

Medical University of South Carolina

MEDICA

MUSC Theses and Dissertations

2016

Statistical Methods for Modeling Count Data with Overdispersion and Missing Time Varying Categorical Covariates

Elizabeth Holly Payne

Medical University of South Carolina

Follow this and additional works at: <https://medica-musc.researchcommons.org/theses>

Recommended Citation

Payne, Elizabeth Holly, "Statistical Methods for Modeling Count Data with Overdispersion and Missing Time Varying Categorical Covariates" (2016). *MUSC Theses and Dissertations*. 409.

<https://medica-musc.researchcommons.org/theses/409>

This Dissertation is brought to you for free and open access by MEDICA. It has been accepted for inclusion in MUSC Theses and Dissertations by an authorized administrator of MEDICA. For more information, please contact medica@musc.edu.

**Statistical Methods for Modeling Count Data with
Overdispersion and Missing Time Varying Categorical Covariates**

By

Elizabeth Holly Payne

A dissertation submitted to the faculty of the Medical University of South
Carolina in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in the College of Graduate Studies.


Department of Public Health Sciences

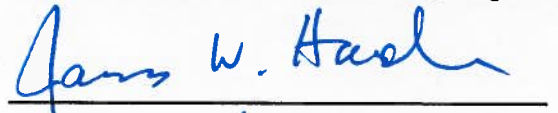
2016

Approved By:


Chairman, Advisory Committee


Mulugeta Gebregziabher


Leonard Egede


James Hardin


Viswanathan Ramakrishnan


Anbesaw Selassie

**Statistical Methods for Modeling Count Data with
Overdispersion and Missing Time Varying Categorical Covariates**

By

Elizabeth Holly Payne

A dissertation submitted to the faculty of the Medical University of South
Carolina in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in the College of Graduate Studies.

Department of Public Health Sciences

2016

Approved By:

Chairman, Advisory Committee

Mulugeta Gebregziabher

Leonard Egede

James Hardin

Viswanathan Ramakrishnan

Anbesaw Selassie

For Jean-Paul Hill, Keith and Beth Payne, Sarah Payne, and Vida Payne

Your constant love and support is everything.

For Natalie Gottlieb, Meagan Herman, Liz Stewart, Katie Morton,

Jennifer Casler, Grace Elliott, Erin Corder, and Brandy Higson

You made school fun and kept me sane the whole time.

And in loving memory of Dr. Holland Payne

I will always be your “can-do girl”.

Psalm 62:7

ACKNOWLEDGEMENTS

This project was supported by CIN 13-418 at the Center of Innovation for Health Equity at the Ralph H. Johnson Medical Center in affiliation with the US Department of Veterans Affairs, P.I.: Leonard Egede, M.D. and Mentor: Mulugeta Gebregziabher. My first and second year studies in the PhD program were supported by Biostatistics Training for Basic Biomedical Research training grant, NIH T32 GM074934, P.I.: Elizabeth Slate, Ph.D. and Mentor: Viswanathan Ramakrishnan, Ph.D.

None of this would be possible without the kind and committed mentorship of Dr. Mulugeta Gebregziabher, Dr. Viswanathan Ramakrishnan, Dr. Leonard Egede, Dr. Anbesaw Selassie, Dr. James Hardin, and June Watson. Thank you for everything.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	3
LIST OF TABLES.....	6
LIST OF FIGURES.....	7
ABSTRACT.....	9
CHAPTER 1. General Introduction.....	11
1.1. Overdispersion.....	12
1.2. Time varying categorical covariates with missing data.....	14
CHAPTER 2. Approaches for dealing with various sources of overdispersion in modeling count data: scale adjustment versus modeling.....	17
2.1. Introduction.....	18
2.2. Statistical models and estimation.....	20
2.3. Simulation study.....	23
2.4. Case studies.....	34
2.5. Discussion.....	37
CHAPTER 3. An empirical approach to determine a threshold for declaring the presence of overdispersion in count data.....	40
3.1. Introduction.....	41
3.2. Statistical models and estimation.....	43
3.3. Simulation.....	47
3.4. Results.....	50
3.5. Motivating real datasets.....	57
3.6. Conclusion.....	59
CHAPTER 4. Latent transition multiple imputation for missing data in time varying categorical covariates.....	63
4.1. Introduction.....	64
4.2. Methods.....	66
4.3. Latent transition multiple imputation.....	71
4.4. Simulation study.....	76

4.5. Data example.....	82
4.6. Discussion.....	88
CHAPTER 5. Dealing with overdispersion in longitudinal models including time varying categorical predictors with missing data.....	90
5.1. Introduction.....	91
5.2. Statistical models and estimation.....	93
5.3. Simulation.....	97
5.4. Data example.....	105
5.5. Discussion.....	112
CHAPTER 6. Concluding Remarks.....	113
6.1. Summary and discussion of all results.....	114
6.2. Future work.....	116
APPENDIX 1. CHAPTER 2 METHODOLOGY.....	118
APPENDIX 2. ADDITIONAL FIGURES AND TABLES FROM CHAPTER 2.....	121
APPENDIX 3. DERIVATION OF LIKELIHOOD CONTRIBUTION BASED ON LATENT STATUS CORRESPONDING TO CHAPTER 4.....	129
APPENDIX 4. ADDITIONAL FIGURES AND TABLES FROM CHAPTER 4.....	131
APPENDIX 5. ADDITIONAL TABLES FROM CHAPTER 5.....	139
REFERENCES.....	141

LIST OF TABLES

Table 2.1. Summary of models with estimated level of overdispersion.....	21
Table 2.2. Percentage of simulations in which true beta was contained in 95% CI by methods.....	33
Table 2.3. Comparison of methods for dealing with overdispersion in the NLST and <i>Salmonella</i> datasets.....	35
Table 2.4. NB-GLMM model comparing comorbidity count with patient demographics in the NLST dataset.....	36
Table 2.5. Summary of methods chosen to deal with overdispersion by cause.....	38
Table 3.1. Percent of simulations at varying levels of overdispersion in which the score test did in fact reject the null hypothesis and affirm the presence of overdispersion in the dataset.....	48
Table 3.2. Percentage of simulations with X_1 Type 1 errors and Type 2 errors and in which parameter coverage included the true parameter given true values of 0.01, 0.041, and 0.92 for the cross-sectional scenario using the unadjusted Poisson model and Poisson GLMM.....	50
Table 3.3. Percentage of simulations with X_1 Type 1 errors and Type 2 errors and in which parameter coverage included the true parameter given true values of 0.01, 0.041, and 0.92 for the cross-sectional scenario using the negative binomial regression model and negative binomial GLMM.....	51
Table 3.4. Percentage of simulations with X_1 Type 1 errors and Type 2 errors and in which parameter coverage included the true parameter given true values of 0.01, 0.041, and 0.92 for the longitudinal scenario using Poisson and negative binomial GLMM.....	55
Table 3.5. Standard error and rate ratio by overdispersion magnitude for NLST and <i>Salmonella</i> datasets.....	58
Table 4.1. Results of LTMI-LTA imputation for 20% and 50% missingness scenarios.....	81
Table 4.2. Missing MNA and A1C covariates by demographics.....	84
Table 4.3. Relationship between Elixhauser score and covariates in Diabetes dataset via LTMI-LTA.....	87
Table 5.1. Results for high outlier scenario via CCA.....	100
Table 5.2. Results for high outlier scenario via LTMI.....	101
Table 5.3. Missing MNA and A1C covariates by demographics.....	107
Table 5.4. Comparison of goodness of fit and dispersion statistics.....	108
Table 5.5. Relationship between Elixhauser score and covariates via CCA by distribution.....	110
Table 5.6. Relationship between Elixhauser score and covariates via LTMI by distribution.....	111

LIST OF FIGURES

Figure 2.1. Mean AIC and BIC values for simulated dataset with one important predictor omitted.....25

Figure 2.2. Mean parameter SE values for simulated dataset with one important predictor omitted.....26

Figure 2.3a. Mean AIC and BIC values for simulated dataset with outliers added (+50).....28

Figure 2.3b. Mean AIC and BIC values for simulated dataset with zero outliers added (20%).....28

Figure 2.4a. Mean parameter SE values for simulated dataset with outliers added (+50).....29

Figure 2.4b. Mean parameter SE values for simulated dataset with zero outliers added (20%)...29

Figure 2.5. Mean AIC and BIC values for simulated dataset with random effect $\gamma \sim N(0, group/10)$31

Figure 2.6. Mean parameter SE values for simulated dataset with random effect $\gamma \sim N(0, group/10)$32

Figure 3.1a. Percentage of simulations with Type 1 errors and in which parameter coverage included the true parameter given a true parameter value of 0.01 in the cross-sectional scenario, for outlier-dependent overdispersion.....52

Figure 3.1b. Percentage of simulations with Type 1 errors and in which parameter coverage included the true parameter given a true parameter value of 0.01 in the cross-sectional scenario, for overdispersion caused by zero inflation.....52

Figure 3.2a. Percentage of simulations with Type 2 errors and in which parameter coverage included the true parameter values of 0.41 and 0.92 in the cross-sectional scenario, for outlier-dependent overdispersion.....53

Figure 3.2b. Percentage of simulations with Type 2 errors and in which parameter coverage included the true parameter values of 0.41 and 0.92 in the cross-sectional scenario, for overdispersion caused by zero inflation.....53

Figure 3.3a. Percentage of simulations with Type 1 errors and in which parameter coverage included the true parameter given a true parameter value of 0.01 in the longitudinal scenario, for outlier-dependent overdispersion.....56

Figure 3.3b. Percentage of simulations with Type 1 errors and in which parameter coverage included the true parameter given a true parameter value of 0.01 in the longitudinal scenario, for overdispersion caused by zero inflation.....56

Figure 3.4a. Percentage of simulations with Type 2 errors and in which parameter coverage included the true parameter given true parameter values of 0.41 and 0.92 in the longitudinal scenario, for outlier-dependent overdispersion.....57

Figure 3.4b. Percentage of simulations with Type 2 errors and in which parameter coverage included the true parameter given true parameter values of 0.41 and 0.92 in the longitudinal scenario, for overdispersion caused by zero inflation.....57

Figure 4.1. Comparison of asymptotic standard errors for all simulated models with count outcome Y	78
Figure 4.2. Comparison of estimated standard errors for all simulated models with count outcome Y	79
Figure 4.3. Comparison of conditional AIC for all imputation methods with count outcome Y	80
Figure 4.4. Mean Elixhauser score by MNA and A1C status over five year time period	82
Figure 4.5. Percentage of missing MNA and A1C statuses over five year time period	83
Figure 4.6. Comparison of parameter standard errors for A1C and MNA predictors with missing observations, by method of dealing with missing data (CCA, LCMI, or LTMI methods)	85
Figure 4.7. Comparison of parameter standard errors for all predictors in the model, by method of dealing with missing data (CCA, LCMI, or LTMI methods)	85
Figure 4.8. Comparison of conditional AIC goodness of fit values	86
Figure 5.1. ASE for CCA and LTMI results	102
Figure 5.2. ESE for CCA and LTMI results	103
Figure 5.3. Bias for CCA and LTMI results	104
Figure 5.4. Conditional AIC for CCA and LTMI results	105
Figure 5.5. Mean Elixhauser score by MNA and A1C status over time	106
Figure 5.6. Percentage of missing MNA and A1C values over time	107
Figure 5.7. Standard errors for covariates in MNA and A1C model scenarios by GLMM distribution and method of addressing missing data	109

ELIZABETH HOLLY PAYNE. Statistical Methods for Modeling Count Data with Overdispersion and Missing Time Varying Categorical Covariates. (Under the direction of MULUGETA GEBREGZIABHER).

ABSTRACT

In studying the association between count outcomes and covariates using Poisson regression, the necessary requirement that the mean and variance of responses are equivalent for each covariate pattern is not always met in real datasets. This violation of equidispersion can lead to invalid inference unless proper alternative models are considered. There is currently no comprehensive and definitive assessment of the different methods of dealing with overdispersion, nor is there a standard approach for determining the threshold of overdispersion such that statistical intervention is necessary. The issue of overdispersion can be further complicated by the presence of missing covariate data in count outcome models. In this dissertation we have (1) compared the performance of different statistical models for dealing with overdispersion, (2) determined an appropriate threshold of the ratio of the Pearson chi-squared goodness of fit statistic to degrees of freedom σ_p such that statistical intervention is necessary to address the overdispersion, (3) developed a latent transition multiple imputation (LTMI) approach for dealing with missing time varying categorical covariates in count outcome models, and (4) compared the performance of LTMI with complete case analysis (CCA) and latent class multiple imputation (LCMI) in addressing missing time varying categorical covariates in the presence of overdispersion. Latent class assignment was determined via both SAS software and random effect modeling, and missing observation imputation was performed using predictive mean matching multiple imputation methods. We utilized extensive simulation studies to assess the performance of the proposed methods on a variety of overdispersion and missingness scenarios. We further demonstrated the application of the proposed models and methods via real data examples.

We conclude that the negative binomial generalized linear mixed model (NB-GLMM) is superior overall for modeling count data characterized by overdispersion. Furthermore, a general threshold for relying on the simple Poisson model for cross-sectional and longitudinal datasets is in cases where $\sigma_p \leq 1.2$. LTMI methods outperform CCA and LCMI in many scenarios, particularly when there is a higher percentage of missingness and data are MAR. Lastly, NB-GLMM is preferable to address overdispersion while LTMI is preferable for imputing covariate observations when jointly considering both issues.

General Introduction

CHAPTER 1

1. Overdispersion

Poisson regression is commonly used to study the association between count outcomes and covariates. However, a necessary requirement of Poisson regression is the underlying assumption that the response mean and variance are equivalent for each covariate pattern. This assumption often does not hold true in models with count outcomes based on real data. It is common that data are more variable than is accounted for under a reference (e.g., Poisson) model. This is called overdispersion (Cox 1983). Overdispersion arises only if the variability a model can capture is limited (for example, because of a functional relationship between mean and variance). This may be the result of population heterogeneity, correlated data, omission of important covariates in the model, or other reasons (Hardin and Hilbe 2007, Rigby, Stasinopoulos et al. 2008). For example, omitted important covariates increase the residual variance estimate because variability that should have been modeled through changes in the mean is now “*picked up*” as error variability if the model includes a dispersion parameter (the Poisson model has no such additional parameter). Another possible source of overdispersion is the presence of excess zeroes (or another value) in the count outcome. Two part (hurdle) and zero-inflated regression models have been developed to work with such data, including zero-inflated Poisson and zero-inflated negative binomial models (Lambert 1992, Long 1997, Tin 2008).

A model for which data are overdispersed can result in misleading inferences and conclusions, as overdispersion can lead to the underestimation of parameter standard errors and falsely increase the significance of beta parameters (McCullagh and Nelder 1983, Breslow 1990, Hilbe 2007, Faddy and Smith 2011). An earlier overview of the issue of overdispersion in both binary and count data can be found in (Hinde and Demétrio 1998) and recently a review of Poisson regression and overdispersion was published by Hayat and Higgins (Hayat and Higgins 2014).

Diagnosing and remedying overdispersion is a complicated process. As a result, numerous methods have been developed in an effort to deal statistically with the issue when modeling count responses. The most effective method will likely vary by situation depending on

the source of the overdispersion. For example, the omission of necessary random effects in a model or their inclusion as fixed effects may increase the residual error in the model and can lead to overdispersion. On the other hand, including random effects in the model can be useful if overdispersion is the result of correlation in the count outcomes (Smith and Heitjan 1993, Booth, Casella et al. 2003, Molenberghs, Verbeke et al. 2007, Yang, Hardin et al. 2007). This approach has been shown to be useful for dealing with overdispersion in complicated settings, such as longitudinal models (Milanzi, Alonso et al. 2012). A straightforward post hoc method of addressing overdispersion is to scale the covariance by various dispersion parameters (McCullagh and Nelder 1983). Two commonly used scales are the deviance statistic and the Pearson chi-squared statistic (Pearson 1900, Hardin and Hilbe 2007). Numerous other models have also been discussed for dealing with overdispersion, including the hurdle (Mullahy 1986) and bivariate (Cameron and Trivedi 1998) Poisson models. Hierarchical Bayesian methods have also been examined for dealing with overdispersion with random prior parameters added to the model to account for additional variability (Dauxois, Druilhet et al. 2006, Aregay, Shkedy et al. 2013).

The negative binomial distribution is a common alternative to the Poisson distribution for modeling data that exhibit overdispersion relative to the Poisson (Cameron 2006, Joe and Zhu 2005, Hilbe 2011). The negative binomial distribution accounts for further variance in count outcomes than the Poisson distribution through an additional gamma-distributed shape parameter to the Poisson rate parameter (Booth, Casella et al. 2003). Negative binomial regression has been shown to be effective in accounting for overdispersion in Poisson outcome models caused by missing covariates (Rigby, Stasinopoulos et al. 2008), outliers (Hilbe 2007) and other population heterogeneity factors, and is commonly used instead of Poisson in these situations (Ramakrishnan and Meeter 1993, Bouche, Lepage et al. 2009, Yau, Wang et al. 2003).

The most appropriate method may vary by situation. To handle it appropriately, the source of overdispersion must be identified. Despite numerous efforts to present a definitive answer to how best to adjust or account for overdispersion in count regression models (Hardin

and Hilbe 2001, Hilbe 2007, Xia, Morrison-Beedy et al. 2012, Hayat and Higgins 2014), as has been recently discussed in R-user group forums, there is no comprehensive and more definitive assessment of the different methods of dealing with overdispersion.

2. Time Varying Categorical Covariates with Missing Data

Missing data in time varying categorical variables are frequently encountered in longitudinal biomedical studies. While there has been progress with missing data methods that deal with longitudinally measured continuous variables, there is still paucity of methods that deal with time varying categorical variables that have missing values. Recently, multiple imputation based on latent class (LCMI) has been proposed to deal with the problem of missing data in time invariant categorical variables (Vermunt et al. 2008, Gebregziabher and DeSantis 2010). However, no extension has been made to address the problem of missing data in time varying categorical variables.

There are four paradigms of missing data analysis: multiple imputation (MI), maximum likelihood (ML), Bayesian methods (BM) and weighted estimating equations (WEE) (Ibrahim and Molenberghs 2009). Conditional repeated measures data have been modeled, for example, using the ML estimates of marginal response probabilities in log-linear models (Lindsey 2000); generalized linear mixed models conditional on random effects (Follman and Wu 1995); fixed-effect subject-specific logistic regression models (Rathouz 2009); joint models including time-to-event data using two-stage semiparametric regression (Ye et al. 2008) or Bayesian methods (Guo and Carlin 2004). For the purposes of these papers, we will focus on MI. This method is widely used for dealing with missing data problems in a wide variety of multivariate and longitudinal biomedical applications (Schafer 1997a, Schafer 1997b, Ibrahim and Molenberghs 2009, Engels and Diehr 2003, Nevalainen et al. 2009, Harel and Zhou 2007, Ferro 2014) and has recently been extended to random forest imputation via machine learning methods (Shah et al. 2014). There are several reasons for its wide usage. First, it is routinely available in most commercial statistical software packages such as SAS. Second, once multiple imputed datasets are obtained, statistical analysis may proceed as if all data were observed with an additional

benefit of obtaining parameter estimates that appropriately account for possible uncertainty in the imputed values (Little and Rubin 2002).

Recent work demonstrated that multiple imputation based on latent class can be used to impute missing categorical covariates (Vermunt et al. 2008, Gebregziabher and DeSantis 2010). Such a latent class based method is relevant because the problem of missing categorical data is ubiquitous in biomedical research. Via an extensive simulation study, Gebregziabher and DeSantis (2010) showed that a latent class-based imputation approach provided unbiased parameter estimates in a highly stratified data model with ignorable and some non-ignorable missing data in time invariant categorical variables. Specifically, they showed that in a generalized linear model framework with missing categorical variables, unbiased and efficient parameter estimates can be recovered utilizing latent class based multiple imputation. However, there are no readily available principled methods to deal with missing data in time varying categorical variables. This paper will seek to extend LCMI to latent transition multiple imputation (LTMI) to impute missing categories of time varying covariates by their latent status. We will consider two ways of identifying latent status for LCMI: via latent transition analysis (LTA) and random effects modeling.

In latent transition analysis (LTA), a hidden Markov model is assumed where at each time point, an unobserved time varying latent variable is inferred from a group of longitudinally observed items (time varying items). Parameter estimation for latent transition methods has been successfully utilized and explored (Chung et al. 2008), as well as applied to longitudinal random effect models involving missing data (Albert and Follmann 2007, Xiaowei et al. 2007, Lee et al. 2014). In LTA, the measurement model at each time point is a latent class model (Lazarsfeld and Henry 1968). All associations among categorical variables are explained by the underlying categorical latent variable. The result of fitting such a model is that for each individual, a latent trajectory that characterizes the missingness process is obtained. Conditional on the latent trajectory (latent status), observations and items are independent; this is known as the conditional independence assumption. At each time point, incomplete categorical data can be imputed

conditional on this latent status. In this paper, we will use LTA to estimate latent status imputation model (LTMI-LTA) and latent class imputation model (LCMI-LCA) from completely observed covariates to implement multiple imputation of missing data in time varying categorical variables.

Complete case analysis (CCA) is a widely used ad-hoc method for dealing with missing covariate data, in which all observations with incomplete data are removed from the dataset prior to analysis. This method may involve a high loss of information. Multiple imputation methods are generally considered superior to CCA, as MI is highly efficient and often demonstrate decreased bias compared to CCA depending on the magnitude and cause of missingness (Van der Heijden et al. 2006, Demissie et al. 2003, White and Carlin 2010). Complete case analysis may be acceptable in situations where missingness is completely at random (Knol et al. 2010) or independent of the outcome given covariates (White and Carlin 2010). Our simulation study and motivating data example include complete case analysis results as a general baseline for making comparison.

In the random effects approach to LCMI, we will fit a generalized linear mixed model to the time varying categorical covariate and the predicted random effects will be classified into groups (quintiles, for example) to identify latent classes. Then, the predicted latent classes will be imputed to the missing values of the time varying covariate. The random effects model may be assumed to come from a homogenous, or one, normal distribution, or it can come from a finite mixture of normal distributions (Verbeke and Lesaffre 1996), leading to the use of the heterogeneity linear mixed model (Komarek et al. 2002). Our simulation study and motivating data example will include LTMI latent transition analysis results (LTMI-LTA) and LTMI based on heterogeneity linear mixed model (LTMI-LMM) results. These will be compared with CCA, LCMI-LCA, and LCMI based on homogeneity linear mixed model (LCMI-LMM) results.

We will study the statistical properties of LTMI and make comparison with complete case analysis and LCMI methods via simulation study and a real motivating dataset. We will then perform a similar analysis and comparison taking the additional issue of overdispersion into consideration.

**Approaches for dealing with various sources of overdispersion in modeling count data:
scale adjustment versus modeling**

CHAPTER 2

1. Introduction

Poisson regression is commonly used to study the association between count outcomes and covariates. However, a restriction of Poisson regression is that the response mean must be equal to the variance. This equidispersion often does not hold true in real data. Often, data are more variable than is accounted for under the Poisson model. This is called overdispersion (Cox 1983). The overdispersion may occur due to population heterogeneity, correlated data, omission of important covariates in the model, outliers or other reasons (Hardin and Hilbe 2007, Rigby, Stasinopoulos et al. 2008). For example, if an important covariate is not measured, the residual variance estimate is increased because variability that should have been modeled through changes in the mean is now "*picked up*" as error variability if the model includes a dispersion parameter. The Poisson model has no such additional parameter. Another possible source of overdispersion is the presence of outliers: for example, excess zeroes (or another value) in the count outcome.

An overdispersed model which assumes equidispersion can result in misleading inferences and conclusions, as overdispersion can lead to the underestimation of parameter standard errors and falsely increase the significance of beta parameters (McCullagh and Nelder 1983, Breslow 1990, Hilbe 2007, Faddy and Smith 2011). An earlier overview of the issue of overdispersion in both binary and count data was published by Hinde and Demetrio (1998.) More recently, a review of Poisson regression and overdispersion was published by Hayat and Higgins (2014).

Diagnosing and correcting overdispersion is a complicated process which is imperative to interpreting count data correctly. As a result, numerous methods have been developed in an effort to deal statistically with the issue when modeling count responses. The most effective method will likely vary based on the source of the overdispersion. For example, the omission of necessary random effects in a model or their inclusion as fixed effects may increase the residual error in the model and can lead to overdispersion. Including random effects in the model can therefore be useful if overdispersion is present as the result of correlation in the count outcomes

(Smith and Heitjan 1993, Booth, Casella et al. 2003, Molenberghs, Verbeke et al. 2007, Yang, Hardin et al. 2007). This approach is particularly useful when dealing with overdispersion in more complicated settings, such as longitudinal models (Milanzi, Alonso et al. 2012). A straightforward post hoc method of addressing overdispersion is to scale the covariance by various dispersion parameters (McCullagh and Nelder 1983). Two commonly used scales are the deviance statistic and the Pearson chi-squared statistic (Pearson 1900, Hardin and Hilbe 2007). Numerous other models have also been discussed for dealing with overdispersion, including the hurdle (Mullahy 1986) and bivariate (Cameron and Trivedi 1998) Poisson models. Bayesian methods have also been examined for dealing with overdispersion with random prior parameters added to the model to account for additional variability (Dauxois, Druilhet et al. 2006, Aregay, Shkedy et al. 2013). Two part (hurdle and zero-inflated) regression models including zero-inflated Poisson models (Lambert 1992, Long 1997, Tin 2008) have been further developed to work with overdispersion caused by excess zeros.

The negative-binomial (NB) distribution is a common alternative to the Poisson distribution for modeling data that exhibit overdispersion relative to the Poisson (Cameron 2006, Joe and Zhu 2005, Hilbe 2011). The NB distribution accounts for further variance in count outcomes than the Poisson distribution through an additional gamma-distributed shape parameter to the Poisson scale parameter (Booth, Casella et al. 2003). NB regression has been shown to be effective in accounting for overdispersion in count data models caused by omitted covariates (Rigby, Stasinopoulos et al. 2008), outliers (Hilbe 2007), and other population heterogeneity factors, and is commonly used instead of Poisson in these situations (Ramakrishnan and Meeter 1993, Bouche, Lepage et al. 2009, Yau, Wang et al. 2003).

Despite numerous efforts to present a definitive answer regarding how best to adjust or account for overdispersion in count regression models (Hardin and Hilbe 2001, Hilbe 2007, Xia, Morrison-Beedy et al. 2012, Hayat and Higgins 2014), as has been recently discussed in R-user group forums, there is no comprehensive approach or more definitive assessment of the different methods for dealing with overdispersion. Moreover, the most appropriate method for dealing with

overdispersion may vary by source. Thus, there is a need to examine the differential performance of existing approaches for dealing with overdispersion with respect to the source of overdispersion. Our investigation is therefore an attempt to fill the gap and provide a comprehensive evaluation of six different approaches using simulation studies that consider three key sources of overdispersion and two case studies.

This chapter is organized in the following manner. Subsequent to the introduction, the statistical models and maximum likelihood estimation are described in section 2. Section 3 provides information about the design and results of the simulation study. Section 4 details the motivating case studies and results, and section 5 provides a discussion of all results as well as future research plans in this area.

2. Statistical models and estimation

2.1 Models

Consider a random variable Y distributed Poisson with variance function $Var(Y) = \mu$. If non-equidispersion relative to the Poisson is present, a variance function accounting for changes in variability can be specified as a scale-adjustment of the Poisson variance function $Var(Y) = \varphi\mu$ with dispersion parameter φ . In this case, if $\varphi = 1$ then there is equidispersion; if $\varphi < 1$ there is underdispersion; and if $\varphi > 1$ there is overdispersion.

Another approach to modeling overdispersion relative to the Poisson is to consider a two-stage model for which $Y|\theta \sim Pois(\theta)$ and θ is a random variable such that $E(\theta) = \mu$ and $Var(\theta) = \sigma^2$. It then follows that $E(Y) = \mu$ and $Var(Y) = \mu + \sigma^2$, allowing for variability that is greater than the mean. When the distribution of θ is assumed to be gamma then Y has a negative-binomial distribution with $E(Y) = k/\lambda = \mu$ and $Var(Y) = \mu + \mu^2/k$.

Another approach is to include random effects in a generalized linear mixed model (GLMM) to deal with overdispersion. For vectors of fixed effect (X_i) and random effect (Z_i) explanatory variables ($i = 1, \dots, n$) the GLMM family is given by,

$$E(Y_i|X_i, Z_i) = g^{-1}(X_i\beta + Z_i\gamma_i) = \mu_i$$

Here, g represents a monotone link function, β is a vector of p fixed coefficients, and γ_i is a vector of unobserved random deviations (assumed to have zero mean) for which the variance will be estimated. When the distribution of Y is assumed to be Poisson and the link function is log, then the GLMM is referred to as the Poisson-GLMM. The variance function for this model with normally distributed random effect is given by $Var(Y_i) = \mu_i + k\mu_i^2$, which is the same as the variance function for NB. Similarly, when the distribution of Y is NB and the link function is log, the GLMM is referred to as the NB-GLMM (McCullagh and Nelder 1989). Because it includes an additional dispersion parameter, the NB-GLMM allows for additional residual overdispersion beyond what is captured for by Poisson-GLMM.

Table 1. Summary of models with estimated level of overdispersion.

Distribution	Abbreviation	Method	Adjustment
Poisson	Poisson	Not adjusted	N/A
Poisson	DS-Poisson	Scale-adjusted	Deviance statistic
Poisson	PS-Poisson	Scale-adjusted	Pearson χ^2 statistic
Poisson	Poisson-GLMM	GLMM	Random effects
Negative-binomial	NB	Unadjusted	Additional parameter
Negative-binomial	NB-GLMM	GLMM	Additional parameter, random effects
Estimated Overdispersion			
<i>Covariate</i>	Source	Deviance/df \pm sd	Pearson χ^2 /df \pm sd
Normal	1 covariate omitted	3.65 \pm 0.66	4.55 \pm 1.22
	2 covariates omitted	15.90 \pm 4.10	52.80 \pm 51.94
Binary	1 covariate omitted	38.31 \pm 1.11	33.93 \pm 1.36
	2 covariates omitted	63.56 \pm 1.95	70.87 \pm 1.41
Uniform	1 covariate omitted	8.14 \pm 0.90	8.48 \pm 0.95
	2 covariates omitted	44.98 \pm 4.52	80.50 \pm 9.71
Normal	Small outliers	2.50 \pm 0.36	25.42 \pm 17.06
	Large outliers	9.42 \pm 1.06	121.32 \pm 59.30
	Lower % zero outliers	2.61 \pm 0.47	1.53 \pm 0.31
	Higher % zero outliers	4.34 \pm 0.54	2.94 \pm 0.45
Binary	Small outliers	2.07 \pm 0.10	7.97 \pm 0.77
	Large outliers	8.50 \pm 0.28	50.64 \pm 3.70
	Lower % zero outliers	1.86 \pm 0.06	1.24 \pm 0.06
	Higher % zero outliers	2.21 \pm 0.04	1.84 \pm 0.06
Uniform	Small outliers	2.16 \pm 0.11	8.88 \pm 0.92
	Large outliers	8.74 \pm 0.30	54.74 \pm 4.12
	Lower % zero outliers	1.69 \pm 0.05	1.13 \pm 0.05
	Higher % zero outliers	2.00 \pm 0.04	1.67 \pm 0.06
Normal	Small variance random effects	5.68 \pm 1.66	8.56 \pm 4.68
	Large variance random effects	18.28 \pm 15.15	81.58 \pm 217.85
Binary	Small variance random effects	2.52 \pm 0.34	3.94 \pm 1.26
	Large variance random effects	7.41 \pm 1.81	19.19 \pm 15.09
Uniform	Small variance random effects	2.28 \pm 0.30	3.54 \pm 1.03
	Large variance random effects	6.68 \pm 1.54	17.13 \pm 11.67

The six different methods considered in this study can be generally classified into two categories: scale adjustment and modeling methods. We considered two scale adjustment methods under the standard Poisson regression (abbreviated simply Poisson): (1) deviance scale-adjusted Poisson regression (DS-Poisson) and (2) Pearson scale-adjusted Poisson regression (PS-Poisson). We also considered three modeling methods, (3) negative-binomial regression (NB), and (4, 5) two GLMM with random intercept, log link, and compound symmetry covariance, with outcomes distributed as Poisson and negative-binomial (Poisson-GLMM, NB-GLMM, respectively). Table 1 gives a summary of the various models we considered, including a description of the particular method and adjustment utilized for addressing overdispersion.

2.2 Estimation and Inference

The parameters of the model that need to be estimated include dispersion, regression coefficients and variance components. Estimation and inference for the model based methods can be accomplished using maximum likelihood (McCullagh and Nelder 1989), quasi-likelihood (Hardin and Hilbe 2001), or pseudo-likelihood (Fitzmaurice, Laird et al. 2004). In our case, the Poisson and NB were estimated via maximum likelihood and the NB-GLMM and Poisson-GLMM were estimated via pseudo likelihood. On the other hand, in the scale based methods, the two most commonly used estimators of dispersion in the literature are the ratio of the model deviance to its corresponding degrees of freedom and the ratio of the Pearson X^2 statistic to its corresponding degrees of freedom. The degrees of freedom are typically given by $n - p$ for a study with sample size n observations and p parameters. When the assumption of equal mean and variance is not violated, these ratios will be equal to 1. Relative to the model, if these ratios are greater than 1 then the data are considered overdispersed. Higher values demonstrate a greater magnitude of overdispersion.

2.3 Model Comparison

Akaike's information criteria (AIC) (Akaike 1974) and Bayesian information criteria (BIC) (Schwarz 1978) were utilized to measure goodness of fit and make comparisons among the different models. Parameter standard errors and the 95% confidence interval coverage for each

parameter were also recorded to determine the level of bias in the standard error estimates compared to the assumed value in the simulation study. These values were then compared across the models to determine which method for dealing with overdispersion resulted in the lowest AIC and BIC values as well as offered standard errors that are closer to the simulated value with nominal 95% confidence interval coverage. This model comparison using multiple criteria is similar in format to work recently published by Xia et al., in which the authors compare Poisson, negative-binomial, and zero-inflated regression methods to model overdispersed and zero-inflated data from an HIV risk reduction intervention study (2012). Gardner et al. also compared Poisson and negative-binomial methods of analyzing overdispersed count outcomes related to psychological datasets (1995), while Ver Hoef and Boveng provide an overview and comparison of these methods for ecologists (2007).

In this paper, we provide a more unified comparison among the many possible approaches to dealing with overdispersion. We also provide a detailed derivation of dispersion in the context of count data fitted using different models under multiple covariate type scenarios (see technical Appendix 1) to complement the simulation and case-studies. SAS 9.4 was utilized in all analyses for both simulated and real datasets, particularly *Proc GENMOD* and *Proc GLIMMIX* packages.

3. Simulation Study

We simulated 1 000 datasets each with a sample size of $n=1\ 000$ random observations generated following scenarios given in (Hilbe 2007) under three distributions of predictor scenarios. Scenarios under a sample size of 500 did not lead to different conclusions (results not shown). Table 1 provides a list of all scenarios with their corresponding measure of overdispersion. After generating overdispersed datasets for these scenarios, analysis was made using the six models for all simulated data from each scenario. Goodness of fit statistics including AIC, BIC, deviance, Pearson statistic and parameter estimates for the regression coefficients corresponding to each covariate with their corresponding variance and 95% confidence interval (CI) coverage were calculated.

3.1 *Covariate dependent overdispersion design*

We considered three different scenarios. In **scenario 1**, each dataset included three normal independent predictors with $x_1 \sim \text{Normal}(1, 2)$, $x_2 \sim \text{Normal}(2, 3)$, and $x_3 \sim \text{Normal}(3, 4)$. In the count regression model, $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ were assigned values (1.0, 0.5, -0.75, and 0.25). These variables were then utilized to create a count response Y using a Poisson error and log link function that ranged from 0 to 3443. The distributions of these variables are illustrated in Appendix 2 Figure 1a-d. In **scenario 2**, binary covariates were derived from the normally distributed covariates described above. Values less than the mean were assigned a value of 0; values greater than or equal to the mean were assigned a value of 1. The $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ for intercept, x_1 , x_2 , and x_3 were assigned to be (1.0, 2.0, 1.5, 1.0) respectively. In **scenario 3**, predictor x_1 was drawn from Uniform(5, 10), x_2 from Uniform(10, 15), and x_3 from Uniform(15, 20). In this case, $(\beta_0, \beta_1, \beta_2, \beta_3)$ for the intercept, x_1 , x_2 , and x_3 were again assigned to be (1.0, 0.5, -0.75, 0.25), respectively. Overdispersion relative to the Poisson was then created in these datasets via the omission of important predictors from the model where (i) predictor x_1 was first removed from the model and (ii) both x_1 and x_2 were removed from the model, creating overdispersion of a higher magnitude. Further details of the methodology are discussed in Appendix 1.

3.2 *Covariate dependent overdispersion results*

When one important predictor was omitted from the model, the mean deviance/df value for the unadjusted Poisson model was 3.65 ± 0.66 and the mean Pearson X^2/df value was 4.55 ± 1.22 , indicating the presence of overdispersion. When two important predictors were omitted, these values increased to 15.90 ± 4.10 and 52.80 ± 51.94 , respectively, indicating overdispersion of greater magnitude. For binary covariates, after the omission of one predictor the mean deviance/df value for the unadjusted Poisson model in binary covariate simulations was 38.31 ± 1.11 , and the mean Pearson X^2/df value was 33.93 ± 1.36 . After the omission of two predictors, the mean deviance/df value increased to 63.56 ± 1.95 , and the mean Pearson X^2/df value increased to 70.87 ± 1.41 . In the scenario where the covariates come from a uniform distribution,

after the omission of one predictor the mean deviance/df value for the unadjusted Poisson model in uniform covariate simulations was 8.14 ± 0.90 , and the mean Pearson X^2/df value was 8.48 ± 0.95 . After the omission of two predictors, the mean deviance/df value increased to 44.98 ± 4.52 , and the mean Pearson X^2/df value increased to 80.50 ± 9.71 (Table 1).

Figure 1 shows the mean AIC and BIC values when one important predictor is omitted, for the normal predictor scenario. The results indicate that the NB and NB-GLMM models have the lowest AIC and BIC that is comparable to the original model without overdispersion. All Poisson regression models exhibited very large values of AIC and BIC, indicating poorer fit to the data compared to the NB models.

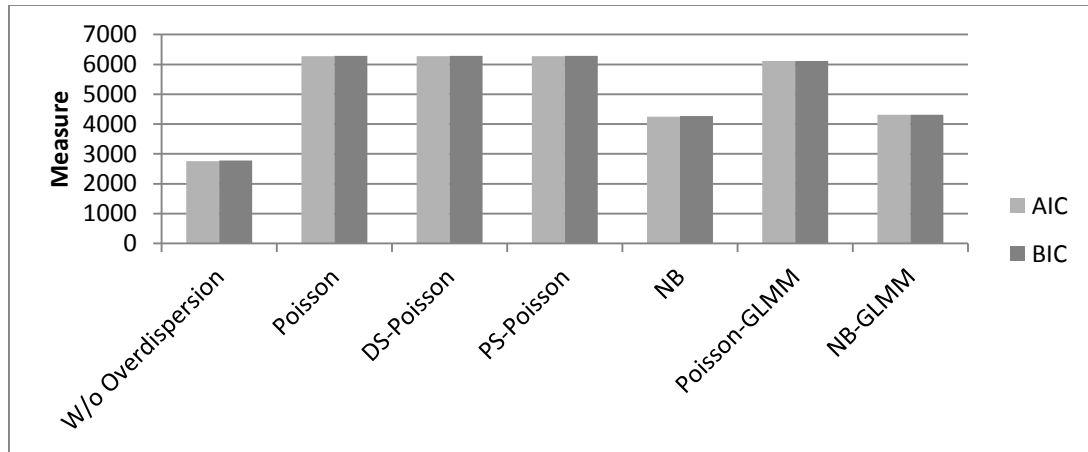


Figure 1. Mean AIC and BIC values for simulated dataset with one important predictor omitted.

Figure 2 shows the mean parameter SE estimates for this simulation. The DS-Poisson and PS-Poisson had much larger SE than the model without overdispersion for the intercept (results are not shown) but the SE estimates for the x_2 and x_3 were generally closer. The consequence of not capturing the overdispersion is a more conservative inference with potential for type II error. On the other hand, the SE estimates for the regression coefficients of x_2 and x_3 in the scale-adjusted models appeared to have moderately increased the SE estimates, especially compared to Poisson and Poisson-GLMM. The NB also appeared to have moderately increased the SE estimates for the coefficients of x_2 and x_3 , thereby accounting for the overdispersion introduced into the data. NB-GLMM gave much higher values here. Not

surprisingly, the 95% CI appeared to follow a similar trend. These results generally hold true irrespective of the type of omitted covariate (binary, uniform or normal). Results for covariate dependent overdispersion resulting from the omission of two important covariates are given in Appendix 2, and are qualitatively similar.

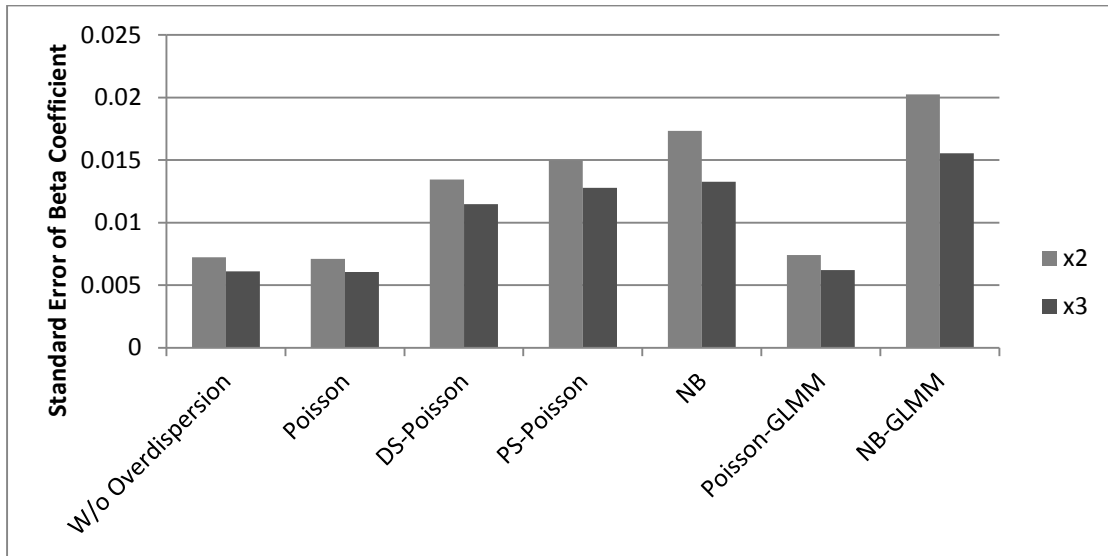


Figure 2. Mean parameter SE values for simulated dataset with one important predictor omitted.

3.3 Outlier dependent overdispersion design

The second scenario for creating overdispersion relative to the Poisson was the addition of either high outliers or excess zero outliers to the count outcome Y . In the first scenario, variable x_1 was left in the model and a random Y value in each group of each simulation was increased by 50 to create outlier dependent overdispersion in the data. This gave 10 total outliers in each dataset containing 1,000 values; i.e. 1% of the data were replaced by outliers. This was followed by an increase in the outliers by 150, which created overdispersion of a higher magnitude. In the second scenario, varying percentages of the outcome were replaced with 0. For binary covariates, the $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ for intercept, x_1 , x_2 , and x_3 were assigned to be (1, 0, 0.5, -0.75, 0.25), respectively. Overdispersion was then created in the datasets as detailed above via the addition of outliers of differing magnitudes or zero values.

3.4 Outlier dependent overdispersion results

The simulated data were analyzed with all variables in the model and overdispersion created via the addition of outliers. After the smaller outliers were added, the mean deviance/df value for the unadjusted Poisson model was 2.50 ± 0.36 and the mean Pearson X^2/df value was 25.42 ± 17.06 , demonstrating the presence of overdispersion. After the addition of larger outliers, these values increased to 9.42 ± 1.06 and 121.32 ± 59.30 , respectively. After the addition of 20% zero outliers, these values were 2.61 ± 0.47 and 1.53 ± 0.31 , respectively. After the addition of 40% zero outliers, these values increased to 4.34 ± 0.54 and 2.94 ± 0.45 , respectively. For binary covariates, after the addition of smaller outliers the mean deviance/df value for the unadjusted Poisson model in binary covariate simulations was 2.07 ± 0.10 , and the mean Pearson X^2/df value was 7.97 ± 0.77 . After the magnitude of the outliers was increased, the mean deviance/df value increased to 8.50 ± 0.28 , and the mean Pearson X^2/df value increased to 50.64 ± 3.70 . After the addition of 40% zero outliers, these values were 1.86 ± 0.06 and 1.24 ± 0.06 , respectively. After the addition of 60% zero outliers, these values increased to 2.21 ± 0.04 and 1.84 ± 0.06 , respectively. In the scenario where the covariates come from a uniform distribution, after the addition of the smaller outliers the mean deviance/df value for the unadjusted Poisson model in uniform covariate simulations was 2.16 ± 0.11 , and the mean Pearson X^2/df value was 8.88 ± 0.92 . After the magnitude of the outliers was increased, the mean deviance/df value increased to 8.74 ± 0.30 , and the mean Pearson X^2/df value increased to 54.74 ± 4.12 (Table 1). After the addition of 40% zero outliers, these values were 1.69 ± 0.05 and 1.13 ± 0.05 , respectively. After the addition of 60% zero outliers, these values increased to 2.00 ± 0.04 and 1.67 ± 0.06 , respectively.

Figures 3a and 3b respectively give the mean AIC and BIC values with smaller outliers (+50) and 20% zero outliers, for the normal predictor scenario. The NB followed by the NB-GLMM model had the lowest mean AIC and BIC values in models with all kinds of outliers showing good fit to the data while the Poisson model variations exhibited generally poor fit to the data.

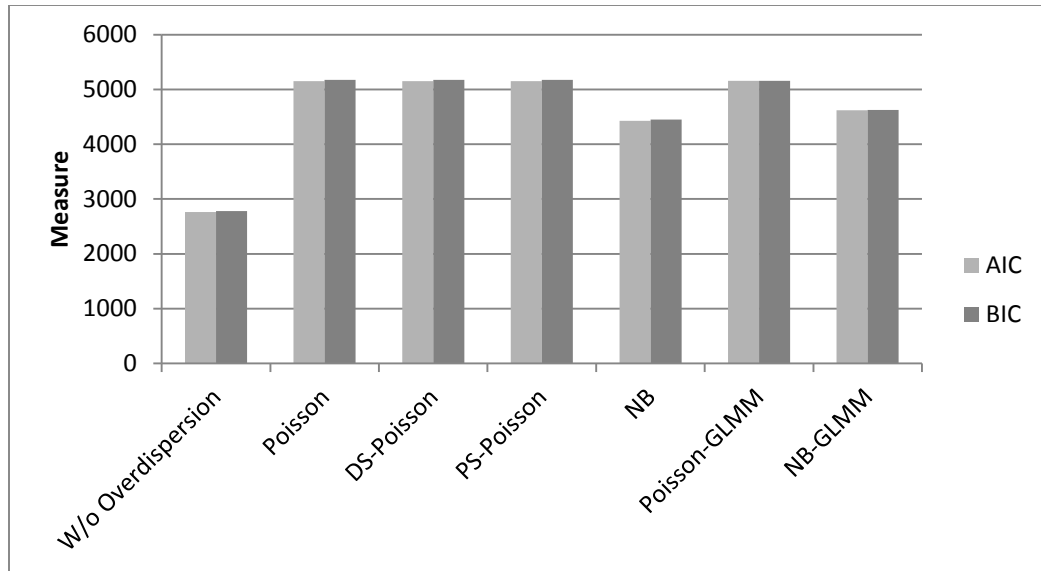


Figure 3a. Mean AIC and BIC values for simulated dataset with outliers added (+50).

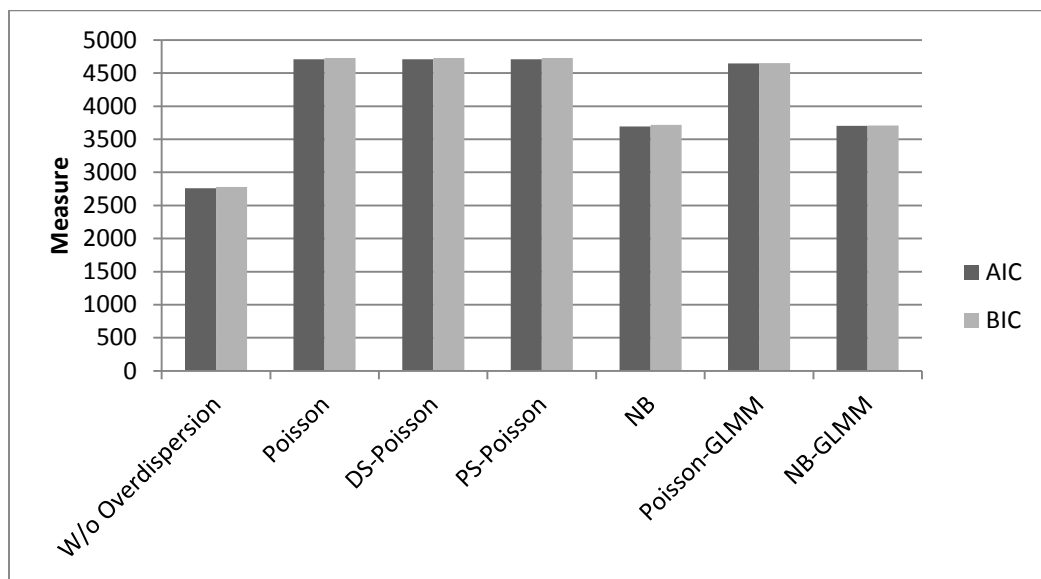


Figure 3b. Mean AIC and BIC values for simulated dataset with zero outliers added (20%).

Figures 4a and 4b show the mean SE estimates for these respective scenarios. The SE estimates for the full Poisson model without overdispersion are provided for comparison. When the covariates are from a normal distribution, the NB model appeared to produce moderately increased SE for the non-zero outliers while the NB-GLMM had somewhat highly increased SE estimates of the regression coefficients of the three covariates. The PS-Poisson model had

particularly high SE estimates for all parameters, while the Poisson and Poisson-GLMM models gave much lower estimates of the SE compared to what would be expected under the simulated dispersed data. The 95% CI appeared to follow the same trend. Therefore, it appears that the NB and DS-Poisson models may be considered superior for dealing with outlier dependent overdispersion in this case, with NB demonstrating better goodness of fit. The NB models gave higher SE for the zero outlier scenarios, while the scale-adjusted Poisson models gave moderately increased SE for both levels of overdispersion magnitude.

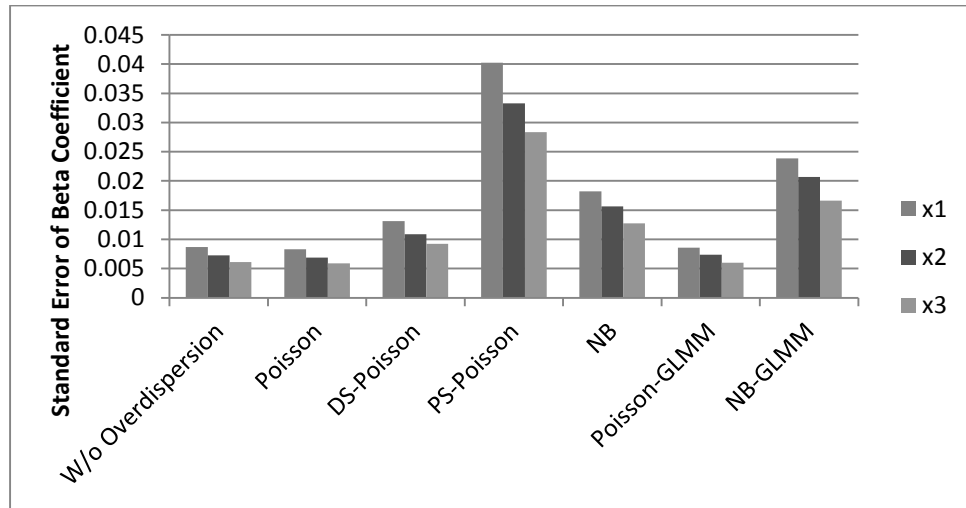


Figure 4a. Mean parameter SE values for simulated dataset with outliers added (+50).

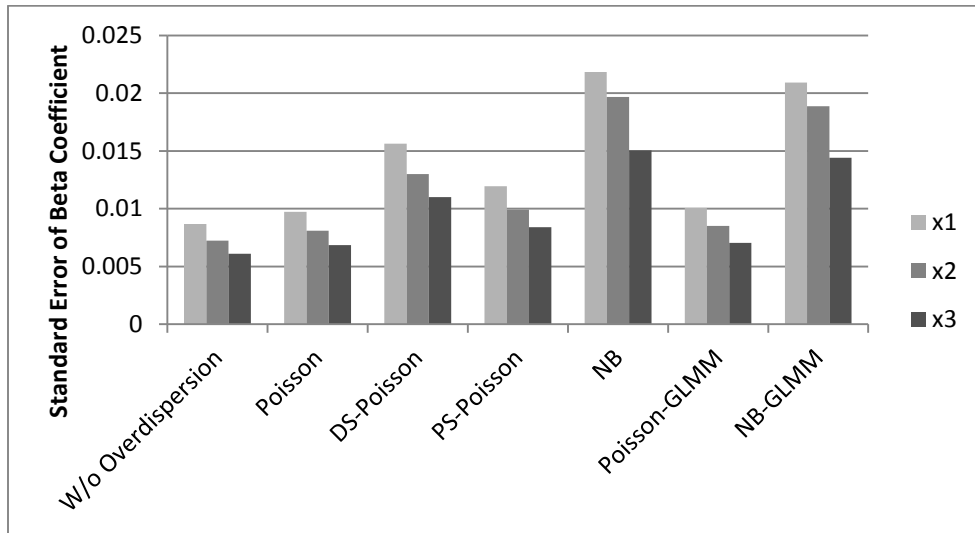


Figure 4b. Mean parameter SE values for simulated dataset with zero outliers added (20%).

The lowest AIC and BIC values for the binary covariate results were given by the NB method. The two GLMM increased the AIC and BIC values, while the Poisson, DS-Poisson, and PS-Poisson gave identical results. NB and DS-Poisson gave moderate SE and 95% CI coverage. The NB-GLMM and PS-Poisson gave higher SE and 95% CI for all covariates, while the original Poisson and Poisson-GLMM gave lower values. Results are similar for the larger level outlier scenarios, given in Appendix 2.

3.5 Random effect dependent overdispersion design

The third scenario for creating overdispersion relative to the Poisson was the addition of a random intercept to the dataset which is then omitted from the model. The data were divided at random into ten groups, such that $g = 1, \dots, 10$. The intercept value was assigned to be 1.0. A random effect γ dependent on each group was added from distribution $N(0, g/10)$ to create a lower magnitude of overdispersion, and from $N(0, g/5)$ to create a higher magnitude of overdispersion. The random effect γ was added to create extra heterogeneity or overdispersion of varying magnitudes. Higher variability of the random effect increases the overdispersion which occurs when it is omitted from the model. Similarly, binary and uniform covariates were created and their β parameters assigned as for the outlier dependent simulations described above.

3.6 Random effect dependent overdispersion results

When the covariates were all normally distributed for the random effect dependent overdispersion with lower magnitude, the mean deviance/df value for the unadjusted Poisson model was 5.68 ± 1.66 and the mean Pearson X^2/df value was 8.56 ± 4.68 . For the higher magnitude of overdispersion, these values increased to 18.28 ± 15.15 and 81.58 ± 217.85 , respectively. For binary predictors, after the addition of the random effect with lesser variability, the mean deviance/df value for the binary covariate simulations was 2.52 ± 0.34 , and the mean Pearson X^2/df value was 3.94 ± 1.26 . After the variability of the random effect was increased, the mean deviance/df value increased to 7.41 ± 1.81 , and the mean Pearson X^2/df value increased to 19.19 ± 15.09 . For uniform distributed predictors, after the addition of the less variable random effects, the mean deviance/df value for the uniform covariate simulations was 2.28 ± 0.30 , and

the mean Pearson X^2/df value was 3.54 ± 1.03 . After the magnitude of the outliers was increased, the mean deviance/df value for the uniform covariate simulations increased to 6.68 ± 1.54 , and the mean Pearson X^2/df value increased to 17.13 ± 11.67 (Table 1).

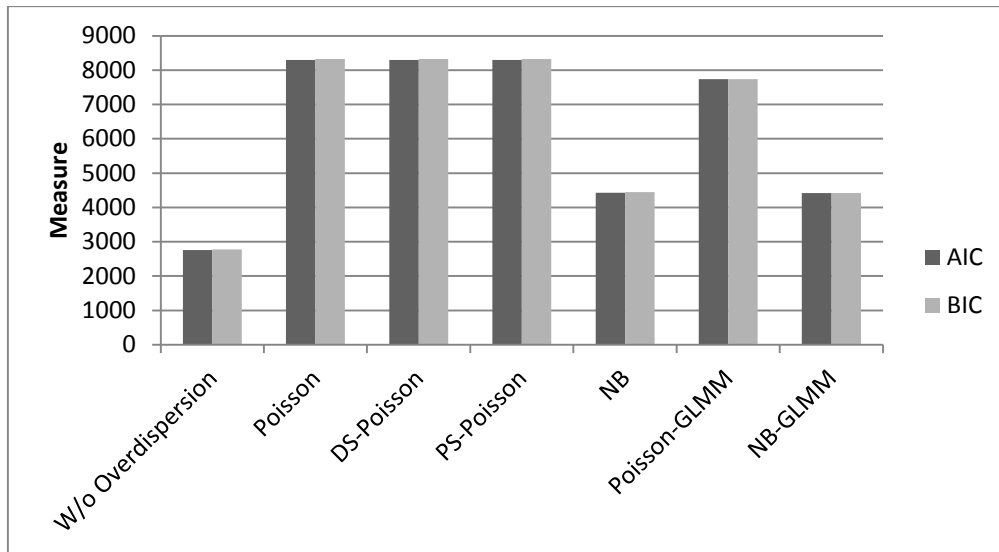


Figure 5. Mean AIC and BIC values for simulated dataset with random effect $\gamma \sim N(0, group/10)$.

Figure 5 shows the mean AIC and BIC values with random effects of smaller variance for the normal predictor scenario. These values before the addition of overdispersion are also included for comparison. The NB-GLMM model had the lowest mean AIC and BIC values, followed by the NB, showing good fit to the data, while the Poisson model variations exhibited poorer goodness of fit.

Figure 6 shows the mean SE estimates for this scenario. Again, the SE estimates for the full Poisson model without overdispersion are provided for comparison. In this scenario, the NB, NB-GLMM, and DS-Poisson models appeared to produce moderately increased SE for both kinds of random intercepts. The PS-Poisson model had particularly high SE estimates for all parameters, while the Poisson and Poisson-GLMM gave much lower estimates of the SE compared to what would be expected under the simulated dispersed data. The 95% CI appeared to follow the same trend. Therefore, it appears that the NB-GLMM may be considered superior in

dealing with overdispersion resulting from these scenarios. Results are similar for the larger variance random effect scenarios, given in Appendix 2.

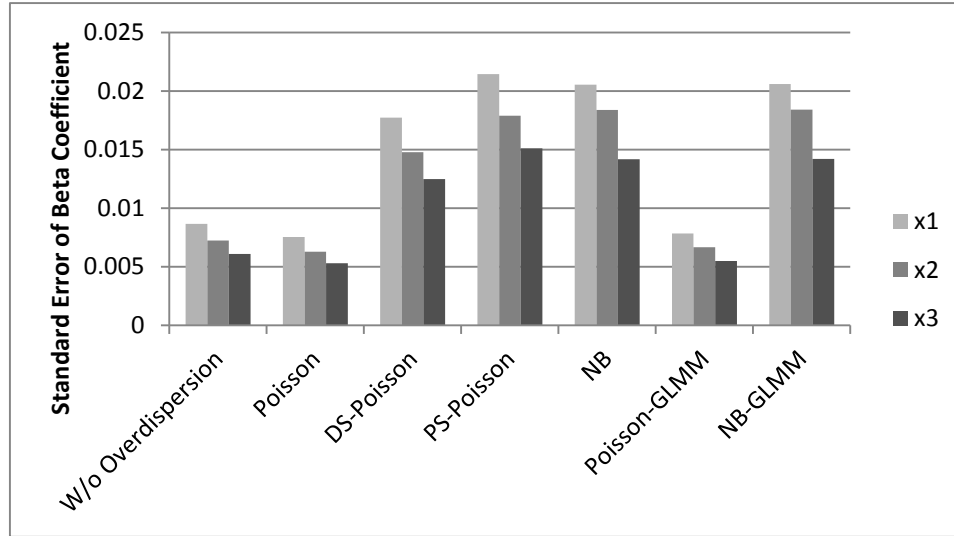


Figure 6. Mean parameter SE values for simulated dataset with random effect $\gamma \sim N(0, group/10)$.

3.7 Confidence interval coverage results

In order to examine the nominal 95% CI coverage of the different methods, we recorded the percentage of estimates from each simulation in which the true beta value was contained in the parameter 95% confidence interval by method, covariate type, and overdispersion type. These results are given in Table 2.

For the covariate dependent overdispersion simulations, the Pearson-scaled Poisson and negative-binomial methods generally gave the highest percentage of coverage closest to the nominal 95%. For the outlier and random effect dependent overdispersion, the Pearson-scaled Poisson and negative-binomial generalized linear mixed model gave the highest percentages of coverage. Other methods were generally unreliable.

Table 2. Percentage of simulations in which true beta was contained in 95% CI by methods.

Simulation	Covariate	Covariate Dependent Overdispersion					
		Poisson	DS-Poisson	PS-Poisson	NB	Poisson-GLMM	NB-GLMM
Normal – 1 cov omitted	X2, X3	26.8, 31.7	47.0, 57.0	51.6, 61.7	80.2, 87.8	28.4, 34.8	86.8, 92.7
Normal – 2 covs omitted	X3	19.8	70.3	90.1	71.9	19.5	89.7
Uniform – 1 cov omitted	X2, X3	28.9, 23.3	67.7, 62.7	68.5, 63.8	38.0, 82.5	29.3, 23.9	45.1, 87.7
Uniform – 2 covs omitted	X3	15.8	81.1	92.3	91.8	16.3	94.1
Binary – 1 cov omitted	X2, X3	29.4, 23.2	98.3, 94.5	97.7, 93.7	97.5, 97.4	29.3, 23.9	96.3, 95.2
Binary – 2 covs omitted	X3	22.0	96.6	97.5	95.3	22.3	95.5
Outlier Dependent Overdispersion							
Normal – outliers (+50)	X1, X2, X3	15.0, 0.6, 31.7	31.6, 1.4, 54.5	99.1, 85.2, 99.5	22.0, 3.4, 36.2	15.5, 1.0, 32.2	33.2, 10.9, 46.4
Normal – outliers (+150)	X1, X2, X3	3.2, 0.0, 9.1	14.0, 0.4, 31.3	94.8, 46.1, 97.3	13.9, 3.0, 20.8	3.1, 0.0, 8.9	53.6, 25.8, 62.1
Normal – low % 0 outliers	X1, X2, X3	39.3, 33.0, 44.9	62.3, 52.3, 62.6	48.5, 40.5, 51.2	96.8, 90.1, 98.5	45.0, 38.1, 49.0	95.8, 87.0, 98.4
Normal – high % 0 outliers	X1, X2, X3	29.7, 24.5, 29.2	55.7, 47.9, 59.1	47.2, 40.8, 48.2	98.4, 89.8, 98.4	33.5, 29.4, 33.6	94.3, 83.6, 97.0
Uniform – outliers (+50)	X1, X2, X3	49.8, 24.9, 45.6	65.2, 39.9, 62.2	94.0, 80.3, 93.9	59.1, 36.2, 55.7	49.9, 24.9, 45.6	93.9, 79.2, 93.5
Uniform – outliers (+150)	X1, X2, X3	20.9, 9.5, 17.9	54.0, 27.0, 49.1	92.6, 73.6, 91.9	38.9, 21.3, 35.7	20.9, 9.5, 17.9	92.2, 75.3, 91.1
Uniform – low % 0 outliers	X1, X2, X3	89.5, 92.7, 89.9	96.7, 98.0, 96.9	92.0, 93.8, 91.8	97.4, 98.2, 97.5	89.9, 92.7, 89.9	98.2, 98.3, 98.7
Uniform – high % 0 outliers	X1, X2, X3	84.0, 84.5, 83.9	94.4, 95.0, 94.9	92.4, 93.2, 92.7	98.8, 98.8, 98.6	84.8, 84.7, 83.2	98.7, 98.8, 98.3
Binary – outliers (+50)	X1, X2, X3	44.9, 15.7, 47.1	55.3, 25.4, 67.5	91.4, 62.0, 93.9	51.3, 25.6, 58.3	45.1, 15.7, 47.1	90.8, 64.3, 93.6
Binary – outliers (+150)	X1, X2, X3	18.8, 11.6, 21.5	47.2, 32.1, 59.3	86.6, 69.3, 94.9	30.9, 20.1, 39.4	18.9, 11.6, 21.5	88.3, 75.0, 93.6
Binary – low % 0 outliers	X1, X2, X3	90.2, 75.5, 57.1	97.6, 91.1, 76.8	92.9, 81.6, 63.2	98.0, 93.6, 74.7	90.3, 76.0, 57.1	98.3, 95.5, 79.3
Binary – high % 0 outliers	X1, X2, X3	84.6, 77.2, 62.0	96.1, 93.2, 84.9	94.1, 90.8, 80.4	99.0, 98.7, 91.4	85.1, 77.5, 62.2	98.8, 98.4, 89.5
Random Effect Dependent Overdispersion							
Normal – $\gamma \sim N(0, g/10)$	X1, X2, X3	22.8, 19.5, 20.1	51.6, 43.5, 47.9	59.2, 50.6, 57.1	83.3, 74.3, 84.7	26.3, 22.8, 22.9	85.4, 76.4, 86.0
Normal – $\gamma \sim N(0, g/5)$	X1, X2, X3	10.7, 9.0, 11.5	43.3, 34.1, 41.2	65.8, 59.3, 65.0	66.1, 58.4, 64.1	12.5, 10.7, 12.2	74.2, 66.3, 76.9
Uniform – $\gamma \sim N(0, g/10)$	X1, X2, X3	67.1, 70.5, 65.3	84.4, 87.5, 83.8	92.6, 94.3, 92.2	88.8, 89.4, 86.5	67.0, 71.2, 65.5	91.0, 90.2, 88.2
Uniform – $\gamma \sim N(0, g/5)$	X1, X2, X3	37.2, 38.3, 33.7	74.6, 79.4, 75.9	91.0, 93.6, 92.1	76.4, 76.5, 75.7	36.2, 37.0, 35.3	83.8, 84.1, 83.2
Binary – $\gamma \sim N(0, g/10)$	X1, X2, X3	64.9, 70.5, 65.6	86.8, 91.0, 84.6	94.9, 95.2, 94.5	88.2, 89.0, 86.8	64.9, 70.1, 63.8	91.0, 90.8, 89.6
Binary – $\gamma \sim N(0, g/5)$	X1, X2, X3	34.3, 35.7, 33.4	76.3, 79.9, 74.3	93.8, 95.4, 93.2	74.3, 74.4, 74.0	35.2, 35.5, 30.8	83.2, 84.9, 82.5

4. Case studies

The motivating case study is a large randomized trial dataset containing some overdispersion which results from any of the three scenarios discussed above. The second case study is a classical small sample example of overdispersion in the literature where the overdispersion could be attributed to population heterogeneity. In both datasets we estimated overdispersion using the Pearson and deviance scales, which have been shown to agree with the other score statistics based test for overdispersion relative to the Poisson and negative-binomial models (Dean and Lawless 1989, Dean 1992, Deng and Paul 2000).

4.1 NLST dataset

The National Lung Screening Trial (NLST) randomized a total of 53,454 current and former smokers into two types of screening for lung cancer (Aberle, Adams et al. 2010). The purpose of this study was to compare lung cancer mortality rates of patients screened with a low-dose CT scan with those screened via chest radiography. Our interest is to examine the relationship between comorbidity count and whether patients were current or former smokers, adjusted for demographic covariates. Eligible participants were 55-74 years old, were either current or former smokers who had quit smoking within the last 15 years, and had a cigarette smoking history of 30 or more pack-years. Patients who were randomized to the CT scan showed a 20% and 6.7% reduction in lung cancer specific and all-cause mortality, respectively, compared with patients who received chest radiography. Demographic information was also collected for these patients to include comorbidity burden, race, gender, age, education status and smoking history.

We applied the six methods of analysis to the NLST dataset. The deviance/df value for the unadjusted Poisson model was 1.35, and the Pearson X^2/df value was 1.26, demonstrating mild overdispersion in the dataset. Table 3 gives the AIC and BIC values, SE, and 95% CI for each of the covariates included in the models.

Table 3. Comparison of methods for dealing with overdispersion in the NLST and *Salmonella* datasets.

NLST						
Value	Poisson	DS-Poisson	PS-Poisson	NB	Poisson-GLMM	NB-GLMM
AIC	158573.56	158573.56	158573.56	157208.53	158087.40	156833.00
BIC	158689.01	158689.01	158689.01	157332.86	158109.90	156856.90
Salmonella						
AIC	177.77	177.77	177.77	140.43	152.85	141.02
BIC	173.55	173.55	173.55	143.10	149.24	136.51

The NB-GLMM had the lowest AIC and BIC values followed by the NB model. The unadjusted and scale-adjusted Poisson models all had higher AIC and BIC values. The scale-adjusted Poisson models have increased the SE for the parameters, particularly the DS-Poisson. The NB, Poisson-GLMM, and NB-GLMM models moderately corrected the SE and the width of the corresponding 95% CI for the parameters. We can conclude that the NB-GLMM may be considered superior in dealing with the overdispersion present in the NLST dataset.

Table 4 gives NB-GLMM results comparing patient comorbidity burden with demographics. Former smokers had a higher comorbidity burden than current smokers (RR=1.11, $p<0.0001$), probably resulting in part from many years of smoking previously. There are also significant differences in comorbidity count based on patient gender, education, race, and age. Female patients had higher comorbidity burden than males (RR=1.08, $p<0.0001$). Patients who did not finish high school had the highest comorbidity burden among educational status (RR=1.24, $p<0.0001$). Non-Hispanic black patients had the highest comorbidity burden among the race groups (RR=1.08, $p=0.0116$). Not surprisingly, the youngest patients had the lowest comorbidity burden among the age categories (RR=0.69, $p<0.0001$).

Table 4. NB-GLMM model comparing comorbidity count with patient demographics in the NLST dataset.

Covariate	Rate Ratio	95% CI	P-Value
Former smoker vs. current smoker	1.11	(1.09, 1.12)	<0.0001
Female vs. Male	1.08	(1.07, 1.10)	
<High school	1.24	(1.15, 1.33)	<0.0001
High school	1.08	(1.02, 1.15)	0.0149
College	0.94	(0.88, 1.01)	0.0795
Graduate school	0.94	(0.88, 1.01)	0.0733
Other education (ref)	--	--	--
NHW	0.90	(0.86, 0.94)	<0.0001
NHB	1.08	(1.02, 1.15)	0.0116
Asian	0.91	(0.84, 0.99)	0.0271
Hispanic/Other (ref)	--	--	--
Age < 57	0.69	(0.67, 0.71)	<0.0001
57 ≤ Age < 60	0.76	(0.75, 0.79)	<0.0001
60 ≤ Age < 65	0.86	(0.84, 0.88)	<0.0001
Age ≥ 65 (ref)	--	--	--

4.2 *Salmonella* dataset

The Ames *Salmonella* dataset is a classic example of the presence of overdispersion in a small dataset (Mortelmans and Zeiger 2000). The variables in this dataset include three different plates, six levels of medication dose on each plate, and a count response of *Salmonella* bacterial colonies (refer to Figure 8 in Appendix 2). The medication dose variable was modeled as a log dose in this analysis (the smallest non-zero dose size of 10 was first added to the variable in order to avoid a log of zero).

The deviance/df value for the unadjusted Poisson model was 4.69, and the Pearson X^2/df value was 5.33, demonstrating the presence of overdispersion in the *Salmonella* dataset. This dataset was analyzed using the six approaches and the results are reported in Table 3. The AIC and BIC values, parameter SE, and 95% parameter CI were included for comparison.

Based on the AIC and BIC criteria, the NB-GLMM demonstrated the best goodness of fit in this overdispersed dataset. The NB, NB-GLMM and Poisson-GLMM also gave SE values that

are higher than those in Poisson but lower than scale-adjusted Poisson. The scale-adjusted Poisson models both appeared to have much larger SE, particularly the PS-Poisson. The 95% CI appeared to follow the same trend. Overall, the NB-GLMM may be considered superior in dealing with overdispersion present in the *Salmonella* dataset based on the AIC and BIC criteria, SE, and 95% CI estimates. This is likely because the overdispersion in this case study was at least in part the result of correlation in bacterial count outcome by plate, which was included in the model via the random effect. In the NB-GLMM model, the log dose variable significantly effects bacterial count outcome (RR=1.13, p=0.0194).

5. Discussion

In this paper, we provide a comprehensive comparative analysis of six different models for dealing with overdispersion caused by different mechanisms when modeling count data. Overall, the negative-binomial models appeared to demonstrate superiority in adjusting for overdispersion in the simulation studies. The NB-GLMM performed best in modeling count of comorbidity data in the motivating NLST study. This model also appeared to deal most effectively with overdispersion in the small *Salmonella* dataset.

Based on our analyses, we conclude that NB-GLMM is superior overall for modeling count data characterized by overdispersion, jointly considering all criteria. The negative-binomial distribution is often used instead of Poisson to account for overdispersion resulting from omitted important covariates and population heterogeneity, among other causes. Therefore, it is reasonable that overdispersion caused by the omission of important predictors, the addition of high or zero outliers to the outcome, and the omission of a random effect would be effectively controlled by using models that are based on the negative-binomial distribution. For example, as in the NB-GLMM, the addition of random effects is shown to be effective in dealing with overdispersion resulting from with-in subject correlation of count outcome.

Our results further demonstrate that the best method for dealing with overdispersion will likely vary by dataset depending on the cause of the overdispersion. The negative-binomial model

may account for overdispersion due to a number of common causes, but it is not ideal in every case. Numerous model options should be considered when overdispersion is an issue.

In order to make application of these results to real datasets, a clinician should first check dispersion via the deviance/df and Pearson X^2/df values to determine whether they are greater than 1. If the count outcome is overdispersed, the clinician should attempt to identify the cause of the overdispersion via testing of parameter significance, identifying excessive high or zero outliers in the outcome, and checking for the presence of random effects in the data. It may be possible to address the issue with simple model adjustments. To address the overdispersion via scale or modeling methods, Poisson and negative-binomial regression should both be considered as in our analysis and compared via parameter standard errors and goodness-of-fit statistics. It should also be determined that the benefit of utilizing the negative-binomial distribution will outweigh the added model complexity. Table 5 gives a summary of the possible overdispersion causes examined in our analysis and our corresponding choices of modeling method.

Table 5. Summary of methods chosen to deal with overdispersion by cause.

Type of Overdispersion	Methods and Comments
Covariate dependent	NB and NB-GLMM performed best overall, jointly considering goodness-of-fit, error, and coverage criteria. NB-GLMM is preferable if the data includes random effects. The scale-adjusted Poisson methods performed fairly well with non-normal covariates and could also be considered.
Outlier dependent: high outliers	NB-GLMM and PS-Poisson performed best overall, jointly considering all criteria. NB-GLMM is preferable if the data includes random effects.
Outlier dependent: zero outliers	NB-GLMM performed best for normal covariate scenarios, jointly considering all criteria. The NB and scale-adjusted Poisson methods performed fairly well with non-normal covariates and could also be considered.
Random effects dependent	NB and NB-GLMM performed best overall, jointly considering all criteria. The DS-Poisson performed fairly well with non-normal covariates and could also be considered. A random effect should be included.

This article illustrates how negative-binomial regression and NB-GLMM can be used to effectively model overdispersed count outcomes. It also showed that simple post hoc scaling in the Poisson model to decrease overdispersion was not consistently effective. Basic scaling does not take the specific cause of the overdispersion into account. Overdispersion may result from a variety of causes, which must be considered to determine the most effective method of dealing with it.

To more thoroughly analyze the options for dealing with overdispersion present in datasets with count outcomes, we plan to examine the performance of these methods in the presence of missing covariate data. Pacheco et al. recently performed a related comparison of various methods for dealing with overdispersion using simulated time-dependent data, including generalized estimating equations models, generalized linear mixed models, and Bayesian methods (Durán Pacheco, Hattendorf et al. 2009). But there are none that address the co-occurrence of both covariate missingness and overdispersion. Future studies need to explore and address how to handle co-occurrence of overdispersion and missing covariate data.

An empirical approach to determine a threshold for declaring the presence of overdispersion in count data

CHAPTER 3

1. Introduction

The assumption of Poisson regression that the conditional mean must be equal to the conditional variance often fails in real data situations. Overdispersion occurs when data have greater conditional variance than is assumed under the Poisson model (Cox 1983), which may result from population heterogeneity, correlation, omission of important covariates in the model, the presence of high or zero outliers, or other reasons (Hardin and Hilbe 2007, Rigby et al. 2008). A Poisson model estimated on overdispersed data can include underestimated standard errors of the parameter estimates. As a consequence, the hypotheses on the regression parameters may be rejected more often than they should be (McCullagh and Nelder 1989, Breslow 1990, Hilbe 2011, Faddy and Smith 2011). We examined overdispersion occurring in real and simulated datasets resulting from outliers, omission of key predictors, and omission of necessary random effects (Payne et al. 2015). We compared six different scaling and modeling methods of analysis via goodness of fit and error statistics. The results showed that negative binomial regression and negative binomial generalized linear mixed models were preferred for dealing with overdispersion resulting from the sources we considered. Scaling methods and unadjusted Poisson regression were less reliable and often produced larger or smaller standard errors than expected.

The two most commonly used estimators of dispersion in the literature are the ratio of the model deviance to its corresponding degrees of freedom and the ratio of the Pearson χ^2 statistic to its corresponding degrees of freedom (McCullagh and Nelder 1989). For a study with sample n and p predictors, the degrees of freedom are typically given by $n - p$. This ratio will equal one when the Poisson assumption or, equivalently, the assumption that the conditional mean and variance are equal, holds. Relative to the model, the data are considered overdispersed if this ratio is greater than one, with greater magnitudes of overdispersion corresponding to higher Pearson χ^2 statistics.

A likelihood ratio test may be used to test the difference of the simple Poisson and a more complex models such as negative binomial regression to assess whether the simpler model

should be rejected (Cameron and Trivedi 1986). The Wald statistic associated with a test of the dispersion parameter in the more complex model may also be used for this assessment (Molla and Muniswamy 2012). Score tests for determining the presence of extra-Poisson variation are also available in many case-specific variations (Gurmu 1991, Dean and Lawless 1989, Lee et al. 2007, Breslow 1990, Collings and Margolin 1985), and may be more appropriate than Wald or likelihood ratio tests since the score test requires only an estimation of the simpler model and provides greater power (Yang et al. 2007). In addition, hypothesis testing of the ratios of negative binomial and Poisson regression log-likelihoods may rely on asymptotic distributions which underestimate the evidence against the base model and thereby provide results which are misleading (Cameron and Trivedi 1998, Dean 1992, Lawless 1987). O'Hara Hines provides an overview of numerous score tests which have been developed to test for overdispersion (1997). Molla and Muniswamy recently demonstrated the superior power of the score test compared to likelihood ratio and Wald tests via an extensive Monte Carlo simulation study (2012).

Currently, one of the most commonly used estimators of dispersion in the literature is the goodness of fit ratio of the Pearson χ^2 statistic to its corresponding degrees of freedom. A decision about whether data are overdispersed is made by checking whether this ratio is bigger than one. The relative variance is defined as the ratio of the variance to the mean and is theoretically comparable to the Pearson χ^2 ratio with its degrees of freedom. One possible rule of thumb suggests that if the relative variance is greater than two, then the data may be considered overdispersed and require statistical intervention (Cameron and Trivedi 1990). In this case, the average of the covariate-pattern specific ratio of the conditional variance to conditional mean of the count outcome is more than two, contradicting the Poisson model. Smaller values in the average of the ratios of conditional variance to conditional mean may still point to an overdispersed model which underestimates the parameter standard errors and requires a more complex modeling strategy than simple Poisson regression (Rodriguez 2015). In some cases,

relative variance tests and curves may be more effective in identifying the presence of overdispersion than score tests (Lambert and Roeder 1995).

In this paper, we examine count outcomes containing overdispersion represented by varying magnitudes of Pearson χ^2 ratios in cross-sectional and longitudinal datasets, to determine the threshold over 1 at which overdispersion may be considered detrimental to data analysis if ignored. We examine scenarios in which overdispersion is the result of either outliers or zero inflation in the count outcome. Results from two real case studies containing varying magnitudes of overdispersion are also considered. This paper is organized in the following manner. Subsequent to the introduction, a description of the statistical models as well as measures and tests of overdispersion is given in section 2. Section 3 provides information about the design of the simulation study. Section 4 provides the results of the simulation study. Section 5 gives a description and results from our real datasets. Section 6 gives a conclusion and discussion based on all results.

2. Statistical Models and Estimation

2.1. Models

For cross-sectional data, let vector $Y = (Y_1, \dots, Y_n)'$ be a response vector with independent and identically Poisson distributed random Y values. The variance function is $Var(Y_i) = \mu_i$ and the probability mass function for the quasi-Poisson is given by

$$f(y_i | \mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \quad (1)$$

with $0 \leq y_i < \infty$ and positive conditional mean parameter μ_i . The conditional variance as a function of the conditional mean is given by $\varphi\mu$, with dispersion parameter φ . There is equidispersion in the dataset when $\varphi = 1$, while if $\varphi < 1$ there is underdispersion, and if $\varphi > 1$ there is overdispersion. The Poisson can be extended to define the generalized Poisson

regression model including covariates for which the conditional mean is $E(Y_i) = \mu_i = \exp(\mathbf{X}'\beta)$ via the following format (Rodriguez 2015):

$$\Pr(Y_i | \mu_i, \beta, k) = \left(\frac{\mu_i}{1 + \varphi\mu_i} \right)^{y_i} \frac{(1 + \varphi y_i)^{y_i - 1}}{y_i!} \exp\left(-\frac{\mu_i(1 + \varphi y_i)}{1 + \varphi\mu_i} \right) \quad (2)$$

with $0 \leq y_i < \infty$. A score test may then assess the parameter φ to determine whether the conditional variance exceeds the conditional mean (refer to Section 2.2).

If $Y | \theta \sim \text{Pois}(\theta)$ and θ is a random variable such that $E(\theta) = \mu$ and $\text{Var}(\theta) = \sigma^2$, then $E(Y) = \mu$ and $\text{Var}(Y) = \mu + \sigma^2$, indicating greater variance compared to the mean; if θ is assumed to be distributed gamma, then Y follows a negative-binomial distribution with

$$E(Y) = \frac{k}{\lambda} = \mu \text{ and } \text{Var}(Y) = \mu + \frac{\mu^2}{k} \text{ (Payne et al. 2015).}$$

Random effects may also be included to deal with overdispersion. For vectors of fixed effect (\mathbf{X}_i) and random effect (\mathbf{Z}_i) for explanatory variables ($i = 1, \dots, n$) the GLMM family is given by,

$$E(Y_i | \mathbf{X}_i, \mathbf{Z}_i) = \mathbf{g}^{-1}(\mathbf{X}_i \beta + \mathbf{Z}_i b_i) = \mu_i \quad (3)$$

Here, β is a vector of p fixed coefficients, \mathbf{g} is a monotone link function, and b_i is a vector of unobserved normally-distributed random deviations with zero mean for which the variance will be estimated. The conditional variance for this model is given by

$$\text{Var}(Y_i) = \mu_i + k\mu_i^2.$$

NB-GLMM allows for greater conditional variance than assumed by the Poisson-GLMM. We have previously showed that NB and NB-GLMM are superior for dealing with overdispersion compared to other models in various scenarios, jointly considering the specified criteria (Payne et al. 2015).

We also consider a generalized linear model setup for longitudinal scenarios. While a general set of predictor variables is allowed, we focus on a scenario including two covariates: a

main predictor variable and possible confounder. Let Y_{ij} be a response, while X_{1ij} and X_{2ij} are covariates of interest at the j^{th} repeated measure for the i^{th} subject ($i = 1, \dots, n, j = 0, \dots, T_i$). Let q_i denote the random effects for each individual i which could be assumed to have a normal distribution with zero mean and covariance G . Let $\beta = (\beta_1, \beta_2)$ be the regression coefficients corresponding to X_{1ij} and X_{2ij} , respectively, and

$$\eta_{ij} = q_i + X_{1ij}\beta_1 + X_{2ij}\beta_2 \quad (4)$$

We can rewrite this in vector form as for the cross-sectional GLMM above:

$$\eta_i = Z_i b_i + X_i \beta \quad (5)$$

where $X_i = (X_{1ij}, X_{2ij})'$, $\eta_i = g(E[Y_{ij} | q_i, \beta])$, g is a monotone link function, Z_i is the random effects design matrix and b_i is the random effects vector for each individual i .

In this paper we address overdispersion resulting from the presence of outliers or zero inflation in the count outcome in both cross-sectional and longitudinal datasets. We consider four methods for analyzing cross-sectional data as in our previous work: unadjusted Poisson regression (Poisson), negative-binomial regression (NB), and two GLMM with random intercept, log link, and compound symmetry covariance, with outcomes distributed as Poisson and negative-binomial (Poisson-GLMM, NB-GLMM, respectively) (Payne et al. 2015). In the longitudinal scenario, we considered GLMM with random intercept to account for individual variability with outcomes distributed as either Poisson or negative-binomial (Poisson-GLMM, NB-GLMM, respectively). SAS 9.4 was utilized in all analyses, particularly the *Proc GENMOD* and *Proc GLIMMIX* packages.

2. 2. Tests and Measures of Overdispersion

A variety of score, Wald, and likelihood ratio tests have been considered to determine when overdispersion is statistically significant. One score statistic (Yang et al. 2009) for testing

whether the dispersion parameter indicates extra-Poisson variation $H_0 : \varphi = 0$ vs. $H_1 : \varphi > 0$ is given by

$$S_1(\hat{\beta}) = \left(\sum_{i=1}^n 2\hat{\mu}_i^2 \right)^{-1} \left(\sum_{i=1}^n ((y_i - \hat{\mu}_i)^2 - y_i) \right)^2 \quad (6)$$

Under the null hypothesis that overdispersion is not present and the data follow an unadjusted Poisson model, the score statistic is distributed according to the χ_1^2 distribution with 1 degree of freedom. We can also write this score statistic as

$$S_2(\hat{\beta}) = \left(\sqrt{2 \sum_{i=1}^n \hat{\mu}_i^2} \right)^{-1} \sum_{i=1}^n ((y_i - \hat{\mu}_i)^2 - y_i) \quad (7)$$

which is asymptotically distributed as a standard normal. It is clear from the structure of this statistic that greater variability between observed and predicted values will increase the magnitude of the score statistic, which implies overdispersion resulting from data heterogeneity or other factors. According to this statistic, we can reject the assumption of equidispersion at a significance level of 0.05 via a one-sided test if score statistic $S_2(\hat{\beta})$ is greater than the 95th percentile of the $N(0,1)$ distribution. This gives us a score statistic cutoff of 1.65 for declaring the presence of overdispersion in large samples. Though this is a useful paradigm, our interest is in determining a general threshold for declaring the presence of overdispersion across datasets using the commonly considered Pearson χ^2 ratio to its degrees of freedom. We will provide a crossover comparison of rejection via score test at each of our considered Pearson χ^2 ratios.

Using our notation, the Pearson χ^2 statistic is defined for the Poisson distribution within the context of GLMs as below (Morel and Neerchal 2012):

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \quad (8)$$

This statistic is commonly utilized to analyze model goodness of fit, and is approximately distributed χ^2_{df} [19]. For a study with sample n and p predictors, the degrees of freedom are typically given by $n - p$. Dispersion parameter σ_p is therefore defined as the ratio of the Pearson χ^2 to its degrees of freedom and is approximately unbiased (Ruoyan 2004) as follows:

$$\sigma_p = \frac{\chi^2}{n - p} \quad (9)$$

Dispersion parameter σ_p will equal one where the assumption of equal mean and variance holds. Our goal is to determine if there is an appropriate threshold for declaring overdispersion requiring statistical intervention via the popular Pearson χ^2 goodness of fit statistic using dispersion parameter σ_p . This value may also be used to determine the presence of underdispersion in datasets, though this is a less common scenario when working with real clinical data.

3. Simulation

3.1. Design

We simulated 200 cross-sectional datasets each with a sample size of 100 random observations, to include a Poisson count outcome and $m = 2$ binary predictor variables X_1 and X_2 according to the model $\log(E(Y_{im} = y | X_{im})) = \alpha + \sum_{m=1}^2 \beta_m X_{im}$ where β is the collection of parameters (β_1, β_2) and $\alpha = 1.0$. Outcome count Y for the i^{th} individual was determined by $\exp(\alpha + \sum_{m=1}^2 \beta_m X_{im})$. We alternated assigning true parameter value $\beta_1 = [0.01, 0.41, 0.92]$ to yield odds ratios of 1.0, 1.5, and 2.5 respectively, and assigned true parameter value $\beta_2 = 0.69$ to yield an odds ratio of 2.0 as a potential confounder.

We then created overdispersion relative to the Poisson in the first scenario via the addition of outliers to the count outcome Y . A random sample of 10% of the Y values in each simulation was increased to create outlier-dependent overdispersion in the data such that running unadjusted Poisson regression or Poisson-GLMM resulted in varying values of $\sigma_p = [1.0, 1.2, 1.3, 1.4, 1.5, 2.0, 2.5, 5.0, 10.0]$. We created a second scenario in which the unadjusted Poisson gave overdispersion magnitudes $\sigma_p = [1.0, 1.2, 1.3, 1.4, 1.5, 2.0, 2.5, 5.0]$ by setting various percentages of the Y outcome variable to zero (we could not achieve 10.0 here).

Recall our discussion of a score test statistic (Yang et al. 2009) to test $H_0 : \varphi = 0$ vs.

$H_1 : \varphi > 0$ presented in Section 2.2. The frequency of rejection of $H_0 : \varphi = 0$ via score test for both outlier-dependent and zero-dependent overdispersion of all magnitudes is given in Table 1.

Table 1. Percent of simulations at varying levels of overdispersion in which the score test did in fact reject the null hypothesis and affirm the presence of overdispersion in the dataset.

σ_p	Outlier Dependent		
	$\beta_1=0.01$	$\beta_1=0.41$	$\beta_1=0.92$
1.0	6.50	6.50	6.50
1.2	55.50	50.50	28.50
1.3	71.50	53.00	47.00
1.4	79.00	71.50	68.00
1.5	95.50	88.00	87.50
2.0	99.50	99.50	98.50
2.5	100.00	100.00	100.00
5.0	100.00	100.00	100.00
10.0	100.00	100.00	100.00
σ_p	Zero Inflation		
	$\beta_1=0.01$	$\beta_1=0.41$	$\beta_1=0.92$
1.0	6.50	6.50	6.50
1.2	46.50	43.50	51.50
1.3	63.50	64.00	56.50
1.4	74.50	77.00	69.00
1.5	90.00	90.00	85.00
2.0	100.00	100.00	99.00
2.5	100.00	100.00	100.00
5.0	100.00	100.00	100.00

σ_p is defined as the ratio of the Pearson χ^2 to its degrees of freedom

Higher percentages of rejection via the score test statistic in simulations indicate overdispersion in the dataset at the given level of σ_p , suggesting that statistical intervention is necessary. From

this table we can see that values of $1.5 \leq \sigma_p \leq 2.0$ result in a percentage of rejection close to the nominal 95% depending on the effect size of β_1 , indicating rejection of $H_0 : \varphi = 0$ and conclusion that the data are overdispersed according to the score test. Values of $\sigma_p < 1.5$ result in lower rejection percentages under both scenarios and therefore do not reject the null hypothesis of equidispersion. Higher effect sizes give slightly more conservative results. At values of $\sigma_p \geq 2.5$, equidispersion is rejected in 100% of cases.

We further simulated 200 longitudinal datasets of the same initial sample size of 100 to include the time-varying Poisson count outcome and two time-varying binary predictor variables

according to the model $\log(E(Y_{ijm} = y | X_{ijm})) = \alpha + \sum_{m=1}^2 \beta_m X_{ijm}$ with data now taken at five

continuous time points $j = 1, 2, \dots, 5$. Again, β is the collection of parameters (β_1, β_2) and

$\alpha = 1.0$. Outcome count Y for the i^{th} individual was now generated using a mean

$\exp(\alpha + \sum_{m=1}^2 \beta_m X_{ijm})$. In the outlier-dependent scenario, random Y values were similarly

increased at baseline each simulation as for the cross-sectional datasets. In the zero-dependent scenario, varying percentages of random Y values were set to 0 over time as for the cross-sectional datasets.

Comparison among models in all scenarios was then made using Type 1 and Type 2 errors, as well as coverage probabilities of β_1 . Type 1 error is determined via the percentage of simulations in which the effect of β_1 is detected though not present, i.e. the percentage of false positives; here we consider datasets with a true β_1 value of 0.01. Type 2 error is determined via the percentage of simulations in which the effect of β_1 is not detected though present, i.e. the percentage of false negatives. These errors are observed for both true β_1 values of 0.41 and

0.92. Coverage probabilities are considered for all values of $\beta_1 = [0.01, 0.41, 0.92]$ and are the percentage of simulations in which parameter 95% confidence intervals contain the true β_1 .

4. Results

4.1. Cross-Sectional Results

Poisson and negative binomial results for both cross-sectional scenarios are given in Tables 2 and 3, respectively, and illustrated in Figures 1a-b and 2a-b by model type and value of β_1 at all considered values of σ_p .

Table 2. Percentage of simulations with X_1 Type 1 errors and Type 2 errors and in which parameter coverage included the true parameter given true values of 0.01, 0.041, and 0.92 for the cross-sectional scenario using the unadjusted Poisson model and Poisson GLMM.

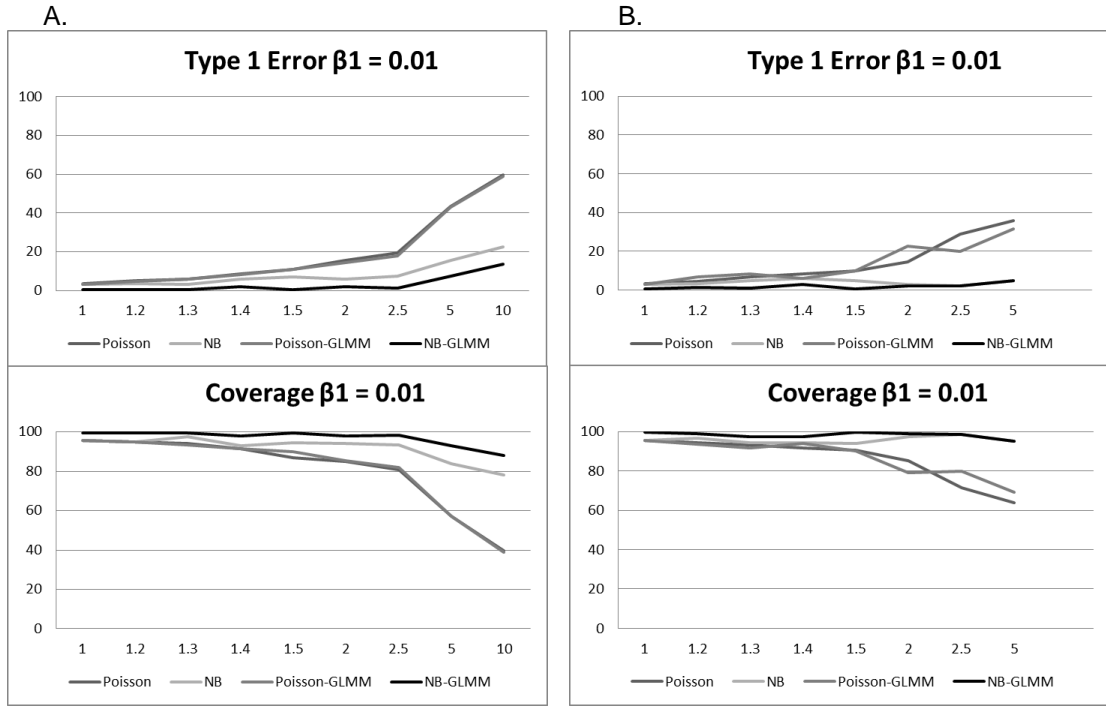
σ_p	Outlier Dependent						Zero Inflation					
	Unadjusted Poisson						Unadjusted Poisson					
	$\beta_1=0.01$		$\beta_1=0.41$		$\beta_1=0.92$		$\beta_1=0.01$		$\beta_1=0.41$		$\beta_1=0.92$	
	Type 1 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 1 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 2 (%)	Coverage (%)
1.0	3.50	95.50	0.50	92.50	0.00	91.50	3.50	95.50	0.50	92.50	0.00	91.50
1.2	5.00	95.00	1.50	88.50	0.00	81.50	4.50	94.50	1.00	89.50	0.00	89.50
1.3	6.00	94.00	1.50	87.50	0.00	76.00	7.00	93.00	2.00	91.00	0.00	88.50
1.4	8.50	91.50	1.50	86.50	0.00	77.00	8.50	91.50	3.00	86.00	0.00	89.00
1.5	11.00	87.00	2.00	83.50	0.00	66.00	10.00	90.50	6.00	83.50	0.00	86.50
2.0	15.50	85.00	6.50	72.00	0.00	59.00	14.50	85.00	11.00	79.50	0.00	86.50
2.5	19.50	81.00	8.50	71.00	0.00	52.50	29.00	71.50	14.50	74.00	0.00	82.00
5.0	43.50	57.00	21.00	53.00	0.00	29.00	36.00	64.00	43.00	56.50	0.00	63.50
10.0	59.50	39.50	23.00	42.50	1.00	23.00	--	--	--	--	--	--
σ_p	Poisson GLMM						Poisson GLMM					
	$\beta_1=0.01$		$\beta_1=0.41$		$\beta_1=0.92$		$\beta_1=0.01$		$\beta_1=0.41$		$\beta_1=0.92$	
	Type 1 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 1 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 2 (%)	Coverage (%)
1.0	3.00	95.50	1.00	93.50	0.00	92.00	3.00	95.50	1.00	93.50	0.00	92.00
1.2	4.50	95.00	2.00	89.50	0.00	81.50	7.00	93.50	2.00	91.50	0.00	90.00
1.3	6.00	93.50	2.50	89.00	0.00	77.50	8.50	91.50	2.50	91.50	0.00	90.50
1.4	8.00	91.50	1.50	87.00	0.00	78.00	6.00	94.00	2.50	90.50	0.00	89.50
1.5	11.00	90.00	2.50	84.00	0.00	67.50	10.00	90.00	4.00	86.50	0.00	89.00
2.0	14.50	85.50	7.50	74.00	0.00	61.50	22.50	79.00	8.50	83.50	0.00	80.50
2.5	18.00	82.00	8.50	71.50	0.00	52.50	20.00	80.00	8.50	78.50	0.00	82.00
5.0	43.00	57.00	22.00	54.00	0.00	30.00	31.50	69.00	53.50	60.50	1.00	65.50
10.0	59.00	39.00	24.00	42.50	1.50	23.50	--	--	--	--	--	--

σ_p is defined as the ratio of the Pearson χ^2 to its degrees of freedom

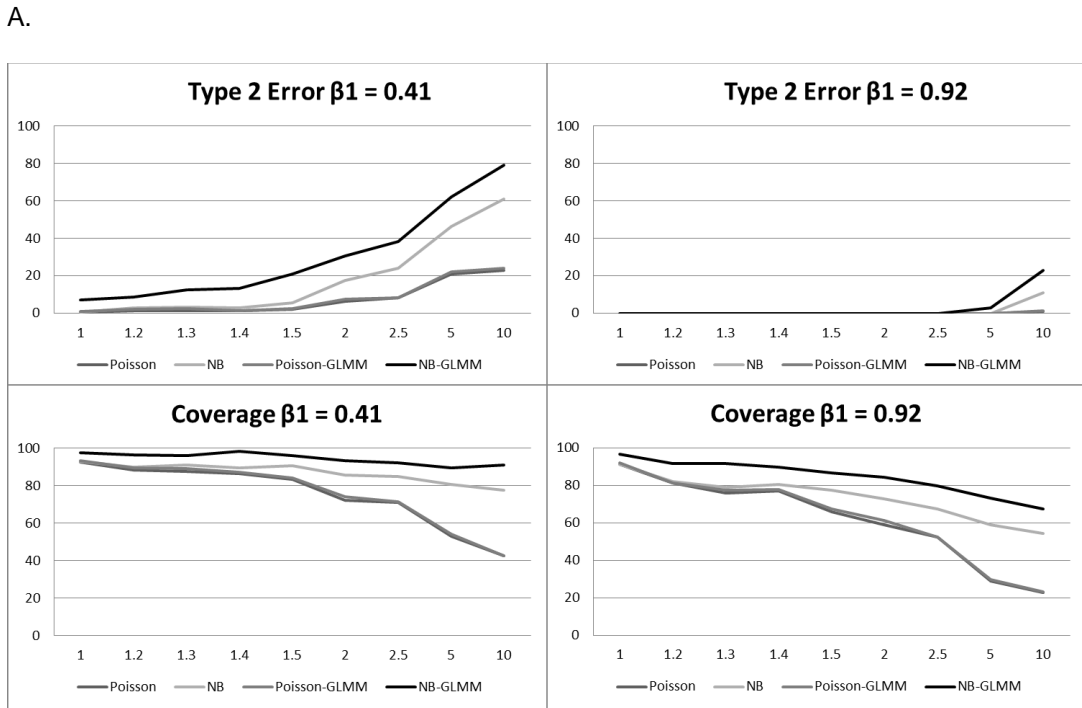
Table 3. Percentage of simulations with X_1 Type 1 errors and Type 2 errors and in which parameter coverage included the true parameter given true values of 0.01, 0.041, and 0.92 for the cross-sectional scenario using the negative binomial regression model and negative binomial GLMM.

σ_p	Outlier Dependent						Zero Inflation					
	Negative Binomial						Negative Binomial					
	$\beta_1=0.01$		$\beta_1=0.41$		$\beta_1=0.92$		$\beta_1=0.01$		$\beta_1=0.41$		$\beta_1=0.92$	
	Type 1 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 1 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 2 (%)	Coverage (%)
1.0	3.00	95.50	0.50	93.00	0.00	91.50	3.00	95.50	0.50	93.00	0.00	91.50
1.2	3.50	95.00	3.00	90.00	0.00	82.00	3.50	96.50	4.50	92.50	0.00	94.00
1.3	3.00	97.50	3.50	91.00	0.00	79.00	5.00	94.50	5.50	93.00	0.00	92.50
1.4	6.00	93.00	3.00	89.50	0.00	80.50	6.00	94.50	8.50	96.00	0.00	93.50
1.5	7.00	94.50	5.50	90.50	0.00	77.50	5.00	94.00	10.00	95.00	0.00	93.50
2.0	6.00	94.00	17.50	85.50	0.00	73.00	3.00	97.50	33.50	97.50	0.00	96.00
2.5	7.50	93.50	24.00	85.00	0.00	67.50	2.00	98.50	60.50	99.00	0.00	95.00
5.0	15.50	84.00	46.50	80.50	0.00	59.00	5.00	95.00	93.00	97.50	34.50	98.50
10.0	22.50	78.00	61.00	77.50	11.00	54.50	--	--	--	--	--	--
σ_p	Negative Binomial GLMM						Negative Binomial GLMM					
	$\beta_1=0.01$		$\beta_1=0.41$		$\beta_1=0.92$		$\beta_1=0.01$		$\beta_1=0.41$		$\beta_1=0.92$	
	Type 1 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 1 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 2 (%)	Coverage (%)
1.0	0.52	99.48	7.33	97.38	0.00	96.88	0.52	99.48	7.33	97.38	0.00	96.88
1.2	0.51	99.49	8.59	96.46	0.00	91.96	1.52	98.99	11.00	96.00	0.00	97.47
1.3	0.50	99.50	12.63	95.96	0.00	91.96	1.02	97.46	12.56	98.49	0.00	97.50
1.4	2.00	98.00	13.50	98.50	0.00	89.95	3.02	97.49	13.00	95.50	0.00	96.97
1.5	0.50	99.50	21.00	96.00	0.00	86.93	0.50	99.50	19.00	98.50	0.00	96.00
2.0	2.00	98.00	30.50	93.50	0.00	84.50	2.00	99.00	43.00	96.50	0.00	97.50
2.5	1.00	98.50	38.50	92.00	0.00	80.00	2.00	98.50	59.50	98.00	0.00	96.00
5.0	7.50	93.00	62.00	89.50	3.00	73.50	5.03	94.97	88.94	96.48	18.00	98.50
10.0	13.50	88.00	79.00	91.00	23.00	67.50	--	--	--	--	--	--

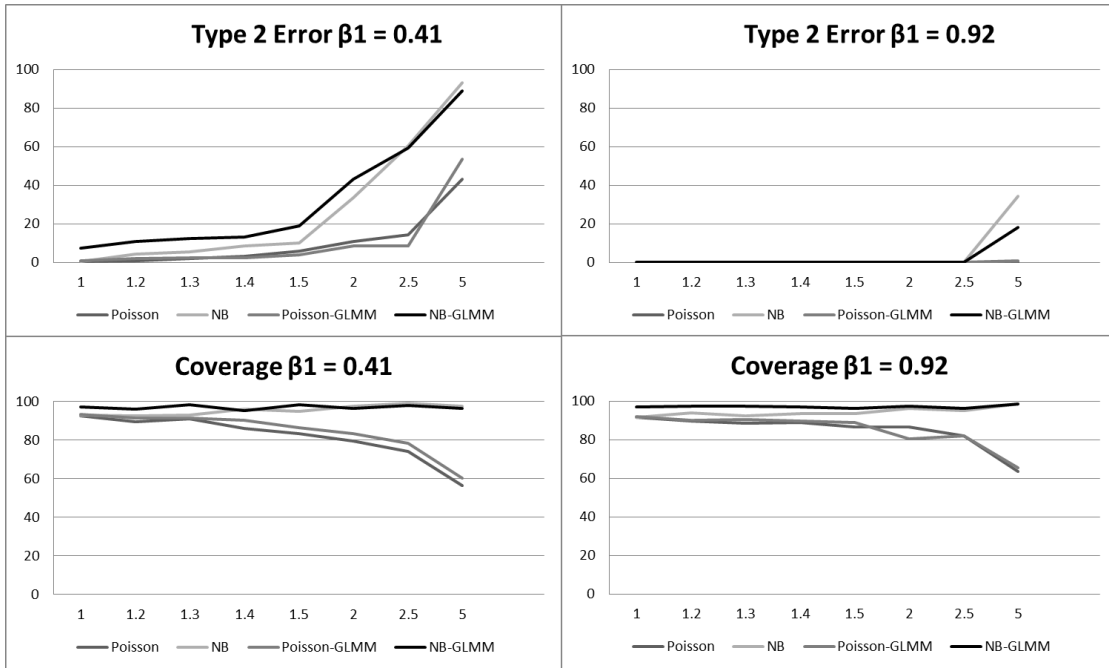
σ_p is defined as the ratio of the Pearson χ^2 to its degrees of freedom



Figures 1a-b. Percentage of simulations with Type 1 errors and in which parameter coverage included the true parameter given a true parameter value of 0.01 in the cross-sectional scenario, for a.) outlier-dependent overdispersion and b.) overdispersion caused by zero inflation.



B.



Figures 2a-b. Percentage of simulations with Type 2 errors and in which parameter coverage included the true parameter given true parameter values of 0.41 and 0.92 in the cross-sectional scenario, for a.) outlier-dependent overdispersion and b.) overdispersion caused by zero inflation.

Increases in magnitude of both outlier-dependent and zero-dependent overdispersion result in increases in Type 1 and Type 2 errors of the β_1 estimates as well as a decrease in coverage probabilities. Not surprisingly, the Type 2 error and coverage probabilities decrease with the higher effect size. Given the Type 1 error results, the unadjusted Poisson regression model and Poisson-GLMM perform fairly well for both scenarios with low overdispersion magnitude, particularly when $\sigma_p \leq 1.2$. The negative binomial regression models have higher tolerance for extra variability, performing well up to $\sigma_p \leq 1.4$. Furthermore, the NB-GLMM gives acceptable results in some cases up to $\sigma_p \leq 5.0$.

It would appear the simple Poisson model may be utilized in cross-sectional cases where $\sigma_p \leq 1.2$. Furthermore, negative binomial regression should be utilized if $1.2 < \sigma_p \leq 1.5$ while NB-GLMM should be utilized for higher values up to $\sigma_p \leq 5.0$.

There is clearly an effect of overdispersion on the models for values of σ_p lower than those picked up by the score test. NB-GLMM also results in the highest Type 2 error of all considered models, suggesting that negative binomial regression may be sufficient in some cases to address overdispersion of higher magnitude in these scenarios. The contrast between negative binomial and Poisson distribution models becomes more obvious as the magnitude of σ_p increases.

4. 2. Longitudinal Results

Results for both longitudinal scenarios are given in Table 4 and illustrated in Figures 3a-b and 4a-b for all considered values of σ_p , calculated under the Poisson-GLMM model.

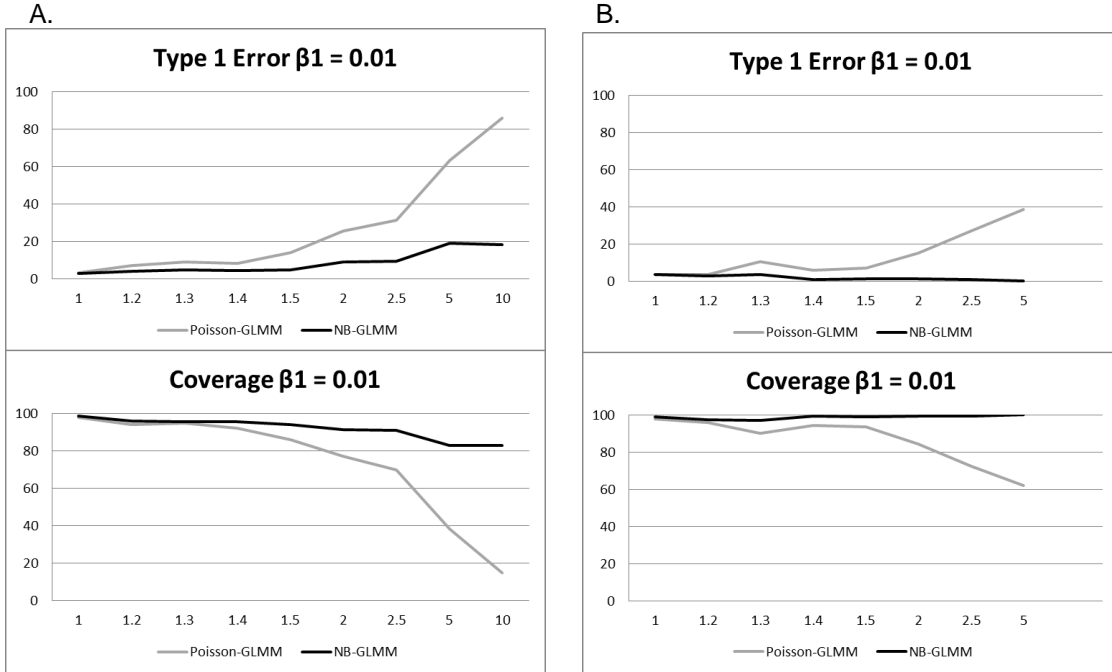
Longitudinal results are similar to those for the cross-sectional analysis. Given the percentage values of the Type 1 errors, the Poisson-GLMM again performs fairly well in addressing both outlier-dependent and zero-dependent overdispersion when $\sigma_p \leq 1.2$.

For larger magnitudes of overdispersion, up to $\sigma_p \leq 2.5$, NB-GLMM performs well. NB-GLMM results in considerably lower Type 1 errors and higher coverage probabilities and comparable Type 2 errors compared to Poisson-GLMM. As the magnitude of σ_p increases, the superiority of the NB-GLMM model becomes more apparent as the difference in errors and coverage increases compared to the Poisson-GLMM. Results become much less reliable when $\sigma_p \geq 5.0$.

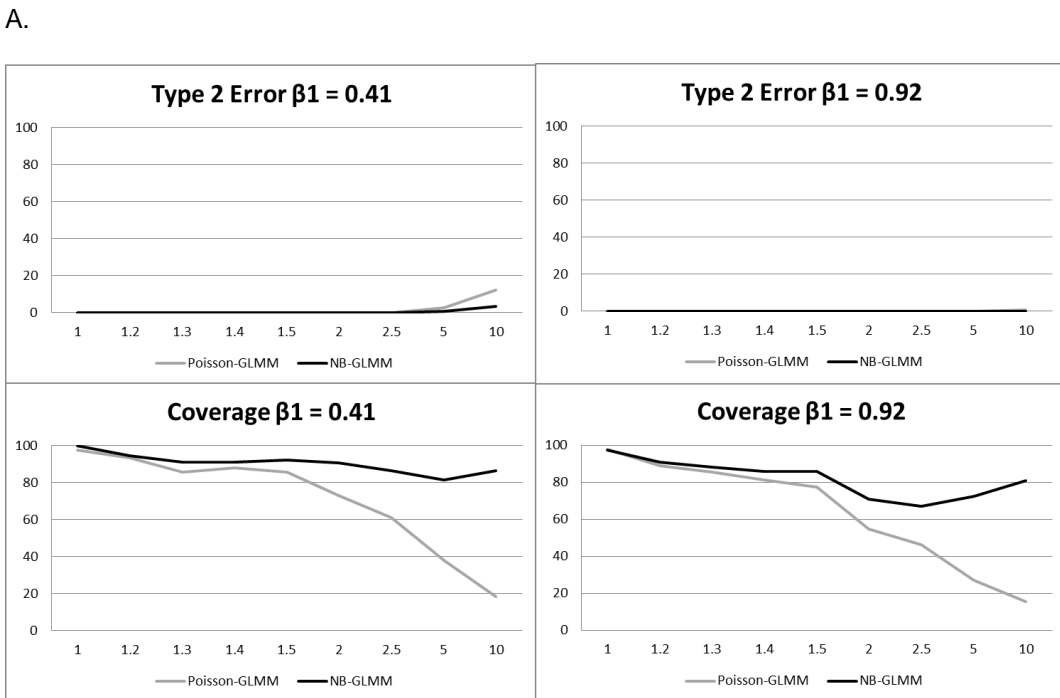
Table 4. Percentage of simulations with X_1 Type 1 errors and Type 2 errors and in which parameter coverage included the true parameter given true values of 0.01, 0.041, and 0.92 for the longitudinal scenario using Poisson and negative binomial GLMM.

σ_p	Outlier Dependent						Zero Inflation					
	Poisson GLMM						Poisson GLMM					
	$\beta_1=0.01$		$\beta_1=0.41$		$\beta_1=0.92$		$\beta_1=0.01$		$\beta_1=0.41$		$\beta_1=0.92$	
	Type 1 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 1 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 2 (%)	Coverage (%)
1.0	3.50	98.00	0.00	97.50	0.00	98.00	3.50	98.00	0.00	97.50	0.00	98.00
1.2	7.00	94.00	0.00	93.50	0.00	89.00	3.50	96.00	0.00	93.00	0.00	90.00
1.3	9.00	95.00	0.00	85.50	0.00	85.50	10.50	90.00	0.00	93.50	0.00	90.00
1.4	8.50	92.00	0.00	88.00	0.00	81.50	6.00	94.50	0.00	86.50	0.00	88.50
1.5	14.00	86.00	0.00	85.50	0.00	77.50	7.00	93.50	0.00	86.50	0.00	89.00
2.0	25.50	77.00	0.00	73.00	0.00	55.00	15.00	84.50	0.00	84.00	0.00	80.50
2.5	31.50	70.00	0.00	61.00	0.00	46.50	27.00	72.50	0.00	68.00	0.00	75.50
5.0	63.50	38.50	2.50	38.00	0.00	27.00	38.50	62.00	8.00	66.00	0.00	56.50
10.0	86.00	15.00	12.00	18.50	0.50	15.50	--	--	--	--	--	--
	Negative Binomial GLMM						Negative Binomial GLMM					
	$\beta_1=0.01$		$\beta_1=0.41$		$\beta_1=0.92$		$\beta_1=0.01$		$\beta_1=0.41$		$\beta_1=0.92$	
	Type 1 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 1 (%)	Coverage (%)	Type 2 (%)	Coverage (%)	Type 2 (%)	Coverage (%)
1.0	3.55	98.82	0.00	100.00	0.00	97.48	3.55	98.82	0.00	100.00	0.00	97.48
1.2	4.02	95.98	0.00	94.44	0.00	90.86	3.00	97.50	0.00	96.00	0.00	92.46
1.3	5.00	95.50	0.00	90.95	0.00	88.38	3.50	97.00	0.00	97.50	0.00	96.00
1.4	4.50	95.50	0.00	90.95	0.00	86.00	1.00	99.50	0.00	97.00	0.00	97.50
1.5	5.00	94.00	0.00	92.00	0.00	86.00	1.50	99.00	0.00	94.50	0.00	98.50
2.0	9.00	91.50	0.00	90.50	0.00	71.00	1.50	99.50	0.00	99.00	0.00	99.50
2.5	9.50	91.00	0.00	86.50	0.00	67.00	1.00	99.50	3.50	98.50	0.00	99.50
5.0	19.00	83.00	0.50	81.50	0.00	72.50	0.00	100.00	75.00	99.50	0.00	100.00
10.0	18.50	83.00	3.50	86.50	0.00	81.00	--	--	--	--	--	--

σ_p is defined as the ratio of the Pearson χ^2 to its degrees of freedom



Figures 3a-b. Percentage of simulations with Type 1 errors and in which parameter coverage included the true parameter given a true parameter value of 0.01 in the longitudinal scenario, for a.) outlier-dependent overdispersion and b.) overdispersion caused by zero inflation.



B.

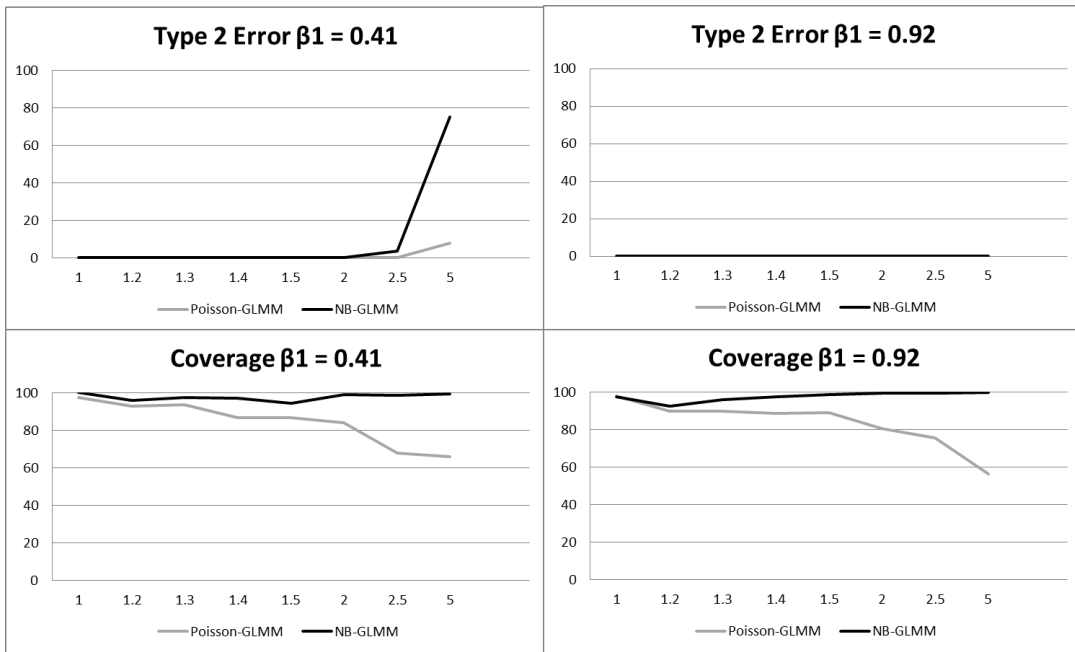


Figure 4a-b. Percentage of simulations with Type 2 errors and in which parameter coverage included the true parameter given true parameter values of 0.41 and 0.92 in the longitudinal scenario.

5. Motivating Real Datasets

5.1 Description

We utilize two real datasets to examine model performance at varying magnitudes of overdispersion. We modify the datasets in order to produce datasets with different levels of overdispersion. The National Lung Screening Trial (NLST) (Aberle et al. 2011) randomized 50,263 non-Hispanic white (NHW) and non-Hispanic black (NHB) patients to compare lung cancer mortality rates between those screened via low-dose CT screening and those given chest radiography. We consider the relationship between patient race predictor (NHB versus NHW) and comorbidity burden outcome, adjusted for assigned treatment group. The dispersion parameter σ_p for the whole cohort is 1.30. When we look into gender based subgroups, the dispersion parameter values for comorbidity burden are 1.25 and 1.36 for male and female patients, respectively. The second example is the classic Ames *Salmonella* dataset that is known for its

highly overdispersed count data (Mortelmans and Zeiger 2000). This classic overdispersed dataset includes a count outcome of bacterial colonies by six levels of medication dose on three different plates. The dispersion parameter σ_p for the whole cohort is 5.33. When we stratify the data by medication dose into low (less than or equal to 33 micrograms) and high (33 or more micrograms), we achieve a dispersion value of 1.99 for the low dose group and 4.18 for the high dose group. Here, we examine the relationship between bacterial colony count outcome and log medication dose predictor.

5.2 Results

All model results are given in Table 5, including rate ratios, AIC goodness-of-fit statistic, standard error of the beta parameters, and parameter p-values.

Table 5. Standard error and rate ratio by overdispersion magnitude for NLST and *Salmonella* datasets.

NLST						
Description	Model	ϕ_p	AIC	RR	SE	P-Value
<i>Male Patients</i>	Poisson	1.25	86772.4	1.127	0.025	<0.0001
	Poisson GLMM		86499.0	1.102	0.025	0.0001
	NB		86070.3	1.127	0.028	<0.0001
	NB GLMM		85861.2	1.200	0.020	0.0005
<i>Whole Cohort</i>	Poisson	1.30	151861.2	1.211	0.017	<0.0001
	Poisson GLMM		151314.1	1.190	0.018	<0.0001
	NB		150147.2	1.211	0.020	<0.0001
	NB GLMM		149741.7	1.190	0.021	<0.0001
<i>Female Patients</i>	Poisson	1.36	64958.8	1.296	0.024	<0.0001
	Poisson GLMM		64709.1	1.279	0.025	<0.0001
	NB		63952.0	1.296	0.029	<0.0001
	NB GLMM		63787.5	1.279	0.030	<0.0001
<i>Salmonella</i>						
Description	Model	ϕ_p	AIC	RR	SE	P-Value
<i>Low Dose</i>	Poisson	1.99	61.5	1.117	0.120	0.3577
	Poisson GLMM		60.1	1.117	0.120	0.3999
	NB		62.6	1.114	0.147	0.4629
	NB GLMM		62.1	1.117	0.120	0.4000
<i>High Dose</i>	Poisson	4.18	81.7	0.824	0.061	0.0014
	Poisson GLMM		69.5	0.824	0.061	0.0244
	NB		73.7	0.824	0.107	0.0708
	NB-GLMM		71.5	0.825	0.062	0.0271
<i>Whole Cohort</i>	Poisson	5.33	171.77	1.119	0.027	<0.0001
	Poisson GLMM		152.85	1.119	0.027	0.0009
	NB		140.43	1.134	0.057	0.0275
	NB GLMM		141.02	1.132	0.047	0.0194

We observe that the negative binomial regression model results in moderately adjusted standard error values and low AIC goodness-of-fit statistics in the NLST datasets, with respective dispersion magnitudes of 1.25, 1.30, and 1.36. The standard errors resulting from the NB model for these dispersion magnitudes are, respectively, 12.00%, 17.65%, and 20.83% higher than those resulting from the simple unadjusted Poisson model. The percent increase in standard error produced by the NB here clearly increases with the level of overdispersion in the dataset. The negative binomial generalized linear mixed models also perform well.

The results are similar among the higher magnitudes of overdispersion in the *Salmonella* datasets. The standard errors resulting from the NB model for dispersion magnitudes of 1.99, 4.18, and 5.33 are 22.50%, 75.41%, and 111.11% higher than those resulting from the simple unadjusted Poisson model, respectively. For the dataset with the highest magnitude of overdispersion, the NB-GLMM gives a more moderate increase of 74.07% compared to the unadjusted Poisson and may be preferable here. Again, the percent increase in standard error appears to correspond with the increase in overdispersion magnitude.

6. Conclusion

We compared Poisson and negative binomial methods via simulation study for analyzing cross-sectional and longitudinal datasets with two binary predictors and count outcome containing overdispersion due to either the addition of outliers or zero inflation. Magnitude of overdispersion was measured by dispersion parameter σ_p , defined as the ratio of the Pearson χ^2 value to its corresponding degrees of freedom $n - p$. Comparison among models was made using Type 1 error with a true β_1 value of 0.01, Type 2 errors using true β_1 values of 0.41 or 0.92, and coverage probability of β_1 for all effect sizes of β_1 .

Results of our analysis demonstrate that the unadjusted Poisson regression and Poisson-GLMM perform fairly well for cross-sectional scenarios when there is low overdispersion magnitude, particularly when $\sigma_p \leq 1.2$. The negative binomial regression model performs well at higher magnitudes of overdispersion under both outlier-dependent and zero-dependent

scenarios, up to $\sigma_p \leq 1.4$. The NB-GLMM gives acceptable results at high magnitudes of overdispersion in some cases up to $\sigma_p \leq 5.0$. Both the Poisson-GLMM and NB-GLMM resulted in more conservative Type 1 errors than their corresponding regression models. The Type 2 errors are higher for negative binomial regression and NB-GLMM compared to the unadjusted Poisson and Poisson-GLMM. The Type 2 error and coverage probability also decreased for higher β_1 effect sizes. NB-GLMM resulted in the highest Type 2 errors overall, so negative binomial regression appears to be sufficient to address the overdispersion in the cross-sectional datasets. Further statistical intervention would be required under the most extreme outlier-dependent overdispersion scenario when $\sigma_p \geq 10.0$, as our results demonstrate that none of our models give reliable results in these cases.

Longitudinal datasets appeared to be somewhat less tolerant of the more moderate levels of overdispersion. NB-GLMM gave more conservative Type 1 errors and higher coverage probabilities than Poisson-GLMM, as well as generally comparable Type 2 errors. Again, the Poisson-GLMM performs well in addressing both outlier-dependent and zero-dependent overdispersion when $\sigma_p \leq 1.2$. For larger magnitudes of overdispersion, up to about $\sigma_p \leq 2.5$, NB-GLMM performs well. The superiority of the NB-GLMM model became more apparent as the overdispersion in the dataset increased. Once again, further statistical intervention may be required when $\sigma_p \geq 5.0$ in longitudinal analysis. Our models addressing both outlier-dependent and zero-dependent overdispersion are less reliable in these cases. In a clinical setting, the covariates included in the model should be reexamined for errors leading to faulty models beyond the issue of overdispersion.

It would appear that a general threshold for relying on the simple Poisson model for cross-sectional and longitudinal datasets is in cases where $\sigma_p \leq 1.2$. For cross-sectional datasets, the negative binomial distribution via NB or NB-GLMM should be utilized if

$1.2 < \varphi_p \leq 1.5$. For higher values of σ_p in these scenarios, NB-GLMM should be utilized up to $\sigma_p \leq 5.0$. However, if $\sigma_p \geq 5.0$ for longitudinal datasets or if $\sigma_p \geq 10.0$ for cross-sectional datasets, the model may not be reliable based on adjustment for overdispersion and should be checked for additional modeling errors.

We also utilized two real cross-sectional datasets to produce varying magnitudes of overdispersion for analysis. We used data from the National Lung Screening Trial (NLST) [1] to examine the relationship between comorbidity count and patient race (NHB to NHW), adjusting for assigned treatment group. The σ_p value for the whole cohort was 1.30, and stratifying by gender gave dispersion values of 1.25 and 1.36 for male and female patients, respectively. According to our simulation results, these levels of σ_p would require statistical intervention via negative binomial regression or NB-GLMM. This was confirmed by decreased goodness-of-fit statistics and moderately adjusted standard errors compared to the unadjusted Poisson model. We also considered higher magnitudes of overdispersion using the Ames *Salmonella* dataset [20], which is a classic example of overdispersion in a dataset and includes measures of medication dose by plate and a count of *Salmonella* bacterial colonies. The σ_p values were 1.99 for observations with medication levels of 33 micrograms or lower, 4.18 for observations with medications of higher than 33 micrograms, and 5.33 for the whole cohort. Our results indicate that these high levels of overdispersion require adjustment via the NB or the NB-GLMM, which is also supported by our analysis. The percent increase in standard errors resulting from the negative binomial models compared to the unadjusted Poisson increased in correspondence with higher magnitudes of overdispersion.

We discussed a score test for overdispersion in Section 2.2 of $H_0 : \varphi = 0$ vs. $H_1 : \varphi > 0$ in which the score statistic has a standard normal distribution under the null hypothesis. This score test suggests that a dataset which results in a score statistic greater than or equal to 1.65 allows us to reject the assumption of equidispersion at a significance level less than or equal to

0.05. In our simulations, this translated into a level of overdispersion given by a value of σ_p at about $1.5 \leq \varphi_p \leq 2.0$ for both overdispersion scenarios dependent on effect size, as demonstrated by the nominal 95% rejection of equidispersion by the score test at these levels. It is clear from our simulations, however, that the presence of outlier-dependent overdispersion is harmful to our analyses and should be addressed at even lower values of σ_p , particularly at $\varphi_p > 1.2$, although the assumption of equidispersion may not be rejected at these levels by the score test.

Latent transition multiple imputation for missing data in time varying categorical covariates

CHAPTER 4

1. Introduction

Missing data in time varying categorical variables are frequently encountered in longitudinal biomedical studies. While there has been progress with missing data methods that deal with longitudinally measured continuous variables, there is still paucity of methods that deal with time varying categorical variables that have missing values. Recently, multiple imputation based on latent class (LCMI) has been proposed to deal with the problem of missing data in time invariant categorical covariates (Vermunt et al. 2008, Gebregziabher and DeSantis 2010). However, no extension has been made to address the problem of missing data in time varying categorical covariates.

Our motivating dataset is a retrospective, longitudinal cohort consisting of veterans with type 2 diabetes who were followed from 2002-2006 (Lynch et al. 2014). In this dataset, the outcome of interest is disease burden measured as a count of comorbidities based on those listed in the Elixhauser comorbidity index, which may range from 0 to 31. In this study two important covariates, medication non-adherence (MNA) and patient blood hemoglobin levels (A1C), which were measured longitudinally, were missing for a substantial number of patients. We use this motivating dataset to develop methodology for handling missing data in time varying categorical covariates.

Recent work demonstrated that multiple imputation based on latent class can be used to impute missing categorical covariates (Vermunt et al. 2008, Gebregziabher and DeSantis 2010). Such a latent class based method is relevant because missing categorical data are ubiquitous in biomedical research and there are no readily available principled methods for handling this problem (Schafer 1997b). Via an extensive simulation study, Gebregziabher and DeSantis (2010) showed that a latent class-based imputation approach provided unbiased parameter estimates in a highly stratified data model with ignorable and some non-ignorable missing data in time invariant categorical variables. Specifically, they showed that in a general random effects model framework with missing categorical variables, unbiased and efficient parameter estimates can be recovered utilizing latent class based multiple imputation. However, there are no studies that

jointly considered multiple imputation and latent transition analysis (LTA) to deal with missing data in time varying categorical covariates. The current paper seeks to extend LCMI to latent transition multiple imputation (LTMI) to impute missing categories of time varying covariates by their latent status.

In LTA, a hidden Markov model is assumed where at each time point, an unobserved time varying latent variable is inferred from a group of longitudinally observed items (time varying items). Parameter estimation for latent transition methods has been successfully utilized and explored (Chung, Lanza et al. 2008), as well as applied to longitudinal random effect models involving missing data (Albert and Follmann 2007, Xiaowei, Shoptaw et al. 2007, Lee, Lee et al. 2014). In LTA, the measurement model at each time point is a latent class model (Lazarsfeld and Henry 1968). All associations among categorical variables are explained by the underlying categorical latent variable. The result of fitting such a model is that for each individual, a latent trajectory that characterizes the missingness process is obtained. Conditional on the latent trajectory (latent transition or status), observations and items are independent; this is known as the conditional independence assumption. At each time point, incomplete categorical data can be imputed conditional on this latent status. In this paper, we will use LTA to estimate the LTMI model from completely observed covariates to implement multiple imputation of missing data in time varying categorical variables.

Complete case analysis (CCA) is a widely used ad-hoc method for dealing with missing covariate data, in which all subjects with incomplete longitudinal data are removed from the dataset prior to analysis. This method may involve a high loss of information. Multiple imputation methods are generally considered superior to CCA, as MI is highly efficient and often demonstrates decreased bias compared to CCA depending on the magnitude and cause of missingness (van der Heijden, T. Donders et al. 2006, Demissie, LaValley et al. 2003, White and Carlin 2010). Complete case analysis may be acceptable in situations where missingness is completely at random (Knol, Janssen et al. 2010) or independent of the outcome given covariates

(White and Carlin 2010). Our simulation study and motivating data example also include complete case analysis results as a general baseline for making comparison.

In finite mixture models, missing data are assigned to one of numerous distinct mixture components or classes, creating groups in the data called clusters. Missing data are generally assigned to a cluster based on maximum likelihood estimation, the most popular method of which is estimation-maximization (EM) (Leisch 2004). EM algorithms have been used to impute missing data in a wide variety of biomedical applications and missingness scenarios (Lipsitz et al. 1999, Stubbendick and Ibrahim 2003, Ibrahim, Chen et al. 1999, McLachlan 1997). Random effects pattern-mixture models have also been applied (Hedeker and Gibbons 1997). Because of the presence of clusters in mixture models, random effects may be assumed to come from not one but a finite mixture of normal distributions (Verbeke and Lesaffre 1996), leading to the use of the heterogeneity linear mixed model (Komarek et al. 2002). Mixture models using EM and Bayesian methods have been successfully used to model clustered longitudinal data (Heinzl and Tutz 2013, Goodman, Li et al. 2013, Wan and Chan 2009, Grunwald, Bruce et al. 2011) and extended to latent class mixture models (Beunckens, Molenberghs et al. 2008). Our simulation study and motivating data example also include LCMI and LTMI heterogeneity linear mixed model applications.

The paper is organized as follows. Section 2 reviews methods and provides discussion of LCMI. Section 3 introduces the LTMI method. Section 4 presents simulation results in terms of goodness of fit, bias and efficiency of LTMI versus CCA and LCMI methods. A description and results of analysis for the motivating dataset are given in Section 5. Section 6 includes a discussion of all results and future research plans in this area.

2. Methods

2.1. Data, Model and Notation

We consider a longitudinal generalized linear model setup to develop an analytic framework for the analysis of missing data in time varying categorical variables. Let Y_{ij} be a

response, while X_{1ij} (subject to missingness) and X_{2ij} (not subject to missingness) are covariates of interest at the j^{th} repeated measure for the i^{th} subject ($i = 1, \dots, n, j = 0, \dots, T_i$). Let q_i denote the random effects for each individual i which could be assumed to have a normal distribution with zero mean and covariance G . Let $\beta = (\beta_1, \beta_2)$ be the regression coefficients corresponding to X_{1ij} and X_{2ij} , respectively, and

$$\eta_{ij} = q_i + X_{1ij}\beta_1 + X_{2ij}\beta_2 \quad (1)$$

We can rewrite this in vector form as:

$$\eta_i = Z_i b_i + X_i \beta$$

where $X_i = (X_{1ij}, X_{2ij})'$, $\eta_i = g(E[Y_{ij} | q_i, \beta])$, g is a monotone link function, Z_i is the random effects design matrix and b_i is the random effects vector for each individual i . When data on all variables are observed, this model can be estimated in several different ways based on how one handles the estimation of the large number of nuisance parameters q_i which could be a source of loss of efficiency. Estimation of a model based on generalized estimating equations (GEE) could be used to make marginal inference on β . With an additional assumption on the distribution of q_i , maximum likelihood methods (eg. pseudo-likelihood, REML) could also be used estimating models yielding inference on β . Under the assumption that q_i are Gaussian, the integral in the specification of the log-likelihood (see Equation 2) could be approximated using Gaussian quadrature to approximate the integral by weighted sums (Breslow and Clayton 1993). However, if some components of X_i are not fully observed, methods used for complete data may lead to biased estimates in these likelihood methods. On the other hand, if the missing data mechanism is characterized as being missing completely at random (MCAR), valid inference

could be made using GEE. The same is true with the maximum likelihood methods (Breslow and Clayton 1993).

Suppose some components of X_i are not fully observed. Let R_{ij} be a missing indicator for covariate X_{ij} which takes values r_{ij} and comes from a distribution parameterized by γ .

Further define the joint density of R_{ij} , Y_{ij} and q_i as $h(Y_{ij}, R_{ij}, q_i)$. This can be factored either

using the selection model $h(Y_{ij}, R_{ij}, q_i) = h(Y_{ij} | q_i) \times h(q_i) \times h(R_{ij} | q_i)$, or the mixture pattern

model $h(Y_{ij}, R_{ij}, q_i) = h(Y_{ij} | R_{ij}, q_i) \times h(q_i) \times h(R_{ij})$. Under the selection model, the full

likelihood based on this joint density is given by,

$$L(\beta, \gamma, G) = \prod_{i=1}^n \prod_{j=1}^{T_i} \int h(y_{ij} | q_i, \beta, G) h(q_i | G) h(r_{ij} | \gamma, Y_{ij}) \partial q_i \quad (2)$$

In this paper, we use the multiple imputation paradigm to efficiently estimate parameters of the random effects data model given in Equation 1. We implement this by using pattern-mixture models which are more theoretically appealing (Rubin 1987, Allison 2001) than selection models for MI. Thus, we define the full likelihood based on pattern-mixture model as

$$L(\beta, \gamma, G) = \prod_{i=1}^n \prod_{j=1}^{T_i} \int h(y_{ij} | r_{ij}, q_i, \beta) h(q_i | G) h(r_{ij} | \gamma) \partial q_i \quad (3)$$

We will use latent transition analysis coupled with multiple imputation to achieve our objectives.

The details are given in Section 3.

2.2. Conditional AIC

The conditional Akaike information criterion (cAIC) has been adopted as a conditional deviation information criterion (Celeux, Forbes et al. 2006) and proposed to choose among mixed effects models when data are clustered, by accounting for shrinkage in the random effects via the effective degrees of freedom (Vaida 2005). The marginal AIC has been shown to be biased when estimating information for random effects models (Greven and Kneib 2010), suggesting that traditional information criterion measures may be inappropriate in these cases.

The conditional AIC has been further corrected and developed, increasing its usefulness (Vaida 2005, Liang, Wu et al. 2008, Greven and Kneib 2010). Further application has also been made to generalized linear mixed models (Donohue, Overholser et al. 2011). The unbiased estimator of the cAIC is given as follows:

$$cAIC = -2\log(g[y | \hat{\theta}(y), \hat{b}(y)]) + 2\phi_0(y)$$

Here, $\phi_0(y) = \sum_{i=1}^n \sum_{j=1}^{T_i} \frac{\partial \hat{y}_{ij}}{\partial y_{ij}}$ and we set $N = \sum_{i=1}^n \sum_{j=1}^{T_i} ij$. Note that y_i is the i^{th} component of y

such that $y_i = (y_{i1}, \dots, y_{i5})'$ and \hat{y}_i is the i^{th} component of the fitted vector $\hat{y} = Z\hat{b} + X\hat{\beta}$. The

partial derivatives $\frac{\partial \hat{y}_{ij}}{\partial y_{ij}}$, with $ij = 1, \dots, N$, can be approximated numerically by

$\{\hat{y}_{ij}(y + he_{ij}) - \hat{y}_{ij}(y)\} / h$ where h is a small number, e_{ij} is the $N \times 1$ vector, the i^{th} component

is equal to 1, and other components are equal to zero (Liang, Wu et al. 2008). In this study, we will compute the cAIC to compare goodness of fit among our Poisson models with random effects and equal N as above, setting $h = 0.0001$ as in (Liang, Wu et al. 2008).

2.3. Multiple Imputation

In multiple imputation (MI) (Rubin 1987), an imputation model is based on the conditional distribution of the missing responses on the observed responses and is used to draw and replace each missing value with a set of plausible values. Each of the complete data sets (after imputation) is then analyzed using a standard method, and the results are later combined to produce parameter estimates, standard errors, and confidence intervals which account for the uncertainty in the imputation.

Several different MI-based approaches have been proposed for a variety of clinical research applications (Rubin 1987, Engels 2003, Nevalainen, Kenward et al. 2009, Harel and Zhou 2007). The expectation-maximization (EM) algorithm finds the maximum likelihood estimate (MLE) of parameters via iterating the E- and M-steps until convergence (Dempster et al. 1977).

Loglinear multiple imputation (LLMI) imputes missing categorical data at time j using a saturated log-linear model based on observed and imputed data from all prior time points, $j = 1, \dots, j-1$ (Schafer 1997a). The saturated log-linear model can also be reformulated as a logistic regression model with interaction terms included (Agresti 2002).

In general, a suitable model is specified for the conditional distribution of the missing responses on the observed responses, $f(Y_{mis} | X_{obs}, \theta)$, where Y_{mis} indicates the missing portion of the data, X_{obs} denotes the observed portion and θ is the vector of parameters. This joint distribution could be expressed as follows with an additional term of the missing indicator, R_{ij} .

$$\begin{aligned} \Pr(R_{ij}, Z_{ij}, X_{ij}, Y_{ij}; \beta_1, \beta_2, \gamma) &= \Pr(Z_{ij}, X_{ij}, Y_{ij}; \beta_1, \beta_2) \times \Pr(R_{ij} | Z_{ij}, X_{ij}, Y_{ij}; \gamma) \\ &= \Pr(Z_{ij}, X_{ij}, Y_{ij}; \beta_1, \beta_2) \times \Pr(\beta_1 | Z_{ij}, Y_{ij}) \times \Pr(R_{ij} | Z_{ij}, X_{ij}, Y_{ij}; \gamma) \end{aligned} \quad (4)$$

The pattern of missingness, missingness mechanism, and whether the distributions for R_{ij} and X_{ij} involve common parameters will determine the method of choice for the first stage. For missing categorical data with monotone missing data pattern, a propensity score method or regression method based on parametric models could be used (e.g., logistic or discriminant analysis) (Schafer 1997a, Little and Rubin 2002). In a propensity score method, a sequence of logistic regression models are estimated with R_{ij} as an outcome and all past observed values as covariates to estimate $f(Y_{mis} | X_{obs}, \theta)$. After fitting this model, imputed values are drawn from the fitted model following three steps; see (Li, Mehrotra et al. 2006, Rubin 1987). In regression methods, a model of the following form is fitted with the observed data for a variable Y_{ij} with missing values.

$$g[E(Y_{ij} | X_{i1}, \dots, X_{ij-1})] = \gamma_0 + \gamma_1 X_{i1} + \dots + \gamma_{j-1} X_{ij-1} \quad (5)$$

After fitting model (5) using observed data from subjects who have not dropped out at the j^{th} visit, γ^* and σ^* are then drawn from the distribution of $\hat{\gamma}$ and $\hat{\sigma}$ to account for the

uncertainty in estimating γ and σ , where σ is the residual standard error of the model. Finally, predicted values based on the estimated model are used to impute the missing values. The predictive mean matching (PMM) method is a similar regression option for semi-parametric data. In this case, after the simulated regression model is run, a value is randomly chosen from among the observed values in which the predicted value is closest to the predicted value of the missing observation (Heitjan and Little 1991, Schenker and Taylor 1996).

2.4. Latent class multiple imputation (LCMI)

Latent class multiple imputation (LCMI) is a method developed by Gebregziabher and DeSantis (2010) for dealing with time invariant missing data in categorical covariates. They implemented LCMI by first fitting the latent class model to the observed data, $x_{i,obs}$, using *Proc LCA* Version 1.1.5. This SAS procedure estimates latent classes measured by categorical indicators when covariates are time invariant. Gebregziabher and DeSantis went on to sample from the posterior probability of time invariant latent class L_i for each individual i given the observed data, $P(L_i = 1 | Y_{i,obs} = y_{i,obs})$, and also sampled from the distribution of the missing data conditional on class, $P(Y_{i,mis} | L_i = 1)$. Finally, they used a full Bayesian MCMC within class posterior sampling approach to impute the missing categorical data, $y_{i,mis}$. Additional technical details and information regarding the implementation of LCMI can be found in (Gebregziabher and DeSantis 2010). A complete discussion of latent transition analysis follows in Section 3.1.

3. Latent transition multiple imputation

3.1. Latent transition model

Let $L_j = (L_{1j}, \dots, L_{Tj})$ represent class membership indicators at time $j = 1, \dots, T$ where observed $l_j = 1, \dots, L$. The vector $Y_j = (Y_{1j}, \dots, Y_{Mj})$ represents the M observed categorical variables where each variable may take on values $k = 1, \dots, C_m$ for every time point, $j = 1, \dots, T$.

The joint probability that the i^{th} individual exhibits the realization of item responses, y_{i1}, \dots, y_{iT} , and observed latent class membership, $l_j = (l_1, \dots, l_T)$ at time j is

$$p(Y_1 = y_{i1}, \dots, Y_T = y_{iT}, L_j = l_j) = \left[\delta_{l_1} \prod_{j=2}^T \tau_{l_j | l_{j-1}}^{(j)} \right] \times \left[\prod_{j=1}^T \prod_{m=1}^M \prod_{k=1}^{C_m} \rho_{mkj | l_j}^{I(y_{imj}=k)} \right] \quad (6)$$

where $\delta_{l_1} = P(L_1 = l_1)$, $\tau_{l_j | l_{j-1}}^{(j)} = P(L_j = l_j | L_{j-1} = l_{j-1})$, and $\rho_{mkj | l_j} = P(Y_{mj} = k | L_j = l_j)$.

This representation assumes that items are conditionally independent within each class of l_j for all time points, $j = 1, \dots, T$. The collection of δ and τ parameters represent latent class prevalence at various time points where the sequence, L_j , constitutes a first order Markov chain for $j = 2, \dots, T$. The latent class prevalence at time j where $j \geq 2$ is calculated as,

$$\delta_{l_j}^{(j)} = P(L_j = l_j) = \sum_{l_1=1}^L \dots \sum_{l_{j-1}=1}^L \delta_{l_1} \prod_{j=2}^T \tau_{l_j | l_{j-1}}^{(j)}$$

The likelihood contribution for the i^{th} individual across all time points becomes,

$$L(\theta; Y_1 = y_{i1}, \dots, Y_T = y_{iT}) = \sum_{l_1=1}^L \dots \sum_{l_{j-1}=1}^L \left[\delta_{l_1} \prod_{j=2}^T \tau_{l_j | l_{j-1}}^{(j)} \right] \times \left[\prod_{j=1}^T \prod_{m=1}^M \prod_{k=1}^{C_m} \rho_{mkj | l_j}^{I(y_{imj}=k)} \right] \quad (7)$$

A more detailed derivation of Equation 7 can be found in Appendix 3. The collection of free parameters $\theta = (\delta, \tau, \rho)$ can be estimated using maximum likelihood, i.e., solving the score equations. This can be accomplished using an EM algorithm (Dempster et al. 1977), which involves iterating between the posterior distribution of latent class conditional on the item responses (Equation 6) and the score equations (Chung, Lanza et al. 2008). For the purposes of multiple imputation, the following posterior probabilities of latent class membership are of interest:

$$\hat{\eta}_{i(l_1, \dots, l_T)} = \frac{\delta_{l_1} \tau_{l_j | l_{j-1}}^{(j)} \times \left[\prod_j \prod_m \prod_k \rho_{mkj | l_j}^{I(y_{imj}=k)} \right]}{\sum_{l_1} \dots \sum_{l_T} \delta_{l_1} \prod_j \tau_{l_j | l_{j-1}}^{(j)} \left[\prod_j \prod_m \prod_k \rho_{mkj | l_j}^{I(y_{imj}=k)} \right]} \quad (8)$$

where $\hat{\eta}_{i(l_1, \dots, l_T)} = p(\mathbf{L}_1 = l_1, \dots, \mathbf{L}_T = l_T \mid y_{i1}, \dots, y_{iT})$. LTA enables us to fit the dynamic imputers model to obtain latent status at each time point for the time varying missing covariates, while multiple imputation is used to impute the missing data. Thus, LTMI will improve upon standard multiple imputation techniques for missing categorical data by first clustering like observations in a time dependent manner based on the latent transition model, and then imputing based on latent status l_j at each time j . To accomplish this, the latent transition model may be expressed as a model for the observed data density, $p(y_{ij,obs}; \theta)$:

$$p(y_{ij,obs}; \theta) = \sum_{l_1=1}^L \dots \sum_{l_T=1}^L \left[\delta_{l_1} \prod_{j=2}^T \tau_{l_j | l_{j-1}}^{(j)} \right] \times \left[\prod_{j=1}^T \prod_{m=1}^M \prod_{k=1}^{C_m} p(y_{im1}, \dots, y_{imj} \mid \mathbf{L}_j = l_j)^{r_{imj}} \right] \quad (9)$$

Here, $r_{imj} = 0$ if the value of y_{imj} is missing and 1 otherwise. Note that r_{imj} represents a realization of the missing data indicator, R_{imj} , so only variables $m = 1, \dots, M$ at time points, $j = 1, \dots, T$ that do not have missing values contribute to the estimation of the model. This results in unbiased parameter estimation due to the assumption of conditional independence of variables given latent status assignment, and leads to a straightforward strategy for status-based multiple imputation. Once the latent transition model is estimated via the EM algorithm, one can easily obtain draws from the distribution of the missing data conditional on the observed data

$p(y_{ij,mis} \mid y_{ij,obs}; \theta)$ and with the conditional independence assumption we get,

$$p(y_{ij,mis} \mid y_{ij,obs}; \theta) = \sum_{l_1=1}^L \dots \sum_{l_T=1}^L \left[\delta_{l_1} \prod_{j=2}^T \tau_{l_j | l_{j-1}}^{(j)} \right] \times \frac{p(y_{imj,obs=k} \mid \mathbf{L}_{ij} = l_{ij})}{p(y_{imj,obs=k})} p(y_{imj,mis=k} \mid \mathbf{L}_{ij} = l_{ij}) \quad (10)$$

Since the first part of Equation 10 is the posterior probability of membership in class l_1, \dots, l_T at times $j = 1, \dots, T$ respectively, given the observed data, then the distribution of $y_{imj,mis} \mid y_{imj,obs}; \theta$ can be rewritten as

$$\hat{\eta}_{i(l_1, \dots, l_T)} \times p(y_{imj,mis} \mid \mathbf{L}_{ij} = l_{ij}) \quad (11)$$

which is equivalent to $\prod_j \prod_m \prod_k p(y_{imj,mis} | L_{ij} = 1_{ij})^{r_{imj}}$ where y_{imj} are the complete data and

r_{imj} is the missingness indicator. Recall that as only the observed data are used to fit the latent

class model, $p(y_{i,mis} | L_{ij} = 1_{ij})$ is equivalent to $\prod_{j=1}^T p(y_{ij} | L_{ij} = 1_{ij})^{1-r_{ij}}$, where y_{ij} are the

complete data and r_{ij} is the missing data indicator.

There is continuing debate on how to determine the number of classes when fitting a latent transition model (LTM). It has been recommended that specifying the number of latent statuses to be sufficiently large enables the LTM to capture the inherent characteristics of the data.

Specifically, it will enable it to pick up the univariate distribution, bivariate association and higher order interactions among the covariates (items) used to fit the imputation model (McLachlan and Peel 2000). Statistical measures such as cAIC alone may not necessarily lead to the best imputation model. Hence we check model bias and efficiency as well as goodness of fit measures to choose an optimal missing data method of analysis.

3.2. Imputation based on LTM

LTM is implemented following the example of Gebregziabher and DeSantis (2010). First, we estimated the latent class model to the observed data, $y_{ij,obs}$. We then sampled from the posterior probability of latent class given the observed data, $P(L_{ij} = 1_{ij} | Y_{ij,obs} = y_{ij,obs})$. We also sampled from the distribution of the missing data conditional on latent class, $P(y_{ij,mis} | L_{ij} = 1_{ij})$.

We finally used a within class posterior sampling approach to impute the missing data, $y_{ij,mis}$.

The latent transition imputers model was estimated using *Proc LTA* Version 1.1.5 (Lanza et al. 2007, Lanza et al. 2008). *Proc LTA* is a SAS procedure for latent transition analysis developed for SAS Version 9.2 for Windows. It is used when the latent variable and the items or covariates are all time varying. After fitting the imputers model dependent on observed time

varying covariates, we used the output posterior probabilities of an individual having a particular latent class at a particular time point to assign latent class status by individual and time point.

We imputed the missing categorical X_1 observations dependent on count outcome Y and latent class. An interaction between Y and latent class was also considered but did not improve results. We ran five imputations each with five iterations following a burn-in of 20 iterations. Following imputation, we used the conditional likelihood to estimate the parameters of the model in Equation 1, which we then used to make model comparison.

3.3. Latent Class Discovery via Heterogeneity Linear Mixed Model

Heterogeneity linear mixed models differ from the homogeneity model described in Equation 1 via the distributional assumptions of the random effects. In Equation 1, q_i denote the random effects for each individual i and are assumed to have a normal distribution with zero mean and covariance G . In a heterogeneity linear mixed model, the q_i random effects are distributed according to a mixture of g normal distributions with mean μ_j and covariance matrices G_j such that $q_i \sim \sum_{j=1}^g \pi_j N(\mu_j, G_j)$ where π_j are mixture weights and $\sum_{j=1}^g \pi_j = 1$. The number of mixture weights must be chosen and should be driven by the data. Further details can be found in (Komarek et al. 2002, Heinzl and Tutz 2013). Latent classes are then chosen to correspond with the random effects mixtures, and multiple imputation methods based on latent class are utilized to impute the missing data. A SAS HetMixed macro that can be used to produce latent class results for inputting into the LTMI-LMM and LCMI-LMM is given in (Komarek et al. 2002). An application of mixture models to latent transition analysis using a real substance abuse dataset is given in (Chung, Park et al. 2005). LTMI results using heterogeneity linear mixed models (LTMI-LMM) to assign latent class are provided for comparison with our LTMI method results using *Proc LTA* (LTMI-LTA).

3.4. Fitting the outcomes model in SAS

We first imputed the missing observations dependent on count outcome and assignment time varying latent class via PMM regression methods. We then utilized *Proc GLIMMIX* in SAS 9.3 to perform longitudinal Poisson regression analysis on our generalized linear mixed models using a log link and random intercept. We included both time varying and time invariant covariates in the model as required. We removed bounds from the covariance parameter estimates to ensure model convergence. We also used the Cholesky root when calculating the random-effects matrix in mixed model equations. This algorithm uses more computing power but provides greater numerical stability, and is particularly useful when the estimated variance of the random effects model is not positive definite. We further utilized the Newton-Raphson method with ridging non-linear optimization method to estimate non-linear parameters in our models. This is an ideal optimization method for small problems with computationally simple Hessian matrices. We output pseudo-likelihood goodness of fit statistics, parameter estimates, and predicted outcomes for analysis using *Proc MIANALYZE*.

4. Simulation study

4.1. Design

We generated time varying longitudinal cohort data with a Poisson count outcome with $m = 2$ categorical predictor variables. We created 300 datasets each containing 200 observations with data taken at five continuous time points $j = 1, 2, \dots, 5$. The data for each subject were generated according to the model, $\log(E(Y_{ijm} = y | X_{ijm})) = \alpha + \sum_{m=1}^2 \beta_m X_{ijm}$ where β is the collection of parameters (β_1, β_2) . We set the intercept $\alpha = 1.0$. In a given stratum, the outcome count Y for the i^{th} individual was generated using a mean given by $\exp(\alpha + \sum_{m=1}^2 \beta_m X_{ijm})$. We considered a time varying categorical exposure variable X_1 with $k = 2$ levels. Time varying covariate X_2 is a potential binary confounder of the relationship

between main exposure, X_1 , and the outcome variable, Y . We assigned parameter $\beta_1 = 0.41$ to yield a rate ratio of 1.5 and $\beta_2 = 0.69$ to yield a rate ratio of 2.0.

After generating complete data according to the above model, data sets with missing exposure X_1 were generated from the cohort with both 20% and 50% proportions of observations with missing data. Observations at the first two time points were left completely observed, with observations first set to missing at $j = 3$ and missingness assigned through $j = 5$. Once an observation was assigned missingness at a time point, the remaining time points of exposure X_1 were also set to missing to make the simulation more accurate to clinical data. We generated data to be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) in the sense of Little and Rubin (2002). Further specification of the missingness within MAR was based on the dependence of the probabilities of missing X_1 on either X_2 , Y , or both X_2 and Y in three different scenarios. Missingness within the MNAR setting was based on the dependence of the probabilities of missing X_1 on interaction between X_1 and Y . We make the assumption that the missingness model is logistic with all the variables as covariates,

$$\text{logit} \left[\text{pr}(M_j = 1) \mid X_{1j}, X_{2j}, Y \right] = \gamma_0 + \gamma_1 X_{1j} + \gamma_2 X_{2j} + \gamma_3 Y + \gamma_4 (Y \times X_{1j})$$

where $M = 1 - R$ is a binary indicator that takes a value of 1 if X_1 is missing and 0 if X_1 is observed. The intercept of the model γ_0 determines the overall proportion of missingness while the other γ parameters are the corresponding log odds ratios of missingness for each variable. For the 20% missing proportion MAR and MNAR data, we assigned γ values via the appropriate variables such that about 5% of the observations were assigned missingness at $j = 3$, 5% at $j = 4$, and 10% at $j = 5$. For the 50% MAR and MNAR proportions, we assigned these values

such that about 10% of the observations were assigned missingness at $j = 3$, 10% at $j = 4$, and 30% at $j = 5$.

4.2. Results

We made comparison among CCA, LCMI, and LTMI methods for imputing missing data in the simulated datasets via cAIC, asymptotic standard errors (ASE), estimated standard errors (ESE), and 95% confidence intervals (CI). Figures 1 and 2 respectively give the ASE and ESE for MCAR, MAR_{x_2} , MAR_y , $MAR_{x_2,y}$, and MNAR models with 20% and 50% proportions of missingness imputed via LCMI and LTMI methods. The values for the dataset with no missingness are included for comparison.

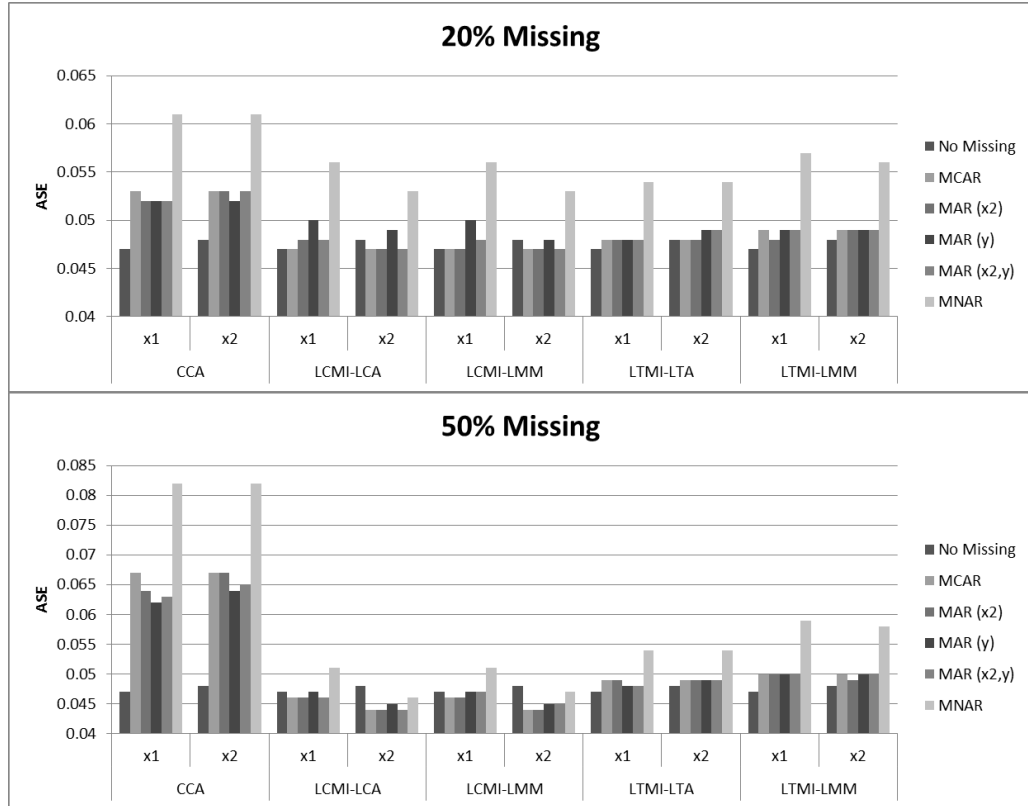


Figure 1. Comparison of asymptotic standard errors for all simulated models with count outcome Y . Results are stratified by percentage of missing data, method of dealing with missingness (CCA, LCMI, or LTMI methods), and type of missingness (MCAR, MAR, MNAR) for predictor X_1 including missingness and complete covariate X_2 .

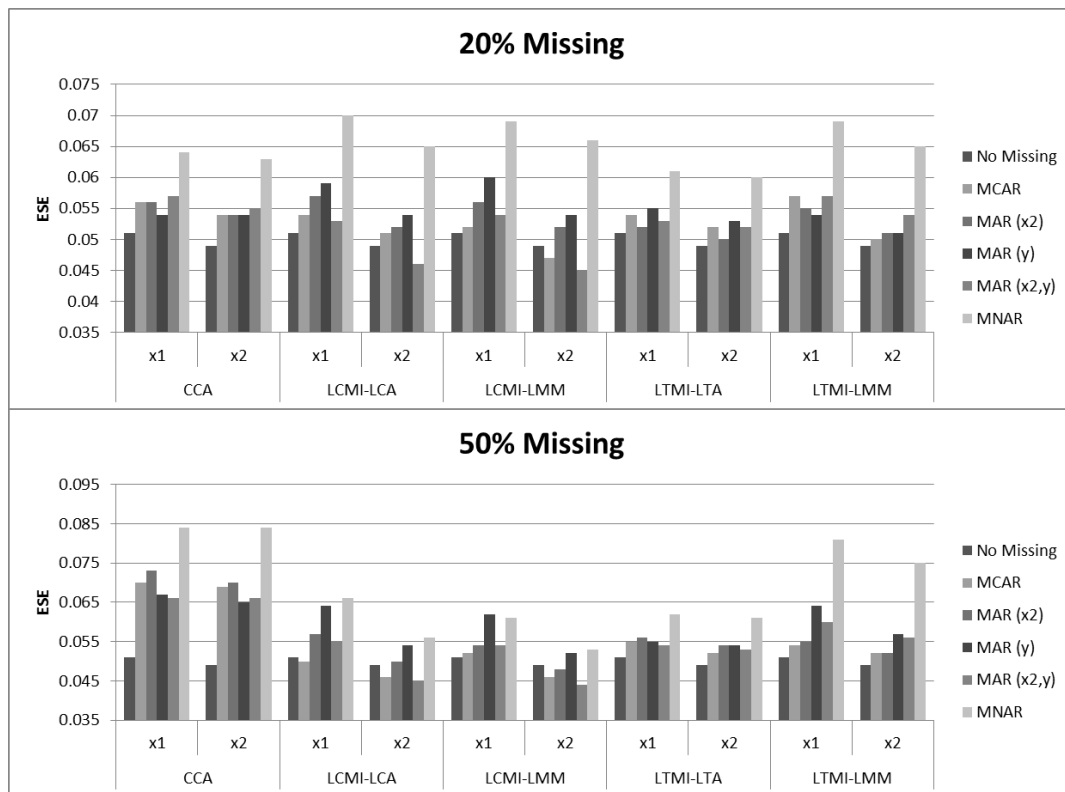


Figure 2. Comparison of estimated standard errors for all simulated models with count outcome Y . Results are stratified by percentage of missing data, method of dealing with missingness (CCA, LCMI, or LTMI methods), and type of missingness (MCAR, MAR, MNAR) for predictor X_1 including missingness and complete covariate X_2 .

Under the 50% missingness scenarios, LTMI-LTA gives moderately adjusted ASE and ESE values for both compared to CCA and LCMI methods. LTMI-LMM appears to inflate the magnitude of the standard errors, while LCMI-LCA and LCMI-LMM give comparable reduced standard error estimates. The contrast is particularly pronounced in the various 50% MAR and MNAR scenarios. Not surprisingly, CCA performs more adequately under the 20% MCAR scenario but gives much higher ASE and ESE in other scenarios. Under the 20% missingness scenarios, the SE are fairly comparable for all latent class methods by type of missingness. Goodness of fit may be assessed via Figure 3, which gives the cAIC for all imputation methods compared to the cAIC for the dataset excluding missingness. Comparable goodness of fit is achieved among the various imputation methods.



Figure 3. Comparison of conditional AIC for all imputation methods with count outcome Y . Results are stratified by percentage of missing data, method of dealing with missingness (CCA, LCMI, or LTMI methods), and type of missingness (MCAR, MAR, MNAR) for predictor X_1 including missingness and complete covariate X_2 .

A full table of results demonstrating the consistent performance of LTMI-LTA, including decreased bias compared to other methods, is given in Table 1. Similar tables for CCA, LCMI-LCA, LCMI-LMM, and LTMI-LMM are given in Appendix 4 Tables 1 - 4. The ratio of the Pearson χ^2 to its degrees of freedom are included in the tables and are equal to about one in all cases, showing no indication of overdispersion in the count outcome.

Table 1. Results of LTMI-LTA imputation for 20% and 50% missingness scenarios.

	No missing	20% Missing				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3229.46	3233.62	3232.55	3234.62	3233.84	3288.38
Pearson χ^2 /df	1.002	1.006	1.004	1.009	1.008	1.042
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.414	0.400	0.403	0.378	0.387	0.286
Mean RR	1.513	1.492	1.496	1.459	1.473	1.331
ASE	0.047	0.048	0.048	0.048	0.048	0.054
ESE	0.051	0.054	0.052	0.055	0.053	0.061
Bias	-0.004	0.010	0.007	0.032	0.023	0.124
Mean 95% CI for β_1	0.321, 0.506	0.305, 0.494	0.310, 0.497	0.280, 0.470	0.291, 0.481	0.181, 0.393
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.696	0.685	0.688	0.670	0.678	0.603
Mean RR	2.006	1.984	1.990	1.954	1.970	1.828
ASE	0.048	0.048	0.048	0.049	0.049	0.054
ESE	0.049	0.052	0.050	0.053	0.052	0.060
Bias	-0.006	0.005	0.002	0.020	0.012	0.087
Mean 95% CI for β_2	0.603, 0.790	0.590, 0.780	0.594, 0.782	0.572, 0.763	0.583, 0.775	0.500, 0.712
	No missing	50% Missing				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3229.46	3237.77	3236.63	3241.10	3238.99	3298.20
Pearson χ^2 /df	1.002	1.009	1.008	1.017	1.014	1.047
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.414	0.383	0.389	0.337	0.360	0.244
Mean RR	1.513	1.467	1.476	1.401	1.433	1.276
ASE	0.047	0.049	0.049	0.048	0.048	0.054
ESE	0.051	0.055	0.056	0.055	0.054	0.062
Bias	-0.004	0.027	0.021	0.073	0.050	0.166
Mean 95% CI for β_1	0.321, 0.506	0.286, 0.480	0.293, 0.484	0.242, 0.428	0.264, 0.453	0.116, 0.327
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.696	0.671	0.676	0.637	0.657	0.566
Mean RR	2.006	1.956	1.966	1.891	1.929	1.761
ASE	0.048	0.049	0.049	0.049	0.049	0.054
ESE	0.049	0.052	0.054	0.054	0.053	0.061
Bias	-0.006	0.019	0.014	0.053	0.033	0.124
Mean 95% CI for β_2	0.603, 0.790	0.574, 0.768	0.580, 0.771	0.543, 0.731	0.561, 0.722	0.440, 0.651

- ASE = the mean of the Asymptotic SE as computed by *Proc MEANS* (reported as mean of ASE)
- ESE = the SD of the estimates of beta as computed by *Proc MEANS* (reported SD Estimate)

5. Data Example

5.1. Description

We used a real data example to demonstrate the application of LCMI and LTMI methods, for dealing with missing categorical covariate data. The outcome is count data and the model used to study the association between the outcome and covariates is Poisson regression. The motivating dataset comes from a study designed to explore relationships between count of comorbidities with MNA and A1C adjusting for demographics such as geographic and racial/ethnic factors in veterans with type 2 diabetes. A total of 892,223 patients participated in this retrospective cohort study with yearly time points from 2002-2006, from which we randomly sampled 10,000 patients with complete outcomes and time invariant covariates. Two covariates of particular interest in this study are medication possession ratio (MPR), a measure of adherence to medication, and patient hemoglobin levels (A1C), a measure of blood sugar control in diabetic patients. The primary outcome was the patient's time varying comorbidity burden measured as an Elixhauser comorbidity count of up to 31 comorbidities. These include medical comorbidities such as cancer, cardiovascular disease, hypertension, and obesity, and mental comorbidities, including depression, psychosis, and substance abuse.

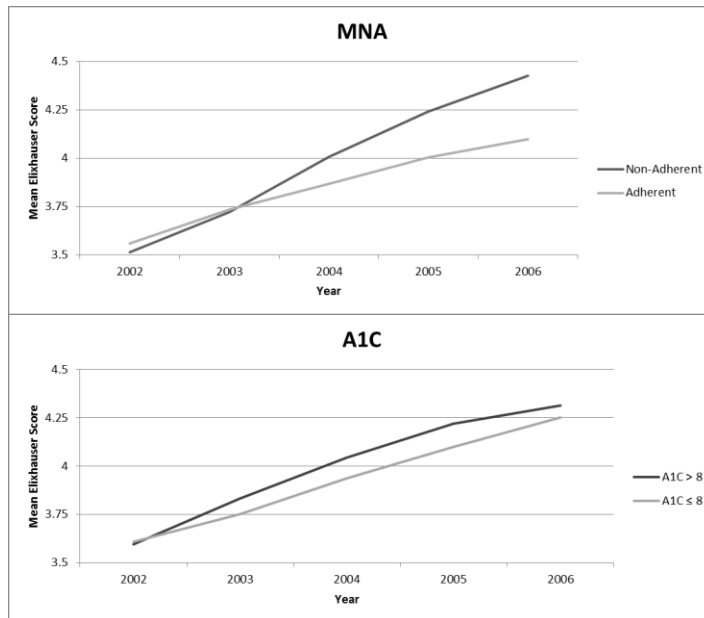


Figure 4. Mean Elixhauser score by MNA and A1C status over five year time period.

Line graphs modeling mean Elixhauser score over time by MNA and A1C status are given in Figure 4. An MPR of less than 0.80 demonstrates medication non-adherence (MNA), while an MPR of 0.80 or higher demonstrates adherence to medication. An A1C of 8.0 or lower suggests normal blood sugar control, while an A1C of greater than 8.0 suggests abnormally high blood sugar control (common in diabetic patients). In many cases, these values are missing for a given patient at one or more time points. Other covariates of interest include patient demographics such as age, gender, race, marital status, and urban or rural living. The data have been previously published in 2014 by Lynch et al. and more information about the study design can be found there.

Plots modeling the percentage of missing MNA and A1C values over time are given in Figure 5.

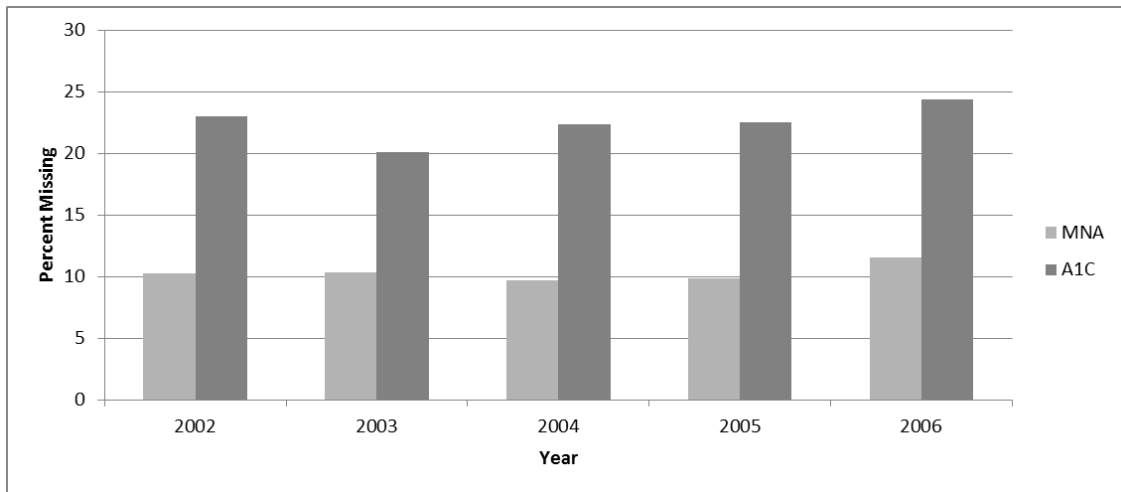


Figure 5. Percentage of missing MNA and A1C statuses over five year time period.

To understand the nature of the missing MNA and A1C values, logistic regression results examining the relationship between the dichotomized missing MNA and missing A1C values and demographic covariates are given in Table 2. There is association between missingness in the MNA and A1C variables and observed variables. As missingness is not expected to depend on the individual patient's MNA or A1C status, the missing mechanism is likely to be MAR.

Table 2. Missing MNA and A1C covariates by demographics.

<i>Covariate</i>	Missing MNA		Missing A1C	
	<i>OR</i>	<i>P-Value</i>	<i>OR</i>	<i>P-Value</i>
Time	1.049	<0.0001	1.057	<0.0001
Age	1.038	<0.0001	1.016	<0.0001
Comorbidity Count	0.880	<0.0001	0.852	<0.0001
<i>Region</i>				
South (reference)				
Northeast	1.818	<0.0001	1.532	<0.0001
Midatlantic	1.560	<0.0001	1.469	<0.0001
Midwest	1.421	<0.0001	2.128	<0.0001
West	1.231	<0.0001	0.975	0.5090
<i>Gender</i>				
Male (reference)				
Female	1.447	0.0004	1.048	0.5557
<i>Race</i>				
NHW (reference)				
NHB	1.481	<0.0001	0.853	<0.0001
Hispanic	10.546	<0.0001	2.116	<0.0001
Other	5.081	<0.0001	0.937	0.0612
<i>Living</i>				
Urban (reference)				
Rural	1.070	0.0505	1.157	<0.0001
<i>Marital Status</i>				
Married (reference)				
Unmarried	1.052	0.1531	0.927	0.0022
<i>Percent Service Connected Disability</i>				
<50% (reference)				
≥50%	0.834	0.0007	0.833	<0.0001

To examine the relationship between disease burden defined as count of patient Elixhauser comorbidities with both time invariant and time varying demographics of interest, we performed Poisson regression with a log link and random individual intercept. We ran CCA as well as LCMi-LCA, LCMi-LMM, LTMI-LTA, and LTMI-LMM via PMM regression imputation methods to make comparison among the methods of dealing with missing time varying categorical covariates. We utilized Proc GLIMMIX in SAS 9.3 to perform our Poisson regression analyses. Method performance is assessed and compared using goodness of fit statistics including cAIC, parameter standard errors, and 95% confidence intervals.

5.2. Results

We compared models via two-class LCMI and LTMI methods for model parsimony and given that our missing covariates are binary. Figure 6 gives the parameter standard error estimates for the missing data parameters A1C and MNA by method, while Figure 7 gives the SE for all parameters included in the models.

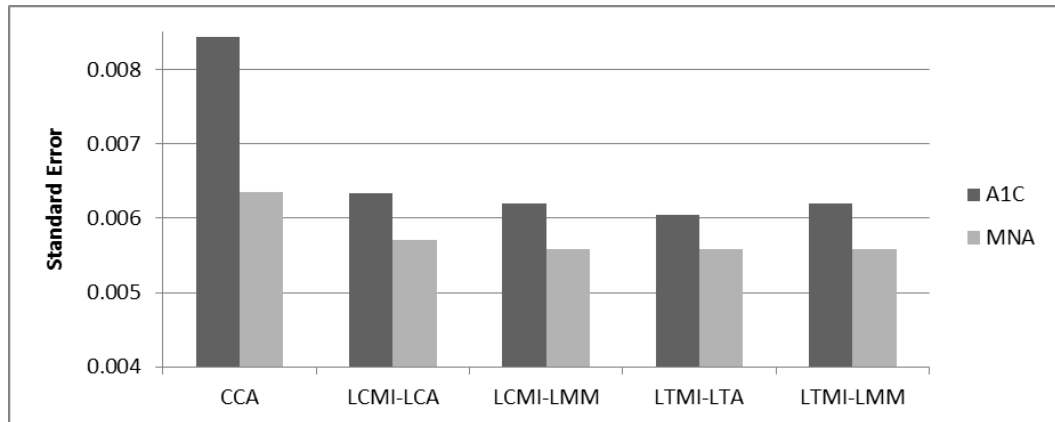


Figure 6. Comparison of parameter standard errors for A1C and MNA predictors with missing observations, by method of dealing with missing data (CCA, LCMI, or LTMI methods).

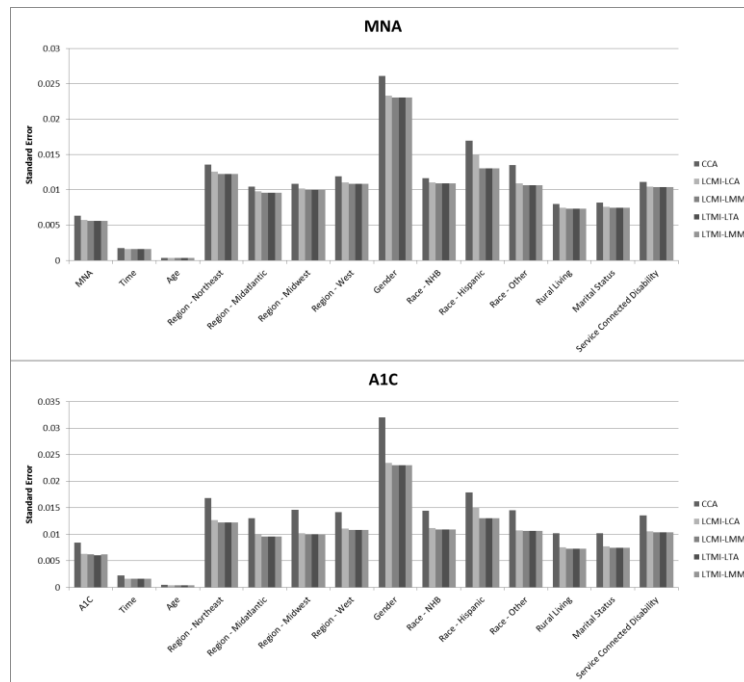


Figure 7. Comparison of parameter standard errors for all predictors in the model, by method of dealing with missing data (CCA, LCMI, or LTMI methods).

The SE estimates were lowest for LTMI-LTA, notably in the case of the covariates with missing data and particularly for the A1C covariate which had a higher percentage of missingness. LCMI-LTA and the LMM methods gave fairly comparable results in terms of parameter standard error. LTMI-LCA and CCA gave higher standard errors for nearly every parameter. The conditional AIC for all latent class models by method are given in Figure 8, and are generally comparable.

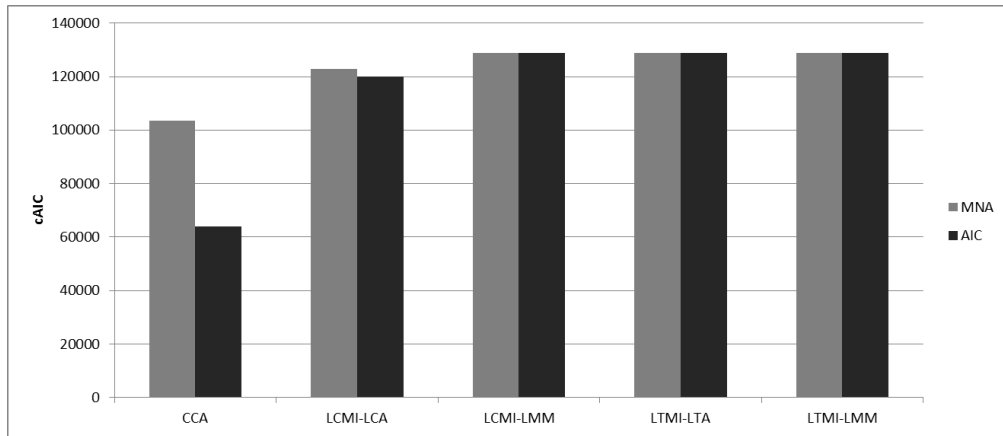


Figure 8. Comparison of conditional AIC goodness of fit values.

Table 3 gives LTMI-LTA results comparing patient Elixhauser score with MNA or A1C, both adjusted for demographics. There are significant differences in score based on medication adherence, A1C level, patient age, region, race, rural living, marital status, and percent of service connected disability. In the model containing MNA value, medically non-adherent patients have borderline higher scores than adherent patients (RR=1.009, p=0.0571), suggesting that patients with higher comorbidity burden may have poorer medication adherence. Patients from the West in the same model have the lowest comorbidity burden compared to patients from the South (RR=0.975, p=0.0101). Non-Hispanic black patients have the highest comorbidity burden among the race groups, compared to non-Hispanic white patients (RR=1.064, p<0.0001), while Hispanic and Other race groups have lower comorbidity burden (respectively, RR=0.964, p=0.0025; RR=0.935, p<0.0001). Unmarried patients have higher Elixhauser scores than married patients (RR=1.043, p<0.0001). Not surprisingly, patients with 50% or higher disability also have a higher comorbidity burden than patients with reduced disability (RR=1.079, p<0.0001). In the

model containing A1C value, patients with abnormally high blood sugar (A1C greater than 8.0) also have a statistically higher comorbidity burden than those with normal blood sugar (RR=1.022, p=0.0002). The other covariate parameters are nearly identical to those in the MNA model. Similar tables can be found for CCA, LCMI-LCA, LCMI-LMM, and LTMI-LMM methods in Appendix 4 Tables 5 – 8, and give comparable results.

Table 3. Relationship between Elixhauser score and covariates in Diabetes dataset via LTMI-LTA.

Covariate	MNA			A1C		
	RR	SE	P-Value	RR	SE	P-Value
<i>Covariate with Missingness</i>						
Medically adherent (reference)						--
Medically non-adherent	1.009	0.006	0.0571			
Normal blood sugar (reference)						--
Abnormally high blood sugar				1.022	0.006	0.0002
Time	1.044	0.002	<0.0001	1.044	0.002	<0.0001
Age	1.002	0.000	<0.0001	1.003	0.000	<0.0001
<i>Region</i>						
South (reference)						
Northeast	0.985	0.012	0.1121	0.986	0.012	0.1173
Midatlantic	1.002	0.010	0.4241	1.002	0.010	0.4290
Midwest	0.997	0.010	0.3770	0.997	0.010	0.3754
West	0.975	0.011	0.0101	0.975	0.011	0.0107
<i>Gender</i>						
Male (reference)						
Female	1.034	0.023	0.0790	1.035	0.023	0.0710
<i>Race</i>						
NHW (reference)						
NHB	1.064	0.011	<0.0001	1.063	0.011	<0.0001
Hispanic	0.964	0.013	0.0025	0.964	0.013	0.0023
Other	0.935	0.011	<0.0001	0.936	0.011	<0.0001
<i>Living</i>						
Urban (reference)						
Rural	1.001	0.007	0.4709	1.000	0.007	0.4757
<i>Marital Status</i>						
Married (reference)						
Unmarried	1.043	0.007	<0.0001	1.043	0.007	<0.0001
<i>Percent Service Connected Disability</i>						
<50% (reference)						
≥50%	1.079	0.010	<0.0001	1.079	0.010	<0.0001

6. Discussion

Complete case analysis is a common approach to missing data analysis, in which the missing data are ignored altogether and only subjects with data available at all time points are utilized. This method is only valid under MCAR. Even if this is a correct assumption, while parameter estimates are unbiased, such an analysis may suffer from loss of power. Another related approach is the all available case analysis approach based on GEE. If the data are missing at random, both complete case analysis and GEE methods yield results with moderate to large bias. However, GEE coupled with propensity score method (commonly called weighted GEE) provides valid estimates of the parameters when the missing mechanism is missing at random (MAR). Researchers also use single imputation techniques such as last (LVCF) or worst (WVCF) value carried forward. But these methods are commonly shown to lead to biased estimates and underestimated variance. The variance underestimation is typical in follow-up studies whose outcome is measured using scale scores (e.g., NIHSS) since worst values often tend to be similar, leading to less variable data set. Single imputation techniques also have an inherent problem of not accounting for uncertainty in the imputed value.

This study proposes a latent transition multiple imputation approach to deal with missing data in time varying categorical covariates. This study is the first to assess and implement LTMI for modeling time varying missing categorical covariate data. We have demonstrated that this method is statistically efficient and leads to unbiased estimates and can be implemented using standard software. In comparing simulated and real data scenarios, parameter standard errors were most efficient in the LTMI-LTA scenarios. In simulation studies, LTMI-LTA outperformed other methods most clearly in the 50% MAR and MNAR scenarios. CCA performed fairly well in the 20% MCAR scenario, and generally produced standard error results of greater magnitude otherwise. LCMI methods produced biased estimates and reduced standard error estimates in the simulations compared to the dataset with no missing data. LTMI-LMM also performed fairly well in simulation studies, though standard error estimates were higher and generally less consistent for this method than LTMI-LTA.

The standard error estimates in the real data analysis were lowest for LTMI-LTA, notably in the case of the A1C and MNA variables with missing data. Goodness of fit was measured and compared via conditional AIC, which is useful for choosing mixed effects models when dealing with clustered data, and results are comparable for all LCMI and LTMI scenarios in both real and simulated data scenarios. LTMI-LTA outperforms other methods for dealing with missing data in time varying categorical covariates, particularly in various non-MCAR scenarios with a higher percentage of missingness, jointly considering all criteria.

LTMI methods are appropriate for imputing time varying categorical variables. Multiple imputation methods account for uncertainty in the imputed categorical variable by utilizing time varying latent classes assigned via observed data. Both LTMI-LTA and LTMI-LMM results are more efficient and less biased than CCA or LCMI methods for imputing categorical missing data over time under some missingness scenarios, particularly under various MAR and some MNAR scenarios including a higher percentage of missingness. Additionally, LTMI methods are computationally inexpensive and the results are easily interpretable for clinicians.

Overdispersion was not an issue here, and therefore the methods we have previously studied for dealing with overdispersion were not applicable (Payne et al. 2015). However, future research will involve examining data scenarios with overdispersed count outcomes and time varying categorical covariates containing missingness. We will examine various methods for dealing with overdispersed data taking into account missingness in the time varying categorical predictor via real data and extensive simulation studies.

Dealing with overdispersion in longitudinal models including time varying categorical predictors with missing data

CHAPTER 5

1. Introduction

The restriction of Poisson regression that the response mean must be equal to the variance often fails in real data situations. Overdispersion occurs when data are more variable than is allowed under the Poisson model (Cox 1983), which may result from population heterogeneity, correlation, omission of important covariates in the model, the presence of high or zero outliers, among other reasons (Hardin and Hilbe 2007, Rigby, Stasinopoulos et al. 2008). An overdispersed model can result in underestimated parameter standard errors and falsely increased beta parameter significance, which may result in misleading inferences and conclusions (McCullagh and Nelder 1983, Breslow 1990, Hilbe 2007, Faddy and Smith 2011). We recently examined overdispersion occurring in real and simulated time invariant datasets resulting from omission of key predictors, high and zero outliers, and omission of necessary random effects (Payne et al. 2015). We compared six different scaling and modeling methods of analysis via goodness of fit and error statistics. The results showed that negative binomial regression and negative binomial generalized linear mixed models were preferred for dealing with overdispersion resulting from the sources we considered, while scaling methods and unadjusted Poisson regression were less reliable and often produced larger or smaller standard errors than expected. However, multiple options should be considered as the optimal method for data analysis may vary based on the source of the overdispersion. In this paper, we extend our comparison to longitudinal datasets which include categorical time varying predictor variables with missing observations.

Missing data is a common statistical issue in longitudinal biomedical studies. Multiple imputation based on latent class (LCMI) is a method previously proposed to deal with missing data in time invariant categorical variables (Vermunt et al. 2008, Gebregziabher and DeSantis 2010), which we have previously extended to address time varying categorical variables via latent transition multiple imputation (LTMI) under Poisson regression (Payne et al. 2016b). Parameter estimation for latent transition methods has been explored (Chung et al. 2008), and applied to longitudinal random effect models involving missing data (Albert and Follmann 2007, Xiaowei et

al. 2007, Lee et al. 2014). In summary, a latent class model is created at each time point (Lazarsfeld and Henry 1968) such that the underlying categorical latent variable explains all associations among categorical variables. A latent trajectory characterizing the missingness process for each individual is thereby obtained, and missing categorical covariates can then be imputed conditional on the assigned latent class status. For LTMI, we fitted the latent status imputers model using PROC LTA Version 1.1.5 (Lanza et al. 2007, Lanza et al. 2008), a SAS procedure for latent transition analysis developed for SAS Version 9.2 for Windows for utilization in scenarios where the latent variable and items of interest are time varying. We then imputed missing observations by sampling from the posterior distribution of the missing data model via predictive mean matching methods. We further considered latent status derived via heterogeneity linear mixed modeling. We then demonstrated the capability of LTMI performance compared to that of complete case analysis and LCMI methods through simulation studies and real data application, particularly in cases where data was missing at random (MAR).

It is not uncommon in real datasets to deal simultaneously with missing predictor data and overdispersion resulting from model specifications. However, there are presently no studies that address the co-occurrence of both time varying categorical covariate missingness and overdispersed count outcomes in models via a comprehensive evaluation. Our investigation therefore extends the approaches we examined previously to deal with overdispersion in Poisson-distributed count data to longitudinal Poisson analysis. We simultaneously address the issue of time varying categorical covariates with missing observations with complete case analysis and LTMI methods. We then make comparison among all of the models to determine superiority of method. We utilize simulation studies that consider outlier dependent overdispersion and make real data application while also addressing the co-occurrence of missingness in important categorical predictors. A study of related issues was recently performed by Zhang et al. (2015), in which researchers address the issue of overdispersion in non-parametric count outcomes with missing data in repeated measures scenarios. In this article, the

researchers extend the Mann–Whitney–Wilcoxon rank sum test to longitudinal data and address missingness via the inverse probability weighted method.

Our real data application comes from a retrospective cohort data example consisting of veterans with type 2 diabetes who were followed from 2002 – 2006. This dataset was previously published by Lynch et al. in 2014, and this study examined the association of various patient demographics with patient comorbidity burden to better understand health disparities among diabetic veterans. The outcome of interest is the patient Elixhauser score, a count of patient comorbidities which may range from 0 to 31. We choose a small subset of patients for analysis such that the count outcome in the dataset is overdispersed. Covariates include dichotomized values for medication non-adherence (MNA) and patient hemoglobin levels (A1C), which are time-varying and missing intermittently or monotonically for many patients.

This paper is organized in the following manner. Subsequent to the introduction, the statistical models and estimation are described in section 2. Section 3 provides information about the design and results of the simulation study. Section 4 details the real data application and results, and section 5 provides a discussion of all results as well as future research plans in this area.

2. Statistical models and estimation

2. 1. Overdispersion and missing covariate data in longitudinal analysis

As in our previous work, we utilize a generalized linear model setup (Payne et al. 2016b) for the analysis of simultaneous occurrence of missing data in time varying categorical variables and overdispersion resulting from model specifications. Let Y_{ij} be a time varying response containing overdispersion. Let X_{ij} be a time varying covariate subject to missingness and Z_{ij} be a time varying covariate not subject to missingness. Let t_{ij} be the time of the j^{th} repeated measure for the i^{th} subject ($i = 1, \dots, n, j = 0, \dots, T_i$) and q_i denote the random effects for each individual i , assumed to have a normal distribution with mean 0 and covariance G . The

regression coefficients corresponding to Z_{ij} , X_{ij} , and t_{ij} respectively are $\beta = (\beta_1, \beta_2, \beta_3)$.

So we have

$$\eta_{ij} = q_i + Z_{ij}\beta_1 + X_{ij}\beta_2 + t_{ij}\beta_3 \quad (1)$$

where $\eta_i = g(E[Y_{ij} | q_i, \beta])$ and g is a monotone link function.

Let a random longitudinal variable Y be distributed Poisson with variance function $Var(Y) = \mu$. If this variable is not equidispersed, a dispersion parameter φ may be utilized as a scale-adjustment to the variance function to account for changes in variability, via $Var(Y) = \varphi\mu$.

If $\varphi = 1$ then there is equidispersion and we can assume equal mean and variance in the Poisson model. If $\varphi < 1$ there is underdispersion in the model, and if $\varphi > 1$ there is overdispersion.

There are a variety of methods available for dealing with overdispersion in datasets, several of which we have utilized previously and compared (Payne et al. 2015). We considered both deviance and Pearson scale adjustment methods in addition to various Poisson and negative-binomial modeling methods. We previously showed that the negative-binomial distribution is effective in dealing with overdispersion resulting from a variety of causes. Here, $Y | \theta \sim Pois(\theta)$ and θ is a random variable such that $E(\theta) = \mu$ and $Var(\theta) = \sigma^2$. We can then say that $E(Y) = \mu$ and $Var(Y) = \sigma^2 + \mu$, such that the variance is greater than the mean.

Variable Y has a negative-binomial distribution when θ is assumed to be gamma; here,

$$E(Y) = \frac{k}{\lambda} = \mu \text{ and } Var(Y) = \mu + \frac{\mu^2}{k}.$$

Another option for picking up extra variability in overdispersed data is to include random effects in a generalized linear mixed model (GLMM). The GLMM family is given by

$$E(Y_i | X_i, R_i) = g^{-1}(X_i \beta + R_i \gamma_i) = \mu_i$$

for vectors of fixed effect (X_i) and random effect (R_i) explanatory variables ($i = 1, \dots, n$).

In this paper we jointly address the issues of missing time varying covariate data and overdispersion resulting from the presence of outliers in the count outcome. We consider GLMM with random intercept to account for individual variability with outcomes distributed as either Poisson or negative-binomial (Poisson-GLMM and NB-GLMM, respectively). We use a combination of latent transition analysis and multiple imputation methods to address the issue of missing data in our analysis prior to dealing with the overdispersion in our count outcome Y_{ij} in simulation study and real data application.

2.2. Latent Transition Multiple Imputation

We previously introduced the LTMI method for imputing missing time-varying categorical covariates and proved its capability via simulation and real data application (Payne et al. 2015). In this model, the joint probability that the i^{th} individual exhibits a categorical response at each time point $\mathbf{Y}_i = y_{i1}, \dots, y_{iT}$, as well as latent class membership at the corresponding time point $\mathbf{l}_i = (l_1, \dots, l_T)$, is given below:

$$p(\mathbf{Y}_i = y_{i1}, \dots, \mathbf{Y}_i = y_{iT}, \mathbf{l}_i = \mathbf{l}_i) = \left[\delta_{l_1} \prod_{j=2}^T \tau_{l_j | l_{j-1}}^{(i)} \right] \times \left[\prod_{j=1}^T \prod_{m=1}^M \prod_{k=1}^{C_m} \rho_{mkj | l_j}^{I(y_{imj}=k)} \right]$$

The likelihood contribution for the i^{th} individual to the whole model across all possible latent classes at each time point is therefore given via

$$L(\theta; \mathbf{Y}_i = y_{i1}, \dots, \mathbf{Y}_i = y_{iT}) = \sum_{l_1=1}^L \dots \sum_{l_{j-1}=1}^L \left[\delta_{l_1} \prod_{j=2}^T \tau_{l_j | l_{j-1}}^{(i)} \right] \times \left[\prod_{j=1}^T \prod_{m=1}^M \prod_{k=1}^{C_m} \rho_{mkj | l_j}^{I(y_{imj}=k)} \right]$$

The collection of free parameters $\theta = (\delta, \tau, \rho)$ can be estimated using maximum likelihood. In order to impute missing observations in our datasets via LTMI methods, we fit a two-class latent class model to the observed data $y_{i,obs}$, and sampled from the posterior probability of latent class given the observed data, $P(\mathbf{K}_i = \mathbf{k} | \mathbf{X}_{i,obs} = x_{i,obs})$. We also sampled from the distribution of the missing data conditional on class, $P(x_{i,mis} | \mathbf{K}_i = \mathbf{k})$. We imputed the missing data $x_{i,mis}$ using

predictive mean matching regression methods for monotone missing class data imputation by fully conditional specification methods.

2. 3. Estimation and Model Comparison

One of the most commonly used estimators of dispersion in the literature is the ratio of the Pearson χ^2 statistic to its corresponding degrees of freedom, typically $n - p$ for a study with sample size n observations and p parameters. Dispersion parameter σ_p is thus defined as:

$$\sigma_p = \frac{\chi^2}{n - p}$$

This ratio will be equal to one when data is equidispersed. When these ratios are greater than one then the data are considered overdispersed, with higher values signaling a greater magnitude of overdispersion. We recently determined that a general threshold for relying on the simple Poisson model for cross-sectional and longitudinal datasets is in cases where $\sigma_p \leq 1.2$.

Negative binomial models should be utilized if $1.2 < \sigma_p \leq 5.0$. If $\sigma_p > 5.0$ for longitudinal datasets or if $\sigma_p \geq 10.0$ for cross-sectional datasets, the model will likely be unreliable (Payne et al. 2016a).

We also computed the conditional AIC (cAIC) goodness of fit statistic for model comparison, which has been proposed to choose among mixed effects models when data is clustered by using the effective degrees of freedom to account for shrinkage in the random effects (Vaida 2005). Traditional information criterion may be inappropriate in these cases, as the marginal AIC has been shown to be biased when estimating information for random effects (Greven and Kneib 2010). The usefulness of the conditional AIC has increased as it has been corrected and developed (Vaida 2005, Liang, Wu et al. 2008, Greven and Kneib 2010) and applied to generalized linear mixed models (Donohue, Overholser et al. 2011).

Mean asymptotic and estimated parameter standard errors, mean bias, and the mean 95% confidence intervals for each parameter were also recorded to determine the predictive

ability of the models compared to the assumed value in the simulation study. These values were compared across the models to determine which method for dealing with both outlier dependent overdispersion and missing categorical time varying predictor data jointly resulted in the lowest cAIC values as well as offered moderately adjusted standard errors and 95% confidence intervals and low bias.

2. 4. Fitting the SAS model

After imputing missing observations via LTMI, we performed generalized linear mixed model analysis using Poisson or negative-binomial distributions, a log link, and a random intercept to account for individual subject variability. We removed bounds from the covariance parameter estimates in order to ensure model convergence and used the Cholesky root when calculating the random-effects matrix in mixed model equations. We estimated non-linear parameters in our models using the Newton-Raphson method with ridging non-linear optimization. We then output results for analysis including pseudo-likelihood goodness of fit statistics, parameter estimates, and predicted outcomes. All analysis was performed using SAS 9.4, particularly the Proc GLIMMIX package.

3. Simulation

3. 1. Design

We simulated 300 longitudinal datasets each with a sample size of $n = 200$ random observations to include a time varying longitudinal Poisson count outcome and two categorical predictor variables. Data was generated at five continuous time points $j = 1, 2, \dots, 5$ according to

the model $\log(E(Y_{ijm} = y | X_{ijm})) = \alpha + \sum_{m=1}^2 \beta_m X_{ijm}$ where β is the collection of parameters

(β_1, β_2) and intercept $\alpha = 1.0$. We considered time varying binary exposure variables X_1 and X_2 . We assigned parameter $\beta_1 = 0.41$ to yield an rate ratio of 1.5 and $\beta_2 = 0.69$ to yield a

rate ratio of 2.0 . Outcome count Y for the i^{th} individual was determined by

$$\exp(\alpha + \sum_{m=1}^2 \beta_m X_{ijm}).$$

We then created overdispersion relative to the Poisson via the addition of outliers to the count outcome Y . Y values greater than 10 in each simulation were chosen at random and increased by 20 to create outlier dependent overdispersion in the data; i.e. about 1% of the data were replaced by high outliers. Following the addition of overdispersion, missing time varying categorical covariate data was added via a variety of missingness scenarios as described below.

3. 2. Missingness scenarios

After generating overdispersed datasets for this scenario, datasets with missing exposure X_1 were generated from the cohort with a 20% and 50% proportion of missingness. We considered various missingness scenarios including data missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) according to a logistic missingness model (Little and Rubin 2002). Observations at $t = 1, 2$ were left completely observed, while missingness was assigned at $t = 3, 4, 5$. Once an observation was assigned missingness at a time point, the remaining time points of exposure X_1 were also set to missing to create monotone missingness. Missingness within MAR was given by three different scenarios, based respectively on the dependence of the probabilities of missing X_1 on X_2 , Y , or both X_2 and Y .

Analysis was then made using Poisson and negative binomial GLMM for all simulated data from each scenario using both complete case analysis and LTMI methods to address missing observations in predictor X_1 . Conditional AIC, dispersion parameter, parameter estimates for the regression coefficients corresponding to each covariate with their corresponding asymptotic and estimated standard errors, bias, and 95% confidence interval (CI) coverage were calculated for comparison among methods.

3. 3. Results

After the addition of outliers to the count outcome, the mean σ_p for the Poisson-GLMM model with no missing data was 1.44 ± 0.16 . Thus we can conclude that overdispersion is present in the full dataset prior to the addition of missingness in the categorical predictor. We then added 50% missing data to variable X_1 via various missingness mechanisms and either performed CCA or used LTMI methods to impute the missing categorical covariate data as described above. The CCA and LTMI results of our analysis are given in Tables 1 and 2, respectively.

According to the high σ_p dispersion statistics, CCA and LTMI Poisson-GLMM are both overdispersed under complete data and the various missingness scenarios. When LTMI methods were utilized to impute missing X_1 values in all scenarios, the result was comparable ASE and ESE values and conditional AIC goodness of fit statistics compared to the models with no missing data. Parameter estimates were again closest to those in the scenarios without missingness in MCAR and MAR cases.

NB-GLMM resulted in moderately adjusted ASE and ESE compared to the Poisson-GLMM in the 50% missing data scenarios. Overdispersion in both cases was effectively addressed via the NB-GLMM method under all missingness scenarios, while Poisson-GLMM did not address the overdispersion. The conditional AIC also demonstrates the superior goodness of fit of NB-GLMM. The NB-GLMM often resulted in comparable parameter and standard error estimates compared to those in the scenario without missingness. CCA and LTMI results for the 20% missing data scenarios are given in Appendix 5 Tables 1 and 2 and give similar results. In general NB-GLMM outperformed Poisson-GLMM in addressing outlier dependent overdispersion after utilizing LTMI methods, jointly considering all criteria.

Table 1. Results for high outlier scenario via CCA.

	No missing or overdispersion	Poisson-GLMM				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3229.34	1939.41	1967.71	1937.85	1961.03	1853.13
σ_p	1.000	1.445	1.488	1.406	1.437	1.370
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.409	0.495	0.508	0.455	0.484	0.345
Mean RR	1.505	1.640	1.662	1.576	1.623	1.412
ASE	0.047	0.068	0.065	0.066	0.065	0.083
ESE	0.048	0.120	0.124	0.117	0.125	0.146
Bias	0.001	0.085	0.098	0.045	0.074	0.065
Mean 95% CI for β_1	0.317, 0.501	0.361, 0.630	0.380, 0.637	0.328, 0.582	0.356, 0.611	0.181, 0.509
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.690	0.814	0.821	0.726	0.703	0.786
Mean RR	1.994	2.257	2.273	2.067	2.020	2.195
ASE	0.048	0.069	0.068	0.066	0.067	0.084
ESE	0.047	0.125	0.120	0.115	0.121	0.143
Bias	0.000	0.124	0.131	0.036	0.013	0.096
Mean 95% CI for β_2	0.597, 0.783	0.678, 0.950	0.687, 0.955	0.596, 0.856	0.572, 0.835	0.621, 0.951
	No missing or overdispersion	NB-GLMM				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3228.55	1953.85	1978.15	1967.89	1985.46	1877.07
σ_p	1.000	0.974	0.974	0.974	0.976	0.968
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.409	0.486	0.500	0.446	0.473	0.333
Mean RR	1.505	1.626	1.649	1.562	1.605	1.395
ASE	0.047	0.095	0.093	0.089	0.091	0.108
ESE	0.048	0.108	0.113	0.106	0.113	0.131
Bias	0.001	0.076	0.090	0.036	0.063	0.077
Mean 95% CI for β_1	0.317, 0.501	0.300, 0.673	0.317, 0.683	0.272, 0.621	0.294, 0.652	0.121, 0.545
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.690	0.807	0.812	0.718	0.692	0.777
Mean RR	1.994	2.241	2.252	2.050	1.998	2.175
ASE	0.047	0.095	0.095	0.090	0.092	0.107
ESE	0.047	0.113	0.109	0.105	0.111	0.129
Bias	0.000	0.117	0.122	0.028	0.002	0.087
Mean 95% CI for β_2	0.597, 0.783	0.619, 0.994	0.625, 0.999	0.541, 0.895	0.510, 0.874	0.566, 0.988

Table 2. Results for high outlier scenario via LTMI.

	No missing or overdispersion	Poisson-GLMM				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3229.34	3886.05	3875.84	3877.86	3874.94	3974.62
σ_p	1.000	1.471	1.462	1.471	1.467	1.536
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.409	0.440	0.460	0.425	0.447	0.277
Mean RR	1.505	1.553	1.584	1.530	1.564	1.319
ASE	0.047	0.051	0.050	0.049	0.049	0.055
ESE	0.048	0.090	0.094	0.089	0.090	0.095
Bias	0.001	0.030	0.050	0.015	0.037	0.133
Mean 95% CI for β_1	0.317, 0.501	0.344, 0.544	0.361, 0.558	0.361, 0.558	0.372, 0.570	0.169, 0.386
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.690	0.767	0.782	0.757	0.777	0.645
Mean RR	1.994	2.153	2.186	2.132	2.175	1.906
ASE	0.048	0.051	0.050	0.049	0.050	0.055
ESE	0.047	0.087	0.091	0.087	0.088	0.094
Bias	0.000	0.077	0.092	0.067	0.087	0.045
Mean 95% CI for β_2	0.597, 0.783	0.669, 0.868	0.682, 0.879	0.680, 0.876	0.694, 0.893	0.537, 0.754
	No missing or overdispersion	NB-GLMM				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3228.55	3909.45	3901.58	3899.76	3897.71	3990.38
σ_p	1.000	0.971	0.972	0.970	0.971	0.977
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.409	0.428	0.444	0.412	0.430	0.264
Mean RR	1.505	1.534	1.559	1.510	1.537	1.302
ASE	0.047	0.069	0.068	0.068	0.069	0.076
ESE	0.048	0.081	0.084	0.079	0.081	0.085
Bias	0.001	0.018	0.034	0.002	0.020	0.146
Mean 95% CI for β_1	0.317, 0.501	0.297, 0.568	0.310, 0.577	0.315, 0.592	0.314, 0.587	0.115, 0.413
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.690	0.757	0.767	0.746	0.762	0.633
Mean RR	1.994	2.132	2.153	2.109	2.143	1.883
ASE	0.047	0.069	0.068	0.069	0.070	0.076
ESE	0.047	0.080	0.082	0.079	0.080	0.085
Bias	0.000	0.067	0.077	0.056	0.072	0.057
Mean 95% CI for β_2	0.597, 0.783	0.623, 0.892	0.632, 0.899	0.632, 0.908	0.637, 0.910	0.485, 0.782

Figures 1 and 2 respectively give the ASE and ESE for MCAR, MAR_{x2} , MAR_y , $MAR_{x2,y}$, and MNAR models with 20% and 50% proportions of missingness analyzed via CCA and LTMI methods.

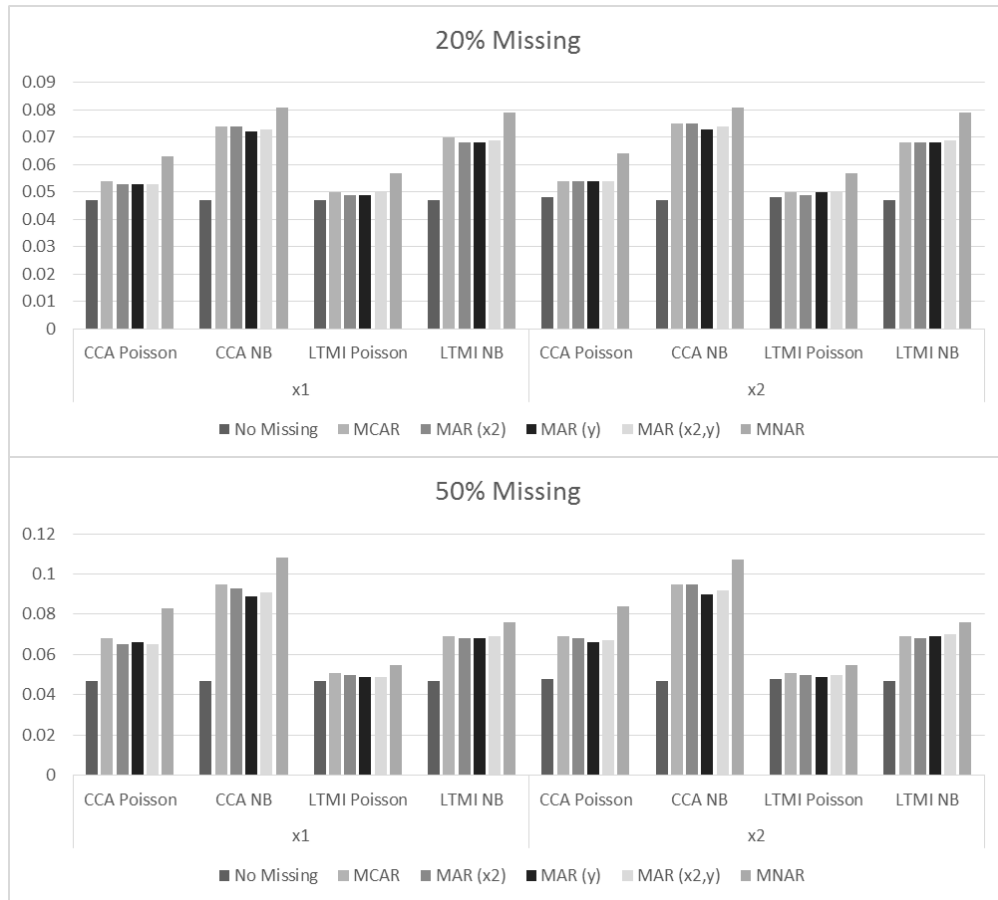


Figure 1. ASE for CCA and LTMI results.

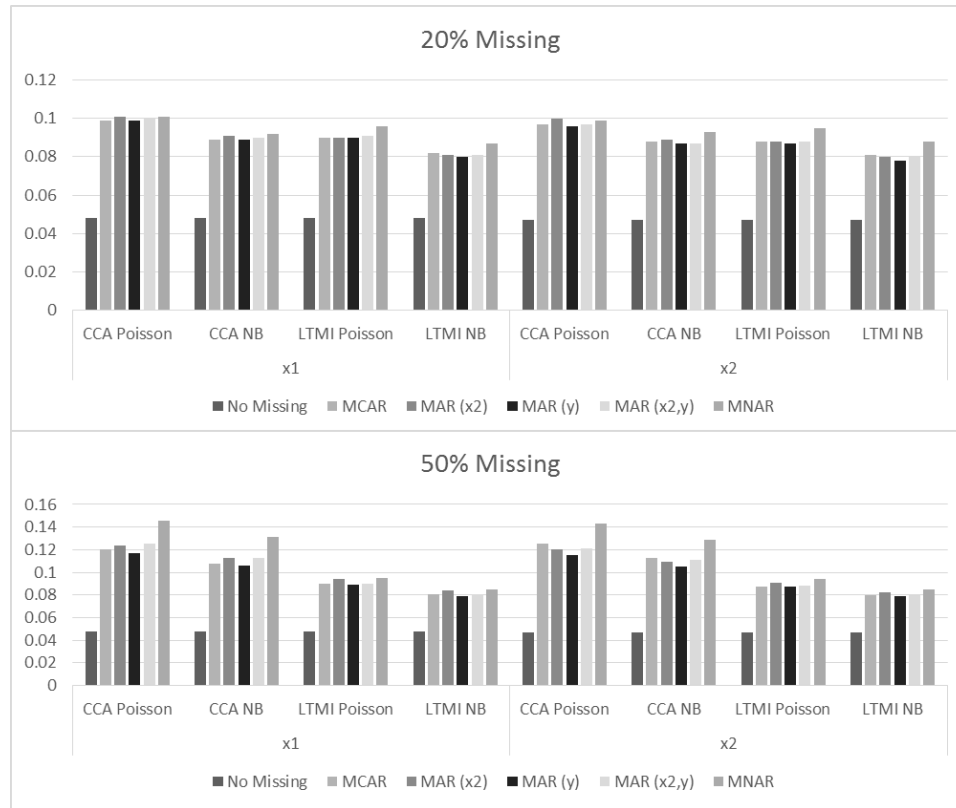


Figure 2. ESE for CCA and LTMI results.

The values for the dataset with no missingness are also included for comparison. LTMI consistently gives reduced ASE and ESE values compared to CCA methods. The contrast is more pronounced in the various 50% MAR and MNAR scenarios. Furthermore, the overdispersion appears to be adequately addressed via the negative binomial regression, with negative binomial providing adjusted errors compared to the Poisson. Figure 3 illustrates the bias for both predictors compared to the true beta values.

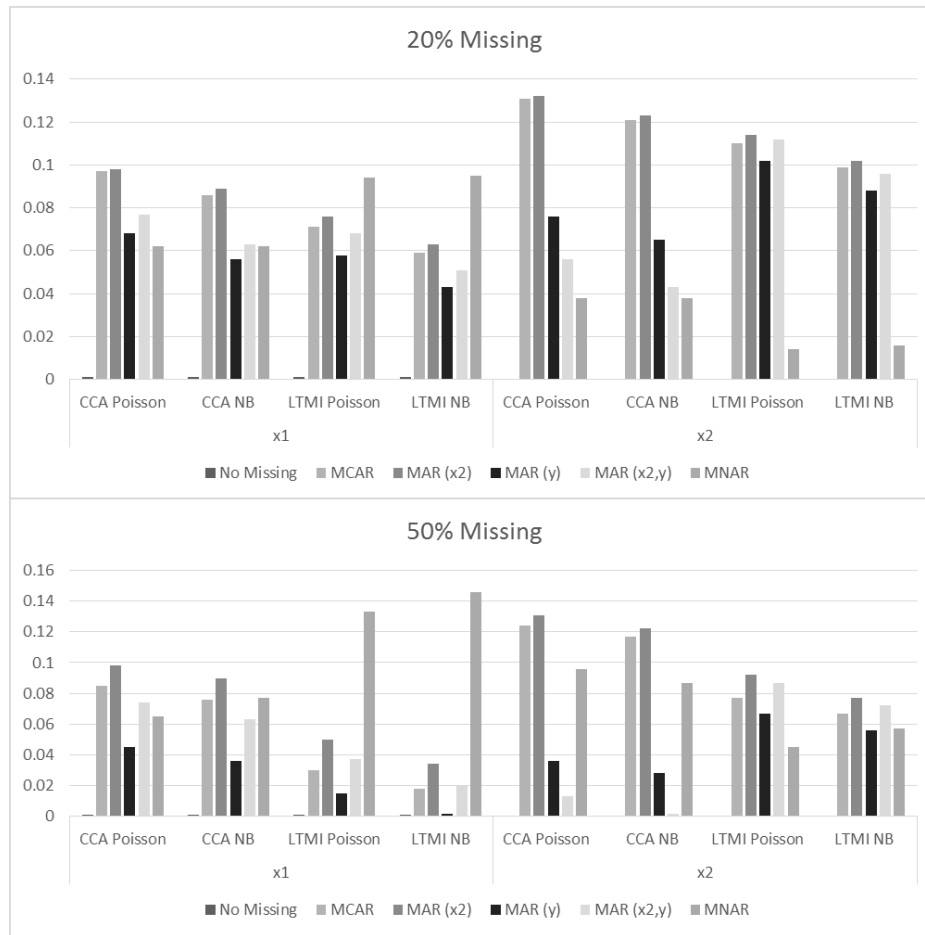


Figure 3. Bias for CCA and LTMI results.

Bias is generally higher for the CCA methods than for LTMI, particularly for variable X_2 in the MAR, MNAR, and the 50% missingness scenarios. The negative binomial also outperforms the Poisson in many cases, particularly among the MAR data and the 50% missingness scenarios. Figure 4 gives the cAIC for all imputation methods compared to the cAIC for the dataset excluding missingness, demonstrating comparable goodness of fit between LTMI NB and the CCA methods. CCA performs most adequately under the MCAR scenarios, jointly considering all criteria.

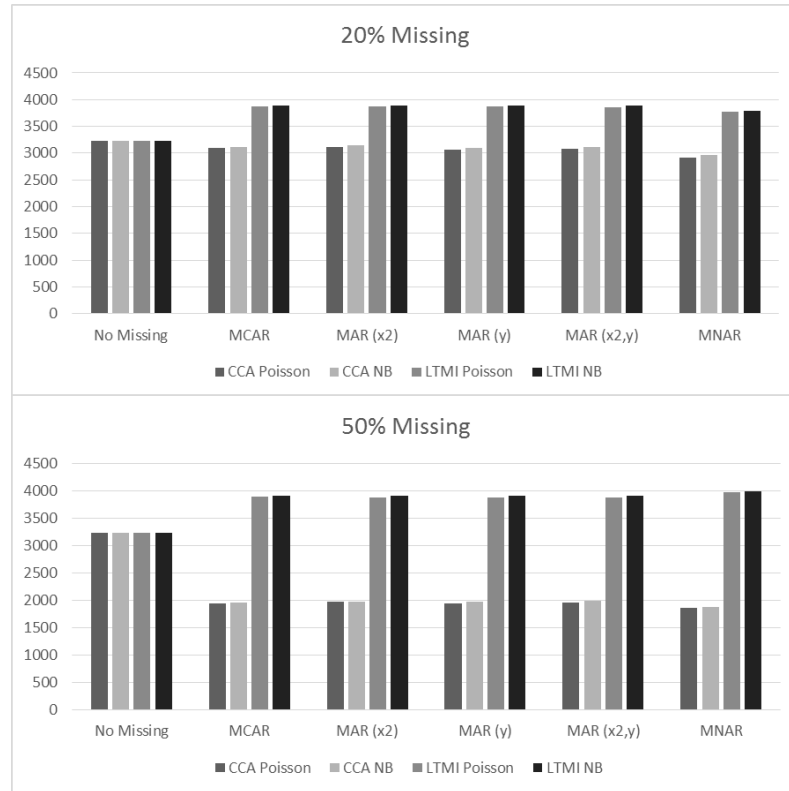


Figure 4. Conditional AIC for CCA and LTMi results.

4. Data Example

4.1. Description

We also used a real data example to compare methods of dealing with overdispersion assuming co-occurrence of missing categorical predictor data. The motivating dataset comes from a study designed to explore relationships between demographics such as geographic and racial/ethnic factors and patient multimorbidity in veterans with type 2 diabetes (Lynch et al. 2014). Our primary outcome is the patient's time varying Elixhauser comorbidity count of up to 31 comorbidities, including both mental and medical comorbidities. A total of 892,223 patients participated in this retrospective cohort study with yearly time points from 2002-2006, from which we took a sample of 40 non-Hispanic white patients, aged 65 or older, with complete outcomes and time invariant covariates as well as at least one Elixhauser score of 13 or more to ensure the presence of overdispersion in the dataset. Our two time varying predictors of interest containing

missingness are patient monthly prescribing reference (MPR), a measure of adherence to medication, and patient hemoglobin levels (A1C), a measure of blood sugar control. An MPR of less than 0.80 demonstrates medication non-adherence (MNA) compared to higher values, and an A1C higher than 8.0 suggests abnormally high blood sugar control. Figure 5 gives line graphs modeling mean Elixhauser score over time by MNA and A1C status.

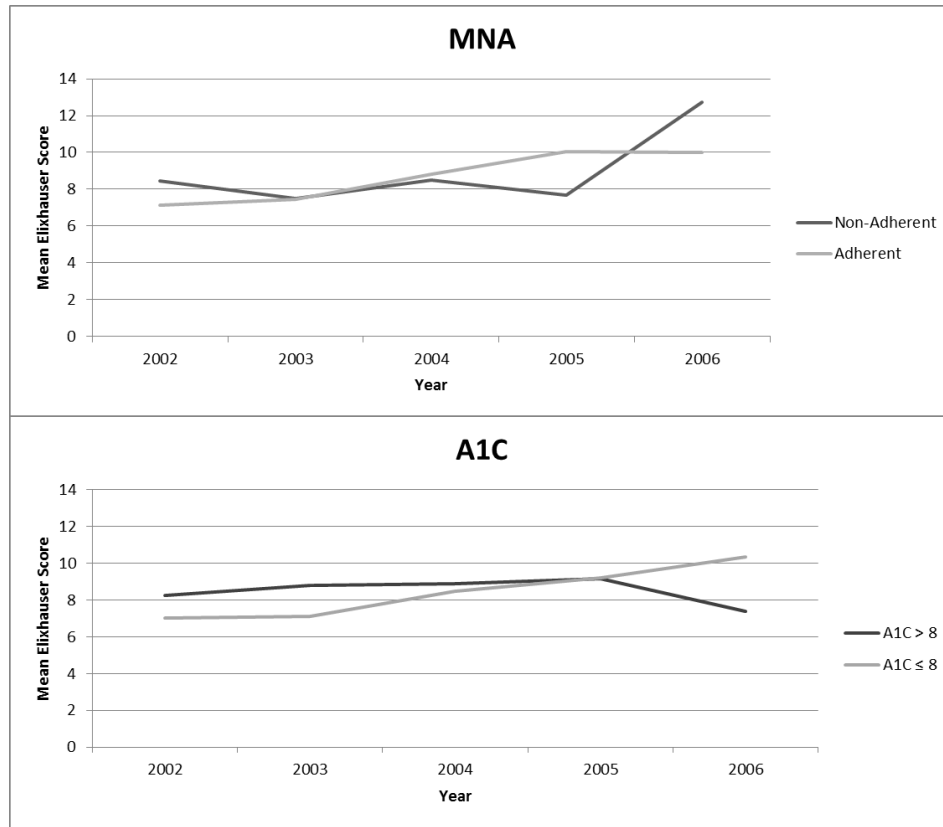


Figure 5. Mean Elixhauser score by MNA and A1C status over time.

Plots modeling the percentage of missing MNA and A1C values over time are given in **Figure 6**, while logistic regression results examining the relationship between the dichotomized missing MNA and missing A1C values and other demographic covariates are given in **Table 3**.

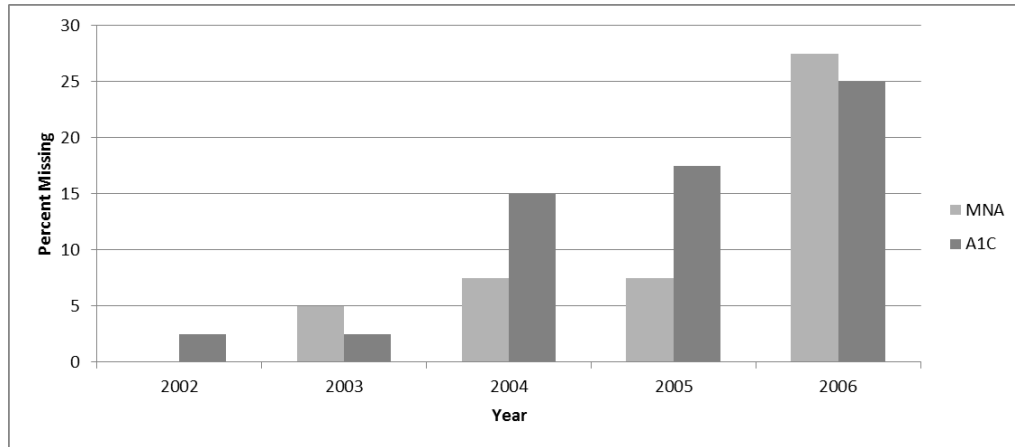


Figure 6. Percentage of missing MNA and A1C values over time.

Table 3. Missing MNA and A1C covariates by demographics.

<i>Covariate</i>	Missing MNA		Missing A1C	
	<i>OR</i>	<i>P-Value</i>	<i>OR</i>	<i>P-Value</i>
Time	3.117	<0.0001	1.896	0.0003
Age	1.139	0.0248	1.011	0.7925
Comorbidity Count	0.915	0.2467	1.121	0.0803
<i>Region</i>				
South (reference)				
Northeast	3.959	0.2144	0.397	0.3617
Midatlantic	32.092	<0.0001	0.504	0.3562
Midwest	6.841	0.0086	1.725	0.2618
West	0.000	0.9990	0.000	0.9991
<i>Gender</i>				
Male (reference)				
Female	2.823	0.3959	10.274	0.0291
<i>Living</i>				
Urban (reference)				
Rural	1.205	0.7573	0.992	0.9936
<i>Marital Status</i>				
Married (reference)				
Unmarried	8.861	0.0006	1.736	0.2375
<i>Percent Service Connected Disability</i>				
<50% (reference)				
≥50%	1.735	0.3305	1.549	0.3234

There is some association between missingness in the MNA and A1C variable by observed time, age, regional, gender, and marital status. Missingness is not expected to depend on the individual patient's comorbidity count; thus we can assume that the missing mechanism is likely to be MAR. We then performed both CCA and LTMI via Proc LTA to impute missing categorical covariates.

We utilized Poisson-GLMM and NB-GLMM methods of analysis on the complete case and imputed diabetes datasets. Model comparison was made using conditional AIC, bias, asymptotic and estimated standard errors, and 95% confidence intervals of parameter estimates.

4. 2. Results

A comparison of conditional AIC and σ_p results for both CCA and LTMI Poisson-GLMM and NB-GLMM methods is presented in Table 4.

Table 4. Comparison of goodness of fit and dispersion statistics.

CCA			
		σ_p	Conditional AIC
MNA	Poisson-GLMM	1.242	533.78
	NB-GLMM	1.054	530.35
A1C	Poisson-GLMM	1.349	521.10
	NB-GLMM	1.014	512.00
LTMI			
		σ_p	Conditional AIC
MNA	Poisson-GLMM	1.253	835.30
	NB-GLMM	1.033	826.70
A1C	Poisson-GLMM	1.258	835.86
	NB-GLMM	1.032	827.06

It is clear that mild overdispersion is present in both scenarios for the Poisson models, while goodness of fit is slightly better in the negative binomial models. Furthermore, Figure 7 gives the standard errors for all parameters included in the models by GLMM distribution and method of dealing with missing data. The LTMI methods result in lower standard error estimates compared to CCA.

Tables 5 and 6 respectively give CCA and LTMI results comparing patient Elixhauser score with MNA or A1C under Poisson and negative binomial regression after adjusting for demographics. Looking at the analysis of MNA under negative binomial distribution and imputing missing data via LTMI, we conclude that there are significant or borderline differences in Elixhauser score based on time, western region, and marital status. With each year increase, patients have higher multimorbidity (RR=1.084, p<0.0001). Patients living in the west have decreased comorbidity burden compared to patients in the south (RR=0.597, p=0.0054). Unmarried patients are also

likely to have higher multimorbidity than married patients (RR=1.134, p=0.0579). Medication non-adherent patients have statistically comparable comorbidity burden compared to medication adherent patients in this cohort (RR=1.041, p=0.2474). In the NB-GLMM model utilizing LTMI containing A1C value, patients with abnormally high blood sugar do not have a statistically higher comorbidity burden than those with normal blood sugar (RR=1.024, p=0.3692). Covariates in other models produce generally comparable parameter estimates.

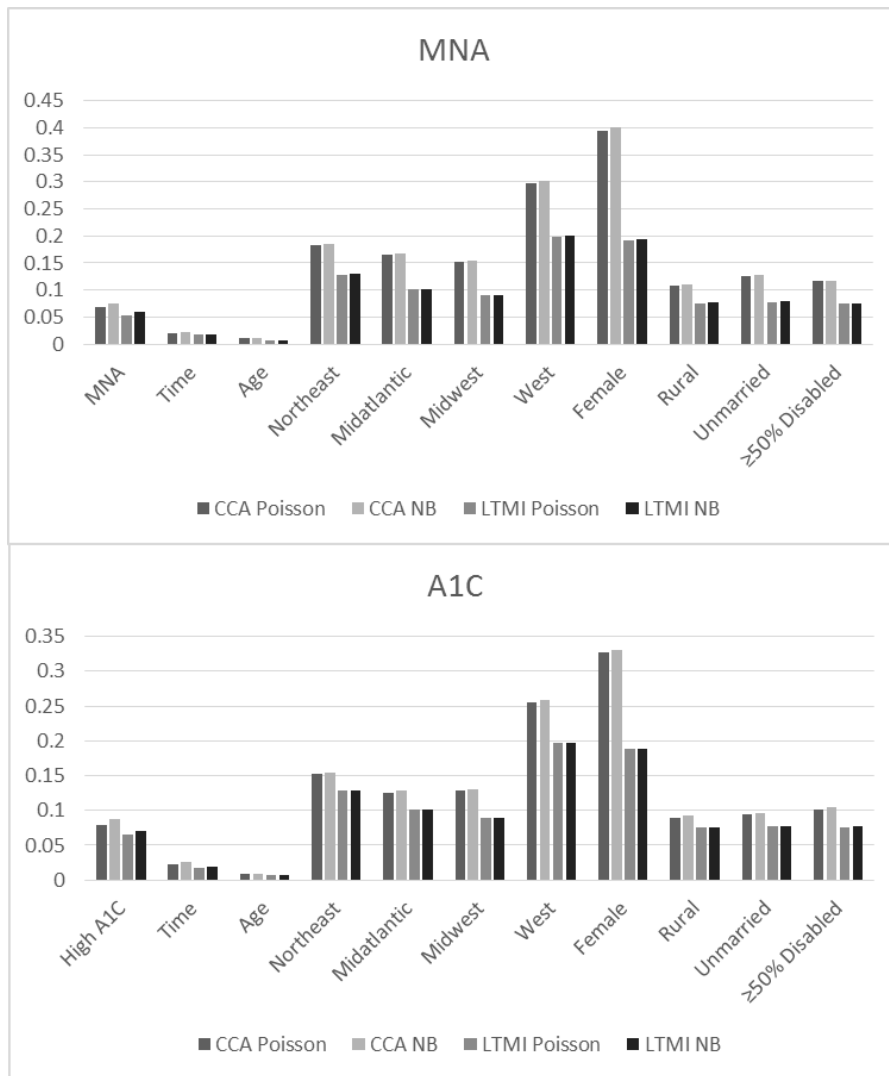


Figure 7. Standard errors for covariates in MNA and A1C model scenarios by GLMM distribution and method of addressing missing data.

Table 5. Relationship between Elixhauser score and covariates via CCA by distribution.

Poisson-GLMM						
Covariate	MNA			A1C		
	RR	SE	P-Value	RR	SE	P-Value
<i>Covariate with Missingness</i>						
Medically adherent (reference)						--
Medically non-adherent	1.063	0.068	0.3714			
Normal blood sugar (reference)						
Abnormally high blood sugar		--		1.159	0.079	0.0633
Time	1.141	0.021	<0.0001	1.145	0.022	<0.0001
Age	1.002	0.012	0.8700	1.009	0.009	0.3429
<i>Region</i>						
South (reference)						
Northeast	0.950	0.182	0.7806	1.010	0.152	0.9487
Midatlantic	1.041	0.166	0.8085	1.108	0.126	0.4192
Midwest	1.053	0.153	0.7361	1.161	0.129	0.2488
West	0.618	0.297	0.1079	0.674	0.255	0.1247
<i>Gender</i>						
Male (reference)						
Female	1.081	0.395	0.8441	0.996	0.327	0.9913
<i>Living</i>						
Urban (reference)						
Rural	1.015	0.109	0.8915	0.965	0.090	0.6944
<i>Marital Status</i>						
Married (reference)						
Unmarried	1.195	0.125	0.1558	1.232	0.095	0.0304
<i>Percent Service Connected Disability</i>						
<50% (reference)						
≥50%	1.044	0.116	0.7128	1.073	0.102	0.4939
NB-GLMM						
Covariate	MNA			A1C		
	RR	SE	P-Value	RR	SE	P-Value
<i>Covariate with Missingness</i>						
Medically adherent (reference)						--
Medically non-adherent	1.060	0.075	0.4367			
Normal blood sugar (reference)						
Abnormally high blood sugar		--		1.182	0.088	0.0605
Time	1.142	0.023	<0.0001	1.148	0.026	<0.0001
Age	1.002	0.012	0.8899	1.009	0.009	0.3453
<i>Region</i>						
South (reference)						
Northeast	0.953	0.185	0.7958	1.014	0.155	0.9297
Midatlantic	1.042	0.168	0.8067	1.116	0.128	0.3943
Midwest	1.052	0.155	0.7453	1.167	0.131	0.2400
West	0.614	0.302	0.1088	0.675	0.259	0.1328
<i>Gender</i>						
Male (reference)						
Female	1.093	0.401	0.8240	1.019	0.330	0.9548
<i>Living</i>						
Urban (reference)						
Rural	1.014	0.110	0.9001	0.963	0.092	0.6784
<i>Marital Status</i>						
Married (reference)						
Unmarried	1.198	0.127	0.1581	1.236	0.096	0.0305
<i>Percent Service Connected Disability</i>						
<50% (reference)						
≥50%	1.046	0.118	0.7032	1.083	0.104	0.4461

Table 6. Relationship between Elixhauser score and covariates via LTMI by distribution.

Poisson-GLMM						
Covariate	MNA			A1C		
	RR	SE	P-Value	RR	SE	P-Value
<i>Covariate with Missingness</i>						
Medically adherent (reference)						--
Medically non-adherent	1.031	0.053	0.2810			
Normal blood sugar (reference)						
Abnormally high blood sugar		--		1.005	0.065	0.4670
Time	1.083	0.017	<0.0001	1.084	0.017	<0.0001
Age	1.003	0.007	0.3571	1.002	0.007	0.3690
<i>Region</i>						
South (reference)						
Northeast	1.044	0.129	0.3687	1.040	0.128	0.3799
Midatlantic	1.046	0.101	0.3274	1.046	0.101	0.3286
Midwest	1.046	0.090	0.3079	1.047	0.089	0.3038
West	0.599	0.198	0.0052	0.598	0.197	0.0049
<i>Gender</i>						
Male (reference)						
Female	1.213	0.191	0.1573	1.231	0.189	0.1360
<i>Living</i>						
Urban (reference)						
Rural	1.044	0.076	0.2873	1.043	0.076	0.2906
<i>Marital Status</i>						
Married (reference)						
Unmarried	1.131	0.078	0.0591	1.131	0.078	0.0587
<i>Percent Service Connected Disability</i>						
<50% (reference)						
≥50%	1.085	0.076	0.1405	1.086	0.076	0.1391
NB-GLMM						
Covariate	MNA			A1C		
	RR	SE	P-Value	RR	SE	P-Value
<i>Covariate with Missingness</i>						
Medically adherent (reference)						--
Medically non-adherent	1.041	0.059	0.2474			
Normal blood sugar (reference)						
Abnormally high blood sugar		--		1.024	0.071	0.3692
Time	1.084	0.019	<0.0001	1.086	0.019	<0.0001
Age	1.003	0.007	0.3492	1.003	0.007	0.3552
<i>Region</i>						
South (reference)						
Northeast	1.050	0.130	0.3557	1.044	0.128	0.3697
Midatlantic	1.051	0.102	0.3126	1.049	0.101	0.3177
Midwest	1.050	0.091	0.2966	1.052	0.090	0.2869
West	0.597	0.200	0.0054	0.598	0.197	0.0050
<i>Gender</i>						
Male (reference)						
Female	1.215	0.194	0.1587	1.243	0.189	0.1260
<i>Living</i>						
Urban (reference)						
Rural	1.044	0.077	0.2893	1.042	0.076	0.2945
<i>Marital Status</i>						
Married (reference)						
Unmarried	1.134	0.079	0.0579	1.132	0.078	0.0571
<i>Percent Service Connected Disability</i>						
<50% (reference)						
≥50%	1.090	0.076	0.1313	1.095	0.077	0.1188

5. Discussion

It is not uncommon in real datasets to deal simultaneously with both missing predictor data and overdispersion resulting from model specifications. This investigation extends the approaches we examined previously to deal with overdispersion in Poisson-distributed count data to longitudinal Poisson analysis including overdispersion, while simultaneously addressing the issue of time varying categorical covariates with missing data in the models. We considered outlier dependent overdispersion and made real data application while also addressing the co-occurrence of missingness in important categorical predictors via both CCA and LTMI methods. We then made comparison among all methods using conditional AIC, bias, and error estimates.

In the simulation study, LTMI consistently gave lower ASE and ESE values for both missingness scenarios compared to CCA methods. The contrast was more pronounced in the various 50% MAR and MNAR scenarios. The bias for both predictors compared to the true beta values was generally higher for the CCA methods than for LTMI, particularly for variable X_2 in the MAR, MNAR, and higher missingness scenarios. The negative-binomial also outperforms the Poisson in many cases, particularly among the MAR datasets and the 50% missingness scenarios, giving moderately adjusted errors and comparable goodness of fit. Outlier dependent overdispersion appears to be adequately addressed via the negative binomial regression.

The real data application gives similar results for analyzing the relationship between Elixhauser score and covariates. LTMI methods are preferred over CCA given the lower standard error estimates produced, while NB-GLMM is preferred over Poisson-GLMM. Jointly considering all results and criteria, we can conclude both that NB-GLMM is superior for analysis of data containing overdispersion in the outcome and also that LTMI is preferred for imputing missing data in time-varying categorical predictors. Therefore, both methods are utilized when analyzing datasets in which both issues are present.

Concluding Remarks

CHAPTER 6

1. Summary and discussion of all results

In Chapter 2, we provide a comprehensive comparative analysis of six different models for dealing with overdispersion caused by different mechanisms when modeling count data. Overall, the negative-binomial models appeared to demonstrate superiority in adjusting for overdispersion in the simulation studies. The NB-GLMM performed best in modeling count of comorbidity data in the motivating NLST study. This model also appeared to deal most effectively with overdispersion in the small *Salmonella* dataset. Based on our analyses, we conclude that NB-GLMM is superior overall for modeling count data characterized by overdispersion, jointly considering all criteria. Simple post hoc scaling in the Poisson model to decrease overdispersion was not consistently effective, as basic scaling does not take the specific cause of the overdispersion into account. Our results further demonstrate that the best method for dealing with overdispersion will likely vary by dataset depending on the cause of the overdispersion. The negative-binomial model may account for overdispersion due to a number of common causes, but it is not ideal in every case. Numerous model options should be considered when overdispersion is an issue.

In Chapter 3, we utilized simulations to compare Poisson and negative binomial methods for analyzing cross-sectional and longitudinal datasets with two binary predictors and count outcome with overdispersion of varying magnitudes resulting from the addition of outliers or zero inflation. Magnitude of overdispersion was measured by dispersion parameter σ_p , defined as the ratio of the Pearson χ^2 value to its corresponding degrees of freedom $n - p$. Comparison among models was made using Type 1 error with a true β_1 value of 0.01, Type 2 errors using true β_1 values of 0.41 or 0.92, and coverage probability of β_1 for all effect sizes of β_1 . It would appear that a general threshold for relying on the simple Poisson model for cross-sectional and longitudinal datasets is in cases where $\sigma_p \leq 1.2$. For cross-sectional datasets, the negative binomial distribution via NB or NB-GLMM should be utilized if $1.2 < \sigma_p \leq 1.5$. For higher values

of σ_p in these scenarios, NB-GLMM should be utilized up to $\sigma_p \leq 5.0$. However, if $\sigma_p > 5.0$ for longitudinal datasets or if $\sigma_p \geq 10.0$ for cross-sectional datasets, the model will likely not be reliable based on adjustment for overdispersion and should be checked for additional modeling errors. Results of our real data application to the NLST and *Salmonella* datasets indicate that these high levels of overdispersion require adjustment via the NB or the NB-GLMM, which is also supported by our analysis. The percent increase in standard errors resulting from the negative binomial models compared to the unadjusted Poisson increased in correspondence with higher magnitudes of overdispersion.

In Chapter 4, we propose a latent transition multiple imputation approach to deal with missing data in time varying categorical covariates in count outcome datasets. This study is the first to assess and implement LTMI for modeling time varying missing categorical covariate data. We have demonstrated that this method is statistically efficient and leads to unbiased estimates and can be implemented using standard software. In comparing simulated and real data scenarios, parameter standard errors were most efficient in the LTMI scenarios using Proc LTA for the dynamic imputers model. In simulation studies, LTMI-LTA outperformed other methods most clearly in the 50% MAR scenarios. Complete case analysis performed fairly well in the 20% MCAR scenario, and generally produced standard error results of greater magnitude otherwise. LCMI methods produced biased estimates and reduced standard error estimates in the simulations compared to the dataset with no missing data. The standard error estimates in the real diabetes analysis were also lowest for LTMI-LTA, notably in the case of the A1C and MNA variables with missing data. Goodness of fit was measured and compared via conditional AIC, which is useful for choosing mixed effects models when dealing with clustered data, and results are comparable for all LCMI and LTMI scenarios in both real and simulated data scenarios. LTMI-LTA outperforms other methods for dealing with missing data in time varying categorical covariates, particularly in various non-MCAR scenarios with a higher percentage of missingness,

jointly considering all criteria. Additionally, LTMI methods are computationally inexpensive and the results are easily interpretable for clinicians.

It is not uncommon in real datasets to deal simultaneously with missing predictor data and overdispersion resulting from model specifications. Our investigation in Chapter 5 extends the approaches we examined previously to deal with overdispersion in time invariant data to longitudinal Poisson analysis including overdispersion, while simultaneously addressing the issue of time varying categorical covariates with missing data in the models. Here, we considered outlier dependent overdispersion and made real data application while also addressing the co-occurrence of missingness in important categorical predictors via both CCA and LTMI methods. We then made comparison among all methods using conditional AIC, bias, and error estimates. In the simulation study, LTMI consistently gave moderately adjusted ASE and ESE values for both missingness scenarios compared to CCA methods. The contrast is again more pronounced in the various 50% MAR scenarios. The negative-binomial also outperforms the Poisson in many cases, particularly among the MAR datasets and several 50% missingness scenarios. There is comparable conditional AIC goodness of fit between the LTMI negative binomial and CCA methods. Application to the real diabetes dataset gives similar results for analyzing the relationship between Elixhauser score and covariates. LTMI methods are preferred over CCA given the lower standard error estimates produced. Jointly considering all results and criteria, we can conclude that NB-GLMM is preferable for analysis of data containing overdispersion in the outcome and that LTMI is preferred for imputing missing data in time-varying categorical predictors.

2. Future work

These analyses may be expanded to include future research in several areas. First, the generalizability of the Pearson X^2/df thresholds of overdispersion may be improved with simulations in which overdispersion in the count outcome results from additional causes, including the removal of important covariates or necessary random effects. Predictors from different distributions could also be considered in addition to the binary, such as normal and

uniform covariates. Thresholds for declaring the presence of overdispersion utilizing the deviance/df value could also be examined in addition to the Pearson X^2/df and compared with the results given here.

We could further consider the ability of LTMI methods to address the issue of missingness in time varying categorical covariates under additional outcome scenarios, including normal and logistic regression models. The joint ability of LTMI and NB methods to address co-occurring overdispersion relative to the Poisson and time varying categorical covariate missingness could also be considered under additional overdispersion scenarios. Lastly, it would be interesting to consider the ability of Bayesian Poisson regression to deal with the presence of overdispersion in datasets, perhaps via Proc MCMC methods in SAS 9.4, and compare with our scale adjustment and modeling techniques.

APPENDIX 1

CHAPTER 2 METHODOLOGY

Throughout this section, we assume estimating the full generalized linear model

$$y = \exp(\beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3)$$

as well as the two reduced generalized linear models

$$y = \exp(\alpha_0 + x_2\alpha_2 + x_3\alpha_3)$$

$$y = \exp(\gamma_0 + x_3\gamma_3)$$

Coefficients of remaining terms are unbiased as discussed in (Neuhaus and Jewell 1993).

If we generate three independent covariates

$$x_1 \sim \text{Bernoulli}(0.5) \quad x_2 \sim \text{Bernoulli}(0.5) \quad x_3 \sim \text{Bernoulli}(0.5)$$

then

$$1.00 + 2.00x_1 + 1.50x_2 + 1.00x_3 \sim \text{Known Discrete}$$

If we define an outcome

$$y \sim \text{Poisson}(\exp(1.00 + 2.00x_1 + 1.50x_2 + 1.00x_3))$$

then $E(y) \approx 58.10$. Estimating a full model should result in unbiased estimates of the parameters

$$\hat{\beta}_0 \approx 1.00 \quad \hat{\beta}_1 \approx 2.00 \quad \hat{\beta}_2 \approx 1.50 \quad \hat{\beta}_3 \approx 1.00$$

Estimating a reduced model in which we leave out x_1 should result in:

$$\hat{\alpha}_0 \approx 2.43 \quad \hat{\alpha}_2 \approx 1.50 \quad \hat{\alpha}_3 \approx 1.00$$

where the constant term can be solved from the discrete distribution. Estimating a reduced model in which we leave out x_1 and x_2 should result in:

$$\hat{\gamma}_0 \approx 3.44 \qquad \hat{\gamma}_3 \approx 1.00$$

for which the constant term can be solved from the discrete distribution.

Similarly, if we define three covariates

$$x_1 \sim \text{Normal}(1,2) \quad x_2 \sim \text{Normal}(2,3) \quad x_3 \sim \text{Normal}(3,4)$$

then a linear combination of these covariates gives the following distribution:

$$1 + 0.50x_1 - 0.75x_2 + 0.25x_3 \sim \text{Normal}\left(\frac{3}{4}, \frac{39}{16}\right)$$

If we define an outcome as

$$y \sim \text{Poisson}(\exp(1 + 0.50x_1 - 0.75x_2 + 0.25x_3))$$

then

$$E(y) = \exp\left[\frac{3}{4} + \left(\frac{39}{16}\right)\left(\frac{1}{2}\right)\right] = \exp\left(\frac{63}{32}\right) \approx 7.16$$

Estimating a full model should result in:

$$\hat{\beta}_0 \approx 1.00 \quad \hat{\beta}_1 \approx 0.50 \quad \hat{\beta}_2 \approx -0.75 \quad \hat{\beta}_3 \approx 0.25$$

Estimating a reduced model in which x_1 is omitted should result in:

$$\hat{\alpha}_0 \approx 1.75 \qquad \hat{\alpha}_2 \approx -0.75 \quad \hat{\alpha}_3 \approx 0.25$$

The constant term can be estimated under constrained maximum likelihood so that it is approximately equal to:

$$\hat{\alpha}_0 \approx \ln \left(\frac{\exp \left[\frac{3}{4} + \left(\frac{39}{16} \right) \left(\frac{1}{2} \right) \right]}{\exp \left[-\frac{3}{4} + \left(\frac{31}{16} \right) \left(\frac{1}{2} \right) \right]} \right) = \frac{63}{32} - \frac{7}{32} = 1.75$$

Estimating a reduced model in which x_1 and x_2 are omitted should result in:

$$\hat{\gamma}_0 \approx 1.09 \qquad \hat{\gamma}_3 \approx 0.25$$

The constant term can be estimated under constrained maximum likelihood so that it is approximately equal to:

$$\hat{\gamma}_0 \approx \ln \left(\frac{\exp \left[\frac{3}{4} + \left(\frac{39}{16} \right) \left(\frac{1}{2} \right) \right]}{\exp \left[\frac{3}{4} + \left(\frac{1}{4} \right) \left(\frac{1}{2} \right) \right]} \right) = \frac{63}{32} - \frac{28}{32} = 1.09375$$

Finally, if we define three covariates

$$x_1 \sim \text{Uniform}(5,10) \quad x_2 \sim \text{Uniform}(10,15) \quad x_3 \sim \text{Uniform}(15,20)$$

and we define an outcome as

$$y \sim \text{Poisson}(\exp(1 + 0.50x_1 - 0.75x_2 + 0.25x_3))$$

then $E(y) \approx 1.81$. Estimating a full model should result in unbiased estimates of the parameters

$$\hat{\beta}_0 \approx 1.00 \quad \hat{\beta}_1 \approx 0.50 \quad \hat{\beta}_2 \approx -0.75 \quad \hat{\beta}_3 \approx 0.25$$

Estimating a reduced model in which we leave out x_1 should result in:

$$\hat{\alpha}_0 \approx 5.00 \qquad \hat{\alpha}_2 \approx -0.75 \quad \hat{\alpha}_3 \approx 0.25$$

where the constant term was obtained from simulation. Estimating a reduced model in which we leave out x_1 and x_2 should result in:

$$\hat{\gamma}_0 \approx -3.86 \qquad \hat{\gamma}_3 \approx 0.25$$

where the constant term was obtained from simulation.

APPENDIX 2

ADDITIONAL FIGURES AND TABLES CORRESPONDING TO CHAPTER 2

The figures in Appendix 2 correspond to those presented in Chapter 2, giving results for methods with larger magnitudes of overdispersion and normal predictors. The distributions of the variables in our normal simulation scenario are illustrated in Appendix 2 Figure 1a-d. A figure of AIC and BIC where two important predictors have been omitted from the model can be found in Figure 2, and the corresponding SE figure is given in Figure 3. Similar figures of AIC and BIC for the models containing larger outliers (+150) and 40% zero outliers are given respectively in Figures 4a and 4b, and the corresponding figures of SE can be found in Figures 5a and 5b. Lastly, Figure 6 gives the AIC and BIC for the random effects model with larger variance, and Figure 7 shows the corresponding SE figure. The results from these additional analyses are qualitatively similar to those presented in the paper. The lowest AIC and BIC values and moderately corrected standard errors are overall generally given by the NB-GLMM .

Appendix 2 also gives a summary of goodness-of-fit results for all covariate, outlier, and random effects dependent overdispersion models in Tables 1, 2a and 2b, and 3, respectively. The NB and NB-GLMM give consistently lower AIC and BIC values compared to other models. In general, the NB and NB-GLMM also give moderate SE and 95% CI coverage, the original Poisson and Poisson-GLMM give lower values, and the scale-adjusted models give mixed results.

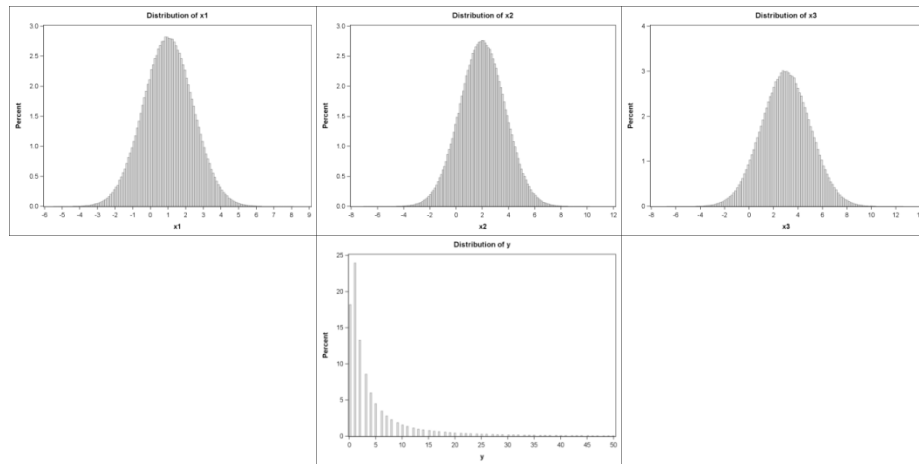


Figure 1a-d. Distributions of normal covariates and response for simulation study.

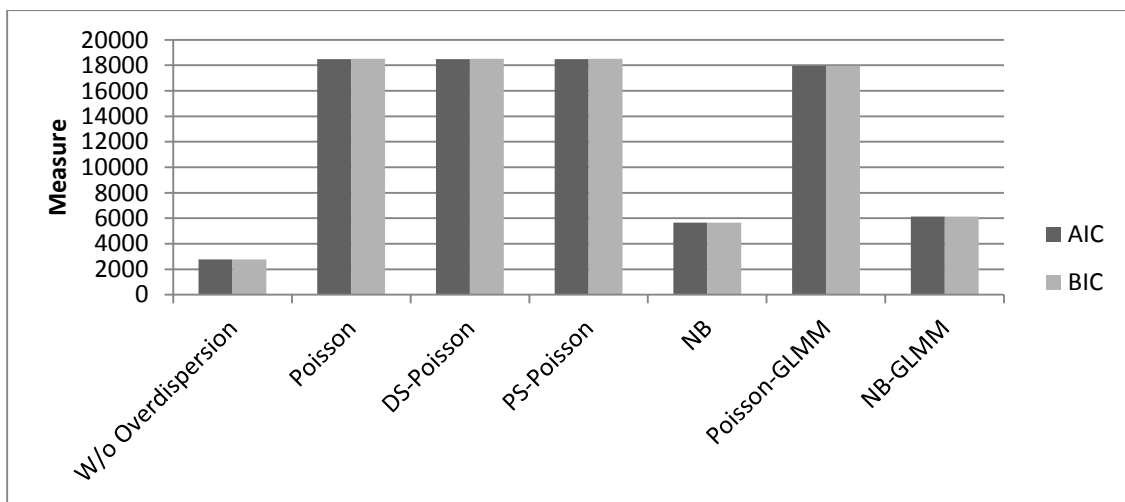


Figure 2. Mean AIC and BIC values for simulated dataset with two important predictors omitted.

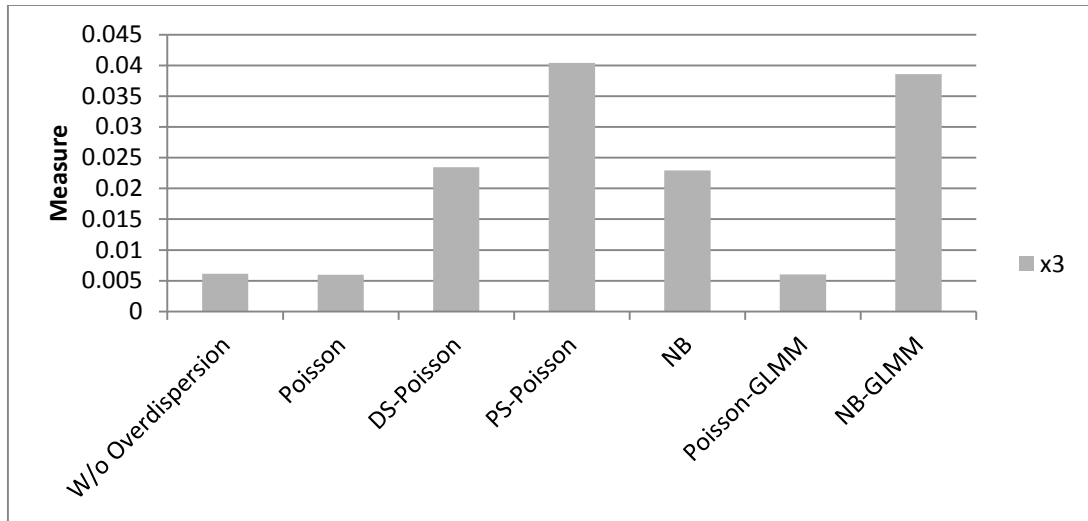


Figure 3. Mean parameter SE values for simulated dataset with two important predictors omitted.

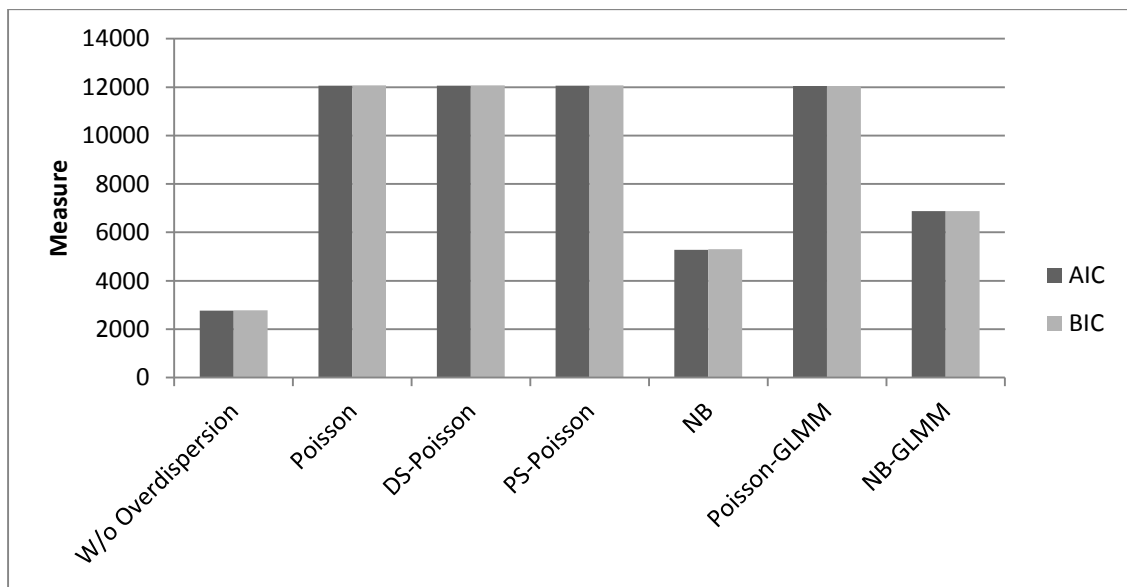


Figure 4a. Mean AIC and BIC values for simulated dataset with outliers added (+150).

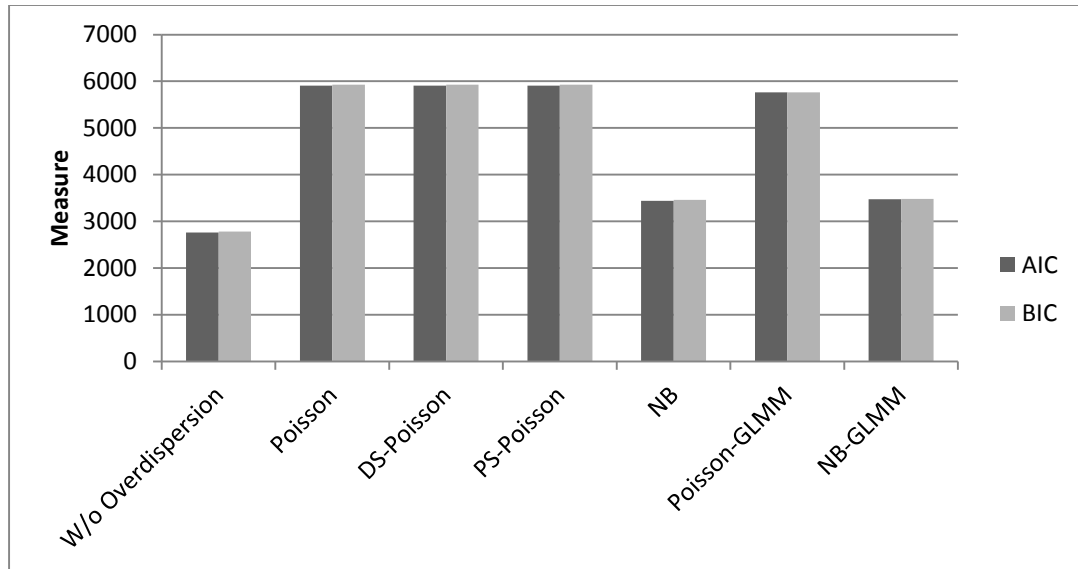


Figure 4b. Mean AIC and BIC values for simulated dataset with zero outliers added (40%).

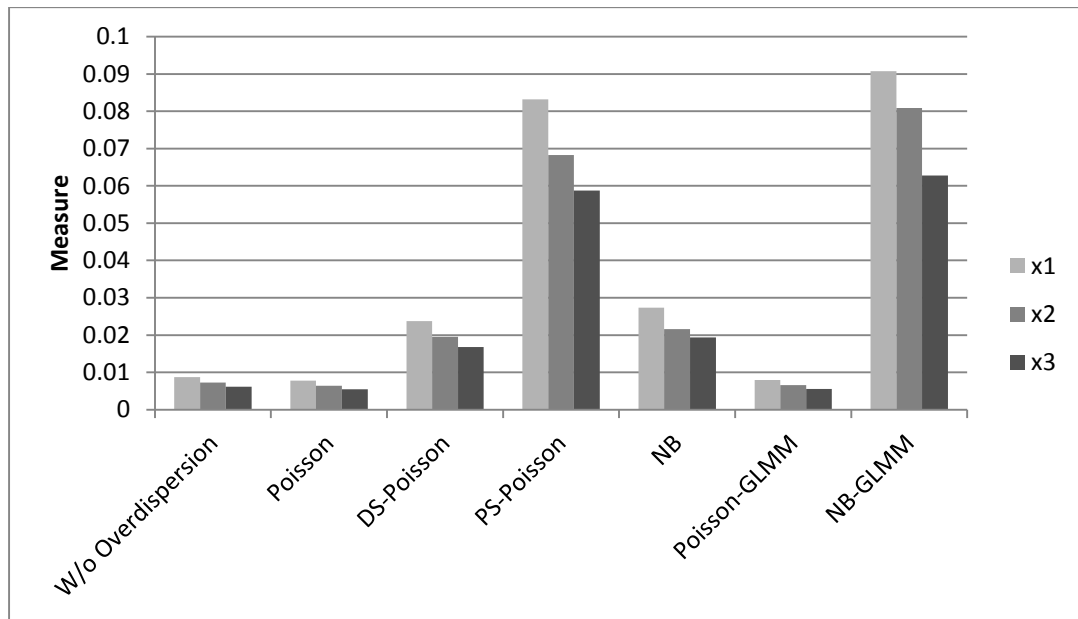


Figure 5a. Mean parameter SE values for simulated dataset with outliers added (+150).

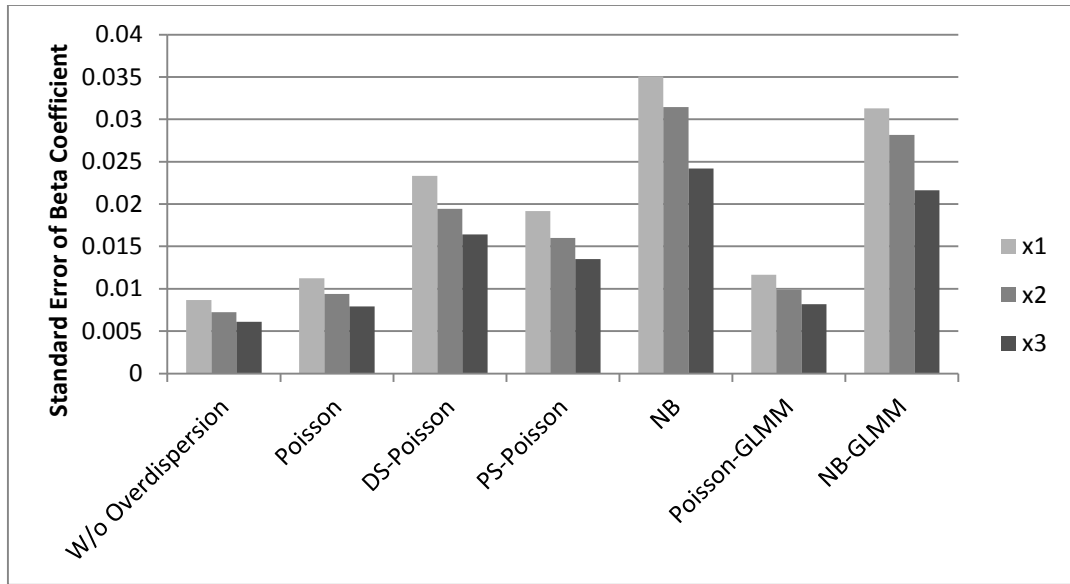


Figure 5b. Mean parameter SE values for simulated dataset with zero outliers added (40%).

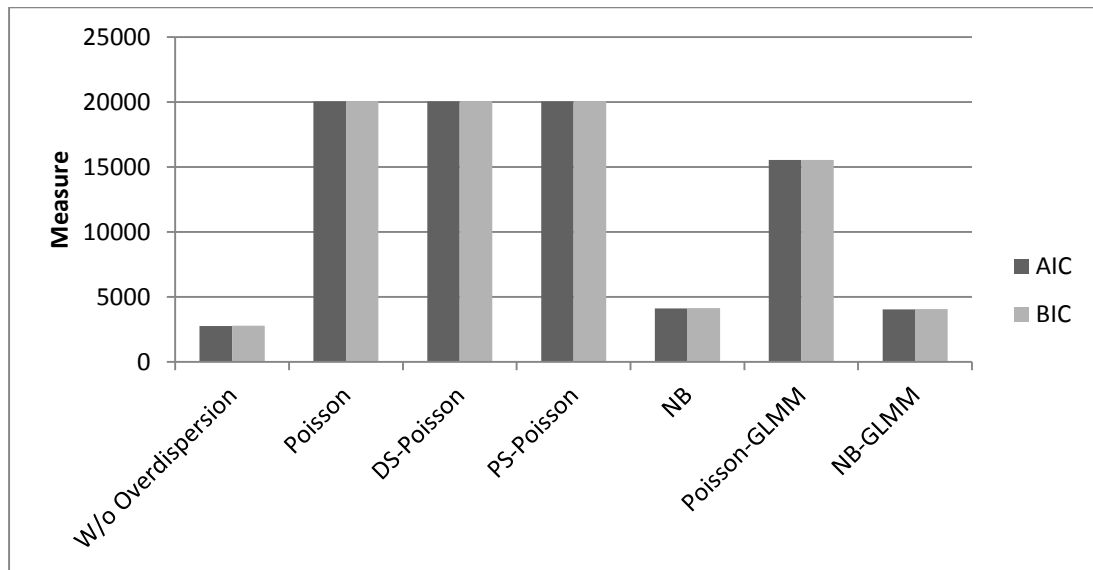


Figure 6. Mean AIC and BIC values for simulated dataset with random effect $\gamma \sim N(0, group/5)$.

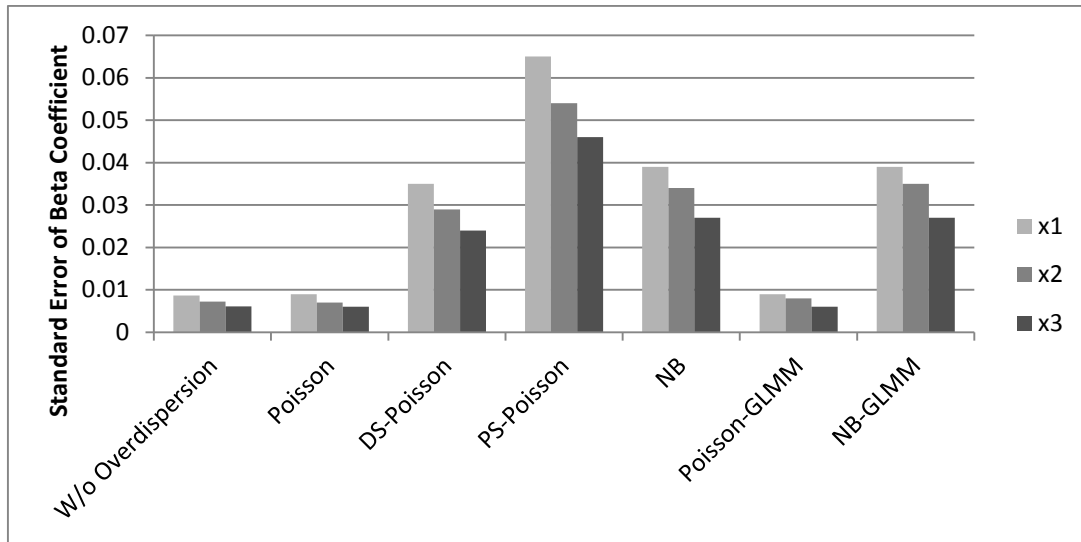


Figure 7. Mean parameter SE values for simulated dataset with random effect $\gamma \sim N(0, group/5)$.

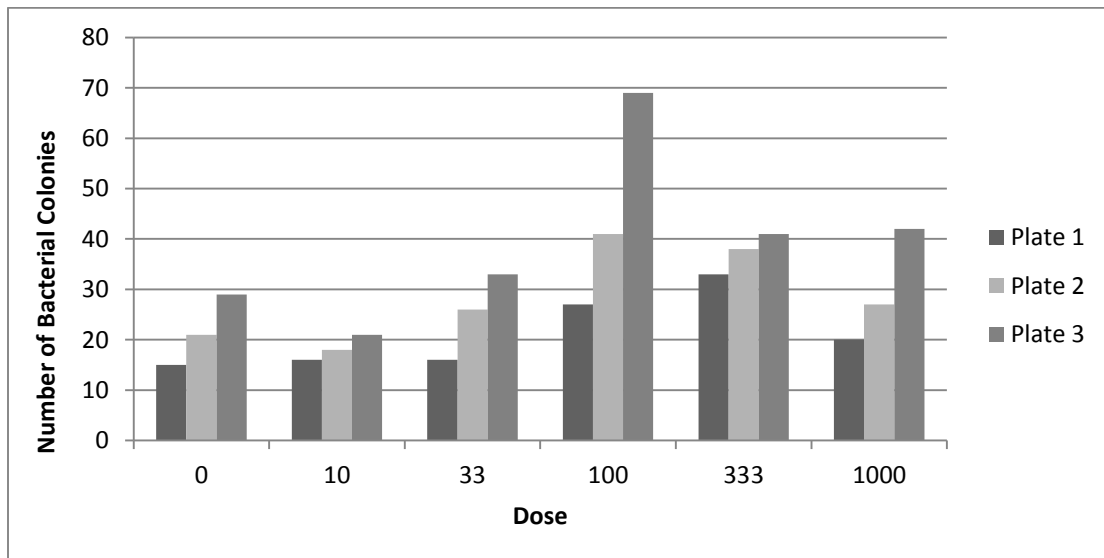


Figure 8. Bacterial count response by dose and plate in *Salmonella* dataset.

Table 1. Comparison of methods for dealing with covariate dependent overdispersion using simulated data by covariate distribution and number of omitted predictors.

<i>Covariate</i>	<i>No.Omitted</i>	<i>Value</i>	Poisson	DS-Poisson	PS-Poisson	NB	Poisson-GLMM	NB-GLMM
Normal	1	AIC	6273.13	6273.13	6273.13	4249.52	6113.79	4315.32
		BIC	6287.86	6287.86	6287.86	4269.15	6115.30	4316.95
	2	AIC	18501.66	18501.66	18501.66	5655.97	17952.10	6133.45
		BIC	18511.47	18511.47	18511.47	5670.70	17953.31	6134.77
Binary	1	AIC	43306.61	43306.61	43306.61	9363.32	42845.28	9386.82
		BIC	43321.34	43321.34	43321.34	9382.95	42846.79	9388.44
	2	AIC	68541.13	68541.13	68541.13	9894.71	67836.79	9899.21
		BIC	68550.94	68550.94	68550.94	9909.43	67838.00	9900.52
Uniform	1	AIC	10582.41	10582.41	10582.41	4166.76	10128.85	4181.82
		BIC	10597.14	10597.14	10597.14	4186.39	10130.37	4183.44
	2	AIC	47351.56	47351.56	47351.56	5937.99	45948.34	5955.97
		BIC	47361.37	47361.37	47361.37	5952.72	45949.55	5957.30

Table 2a. Comparison of methods for dealing with outlier dependent overdispersion using simulated data by covariate distribution and magnitude of outliers.

<i>Covariate</i>	<i>Outlier</i>	<i>Value</i>	Poisson	DS-Poisson	PS-Poisson	NB	Poisson-GLMM	NB-GLMM
Normal	+50	AIC	5153.42	5153.42	5153.42	4427.09	5156.79	4622.05
		BIC	5173.05	5173.05	5173.05	4451.63	5158.60	4623.99
	+150	AIC	12055.89	12055.89	12055.89	5278.41	12051.44	6874.90
		BIC	12075.52	12075.52	12075.52	5302.95	12053.26	6876.80
Binary	+50	AIC	4989.94	4989.94	4989.94	4192.88	4993.94	5313.65
		BIC	5009.57	5009.57	5009.57	4217.42	4995.76	5315.47
	+150	AIC	11405.51	11405.51	11405.51	5075.12	11409.51	7613.69
		BIC	11425.14	11425.14	11425.14	5099.66	11411.32	7615.51
Uniform	+50	AIC	4967.39	4967.39	4967.39	4099.04	4971.39	5306.66
		BIC	4987.03	4987.03	4987.03	4123.58	4973.21	5308.47
	+150	AIC	11534.20	11534.20	11534.20	4963.86	11538.20	7610.13
		BIC	11553.83	11553.83	11553.83	4988.40	11540.02	7611.94

Table 2b. Comparison of methods for dealing with outlier dependent overdispersion using simulated data by covariate distribution and percentage of excess zeros.

Covariate	Outlier	Value	Poisson	DS-Poisson	PS-Poisson	NB	Poisson-GLMM	NB-GLMM
Normal	Lower %	AIC	4708.91	4708.91	4708.91	3694.7	4647.39	3703.74
		BIC	4728.54	4728.54	4728.54	3719.24	4649.2	3705.66
	Higher %	AIC	5905.90	5905.90	5905.90	3437.05	5763.10	3477.61
		BIC	5925.53	5925.53	5925.53	3461.58	5764.91	3479.54
Binary	Lower %	AIC	3596.27	3596.27	3596.27	3477.75	3599.03	3488.81
		BIC	3615.90	3615.90	3615.90	3502.29	3600.85	3490.73
	Higher %	AIC	3369.53	3369.53	3369.53	2886.92	3369.42	2893.47
		BIC	3389.16	3389.16	3389.16	2911.46	3371.24	2895.40
Uniform	Lower %	AIC	3355.53	3355.53	3355.53	3278.80	3358.50	3297.99
		BIC	3375.16	3375.16	3375.16	3303.34	3360.32	3299.92
	Higher %	AIC	3112.77	3112.77	3112.77	2739.91	3113.16	2743.07
		BIC	3132.40	3132.40	3132.40	2764.44	3114.98	2744.99

Table 3. Comparison of methods for dealing with lower random effect dependent overdispersion using simulated data by covariate distribution and magnitude of outliers.

Covariate	γ	Value	Poisson	DS-Poisson	PS-Poisson	NB	Poisson-GLMM	NB-GLMM
Normal	$N(0, g/10)$	AIC	8301.66	8301.66	8301.66	4424.69	7740.06	4418.28
		BIC	8321.29	8321.29	8321.29	4449.23	7741.88	4420.39
	$N(0, g/5)$	AIC	20037.57	20037.57	20037.57	4098.81	15522.58	4044.95
		BIC	20057.20	20057.20	20057.20	4123.35	15524.40	4047.06
Binary	$N(0, g/10)$	AIC	5410.88	5410.88	5410.88	4586.73	5327.46	4582.96
		BIC	5430.51	5430.51	5430.51	4611.27	5329.28	4585.08
	$N(0, g/5)$	AIC	10269.75	10269.75	10269.75	5377.65	9717.36	5418.19
		BIC	10289.38	10289.38	10289.38	5402.19	9719.18	5420.31
Uniform	$N(0, g/10)$	AIC	5056.44	5056.44	5056.44	4374.92	4983.67	4364.20
		BIC	5076.07	5076.07	5076.07	4399.46	4985.49	4366.32
	$N(0, g/5)$	AIC	9427.43	9427.43	9427.43	5155.89	8931.69	5130.26
		BIC	9447.07	9447.07	9447.07	5180.43	8933.50	5132.38

APPENDIX 3

DERIVATION OF LIKELIHOOD CONTRIBUTION BASED ON LATENT STATUS

CORRESPONDING TO CHAPTER 4

Recall that we have defined $L_j = (l_1, \dots, l_T)$ to represent class membership indicators at time $j = 1, \dots, T$ where $l_j = 1, \dots, L$. The vector $Y_j = (Y_{1j}, \dots, Y_{Mj})$ represents the M observed categorical variables where each variable may take on values $k = 1, \dots, C_m$ for every time point, $j = 1, \dots, T$. To derive the likelihood equation for LTMI, we must begin with the basic joint probability equation below:

$$p(Y_1 = y_{i1}, \dots, Y_T = y_{iT}; L_j = l_j) = p(L_j = l_j) \times p(Y_1 = y_{i1}, \dots, Y_T = y_{iT} | L_j = l_j)$$

We define $\delta_{l_1} = P(L_1 = l_1)$ as the probability that the latent class is l_1 at the first time point. This implies that the sum of the probabilities of the observation being assigned to a particular latent class $l_j = 1, \dots, L$ at given time point $j = 1, \dots, T$ is given as follows:

$$\delta_{l_1}^{(j)} = P(L_j = l_j) = \sum_{l_1=1}^L \dots \sum_{l_{j-1}=1}^L \delta_{l_1} \prod_{j=2}^T \tau_{l_j | l_{j-1}}^{(j)}$$

With $\tau_{l_j | l_{j-1}}^{(j)} = p(L_j = l_j | L_{j-1} = l_{j-1})$ defined as the probability that the latent class at time j is any l_j given the latent class assignments at previous time points. We then define the probability that the i^{th} individual has categorical response $k = 1, \dots, C_m$ for each variable $m = 1, \dots, M$ across time points $j = 1, \dots, T$ given the time varying latent class assignment l_j as:

$$p(Y_1 = y_{i1}, \dots, Y_T = y_{iT}; L_j = l_j) = \prod_{j=1}^T \prod_{m=1}^M \prod_{k=1}^{C_m} \rho_{mkj | l_j}^{I(y_{imj}=k)}$$

where $\rho_{mkj|l_j} = p(Y_{mj} = k | L_j = l_j)$. Therefore, the joint probability that the i^{th} individual exhibits a specific categorical response at each time point y_{i1}, \dots, y_{iT} , and latent class membership at the corresponding time point, $l_j = (l_1, \dots, l_T)$, is given as follows:

$$p(Y_1 = y_{i1}, \dots, Y_T = y_{iT}, L_j = l_j) = \left[\delta_{l_1} \prod_{j=2}^T \tau_{l_j | l_{j-1}}^{(j)} \right] \times \left[\prod_{j=1}^T \prod_{m=1}^M \prod_{k=1}^{C_m} \rho_{mkj|l_j}^{I(y_{imj}=k)} \right]$$

We can then calculate the likelihood contribution for the i^{th} individual to the whole model across all possible latent classes at each time point. This is the joint probability that the i^{th} individual exhibits a specific categorical response for each variable at each time point and is also assigned to a specific latent class at each time point, and is given below:

$$L(\theta; Y_1 = y_{i1}, \dots, Y_T = y_{iT}) = \sum_{l_1=1}^L \dots \sum_{l_{j-1}=1}^L \left[\delta_{l_1} \prod_{j=2}^T \tau_{l_j | l_{j-1}}^{(j)} \right] \times \left[\prod_{j=1}^T \prod_{m=1}^M \prod_{k=1}^{C_m} \rho_{mkj|l_j}^{I(y_{imj}=k)} \right]$$

APPENDIX 4

ADDITIONAL FIGURES AND TABLES CORRESPONDING TO CHAPTER 4

Table 1. Results of CCA imputation for 20% and 50% missingness scenarios.

	No missing	20% Missing				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3229.46	2593.84	2596.56	2530.56	2541.12	2524.03
Pearson χ^2 /df	1.002	1.002	1.001	0.926	0.940	0.954
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.414	0.410	0.414	0.384	0.394	0.310
Mean RR	1.513	1.507	1.513	1.468	1.483	1.363
ASE	0.047	0.053	0.052	0.052	0.052	0.061
ESE	0.051	0.056	0.056	0.054	0.057	0.064
Bias	-0.004	0.000	-0.004	0.026	0.016	0.100
Mean 95% CI for β_1	0.321, 0.506	0.307, 0.514	0.312, 0.516	0.282, 0.485	0.292, 0.496	0.209, 0.428
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.696	0.693	0.697	0.638	0.621	0.653
Mean RR	2.006	2.000	2.008	1.893	1.861	1.921
ASE	0.048	0.053	0.053	0.052	0.053	0.061
ESE	0.049	0.054	0.054	0.054	0.055	0.063
Bias	-0.006	-0.003	-0.007	0.052	0.069	0.037
Mean 95% CI for β_2	0.603, 0.790	0.588, 0.798	0.593, 0.801	0.535, 0.741	0.517, 0.725	0.532, 0.774
	No missing	50% Missing				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3229.46	1632.64	1622.03	1560.37	1579.51	1580.24
Pearson χ^2 /df	1.002	1.005	1.004	0.887	0.925	0.992
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.414	0.410	0.416	0.360	0.385	0.302
Mean RR	1.513	1.507	1.516	1.433	1.470	1.353
ASE	0.047	0.067	0.064	0.062	0.063	0.082
ESE	0.051	0.070	0.073	0.067	0.066	0.084
Bias	-0.004	0.000	-0.006	0.050	0.025	0.108
Mean 95% CI for β_1	0.321, 0.506	0.279, 0.541	0.290, 0.543	0.238, 0.483	0.260, 0.509	0.141, 0.462
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.696	0.692	0.699	0.594	0.569	0.713
Mean RR	2.006	1.998	2.012	1.811	1.766	2.040
ASE	0.048	0.067	0.067	0.064	0.065	0.082
ESE	0.049	0.069	0.070	0.065	0.066	0.084
Bias	-0.006	-0.002	-0.009	0.096	0.121	-0.023
Mean 95% CI for β_2	0.603, 0.790	0.559, 0.825	0.568, 0.830	0.469, 0.719	0.441, 0.697	0.552, 0.874

- ASE = the mean of the Asymptotic SE as computed by *Proc MEANS* (reported as mean of ASE)
- ESE = the SD of the estimates of beta as computed by *Proc MEANS* (reported SD Estimate)

Table 2. Results of LCMI-LCA imputation for 20% and 50% missingness scenarios.

	No missing	20% Missing				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3229.46	3248.57	3257.22	3240.26	3232.65	3300.96
Pearson χ^2 /df	1.002	1.022	1.022	1.011	0.999	1.061
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.414	0.299	0.311	0.315	0.391	0.100
Mean RR	1.513	1.349	1.365	1.370	1.478	1.105
ASE	0.047	0.047	0.048	0.050	0.048	0.056
ESE	0.051	0.054	0.057	0.059	0.053	0.070
Bias	-0.004	0.111	0.099	0.095	0.019	0.310
Mean 95% CI for β_1	0.321, 0.506	0.206, 0.393	0.216, 0.406	0.215, 0.415	0.297, 0.486	-0.012, 0.212
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.696	0.590	0.600	0.604	0.663	0.441
Mean RR	2.006	1.804	1.822	1.829	1.941	1.554
ASE	0.048	0.047	0.047	0.049	0.047	0.053
ESE	0.049	0.051	0.052	0.054	0.046	0.065
Bias	-0.006	0.100	0.090	0.086	0.027	0.249
Mean 95% CI for β_2	0.603, 0.790	0.498, 0.682	0.507, 0.693	0.507, 0.700	0.571, 0.755	0.336, 0.546
	No missing	50% Missing				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3229.46	3265.24	3280.81	3256.76	3245.24	3292.48
Pearson χ^2 /df	1.002	1.037	1.037	1.026	1.005	1.058
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.414	0.204	0.220	0.209	0.306	-0.011
Mean RR	1.513	1.226	1.246	1.232	1.358	0.989
ASE	0.047	0.046	0.046	0.047	0.046	0.051
ESE	0.051	0.050	0.057	0.064	0.055	0.066
Bias	-0.004	0.206	0.190	0.201	0.104	0.421
Mean 95% CI for β_1	0.321, 0.506	0.113, 0.296	0.129, 0.312	0.114, 0.304	0.213, 0.399	-0.013, 0.091
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.696	0.505	0.515	0.509	0.578	0.355
Mean RR	2.006	1.657	1.674	1.664	1.782	1.426
ASE	0.048	0.044	0.044	0.045	0.044	0.046
ESE	0.049	0.046	0.050	0.054	0.045	0.056
Bias	-0.006	0.185	0.177	0.181	0.112	0.335
Mean 95% CI for β_2	0.603, 0.790	0.418, 0.591	0.429, 0.602	0.421, 0.597	0.490, 0.666	0.264, 0.446

- ASE = the mean of the Asymptotic SE as computed by *Proc MEANS* (reported as mean of ASE)
- ESE = the SD of the estimates of beta as computed by *Proc MEANS* (reported SD Estimate)

Table 3. Results of LCMI-LMM imputation for 20% and 50% missingness scenarios.

	No missing	20% Missing				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3229.46	3249.22	3259.85	3241.80	3232.67	3304.41
Pearson χ^2 /df	1.002	1.020	1.023	1.012	0.999	1.063
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.414	0.303	0.314	0.316	0.390	0.106
Mean RR	1.513	1.354	1.369	1.372	1.477	1.112
ASE	0.047	0.047	0.047	0.050	0.048	0.056
ESE	0.051	0.052	0.056	0.060	0.054	0.069
Bias	-0.004	0.107	0.096	0.094	0.020	0.304
Mean 95% CI for β_1	0.321, 0.506	0.210, 0.397	0.220, 0.407	0.216, 0.415	0.296, 0.484	-0.006, 0.218
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.696	0.595	0.603	0.604	0.662	0.446
Mean RR	2.006	1.813	1.828	1.829	1.939	1.562
ASE	0.048	0.047	0.047	0.048	0.047	0.053
ESE	0.049	0.047	0.052	0.054	0.045	0.066
Bias	-0.006	0.095	0.087	0.086	0.028	0.244
Mean 95% CI for β_2	0.603, 0.790	0.504, 0.687	0.511, 0.695	0.508, 0.700	0.571, 0.754	0.341, 0.551
	No missing	50% Missing				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3229.46	3265.74	3280.63	3258.22	3244.47	3296.56
Pearson χ^2 /df	1.002	1.036	1.037	1.025	1.004	1.058
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.414	0.213	0.224	0.218	0.311	-0.003
Mean RR	1.513	1.237	1.251	1.244	1.365	0.997
ASE	0.047	0.046	0.046	0.047	0.047	0.051
ESE	0.051	0.052	0.054	0.062	0.054	0.061
Bias	-0.004	0.197	0.186	0.192	0.179	0.413
Mean 95% CI for β_1	0.321, 0.506	0.121, 0.305	0.132, 0.315	0.123, 0.314	0.217, 0.404	-0.106, 0.101
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.696	0.514	0.518	0.518	0.585	0.363
Mean RR	2.006	1.672	1.679	1.679	1.795	1.438
ASE	0.048	0.044	0.044	0.045	0.045	0.047
ESE	0.049	0.046	0.048	0.052	0.044	0.053
Bias	-0.006	0.176	0.172	0.172	0.105	0.327
Mean 95% CI for β_2	0.603, 0.790	0.427, 0.601	0.432, 0.605	0.429, 0.607	0.497, 0.673	0.271, 0.456

- ASE = the mean of the Asymptotic SE as computed by *Proc MEANS* (reported as mean of ASE)
- ESE = the SD of the estimates of beta as computed by *Proc MEANS* (reported SD Estimate)

Table 4. Results of LTMI-LMM imputation for 20% and 50% missingness scenarios.

	No missing	20% Missing				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3229.46	3230.34	3236.55	3231.60	3228.29	3273.66
Pearson χ^2 /df	1.002	1.001	1.007	1.002	1.001	1.027
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.414	0.406	0.406	0.405	0.402	0.335
Mean RR	1.513	1.501	1.501	1.499	1.495	1.398
ASE	0.047	0.049	0.048	0.049	0.049	0.057
ESE	0.051	0.057	0.055	0.054	0.057	0.069
Bias	-0.004	0.004	0.004	0.005	0.008	0.075
Mean 95% CI for β_1	0.321, 0.506	0.311, 0.501	0.311, 0.501	0.309, 0.501	0.305, 0.498	0.223, 0.448
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.696	0.692	0.691	0.690	0.690	0.644
Mean RR	2.006	1.998	1.996	1.994	1.994	1.904
ASE	0.048	0.049	0.049	0.049	0.049	0.056
ESE	0.049	0.050	0.051	0.051	0.054	0.065
Bias	-0.006	-0.002	-0.001	0.000	0.000	0.046
Mean 95% CI for β_2	0.603, 0.790	0.597, 0.788	0.596, 0.787	0.594, 0.786	0.594, 0.786	0.534, 0.755
	No missing	50% Missing				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3229.46	3231.48	3225.08	3228.12	3234.17	3278.25
Pearson χ^2 /df	1.002	1.003	0.999	1.001	1.006	1.029
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.414	0.412	0.416	0.408	0.404	0.350
Mean RR	1.513	1.510	1.516	1.504	1.498	1.419
ASE	0.047	0.050	0.050	0.050	0.050	0.059
ESE	0.051	0.054	0.055	0.064	0.060	0.081
Bias	-0.004	-0.002	-0.006	0.002	0.006	0.060
Mean 95% CI for β_1	0.321, 0.506	0.314, 0.510	0.318, 0.513	0.309, 0.507	0.306, 0.502	0.233, 0.467
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.696	0.694	0.699	0.694	0.691	0.656
Mean RR	2.006	2.002	2.012	2.002	1.996	1.927
ASE	0.048	0.050	0.049	0.050	0.050	0.058
ESE	0.049	0.052	0.052	0.057	0.056	0.075
Bias	-0.006	-0.004	-0.009	-0.004	-0.001	0.034
Mean 95% CI for β_2	0.603, 0.790	0.597, 0.792	0.602, 0.796	0.596, 0.792	0.594, 0.788	0.542, 0.771

- ASE = the mean of the Asymptotic SE as computed by *Proc MEANS* (reported as mean of ASE)
- ESE = the SD of the estimates of beta as computed by *Proc MEANS* (reported SD Estimate)

Table 5. Relationship between Elixhauser score and covariates in Diabetes dataset via CCA.

Covariate	MNA			A1C		
	RR	SE	P-Value	RR	SE	P-Value
<i>Covariate with Missingness</i>						
Medically adherent (reference)				--		
Medically non-adherent	1.017	0.006	0.0076			
Normal blood sugar (reference)		--				
Abnormally high blood sugar				1.008	0.008	0.3299
Time	1.045	0.002	<0.0001	1.048	0.002	<0.0001
Age	1.003	0.000	<0.0001	1.003	0.000	<0.0001
<i>Region</i>						
South (reference)						
Northeast	0.993	0.014	0.6110	1.030	0.017	0.0782
Midatlantic	1.005	0.010	0.6299	1.022	0.013	0.0950
Midwest	0.998	0.011	0.8684	1.017	0.015	0.2345
West	0.970	0.012	0.0131	0.989	0.014	0.4478
<i>Gender</i>						
Male (reference)						
Female	1.040	0.026	0.1408	1.024	0.032	0.4603
<i>Race</i>						
NHW (reference)						
NHB	1.058	0.012	<.0001	1.073	0.014	<.0001
Hispanic	1.006	0.017	0.7039	0.992	0.018	0.6560
Other	0.904	0.013	<.0001	0.937	0.015	<.0001
<i>Living</i>						
Urban (reference)						
Rural	1.007	0.008	0.4015	1.004	0.010	0.7192
<i>Marital Status</i>						
Married (reference)						
Unmarried	1.038	0.008	<0.0001	1.038	0.010	0.0003
<i>Percent Service Connected Disability</i>						
<50% (reference)						
≥50%	1.078	0.011	<0.0001	1.064	0.014	<0.0001

Table 6. Relationship between Elixhauser score and covariates in Diabetes dataset via LCMI-LCA.

<i>Covariate</i>	MNA			A1C		
	<i>RR</i>	<i>SE</i>	<i>P-Value</i>	<i>RR</i>	<i>SE</i>	<i>P-Value</i>
<i>Covariate with Missingness</i>						
Medically adherent (reference)					--	
Medically non-adherent	1.008	0.006	0.0831			
Normal blood sugar (reference)					--	
Abnormally high blood sugar				1.023	0.006	0.0002
Time	1.044	0.002	<0.0001	1.045	0.002	<0.0001
Age	1.003	0.000	<0.0001	1.003	0.000	<0.0001
<i>Region</i>						
South (reference)						
Northeast	0.987	0.013	0.1526	0.996	0.013	0.3684
Midatlantic	1.002	0.010	0.4119	1.008	0.010	0.2042
Midwest	0.996	0.010	0.3389	0.999	0.010	0.4666
West	0.976	0.011	0.0136	0.976	0.011	0.0170
<i>Gender</i>						
Male (reference)						
Female	1.042	0.023	0.0381	1.043	0.023	0.0351
<i>Race</i>						
NHW (reference)						
NHB	1.062	0.011	<0.0001	1.059	0.011	<0.0001
Hispanic	0.976	0.015	0.0576	0.973	0.015	0.0385
Other	0.931	0.011	<0.0001	0.931	0.011	<0.0001
<i>Living</i>						
Urban (reference)						
Rural	1.004	0.007	0.2898	1.005	0.008	0.2481
<i>Marital Status</i>						
Married (reference)						
Unmarried	1.044	0.008	<0.0001	1.041	0.008	<0.0001
<i>Percent Service Connected Disability</i>						
<50% (reference)						
≥50%	1.080	0.010	<0.0001	1.076	0.011	<0.0001

Table 7. Relationship between Elixhauser score and covariates in Diabetes dataset via LCMI-LMM.

Covariate	MNA			A1C		
	RR	SE	P-Value	RR	SE	P-Value
<i>Covariate with Missingness</i>						
Medically adherent (reference)					--	
Medically non-adherent	1.008	0.006	0.0813			
Normal blood sugar (reference)						
Abnormally high blood sugar		--		1.033	0.006	<0.0001
Time	1.044	0.002	<0.0001	1.044	0.002	<0.0001
Age	1.002	0.000	<0.0001	1.003	0.000	<0.0001
<i>Region</i>						
South (reference)						
Northeast	0.985	0.012	0.1123	0.986	0.012	0.1247
Midatlantic	1.002	0.010	0.4236	1.002	0.010	0.4105
Midwest	0.997	0.010	0.3774	0.997	0.010	0.3948
West	0.975	0.011	0.0101	0.975	0.011	0.0114
<i>Gender</i>						
Male (reference)						
Female	1.034	0.023	0.0793	1.035	0.023	0.0677
<i>Race</i>						
NHW (reference)						
NHB	1.064	0.011	<0.0001	1.062	0.011	<0.0001
Hispanic	0.964	0.013	0.0025	0.965	0.013	0.0027
Other	0.935	0.011	<0.0001	0.936	0.011	<0.0001
<i>Living</i>						
Urban (reference)						
Rural	1.001	0.007	0.4719	1.001	0.007	0.4704
<i>Marital Status</i>						
Married (reference)						
Unmarried	1.043	0.007	<0.0001	1.042	0.007	<0.0001
<i>Percent Service Connected Disability</i>						
<50% (reference)						
≥50%	1.079	0.010	<0.0001	1.079	0.011	<0.0001

Table 8. Relationship between Elixhauser score and covariates in Diabetes dataset via LTMI-LMM.

<i>Covariate</i>	MNA			A1C		
	<i>RR</i>	<i>SE</i>	<i>P-Value</i>	<i>RR</i>	<i>SE</i>	<i>P-Value</i>
<i>Covariate with Missingness</i>						
Medically adherent (reference)						--
Medically non-adherent	1.010	0.006	0.0408			
Normal blood sugar (reference)			--			
Abnormally high blood sugar				1.033	0.006	<0.0001
Time	1.044	0.002	<0.0001	1.044	0.002	<0.0001
Age	1.002	0.000	<0.0001	1.003	0.000	<0.0001
<i>Region</i>						
South (reference)						
Northeast	0.985	0.012	0.1122	0.986	0.012	0.1254
Midatlantic	1.002	0.010	0.4231	1.002	0.010	0.4101
Midwest	0.997	0.010	0.3785	0.997	0.010	0.3950
West	0.975	0.011	0.0102	0.975	0.011	0.0114
<i>Gender</i>						
Male (reference)						
Female	1.033	0.023	0.0799	1.035	0.023	0.0683
<i>Race</i>						
NHW (reference)						
NHB	1.064	0.011	<0.0001	1.062	0.011	<0.0001
Hispanic	0.964	0.013	0.0024	0.965	0.013	0.0027
Other	0.934	0.011	<0.0001	0.936	0.011	<0.0001
<i>Living</i>						
Urban (reference)						
Rural	1.001	0.007	0.4705	1.001	0.007	0.4694
<i>Marital Status</i>						
Married (reference)						
Unmarried	1.043	0.007	<0.0001	1.042	0.007	<0.0001
<i>Percent Service Connected Disability</i>						
<50% (reference)						
≥50%	1.079	0.010	<0.0001	1.079	0.010	<0.0001

APPENDIX 5

ADDITIONAL TABLES CORRESPONDING TO CHAPTER 5

Tables giving Poisson-GLMM and NB-GLMM results for the 20% missingness scenarios are given below. They demonstrate results comparable to those in the manuscript with 20% missingness.

Table 1. Results for high outlier scenario via CCA.

	No missing or overdispersion	Poisson-GLMM				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3229.34	3093.89	3118.75	3062.57	3075.42	2923.05
σ_p	1.000	1.437	1.459	1.384	1.403	1.313
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.409	0.507	0.508	0.478	0.487	0.348
Mean RR	1.505	1.660	1.662	1.613	1.627	1.416
ASE	0.047	0.054	0.053	0.053	0.053	0.063
ESE	0.048	0.099	0.101	0.099	0.100	0.101
Bias	0.001	0.097	0.098	0.068	0.077	0.062
Mean 95% CI for β_1	0.317, 0.501	0.402, 0.612	0.404, 0.612	0.374, 0.582	0.382, 0.592	0.224, 0.472
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.690	0.821	0.822	0.766	0.746	0.728
Mean RR	1.994	2.273	2.275	2.151	2.109	2.071
ASE	0.048	0.054	0.054	0.054	0.054	0.064
ESE	0.047	0.097	0.100	0.096	0.097	0.099
Bias	0.000	0.131	0.132	0.076	0.056	0.038
Mean 95% CI for β_2	0.597, 0.783	0.714, 0.928	0.716, 0.929	0.660, 0.872	0.640, 0.852	0.603, 0.853
	No missing or overdispersion	NB-GLMM				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3228.55	3118.43	3141.09	3104.14	3108.49	2971.05
σ_p	1.000	0.971	0.971	0.964	0.965	0.959
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.409	0.496	0.499	0.466	0.473	0.348
Mean RR	1.505	1.642	1.647	1.594	1.605	1.416
ASE	0.047	0.074	0.074	0.072	0.073	0.081
ESE	0.048	0.089	0.091	0.089	0.090	0.092
Bias	0.001	0.086	0.089	0.056	0.063	0.062
Mean 95% CI for β_1	0.317, 0.501	0.350, 0.642	0.353, 0.645	0.325, 0.608	0.330, 0.617	0.190, 0.506
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.690	0.811	0.813	0.755	0.733	0.728
Mean RR	1.994	2.250	2.255	2.128	2.081	2.071
ASE	0.047	0.075	0.075	0.073	0.074	0.081
ESE	0.047	0.088	0.089	0.087	0.087	0.093
Bias	0.000	0.121	0.123	0.065	0.043	0.038
Mean 95% CI for β_2	0.597, 0.783	0.664, 0.958	0.666, 0.961	0.612, 0.897	0.588, 0.877	0.570, 0.886

Table 2. Results for high outlier scenario via LTMI.

	No missing or overdispersion	Poisson-GLMM				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3229.34	3867.38	3865.26	3866.97	3863.95	3765.62
σ_p	1.000	1.454	1.452	1.457	1.454	1.542
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.409	0.481	0.486	0.468	0.478	0.316
Mean RR	1.505	1.618	1.626	1.597	1.613	1.372
ASE	0.047	0.050	0.049	0.049	0.050	0.057
ESE	0.048	0.090	0.090	0.090	0.091	0.096
Bias	0.001	0.071	0.076	0.058	0.068	0.094
Mean 95% CI for β_1	0.317, 0.501	0.386, 0.581	0.391, 0.583	0.372, 0.567	0.381, 0.579	0.198, 0.424
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.690	0.800	0.804	0.792	0.802	0.676
Mean RR	1.994	2.226	2.234	2.208	2.230	1.966
ASE	0.048	0.050	0.049	0.050	0.050	0.057
ESE	0.047	0.088	0.088	0.087	0.088	0.095
Bias	0.000	0.110	0.114	0.102	0.112	0.014
Mean 95% CI for β_2	0.597, 0.783	0.702, 0.898	0.708, 0.901	0.690, 0.886	0.704, 0.902	0.557, 0.781
	No missing or overdispersion	NB-GLMM				
		MCAR	MAR (x2)	MAR (y)	MAR (x2,y)	MNAR
Conditional AIC	3228.55	3894.29	3893.44	3891.91	3889.28	3787.40
σ_p	1.000	0.971	0.971	0.970	0.970	0.979
X1 ($\beta_1 = 0.41$)						
Mean β_1	0.409	0.469	0.473	0.453	0.461	0.315
Mean RR	1.505	1.598	1.605	1.573	1.586	1.370
ASE	0.047	0.070	0.068	0.068	0.069	0.079
ESE	0.048	0.082	0.081	0.080	0.081	0.087
Bias	0.001	0.059	0.063	0.043	0.051	0.095
Mean 95% CI for β_1	0.317, 0.501	0.337, 0.605	0.341, 0.605	0.319, 0.586	0.328, 0.598	0.166, 0.486
X2 ($\beta_2 = 0.69$)						
Mean β_2	0.690	0.789	0.792	0.778	0.786	0.674
Mean RR	1.994	2.201	2.208	2.177	2.195	1.962
ASE	0.047	0.068	0.068	0.068	0.069	0.079
ESE	0.047	0.081	0.080	0.078	0.080	0.088
Bias	0.000	0.099	0.102	0.088	0.096	0.016
Mean 95% CI for β_2	0.597, 0.783	0.656, 0.923	0.660, 0.924	0.639, 0.906	0.651, 0.922	0.513, 0.831

REFERENCES

- Aberle, D. R., A. M. Adams, et al. (2010). "Baseline characteristics of participants in the randomized national lung screening trial." *J Natl Cancer Inst* **102**(23): 1771-1779.
- Aberle, D. R., A. M. Adams, et al. (2011). "Reduced lung-cancer mortality with low-dose computed tomographic screening." *N Engl J Med* **365**(5): 395-409.
- Agresti, A. *Categorical Data Analysis*, 2nd Edition. Hoboken, NJ: John Wiley & Sons Inc., 2002.
- Akaike, H. (1974). "A new look at the statistical model identification." *IEEE Transactions on Automatic Control* **19**(6): 716-723.
- Albert P.S., D.A. Follmann (2007). "Random effects and latent processes approaches for analyzing binary longitudinal data with missingness: a comparison of approaches using opiate clinical trial data." *Stat Methods Med Res.* **16**(5): 417-39.
- Allison, P. *Missing Data*. Thousand Oaks, CA: Sage Publications, 2001.
- Aregay, M., Z. Shkedy, et al. (2013). "A hierarchical Bayesian approach for the analysis of longitudinal count data with overdispersion: A simulation study." *Computational Statistics & Data Analysis* **57**(1): 233-245.
- Beunckens, Caroline, Geert Molenberghs, Geert Verbeke, and Craid Mallinckrodt (2008). "A Latent-Class Mixture Model for Incomplete Longitudinal Gaussian Data". *Biometrics* **64**: 96-105.
- Booth, J. G., G. Casella, et al. (2003). "Negative binomial loglinear mixed models." *Statistical Modelling* **3**(3): 179-191.
- Bouche, G., B. Lepage, et al. (2009). "[Application of detecting and taking overdispersion into account in Poisson regression model]." *Rev Epidemiol Sante Publique* **57**(4): 285-296.
- Breslow, N. (1990). "Tests of Hypotheses in Overdispersed Poisson Regression and other Quasi-Likelihood Models." *Journal of the American Statistical Association* **85**(410): 565-571.
- Breslow, N., Clayton, D. (1993). "Approximate inference in generalized linear mixed models." *J. Amer. Statist. Assoc.* **88**: 9-25.
- Cameron, A. C. "Advances in Count Data Regression Talk for the Applied Statistics Workshop", March 28, 2009. <http://cameron.econ.ucdavis.edu/racd/count.html>.
- Cameron, A. C. and P. K. Trivedi (1998). *Regression analysis of count data*. Cambridge, Cambridge University Press.
- Cameron, AC and Trivedi, PK (1986). "Econometric models based on count data: comparison and application of some estimators and tests". *Econometrics.* **1**(1): 29-53.
- Cameron AC, Trivedi PK (1990). "Regression-based Tests for Overdispersion in the Poisson Model". *Journal of Econometrics.* **46**: 347-364.

- Celeux G, Forbes F, Robert CP, Titterington DM (2006). "Deviance information criteria for missing data models". *Bayesian Analysis*. **1**: 651–74.
- Chung H, Lanza ST, L. E. (2008). "Latent transition analysis: Inference and estimation". *Statistics in Medicine* **27**: 1834-1854.
- Chung, Hwan, YouSung Park, and Stephanie T. Lanza (2005). "Latent transition analysis with covariates: pubertal timing and substance use behaviours in adolescent females". *Stat Med* **24**:2895-2910.
- Collings, B. J. and B. H. Margolin (1985). "Testing goodness of fit for the Poisson assumption when observations are not identically distributed". *Journal of the American Statistical Association*. **80**(390): 411-418.
- Cox, D. R. (1983). "Some remarks on overdispersion." *Biometrika* **70**(1): 269-274.
- Dauxois, J.Y., P. Druilhet, et al. (2006). "A bayesian choice between poisson, binomial and negative binomial models." *Test* **15**(2): 423-432.
- Demissie S., et al. (2003). "Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model." *Stat Med*. **22**(4): 545-57.
- Dempster, A.P. N.M. Laird, D.B. Rubin. (1977) "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)*. **39**(1): 1-38.
- Deng, D., and S.R. Paul (2000). "Score tests for zero inflation in generalized linear models". *Canadian Journal of Statistics*, **28**: 563-570.
- Dean, C. (1992). "Testing for overdispersion in Poisson and binomial regression models". *Journal of the American Statistical Association*, **87**(418): 451-457.
- Dean, C., and J.F. Lawless (1989), "Tests for detecting overdispersion in Poisson regression models", *Journal of the American Statistical Association*. **84**: 467-472.
- Donohue, M. C., R. Overholser, et al. (2011). "Conditional Akaike information under generalized linear and proportional hazards mixed models." *Biometrika* **98**(3): 685-700.
- Durán Pacheco, G., J. Hattendorf, et al. (2009). "Performance of analytical methods for overdispersed counts in cluster randomized trials: Sample size, degree of clustering and imbalance." *Stat Med* **28**(24): 2989-3011.
- Engels, Jean Mundahl and Paula Diehr (2003). "Imputation of missing longitudinal data: a comparison of methods". *Journal of Clinical Epidemiology* **56**: 968–976.
- Faddy, M. J. and D. M. Smith (2011). "Analysis of count data with covariate dependence in both mean and variance." *Journal of Applied Statistics* **38**(12): 2683-2694.
- Ferro, M.A. (2014). "Missing data in longitudinal studies: cross-sectional multiple imputation provides similar estimates to full-information maximum likelihood." *Ann Epidemiol*. **24**(1): 75-7.

- Fitzmaurice, G. M., N. M. Laird, et al. (2004). *Applied longitudinal analysis*. Hoboken, N.J.; [Great Britain], Wiley-Interscience.
- Follman, Dean, and Margaret Wu (1995). "An Approximate Generalized Linear Model With Random Effects for Informative Missing Data". *Biometrics* 51: 151-168.
- Gardner, W., E. P. Mulvey, et al. (1995). "Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models." *Psychol Bull* 118(3): 392-404.
- Gebregziabher, M., DeSantis, S. (2010). "A latent class based multiple imputation approach for missing categorical data." *Journal of Statistical Planning and Inference*. 140(11): 3252–3262.
- Goodman MS, Li Y, Stoddard AM, Sorensen G. (2013). "Analysis of Ordinal Outcomes with Longitudinal Covariates Subject to Missingness." *J Appl Stat*. 41(5): 1040-1052.
- Greven, S. and T. Kneib (2010). "On the behaviour of marginal and conditional AIC in linear mixed models." *Biometrika* 97(4): 773-789.
- Grunwald, G.K., S.L. Bruce, L. Jiang, M. Strand, N. Rabinovitch (2011). "A statistical model for under- or overdispersed clustered and longitudinal count data." *Biom J*. 53(4): 578-94.
- Guo, Xu, and Bradley P. Carlin (2004). "Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages" *The American Statistician*, 58(1): 16 – 24.
- Gurmu, S (1991). "Tests for Detecting Overdispersion in the Positive Poisson Regression Model". *American Statistical Association*. 9(2) 215-222.
- Hardin, J. and J. M. Hilbe (2001, 2007). *Generalized linear models and extensions*. College Station, Tex., Stata Press.
- Harel, Ofer and Xiao-Hua Zhou (2007) "Multiple imputation: Review of theory, implementation and software". *Stat Med*. 26: 3057–3077.
- Hayat, M. J. and M. Higgins (2014). "Understanding Poisson regression." *J Nurs Educ* 53(4): 207-215.
- Hedeker, Donald, and Robert D. Gibbons (1997). "Application of random-effects pattern-mixture models for missing data in longitudinal studies." *Psychological Methods* 2(1): 64-78.
- Heinzl, Felix, and Gerhard Tutz. (2013). "Clustering in linear mixed models with approximate Dirichlet process mixtures using EM algorithm." *Statistical Modelling* 13(1): 41–67.
- Heitjan, DF and Rubin, DB (1991). "Ignorability and Coarse Data". *Ann Statist* 19(4): 2244-53.
- Hilbe, J. M. (2007, 2011). *Negative binomial regression*. Cambridge, Cambridge University Press.
- Hinde, John and Clarice G.B. Demetrio (1998). "Overdispersion: Models and estimation." *Computational Statistics and Data Analysis* 27: 151-170.

- Ibrahim J.G., M.H. Chen, S.R. Lipsitz (1999). "Monte Carlo EM for missing covariates in parametric regression models." *Biometrics*. **55**(2): 591-6.
- Ibrahim, J.G., G. Molenberghs, (2009). "Missing data methods in longitudinal studies: A review." *Test* **18**(1), 1-43.
- Joe, H. and R. Zhu (2005). "Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution." *Biom J* **47**(2): 219-229.
- Knol, MJ, et al. (2010). "Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example." *J Clin Epidemiol*. **63**(7): 728-36.
- Komarek, A, G Verbeke, G Molenberghs. "A SAS-Macro for Linear Mixed Models with Finite Normal Mixtures as Random-Effects Distribution, version 1.1." Katholieke Universiteit Leuven Biostatistisch Centrum. April 2002.
- Lambert D. (1992). "Zero-inflated Poisson regression, with an application to defects in manufacturing". *Technometrics*. **34**(1): 1-14.
- Lambert, D. and Roeder, K. (1995). "Overdispersion diagnostics for generalized linear models". *Journal of the American Statistical Association*. **90**(432): 1225-1236.
- Lanza, S. T., Collins, L. M., Lemmon, D. R., Schafer, J. L. (2007). "PROC LCA: A SAS procedure for Latent Class Analysis." *Structural Equation Modeling* **14**(4): 671-694.
- Lanza, S. T., Lemmon, D. R., Schafer, J. L., Collins, L. M., 2008. *PROC LCA & PROC LTA Users' Guide (Version 1.1.5)*. University Park: The Methodology Center, Penn State.
- Lawless JF (1987). "Regression Methods for Poisson Process Data. American Statistical Association". **82**(399): 808-815.
- Lazarsfeld, Paul F. and Neil W. Henry. *Latent Structure Analysis*. Boston: Houghton Mifflin Company, 1968.
- Lee M, Lee K, Lee J. (2014). "Marginalized transition shared random effects models for longitudinal binary data with nonignorable dropout." *Biom J* **56**(2): 230-42.
- Lee S, Park C, Bynng SK (2007). "Tests for detecting overdispersion in poisson models". *Communications in Statistics - Theory and Methods*. **24**(9): 2405-2420.
- Leisch F (2004). "FlexMix: A general framework for finite mixture models and latent class regression in R." *Journal of Statistical Software* **11**(8). URL <http://www.jstatsoft.org/v11/i08/>.
- Li, Xiaoming, Devan V. Mehrotra, and John Barnard (2006). "Analysis of incomplete longitudinal binary data using multiple imputation". *Stat Med* **25**: 2107-2124.
- Liang, H., H. Wu, et al. (2008). "A note on conditional AIC for linear mixed-effects models." *Biometrika* **95**(3): 773-778.

- Lindsey, J.K. (2000) "Obtaining marginal estimates from conditional categorical repeated measurements models with missing data". *Stat Med* **19**: 801-809.
- Lipsitz SR, Ibrahim JG, Chen MH, Peterson H. *Stat Med.* (1999). "Non-ignorable missing covariates in generalized linear models." **18**(17-18): 2435-48.
- Little, R., D. Rubin. *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley & Sons Inc, 2002.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, Sage Publications.
- Lynch C.P., M. Gebregziabher, R.N. Axon, K.E. Hunt, E. Payne, L.E. Egede. (2014). "Geographic and Racial/Ethnic Variations in Patterns of Multimorbidity Burden in Patients with Type 2 Diabetes." *J Gen Intern Med.* 2014 Aug 16. [Epub ahead of print]
- McLachlan G.J. (1997). "On the EM algorithm for overdispersed count data." *Stat Methods Med Res.* **6**(1):76-98.
- McLachlan, Geoffrey, and David Peel. *Finite Mixture Models*. Hoboken, NJ: Wiley and Sons, Inc. 2000.
- McCullagh, P. and J. A. Nelder (1983, 1989). *Generalized linear models*. London ; New York, Chapman and Hall.
- Milanzi, Elasma, Ariel Alonso and Geert Molenberghs (2011). "Ignoring overdispersion in hierarchical loglinear models: Possible problems and solutions." *Stat Med* **31**(14): 1475-1482.
- Molenberghs, G., G. Verbeke, et al. (2007). "An extended random-effects approach to modeling repeated, overdispersed count data." *Lifetime Data Anal* **13**(4): 513-531.
- Molla, DT and Muniswamy B (2012). "Power of tests for overdispersion parameter in negative binomial regression model". *IOSR Journal of Mathematics.* **1**(4): 29-36.
- Morel, JG, Neerchal, NK. *Overdispersion Models in SAS*. Cary, NC: SAS Institute, Inc. 2012.
- Mortelmans, K. and E. Zeiger (2000). "The Ames *Salmonella*/microsome mutagenicity assay." *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **455**(1-2): 29-60.
- Mullahy, J. (1986). "Specification and testing of some modified count data models." *Journal of Econometrics* **33**(3): 341-365.
- Neuhaus, JM and NP Jewell (1993). "A Geometric Approach to Assess Bias Due to Omitted covariates in Generalized Linear Models." *Biometrika* **80**(4): 807-815.
- Nevalainen, Jaakko, Michael G. Kenward and Suvi M. Virtanen (2009). "Missing values in longitudinal dietary data: A multiple imputation approach based on a fully conditional specification". *Stat Med Wiley InterScience*.

- O'Hara Hines, RJ (1997). "A comparison of score tests for overdispersion in generalized linear models". *Journal of Statistical Computation and Simulation*. **58**(1): 323-342.
- Payne EH, Hardin JW, Egede LE, Ramakrishnan V, Selassie A, Gebregziabher M (2015). "Approaches for dealing with various sources of overdispersion in modeling count data: scale adjustment versus modeling". *Stat Methods Med Res*.
- Payne EH, Hardin JW, Egede LE, Ramakrishnan V, Selassie A, Gebregziabher M (2016a). "An Empirical Approach to Determine a Threshold for Identification of Overdispersion in Count Data". Under review.
- Payne EH, Hardin JW, Egede LE, Ramakrishnan V, Selassie A, Gebregziabher M (2016b). "Latent transition multiple imputation for missing data in time varying categorical variables". Under review.
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. London.
- Ramakrishnan, V. and D. Meeter (1993). "Negative binomial crosstabulations, with applications to abundance data". *Biometrics*, **49**: 195-207.
- Rathouz, Paul J. (2009). "Fixed effects models for longitudinal binary data with drop-outs missing at random". University of Chicago Technical Report.
- Rigby, R. A., D. M. Stasinopoulos, et al. (2008). "A framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution." *Computational Statistics & Data Analysis* **53**(2): 381-393.
- Rodriguez, G. "Models for Over-Dispersed Count Data." *Generalized Linear Models*. Princeton University, 2015. Web. 14 Oct. 2015.
- Rubin, D. Multiple Imputation for Nonresponse in Surveys. Hoboken, NJ: John Wiley & Sons Inc, 1987.
- Ruoyan, M. *Estimation of Dispersion Parameters in GLMs with and without Random Effects*. Stockholm University: Mathematical Statistics. 2004.
- Schafer, J. Analysis of Incomplete Multivariate Data. London, UK: Chapman and Hall, 1997a.
- Schafer, J. (1997b). "Imputation of missing covariates under a general linear mixed model." Technical report: Dept. of Statistics, Penn State University.
- Schenker, N and Taylor, JMG (1996). "Partially parametric techniques for multiple imputation". *Computational Statistics and Data Analysis*. **22**(4): 425-446.
- Schwarz, G. (1978). "Estimating the Dimension of a Model." *The Annals of Statistics* **6**(2): 461-464.

- Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am J Epidemiol* **179**(6): 764-74.
- Smith, P.J., and D.F. Heitjan (1993). "Testing and adjusting for departures from nominal dispersion in generalized linear models". *Applied Statistics* **42**: 31-41.
- Stubbendick AL, Ibrahim JG. *Biometrics*. (2003). "Maximum likelihood methods for nonignorable missing responses and covariates in random effects models." **59**(4): 1140-50.
- Tin, A. "Modeling Zero-Inflated Count Data with Underdispersion and Overdispersion". SAS Global Forum, Statistics and Data Analysis; 2008.
- Vaida, F. (2005). "Conditional Akaike information for mixed-effects models." *Biometrika* **92**(2): 351-370.
- Van der Heijden G.J., et al. (2006). "Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example." *J Clin Epidemiol*. **59**(10): 1102-9.
- Ver Hoef, J. M. and P. L. Boveng (2007). "Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?" *Ecology* **88**(11): 2766-2772.
- Verbeke, G. and E. Lesaffre. (1996). "A linear mixed-effects model with heterogeneity in the random-effects population." *Journal of the American Statistical Association* **91**(433): 217–21.
- Vermunt, J., van Ginkel, J., van der Ark, L., Sijtsma, K. (2008). "Multiple imputation of incomplete categorical data using latent class analysis." *Sociological Methodology* **38**(1), 369-397.
- Wan, W.Y., J.S. Chan (2009). "A new approach for handling longitudinal count data with zero-inflation and overdispersion: poisson geometric process model." *Biom J*. **51**(4):556-70.
- White, IR, Carlin JB. (2010). "Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values." *Stat Med*. **29**(28): 2920-31.
- Xia, Y., D. Morrison-Beedy, et al. (2012). "Modeling Count Outcomes from HIV Risk Reduction Interventions: A Comparison of Competing Statistical Models for Count Responses." *AIDS Research and Treatment* **2012**: 1-11.
- Xiaowei Yang, Shoptaw S, Kun Nie, Juanmei Liu, Belin TR (2007). "Markov transition models for binary repeated measures with ignorable and nonignorable missing values." *Stat Methods Med Res*. **16**(4): 347-64.
- Yang, Z., J. W. Hardin, et al. (2007). "Testing approaches for overdispersion in poisson regression versus the generalized poisson model." *Biom J* **49**(4): 565-584.
- Yang Z, Hardin JW, Addy CL (2009). "A score test for overdispersion in Poisson regression based on the generalized Poisson-2 model". *Journal of Statistical Planning and Inference*. **139**(4): 1514-21.

Yau, K. K. W., K. Wang, et al. (2003). "Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros." *Biometrical Journal* **45**(4): 437-452.

Ye, Wen, Xihong Lin, and Jeremy M. G. Taylor (2008). "Semiparametric Modeling of Longitudinal Measurements and Time-to-Event Data—A Two-Stage Regression Calibration Approach". *Biometrics* **64**: 1238–1246

Zhang H, He H, Lu N, Zhu L, Zhang B, Zhang Z, Tang L. A non-parametric model to address overdispersed count response in a longitudinal data setting with missingness. *Stat Methods Med Res.* 2015 May 5.