the dropout process is unrelated to the processes under investigation. However, this may not be the case for longitudinal studies of human behaviour. In a study of sexual behaviour of adolescents, Ghosh and Tu (2009) hypothesized that dropouts may be associated with traits that can be characterized by a lack of discipline. These traits may be related not only to the dropout process, but

also may influence sexual behaviour, motivating the joint analysis of the zero-inflated count outcome and dropout process. Similar arguments may be justified in the context of criminal behaviour of juvenile offenders. As well, when a subject is in a secure facility it is unlikely that they will miss a scheduled interview. Given that it is well known that failure to accommodate informative dropouts may lead to suspect inference (Wu and Carroll 1988; Little 1995; Wu 2007), extending the Bayesian framework of Ghosh and Tu (2009) to accommodate several zero-inflated count outcomes and the effects due to time spent in a secure facility, both on the outcomes and the dropout process, is warranted.

Moreover, in the Pathways to Desistance data set, there are various patterns of missing data, including intermittently missing patterns corresponding to non-responses for a particular question and missed interviews. In cases where several longitudinal outcomes are jointly considered, methods for handling missing data need to account for multiple sources of correlation. Luo et al. (2016) proposed Bayesian methods using multiple imputation for missing multivariate longitudinal data of various types. Under multiple imputation, uncertainty concerning the imputed values is addressed by generating $M > 1$ sets of imputed values for the missing values in the data set as draws from the predictive distribution. Inference across the imputed data sets can be obtained using Rubin's multiple imputation rules (Rubin, 1987). Here, the authors utilized underlying normal variable models for binary, ordinal and continuous data. In other settings, gamma frailty models for underlying Poisson variables could be utilized for count and discrete event time outcomes (Dunson and Herring, 2005). Accommodating zero-inflated count data within such a framework would require some additional computational algorithmic developments.

Assumptions concerning the processes generating missing data may be untestable, but inference can be sensitive to the particular assumptions made. Linero and Daniels (2015) developed a Bayesian framework for continuous-valued longitudinal outcomes which accommodates a sensitivity analysis. The use of such sensitivity analyses is important in the broader context of longitudinal studies.

References

Dunson, D. B., & Herring, A. H. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics* **6**, 11-25.

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

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733-760.

Ghosh, P., and Tu, W. (2008). Assessing sexual attitudes and behaviors of young women: a joint model with nonlinear time effects, time varying covariates, and dropouts. *Journal of the American Statistical Association* **103**, 1496.

Lee, C. H., Luo, X., Huang, C. Y., DeFor, T. E., Brunstein, C. G., & Weisdorf, D. J. (2015). Nonparametric methods for analyzing recurrent gap time data with application to infections after hematopoietic cell transplant. *Biometrics*. doi: 10.1111/biom.12439

Li, Y., Wileyto, E. P., & Heitjan, D. F. (2011). prediction of individual long-term outcomes in Smoking Cessation Trials Using Frailty Models. *Biometrics* **67**, 1321-1329.

Linero, A. R., & Daniels, M. J. (2015). A flexible Bayesian approach to monotone missing data in longitudinal studies with nonignorable missingness with application to an acute schizophrenia clinical trial. *Journal of the American Statistical Association* **110**, 45-55.

Luo, S., Lawson, A. B., He, B., Elm, J. J., & Tilley, B. C. (2016). Bayesian multiple imputation for missing multivariate longitudinal data from a Parkinson's disease clinical trial. *Statistical Methods in Medical Research* **25**, 821-837.

Mauguen, A., Rachet, B., Mathoulin-Pélissier, S., MacGrogan, G., Laurent, A., & Rondeau, V. (2013). Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models. *Statistics in Medicine* **32**, 5366-5380.

Peña, E. A., Strawderman, R. L., & Hollander, M. (2001). Nonparametric estimation with recurrent event data. *Journal of the American Statistical Association* **96**, 1299-1315.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67**, 819-829.

Rubin, D. B. (2004). Multiple imputation for nonresponse in surveys (Vol. 81). New York, NY: John Wiley & Sons.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 583-639.

Wall, M. M., & Li, R. (2009). Multiple indicator hidden Markov model with an application to medical utilization data. *Statistics in Medicine* **28**, 293-310.

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

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**, 214-224.

Appendix A

Supplementary Material for Chapter 2

Table A.1: Summary of the eight outcomes analyzed

Outcome	Prohibited in a secure facility?	Type
Carried a gun	yes	bounded count
Sold marijuana	no	bounded count
Sold other drugs	no	bounded count
Drove drunk	yes	count
Aggressive I	yes	count
Set fire		
Forced someone to have sex		
Killed someone		
Shot someone, bullet hit		
Shot at someone, no hit		
Robbery with weapon		
Aggressive II	no	count
Destroyed/damaged property		
Beat up someone, serious injury		
In a fight		
Beat someone as part of gang		
Income I	yes	count
Broke in to steal		
Shoplifted		
Used check/credit card illegally		
Stole car or motorcycle		
Carjack		
Paid for sex		
Broke into car to steal		
Income II	no	count
Bought/received/sold stolen property		
Robbery no weapon		

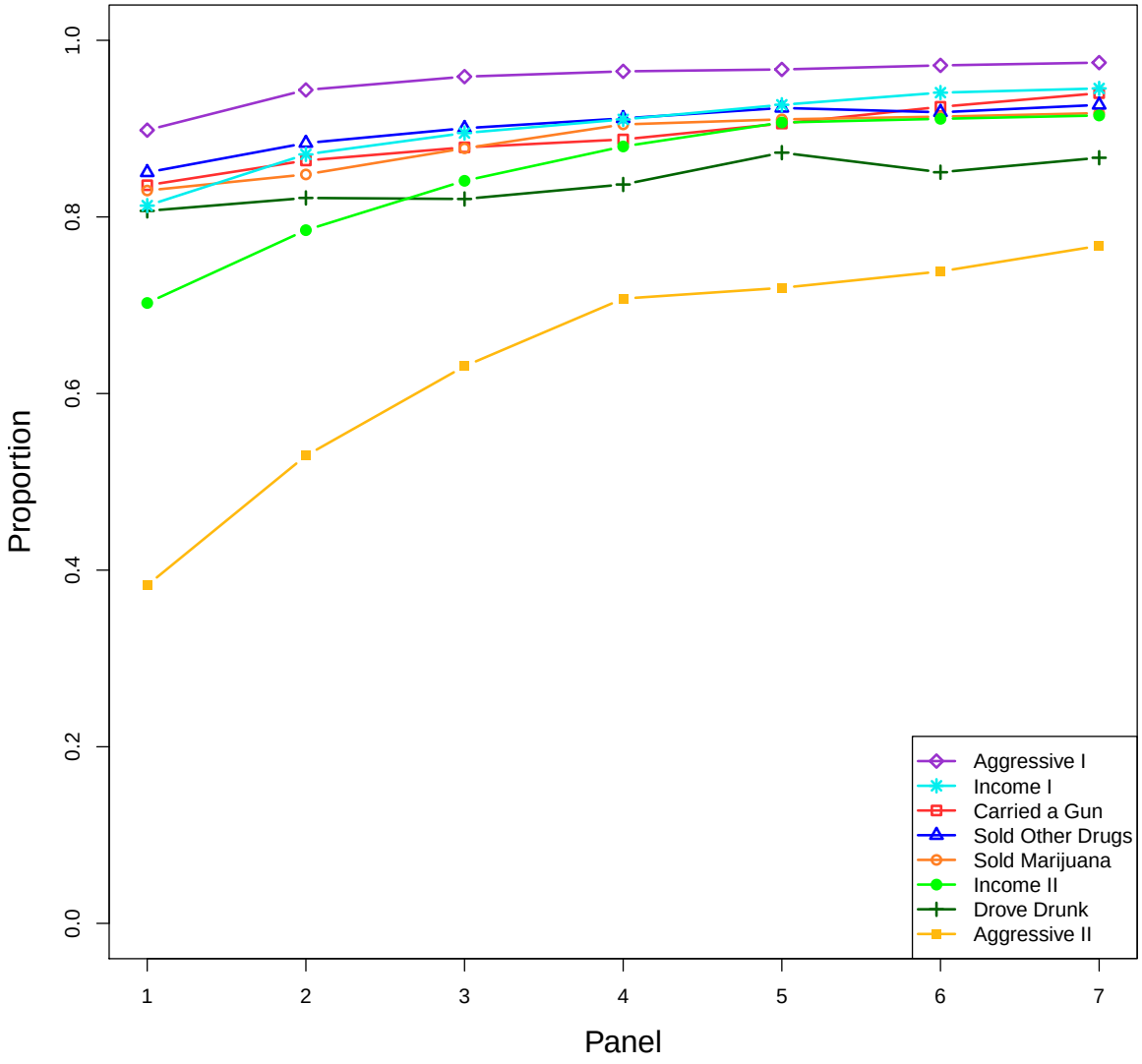

Figure A.1: Proportion of zero counts over time for the eight outcomes analyzed.

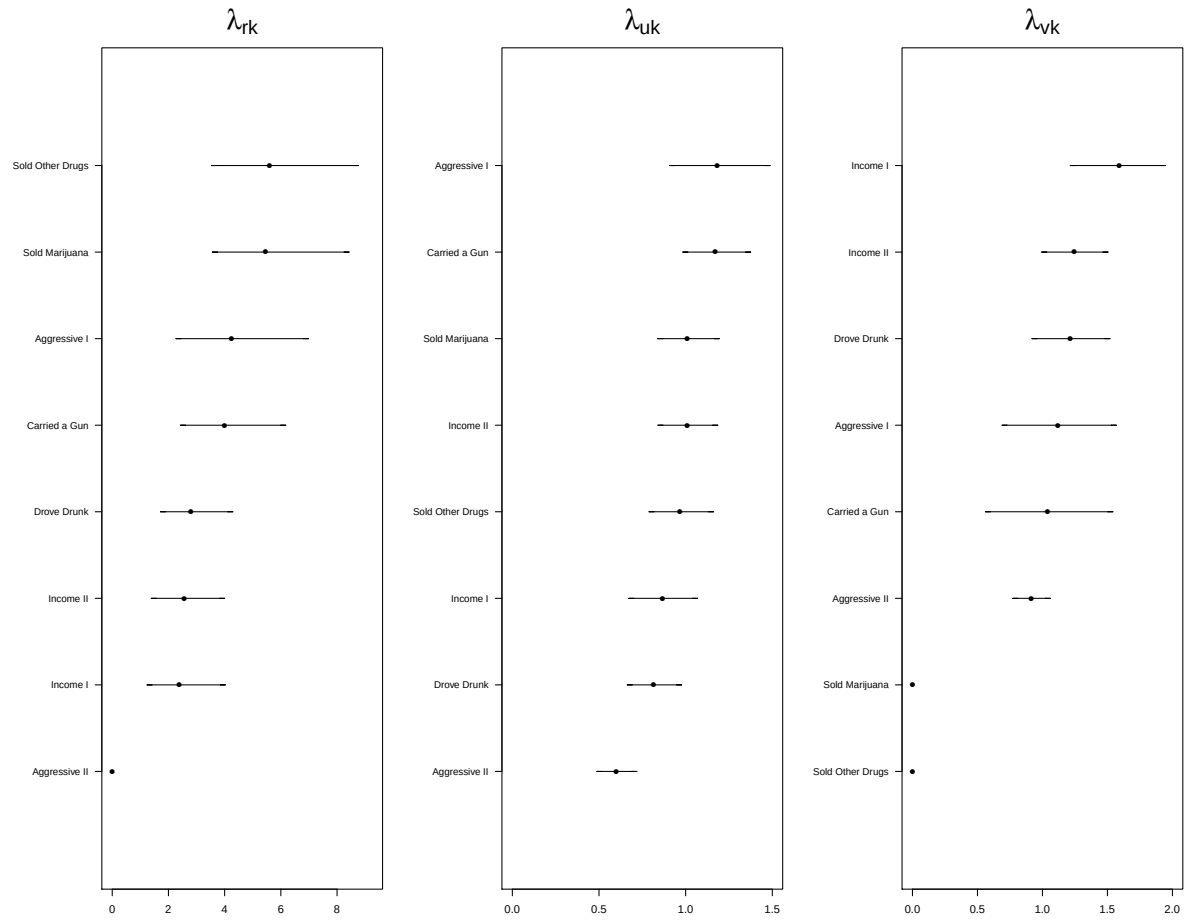

Figure A.2: Posterior medians and 95% credible intervals for the factor loading parameters for the probability of a non-engager (λ_{rk}, left), the probability of a structural zero (λ_{uk}, middle) and mean of the standard count distribution (λ_{vk}, right).

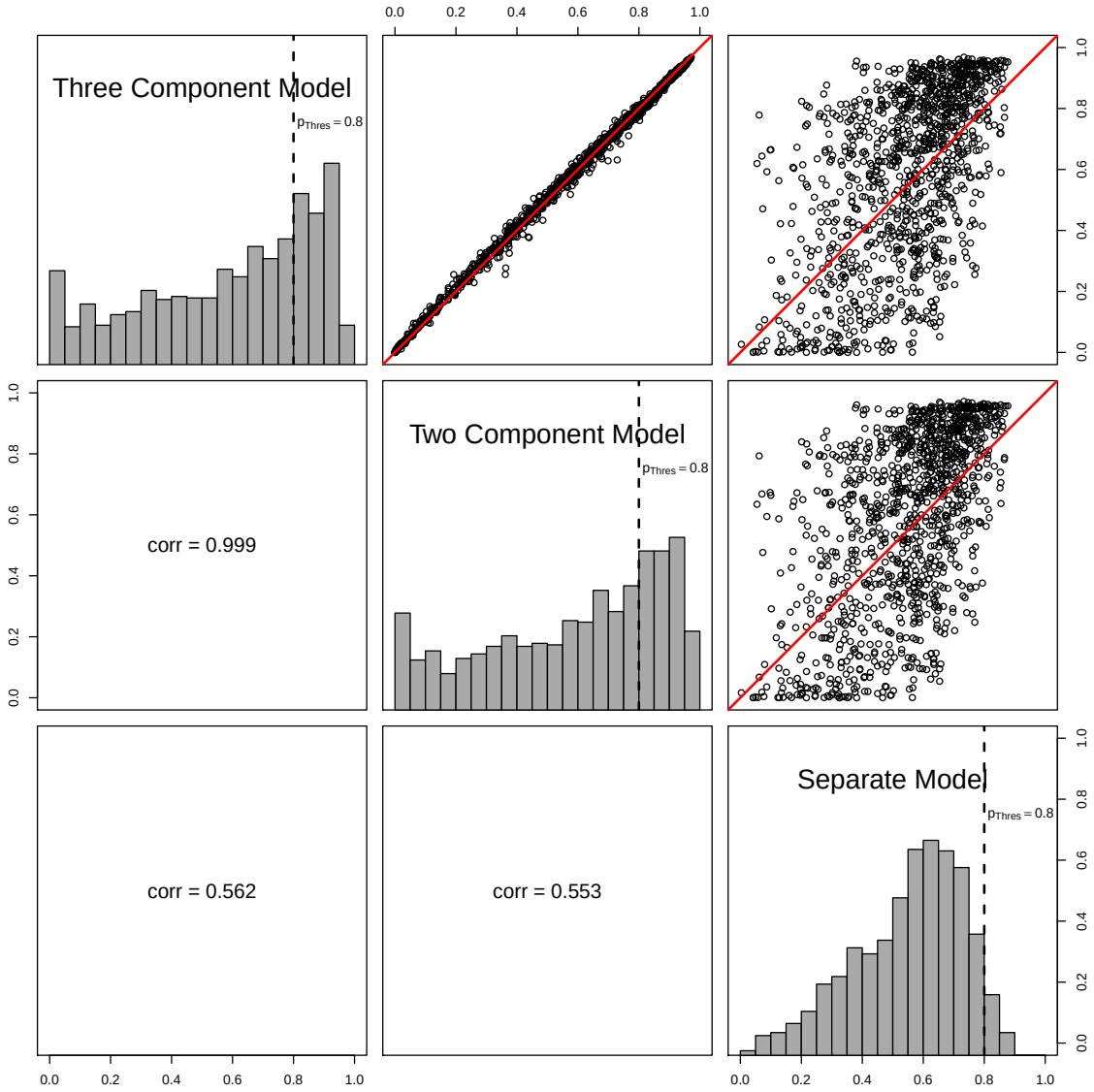

Figure A.3: Comparison of posterior median for the probability of not offending during panel $T_i + 1$ for all individuals.

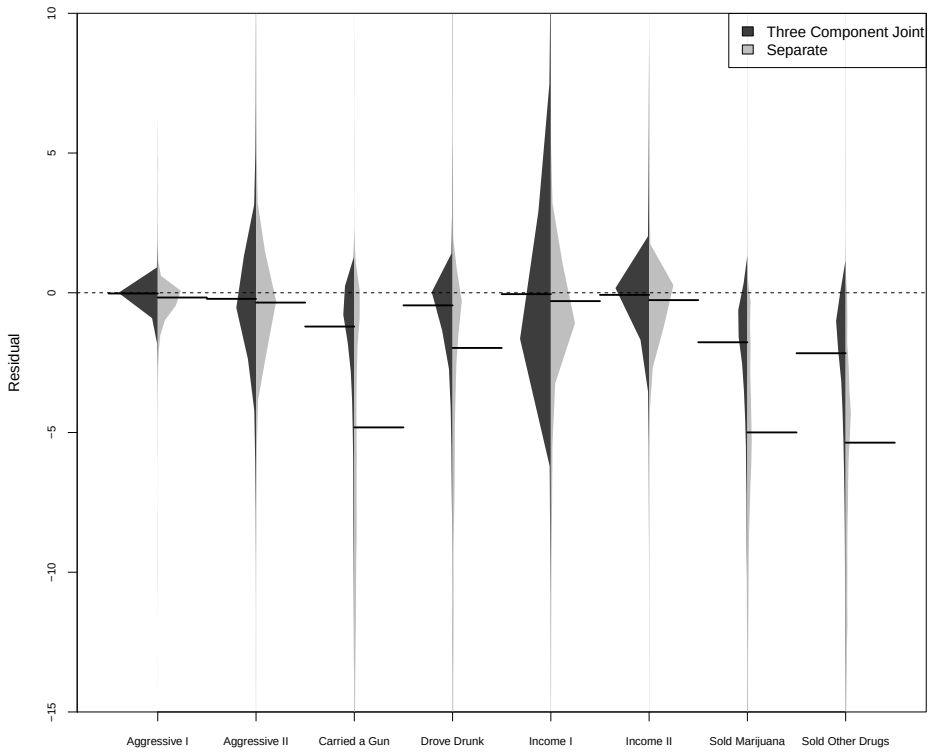

Figure A.4: Comparison of distribution of residuals. The median is denoted by a black line.

Appendix B

Strategies to Consider in the Design of Recurrent Event Studies

Planning studies that are expected to yield zero-inflated Poisson data can be challenging when heaping may be a significant concern. Opportunities to elicit from individuals more accurate data that is less burdensome on participants to record may include obtaining information of presence/absence of events at successive, closely timed assessments rather than a precise count of the number of events that occurred in some window of observation. In this case a key question will be how often such presence/absence data should be collected in order to obtain as high efficiency in the analysis as provided by accurately rendered counts. We describe how practitioners can come to terms with these issues to derive an efficient study design by adapting the methodology utilized in our simulations study.

Selection of Families of Plausible Heaping Distributions: We need to establish a set of relevant heaping distributions that are characterized by a set of parameters. This requires careful examination of the both the data collection process and the raw data. The data collection process and the wording of the survey questions may influence at which points in the distribution one should expect to observe heaping. For example, in the Pathways in Desistance study, at the follow up interview subjects where asked to indicate which months they engaged in an activity

and how many times they engaged in this activity since the baseline interview. As expected, we observed peaks at multiples of 30. Examining plots of the raw data will help identify plausible heaping points. It is important to look for evidence of heaping using the raw data on the scale of the response offered by the survey question. Using smoking cessation data, Bar and Lillard (2012) illustrated that heaping may be obscured by transformations of the raw data. In the criminal behaviour application, we identified three plausible families of heaping distributions based on (i) rounding to multiples of 5, (ii) a proportional odds model where the observed data are coarser for larger values of the true count, and (iii) change points for different levels of coarsening.

Identification of Heaping Distribution which Best Mimics Heaping Observed in Pilot Data: Fit an appropriate regression model to aggregate count pilot data, ignoring any apparent heaping, to obtain rough values for parameter estimates. As in our simulation study, we will subsequently view this as the true data generating model. Using these parameter values, we generate R replicate data sets representing accurately recorded aggregate count data. For each of the families of relevant heaping distributions, we generate R replicate data sets of rounded counts where the values of the heaping parameters reflect hypotheses concerning the heaping mechanisms. We visually compare the distributions of the observed pilot data and the simulated heaped data, averaged over the replicate data sets. Based on this, we select the family of heaping distributions which best mimics the patterns of heaping observed in the pilot data. In Figure B.1, we display the comparison of the distributions of observed data and that of the simulated heaped count data for the three families of plausible heaping distributions considered for criminal behaviour study. Here, we identified the heaping distribution with change points for different levels of coarsening as the most appropriate. Next, tune the values of the heaping parameters for the selected heaping distribution using a measure of discrepancy

$$D = \sum_{y=0}^{L} \frac{|(\ \%\ \text{obs. counts} = y) - (\%\ \text{simulated counts} = y, \text{avg'd over } R \text{ sims.})|}{2}$$

where L is enough large so that the probability of a count exceeding this threshold is of small order. We utilize parameter values that minimize the measure of discrepancy to represent the observed data in the prospective study, denoted H_I in the our simulations. In our motivating example, the measure of discrepancy for H_I as presented in Table 3.1 were 0.089 and 0.120 for DD and AGG, respectively. This represents an improvement in goodness-of-fit, relative to preliminary parameter values for this heaping distribution, displayed in the final row of Figure B.1 ($D = 0.097$ for DD and $D = 0.128$ for AGG). In addition to a heaping distribution that closely resembles the pilot study data, a worst case scenario with more pronounced heaping should be considered. Comparison of the ABIAS for the parameters obtained from the simulated true count and heaped count data provide an indication of the extent to which heaping of the data may introduce bias.

Investigating Efficient Choices of Timing of Longitudinal Presence/Absence Data by Simulation: Using the simulated true count data, we derive R replicate data sets of accurately reported presence/absence data collected at periodic assessment points within the window of observation. We consider several monitoring schemes where we vary the length of time between assessments points, including the most frequent collection schedule plausible given the resource allocation for the prospective study. By comparing the ASE for parameters obtained from the simulated true count data and the simulated presence/absence data collected at varying frequencies, we can determine how often the longitudinal binary should be collected in order to obtain as high (or nearly as high) efficiency as the analysis as provided by accurately reported counts.

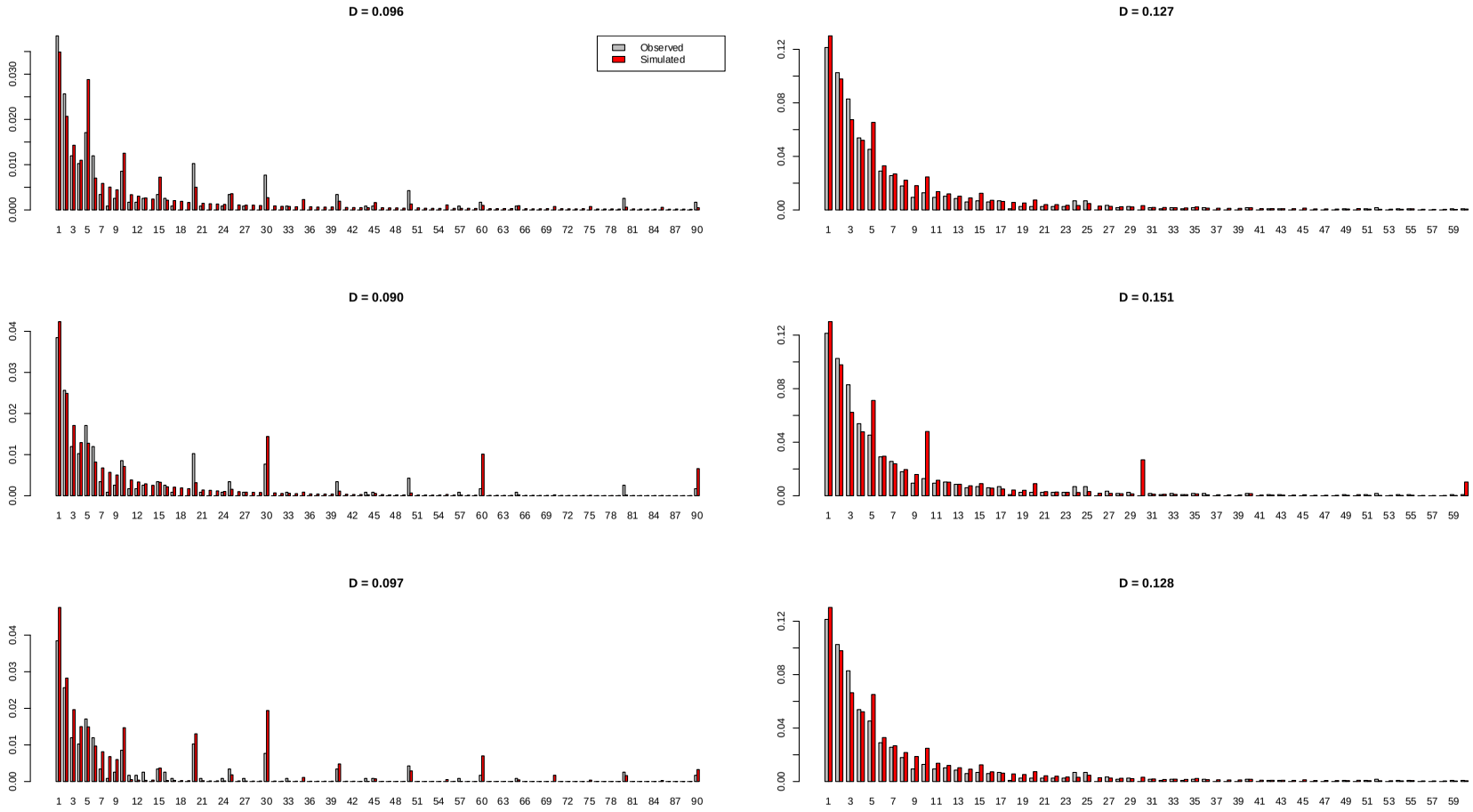

Figure B.1: Comparison of distributions of non-zero counts between simulated rounded count data, averaged over the 500 replicate data sets, and the Pathways to Desistance data for DD (left column) and AGG (right column). The first row corresponds to a heaping distribution where true counts are rounded to multiples of 5, the second row corresponds to heaping distribution based on a proportional odds model and the third row corresponds to heaping distribution H_I with different parameter values. D denote the value of the measure of discrepancy.

Appendix C

Algorithm for Generation of Simulated Data in Chapter 4

In our simulations, the offending pattern is generated using the following procedure.

(1) Set $m = 1$ and at the mth replication, generate $\boldsymbol{r}^{(m)} = (r_1^{(m)}, \ldots, r_N^{(m)})' \sim N(\mathbf{0}, \mathbf{I})$, $\boldsymbol{v}^{(m)} = (v_1^{(m)}, \ldots, v_N^{(m)})' \sim N(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{b_k}^{(m)} = (b_{1k}^{(m)}, \ldots, b_{Nk}^{(m)})' \sim MVN(\mathbf{0}, \sigma_{b_k}^2 \mathbf{I})$ for $k = 1, \ldots, K$.

(2) Generate $s_{ik}^{(m)}$, markers for the outcome-specific non-engagers where (4.2) is given by

$$p_{ik}^{(m)} = \{1 + \exp(-\beta_{p_{0k}} - \beta_{p_{1k}} x_i - \lambda_{r_k} r_i^{(m)})\}^{-1}$$

for $i = 1, \ldots, N$, $k = 1, \ldots, K$.

(3) For outcome-specific engagers, generate the presence/absence of at least one event at the first month where (4.5) is expressed as

$$\zeta_{i1k}^{(m)} = 1 - \exp\{-\exp(\beta_{\mu_{0k}} + \beta_{\mu_{1k}} x_i + \log(z_{i1k}) + \lambda_{v_k} v_i^{(m)} + b_{ik}^{(m)}))\}$$

Note that we assume $q_{i1}^{(m)} = 0$ for all $i = 1, \ldots, N$ as the offending process can only resolve following at least one event.

(4) If $y_{i1k}^{(m)} = 0$ for all $k = 1, \ldots, K$ then set $q_{i2}^{(m)} = 0$, otherwise generate $q_{i2}^{(m)}$ where (4.7) is defined as

$$\phi_{i2}^{(m)} = \exp\{-\exp(\beta_{\phi_0} + \beta_{\phi_1} x_i))\}$$

(5) For outcome-specific engagers, given that $q_{i2}^{(m)} = 0$, generate the presence/absence of at least one event at the second month from a Bernoulli($\zeta_{i2k}^{(m)}$) random variable.

(6) Repeat steps (4) and (5) for $j = 3, \ldots, T = 84$.

Curriculum Vitae

Name:	Erin Lundy
Post-Secondary Education and Degrees:	Carleton University Ottawa, ON 2006 - 2010 B.Math McGill University Montreal, QC 2010 - 2012 M.Sc. University of Western Ontario London, ON 2012 - 2016 Ph.D.
Honours and Awards:	Dean's Doctoral Scholarship 2012 - 2014 Queen Elizabeth II Scholarship 2014 - 2016
Related Work Experience:	Teaching Assistant The University of Western Ontario 2012 - 2014 Statistical Consultant The University of Western Ontario 2014 - 2015

Publications:

Dean, C. B. and Lundy, E. R. (2016). Overdispersion. Wiley StatsRef: Statistics Reference Online. 1-9.