Title
Addressing Spatial Dependence and Missing Data in Dental Research

Author
Clague, Jason Scott

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Addressing Spatial Dependence and Missing Data in Dental Research

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

by

Jason S. Clague

2019

ABSTRACT OF THE DISSERTATION

Addressing Spatial Dependence and Missing Data in Dental Research

by

Jason S. Clague

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2019

Professor Thomas R. Belin, Chair

Dental data and the way dental data are collected present interesting and challenging statistical issues that can complicate analyses and if not addressed appropriately lead to misleading conclusions. By their nature, dental data suggest the possibility of spatial correlation between neighboring teeth. However, it is not uncommon for statistical methodology assuming independent observations to be employed, which results in unwarranted precision in the inferences drawn from the model. Another obstacle in the analysis of dental data is that teeth can be missing, a pervasive issue that can have an outsized influence on research conclusions. Additionally, collection of dental data on oral health behaviors (OHBs) occurs during dental visits, making the data subject to recall bias. One proposed solution to mitigate recall bias is the use of ecological momentary assessments (EMAs), administered in real time, through surveys on the phone and asking them about recent OHBs. However, this approach inevitably results non-response and consequent missing data, along with questions regarding the optimal number of survey questions per EMA to minimize subject fatigue. Research in Bayesian spatial data analysis (Banerjee, Carlin, and Gelfand, 2014) and missing data (Rubin, 1987; Little and Rubin,

2002) have shown, through application in medical research, the ability of these methods to control for spatially correlated data and use data imputation methods to draw more accurate conclusions when missing data are present. We endeavor to apply and advance these methods in a dental data setting to answer research questions regarding patterns and underlying mechanisms of dental decay in methamphetamine users.

The dissertation of Jason S. Clague is approved.

Vivek Shetty

Sudipto Banerjee

Ronald Brookmeyer

Thomas R. Belin, Committee Chair

University of California, Los Angeles

2019

To Partow for completing this journey with me

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

My experience at UCLA has been positive and impactful thanks to my mentors Dr. Thomas R. Belin of Biostatistics and Dr. Vivek Shetty of Dentistry. Dr. Belin allowed me to explore the field of statistics with autonomy and freedom. Ultimately, this opportunity resulted in my having substantial influence in the material presented in this dissertation. However, this exploration was enhanced by Dr. Belin's expertise and knowledge. Dr. Belin's perspective and philosophical ideas concerning data analysis, study design, and science have had a lasting impact on my own view regarding these topics. Dr. Shetty was also instrumental in my experience. He often would engage me in thoughtful conversations about philosophy, technology, and the future of my field. These talks have shaped my own perception of events and thinking. My affiliation with Dr. Shetty and UCLA Dentistry, with the help of Dr. David Wong and Ms. Muneeza Irfani, aided me in securing outside funding from the NIDCR, a division of the NIH. My research was funded by the T90 training grant and the F31 fellowship.

In addition to my research, I will forever be thankful for my experience as a teaching assistant for Dr. Ronald Brookmeyer. My five years of experience as Dr. Brookmeyer's TA for Biostatistics 100b taught me to deconstruct statistical ideas and explain them in numerous ways to non-mathematical students. This way of thinking has strengthened my own conceptual understanding of statistical topics. Most importantly, I met my wife, Partow, a source of support, inspiration, and motivation during my second year at UCLA. She has been instrumental in keeping me grounded, motivated, and happy during this experience. I could not have accomplished this feat without her as she is my best friend, partner, and love. I would also like to thank my family, David, Luanne, and Angela (Tom, Philo, and Allie as well) for always

supporting me and encouraging me. Additionally, the moral support of my newfound family members, Ali, Minoo, Kazem, Farnaz, and Reza, is something I will always be grateful for. Lastly, I would like to thank my grandfather Donald Cooks for being my role model and sharing his thoughts, experience, and life with me.

VITA

Education:

University of California Los Angeles, Fielding School of Public Health

Los Angeles, CA

PhD Candidate, Biostatistics, GPA: 3.94 Expected 2019

University of California Los Angeles, Fielding School of Public Health

Los Angeles, CA

MS, Biostatistics, GPA: 3.92 2014

New York University, College of Arts and Science

New York, NY

BA, Mathematics, GPA: 3.82 2012

Grants:

2017-2019 Two-Year Predoctoral Fellowship Award - National Institute of Health (Ruth L. Kirschstein Award – F31)

2014 – 2017 Three-year Predoctoral Training Grant for Interdisciplinary Research in Dentistry - National Institute of Health (T-90)

2012 – 2014 Regents Stipend: Awarded to top UCLA graduate students

Travel Awards:

2016 Travel and Conference Award to Attend the 2016 mHealth Training Institute, Mobile-Sensor Data to Knowledge

2015 Travel Award to Present at Association for Psychological Science Convention for Interdisciplinary Research in Dentistry and Psychology, National Institute of Dental and Craniofacial Research

Teaching Awards:

2018 Teaching Assistant of the Year, Fielding School of Public health

2017 Teaching Assistant of the Year, Fielding School of Public health

2016 Teaching Assistant of the Year, Fielding School of Public health

2015 Teaching Assistant of the Year, Fielding School of Public health

2014 Teaching Assistant of the Year, Fielding School of Public health

Published Abstracts:

"A Conditionally Autoregressive-Spatial Model for Clustered Count Data Controlling for Two Classes of Neighbor Relations" Joint Statistical Meeting, 2017 (Clague J, Belin T, Shetty V)

"Machine Learning Algorithms Identify Covert Methamphetamine Users from Caries Patterns" American Association for Dental Research, 2017 (Clague J, Morrison D, Kotlerman S, Roychowdhury V, Shetty V)

1 Introduction

Dental data and the way dental data are collected present interesting and challenging statistical issues that can complicate analyses and if not addressed appropriately lead to misleading conclusions. By their nature, dental data suggest the possibility of spatial correlation between neighboring teeth. However, it is not uncommon for statistical methodology assuming independent observations to be employed, which results in unwarranted precision in the inferences drawn from the model. Another obstacle in the analysis of dental data is that teeth can be missing, a pervasive issue that can have an outsized influence on research conclusions. Additionally, collection of dental data on oral health behaviors (OHBs) occurs during dental visits, making the data subject to recall bias. One proposed solution to mitigate recall bias is the use of ecological momentary assessments (EMAs), administered in real time, through surveys on the phone and asking them about recent OHBs. However, this approach inevitably results non-response and consequent missing data, along with questions regarding the optimal number of survey questions per EMA to minimize subject fatigue. Research in Bayesian spatial data analysis (Banerjee, Carlin, and Gelfand, 2014) and missing data (Rubin, 1987; Little and Rubin, 2002) have shown, through application in medical research, the ability of these methods to control for spatially correlated data and use data imputation methods to draw more accurate conclusions when missing data are present. We endeavor to apply and advance these methods in a dental data setting to answer research questions regarding patterns and underlying mechanisms of dental decay in methamphetamine users.

1.1 Controlling for Correlated Dental Data Having Multiple Neighbor Relations Using Spatial Random Effects with a Conditionally Autoregressive Prior Distribution

Dental caries data, where each tooth is a count variable expressing the number of decayed, missing, or filled surfaces (DMFS index), exhibit a complex and patterned correlation matrix where cross-mouth and neighboring teeth exhibit strong correlations. Historically, researchers have often used models that assume independence between teeth or employ data aggregation methods to simplify analysis, such as mouth-wide counts of decayed, missing, and filled surfaces (DMFS) or counts of decayed, missing, and filled teeth (DMFT). These approaches finesse the correlation structure within the mouth but might represent an undesirable oversimplification. To account for this neighbor correlation, models with spatial random effects can be useful. Specifically, neighbor relations among teeth can be accounted for using a conditionally autoregressive (CAR) prior distribution on the spatial random effects (Besag et al., 1991). To fit models incorporating CAR structure, posterior sampling will be done utilizing Hamiltonian Mote Carlo (HMC) techniques (Hoffman and Gelman, 2014) in STAN open source software.

1.2 Missing Data Methods for Spatially Correlated Dental Data

Missing teeth, which can occur for a number of reasons, present challenges for modeling DMFS index data. A standard approach is to assign a maximal DMFS index of 4 or 5 because all surfaces of the tooth are marked as missing. However, this rating gives missing teeth an outsized influence on model parameter estimates and may not be on the same scale as decayed and filled teeth, which only occasionally exceed a DMFS index of 3. A possible reason for decayed and

filled teeth often not achieving a DMFS index of 4 or 5 might be because the tooth becomes missing after 2 or 3 surfaces are decayed or filled. This demonstrates a shortcoming of the DMFS index measure to represent dental decay. An alternative strategy is to treat missing teeth as missing data and utilize imputation strategies that allow for spatial stratification to accommodate the observed decay and missingness patterns in the mouth. From previous research, it appears that dental decay is both local and has a tendency to be symmetric. Decay often starts in the molars when decay throughout the mouth is mild, progressing to the premolars as the mouth wide decay is more moderate, and reach the front teeth when dental decay is severe. However, there are individuals who have unusual occurrences of missing teeth accompanied by minimal decay that does not adhere to a prevalent pattern. Such discrepancies suggest that the teeth might be missing due to different underlying mechanisms, pointing to a role for different imputation models. We propose use of a pattern-mixture model specifically designed for the observed spatial stratification, progression levels of decay and missingness, and general patterns of missing data observed in the mouth (Rubin, 1977a; Little, 2002). Additionally, we will develop models that assume ignorability and non-ignorability of the missing data to assess the robustness of the conclusions drawn. Posterior sampling will be performed using Just Another Gibbs Sampler (JAGS) open source software developed by Martyn Plummer and publicly released in 2003.

1.3 Orthogonal Array Study Designs Used for Ecological Momentary Assessments in Mobile Health with Missing Data

Dental data, including specific oral health behaviors (estimated frequency of monthly brushing/flossing/soda consumption over the last year), are often collected during routine dental visits. Requiring the subject to recall such granular data make these data subject to recall bias. Such bias can have a substantial effect on parameter estimates and resulting conclusions drawn from statistical models. To mitigate recall bias, dental researchers are experimenting with mobile health technologies, which allow for ecological momentary assessments, in real time, by administering questionnaires using the subject's phone. For the sake of limiting response burden, we will use the orthogonal array study design (Taguchi, 1987) in order to limit the sample size while attempting to control for main effects and interaction terms. Missing data and non-response are inevitable and due to the specific study design, the missing-data patterns will be limited. We propose using a pattern-mixture model (Little, 1992) to accommodate and leverage unique aspects of our orthogonal array study design in order to impute missing values and ascertain accurate inference (Rubin, 1977a). Other missing data models will be implemented to control for interaction terms and compared to the specially designed pattern-mixture model. All posterior sampling will be performed in JAGS open source software.

2 Background and Literature Review

Some background information on dental and statistical topics is necessary for context before proceeding to development and application of our methodologies. This chapter will be divided into four subsections, in section 2.1 we will review and discuss dental data, the challenges encountered in these data, and the models often used in the dental literature. In section 2.2 we then offer an introduction to the CAR prior distribution by starting from its connection to Markov random fields, lattice data, and dental data as lattice data in an effort to control for the complex correlation structures in these data. HMC sampling techniques will also be discussed. Next, in section 2.3, we provide an overview of issues stemming from missing teeth, which are abundant in our data, and methods to better model data that have missing teeth. We then offer an introduction to how mobile health technology can assist dental research in mitigating recall bias when collecting data on OHBs, possible study designs to help reduce missing data, and statistical methodology that may be useful to analyze these data in section 2.4.

2.1 Dental Data and Statistical Models Used in Dental Research

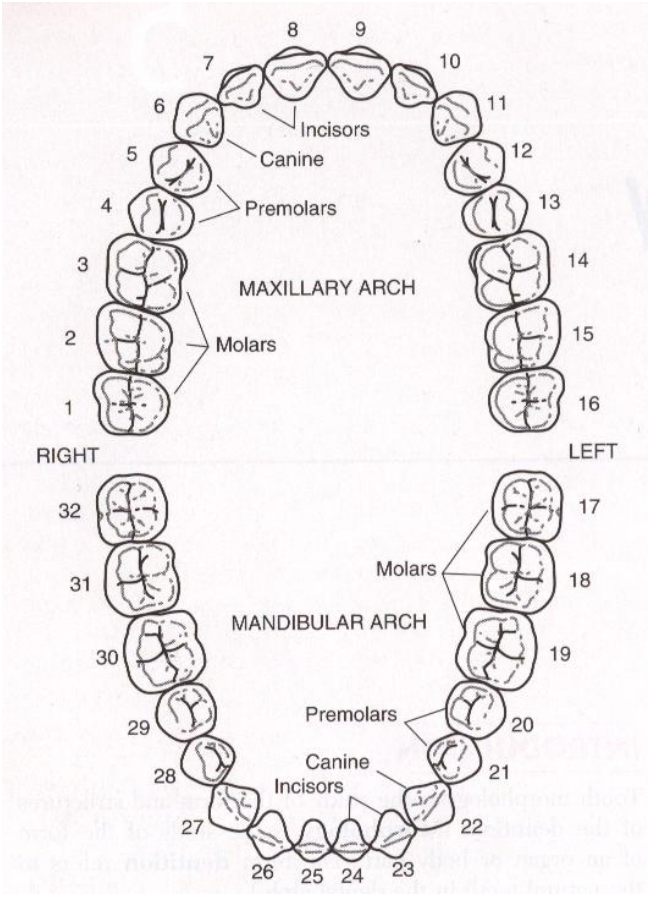
Dental data present several obstacles that make it difficult to model, which include strong correlation among local neighboring teeth, correlation among cross-mouth teeth, missing teeth, and the DMFS index measure used to represent the dental health of a tooth. Observed correlation in dental data are not often accounted for or are avoided entirely through aggregating the data using averages or total mouth counts of DMFS. This section will furnish an introduction to dental data, the aforementioned issues in these data, and the methodology commonly used to

analyze these data as a starting point before exploring methodology to account for these challenges.

2.1.1 Introduction to Dental Data and DMFS Index

In the collection of dental data, individual teeth are numbered 1 through 32 starting with the molars on the upper right, cycling through all of the upper teeth then cycling back through the lower teeth starting on the left and ending on the right, with four tooth types are represented: molars (1-3, 14-16, 17-19, and 30-32), premolars (4-5, 12-13, 15-16, and 28-29), canines (6, 11, 22, and 27), and incisors (7-10 and 23-26). The numbering scheme can be viewed in Figure 1:

Figure 1: Tooth Numbering Scheme



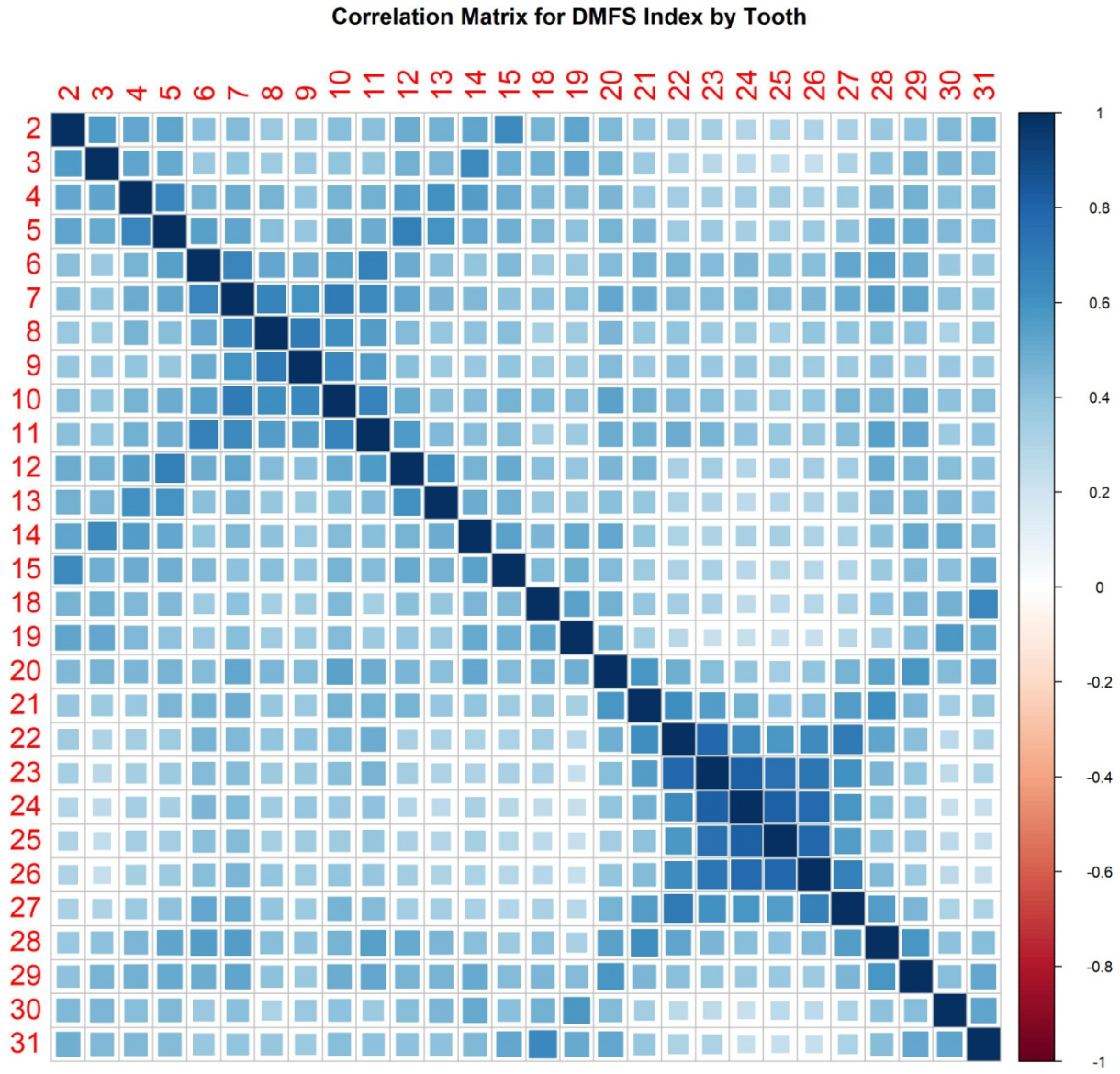
(Taken from https://buildinggreatsmiles.com/blog/what-tooth-number-is-this-tooth/)

It is common for the third molars (1, 16, 17, and 32), also known as wisdom teeth, to be removed from the dataset as they are often missing, which results in 28 teeth, or outcomes, per subject. Data are then collected on each tooth surface, where molars and premolars have 5 surfaces and canines and incisors 4. These data are indicator variables, which are equal to 1 if the surface is decayed, missing, or filled. This measure is called DMFS (Bödecker, 1939; Darby and Walsh, 2003). Note that missing teeth, if all surfaces are summed (DMFS index), receive a maximal DMFS index of 4 or 5 depending on tooth type due to all surfaces of the tooth being recorded as missing. In contrast, a tooth with two decayed surfaces yields a DMFS index of 2. This

difference makes clear that missing teeth can be influential, suggesting a need for special consideration that is not commonly incorporated in the dental literature. Additionally, in our data, decayed and filled teeth often do not reach a DMFS index above 3. This may imply that the tooth becomes missing when 2 or 3 surfaces are decayed/filled. This suggests that decayed/filled teeth are on a different scale than missing teeth even though both use the DMFS index measure (decayed/filled teeth measured from 0 to 3 and missing teeth from 4 to 5). The consequences of this have not received substantial attention in the literature, pointing to a need for further review as well as methodology that accounts for the influence attributed to missing teeth.

The treatment of the outcome variable has profound implications regarding which statistical methods should be used in an analysis. The primary determination is to either use mouth level data, where information is lost due to averaging or summing over teeth resulting in total number of DMFS or DMFT in each subject's mouth, or intra-oral data, which is obtained by summing total DMFS for each tooth (DMFS index) resulting in 28 count variables for each subject. Intra-oral data provide more information, but result in correlation among neighboring teeth (Todem, 2012). The resulting correlation in intra-oral data, to be described in more detail later, can be seen in Figure 2:

Figure 2: Correlation Matrix



Noting that the third molars (1, 16, 17, and 32) are not included, the red numbers represent the corresponding tooth number in Figure 1 and the blue coloration represents positive correlation between the teeth using DMFS index. From this correlation matrix, there is evident complexity in the relationships that are present. There is a local correlation between direct neighbors, teeth on the same jaw, and a tooth type correlation (molars, premolars, canines, and incisors).

Additionally, there is a symmetric correlation, evidenced in the correlation matrix by the seemingly orthogonal lines drawn through the main diagonal line, where the cross-mouth tooth, on the same jaw, exhibits a strong correlation. For instance, tooth 2's cross-mouth neighbor is tooth 15. From these observed patterns, it is clear that it would be highly questionable to assume independence across teeth and that care should be taken when modeling intra-oral data.

2.1.2 Statistical Methods for Mouth Level Data

Mouth level data, expressed as a count variable, represents total DMFS of the subject's mouth or the total teeth experiencing a DMFS event present in the subject's mouth (DMFT). When represented as a count variable, it is analyzed as either a bounded or unbounded counts. For unbounded counts, natural starting points would be to assume the underlying distribution to be either Poisson or Negative-Binomial. The Negative-Binomial is considered due to the overdispersion observed in dental data, which, if not controlled for, can lead to inaccurate estimates of standard errors and changes in deviance associated with model terms that will be too large (Hinde and Demetrio, 1998). For bounded counts, the binomial distribution is used (Hall, 2000). It should be noted that the Poisson and Negative-Binomial provide a good approximation to the binomial model when the number of trials is sufficiently large. Additionally, the binomial regression model is susceptible to overdispersion.

2.1.3 Statistical Methods for Intra-Oral Data

Intra-oral data include tooth (DMFS index) and tooth-surface level (DMFS) data. Such granular data provide information regarding the symmetry of dental decay patterns in the mouth, whether tooth type (molars, premolars, canines, and incisors) and tooth surfaces (facial, lingual, occlusal,

mesial, distal, and incisal) develop dental caries with the same frequency and exhibit unique correlations within and between teeth. The analysis of these data point to a need for the statistical model to account for this more complex correlation structure in order to yield accurate inferences, making it natural to consider generalized linear mixed effects models and generalized estimating equation models for the analysis (Todem, 2012). If tooth level data are used, DMFS index, each tooth represents a count variable, which enables the previously discussed count distributions to be used as the likelihood distribution. However, analyses of tooth-surface level data are binary and suggest the relevance of regression models like logistic or probit regression. In our work, the DMFS index measure will be used for scientific reasons as we are more interested in relationships between teeth rather than surfaces. Additionally, DMFS level data create a host of issues with identifiability and computational complexity. Analysis of periodontal pockets, gum tissue, encounters many of these same obstacles as each tooth has four pockets. Identifiability issues and modeling complexity create significant obstacles and scientific inference is based on individual pockets opposed to entire teeth (Reich, Hodges, and Carlin, 2012; Reich, Bandyopadhyay, and Bondell, 2007).

2.2 Spatial Statistics: Conditionally Autoregressive Prior Distribution

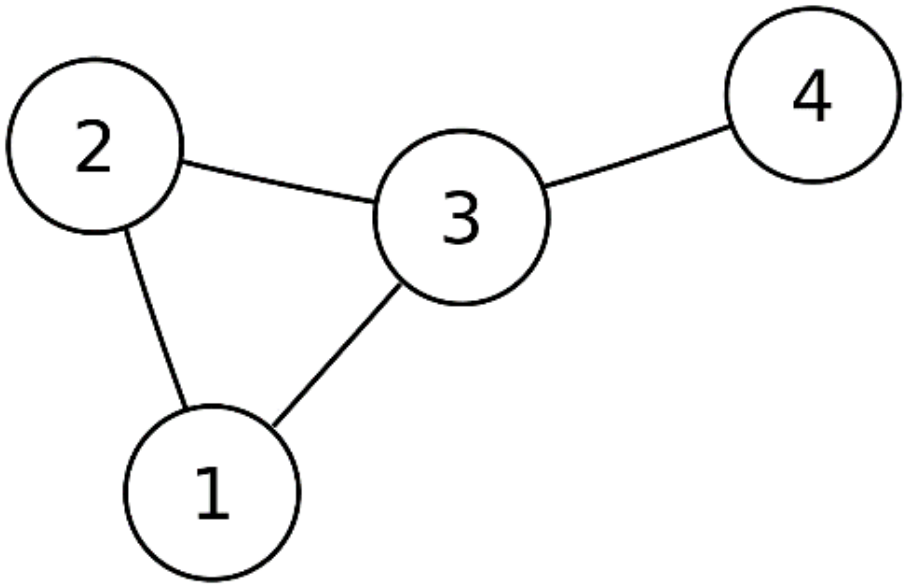
Intra-oral dental data contain many complex relationships between teeth. Two such relationships, discussed previously, include the correlation between adjacent and symmetric, or cross-mouth, neighboring teeth. Methodology often employed to analyze dental data either assumes independence, which ignores the spatial relationships present among teeth, or accounts for some

correlation, but fails to control for more nuanced spatial patterns (Todem, 2012). One possible solution to account for the local neighbor and cross-mouth symmetry correlation observed in dental data is the use of spatial random effects with a conditionally autoregressive (CAR) prior distribution. The CAR prior distribution includes a smoothing component where teeth designated as neighbors have their DMFS index values influence the models estimates. The neighbor smoothing of the CAR prior distribution might help make outliers less influential, which, in a dental setting, would possibly limit the influence of missing teeth. However, to justify this, we must show how tooth level dental data can be modeled as a Markov random field, which can then be expressed through the CAR prior distribution.

2.2.1 Introduction to Markov Random Fields

An undirected graphical model is a multivariate distribution coupled with an undirected graph that encodes conditional independence relations implied by this distribution (Wasserman, 2013). Dental data can be modeled as an undirected graphical model where, within each subject's mouth, the teeth are a random vector T = $(T_2, \ldots, T_{15}, T_{18}, \ldots, T_{31})$ with a multivariate distribution P_T. To give them a conditional independence structure, the teeth are represented in an undirected graph by a set V of vertices and a set of E edges where pairs of vertices are connected if they are designated as neighbors. More formally, tooth i and j, both vertices, are adjacent, written as $T_i \sim T_j$, if there is an edge between them. The resulting collection of edges and vertices, G = (V, E), is the undirected graph. The formal definition of a graph is a set of objects, called vertices or nodes, which are connected together, where all the edges are bidirectional. An example of an undirected graph can be seen below in Figure 3:

Figure 3: Undirected Graph



Undirected graphical models are also referred to as Markov random fields or Markov networks, due to the Markov properties. These Markov properties represent conditional independence relations among the random variables and presented below:

1. Pairwise Markov Property: nodes i and j are independent given all other nodes:

2. Local Markov Property: node i is independent from all other nodes given its neighbors (Ne):

3. Global Markov Property: If the graph is broken up into three disjoint subsets of vertices A, B, and C such that C separates A and B, then the vertices in A are conditionally independent from those in B given C:

Where the global Markov property implies the local Markov property and the local Markov property implies the pairwise Markov property. However, the pairwise Markov property implies the global Markov property, making all three properties equivalent, only if the assigned distribution, P_T, does not assign zero probability to any assignment of the variables. This allows for easier construction of graphs based on pairwise relationships, which tend to be more intuitive. Global conditional independence statements can be extracted after it has been constructed. The following two sections cover Brook's Lemma and the Hammersley-Clifford Theorem and provide methods for retrieving the global joint distribution given the local conditional relations.

2.2.2 Brook's Lemma

Brook's Lemma demonstrates that given the full conditional distributions, $p(T_i|T_j, j \neq i)$ for i = 1, … ,n, the joint distribution, $p(T_1, \ldots, T_n)$, is uniquely determined (Brook, 1964). It allows for retrieval of the unique joint distribution determined by the full conditional distributions (Banerjee, Carlin, and Gelfand, 2014). More mathematically, Brook's Lemma notes that with strictly positive probability or $p(T) > 0$:

$$p(T_1, \ldots, T_n) = \frac{p(T_1|T_2, \ldots, T_n)}{p(T_{10}|T_2, \ldots, T_n)} \frac{p(T_2|T_{01}, T_3, \ldots, T_n)}{p(T_{20}|T_{01}, T_3, \ldots, T_n)} \cdots \frac{p(T_n|T_{10}, \ldots, T_{n-1,0})}{p(T_{n0}|T_{10}, \ldots, T_{n-1,0})} p(T_{10}, \ldots, T_{n0})$$

Where $T_0 = (T_{10}, \ldots, T_{n0})^T$ is any is any fixed point in the support of $p(T_1, \ldots, T_n)$, which means that $p(T_1, \ldots, T_n)$ is determined by the full conditional distributions and, due to $T_0 = (T_{10}, \ldots, T_{n0})^T$ being a constant, is determined up to a proportionality constant. Care must be taken to ensure that the full conditional distributions are compatible (Arnold and Strauss, 1991) and to assess if the joint distribution is proper even when the full conditional distributions are proper (Casella and George, 1992). Simplification of the full conditional distributions,

$p(T_i|T_j, j \neq i) = p(T_i|T_j : j \in Ne(i))$ where the set Ne refers to the neighbors of i, can be achieved using the Markov properties discussed previously, the local Markov property in particular, if we have a Markov random field.

2.2.3 The Hammersley-Clifford Theorem

The Hammersley-Clifford Theorem (Besag, 1974; Clifford, 1990) facilitates a connection between Markov random fields and the joint distribution. Additionally, it has been shown that this theorem allows the full conditional distributions to be specified locally, which means it is necessary to condition on local variables as opposed to all variables (Cressie, 1993). To represent $p(T_1, \ldots, T_n)$, the joint distribution of the random vector of teeth, it must be observed that there is no topological ordering associated with the undirected graph, which means the chain rule cannot be used for conditional probability distributions (Murphy, 2012). Instead, potential functions are associated with each maximal clique in the graph. A clique is a set of vertices such that each vertex is a neighbor of every other element, and a maximal clique is a clique such that the addition of another vertex makes it no longer a clique (Banerjee, Carlin, and Gelfand, 2014). For instance, in Figure 3 the maximal cliques are (1, 2, 3) and (3, 4). The potential function of order k for clique c, denoted as $\varphi_c^k(T_c|\theta_c)$, can be any non-negative function of k arguments that is exchangeable in these arguments. Next, it is important to define a Gibbs distribution: $p(T_1, \ldots, T_n)$ is a Gibbs distribution if it is a function of the T_i only through potentials on cliques, which is represented as

$$p(T_1, \ldots, T_n) \propto \exp\left(\gamma \sum_k \sum_{\alpha \epsilon M_k} \varphi^{(k)}(T_{\alpha 1}, \ldots, T_{\alpha k})\right)$$

where $\varphi^{(k)}$ is a clique or order k, M_k is the collection of all subsets of size k from $\{1, 2, \ldots, n\}$, α indexes this set, and $\gamma > 0$. The Gibbs distribution can also be written:

$$p(T|\theta) = \frac{1}{Z(\theta)} exp\left(-\sum_{c} E(T_c|\theta_c)\right)$$

where $E(T_c) > 0$ is the energy associated with the variables in clique c. High probability states correspond to low energy configurations (Murphy, 2012). The importance of the Gibbs distribution lies in the Hamersley-Clifford Theorem: A positive distribution $P(T) > 0$, known as the positivity condition (Hammersley and Clifford, 1971), satisfies the conditional independence properties (Markov properties) of an undirected graph G if and only if p can be represented as a product of potential functions, one per maximal clique:

$$p(T|\theta) = \frac{1}{Z(\theta)} \prod_{c \epsilon C} \varphi_c(T_c|\theta_c)$$

$$Z(\theta) \triangleq \sum_{T} \prod_{c \epsilon C} \varphi_c(T_c|\theta_c)$$

Essentially, a Markov random field's joint distribution can be written as a Gibbs distribution (Koller and Friedman, 2009), and a Gibbs distribution is a Markov random field (Geman and Geman, 1984). Note that maximal cliques can be used in the potentials or, to restrict complexity, pairwise or individual groupings can be used. Commonly, for simplicity, instead of using maximal cliques, k is set to equal either to k = 1, known as an independence model, or to k = 2, called a pairwise Markov random field. Spatial structure is achieved when k = 2, where a vertex and its neighbor are coupled in a potential function.

2.2.4 Gaussian Markov Random Field

Assigning p($T_1, \ldots, T_n$) as a multivariate normal distribution is a common choice due to its familiarity and computational convenience. In order for the multivariate normal distribution, with precision matrix Q and $\mu = 0$, to accommodate the conditional independence relation implied from the graph G (Rue and Held, 2005):

$$p(T) \propto |Q|^{\frac{1}{2}} exp\left(-\frac{1}{2} T^T Q T\right)$$

$$T_i \perp T_j | T_{-ij} \; if \; and \; only \; if \; Q_{ij} = 0$$

which leads to the following definition:

A random vector T = $(T_1, \ldots, T_n)^T$ is called a Gaussian Markov random field (GMRF) with respect to the graph G = (V, E) with mean μ and precision matrix Q if and only if its density has the form

$$p(T) = (2\pi)^{-\frac{n}{2}} |Q|^{\frac{1}{2}} exp\left(-\frac{1}{2}(T-\mu)^T Q (T-\mu)\right)$$

and

$$Q_{ij} \neq 0 \text{ if and only if } \{i,j\} \in E \; for \; all \; i \neq j$$

Additionally, the normal distribution satisfies the positivity condition, which makes all the aforementioned Markov properties equivalent and allows for the use of Brook's Lemma and the Hammersley-Clifford Theorem. A pairwise GMRF uses the following potentials for the Hammersley-Clifford Theorem (Murphy, 2012):

The joint distribution:

$$p(T|\theta) \propto \exp\left((Q\mu)^T T - \frac{1}{2} T^T Q T\right)$$

can be rewritten as:

$$p(T|\theta) \propto \prod_{i \sim j} \varphi_{ij}(T_i, T_j) \prod_j \varphi_j(T_j)$$

$$\varphi_{ij}(T_i, T_j) = exp\left(-\frac{1}{2} T_i Q_{ij} T_j\right)$$

$$\varphi_j(T_j) = exp\left(-\frac{1}{2} Q_{jj} T_j^2 + Q_{jj} \mu_j T_j\right)$$

If $Q_{ij} = 0$, then there is no pairwise term connecting i and j. Using the factorization theorem, the Markov conditional independence relations hold. The above choice of potential is intuitive given the following factorization based on assuming $\mu = 0$:

$$p(T) \propto \exp\left(-\frac{1}{2} T^T Q T\right) = exp\left(-\frac{1}{2}\left(\sum_j Q_{jj} T_j^2 + \sum_{i \sim j} Q_{ij} T_i T_j\right)\right)$$

$$= \prod_{i \sim j} exp\left(-\frac{1}{2} T_i Q_{ij} T_j\right) \prod_j \exp\left(-\frac{1}{2} Q_{jj} T_j^2\right)$$

2.2.5 Conditionally Autoregressive Prior Distribution

Utilizing the aforementioned graphical models, conditionally autoregressive (CAR) models provide a natural way of accounting for spatial correlation. The CAR specification is a graphical model and assumes a Markov Random Field model where the conditional distribution of each unit is dependent on predefined neighbors. The CAR model was introduced by Besag (1974 and 1991) where the compatible full conditional distributions, in the zero-centered Gaussian case, are specified:

$$\varphi_i | \varphi_j, j \neq i \sim N\left(\sum_{j=1}^{n} b_{ij} \varphi_j, \tau_i^2\right)$$

where τ_i^2 is a spatially varying precision parameter and $b_{ii} = 0$. Using Brook's Lemma, we can rewrite these full conditional distributions as a joint distribution (Gelfand and Vounatsou, 2003):

$$f(\varphi|\tau) \propto \exp\left\{-\frac{1}{2}\varphi^T D^{-1}(I-B)\varphi\right\}$$

the joint density represents a Gaussian kernel with mean 0 and covariance matrix $\sum = (I -$ $B-1D$ where $B=bij$, $bii=0$, and D a diagonal matrix with $Dii=\tau i2$. This joint density can be rewritten, given a spatial proximity matrix or adjacency matrix W where $W_{ij} = 1$ when i and j are neighbors and 0 otherwise, but assignment of other values to W_{ij} are done in practice:

$$f(\varphi|\tau) \propto \exp\left(-\frac{1}{2\tau^2}\varphi^T(D_w - W)\varphi\right)$$

where D_w is a diagonal matrix with $(D_w)_{ii} = w_{i+} = \sum_j w_{ij}$.

To make the covariance matrix symmetric requires:

$$\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2}$$

If $b_{ij} = \frac{w_{ij}}{w_{i+}}$ and $\tau_i^2 = \frac{\sigma^2}{w_{i+}}$, then the condition is satisfied and the full conditional distributions take the form:

$$\varphi_i|\varphi_j, j \neq i \sim N\left(\sum_j \frac{w_{ij}}{w_{i+}}\varphi_j, \frac{\sigma^2}{w_{i+}}\right)$$

which results in spatial smoothing determined by a weighted average of the neighboring values determined by the W matrix. This neighbor weight matrix, W, is highly flexible and influential in the strength and units that contribute to the spatial smoothing (Earnest, et al., 2007). However, $(D_w - W)1 = 0$, which implies that the covariance matrix is singular and the distribution is improper. The improper specification is widely used as a prior distribution for spatial random effects because it gives an intuitive interpretation where each unit is smoothed by the average of its neighbors (Banerjee, Carlin, and Gelfand, 2014). However, if a proper prior distribution is

desired, that can be accomplished by changing the covariance matrix so that $D_w - \rho W$ where $\rho \in \left(\frac{1}{\lambda_1}, \frac{1}{\lambda_n}\right)$ and λ_i are the ordered eigenvalues of $D_W^{-\frac{1}{2}} W D_W^{-\frac{1}{2}}$(Cressie, 1993; Sun, et al., 2000). However, we can relax this range to $|\rho| < 1$ using the results of the Gershgorin disk theorem (Golub and Van Loan, 2012; Horn and Johnson, 1990) and its application to a specific covariance matrix (Carlin and Banerjee, 2003). The full conditional distributions now take the form:

$$\varphi_i | \varphi_j \, , j \neq i \sim N\left(\rho \sum_j \frac{w_{ij}}{w_{i+}} \varphi_j, \frac{\sigma^2}{w_{i+}}\right)$$

where spatial smoothing is now reflected by a proportion, ρ, of the weighted averages of the neighboring values. With these changes, the model can now be applied to spatially correlated data. The joint distribution, is now proper, is given by:

$$\varphi \sim \mathrm{N}(0, [\tau D(I - \rho B)]^{-1})$$

Use of a CAR prior distribution on spatial random effects influences the fixed effect means and variance inflation factors (Reich, Hodges, and Zadnik, 2006). The CAR model, when specified with a Gaussian distribution, is a Gaussian Markov Random Field, as discussed in the previous section (Rue and Held, 2005). There also exist non-Gaussian data that exhibit spatial correlation. In these situations, with binary or categorical outcomes, CAR priors are still useful.

2.2.6 Application of CAR model to Dental Caries Data

Bandyopadhyay, Reich, and Slate (2011) utilized a CAR model to analyze data from N = 100 subjects who were Type-2 diabetic Gullah-speaking African Americans. In this analysis, teeth,

the outcome variable, was treated as a count variable using the DMFS index measure. To account for possible overdispersion relative to a Poisson model, a Beta-Binomial model with spatial random effects was used, leading to the following framework for person i and tooth s:

$$(\varphi_i(1),\ldots,\varphi_i(n))^T \sim CAR(\rho,\tau)$$

$$logit\big(\mu_i(s)\big) = X_i^T\beta + \varphi_i(s)$$

$$p_i(s) \sim Beta(\theta\mu_i(s),\theta[1-\mu_i(s)])$$

$$y_i(s) \sim Binomial(n_s,p_i(s))$$

The CAR prior distribution for the spatial random effects is proper; incorporating an adjacency matrix was specified to denote the tooth directly above/below and the locally adjacent teeth neighbors. Model comparisons were performed using DIC between Binomial and Beta-Binomial models with and without spatial random effects. Models with spatial random effects consistently outperformed those without and the Beta-Binomial regression models also outperformed Binomial regression models. Inclusion of spatial random effects, on average, resulted in a decrease in DIC by approximately 1,000. The Beta-Binomial models outperformed their analogous Binomial model by a 10,000 DIC difference (decrease). Large decreases in DIC suggest that the Beta-Binomial model's ability to control for overdispersion and spatial random effects with CAR prior distributions may be useful in modeling dental data.

2.2.7 Alternative Models for Spatial Data

Using spatial random effects with a CAR prior distribution can offer a simple way to control for nuanced patterns of correlation. However, these models sometimes experience convergence, identifiability, and computational issues. Alternative models will be considered and run for comparison purposes. One of these models will be a Hierarchical Bayesian Model that allows for

high-order, spatially stratified interaction terms with prior distributions that increase shrinkage to 0 as the order of the interaction term increases. The other model is a Smoothed Analysis of Variance (SANOVA) that also allows for high-order, spatially stratified interaction terms, but implements a different method for shrinking coefficient estimates. Both models appear to have better convergence, identifiability, and computational efficiency than the models using spatial random effects with a CAR prior distribution.

The Hierarchical Bayesian model was proposed by Rubin, Schafer, and Schenker while working with the U.S. Census Bureau's post-enumeration survey (Rubin, Schafer, and Schenker, 1988). In this analysis, they developed a pattern-mixture model to permit high-order interaction terms for multinomial categorical data:

$$\log(\theta_{ijk..p}) = \mu + \mu_{1(i)} + \mu_{2(j)} + \cdots + \mu_{8(p)} + \mu_{12(ij)} + \mu_{13(ik)} + \cdots + \mu_{123..8(ijk..p)}$$

where $\theta_{ijk..p}$ is the probability an observation falls in cell ijk…p and the $\mu's$ represent the main effects, one way, two way, three way, and higher-order interaction terms. The prior distributions placed on the interaction terms allow for increased smoothing the higher the order the interaction term:

$$\mu_i \sim N(0, \sigma^2)$$

$$\mu_{ij} \sim N\left(0, \frac{\sigma^2}{\tau}\right)$$

$$\mu_{ijk} \sim N\left(0, \frac{\sigma^2}{\tau^2}\right)$$

.

.

$$\mu_{ijk..p} \sim N\left(0, \frac{\sigma^2}{\tau^7}\right)$$

For $\sigma > 0$ and $\tau > 1$. The τ term is a scalar parameter estimated by computing posterior probabilities over a grid of values and selecting the posterior mean with best model fit (WAIC). This model allows for high-order interaction terms to be included, but smooths them towards 0, with the strength of the smoothing increasing with the order of the interaction term. In the dental setting, the outcome variable is no longer a categorical variable like in the work Rubin, Schafer, and Schenker did, but a count variable. The adaptation is straightforward by maintaining the mean structure and the prior distribution scheme, but changing the link function and the likelihood. The spatial stratification mentioned previously is achieved in the interaction terms. The interaction terms are composed of the main effects and indicator variables for tooth location. These tooth location variables, 14 in total with one dropped as the reference group, are constructed to identify jaw (lower or upper), right or left side, and tooth type (e.g., upper, left molar).

The other alternative spatial model, considered here, the SANOVA model, is based on the Analysis of Variance (ANOVA) model. In ANOVA, the model accounts for variance in the outcome variable using individual factors, expressed as a matrix A_1 having dimensions cn x M where c are the number of cells, n the observations per cell, and M the number of columns for main effects, along with interactions among those factors, which are represented by a matrix

(A_2) with dimension cn X N where N is the number of interaction terms. The design matrix is defined as $X = [A_1|A_2]$ with regression parameters $\left(\frac{\theta_1}{\theta_2}\right)$ yielding a linear model $y = [A_1|A_2]\left[\frac{\theta_1}{\theta_2}\right] + \epsilon$ where ϵ is normally distributed with mean 0 and covariance $\frac{1}{\gamma_0} I_{cn}$.

SANOVA has the same mean structure, but adds a smoothing model (Hodges, Cui, Sargent, Carlin, 2007):

$$0_N = [0_{NxM}|I_N]\left[\frac{\theta_1}{\theta_2}\right] + \delta$$

Where δ is an N-variate normal distribution with mean 0 and diagonal covariance matrix $diag(\varphi_1, \ldots, \varphi_N)^{-1}$. The φ_k determine the smoothing, or shrinkage, of their respective θ_k. Each φ_k need not be distinct and can be used for multiple regression parameters. For instance, a set of distinct constraint case precisions $\{\eta_0, \eta_1, \ldots, \eta_s\}$, $s \leq N$and a deterministic assignment function j(k) yields $\varphi_k = \eta_{j(k)}$. This makes it so the interaction parameters are smoothed by the underlying $\eta_{j(k)}$. Prior distributions on these smoothing parameters determine the degree of shrinkage. A gamma prior distribution is placed on the shrinkage parameters. The spatial stratification is achieved in SANOVA using the same structure as the Hierarchical Bayesian Model. The difference between the two models is the method of smoothing the interaction terms.

2.2.8 Model Selection Information Criteria

It can be desirable to estimate the predictive accuracy of models, while correcting for the bias in the predictive accuracy measure, and to compare them based on this measure for model selection purposes. Commonly used information criteria, including AIC (Akaike, 1973), DIC

(Spiegelhalter et al., 2002; van der Linde, 2005), and WAIC (Watanabe, 2010), will be reviewed in this section. Each attempts to estimate within-sample predictive accuracy while adjusting for bias. The Akaike Information Criterion (AIC), for a model with k parameters, is calculated:

$$AIC = -2\log p\left(y|\hat{\theta}_{mle}\right) + 2k$$

AIC utilizes the maximum likelihood estimate (MLE) for the parameters in its calculation and its bias correction is intuitive for linear models with flat priors. However, the effective number of parameters changes when more complex hierarchical structures and informative priors are used. The Deviance Information Criterion (DIC) replaces the MLE estimate of the parameters with the posterior mean $\hat{\theta}_{DIC} = E(\theta|y)$ and is calculated using a data-based bias correction:

$$DIC = -2\log p\left(y|\hat{\theta}_{DIC}\right) + 2p_{DIC}$$

$$p_{DIC} = 2\left(\log p\left(y|\hat{\theta}_{DIC}\right) - E_{post}(\log p(y|\theta))\right)$$

p_{DIC} is calculated using simulations θ_s, s = 1,…,S:

$$2\left(\log p\left(y|\hat{\theta}_{DIC}\right) - \frac{1}{S}\sum_{s=1}^{S}\log p(y|\theta_s)\right)$$

By using the mean of the posterior and the data-driven bias correction, DIC represents a more Bayesian version of AIC (Gelman, Hwang, and Vehtari, 2014). However, if the posterior distribution is not summarized well by the mean, DIC can gives nonsensical results. The Watanabe-Akaike Information Criterion (WAIC) avoids using pointwise parameter estimates by computing the log pointwise predictive density (lppd) where $p_{post}(\theta) = p(\theta|y)$:

$$WAIC = -2(lppd - p_{WAIC})$$

$$lppd = log\left(\prod_{i=1}^{n} p_{post}(y_i)\right) = \sum_{i=1}^{n} log\int p(y_i|\theta)p_{post}(\theta)d\theta$$

This expression can be evaluated using draws from $p_{post}(\theta)$, with s = 1, … ,S:

$$lppd = \sum_{i=1}^{n} log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^s)\right)$$

The bias correction is given by:

$$p_{WAIC} = \sum_{i=1}^{n} var_{post}(\log p(y_i|\theta)) = \sum_{i=1}^{n} V_{s=1}^{S}(\log p(y_i|\theta_s))$$

where $V_{s=1}^{S}$ is the sample variance $V_{s=1}^{S} a_s = \frac{1}{S-1}\sum_{s=1}^{S}(a_s - \bar{a})^2$. WAIC has the advantage of averaging over the posterior as opposed to conditioning on a point estimate as AIC and DIC do, which allows it to work with simple and complex hierarchical models having both informative and uninformative priors. Additionally, WAIC performs well when the mean is not a good representation of the posterior. However, WAIC relies on a data partition, which can be difficult for structured models. Despite this drawback of WAIC, Gelman et al. assert that WAIC should still be used in this setting and that WAIC provides the most philosophically Bayesian option of the information criterion (Gelman, Hwang, and Vehtari, 2014). WAIC calculations can be easily computed using the "loo" R package by Vehtari et al. (Vehtari, Gelman, and Gabry, 2016; Vehtari, Gelman, and Gabry, 2017).

2.2.9 Hamiltonian Monte Carlo

With complex models and correlation structures used to model dental data for methamphetamine, challenges such as high dimensionality and correlated parameters arise. These issues, for random walk Metropolis-Hastings and Gibbs sampling algorithms (Metropolis et al., 1953; Geman and Geman, 1984), can lower the effective sample size and make it more difficult to find the stationary distribution. Additionally, the geometric shape of the target distribution's sampling region can be complex, with sharp edges or non-convex regions, which

can result in biased parameter estimates because algorithms used in Bayesian analysis might avoid these regions of the sampling space (Betancourt, 2017). A possible solution to these issues is to utilize Hamiltonian Monte Carlo (HMC) (Neal, 1993, 2011; Duane et al., 1987), which uses the gradient of the log-posterior to place a vector field on the sampling space through the use auxiliary momentum parameters, effectively simulating Hamiltonian dynamics. This hastens convergence to the stationary distribution in high-dimensional situations, reduces the issues presented by correlated parameters, and allows the algorithm to sample difficult geometries of the typical set (areas of interest in the sample space). In HMC, the step size and step count parameters require specification. These parameters have a large influence on the efficiency of the algorithm. The open source HMC software Stan (Stan Development Team, 2013) implements a version of the HMC algorithm called the No-U-Turn sampler (NUTS) (Hoffman and Gelman, 2014). The NUTS sampler runs an HMC algorithm that eliminates the need to specify the number of steps and provides an automated scheme for tuning the step size parameter through dual averaging (Nesterov, 2009). This makes the NUTS HMC algorithm in Stan less prone to user error.

2.3 Missing Teeth

Caries status, a key component of oral health, is often collected as the count of decayed, missing, or filled surfaces (DMFS). Missing teeth, which can occur for a number of reasons, present challenges for modeling DMFS index data. Commonly used fixes, such as aggregating data approaches, treat all surfaces as adding to the total mouth DMFS count, which can result in a loss of information and nuance, or granularity, of scientific conclusions. Inclusion of spatial random effects can assist in accounting for spatial correlation present in the dental data and allow for use

of the DMFS index measure. However, missing teeth and decayed or filled teeth may be the result of separate generative processes, and spatial random effects may not do enough to compensate for this.

From our group's previous work it was observed that teeth that are decayed or filled only reach DMFS index between 0 and 3, whereas missing teeth are given a count of 4 or 5 depending on tooth type (Clague, Belin, and Shetty, 2017). It stands to reason that many teeth become missing before all surfaces are decayed or filled. Allowing the missing teeth a DMFS index that is only occasionally attained by decayed/filled teeth raises questions for interpretation, especially if missing teeth only had a DMFS index of 2 or 3 moments before becoming missing. Such a framework permits missing teeth to have an outsized influence on parameter estimates and to be scored on a different scale even though both decayed/filled and missing teeth use DMFS index. We propose not assigning missing teeth a maximal DMFS index of 4 or 5, but rather treating them as missing. One possible approach would allow an imputation model, based on decayed and filled teeth, to impute values for missing teeth and place them on the same DMFS index scale. Additionally, multiple models will be constructed to account for spatial relations among teeth and different mechanisms for missing data.

2.3.1 Missing Data

When data are missing, the underlying missing data process (missing data mechanism) should be considered, either to be modeled explicitly or implicitly as when the missing data mechanism is assumed to be ignorable. Rubin formalized this notation by setting β to be the parameter for the data and θ the parameter of the missing data, $U = (U_1, \ldots, U_k)$ a vector random variable with

probability density function f_β, and $M = (M_1, \ldots, M_k)$ a vector random variable as an indicator for missingness (Rubin, 1976; Rubin, 2004). The probability that the missing data, M, take the values $m = (m_1, \ldots, m_k)$ given that U take on the values $u = (u_1, \ldots, u_k)$ is $g_\theta(m|u)$, which is the process that causes or generates the missing data where $m_i = 1$ if the data point is missing and 0 otherwise. Thus, the observed data are represented in a vector random variable $V = (V_1, \ldots, V_k)$ where $V_i = u_i$ if $m_i = 0$ and $V_i = *$ if $m_i = 1$. The values of V are what is observed, but it is desired to make inference using the values of U. Imputation techniques can be useful if missigness indicators, M, hide true values that are meaningful for the analysis (Little and Rubin, 2002).

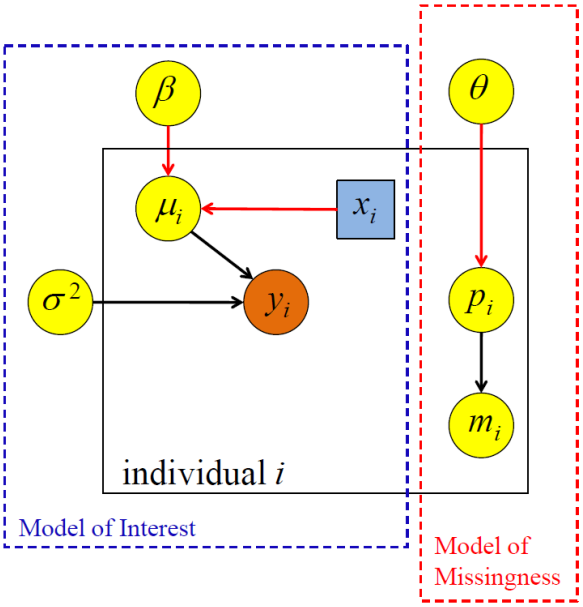
2.3.2 Missing Data Mechanisms

The process that generates the missing data, or the missing data mechanism: $g_\theta(m|u)$, can have several properties that influence the methods used to impute the missing data. These three missing data mechanism assumptions are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Rubin, 1976 and 1977a; Little and Rubin, 2002, 1989; Schafer, 1997). If the missing-data mechanism is MCAR, it means that the missingness does not depend on the values of the data U:

$$g(M|U, \theta) = g(M|\theta) \; for \; all \; U \; and \; \theta$$

This does not necessarily imply that the data pattern is random, but rather that the missingness does not depend on the data. The model shown below is a graphical model for the MCAR assumption, where U = (Y, X) and it is assumed that X is fully observed and Y contain missing values. In Figure 4, y_i is missing, hence the orange coloration of the circle, and x_i is observed, which is signified by the blue shading (Best and Mason, 2012):

Figure 4: Missing Completely At Random

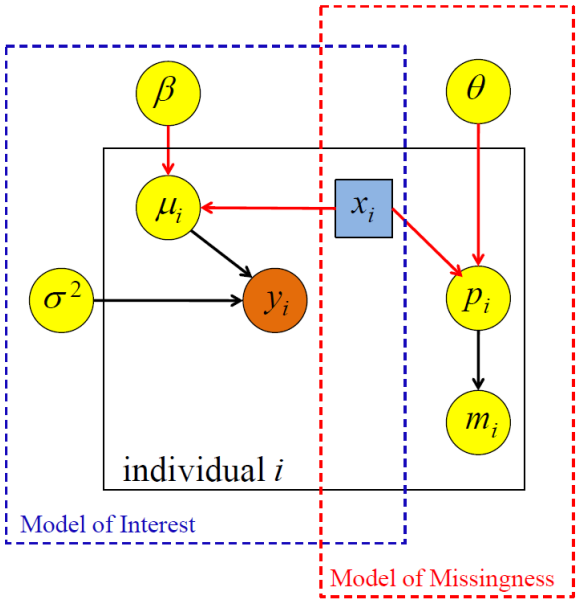


The MAR assumption incorporates for more flexibility, allowing the missingness to depend on the observed values:

$$g(M|U,\theta) = g(M|V,\theta) \; for \; all \; V \; and \; \theta$$

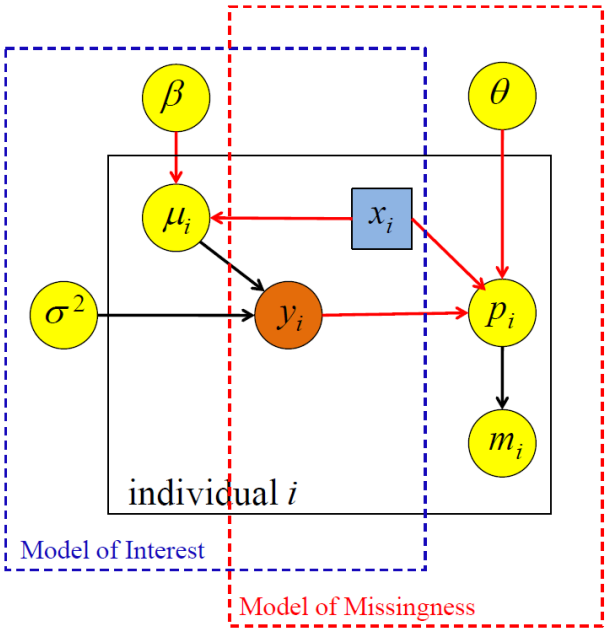
The graphical model representation can be seen in Figure 5 (Best and Mason, 2012):

Figure 5: Missing At Random



If the data are MAR and the parameters in the missingness model and data model are distinct, then the missing data mechanism is deemed ignorable (Rubin, 1976 and 1978), which allows for use of imputation methods that depend on the observed data where the missing data mechanism no longer needs to be modeled for accurate inference. The missingness mechanism is MNAR and considered non-ignorable if the distribution of M depends on the missing values in U (the values of U that were missing), replaced with *, in V so that the missing data mechanism remains as it originally was: $g(M|U,\theta)$. The model set up can be seen in Figure 6 (Best and Mason, 2012):

Figure 6: Missing Not At Random



2.3.3 Methods for Missing Data

Missing data methods will now be introduced where this review will be primarily based on the Little and Rubin (2002) and Schafer and Graham (2002). Missing data methodology can be categorized into four groups:

1. One expedient way to deal with missing data is to discard units that have any missing data, leaving only complete data for the analysis (Nie et. al., 1975). Using only complete units is easy and might work in settings where there is a small amount of missing data. Otherwise, it can lead to bias when data are MAR or MNAR. Thus, using complete units generally only works in the MCAR setting or when there is a small amount of missing data. However, deletion when there is even a modest amount of missing data can lead to a large percentage of units being removed from the data set, which reduces sample size and, thus, statistical power. Pair-wise deletion is also possible, and if the data are MCAR,

it yields more powerful, unbiased estimates (Graham, Hofer, and Piccinin, 1994). However, if there is multicollinearity, the correlation matrix may not be positive definite, giving rise to estimation issues. Additionally, pairwise deletion can result in each correlation having a different sample.

2. Weighting procedures are often times used in survey data without missing data where the units are weighted by their design weights. A theoretical rationale favors use of weights that are inversely proportional to the probability of the unit being selected. Such procedures are adapted for missing data by modifying the weights in an attempt to adjust for the missing data as if it were part of the study design. Implementing weighting procedures can help reduce bias in some non-MCAR situations when complete case analysis is performed.

3. Single imputation methods place a value where there is missing data and the complete data set is then analyzed. A common imputation method is called hot deck imputation, which uses observed units as substitutes for missing values. Another example of hot deck imputation would be mean imputation where means from observed values are substituted for missing values. For longitudinal studies, the last observed value carried forward is sometimes used. Both the mean imputation and last value carried forward can result in bias in estimates of quantities of interest (Carpenter, Kenward, Evans, and White 2004; Cook, Zeng, and Yi 2004; Jansen et al., 2006).

Another method of replacing missing values is regression imputation, which replaces missing values using predicted values from the regression on the observed variables. However, this method has drawbacks as estimated standard errors of the regression coefficients, in ordinary and weighted least squares, of the imputed data will tend to be smaller than would be suggested by actual uncertainty (Little, 1992). Imputation error is defined as the error in the estimates due to uncertainty about the imputed value (Fichman and Cummings, 2003). Little and Rubin suggest that stochastic regression imputation techniques are preferable (Little and Rubin, 2002). However, single imputation still underestimates the standard errors due to variability in imputation.

Using these imputation strategies improves on deletion, but improvement of imputation accuracy does not necessarily result in valid statistical inference (Rubin, 1996). Even if the conditional mean imputation strategies discussed above result in more accurate imputations for missing values, not properly estimating the uncertainty in those imputations and consequent underestimation of standard errors results in rejection rates that are less than or equal to nominal levels (Rubin, 1987 and 1996). With conditional mean imputation, confidence validity is not achieved due to standard errors being too small, which creates issues for valid statistical inference. Additionally, some of the previously discussed imputation methods can distort data distributions and relationships between variables (Rubin, 1987).

4. Multiple imputation is a method that retains the positive aspects of conditional mean imputation, but accounts for imputation uncertainty, which allows for valid inference

(Rubin, 1987). In multiple imputation, each missing value is replaced by one of m values, which results in m complete data sets. These data sets are then analyzed using complete-data methods. It is expected that the results will vary and can be combined to obtain overall estimates and corresponding standard errors that include appropriate missing data uncertainty. Simulation studies document the efficiency of multiple imputation, illustrated by an example where with m = 10 imutations, 50% missing data yields 95% efficiency (Rubin, 1987). It is common for multiple imputation analyses assume that missing data are MAR, although it is possible to utilize multiple imputation in a setting where missing data are assumed MNAR (Glynn et al., 1993; Verbeke and Molenberghs, 2000).

5. Maximum likelihood procedures, particularly using the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977), are often applied to missing-data problems. They work by defining a model for the observed data and utilizing the likelihood (or posterior distribution) under that model to develop inferences for model parameters, potentially using transformations to link parameters of marginal distributions with parameters of conditional distributions (Little and Rubin, 1989 and 2002; Little and Schenker, 1995). These methods provide consistent and efficient parameter estimates under MAR conditions. Given that the MAR assumption is met, other parameter estimates, such as standard errors, are consistent and, additionally, if the model for the complete data is valid, then the marginal distribution of the observed data yields the correct likelihood for the unknown parameters (Little and Rubin, 2002). These properties

are a marked improvement over, conditional mean methods using ordinary least squares, discussed previously.

2.3.4 Multiple Imputation

Due to the extensive use of imputation methods in this dissertation, a closer examination of multiple imputation is indicated. Multiple imputation induces a distribution of possible values of functions of available data (i.e., statistics) by replacing missing values with two or more plausible values (Rubin 1977b; Rubin 1978a). Complete-data methods are then run on these data sets and the estimates are combined using a strategy that can be expected to work reasonably well in large samples. By doing this, the between-imputation variance in quantities of interest can be assessed and included in estimate standard error. Rubin discusses the statistical validity of multiple imputation, where he describes that the validity of resulting estimands relies on estimates of the first two moments of distributions of quantities of interesting being approximately unbiased and that confidence intervals and hypothesis testing must be statistically valid when applied to complete data.

As explained in Rubin (1996), statistical validity, in this context is based on Neyman's definition of confidence intervals, coefficients, and limits (Neyman, 1934). Here Neyman defines confidence limits as statistics defining an interval where, in repetition, the estimand lies in the interval with probability greater than or equal to the confidence coefficient. Shorter intervals are preferable. Randomization validity requires agreement between nominal and actual levels. Thus, it is possible that multiple imputation, being statistically valid, would yield a shorter interval with coverage greater than 95%+, which is preferable to a broader interval with exactly 95%

coverage, which would also have randomization validity. Statistical validity is especially important to consider when thinking about the objectives of missing data problems. The objective is not optimal point prediction or judging an imputation method by its ability to recreate individual missing values, but valid statistical inference.

Multiple imputation inference, for infinite m imputations, can be randomization valid if the complete data inference is randomization valid and the imputation is proper. Rubin (1987) gives a precise definition of proper multiple imputation, characterizing situations where multiple imputations using "Rubin's Rule" for combining estimates yields randomization-valid inferences, which relies on a consistent, asymptotically normal estimator of the unknown parameter and a weakly unbiased estimator of its asymptotic variance in sufficiently regular models. Rubin (1987) concludes that if imputations are drawn as approximate repetitions from a Bayesian posterior of the missing data under the posited response mechanism and an appropriate model for the data, then in large samples the imputation method is apt to be proper. Care must be taken in choosing the covariates in the multiple imputation model. If the outcome variable, Y, is correlated with a variable X that is to be included in the complete data analysis and X is not used in the multiple imputation of Y, then the multiply imputed data will yield estimates of the correlation between X and Y biased towards zero, which can make it improper.

Even if some mildly important variables are left out of the multiple imputation framework, the repeated-imputation inferences can still be confidence valid (statistically valid). Ultimately, it needs to be understood that multiple imputation does not create information, but represents observed information so that complete data tools and methods can be used. Other methodologies

similarly are based on simulating data points, including the jackknife and bootstrap (Miller, 1974; Efron and Tibshirani, 1993), data augmentation (Tanner and Wong, 1987), the Gibbs sampler (Gelfand and Smith, 1990; Gelman and Rubin, 1992), and sampling importance resampling methods (Rubin, 1987 [discussion of Tanner and Wong, 1987]; Smith and Gelfand, 1992). However, multiple imputation can offer advantages over these other methods due to multiple imputation only simulating the missing data where the rest of the inference rests on complete data methods. Multiple imputation also demonstrates improved efficiency of estimation and more accurate inference than single imputation methods with little additional computing (Little, 1988). Ultimately, multiple imputation is more flexible than reweighting or replication for data when complex patterns of nonresponse are present.

2.3.5 Missing Data and Imputation in a Bayesian Setting

In the Bayesian setting, missing data are treated as additional unknown quantities where a posterior distribution can be estimated (Gelman et. al., 2013). Therefore, in the data set where there are missing values, marked as NA, a number of Bayesian approaches are built on replacing the NAs with simulated draws from a predictive distribution. When the missing data are ignorable (MCAR or MAR) it means the posterior distribution is conditionally independent, given the observed data, of the indicator variable for missingness status M. The conditional independence implies that there is no need to model the missing data mechanism $g(M|U,\theta)$. Rather, a complete data model alone would be constructed on the data, including both the missing and observed data, yielding valid regression parameter estimates. In the case where the data are MNAR the target posterior distribution depends on the missing data indicator variable M and jointly modeling the missing data indicator and the data are necessary for valid inference:

$$P(Y_{obs}, Y_{miss}, M|X, \beta, \theta)$$

In both the ignorable and non-ignorable case, how the parameter that replaces the NA in the data set is estimated is clear with a review of the Gibbs sampling algorithm (Gelfand and Smith, 1990) which built on other Markov-chain Monte Calro (MCMC) techniques (Geman and Geman, 1984):

0. Set (x_0, y_0) to a starting value

1. Sample $x_1 \sim p(x|y_0)$ then use this value to sample $y_1 \sim p(y|x_1)$

This results in the sample (x_1, y_1)

.

.

k. Sample $x_k \sim p(x|y_{k-1})$ then use this value to sample $y_k \sim p(y|x_{k-1})$

(x_k, y_k)

In the case of missing data, the simulated values put in place for the NAs would naturally fit into a Gibbs sampling algorithm, being modeled and updated until convergence, effectively imputing a value based on the regression coefficient parameters and the mean structure of the model. It should be noted that when there are missing data, the software package Just Another Gibbs Sampler (JAGS) (Plummer, 2003) automatically generates simulated values in place of the NAs. The software package Stan, however, does not, instead requiring the user to manually specify initial values for missing data and place them in the data. Even though this method in Stan is feasible, modeling with missing data is further complicated due to the program being premised on model parameters being continuous. Thus, if the likelihood distribution is anything other than a continuous distribution, the program will report back an error and not allow a vector of

continuous parameters and non-continuous data to be modeled by a non-continuous likelihood distribution. For analysis where there are missing data and the variable or variables with missing data are not all continuous, Stan is not a viable option for posterior sampling. JAGs offers a simple alternative as it uses Gibbs sampling and can easily accommodate this situation.

2.3.6 Selection Models

As previously mentioned, when the missing data are assumed to be MNAR the missing data indicator variable and the complete data must be modeled together:

$$P(Y_{obs}, Y_{miss}, M|X, \beta, \theta)$$

The joint distribution can be factored into a selection model (Rubin, 1974):

$$P(Y_{obs}, Y_{miss}, M|X, \beta, \theta) = P(Y_{obs}, Y_{miss}|X, \beta)P(M|\theta, X, Y_{obs}, Y_{miss})$$

Where the missing indicator variable, M, is modeled with either a logistic or probit model and the complete data model is the likelihood distribution that best represent the data. This Selection model is convenient due to the parameters of interest being modeled directly and requiring no further transformations or integration: $P(Y_{obs}, Y_{miss}|X, \beta)$.

2.3.7 Pattern-Mixture Models

Building on perspectives developed in Rubin (1976 and 1977) on nonignorable missingness, pattern-mixture models represent an alternative way of factoring the joint distribution of the complete data and the missing data indicator variable (Little, 1992; Little and Rubin, 2002):

$$P(Y_{obs}, Y_{miss}, M = m|X, \beta, \theta) = P(Y_{obs}, Y_{miss}|X, \beta_m, M = m)P(M = m|\theta, X)$$

The indicator variable M can both be an indicator for missing/observed or it can be a categorical variable for pattern of missing data, which will be discussed later and provides additional flexibility if the pattern of missingness is informative to the value of the missing value. To parse the parameters of interest, the patterns of missing data must be integrated out to get the marginal distribution of Y:

$$P(Y_{obs}, Y_{miss}|X, \beta, \theta) = \sum_{m} P(Y_{obs}, Y_{miss}|X, \beta_m, M = m) P(M = m|\theta, X)$$

The pattern-mixture model offers a flexible, hierarchical model structure that allows for specification of a missing indicator variable or a categorical variable to specify missing-data patterns. Each pattern of missing data is modeled separately, estimating its own regression parameters, and imputing the missing data based on those pattern specific regression parameters. The pattern specific parameters are then combined by integrating over the pattern variable, as shown above, to get generalized regression parameter estimates. Modeling for both the pattern-mixture and selection model should be followed by a sensitivity analysis where prior distribution and assumptions are changed or perturbated to assess the sensitivity of the parameters.

2.4 Mobile Health Data and Ecological Momentary Assessments with Missing Data

Collection of dental data, specifically data on oral health behaviors and dental health outcomes, often occurs during a dental visit. This is not ideal in the dental research settings as oral health behavior data are vulnerable to recall bias. Technology may provide a way to mitigate the effects of recall bias through implementation of an ecological momentary assessment (EMA) where study participants are given a phone at the outset of the study, which contains an “app” that

prompts them at predetermined times with a set of survey questions regarding their oral health behaviors. These data are collected closer in time to the actual behaviors, which can be expected to facilitate more accurate recollection. Additionally, if mobile health technology is further developed, it will allow health providers to recognize emerging conditions and initiate timely referrals and interventions (Kumar et al., 2015). However, there are newfound challenges when implementing this technology. Missing data due to nonresponse are likely to occur. We propose using an orthogonal-array study design to minimize the number of subjects needed for the study to meet budgetary constraints. Use of orthogonal arrays allow for efficient study design, but a reduction in degrees of freedom compared to a full factorial design. In addition to controlling for missing data, estimation of parameters for interaction effects might prove challenging. New statistical methodology, based on the pattern-mixture model framework and model based approaches, will need to be developed for analysis of the orthogonal array study design.

2.4.1 Study Design: Randomization and Full Factorial Design

Experimentation is an integral part of academic research and science more generally, but is susceptible to several biases. These susceptibilities include, but are not limited to, systematic bias (e.g. Treatment A administered in January and treatment B in November), selection bias (e.g. Selecting more fit study participants for the treatment group in an athletic performance study), and accidental bias (e.g. Using the first 20 people in line for treatment A and the last 20 for treatment B). There is also a chance of cheating by the experimenter or other people involved in the experiment. R.A. Fisher proposed the use of randomized trials to mitigate such biases and to help control for possible confounding variables by balancing the distributions of background characteristics across treatment groups (Fisher, 1926 and 1935).

In randomized experiments, there can be more than a single factor (for example: treatments A and B) of interest under study as well as several nuisance factors that need to be controlled to yield generalizable results. In situations where there are multiple factors, Fisher noted that letting one factor vary while holding the others constant was inefficient and potentially misleading (1935). A more efficient way of performing the experiment is through the factorial design where the cells are made up of all possible combinations of levels of the factors in the study (Cochran and Cox, 1957). The full factorial design allows for the study of each factor and interactions between the factors on the response variable (Box et al., 1978). Study subjects can be randomly assigned to cells of the full factorial design.

2.4.2 Randomized Block Design

If there are a large number of nuisance factors, a randomized block design can be implemented. In the randomized block design, a heterogeneous sample is collected (a random sample so it can be generalized to the population) and this sample is then grouped into homogenous subgroups, based on the nuisance factors, before they are assigned to the cells or treatment factor levels (Lawson, 2010). The formation of these homogenous groups is called blocking. Assigning treatment factor levels or cells randomly within these smaller, homogenous blocks has the same effect as only using homogeneous experimental units, but permits conclusions to be generalized to the whole heterogeneous sample in the study. Blocking can also assist in balancing covariates between cells or treatment factor levels.

Full factorial designs are efficient, allow for interactions or joint effects, of factors to be detected, and require fewer test units than studying each factor individually with the others held constant. However, as the number of factors increases, the number of required experimental units also increases. Some experimenters attempt to remedy this issue by assigning only one experimental unit per cell and reducing each factor to two levels. This can still lead to large sample size needs. Common fixes include removing factors or varying a single factor and holding the others constant, but these methods lose information and can be inefficient (Cochran and Cox, 1957). Carefully planned fractional factorial experiments can permit the desired tests while reducing the required sample size as a possible fix. Note that some study designs to be covered later, Latin squares and orthogonal array designs, fall under the fractional factorial design category.

2.4.3 Latin Squares

In the situation where there is a single factor of interest, several nuisance factors or variability that needs to be controlled, as well as a need for efficiency, the Latin square design lends itself as a solution (Cox, 1958). The classic Latin square design uses two nuisance factors as blocking variables, which are divided into a tabular grid with the property that each row and each column receive each treatment exactly once (Box, Hunter, and Hunter, 1978). If there are more than two nuisance factors or sources of variation, Graeco-Latin square design allows for three nuisance factors and hyper-Graeco-Latin square design four. Some disadvantages of the Latin square design are the number of levels of the blocking variables must equal the number of levels in the treatment factor, and the designs assumes no interaction between the blocking factors and no interaction between the blocking factors and the treatment.

2.4.4 Orthogonal Arrays

Orthogonal array designs represent a generalized version of the Latin square design (Kacker, Lagergren, and Filliben, 1991) allowing for multifactor experiments (Taguchi, 1986; Kishen, 1942; Rao, 1946). Taguchi denotes the arrays $L_N(s^m)$ with s elements in a N x m matrix where the L stands for Latin square and demonstrates that orthogonal arrays are generalized Latin squares. The columns of the experimental matrix can be viewed as factors, the entries in the columns being the test levels of those factors, and the rows are the runs. Orthogonal arrays have the property that every pair of columns have the ordered pairs of elements appear the same number of times as seen in Figure 7:

Figure 7: Orthogonal Array

1	1	1
0	0	1
1	0	0
0	1	0

As a result, this design is balanced to allow for all levels of the factors to be considered equally, which permits the factors to be evaluated independently, uncorrelated, of each other despite the fractionality of the design (Bose, 1947; Bose and Bush, 1952). The main effects and two-way interactions are considered while higher-order interactions are assumed to be nonexistent. Additionally, the flexibility of this design allows for any column to be removed and the table remains an orthogonal array, which permits the generation of many types of orthogonal arrays

(Taguchi, 1987). Orthogonal arrays remain orthogonal arrays under row permutation, column permutation, and elements within a column permuted as well. Orthogonal arrays accordingly offer efficiency, flexibility, and generalizability.

2.4.5 Missing Data in Latin Squares

Inevitably, even after utilizing the orthogonal array study design to minimize the number of required subjects and questions asked in each EMA, the EMAs will yield missing data. The EMAs will consist of preset questions split into multiple surveys. Each of these surveys has a different number of questions, which will be administered over the course of the study. The number of questions in each survey will be different each week and will be controlled for using an orthogonal array study design. Possible missing-data patterns we expect include complete non-response (survey is sent and not opened or taken), partial missing data with the first sets of questions answered, but few or none of the surveys are taken in the mid to later weeks, the reverse of the previous situation where few to no surveys taken in the early weeks, but compliance in later weeks, and missing survey randomly distributed through the weeks.

In considering missing data in an orthogonal array study design, it is relevant to see how missing data is addressed in Latin squares as literature on missing data in orthogonal arrays is sparse. Initially, it was proposed that if a row or column had a missing values, the entire row or column could be removed so that complete-data methods could be used for analysis (Allan and Wishart, 1930; Yates, 1936). However, this method proved unappealing as it discarded data and induced bias. Alan and Wishart (1930) also suggest imputing the missing data by estimating the missing

value from the minimum error sum where the errors were the difference between the actual values and those predicted by a fitted linear model.

Yates (1933) applied the method of minimizing the error sum of square to other designs with more missing data. Yates (1936) later solved the one row/column missing Latin square problem using the missing-plot technique to derive an unbiased error mean square. When the missing data problem extended beyond one row or column, Yates and Hale (1939) solved the problem using the least-squares normal equations. Other methodology addresses the missing data issue by 1) determining the bias following the missing plot technique, 2) using analysis of covariance, and 3) using the exact approach with a general regression significance test (Ott and Longnecker, 2010). A rigorous discussion of multiple imputation in Latin square design is difficult to find in the literature. Comparisons of methods for randomized complete block design (RCBD), balanced incomplete block design (BIBD), and unbalanced incomplete block design (UIBD) showed that multiple imputation performs comparatively to the exact methods (Altinisik, 2013). This suggests that multiple imputation, engineered to account for the study design and missing-data patterns, will yield statistically valid inference comparable to other missing data methods implemented in Latin square study designs.

2.4.6 Pattern-Mixture Model for EMA Missing Data

An alternative solution to this problem is to implement a pattern-mixture model specifically designed for EMA data. As mentioned previously, we expect to see several distinct patterns of missing data in our EMAs: 1) non-response, all the data missing for all surveys 2) early compliance, reflected in compliance in the first few weeks but where all or most of the later data

from weeks are missing 3) late compliance, reflected in compliance in the later weeks, but where all or most of the data from earlier weeks are missing 4) random, where surveys appear to be missing unsystematically 5) systematic, where surveys appear to always be missing when administered on certain day or days. These patterns could be informative in an imputation process, and controlling for them might ultimately result in more accurate inference. Additionally, the missing data mechanism may be MNAR, which a pattern-mixture model could address. Therefore, pattern-mixture models are a possible solution for handling the expected missing data in the EMAs.

2.4.7 Time Series Alternatives

An alternative set of models that can be applied to missing EMA data are time series models in a Bayesian framework (West and Harrison, 1997). The Bayesian framework would assume MAR missing data for the imputation process while running the time series model. The reason that time series models may offer a possible solution and a point of comparison with the pattern-mixture model is their ability to incorporate the values, errors, and variance of other time points in the estimation of the time point of interest. For instance, in the EMA setting, the answers of one question influencing the model when it is assessing other questions at a different time point. One of the most basic time series models is an aurtoregressive model:

$$y_t = \alpha + \beta y_{t-1} + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

This is an AR(1) model where the 1 represents dependence on 1 previous observation. Allowing previous observations to influence the mean structure is not the only possibility. As seen in the

Autoregressive Conditionally Heteroschedastic model, previous observations can be used to influence the variances as well:

$$y_t = \mu + a_t$$

$$a_t = \sigma_t \varepsilon_t$$

$$\sigma_t = \alpha_0 + \alpha_1 a_{t-1}^2 \quad \alpha_0, \alpha_1 > 0$$

$$\varepsilon_t \sim N(0,1)$$

$$a_{t-1}^2 = (y_{t-1} - \mu)^2$$

This is a ARCH(1) model and it allows for the variance to change over time, adjusting to previous outcomes. These two, basic models are illustrative of a wealth of other models that can be used to influence the mean and variance structure of the model and make them dependent on previous observations.

2.4.8 Controlling for Interaction Terms

As stated previously, the orthogonal array allows for testing of low-order interaction terms while assuming that high-order interaction terms are negligible. However, given missing data and the corresponding reduced information for parameter estimation, the estimation of interaction term parameters may prove challenging. In order to estimate parameters for interaction terms given the degrees of freedom afforded in the orthogonal array design, either regularization or clever statistical methods must be used. One strategy that offers regularization and smoothing of the interaction terms is the approach taken by Rubin, Schafer, and Schenker, (1988) where the prior distribution for each interaction term is centered around zero with a smaller variance for each

increase in order. This results in high-order interaction terms being strongly smoothed to zero with lower-order terms estimated with more modest smoothing. By implementing this modeling strategy, high-order interaction terms are still accommodated but will only have a substantial impact on the analysis if there is information in the data that can inform the estimation of the parameter. This method will be discussed in more detail later. An alternative strategy is use of Tukey's Test of Additivity (Tukey, 1949). This test is designed to assess if factor variables are additively related to the expected value of the response variable while only costing a single degree of freedom. This modeling strategy will also be discussed later in greater detail. Ultimately, these methods will be utilized with incomplete data to control for the possibility of interaction effects.

3 Spatial Models for Clustered Count Data Controlling for Multiple Classes of Neighbor Relations

Dental caries data contain complex correlations stemming from the spatial structure of the mouth. Experience suggests that the tooth-specific count variable that has often been used to summarize dental caries, namely the total number of decayed, missing, or filled surfaces (DMFS) on a tooth (DMFS index), can be over-dispersed relative to Poisson-model variation. Using a sample of 571 methamphetamine users from Los Angeles, we model DMFS index using alternative models to account for possible over-dispersion. First we consider a Beta-Binomial model that retains conditional independence across teeth. We also investigate three spatial models that accommodate tooth-level correlation: a model with conditionally autoregressive (CAR) prior distribution accounting for nearest neighbor and cross-mouth correlation at the tooth level, a smoothed analysis of variance (SANOVA) approach that implicitly reflects interaction effects through a shrinkage-estimation strategy rather than either including or excluding fixed interaction effects, and a hierarchical Bayesian model analogous to ridge regression that smoothes higher –order interaction terms more than lower-order terms, potentially accommodating spatial stratification within the mouth as well. Model fit and inference were compared to assess model performance and compatibility of conclusions with observed data. By utilizing these more flexible spatial models and accounting for the spatial

structure of dental data, we anticipate being able to develop a better understanding of the underlying mechanisms of dental decay in methamphetamine users.

3.1 Introduction

Dental caries present both challenges and opportunities for representing within-individual and between-individual patterns of association. Experience and intuition suggest that patterns of dental caries might be expected to exhibit spatial structure, whether due to similar exposures of teeth to sources of decay or to patterns indexed by an individual's dental hygiene habits. Building on methods that have been utilized to analyze dental data and other types of data, this chapter seeks to provide insight into how spatial statistical can advance understanding of patterns in analyzing dental caries data.

3.1.1 Motivating Example

The present work is motivated by a study where data (Shetty et al., 2010, 2015, 2016; Clague, Belin, and Shetty, 2017) were collected from 571 methamphetamine users in Los Angeles County recruited over a 2-year period from dental clinics associated with two large community health centers: (1) a clinic affiliated with AIDS Project, Los Angeles (APLA), a non-profit organization based in Los Angeles, which primarily serves a socio-demographically diverse group of individuals with HIV/AIDS, and (2) the Mission Community Hospital (Mission) in the San Fernando Valley, a public/private hospital that caters to a large, underserved migrant population. Approximately 69% of the subjects were recruited at the APLA clinic and the remainder at the Mission clinic. Subjects were at least 18 years of age, spoke English or Spanish, had described themselves as methamphetamine users, submitted responses to an extensive 10-

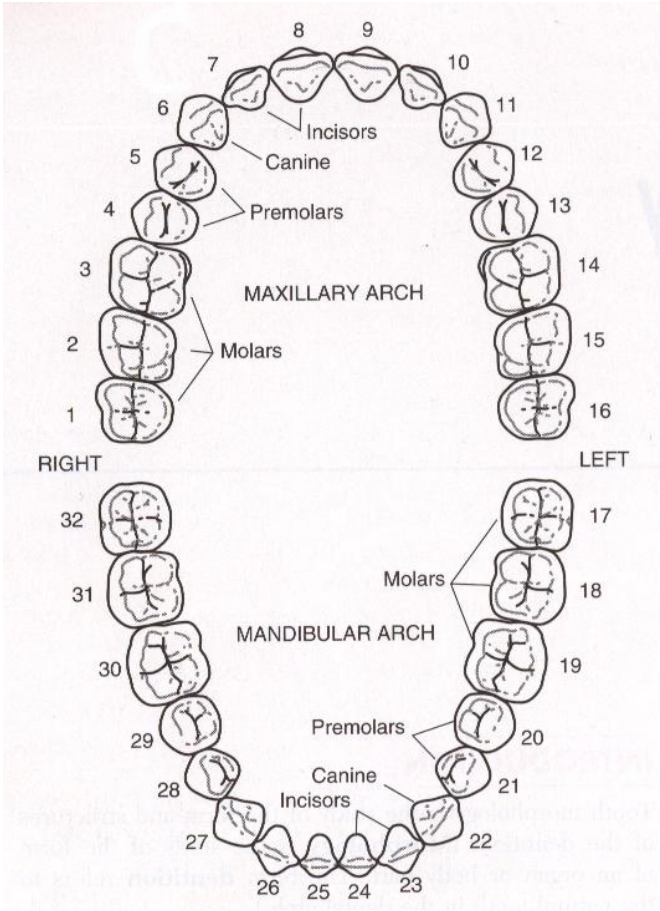
year drug history questionnaire, used methamphetamine in the past 30 days, and were able to undergo a detailed dental examination and psychosocial assessments. Additional details on the data collection process can be found in earlier published articles on the study (Shetty, Harrell, Clague, Murphy, Dye, and Belin, 2016).

The majority of the study sample (n = 571) was male (n = 460), African-American and/or Latino (42.3% and 31.2% respectively), older than 30 years (Mean age = 44.5 years, SD = 9.6), and most had completed high school (n=401). A large subset of the study participants (n=147) were HIV positive. Many of the methamphetamine users were current cigarette smokers (68.8%). Based on self-reported history, 64% of the participants used methamphetamine frequently (reported using 15 or more days a month on average) and, on average, had used meth for at least 10 of the preceding months (Mean = 10.1, SD=2.1). The average age of methamphetamine use initiation was 28.5 years (SD=10.5) and smoking was the most common mode of use, with 53% reporting smoking as their exclusive mode. The study participants consumed an average of 3.5 sugary drinks a day (Shetty, Harrell, Murphy, Vitero, Gutierrez, Belin, Dye, and Spolsky, 2015; Shetty, Harrell, Clague, Dye, and Belin, 2016; Clague, Belin, and Shetty, 2017).

Dental decay, the outcome variable, was represented using the DMFS index (Darby and Walsh, 2003), which is the total count of decayed, missing, or filled surfaces per tooth. Each tooth has a maximal count of 4 (Canine and Incisors) or 5 (Molars and Premolars) on the DMFS index depending on tooth type. This data structure represents a nested sampling structure on teeth of different types within individuals. Based on the implicit understanding that third-molar extraction is often motivated by orthodontic considerations having little to do with dental caries,

it is customary in the dental literature (Shetty, Harrell, Clague, Dye, and Belin, 2016) to remove the third molars (tooth numbers 1, 16, 17, and 32) from the analysis, leaving in 28 teeth, or outcomes, per subject. A visual representation of how teeth are numbered in the mouth can be seen in Figure 8 (Khodai, 2012). With 571 subjects, there were a total of 15,988 teeth in the dataset. Additionally, missing teeth were assigned a DMFS index of 4 or 5, depending on the tooth type, because all surfaces were recorded as missing.

Figure 8: Tooth Numbering Scheme



(Taken from https://buildinggreatsmiles.com/blog/what-tooth-number-is-this-tooth/)

3.1.2 Modeling Counts of Adverse Dental Outcomes Accounting for Spatial Correlation Patterns

Zhang et al. investigated spatial patterns in dental decay controlling for correlation within quadrants, neighboring teeth, and exploring symmetries in patterns of dental decay that are widely believed by dentists to exist (Todem, 2008; Zhang, Todem, Kim, and Lesaffre, 2011). To account for correlation patterns observed in dental data, we consider various sampling models, which include use of spatial random effects with a conditionally autoregressive (CAR) prior distribution, Smoothed ANOVA (SANOVA), and a hierarchical Bayesian model that uses normal prior distributions centered at zero. Each model invokes a shrinkage-estimation strategy that allows for the spatial structure of the mouth to be incorporated into the model. SANOVA and the hierarchical Bayesian model control for the spatial structure of dental data by utilizing high-order interaction terms for spatial stratification. Spatial stratification is achieved by creating interaction terms between the main effects and indicator variables for tooth location. Both models smooth the interaction terms to zero to accommodate the high-dimensionality of the model. The hierarchical model employs priors on the interaction terms that increase the degree of shrinkage as the order of the interaction term increases.

The CAR model utilizes and an autoregressive smoothing strategy where spatial random effects are given a CAR prior distribution, which enables estimates of true caries status to be improved through spatial smoothing dependent on the neighboring teeth. Specifically, outlier teeth are smoothed to their neighbors, which possibly provide a better representation of the oral health of the mouth. Additionally, missing teeth, which receive a full DMFS index of 4 or 5 depending on

tooth type, which can have a sizeable influence on the parameter estimates, are smoothed to their neighbors. This may be helpful in understanding the reason for the missing tooth as teeth often do not decay in isolation.

Neighboring correlation in periodontal, gum, disease data has also been suspected and was explored utilizing the CAR prior specification (Reich and Bandyopaghyay, 2010; Reich, Hodges, and Carlin, 2007). To our knowledge, Bandyopadhyay, Reich, and Slate are the only previous investigators who have controlled for spatial correlations using spatial random effects with a CAR prior distribution with DMFS index as the outcome (Bandyopadhyay, Reich, and Slate, 2011). In their analysis, the influence of local neighbors, the teeth one spot removed, and the tooth directly above or below are noted. For instance, the "neighbors" for tooth 3 were 1, 2, 4, 5, and 30. Additionally, the Beta-Binomial model with reparametrized mean had noticeable improvements in model fit when compared to other models for DMFS index data. We use their Beta-Binomial parameterization in our analysis as one of several models tested.

Due to the technical complexities involved in estimation and inference with spatial models, we review the basic specification of SANOVA, the hierarchical Bayesian model, and the CAR prior distribution. In the discussion of the CAR prior distribution, the construction of the neighbor relations, expressed through the adjacency matrix, for our data will be explained. With all the necessary methodology behind the spatial models reviewed, the Beta-Binomial regression parameterization used by Bandyopadhyay, Reich, and Slate will be presented. Prior distribution specification and model selection criteria will then be explained. All computation was done using open source software R and Stan, which utilizes Hamiltonian Monte Carlo (HMC) sampling

methods. Before discussing the results of the models on the aforementioned Methamphetamine data, the models will be run on simulated datasets to assess parameter convergence, model fit, and run times.

3.2 Spatial Models

3.2.1 Smoothed Analysis of Variance

Analysis of variance (ANOVA) accounts for variance in the outcome variable using individual factors and interaction terms between those factors. Written as a linear model, ANOVA can be expressed with a main effects matrix (A_1), which is cn x M matrix where c are number of cells, n are the observations per cell, and M is the number of columns for the main effects, and an interactions matrix (A_2), which is a cn x N matrix where N is the number of interaction terms. The typical design matrix, X, is defined as $X = [A_1|A_2]$ and the regression parameters split $\left(\frac{\theta_1}{\theta_2}\right)$ so that θ_1 are the model parameters for the main effects and θ_2 the parameters for the interaction terms (with M + N total regression parameters), which yields a linear model $y = [A_1|A_2]\left[\frac{\theta_1}{\theta_2}\right] + \epsilon$ where ϵ is normally distributed with mean 0 and covariance $\frac{1}{\gamma_0} I_{cn}$. It can be useful to include interaction terms between in the model, but are removed if not significant. However, it is not ideal to assume that any interaction term is absent or unimportant. Rather, it would be preferable if all interaction terms were included, but the contributions of the interactions with small effects were small and mostly removed, those with medium effect retained partly or fractionally included, and those with large effect almost entirely retained.

Smoothed analysis of variance (SANOVA), proposed by Hodges, Cui, Sargent, and Carlin (2007), is a method that neither includes nor excludes interaction terms, but implements shrinkage so that all interaction terms are retained and the data allow for those with more effect to have greater contribution. This is achieved using the same linear model as ANOVA:

$$y = [A_1|A_2]\left[\frac{\theta_1}{\theta_2}\right] + \epsilon$$

Where ϵ is a Normal distribution with mean 0 and covariance $\frac{1}{\eta_0} I_{cn}$. However, in SANOVA, the interaction term parameters are smoothed:

$$0_N = [0_{NxM}|I_N]\left[\frac{\theta_1}{\theta_2}\right] + \delta$$

Where δ is an N-variate Normal distribution with mean 0 and diagonal covariance matrix $diag(\varphi_1, \ldots, \varphi_N)^{-1}$. The φ_k determine the smoothing, or shrinkage, of their respective θ_k. Each φ_k need not be distinct and can be used for multiple regression parameters. For instance, a set of distinct constraint case precisions $\{\eta_0, \eta_1, \ldots, \eta_s\}$, $s \leq N$and a deterministic assignment function j(k) yields $\varphi_k = \eta_{j(k)}$. This makes it so the interaction parameters are smoothed by the underlying $\eta_{j(k)}$. Prior distributions on these smoothing parameters determine the degree of shrinkage. A gamma prior distribution is placed on the shrinkage parameters.

Applying this model to the DMFS index data, A_1 contains indicator variables for socioeconomic and socio-demographic variables (Race, Age (over/under 30 years of age), Education, Smoking Status, and Marital Status) and thirteen spatial indicators for location of the tooth based on tooth type, jaw, and right/left side of the mouth. The interaction matrix, A_2 , is composed of two, three, and four way interaction terms between the main effects and the spatial location indicator

variables. One of the spatial indicator variables is present in each interaction term allowing for the model to be spatially stratified. Each order of interaction has its own set of smoothing variance parameters, $\eta_{j(k)}$, and there are thirteen for each order so that the coefficients for the interaction terms are smoothed by tooth location. The SANOVA model presented can be easily adapted to the DMFS index data by changing the likelihood and link function of the model to count based distributions (Poisson, Negative-Binomial, and Beta-Binomial).

3.2.2 Hierarchical Bayesian Model

Rubin, Schafer, and Schenker, working with the U.S. Census Bureau's post-enumeration survey, developed a nonignorable model to impute missing categorical data (Rubin, Schafer, and Schenker, 1988). The method foreshadowed the related development of pattern-mixture models (Little, 1992). To permit high-order interaction terms, they utilized a hierarchical Bayesian imputation model:

$$\log(\theta_{ijk..p}) = \mu + \mu_{1(i)} + \mu_{2(j)} + \cdots + \mu_{8(p)} + \mu_{12(ij)} + \mu_{13(ik)} + \cdots + \mu_{123..8(ijk..p)}$$

Where $\theta_{ijk..p}$ is the probability an observation falls in cell ijk…p and the $\mu's$ represent the main effects, one way, two way, three way, and high-order interaction terms. The prior distributions placed on the interaction terms allow for increased smoothing the higher the order the interaction term:

$$\mu_i \sim N(0, \sigma^2)$$

$$\mu_{ij} \sim N\left(0, \frac{\sigma^2}{\tau}\right)$$

$$\mu_{ijk} \sim N\left(0, \frac{\sigma^2}{\tau^2}\right)$$

.

.

$$\mu_{ijk..p} \sim N\left(0, \frac{\sigma^2}{\tau^7}\right)$$

For $\sigma > 0$ and $\tau > 1$. This hierarchical Bayesian model permits high-order interaction terms to be in the model, but smoothes them heavily toward zero. Additionally, minimizing the number of parameters present in the model is accomplished by having a common σ^2 and setting τ as a constant.

In a dental setting, the outcome is no longer categorical, but a count variable. Changing the likelihood to accommodate the count data, while keeping the model structure the same, allows for spatial stratification by location in the mouth with high-order interaction terms. This spatial stratification is achieved through interaction terms between the main effects and thirteen spatial indicator variables as described in the previous SANOVA section. The mean structure between the hierarchical model and SANOVA are identical, but the two models differ in how they smooth the spatially stratified interaction terms and the number of parameters in each model. Comparisons between this model and the SANOVA will be assessed as the hierarchical model is simpler to program, computationally more efficient, and, intuitive, where high-order interaction terms are more significantly penalized, or shrunk, as there is likely less information in the data supporting their inclusion in the model. The τ parameter was determined using a predefined

grid and the Watanabe information criterion (WAIC) (Watanabe, 2010; Gelman, Hwang, and Vehtari, 2014) for model fit comparisons on the simulation data. Ultimately, τ = 1.1 had the best performance in model fit (WAIC) and was used on the DMFS index data.

3.2.3 CAR Prior Distribution

The CAR prior distribution, an approach introduced by Besag, York, and Mollie (1991) , is expressed in full conditionals distributions using spatial random effects. For illustrative purposes, let the spatial random effects be given as a random quantity $\varphi_s = \left(\varphi_{s(1)}, \varphi_{s(2)}, \ldots, \varphi_{s(n)}\right)^T$. Where $\varphi_{s(t)}$ is a spatial random effect for tooth t (t = 1, … , 28) in subject s (s = 1, …, 571):

$$\varphi_{s(i)} | \varphi_{s(j)}, j \neq i \sim N\left(\alpha \sum_{j=1}^{n} b_{ij} \varphi_{s(j)}, \tau^{-1}\right)$$

Where τ^{-1} is a spatially varying precision parameter, which is common for all subjects and teeth. The b_{ij} are weights, often equal to $\frac{1}{total\ number\ of\ neighbors\ for\ tooth\ i}$ for teeth designated as neighbors of tooth i, zero otherwise, and $b_{ii} = 0$. Using Brook's Lemma, we can rewrite these full conditionals as a joint distribution:

$$\varphi_s \sim \mathrm{N}(0, [\tau D(I - \alpha B)]^{-1})$$

D_{nxn} is a diagonal matrix with entries representing the number of neighbors for location i, I_{nxn} is the identity matrix, α_{1x1} is the spatial dependence parameter, W_{nxn} is known as the adjacency

matrix where $w_{ii} = 0$ and $w_{ij} = 1$ if i is a neighbor of j, otherwise $w_{ij} = 0$, and $B = D^{-1}W$. The adjacency matrix W is manually specified and the diagonal entries of D are equal to the sum of the rows of W. The spatial parameters τ and α receive their own prior distributions. To simplify the joint distribution:

$$\varphi_s \sim \mathrm{N}(0,[\tau D(I - \alpha D^{-1}W)]^{-1}) = \mathrm{N}(0,[\tau(D - \alpha W)]^{-1})$$

The CAR prior assumes a lattice structure with neighbor relations among sites, also known as a Markov random field, which is evident in the full conditionals representation. The spatial random effect for observation i is a weighted average of its designated neighbors, but this weighted average is augmented by α, which makes the spatial random effect for subject i a proportion of the weighted average of its neighbors. To ensure that the joint distribution is proper it is required that $|\alpha| < 1$ because, for the joint distribution to be proper, the covariance matrix must be nonsingular (Gelfand and Vounatsou, 2003). If $\alpha = 1$, $(D - W)1 = 0$, given that 1 is a vector n x 1 of 1's, which implies that the covariance is now singular. However, by setting $\alpha = 1$ the full conditional distributions now have an intuitive interpretation: the spatial random effect for observation i is a weighted average of its specified neighbors. This version of the CAR prior distribution, called the intrinsically autoregressive (IAR) prior distribution, is desirable and can still be used as long as it is relegated to a prior distribution rather than on the data directly due to its being improper (Banerjee, Carlin, and Gelfand, 2014). The IAR joint distribution is often expressed:

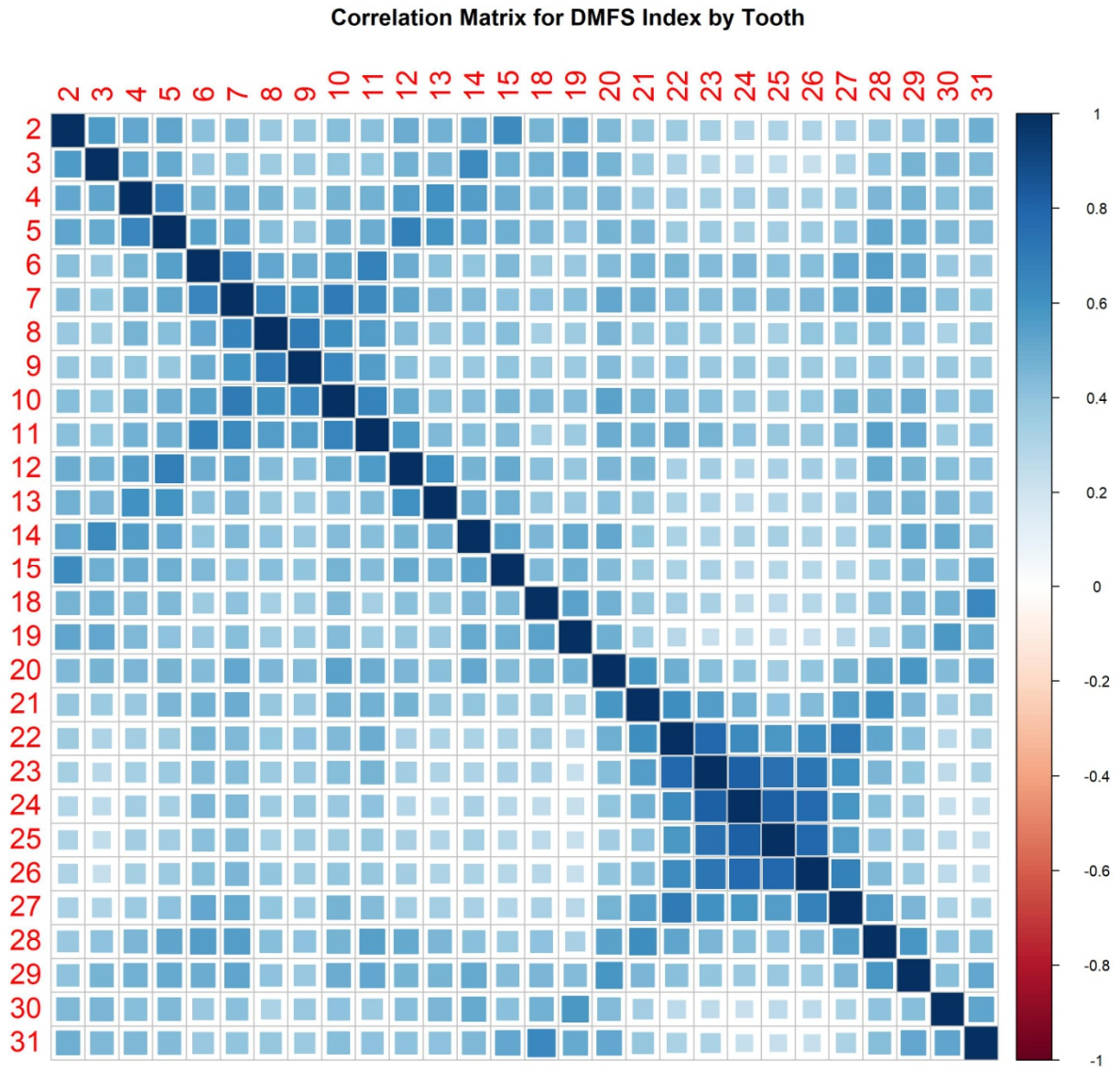
$$p(\varphi_s|\tau) \propto \tau^{\left(\frac{n-G}{2}\right)} e^{\left(\frac{-1}{2}\varphi_s{}^T(\tau Q)^{-1}\varphi_s\right)}$$

Here Q = $(D - W)$, meaning that this normal kernel has precision τQ, where q_{ii} is equal to the number of neighbors for location i and q_{ij} is -1 if location i and j are neighbors, 0 otherwise. Additionally, G represents the number of teeth that have no neighbors, but G = 0 for our neighbor specification. Expressing the IAR prior distribution in this way is done for simplicity, and the rest of this chapter will continue to do so. Additionally, we choose to use the IAR prior distribution specification due to all preliminary runs of the proper CAR prior distribution resulting in α's posterior distribution being tightly clustered near 1.

3.2.4 Specification of Adjacency Matrix and Model Parameters to Account for Salient Spatial Associations Correlation Patterns

The presence of a complex correlation structure was evident in the data from the methamphetamine study after constructing a correlation matrix and representing the magnitudes of correlation using a color scale in Figure 9. In Figure 9, the numbers correspond with the tooth numbering in Figure 8 where teeth 1, 16, 17, and 32 are not included and the blue shading indicates a positive correlation between the DMFS index for each tooth in the sample. This correlation structure evidenced three strong relationships: (1) local neighboring teeth on the same jaw (2) cross-mouth neighbor relationship on the same jaw (3) strong correlation amongst all 4 incisors on the same jaw.

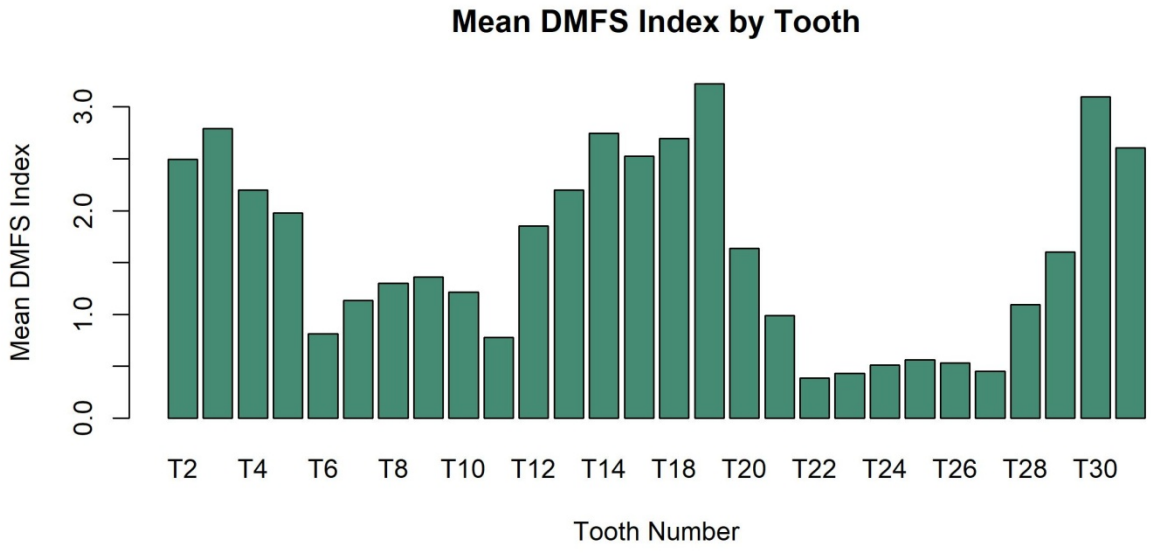
Figure 9: Correlation Matrix



Correlation among teeth with the same tooth type is clearly evident but is weaker than the other three patterns. A similarly complex correlation structure between local and cross-mouth periodontal pockets was observed in an examination of periodontal disease (Reich, Bandyopadhyay, and Bondell, 2013). This Such a pattern could be explained by soft tissue (gum tissue) experiencing disease before the tooth experiences a DMFS event. Further similarity

between neighboring teeth can be seen in Figure 10, which suggests that interpretation of the data could benefit from spatial smoothing, with local neighbor teeth providing useful input for local smoothing. However, it is also evident that including the cross-mouth tooth on the same jaw in the smoothing is important as well. For instance, tooth 3 can be smoothed using teeth 2 and 4, but both show DMFS indexes that are disproportionately below tooth 3's mean. The inclusion of tooth 14, the cross-mouth neighbor of tooth 3, in the smoothing of tooth 3 could presumably improve the precision of estimation of parameters relating to tooth 3. It is important, when constructing the adjacency matrix, to be thoughtful in the characterization of neighbor relations among teeth due to the strong influence such modeling choices can have on the estimation of the spatial random effect and means of the fixed effects (Reich, Hodges, and Zadnik, 2006; Earnest, Morgan, Mengersen, and Ryan, 2007).

Figure 10: Mean DMFS Index by Tooth



The adjacency structure for the caries experience can be found in Table 1. Tooth numbering can be seen in Figure 8 for reference. This neighborhood structure accounts for the three most

prominent correlation patterns mentioned previously and allows second molars that contact vertically (2, 15, 18, and 31) to be neighbors.

Table 1: Adjacency Structure

Tooth Number	Neighbors	Tooth Number	Neighbors
2	3, 15, 31	18	15, 19, 31
3	2, 4, 14	19	18, 20, 30
4	3, 5, 13	20	19, 21, 29
5	4, 6, 12	21	20, 22, 28
6	5, 7, 11	22	21, 23, 27
7	6, 8, 9, 10	23	22, 24, 25, 26
8	7, 9, 10	24	23, 25, 26
9	7, 8, 10	25	23, 24, 26
10	7, 8, 9, 11	26	23, 24, 25, 27
11	6, 10, 12	27	22, 26, 28
12	5, 11, 13	28	21, 27, 29
13	4, 12, 14	29	20, 28, 30
14	3, 13, 15	30	19, 29, 31
15	2, 14, 18	31	2, 18, 30

3.3 Likelihood, Prior Specification, and Simulation Study

3.3.1 Beta-Binomial Model

The DMFS index represents a count variable, suggesting the possibility of using a Poisson, Negative-Binomial, or Beta-Binomial distribution to model observed outcomes. However, due to the truncated nature of the data, and possible issues with over-dispersion, the beta-binomial model is appealing. Therefore, we utilize the specification proposed by Bandyopadhyay, Reich, and Slate (2011) where the DMFS index for tooth t of subject s is represented as $y_s(t): s = 1,2,\ldots,571, t = 1,2,\ldots,28$ and $n_t = 4$ for canines and incisors and 5 for molars and premolars:

$$(\varphi_s(1),\ldots,\varphi_s(28))^T \sim CAR(\tau)$$

$$logit\big(\mu_s(t)\big) = X_{s(t)}^T\beta + \varphi_s(t)$$

$$p_s(t) \sim Beta(\theta(t)\mu_s(t), \theta(t)[1-\mu_s(t)])$$

$$y_s(t) \sim Binomial(n_t, p_s(t))$$

$X_{s(t)}$ represents the vector of fixed effects for subject s and tooth t, which include tooth and subject level covariates. Additionally, given this parametrization of the Beta-Binomial model, the over-dispersion parameter is $\frac{\theta(t)+n_t}{\theta(t)+1} \in (1, n_t)$ where n_t will change based on tooth type (4 or 5). Additionally, this model, although presented in the CAR model specification, will be utilized for the SANOVA and the hierarchical Bayesian model as well.

3.3.2 Choice of Prior Distributions

We specify prior distributions drawing on previous research done by our group (Shetty et al., 2010, 2015, 2016; Clague, Belin, and Shetty, 2017). We assign normal distributions centred at 0 with variance of 1 to the priors for the coefficients of the fixed effects, including the intercept, which implies the density of the associated odds ratio for a one-unit difference in the outcome

centered at 1 with 95% interval (0.14, 7.4). The range of odds ratios implied by the prior distributions seemed responsible given previous research where no odds ratio was observed below 0.2 or above 5. Both the Negative-Binomial and Beta-Binomial distribution have an over dispersion parameter for each unique tooth (28 parameters). Each was initially assigned a uniform(0, 10) prior distribution, but due to identifiability and convergence concerns a gamma(1,1) was used. The CAR prior distribution smoothing parameter, denoted by τ, was taken to have a uniform(0, 1.5) prior distribution after experimenting with gamma(0.001, 0.001) and a variety of normal distributions centered at 0 with a grid of variance values but encountering convergence issues with these prior distributions. The SANOVA precision parameters were specified with a gamma(0.001, 0.001) prior distribution, and the hierarchical Bayesian model utilized the same distribution for its shared variance parameter. The τ value in the hierarchical model was estimated by fitting the model across a grid of fixed values and 1.1 was identified using WAIC as providing the best fit to the data.

The HMC sampling was done for 3,000 iterations after 1,500 initial burn-in iterations with four chains having arbitrary starting values. Autocorrelation plots were examined and the Gelman-Rubin $\hat{R}$ diagnostic (Gelman and Rubin, 1992) was used to assess convergence of the chains to a stationary distribution. The SANOVA and hierarchical Bayesian model had no concerning convergence or identifiability issues. However, the CAR model experienced convergence issues when the τ smoothing parameter was not constrained. After restricting τ to fall between 0 and 1.5, it no longer experienced identifiability or convergence issues.

3.3.3 Simulation Study

Simulated data were generated under known models to assess each model's ability to converge to the correct parameter values. Evaluations also recorded time until convergence. Data were generated based on the mean structure incorporating linear dependence on the aforementioned socioeconomic and socio-demographic variables and 13 tooth-location indicator variables with and without interaction terms using a Binomial likelihood. Multiple data sets were generated from this base model with various changes including: random values for tooth parameters, tooth parameter values chosen to reproduce patterns seen in Figure 10, heteroscedasticity in sampling error, outliers, and data generated with random effects with a CAR prior distribution using the adjacency/neighbor matrix specified previously. Data were generated at two sample size levels, N = 100 subjects (2800 teeth) and 1000 subjects (28000 teeth), to assess sensitivity to sample size of the models. The models were run on these simulated data sets and given a 1000 iteration run time.

The models (CAR, SANOVA, and Hierarchical Model), using Beta-Binomial likelihood distributions converged to the correct parameter values for their respective generative models. In situations where the generative model was dissimilar, for example, data generated by a main-effects model with interactions being modeled with main effects and spatial random effects utilizing a CAR prior distribution, the main effects converged to the correct parameter values. However, the CAR model did experience convergence issues in every situation and required tuning of the smoothing parameter τ to achieve convergence. Model fit was assessed for each dataset using WAIC and posterior predictive checks comparing the distribution of counts generated by the models to the simulated data (Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin, 2013) . SANOVA and the hierarchical model never exhibited a substantial difference in

WAIC on any of the simulated datasets. The CAR model underperformed in WAIC on the multiple linear regression with interactions, interactions and outliers, and interactions having tooth levels representing Figure 3 (For the simulated data generated by a model with interaction terms: $WAIC_{Hierarchical}$ = 42730.9, $\text{WAIC}_{\text{SANOVA}}$ = 42774.9, $WAIC_{CAR}$ = 63107.7). The CAR model did not outperform the other models on the CAR generated data ($WAIC_{Hierarchical}$ = 43223.2, $\text{WAIC}_{\text{SANOVA}}$ = 43095.4, $WAIC_{CAR}$ = 43587.3). Run times were recorded for each model on the generated datasets containing 28000 observations. The hierarchical model consistently ran more quickly (14 hours more quickly on the data with N = 28000) than the CAR and SANOVA model as expected due to having fewer parameters.

3.4 Data Analysis and Conclusion

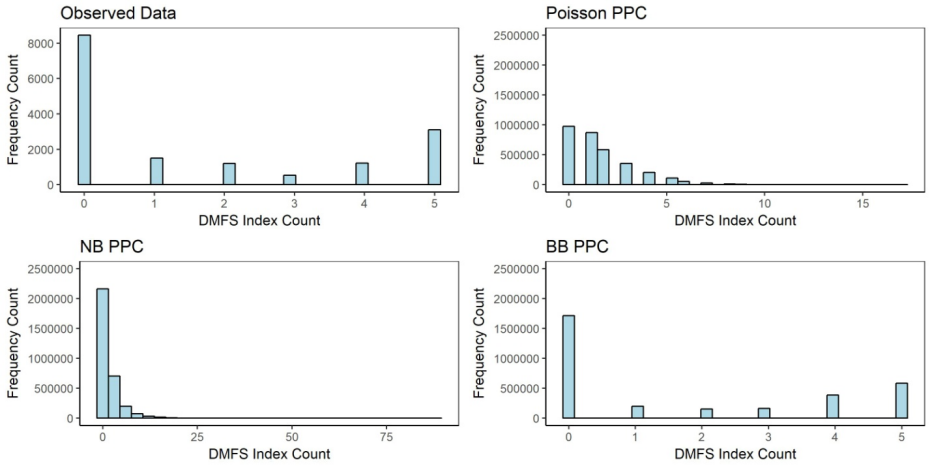
3.4.1 Data Analysis and Findings

A variety of models, including the aforementioned Beta-Binomial, Poisson, and Negative-Binomial models, were run to assess the importance of controlling for overdispersion in the data and accounting for the truncated number of trials for each tooth (4 or 5 total surfaces depending on tooth type). Model fit was assessed using WAIC and posterior predictive checks were used to compare the distribution of counts generated by the models to the observed data. For comparison purposes, 3 sets of models were run on the data. The first set of models had only fixed-effect covariates, which include the thirteen spatial indicator variables and the aforementioned socioeconomic and socio-demographic variables using Binomial (model 1), Poisson (model 2), Negative-Binomial (model 3), and Beta-Binomial (model 4) regression models. The next set of models added random effects for subject and tooth type to the previous set of models to reflect correlation amongst teeth nested in the same mouth and teeth within the same tooth type (Model

5, model 6, model 7, model 8, model 9, and model 10). The final grouping of models included spatial components, CAR, SANOVA, and Hierarchical Bayesian model using a Beta-Binomial regression model, which were detailed previously in this chapter (model 11, model 12, and model 13).

Across all model groupings, the Beta-Binomial regressions very substantially outperformed the other models in WAIC. In the fixed effects only grouping, the Beta-Binomial regression (WAIC = 36,149) had no close competitor $(WAIC_{poisson} = 56{,}574, WAIC_{Negative\ Binomial} = 50868)$. Examination of the posterior predictive checks revealed that Poisson and Negative-Binomial models regularly generated DMFS indexes that exceeded 4 or 5 whereas the Beta-Binomial model only generated values within the 0 to 5 range due to its constrained number of trials (maximum value generated: Poisson = 14, NB = 52, and BB = 5). Below in Figure 11 are the posterior predictive checks for the Beta-Binomial, Poisson, and Negative-Binomial models:

Figure 11: Distribution of DMFS Index (Base Models)



The posterior predictive checks reproduce the distribution of DMFS index implied by fitted model parameters across 200 generated data sets using parameters from every 30th iteration. A second comparison was done where the proportion of observations at 0, (1, 2, or 3), (4 or 5), and 5+ for each generated data set was calculated, plotted, and compared to the true, observed data in Figure 12 below:

Figure 12: Proportion of Simulations in Each Category (Base Models)

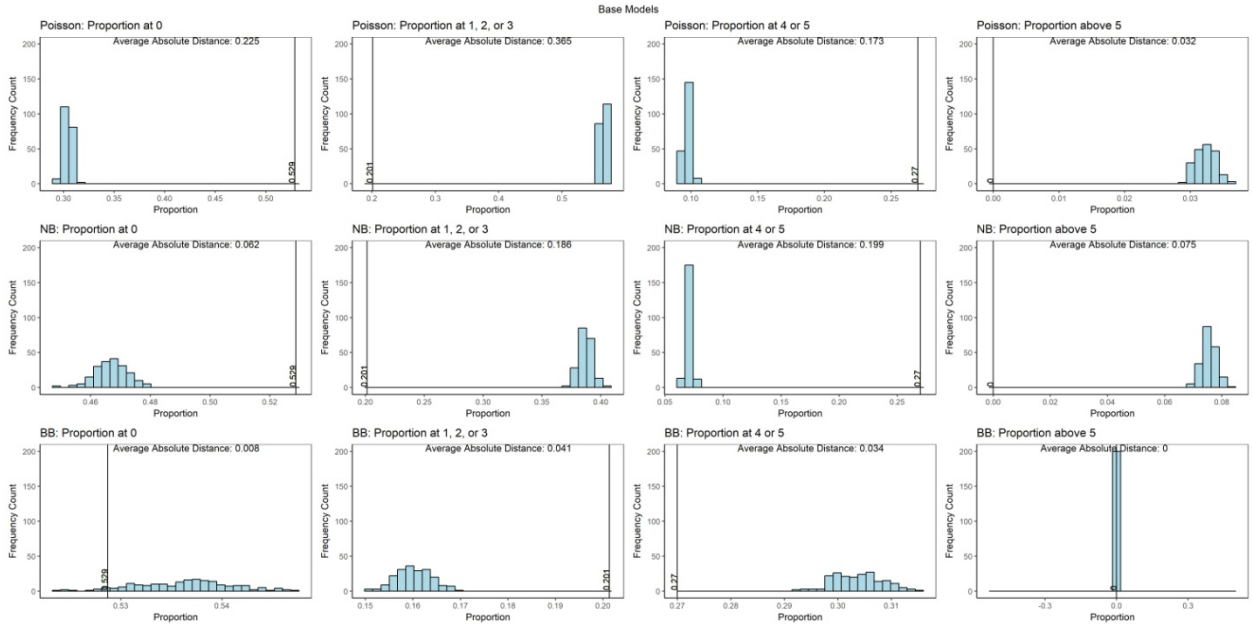
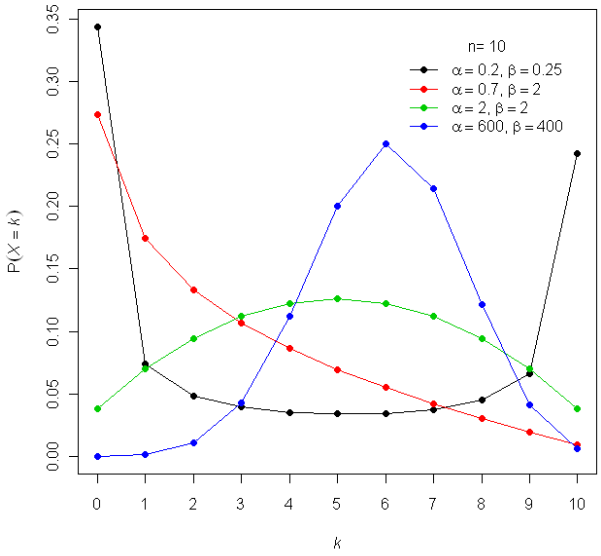


Figure 12 demonstrates that across generated data sets, the Beta-Binomial model outperforms all other models in each category. Even with the large, maximal counts generated by the Poisson and Negative-Binomial models, ultimately, the distribution of the observed data and the flexibility of the Beta-Binomial model are likely the reason for its superior fit. Examining the "y" distribution in Figures 11, the observed data exhibit a "U" shape where the counts for 0 and 5 are larger than the intermediate counts. The Poisson and Negative-Binomial models struggle to fit this distribution because they are unimodal distributions, placing majority of the density around 0

and do not have a natural truncation at 5. However, the Beta-Binomial can create this "U" shape when both the α and β parameters are less than 1, as displayed in Figure 13.

Figure 13: Beta-Binomial Distribution

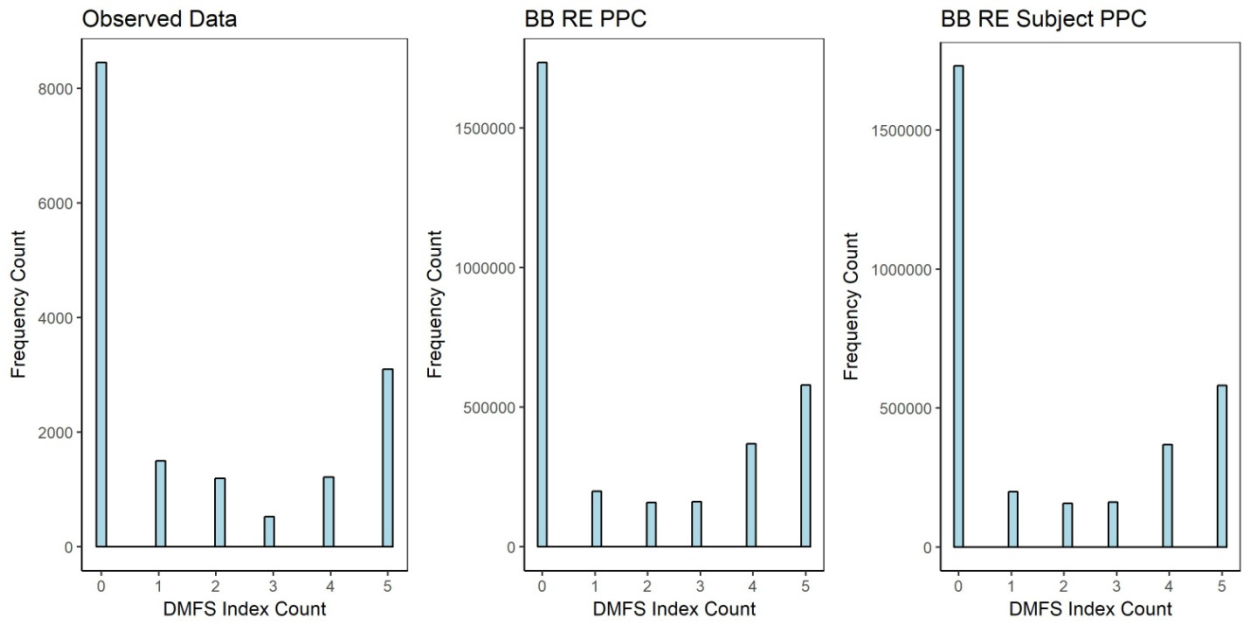


Parsing of the α and β parameters in the posterior predictive generating models found that the maximum values were 0.64 and 0.69, with mean of 0.14 (sd = 0.14) and 0.24 (sd = 0.12), respectively for each parameter. Additionally, 85% of the time, the generated α was lower in value than the generated β. These parameter estimates generate data resulting in a bimodal distribution with more density near 0 than 5, which is precisely how the observed data are distributed and explains why the Beta-Binomial model had such a large improvement in model fit (WAIC) over the Poisson and Negative-Binomial models.

The addition of random effects for subject and tooth location greatly improved model fit. The Poisson and Negative-Binomial models evidenced a large change in WAIC compared to the fixed effects only models $\left(WAIC_{poisson\ RE} = 45{,}779\ , WAIC_{Negative\ Binomial\ RE} = 44{,}900\right)$.

However, the Beta-Binomial model continued to substantially outperform all other models ($WAIC_{Beta-Binomial\ RE} = 28{,}536$). The Beta-Binomial model was then run with just the subject random effect and again with just the tooth type random effect. The subject only random effect model evidenced a large improvement ($WAIC_{Beta-Binomial\ RE\ Subject} = 28{,}724$) over the fixed effects only model and was not meaningfully different from the model with both subject and tooth random effects. Figure 14 below is a plot of 200 generated datasets for the tooth-type and subject random effect and subject only random effect Beta-Binomial models:

Figure 14: Distribution of DMFS Index (Random Effect Models)



From Figure 14, there is little difference in the shape of the posterior predictive checks between the two random effect models. Figure 15 below shows the proportion of each generated dataset in each category as done previously:

Figure 15: Proportion of Simulations in Each Category (Random Effect Models)

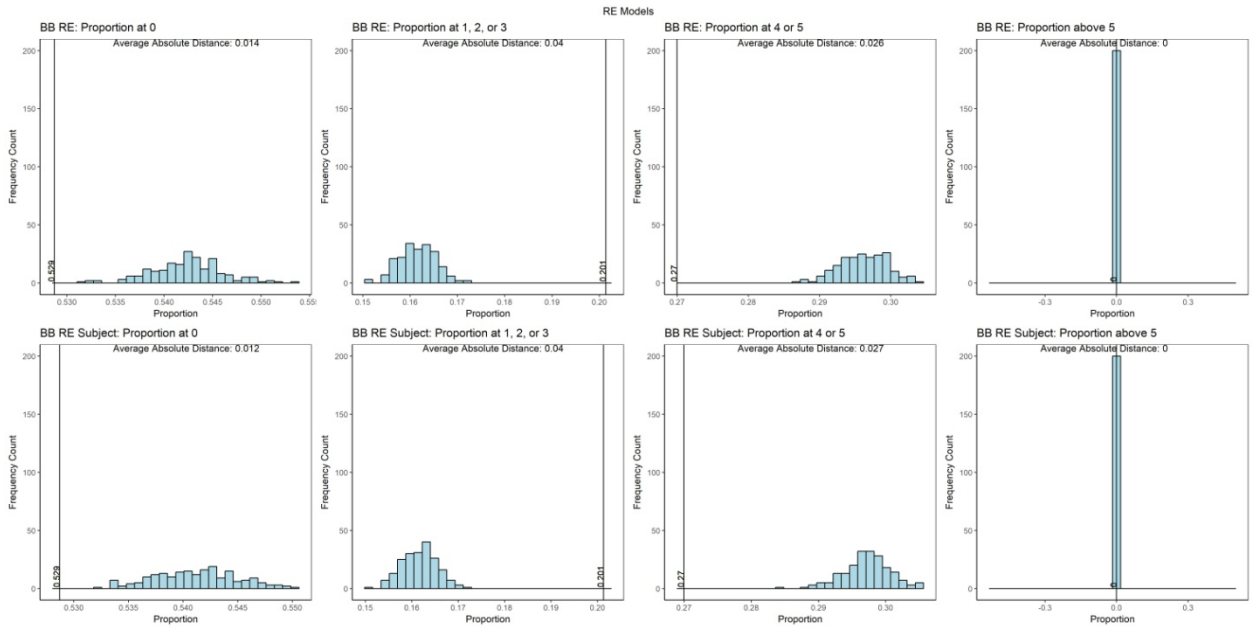


Figure 15 further solidifies that there is little difference in performance when adding a tooth-type random effect to the model over the subject only random effect. Additionally, the model with only subject level random effects improved model runtime by 4,714 seconds (78.6 minutes) over the model that included tooth-level random effects in addition to subject level random effects, which advantages the subject-only random effect model. This suggests that there is substantial variation between subjects and that the inclusion of subject level random effects is necessary to account for this variation.

Based on these results, and with the over-dispersion parameter of the Beta-Binomial models, $\frac{\theta+n_S}{\theta+1}$, all greater than 2 for each tooth type demonstrating the presence of over-dispersion, the Beta-Binomial was the only model considered for the spatial grouping. The following models are the spatial models and best performing non-spatial with results in Table 2:

(1) Beta-Binomial regression with no random effects/fixed effects only

(2) Beta-Binomial with subject level random effects

(3) Beta-Binomial hierarchical model with subject level random effects

(4) Beta-Binomial SANOVA model with subject level random effects

(5) Beta-Binomial with CAR model

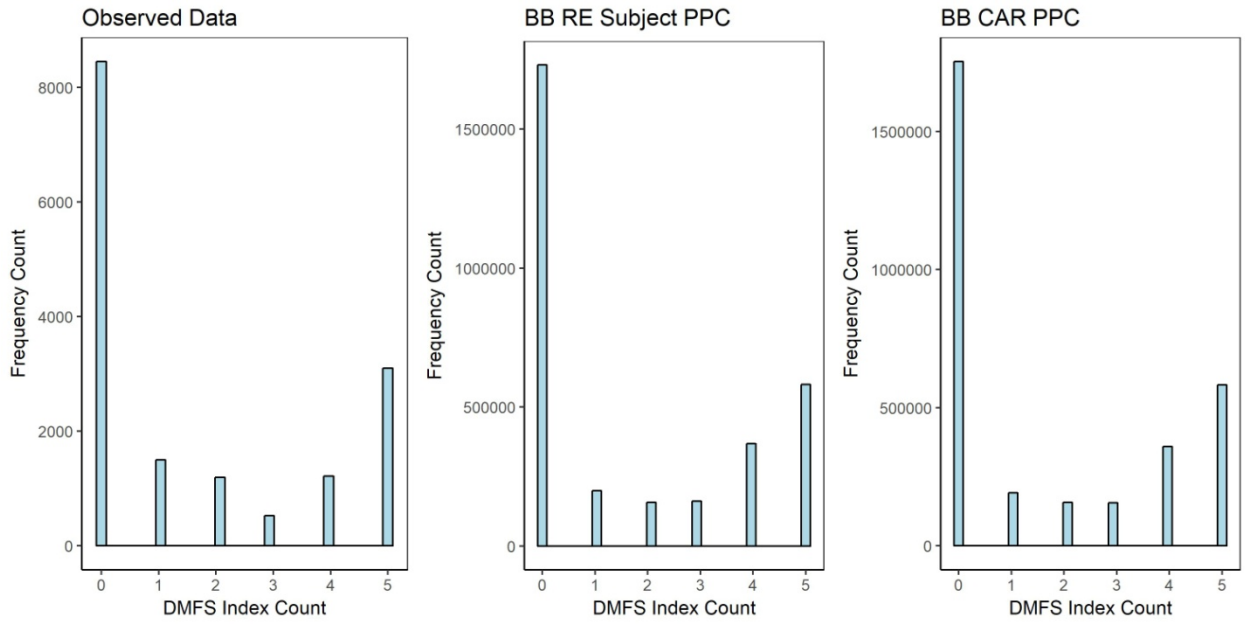
Table 2: WAIC and Runtimes for Final Models

Model	WAIC	Run Time in Seconds
1	36,149	6,839
2	28,724	29,033
3	28,698	73,743
4	28,654	109,099
5	25,110	57,193

Runtimes increased with model complexity leading to SANOVA spatial model having the longest runtime. The hierarchical and the SANOVA models were run with and without the subject level random effects and inclusion of subject level random effects improved model fit considerably. Additionally, the CAR model was run with an adjacency matrix that only included neighbor specifications for teeth directly adjacent/touching each other on the same jaw, but our neighbor/adjacency specification including the cross-mouth tooth and all the incisors was superior in WAIC. Comparisons of WAIC described by Vehtari et al. (2016, 2017) were performed between all models, which resulted in the CAR model being superior. There was not a substantial difference in model fit between the subject level random effects, SANOVA, and the

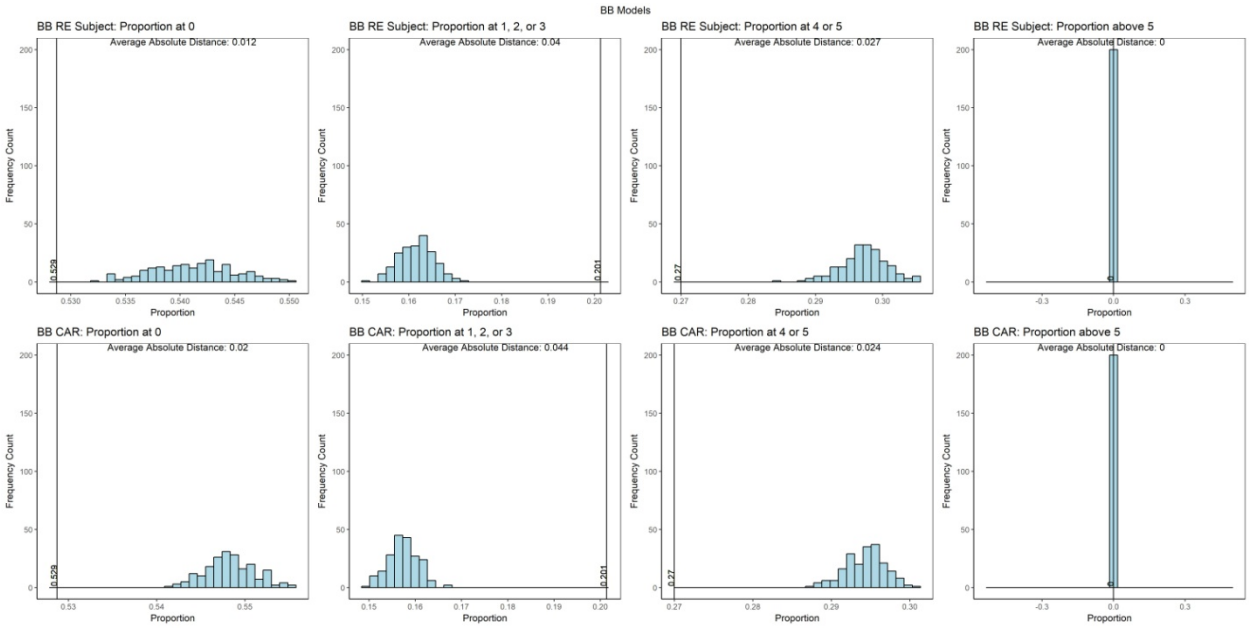
hierarchical model. However, examination of the posterior predictive checks show little difference between the CAR model and the subject level random effect model Figure 16:

Figure 16: Distribution of DMFS Index for Random Effect Comparison



And Figure 17 shows the proportion distributions:

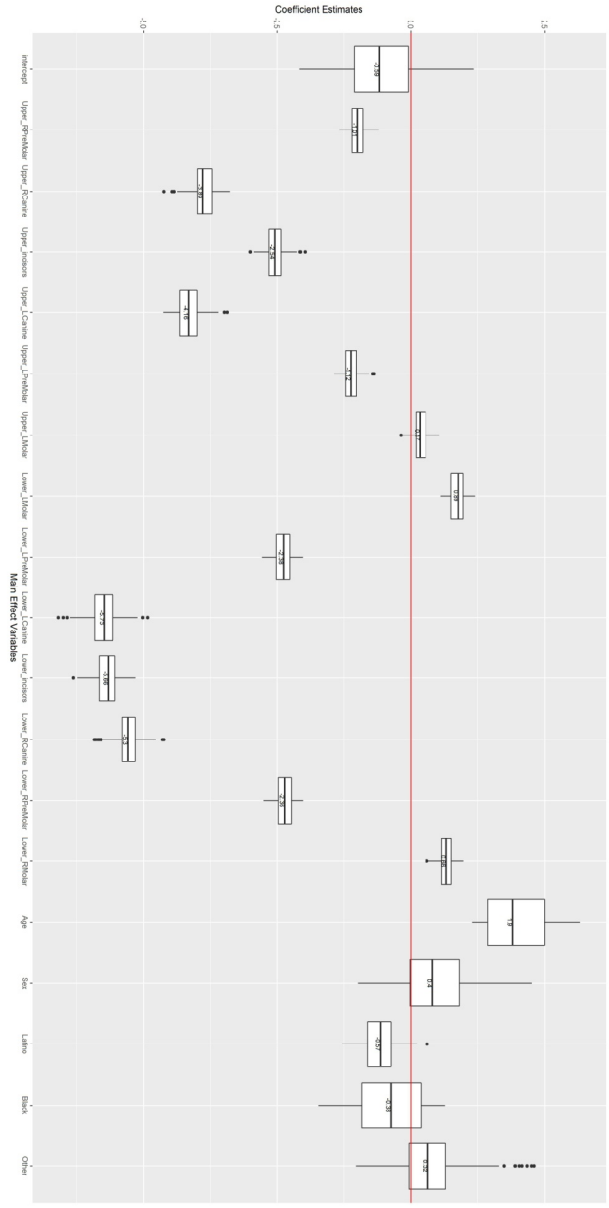
Figure 17: Proportion of Simulations in Each Category for Random Effect Comparison



The two models exhibit similar posterior simulation distributions, but the starkest difference between the two models being a 3614 difference in WAIC. Figure 11 displays a boxplot representation of the posterior estimates of the main effect parameters for the Beta-Binomial CAR model:

Figure 18:

Figure 18: Coefficient Posterior Distributions



The fixed effects, because of the logit link function, can be interpreted as multiplier of the odds of having a DMFS value one point higher on a given tooth. The results reflect what would be expected based on our group's previous analysis where the molars exhibit the highest probability of a DMFS event followed by the pre-molars and the canine and incisors a much lower

probability. Additionally, teeth on the madibular jaw (i.e. the lower jaw) have a noticeably lower probability of a DMFS event. Age, which was center and scaled, has a positive coefficient, indicating that greater age was associated with higher DMFS score on a tooth. Most models showed that females are more likely to have DMFS event, as fewer women use methamphetamine and tended in this sample to be more extreme in their use of the drug. Race/ethnicy variables generally suggested that African-Americans and Latinos evidence lower probabilities for a DMFS event than White and Others controlling for other factors. Ultimately, although the parameter estimates varied slightly, but the conclusions ultimately drawn from the models were similar.

3.4.2 Conclusion

In this chapter, we analyzed spatially correlated count data utilizing a Beta-Binomial parameterization that controls for over-dispersion, and we experimented with ways to control for the spatial nature of dental data. The flexibility of the Beta-Binomial model to create a "U" shape in the aggregated count distribution of DMFS index led to it outperforming competing Poisson and Negative-Binomial models. In situations where the density of data values is greatest at the extremes of the range of possible values, with a lesser amount of data distributed across intermediate values, as in dental DMFS index data, the Beta-Binomial model should be considered. Additionally, inclusion of subject level random effects, inducing a compound-symmetry correlation structure among teeth within the same mouth, greatly improved model fit and demonstrated the evidence in the data of a large amount of variation between subjects in the overall level of dental decay. Experimentation with spatial modelling, specifically using spatial random effects with a CAR prior distribution, resulted in better model fit. This appears to be due

to the structure of the CAR prior adjacency matrix allowing for a refinement of the compound symmetry and limiting the influential neighbors to adjacent teeth on the same jaw and cross-mouth neighbors. Ultimately, due to the strong spatial correlation and over-dispersion relative to the Poisson model present in dental caries data using the DMFS index, the Beta-Binomial regression models containing spatial random effects with a CAR prior distribution controlling for adjacent and cross-mouth neighbors showed improvements over other models and suggests that models similar to it should be considered in the analysis of dental count data when computational time permits. If computational or technical ability is a limiting factor, subject level random effect should be included in place of the CAR prior distribution spatial random effects for improvement over fixed-effect only models.

4 A Pattern-Mixture Model for Spatially Correlated Clustered Count Data with Application to Dental Caries Data

Missing teeth are regularly encountered in the analysis of dental data due to decay or dental procedures. A standard approach to missing teeth is to assign maximal scores on the tooth-specific count variable used to summarize dental caries, namely the total number of decayed, missing, or filled surfaces (DMFS) on each tooth (DMFS index) (Bodecker, 1939; Darby and Walsh, 2003), due to all surfaces being missing. These maximal scores are often not achieved by filled and decayed teeth, resulting in an elevated and possibly disproportionate level of influence for missing teeth in subsequent analyses. Using a sample of 571 methamphetamine users from Los Angeles, we model DMFS index accounting for missing teeth more explicitly by modeling the mechanism giving rise to missing data. When controlling for the missing data mechanism, both selection and pattern-mixture models were utilized to assess the sensitivity of underlying assumptions. A pattern-mixture model was constructed to allow for spatial stratification and control for different patterns of dental decay observed in the data. Posterior predictive distributions, were evaluated to assess overall fit, inference, and compatibility of conclusions with simulated data. Ultimately, by controlling for the spatial structure of dental data and the missing data mechanism, the pattern-mixture model with spatial stratification resulted in inferences more consistent with dental decay patterns in the mouth and should be considered when dental caries data are analyzed.

4.1 Introduction

Dental caries, which arise from a nested sampling structure on teeth of different types within individuals, were shown in our previous work to exhibit a complex spatial structure and be more accurately modeled utilizing spatial statistics methodology than invoking models without spatial structure. Additionally, based on our groups previous work, the Beta-Binomial model evidenced superior performance for modeling these data due to the bimodal nature of the DMFS index distribution. However, even with spatial relationships, better-tailored likelihoods, and subject-level random effects improving overall model fit, it was noted that many subjects had missing teeth. Missing teeth can play a key role in parameter estimates because each tooth represents a count variable describing the number of decayed, missing, or filled surfaces (DMFS) on each tooth (DMFS index), and standard practice has been to associate maximal scores with missing teeth due to all surfaces being missing. Some dentists have advocated for changes to the DMFS index to reduce the influence of missing teeth (Broadbent and Thomson, 2006). However, building on methods that have been utilized to analyze dental data (Todem, 2012; Bandyopadhyay, Reich, and Slate, 2011), this chapter seeks to handle missing teeth by modeling the missing-data mechanism more explicitly and to determine which missing data approach results in more accurate inference and conclusions for dental caries data.

The data for the motivating example were collected from 571 methamphetamine users in Los Angeles County recruited over a 2-year period from dental clinics associated with two large community health centers: a) the AIDS Project, Los Angeles (APLA) center that primarily serves a sociodemographically diverse group of individuals with HIV/AIDS, and b) the Mission Community Hospital (Mission) in the San Fernando Valley that caters to a large, underserved

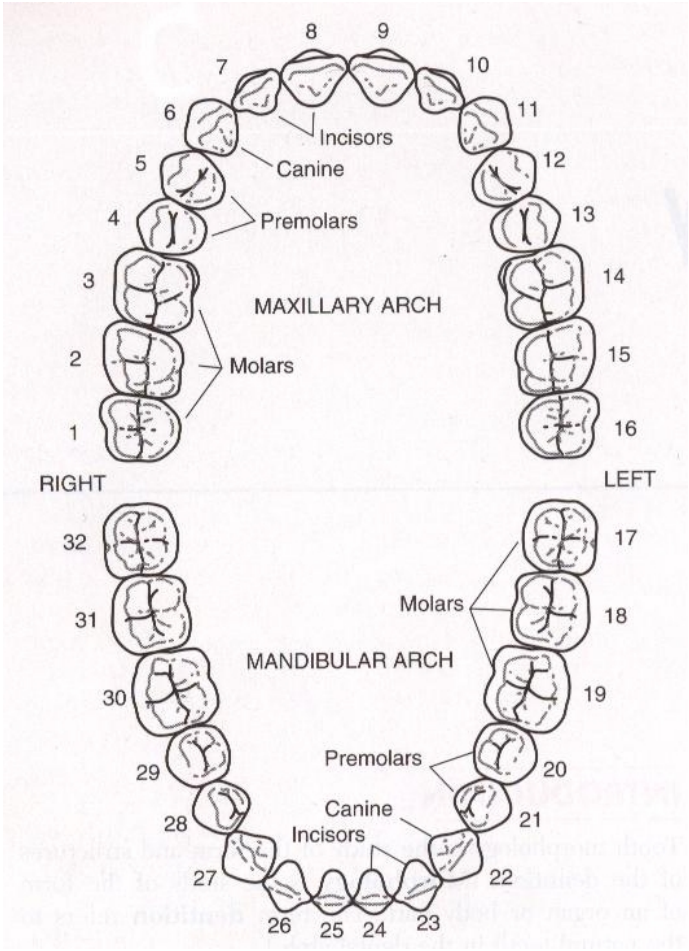
migrant population. Approximately 69% of the subjects were recruited at the APLA clinic and the remainder at the Mission clinic. Subjects were at least 18 years of age, spoke English or Spanish, had described themselves as methamphetamine users, submitted to an extensive 10-year drug history questionnaire, used methamphetamine in the past 30 days, and were able to undergo a detailed dental examination and psychosocial assessments. Additional details on the data collection process can be found in related reports (e.g. Shetty, Harrell, Clague, Murphy, Dye, and Belin, 2016).

The majority of the study sample (n = 571) was male (n = 460), African-American or Hispanic (42.3% and 31.2% respectively), older than 30 years of age (Mean age = 44.5 years, SD = 9.6), and had completed high school (n=401). A large subset of the study participants (n=147) were HIV positive. Many of the methamphetamine users were current cigarette smokers (68.8%). Based on self-reported history, 64% of the participants used methamphetamine frequently (reported using 15 or more days a month on average) and, on average, had used meth for at least 10 of the preceding months (Mean = 10.1, SD=2.1). The average age of methamphetamine use initiation was 28.5 years (SD=10.5) and smoking was the most common mode of use, with 53% reporting smoking as their exclusive mode. The study participants consumed an average of 3.5 sugary drinks a day (SD = 2.2).

Dental decay, the outcome variable, was represented using the DMFS index, which is the total count of decayed, missing, or filled surfaces per tooth. Each tooth has a maximal count of 4 (Canine and Incisors) or 5 (Molars and Premolars) depending on tooth type (Darby and Walsh, 2003). Based on the implicit understanding that third-molar extraction is often motivated by

orthodontic considerations having little to do with dental caries, it is customary in the dental literature to remove the third molars (tooth numbers 1, 16, 17, and 32) from the analysis, resulting in 28 teeth, or outcomes, per subject. A visual representation of how the mouth is numbered can be seen in Figure 19. With 571 subjects, there were a total of 15,988 teeth in the dataset. Additionally, missing teeth are assigned a count of 4 or 5, depending on the tooth type, based on all surfaces having been recorded as missing.

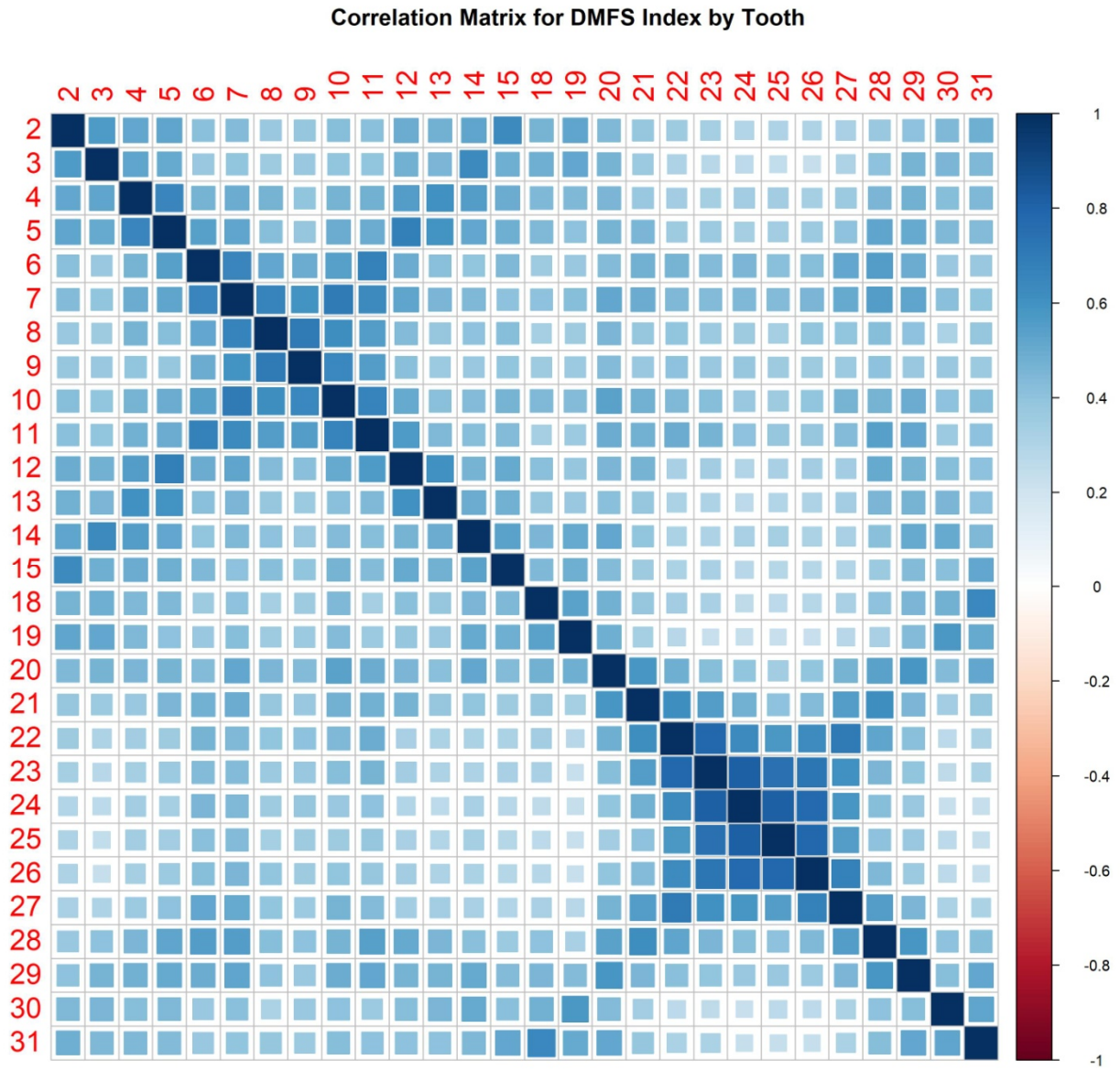
Figure 19: Tooth Numbering Scheme



(Taken from https://buildinggreatsmiles.com/blog/what-tooth-number-is-this-tooth/)

The presence of correlation among DMFS indexes on teeth in the same mouth was evident from previous analysis of these data where missing teeth were treated as maximal DMFS index Figure 20. In Figure 20, the numbers correspond with the tooth numbering in Figure 1 where third molars, teeth 1, 16, 17, and 32, are not included and the blue shading indicates a positive correlation between the counts of DMFS index for each tooth in the sample. This correlation structure evidenced three strong relationships: (1) local neighboring teeth on the same jaw (2) cross-mouth neighbor relationship on the same jaw (3) strong correlation among all 4 incisors on the same jaw. These spatial relationships were important in our previous analysis when controlling for dental decay patterns.

Figure 20: Correlation Matrix



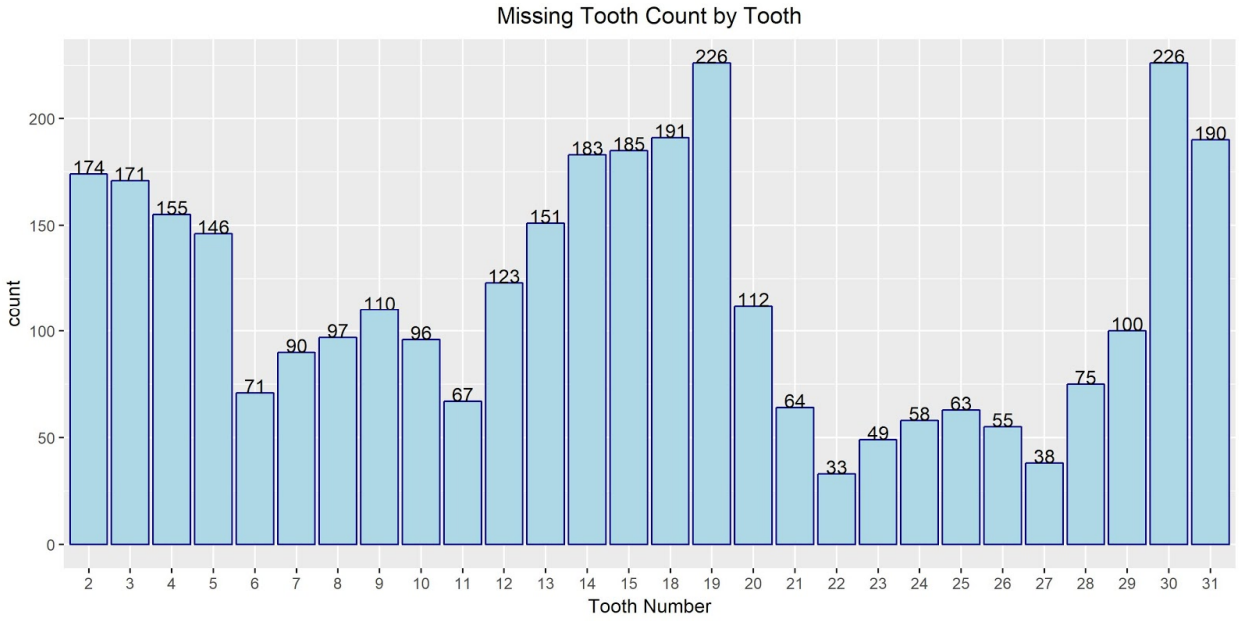
The evidence of dental decay patterns coupled with the assumption that at least one pattern of missingness results from teeth being lost due to dental decay make these patterns informative for the pattern-mixture models to be used later.

With the introduction of the data, DMFS index data, and the correlation observed among teeth, the remainder of the chapter will proceed as follows: 1) Introduction to the missing data present in the methamphetamine data set and a review of assumptions made about the missing data mechanism, 2) Review of Bandyopadhyay, Reich, and Slate's (2011) parameterization of the Beta-Binomial model and its flexibility allowing for a superior model fit when applied to dental caries data, 3) Overview of missing data, selection, and pattern-mixture models, 4) Prior specification for all models, programming specifics, and simulated data to assess impact of modeling and prior choices, 5) Implementation of models on methamphetamine data set , and 6) Discussion and conclusions.

4.1.1 Influence of Missing Teeth

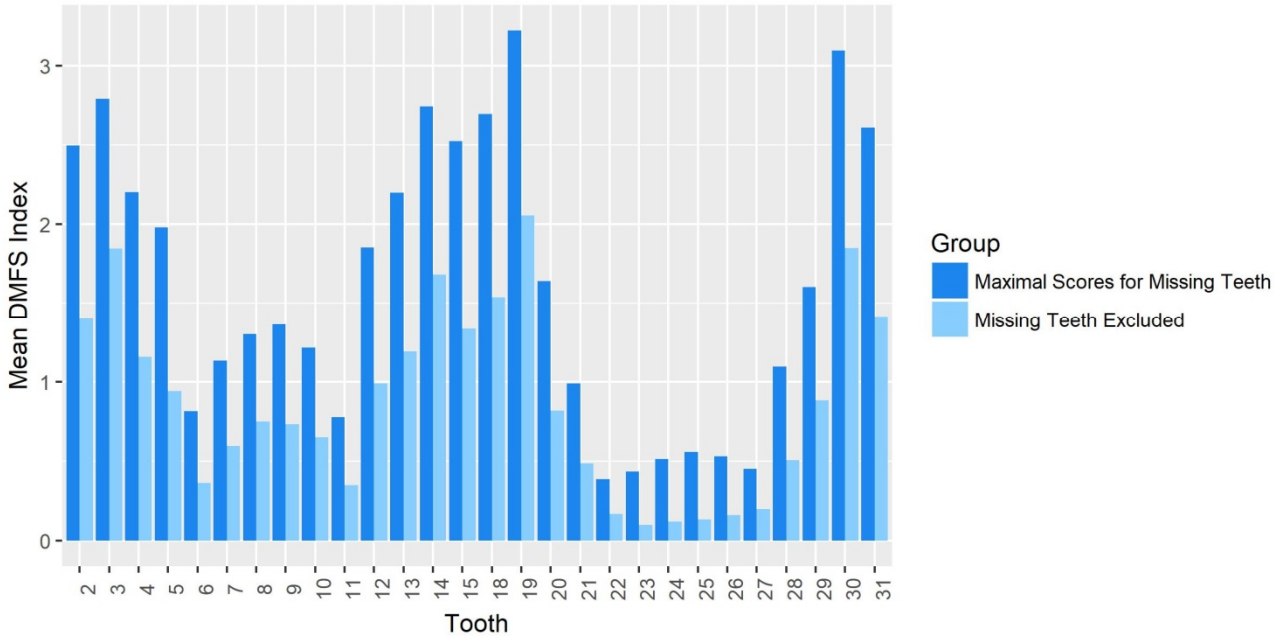
To assess the importance of missing teeth, we begin by exploring the prevalence of missing teeth in the available methamphetamine-using population must be investigated. Below in Figure 21 are the counts of missing a specific tooth, where the numbers on the x-axis correspond to those in Figure 19 with third molars removed, revealing evidence of a large number of missing teeth in the sample.

Figure 21: Missing Tooth Count by Tooth



Subjects average 5.78 (SE = 0.29) overall missing teeth with 2.71 (SE = 0.11) attributed to molars and 1.62 (SE = 0.07) premolars. Canines, 0.37 (SE = 0.04), and incisors, 1.08 (SE = 0.05), experienced missing teeth with much less regularity. The importance of missing teeth, contributing to the average DMFS index for a tooth, is evident in Figure 22 where the mean DMFS index for each tooth is compared when missing teeth are given maximal scores and when they are removed.

Figure 22: Mean DMFS Index



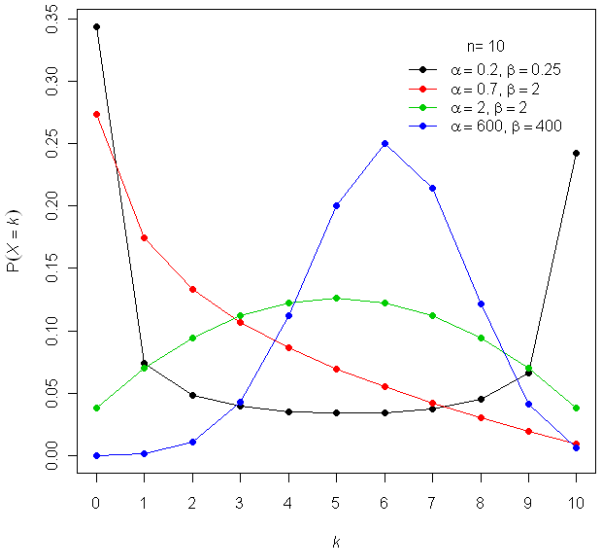
It is clear from Figure 22 that the means drop considerably, showing that missing teeth have a large influence on mean DMFS index given their maximal scores. In our motivating dataset of 15,988 teeth (571 subjects with 28 teeth per mouth), 3,299 teeth were recorded as missing. Of the teeth that were observed, they had the following distribution: 8,453 (DMFS Index = 0), 1,500 (DMFS Index = 1), 1,195 (DMFS index = 2), 524 (DMFS index = 3), 389 (DMFS Index = 4), and 628 (DMFS Index = 5). This resulting distribution is "U" shaped, being bimodal with density at the ranges of the count. Allocating the 3,299 missing teeth to this distribution would clearly be impactful, especially if they are assigned maximal scores of 4 and 5.

4.1.2 Beta-Binomial Model

Previously, it was shown that the Beta-Binomial model outperformed Poisson and Negative-Binomial models on dental caries data. The improved model fit appeared to be attributable to the flexibility and natural truncation of the Beta-Binomial model and its ability to create a "U" shape

when α and β parameters are less than 1. The "U" shape enables the Beta-Binomial model to fit bimodal distributions when substantial amounts of the density are placed at the count ranges. This is particularly important when modeling DMFS index data where missing teeth are assigned maximal scores of 4 or 5 depending on tooth type as shown in Figure 23:

Figure 23: Beta-Binomial Distribution



This figure clearly demonstrates the ability of the Beta-Binomial distribution to achieve the "U" shape, which may be important for modeling dental caries data. However, the distribution modeled in the previous chapter was based on missing teeth receiving maximal DMFS index. In this chapter, we aim to use missing data methods to control for the missing-data mechanism, which could greatly differ from the maximal score assignment and possibly change the distribution of the posterior predictive checks for tooth count for DMFS index. Poisson and Negative-Binomial models may benefit from imputation in model fit and this will be assessed in the simulation data as well as the real data using posterior predictive checks.

This Beta-Binomial model considered here anticiaptes the data being bounded, or truncated, between 0 and a maximal sore of 4 or 5 with possible over-dispersion in a Poisson model. Therefore, we utilize the specification proposed by Bandyopadhyay, Reich, and Slate (2011) where DMFS index for tooth t of subject s is represented as $y_s(t): s = 1,2,\ldots,571, t = 1,2,\ldots,28$ and $n_t = 4$ for canines and incisors and 5 for molars and premolars:

$$logit(\mu_s(t)) = X_{s(t)}^T \beta$$

$$p_s(t) \sim Beta(\theta(t)\mu_s(t), \theta(t)[1 - \mu_s(t)])$$

$$y_s(t) \sim Binomial(n_t, p_s(t))$$

$X_{s(t)}$ represents the vector of fixed effects for subject s and tooth t, which include tooth and subject level covariates. Additionally, given this parameterization of the Beta-Binomial model, the over-dispersion parameter is $\frac{\theta(s)+n_t}{\theta(s)+1} \in (1, n_t)$ where n_t will change based on tooth type (4 or 5).

4.2 Missing Data

When handling a missing data problem, it is important to consider the mechanism that induces the missing data for both modeling and imputation considerations (Little and Rubin, 2002; Rubin, 2004). To better understand the mechanism that leads to missing data, we consider the joint distribution of the outcome variable $Y = (Y_1, Y_2, \ldots, Y_n)$ and a missing/observed indicator variable $M = (M_1, M_2, \ldots, M_n)$ where $M_i = 1$ if tooth i is missing and 0 otherwise:

$$P(Y, M|X, \beta, \theta)$$

In this formula, $Y = \{Y_{obs}, Y_{mis}\}$ is composed of the observed and the missing data, β are the parameters of the model of interest, θ are the parameters for this missing-data mechanism function, and X are the observed covariates. Conditional independence and factorizations of this joint conditional distribution are dependent on assumptions made about the missing-data mechanism, which ultimately impacts the modeling of the data.

4.2.1 Missing Completely at Random

Under Rubin's classification system (Rubin, 1976), the missing-data mechanism can be either missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). This assumption made about the missing-data mechanism directly affects how the joint distribution, specified in the previous section, is modeled.

In a missing data situation where outcome values, Y, are MCAR, the probability of missing data in Y is unrelated to other measured or unmeasured variables, and underlying values of the incomplete data in Y. This implies that a parameter, independent of other parameters in the model and independent of the observed and unobserved data, generated the missing data:

$$P(Y, M|X, \beta, \theta) = P(Y_{obs}, Y_{mis}|X, \beta)P(M|\theta, X, Y_{obs}, Y_{mis}) = P(Y_{obs}, Y_{mis}|X, \beta)P(M|\theta)$$

$$\int P(Y_{obs}, Y_{mis}|X, \beta)P(M|\theta)dY_{mis} = P(Y_{obs}, |X, \beta)P(M|\theta)$$

The MCAR assumption creates the equivalence $P(M|\theta, X, Y_{obs}, Y_{mis}) = P(M|\theta)$ because the missingness indicator M is not dependent on any observed or unobserved variables. Missing data in the MCAR case are unsystematic and the observed data can be viewed as a random subsample of the hypothetically complete data. MCAR is considered a strict assumption because it requires that missingness be unrelated to the any of the study variables and often this is unlikely to be satisfied in non-simulated data unless imposed in the design of data collection.

4.2.2 Missing at Random

A less strict assumption for the missing-data mechanism, is the missing at random (MAR) assumption. Under the MAR assumption the probability of missing data in Y is fully explained by the observed data and not dependent on the underlying values of the incomplete data in Y allowing for the equivalence $P(M|\theta, X, Y_{obs}, Y_{mis}) = P(M|\theta, X, Y_{obs})$:

$$P(Y, M|X, \beta, \theta) = P(Y_{obs}, Y_{mis}|X, \beta)P(M|\theta, X, Y_{obs}, Y_{mis}) = P(Y_{obs}, Y_{mis}|X, \beta)P(M|\theta, X, Y_{obs})$$

$$\int P(Y_{obs}, Y_{mis}|X, \beta)P(M|\theta, X, Y_{obs})dY_{mis} = P(Y_{obs}, |X, \beta)P(M|\theta, X, Y_{obs})$$

MCAR and MAR assumptions are related to the concept of ignorable missing data (Rubin, 1976 and 1978; Mealli and Rubin, 2015). From a Bayesian perspective, if the missing data are ignorable, which could occur with either an MCAR or MAR mechanism, it means that the posterior distribution is conditionally independent, given the observed data, of the indicators for missingness status (Mason, 2009). The conditional independence implies that there is no need to explicitly model the missing-data mechanism represented by $P(R|\theta)$ under MCAR and $P(R|\theta, X, Y_{obs})$ under MAR to achieve valid Bayesian or likelihood-based inference.

For analysis of our dental caries data, we start by assuming the missingness is ignorable and model the data, including the imputation process, without jointly modeling the missing-data mechanism. In Bayesian analysis this is done by replacing the missing data with parameters and then running the model as if the data were fully observed (Gelman et al., 2013). This assumes that imputation model and the model run on the data are congenial (Meng, 2004). Additionally, this is the akin to training the model on the complete data and using observations from the posterior distribution to fill in the missing values. In the methamphetamine study data, the outcome variable, DMFS index, is the only data experiencing missing values. The predictor variables, which are fully observed, include age, sex, race/ethnicity, and tooth type/location in the mouth. The tooth type/location variables are indicator variables for tooth type (molar, premolars, canines, and incisors), jaw (mandibular or maxillary), and left or right side of the mouth, which results in 14 indicators variables with the upper (maxillary jaw) right molars being the reference group. These same covariates will be used in all models in this chapter. Poisson, Negative-Binomial, and Beta-Binomial models will be run and compared on both simulated and actual study data. These models will represent the results under the MCAR and MAR assumptions. The next three sections will discuss the possibility that the missing-data mechanism is non-ignorable along with modeling techniques that can be used in this situation.

4.2.3 Missing Not at Random

When the target posterior distribution depends on the missingness indicators, joint modeling for the data, and the indicators is necessary for valid inference. This scenario is known as a non-ignorable missing-data mechanism (Little and Rubin, 2002). The concept of non-ignorable

missing data, where the missing-data mechanism must be jointly modeled with the data for valid inference, occurs when the missing data are missing not at random (MNAR). Formally, missing data are MNAR when the probability of missingness is systematically related to the hypothetical values that are missing. Therefore, there is no way to simplify the original factorization of the joint distribution that is free of dependence on unobserved quantities:

$$P(Y_{obs}, Y_{mis}, M|X, \beta, \theta)$$

The following two sections discuss ways of factoring the joint distribution to allow for modeling and inference to satisfy scientific questions while controlling for the missing-data mechanism.

4.2.4 Selection Models

In situations where the missing data are MNAR, the missing-data mechanism must be jointly modeled with the data. The selection model (Rubin, 1974) is one way of factoring the joint distribution of Y and R:

$$P(Y_{obs}, Y_{mis}, M|X, \beta, \theta) = P(Y_{obs}, Y_{mis}|X, \beta)P(M|\theta, X, Y_{obs}, Y_{mis})$$

Often, the missing-data mechanism, $P(M|\theta, X, Y_{obs}, Y_{mis})$, is a logistic or probit model and $P(Y_{obs}, Y_{mis}|X, \beta)$ represents the model for the complete data. This model will be adapted to our dental caries data using a logistic regression for the observed/missing indicator and poisson, negative-binomial, and beta-binomial models for the complete data.

4.2.5 Pattern-Mixture Models

Pattern-mixture models (Little, 1992) are an alternative way of jointly modeling (R, Y) where the complete data distribution is a mixture of observed and missing data distributions:

$$P(Y_{obs}, Y_{mis}, M|X, \beta, \theta) = P(Y_{obs}, Y_{mis}|X, \beta, M)P(M|\theta, X)$$

where $P(M|\theta, X)$ is usually a logistic or probit model (we will be using a logistic model for our analysis) and $P(Y_{obs}, Y_{mis}|X, \beta, M)$ is a model that fits the observed data (poisson, negative-binomial, or beta-binomial in our case) .

4.2.6 Alternative Representation of Selection and Pattern-Mixture Models

Both selection and pattern-mixture models were presented in the previous section using the observed/missing indicator variable M. However, M can also be a categorical variable where the different values represent a separate pattern of missing data. This implies that pattern of the missing data is informative about the true value of the missing observation and should be accounted in the analysis. The joint model is similar to the previous formulation, but specific to the pattern of missingness M = m:

$$P(Y_{obs}, Y_{mis}, M|X, \beta, \theta) = P(Y_{obs}, Y_{mis}|X, \beta_m, M = m)P(M = m|\theta, X)$$

This makes the regression parameters pattern-dependent, and a drawback of this approach is that the marginal density of the full data is not explicitly represented. Rather, a mixture of the missingness pattern distributions is required to get the full data distribution:

$$P(Y_{obs}, Y_{mis}|X, \beta, \theta) = \sum_{m} P(Y_{obs}, Y_{mis}|X, \beta_m, M = m) P(M = m|\theta, X)$$

This is similar to the case when M is the observed/missing indicator variable, but requires more summed values to account for all missingness patterns. In dental caries data, the patterns of missigness appear to be informative as teeth decay in certain patterns (Figure 2) and with different levels of severity. The pattern and severity of missing teeth can accordingly inform the modeling process. Prior to any analysis, it was important to assess the different decay patterns in the data. Based on prior research, it has been found that molars are the most likely to experience a DMFS event (Demirci, Tuncer, and Yuceokur, 2010). Premolars have a high probability of DMFS event, but not as high as molars. Canines are the least likely to have a DMFS event and incisors a little more likely than canines. This suggests that dental decay has a natural progression in severity based on which teeth are experiencing DMFS events and, ultimately, tooth loss due to decay.

We separate the data into 5 categories of missing teeth: 1) No missing teeth (N = 128), 2) Molars only tooth type missing (N = 114), 3) Molars and premolars both experiencing missing, but not canine or incisors (N = 116), 4) Molars, premolars, and incisors have missing (N = 57), 5) All tooth types experiencing missing teeth (N = 86). These categories account for 501 out of 571 subjects. The remaining 70 subjects experienced missing teeth in: premolars only (N = 18), canines only (N = 2), incisors only (N = 8), molars and canines (N = 1), molars and incisors (N = 21), premolars and canines (N = 1), premolars and incisors (N = 4), molars, premolars, and canines (N = 14), and molars, canines, and incisors (N = 1).

Further assessment of the 70 subjects showed that the number of non-missing teeth experiencing a DMFS index of 4 or 5 and the number of teeth missing could be used to categorize these data into the already existing 5 groups. The single missing tooth type categories demonstrated similar characteristics to the no missing teeth group (missing only incisors and only canines) and the molars only group (missing only premolars). The two missing tooth types were similar to the molars-only group (molar and canine, molar and incisors) and the no-missing-teeth group (premolars and canines, premolars and incisors). Subjects experiencing loss in the molars, premolars, and canines were most similar to the molars, premolars, and incisors group, but the molars, canines, and incisors subject was placed in the molars only group. With these missing-data patterns established, to address identifiability issues, M patterns were grouped into categories to reflect plausible exchangeability assumptions; this was done three different ways, characterizing three different pattern mixture models as seen in Table 3:

Table 3: Missing-Data Patterns

Model	Exchangeable Subgroups
Pattern-Mixture Model 1	• No missing teeth • Missing teeth in either canines or incisors • Missing teeth in premolars and either canines or incisors

Pattern-Mixture Model 2	• Missing teeth in either premolars or molars • Missing teeth in molars and incisors and/or canines
Pattern-Mixture Model 3	• Missing teeth in both molars and premolars • Missing teeth in molars, premolars, and incisors or canines • Missing teeth present in all tooth types

These patterns will be used in our analysis for the pattern-mixture models that account for missing-data patterns.

4.2.7 Averaging Mixture Method for Pattern-Mixture Model Parameter Estimates

Overall population parameter estimates were determined using averages across patterns (Little, 1992 and 1995; Hogan and Laird, 1998). When the quantity $P(m|\theta, X)$ is modeled using the proportion of observations in the pattern m, then the overall population or mixed parameter estimates are $\widehat{\bar{\beta}} = \widehat{\beta_a} + \widehat{\pi_d}\hat{\beta}^D$ where $\widehat{\beta_a}$ is the parameter estimate for the base pattern, $\widehat{\pi_d}$ the proportion of observations in the non-base pattern, and $\hat{\beta}^D = (\widehat{\beta_d} - \widehat{\beta_a})$ the regression parameter for the non-base pattern. The variance estimate defined as $\hat{V}\left(\widehat{\bar{\beta}}\right) = \hat{V}\left(\widehat{\bar{\beta}}\right)_F + \frac{n_d n_c}{N^3}\left(\hat{\beta}^D\right)^2$ where $\hat{V}\left(\widehat{\bar{\beta}}\right)_F$ is the variance of $\widehat{\bar{\beta}}$ treating the sample proportions as known. The variance estimate is obtained using the delta method described by Hogan and Laird (1998). Utilizing this

method simplifies the computation and is equivalent to the mixing of parameters between missingness patterns using the proportion of observations in each pattern as weights and will be implemented for our pattern-mixture models in this chapter.

4.3 Simulation Study

4.3.1 Prior Distributions

Every model run had identical mean structure and the previously mentioned covariates. Imputation models were the same except for the inclusion of either the M or Y variable for pattern-mixture or selection models. Covariate parameters were assigned a N(0, 1) prior distribution and the over dispersion parameters in the negative-binomial and beta-binomial models were run using a uniform distribution between 0 and 10 initially and then a grid of gamma distributions to assess stability of the model. It should be noted that the beta-binomial model achieves more computational stability when using the rjags open source package 'mix' module, which contains a built in dbetabin distribution. This distribution allows for the Beta-Binomial regression parameterization presented previously.

To assess sensitivity of underlying assumptions, the precision on the prior distributions were changed to allow for more flexibility. In particular the overdisperson parameters were given gamma distributions with varying parameter values to allow for more flexibility and make them more informative, smoothing them to 0. The regression parameters were run with larger variance values. All computation was done with open source software R and JAGS (Plummer, 2003) where there were 4 chains having arbitrary starting values with 5,000 total sampling iterations and a burn-in of 2,500 iterations. Autocorrelation plots were examined and the Gelman- Rubin

diagnostic (Gelman and Rubin, 1992) was used to assess convergence of the chains to the same stationary distribution. Model fit in the simulated data was assessed by comparing the parity about zero of the model parameters to the true parameters, the posterior distributions of DMFS index for each model, and plotting the difference between the true probability and the estimated probability of a DMFS experience.

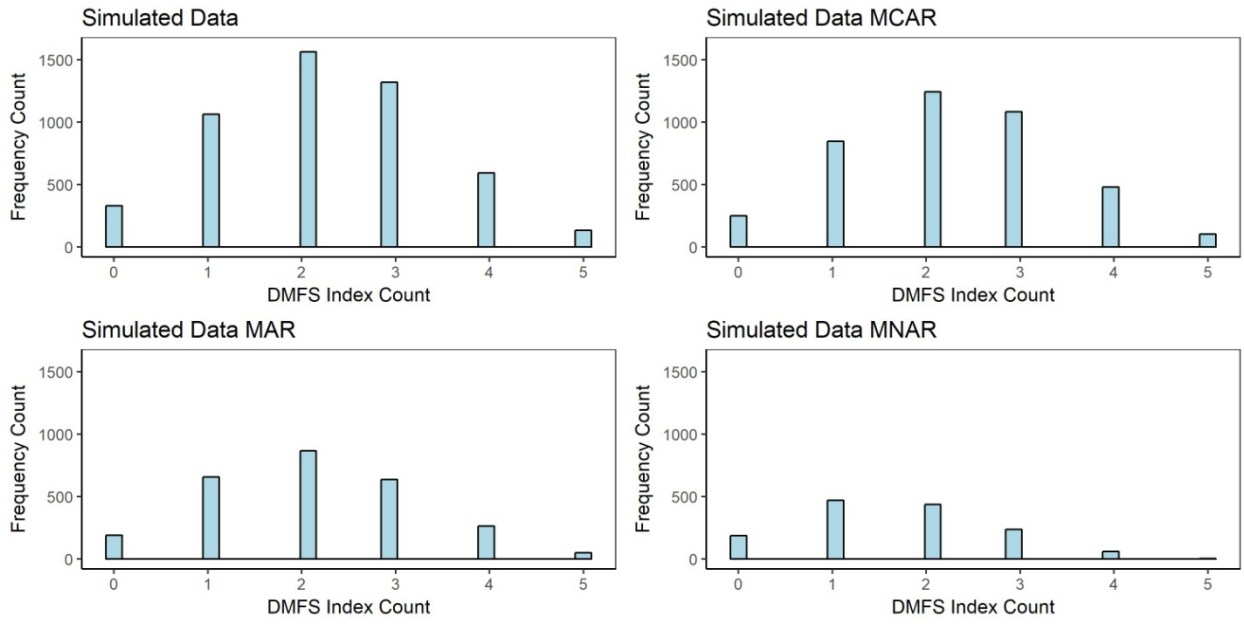
4.3.2 Simulation Study Results

To assess the importance of controlling for the missing-data mechanism, data were simulated and the previously described models, except for the spatial stratified MNAR models with multiple patterns, were fit to these data. The data were generated using a binomial distribution where the generated data was restricted between 0 and 4 or 5 depending on the tooth type variable. This dataset was then used to create new data sets with missing data using MCAR, MAR, and MNAR missing-data mechanisms described below.

Under the MCAR assumption, missing data were generated using a fixed 20.6% probability that the tooth was missing. The value of this fixed probability comes from the proportion of missing teeth in the methamphetamine dataset. This probability was applied to all teeth uniformly regardless of tooth type, position in the mouth, neighboring teeth, or any subject or tooth level covariates. The dataset with an MAR assumption was constructed with the probability of the tooth being missing being dependent on the observed covariates as well as 14 indicator variables specifying tooth type, jaw, and right or left side of the mouth. An example of one of these indicators would be the upper, right, molar. The parameters selected to determine probability of missingness were created based on previous work (Shetty et. al., 2016) and emulating Figure 21.

Under the MNAR assumption the probability of missingness was determined by the observed covariates replicated from the MAR model and a positive association with the value of the tooth. Figure 24 provides a side-by-side comparison of the distribution of teeth with DMFS index counts for the simulated data:

Figure 24: Simulated Datasets' Distributions



In Figure 24, it is apparent that the MCAR data and the complete data distributions are similar as each tooth had the same probability of being missing, in the MCAR case, which means that the MCAR distribution only differs in count. The MAR distribution is also similar to the complete data, but has a larger portion of teeth with DMFS index of 4 and 5 missing. The MNAR distribution, due to the positive association between probability of missing and the tooth's DMFS index, has a large proportion of missing teeth with DMFS index of 2, 3, 4, and 5, shifting the mass of the observed data toward the lower end of the scale.

All models were run for 5,000 iterations with a burn-in of 2,500 iterations and 4 chains on these data. Model fit for models run on the complete data were determined using DIC, posterior predictive checks, and comparison of parameter estimates to the true parameter values. Models run on datasets with missing data were assessed using posterior predictive checks, comparisons of the parameter estimates to the complete data model and true parameter values to assess accuracy of the model inference.

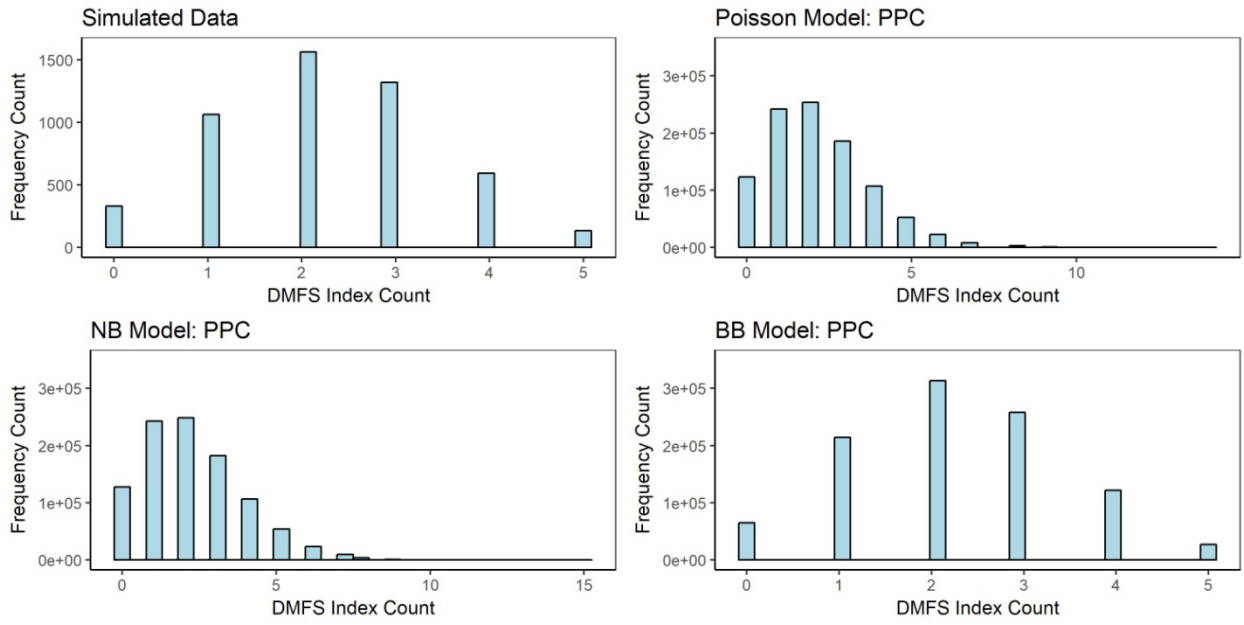
For the complete data, using DIC, the Negative-Binomial (Model 15) and Beta-Binomial (Model 16) models significantly outperformed the Poisson model (Model 14) ($DIC_{Poisson}$ = 33,822, $DIC_{Negative-Binomial}$ = 15,824 , $DIC_{Beta-Binomial}$ = 15,322) with the Beta-Binomial model slightly outperforming the Negative-Binomial model. The Beta-Binomial model demonstrates superior performance in parameter estimation, as seen in the Table 4:

Table 4: Coefficient Estimates for Base Models (No Missing Data)

	True Value	Poisson	Negative-Binomial	Beta-Binomial
intercept	0.2	0.79 (0.7, 0.87)	0.79 (0.7, 0.88)	0.24 (0.11, 0.35)
Age	0.3	0.15 (0.11, 0.19)	0.15 (0.11, 0.19)	0.29 (0.24, 0.34)
Sex	-0.2	-0.09 (-0.14, -0.05)	-0.09 (-0.14, -0.04)	-0.2 (-0.28, -0.14)
Latino	-0.2	-0.1 (-0.14, -0.05)	-0.1 (-0.15, -0.05)	-0.2 (-0.27, -0.13)
Black	-0.3	-0.16 (-0.21, -0.11)	-0.16 (-0.21, -0.1)	-0.31 (-0.38, -0.24)
Other	0.25	0.09 (0.02, 0.15)	0.09 (0.02, 0.15)	0.19 (0.09, 0.28)
Molars 2	0.2	0.28 (0.19, 0.37)	0.28 (0.19, 0.38)	0.13 (-0.01, 0.28)
Molars 3	0.1	0.18 (0.09, 0.28)	0.19 (0.09, 0.29)	-0.08 (-0.22, 0.05)
Molars 4	0.4	0.37 (0.28, 0.47)	0.37 (0.28, 0.47)	0.37 (0.22, 0.52)
Premolars 1	-0.1	0.2 (0.11, 0.3)	0.2 (0.11, 0.3)	-0.05 (-0.19, 0.1)
Premolars 2	-0.2	0.11 (0.02, 0.2)	0.11 (0.01, 0.21)	-0.24 (-0.38, -0.09)
Premolars 3	-0.1	0.15 (0.06, 0.25)	0.15 (0.05, 0.26)	-0.15 (-0.29, -0.001)
Premolars 4	-0.1	0.18 (0.1, 0.28)	0.19 (0.09, 0.29)	-0.09 (-0.23, 0.06)
Canine 1	-0.5	-0.22 (-0.35, -0.09)	-0.21 (-0.35, -0.08)	-0.43 (-0.61, -0.26)
Canine 2	-0.55	-0.26 (-0.4, -0.14)	-0.27 (-0.4, -0.13)	-0.51 (-0.69, -0.33)
Canine 3	-0.3	-0.1 (-0.23, 0.02)	-0.11 (-0.24, 0.03)	-0.23 (-0.41, -0.04)
Canine 4	-0.4	-0.19 (-0.32, -0.07)	-0.19 (-0.32, -0.06)	-0.38 (-0.55, -0.19)
Incisors 1	-0.5	-0.23 (-0.32, -0.14)	-0.23 (-0.32, -0.14)	-0.45 (-0.58, -0.33)
Incisors 2	-0.4	-0.21 (-0.3, -0.13)	-0.21 (-0.31, -0.12)	-0.43 (-0.55, -0.3)

The Poisson and Negative-Binomial Models invert the parity on four parameters while the Beta-Binomial Model inverted the parity on a single parameter. The Beta-Binomial model's ability to achieve accurate parameter estimates and a lower DIC than the Poisson and Negative-Binomial models is likely due to the increased flexibility of the model and limiting of outlier points by restricting the count to be less than 5. To further assess model fit, 200 generated sample data sets were constructed using the parameters from evenly spaced iterations of the Gibbs Sampling procedure (every 25^{th} iteration, with the evidence suggesting that there was not significant autocorrelation between samples as determined by the autocorrelation function). Figure 25 below displays the posterior predictive checks for the three models compared to the simulated data:

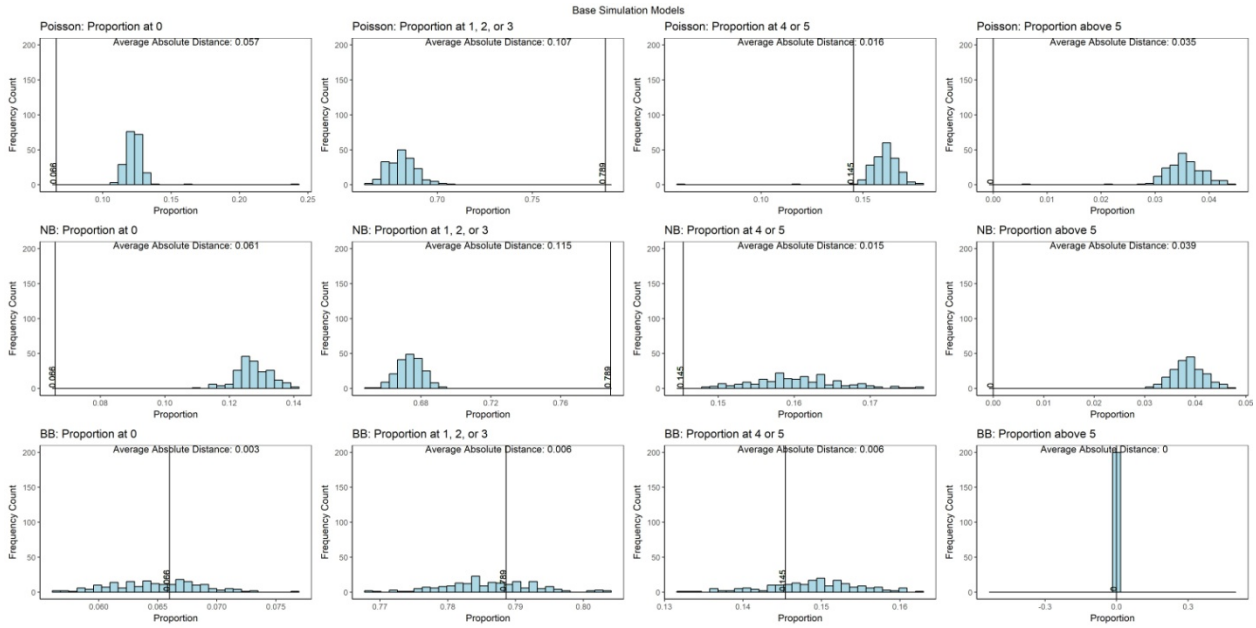
Figure 25: Simulated and Count Distribution (Base Models)



As can be observed from Figure 25, both the Poisson and Negative-Binomial models place most of the density on the 1 and 2 DMFS index while also placing a non-trivial amount of density on count values beyond the tooth's maximal count (4 or 5 depending on the tooth type). The Beta-Binomial model almost perfectly replicates the simulated data's distribution shape and has a natural truncation that occurs due to the maximal count of the tooth being a necessary input to create a ceiling for the Binomial distribution. To equalize the scale of the posterior predictive checks, each of the 200 generated datasets had the proportion of observations at 0, (1, 2, or 3), (4 or 5), and 5+ calculated. This calculation resulted in 200 proportions being calculated for each category. The four proportions were also calculated for the true simulated data and plotted as a vertical line on Figure 26 below where a histogram is plotted for each category and each row

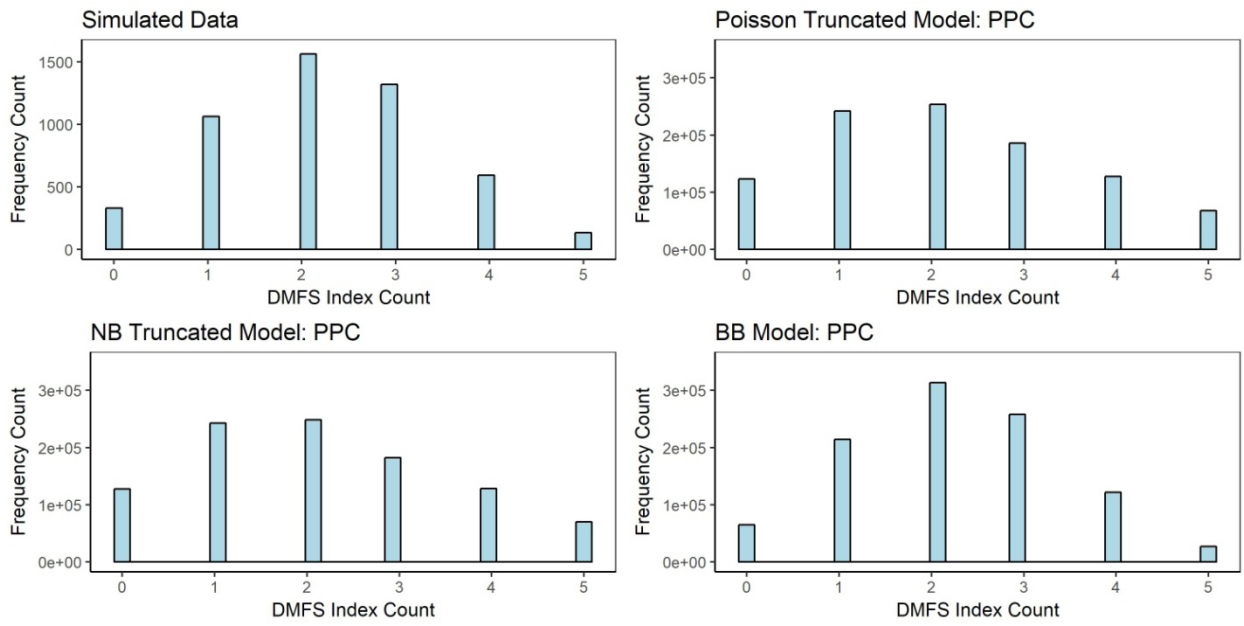
represents a different model to assess the difference between the true proportion in each category and the distribution of the models' proportions in that same category:

Figure 26: Proportion of Posterior Predictive Checks in Each Category (Base Models)



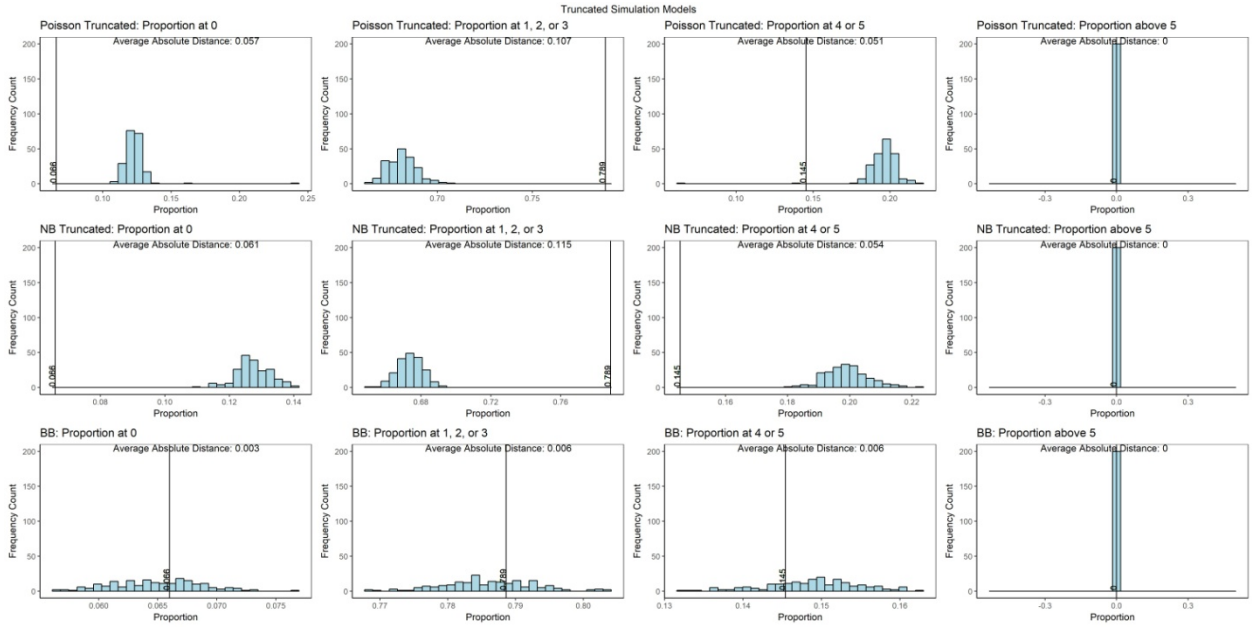
It is clear that the true proportion of observations in each category is contained within the Beta-Binomial model's distribution of proportions. The Negative-Binomial and Poisson models places more density at 0, (4 or 5), and 5+ and less density at (1, 2, or 3), which shows that it is possible that the Beta-Binomial model appears is flexible enough to account for salient features of the simulated DMFS index distribution, but the Negative-Binomial and Poisson models are unable to do so. One advantage the Beta-Binomial Model has is that it truncates the data at the maximal counts, which eliminates the possibility for generated data to exceed the maximal tooth count. However, even with truncation at the maximal tooth count for the Poisson and Negative-Binomial model, the Beta-Binomial model fits the simulated data more closely as seen below in Figure 27:

Figure 27: Simulated and Count Distribution (Truncated Models)



Visually, the models all appear to be more like the simulated data, but with the truncated Poisson and Negative-Binomial placing more of the density at 0, 1, 4, and 5. This allocation of the generated data are more visible in the distribution of the proportions for observations at 0, (1, 2, or 3), (4 or 5), and 5+ for each of the models seen below in Figure 28:

Figure 28: Proportion of Posterior Predictive Checks in Each Category (Truncated Models)



Truncation of the Poisson and Negative-Binomial Models reduce the 5+ difference to zero, but increase the (4 or 5) category's error. All other categories remain the same as they are not affected because all 5+ observations are reassigned values of 4 or 5 depending on that tooth's maximal count. The Beta-Binomial continues to outperform the Poisson and Negative-Binomial models even with truncation, apparently due to the greater flexibility of the Beta-Binomial distribution to model truncated count data.

The corresponding graphics (not shown) of the remaining models run on the simulated data were qualitatively similar to the previously shown graphs. Parameter estimates are reported below.

The differences between the models continued to be evident in the MCAR simulated data as seen below in Table 5:

Table 5: Coefficient Estimates for Base Models (MCAR)

	True Value	Poisson	Negative-Binomial	Beta-Binomial
intercept	0.2	0.79 (0.69, 0.89)	0.79 (0.7, 0.88)	0.25 (0.08, 0.39)
Age	0.3	0.15 (0.11, 0.19)	0.15 (0.1, 0.2)	0.31 (0.23, 0.37)
Sex	-0.2	-0.08 (-0.13, -0.02)	-0.08 (-0.13, -0.03)	-0.17 (-0.25, -0.09)
Latino	-0.2	-0.1 (-0.15, -0.05)	-0.1 (-0.15, -0.05)	-0.21 (-0.28, -0.13)
Black	-0.3	-0.16 (-0.22, -0.1)	-0.16 (-0.22, -0.1)	-0.31 (-0.39, -0.23)
Other	0.25	0.09 (0.01, 0.15)	0.08 (0.01, 0.16)	0.19 (0.08, 0.3)
Molars 2	0.2	0.27 (0.17, 0.38)	0.28 (0.18, 0.38	0.11 (-0.06, 0.27)
Molars 3	0.1	0.17 (0.06, 0.27)	0.17 (0.07, 0.27)	-0.15 (-0.28, 0.03)
Molars 4	0.4	0.35 (0.25, 0.45)	0.35 (0.25, 0.46)	0.29 (0.14, 0.46)
Premolars 1	-0.1	0.2 (0.09, 0.23)	0.2 (0.09, 0.31)	-0.07 (-0.22, 0.11)
Premolars 2	-0.2	0.08 (-0.03, 0.19)	0.09 (-0.02, 0.2)	-0.31 (-0.46, -0.15)
Premolars 3	-0.1	0.13 (0.03, 0.24)	0.14 (0.03, 0.25)	-0.21 (-0.37, -0.04)
Premolars 4	-0.1	0.16 (0.05, 0.27)	0.16 (0.06, 0.27)	-0.16 (-0.31, 0.01)
Canine 1	-0.5	-0.23 (-0.38, -0.08)	-0.23 (-0.37, -0.08)	-0.48 (-0.67, -0.3)
Canine 2	-0.55	-0.27 (-0.42, -0.13)	-0.27 (-0.41, -0.12)	-0.53 (-0.74, -0.31)
Canine 3	-0.3	-0.13 (-0.28, 0.01)	-0.13 (-0.27, 0.01)	-0.28 (-0.48, -0.09)
Canine 4	-0.4	-0.23 (-0.38, -0.09)	-0.23 (-0.37, -0.09)	-0.47 (-0.66, -0.28)
Incisors 1	-0.5	-0.2 (-0.3, -0.11)	-0.2 (-0.3, -0.1)	-0.43 (-0.56, -0.28)
Incisors 2	-0.4	-0.22 (-0.32, -0.12)	-0.22 (-0.32, -0.12)	-0.46 (-0.6, -0.31)

Parameter estimates demonstrate more variability, which is expected with a smaller sample size due to the missing data. However, the Beta-Binomial continues to achieve better sign parity than the other two models, and the posterior predictive checks and proportion distributions in the four categories more similar to the true simulated data. Selection models evidenced similar parameter estimates in Table 6:

Table 6: Coefficient Estimates for Selection Models (MCAR)

	True Value	Poisson	Negative-Binomial	Beta-Binomial
intercept	0.2	0.8 (0.69, 0.89)	0.8 (0.71, 0.89)	0.23 (0.09, 0.37)
Age	0.3	0.15 (0.1, 0.2)	0.15 (0.1, 0.2)	0.3 (0.23, 0.37)
Sex	-0.2	-0.08 (-0.13, -0.03)	-0.08 (-0.13, -0.03)	-0.17 (-0.25, -0.09)
Latino	-0.2	-0.1 (-0.15, -0.05)	-0.1 (-0.15, -0.05)	-0.2 (-0.27, -0.12)
Black	-0.3	-0.16 (-0.22, -0.1)	-0.16 (-0.22, -0.1)	-0.31 (-0.4, -0.22)
Other	0.25	0.08 (0.01, 0.15)	0.08 (0.01, 0.15)	0.19 (0.08, 0.3)
Molars 2	0.2	0.27 (0.17, 0.37)	0.27 (0.17, 0.37)	0.12 (-0.03, 0.28)
Molars 3	0.1	0.16 (0.06, 0.27)	0.16 (0.06, 0.26)	-0.12 (-0.29, 0.03)
Molars 4	0.4	0.35 (0.25, 0.45)	0.35 (0.25, 0.45)	0.32 (0.15, 0.48)
Premolars 1	-0.1	0.2 (0.09, 0.3)	0.19 (0.09, 0.3)	-0.05 (-0.22, 0.12)
Premolars 2	-0.2	0.08 (-0.03, 0.18)	0.08 (-0.03, 0.18)	-0.28 (-0.45, -0.13)
Premolars 3	-0.1	0.13 (0.03, 0.24)	0.13 (0.02, 0.24)	-0.18 (-0.34, -0.02)
Premolars 4	-0.1	0.16 (0.05, 0.26)	0.15 (0.05, 0.25)	-0.14 (-0.3, 0.03)
Canine 1	-0.5	-0.23 (-0.38, -0.08)	-0.24 (-0.38, -0.1)	-0.46 (-0.68, -0.25)
Canine 2	-0.55	-0.27 (-0.42, -0.13)	-0.27 (-0.43, -0.13)	-0.52 (-0.73, -0.32)
Canine 3	-0.3	-0.13 (-0.28, 0.01)	-0.13 (-0.27, 0.01)	-0.27 (-0.49, -0.06)
Canine 4	-0.4	-0.23 (-0.38, -0.09)	-0.23 (-0.38, -0.1)	-0.44 (-0.64, -0.24)
Incisors 1	-0.5	-0.2 (-0.3, -0.1)	-0.21 (-0.3, -0.11)	-0.4 (-0.54, -0.26)
Incisors 2	-0.4	-0.23 (-0.33, -0.12)	-0.23 (-0.32, -0.13)	-0.44 (-0.59, -0.3)

The parameter estimates in the selection models (Model 17, Model 18, Model 19) are only slightly different than their base model counterparts with similar inversions in parity from the true values with the Beta-Binomial model still having the fewest parity inversions. The logistic regression portion of the selection models had similar parameter values where the intercept term had a 95% credible interval that did not contain 0, but all other parameters had 95% credible intervals that did. This result is what would be expected from a selection model applied on data where the missingness probabilities were generated in the MCAR setting. The results of the pattern-mixture model (Model 20, Model 21, Model 22) are presented below with standard deviation estimates:

Table 7: Coefficient Estimates for Pattern-Mixture Models (MCAR)

	True Value	Poisson	Negative-Binomial	Beta-Binomial
intercept	0.2	0.79 (0.21)	0.8 (0.21)	0.23 (0.21)
Age	0.3	0.15 (0.2)	0.15 (0.2)	0.3 (0.2)
Sex	-0.2	-0.08 (0.2)	-0.07 (0.2)	-0.17 (0.2)
Latino	-0.2	-0.1 (0.2)	-0.1 (0.2)	-0.2 (0.21)
Black	-0.3	-0.16 (0.2)	-0.16 (0.2)	-0.31 (0.2)
Other	0.25	0.08 (0.2)	0.09 (0.21)	0.18 (0.21)
Molars 2	0.2	0.27 (0.21)	0.27 (0.21)	0.12 (0.22)
Molars 3	0.1	0.16 (0.2)	0.16 (0.2)	-0.12 (0.22)
Molars 4	0.4	0.35 (0.2)	0.35 (0.21)	0.31 (0.22)
Premolars 1	-0.1	0.2 (0.21)	0.19 (0.21)	-0.04 (0.21)
Premolars 2	-0.2	0.08 (0.2)	0.08 (0.21)	-0.28 (0.21)
Premolars 3	-0.1	0.14 (0.2)	0.13 (0.21)	-0.18 (0.21)
Premolars 4	-0.1	0.16 (0.2)	0.15 (0.21)	-0.13 (0.21)
Canine 1	-0.5	-0.23 (0.21)	-0.23 (0.21)	-0.46 (0.22)
Canine 2	-0.55	-0.26 (0.21)	-0.27 (0.21)	-0.53 (0.22)
Canine 3	-0.3	-0.13 (0.21)	-0.13 (0.21)	-0.26 (0.22)
Canine 4	-0.4	-0.23 (0.21)	-0.23 (0.21)	-0.44 (0.22)
Incisors 1	-0.5	-0.2 (0.2)	-0.21 (0.21)	-0.4 (0.21)
Incisors 2	-0.4	-0.22 (0.2)	-0.22 (0.2)	-0.44 (0.21)

The results reinforce previously presented findings where the Beta-Binomial model achieves more accurate parameter estimates than the other two models. However, the parameter estimates all have similar standard deviation estimates, and these are larger than the previous models. This is due to the mixture of the variances between the two patterns and the lack of information in the data for the missing teeth. Due to estimates for the missing teeth being heavily reliant on the prior distribution, the variance for those parameters is large and increases the overall variance for the combined estimate. However, in situations where the missing data are generated using the MCAR assumption, the pattern-mixture and selection models result in parameter estimates that are comparable to the base models without controlling for the missing-data mechanism.

Outcome data were then removed using a linear regression model, as described at the beginning of section 4.3.2, resulting in MAR data with the following table reporting the results without controlling for the missing-data mechanism:

Table 8: Coefficient Estimates for Base Models (MAR)

	True Value	Poisson	Negative-Binomial	Beta-Binomial
intercept	0.2	0.81 (0.7, 0.93)	0.81 (0.7, 0.93)	0.25 (0.1, 0.46)
Age	0.3	0.15 (0.09, 0.2)	0.15 (0.09, 0.21)	0.29 (0.2, 0.4)
Sex	-0.2	-0.11 (-0.18, -0.04)	-0.11 (-0.18, -0.04)	-0.23 (-0.32, -0.13)
Latino	-0.2	-0.09 (-0.15, -0.02)	-0.09 (-0.15, -0.02)	-0.17 (-0.27, -0.08)
Black	-0.3	-0.17 (-0.24, -0.09)	-0.16 (-0.24, -0.09)	-0.32 (-0.4, -0.21)
Other	0.25	0.09 (0.01, 0.18)	0.1 (0.01, 0.18)	0.21 (0.07, 0.34)
Molars 2	0.2	0.23 (0.08, 0.38)	0.23 (0.07, 0.39)	0.04 (-0.2, 0.25)
Molars 3	0.1	0.26 (0.11, 0.41)	0.25 (0.1, 0.42)	0.1 (-0.11, 0.33)
Molars 4	0.4	0.34 (0.19, 0.48)	0.34 (0.19, 0.49)	0.31 (0.1, 0.53)
Premolars 1	-0.1	0.2 (0.07, 0.33)	0.2 (0.06, 0.34)	-0.02 (-0.21, 0.16)
Premolars 2	-0.2	0.11 (-0.02, 0.24)	0.11 (-0.03, 0.25)	-0.22 (-0.39, -0.03)
Premolars 3	-0.1	0.15 (0.02, 0.28)	0.15 (0.02, 0.28)	-0.13 (-0.32, 0.06)
Premolars 4	-0.1	0.17 (0.04, 0.3)	0.17 (0.05, 0.3)	-0.09 (-0.27, 0.09)
Canine 1	-0.5	-0.17 (-0.33, -0.02)	-0.17 (-0.33, -0.01)	-0.32 (-0.52, -0.11)
Canine 2	-0.55	-0.26 (-0.42, -0.1)	-0.27 (-0.44, -0.1)	-0.5 (-0.72, -0.26)
Canine 3	-0.3	-0.13 (-0.3, 0.03)	-0.13 (-0.3, 0.04)	-0.24 (-0.48, -0.01)
Canine 4	-0.4	-0.2 (-0.35, -0.05)	-0.2 (-0.36, -0.04)	-0.37 (-0.6, -0.17)
Incisors 1	-0.5	-0.24 (-0.35, -0.13)	-0.24 (-0.35, -0.13)	-0.45 (-0.6, -0.3)
Incisors 2	-0.4	-0.22 (-0.33, -0.11)	-0.22 (-0.34, -0.11)	-0.43 (-0.58, -0.26)

As seen in Table 8, the Beta-Binomial model outperforms the Poisson and Negative-Binomial models. The Beta-Binomial model achieves the correct parity for every coefficient. The Poisson and Negative-Binomial models continue to invert the signs of several coefficients. Additionally, compared to the models not controlling for the missing-data mechanism in the MCAR and complete data setting, the 95% intervals are wider for the parameters, which expresses an

increase in uncertainty and is expected due to the increase in missing data in the MAR simulated dataset.

A selection model run on these data achieved comparable results and parameter estimates:

Table 9: Coefficient Estimates for Selection Models (MAR)

	True Value	Poisson	Negative-Binomial	Beta-Binomial
intercept	0.2	0.64 (0.45, 0.82)	0.63 (0.45, 0.81)	0.29 (0.11, 0.46)
Age	0.3	0.16 (0.08, 0.24)	0.17 (0.08, 0.26)	0.28 (0.2, 0.37)
Sex	-0.2	-0.12 (-0.23, -0.01)	-0.12 (-0.23, -0.01)	-0.23 (-0.34, -0.13)
Latino	-0.2	-0.12 (-0.22, -0.03)	-0.13 (-0.23, -0.02)	-0.18, -0.28, -0.08)
Black	-0.3	-0.2 (-0.3, -0.09)	-0.2 (-0.31, -0.09)	-0.31 (-0.43, -0.2)
Other	0.25	0.03 (-0.13, 0.18)	0.03 (-0.13, 0.19)	0.2 (0.06, 0.34)
Molars 2	0.2	0.25 (0.12, 0.47)	0.26 (0.12, 0.49)	0.001 (-0.26, 0.27)
Molars 3	0.1	0.21 (-0.02, 0.43)	0.21 (-0.03, 0.44)	0.07 (-0.19, 0.32)
Molars 4	0.4	0.46 (0.22, 0.69)	0.47 (0.22, 0.72)	0.27 (0.02, 0.51)
Premolars 1	-0.1	0.18 (-0.02, 0.38)	0.19 (-0.03, 0.4)	-0.06 (-0.29, 0.15)
Premolars 2	-0.2	0.16 (-0.05, 0.35)	0.16 (-0.06, 0.36)	-0.25 (-0.45, -0.02)
Premolars 3	-0.1	0.2 (-0.01, 0.41)	0.21 (-0.01, 0.42)	-0.16 (-0.37, 0.05)
Premolars 4	-0.1	0.08 (-0.14, 0.3)	0.09 (-0.14, 0.3)	-0.12 (-0.32, 0.08)
Canine 1	-0.5	-0.2 (-0.45, 0.04)	-0.2 (-0.46, 0.06)	-0.36 (-0.62, -0.12)
Canine 2	-0.55	-0.29 (-0.53, -0.05)	-0.3 (-0.55, -0.05)	-0.53 (-0.78, -0.28)
Canine 3	-0.3	-0.18 (-0.46, 0.08)	-0.19 (-0.47, 0.08)	-0.29 (-0.55, -0.02)
Canine 4	-0.4	-0.21 (-0.46, 0.08)	-0.22 (-0.47, 0.03)	-0.4 (-0.63, -0.17)
Incisors 1	-0.5	-0.27 (-0.44, -0.1)	-0.27 (-0.45, -0.09)	-0.48 (-0.65, -0.3)
Incisors 2	-0.4	-0.26 (-0.43, -0.08)	-0.26 (-0.44, -0.07)	-0.45 (-0.63, -0.26)

The parameter estimates for the selection model, Table 9, differ little from those obtained from the selection model run on the MCAR dataset. However, the parameter estimates for the logistic regression are different, and their 95% intervals do not contain zero for most estimates. This is expected due to the missing data probabilities being generated using a logistic regression and this

model attempting to model those parameters. The pattern-mixture model framework results are displayed below:

Table 10: Coefficient Estimates for Pattern-Mixture Models (MAR)

	True Value	Poisson	Negative-Binomial	Beta-Binomial
intercept	0.2	0.81 (0.47)	0.83 (0.47)	0.23 (0.49)
Age	0.3	0.16 (0.47)	0.14 (0.46)	0.29 (0.47)
Sex	-0.2	-0.12 (0.47)	-0.11 (0.47)	-0.24 (0.47)
Latino	-0.2	-0.07 (0.46)	-0.1 (0.47)	-0.15 (0.47)
Black	-0.3	-0.16 (0.47)	-0.17 (0.48)	-0.31 (0.47)
Other	0.25	0.09 (0.47)	0.1 (0.47)	0.22 (0.48)
Molars 2	0.2	0.23 (0.47)	0.23 (0.47)	0.04 (0.48)
Molars 3	0.1	0.26 (0.48)	0.25 (0.47)	0.1 (0.49)
Molars 4	0.4	0.33 (0.47)	0.34 (0.48)	0.29 (0.48)
Premolars 1	-0.1	0.2 (0.47)	0.21 (0.47)	-0.02 (0.48)
Premolars 2	-0.2	0.12 (0.47)	0.12 (0.47)	-0.21 (0.48)
Premolars 3	-0.1	0.15 (0.48)	0.14 (0.47)	-0.12 (0.47)
Premolars 4	-0.1	0.16 (0.47)	0.17 (0.48)	-0.09 (0.48)
Canine 1	-0.5	-0.17 (0.47)	-0.18 (0.48)	-0.31 (0.48)
Canine 2	-0.55	-0.27 (0.47)	-0.26 (0.47)	-0.49 (0.47)
Canine 3	-0.3	-0.14 (0.47)	-0.12 (0.47)	-0.24 (0.49)
Canine 4	-0.4	-0.18 (0.48)	-0.19 (0.48)	-0.37 (0.48)
Incisors 1	-0.5	-0.23 (0.46)	-0.25 (0.47)	-0.44 (0.47)
Incisors 2	-0.4	-0.23 (0.47)	-0.22 (0.48)	-0.42 (0.47)

Parameter estimates continue to be similar to the true values and little difference is observed compared to the previous pattern-mixture model run on the MCAR dataset. However, standard deviation parameter estimates have more than doubled for all parameters, which indicate that the increase in percent of missing data in the MAR dataset has a large influence of the mixed variance estimate. The posterior predictive checks and the distribution of the proportion in the four categories were similar to the ones displayed previously for the models run on the MAR data.

The MNAR data set differs from the MCAR and MAR datasets in distribution as there are far fewer scores of 4 or 5 in the simulated data. This is due to a positive coefficient on the DMFS index value for the tooth in the logistic regression, which increases the probability of missingness for large DMFS index values. This is evident in the models that do not control for the missing-data mechanism:

Table 11: Coefficient Estimates for Base Models (MNAR)

	True Value	Poisson	Negative-Binomial	Beta-Binomial
intercept	0.2	0.63 (0.45, 0.8)	0.63 (0.46, 0.82)	-0.13 (-0.34, 0.08)
Age	0.3	0.16 (0.08, 0.24)	0.17 (0.09, 0.25)	0.27 (0.15, 0.38)
Sex	-0.2	-0.12 (-0.22, -0.004)	-0.12 (-0.22, -0.001)	-0.21 (-0.35, -0.06)
Latino	-0.2	-0.13 (-0.23, 0.03)	-0.13 (-0.23, -0.02)	-0.2 (-0.33, -0.07)
Black	-0.3	-0.2 (-0.31, -0.09)	-0.2 (-0.31, -0.09)	-0.32 (-0.45, -0.18)
Other	0.25	0.03 (-0.14, 0.18)	0.03 (-0.13, 0.18)	0.05 (-0.15, 0.25)
Molars 2	0.2	0.26 (0.04, 0.47)	0.26 (0.03, 0.48)	0.07 (-0.23, 0.38)
Molars 3	0.1	0.21 (-0.001, 0.43)	0.2 (-0.02, 0.42)	0.01 (-0.27, 0.3)
Molars 4	0.4	0.46 (0.23, 0.71)	0.47 (0.22, 0.7)	0.51 (0.15, 0.87)
Premolars 1	-0.1	0.19 (-0.03, 0.39)	0.18 (-0.02, 0.39)	-0.04 (-0.29, 0.22)
Premolars 2	-0.2	0.15 (-0.06, 0.36)	0.15 (-0.05, 0.35)	-0.09 (-0.36, 0.2)
Premolars 3	-0.1	0.21 (0.01, 0.41)	0.2 (-0.01, 0.41)	-0.01 (-0.28, 0.27)
Premolars 4	-0.1	0.09 (-0.12, 0.3)	0.09 (-0.14, 0.3)	-0.2 (-0.48, 0.07)
Canine 1	-0.5	-0.2 (-0.45, 0.04)	-0.2 (-0.46, 0.05)	-0.31 (-0.6, -0.02)
Canine 2	-0.55	-0.29 (-0.52, 0.06)	-0.3 (-0.55, -0.06)	-0.43 (-0.74, -0.14)
Canine 3	-0.3	-0.19 (-0.46, 0.08)	-0.19 (-0.46, 0.08)	-0.26 (-0.62, 0.06)
Canine 4	-0.4	-0.22 (-0.47, 0.02)	-0.22 (-0.47, 0.04)	-0.33 (-0.65, -0.01)
Incisors 1	-0.5	-0.26 (-0.43, -0.09)	-0.27 (-0.44, -0.1)	-0.41 (-0.64, -0.21)
Incisors 2	-0.4	-0.26 (-0.42, -0.08)	-0.26 (-0.43, -0.09)	-0.4 (-0.6, -0.18)

In Table 11 the Beta-Binomial model reverses the sign of the intercept, but continues to have more parameters with the correct parity compared to the Poisson and Negative-Binomial Models. The 95% intervals on average are also wider, which is expected with the MNAR data set

containing slightly more missing data than the MAR dataset. When controlling for the missing-data mechanism, there is a clear improvement in parameter estimates in the selection model:

Table 12: Coefficient Estimates for Selection Models (MNAR)

	True Value	Poisson	Negative-Binomial	Beta-Binomial
intercept	0.2	0.64 (0.45, 0.82)	0.63 (0.45, 0.81)	0.29 (0.11, 0.46)
Age	0.3	0.16 (0.08, 0.24)	0.17 (0.08, 0.26)	0.28 (0.2, 0.37)
Sex	-0.2	-0.12 (-0.23, -0.01)	-0.12 (-0.23, -0.01)	-0.23 (-0.34, -0.13)
Latino	-0.2	-0.12 (-0.22, -0.03)	-0.13 (-0.23, -0.02)	-0.18, -0.28, -0.08)
Black	-0.3	-0.2 (-0.3, -0.09)	-0.2 (-0.31, -0.09)	-0.31 (-0.43, -0.2)
Other	0.25	0.03 (-0.13, 0.18)	0.03 (-0.13, 0.19)	0.2 (0.06, 0.34)
Molars 2	0.2	0.25 (0.12, 0.47)	0.26 (0.12, 0.49)	0.001 (-0.26, 0.27)
Molars 3	0.1	0.21 (-0.02, 0.43)	0.21 (-0.03, 0.44)	0.07 (-0.19, 0.32)
Molars 4	0.4	0.46 (0.22, 0.69)	0.47 (0.22, 0.72)	0.27 (0.02, 0.51)
Premolars 1	-0.1	0.18 (-0.02, 0.38)	0.19 (-0.03, 0.4)	-0.06 (-0.29, 0.15)
Premolars 2	-0.2	0.16 (-0.05, 0.35)	0.16 (-0.06, 0.36)	-0.25 (-0.45, -0.02)
Premolars 3	-0.1	0.2 (-0.01, 0.41)	0.21 (-0.01, 0.42)	-0.16 (-0.37, 0.05)
Premolars 4	-0.1	0.08 (-0.14, 0.3)	0.09 (-0.14, 0.3)	-0.12 (-0.32, 0.08)
Canine 1	-0.5	-0.2 (-0.45, 0.04)	-0.2 (-0.46, 0.06)	-0.36 (-0.62, -0.12)
Canine 2	-0.55	-0.29 (-0.53, -0.05)	-0.3 (-0.55, -0.05)	-0.53 (-0.78, -0.28)
Canine 3	-0.3	-0.18 (-0.46, 0.08)	-0.19 (-0.47, 0.08)	-0.29 (-0.55, -0.02)
Canine 4	-0.4	-0.21 (-0.46, 0.08)	-0.22 (-0.47, 0.03)	-0.4 (-0.63, -0.17)
Incisors 1	-0.5	-0.27 (-0.44, -0.1)	-0.27 (-0.45, -0.09)	-0.48 (-0.65, -0.3)
Incisors 2	-0.4	-0.26 (-0.43, -0.08)	-0.26 (-0.44, -0.07)	-0.45 (-0.63, -0.26)

The sign of the intercept for the Beta-Binomial model no longer is inverted, and the other parameter estimates are close to the true values in Table 12. It should be noted that the selection and pattern-mixture models include the term that increases the probability of missingness as the value of y increases. Inclusion of this parameter in the models perfectly mirrors the model that simulated the missing data and makes these models oracle models. The Pattern-Mixture model results, Table 13, show that the Beta-Binomial model continues to struggle with the sign of the intercept term:

Table 13: Coefficient Estimates for Pattern-Mixture Models (MNAR)

	True Value	Poisson	Negative-Binomial	Beta-Binomial
intercept	0.2	0.64 (0.73)	0.61 (0.73)	-0.1 (0.71)
Age	0.3	0.16 (0.72)	0.16 (0.71)	0.26 (0.72)
Sex	-0.2	-0.12 (0.72)	-0.11 (0.72)	-0.22 (0.72)
Latino	-0.2	-0.13 (0.71)	-0.11 (0.72)	-0.22 (0.73)
Black	-0.3	-0.2 (0.72)	-0.17 (0.72)	-0.33 (0.71)
Other	0.25	0.05 (0.73)	0.02 (0.72)	0.05 (0.73)
Molars 2	0.2	0.26 (0.72)	0.26 (0.71)	0.06 (0.75)
Molars 3	0.1	0.2 (0.74)	0.2 (0.74)	-0.01 (0.74)
Molars 4	0.4	0.47 (0.72)	0.46 (0.73)	0.47 (0.74)
Premolars 1	-0.1	0.18 (0.71)	0.19 (0.71)	-0.07 (0.74)
Premolars 2	-0.2	0.17 (0.73)	0.15 (0.73)	-0.12 (0.74)
Premolars 3	-0.1	0.2 (0.71)	0.19 (0.73)	-0.01 (0.75)
Premolars 4	-0.1	0.09 (0.72)	0.07 (0.73)	-0.22 (0.73)
Canine 1	-0.5	-0.19 (0.74)	-0.19 (0.73)	-0.34 (0.73)
Canine 2	-0.55	-0.29 (0.73)	-0.31 (0.74)	-0.47 (0.72)
Canine 3	-0.3	-0.17 (0.73)	-0.19 (0.74)	-0.32 (0.75)
Canine 4	-0.4	-0.22 (0.74)	-0.23 (0.76)	-0.33 (0.74)
Incisors 1	-0.5	-0.26 (0.73)	-0.28 (0.49)	-0.41 (0.73)
Incisors 2	-0.4	-0.25 (0.71)	-0.27 (0.72)	-0.4 (0.74)

The standard deviation estimates also increased compared to the pattern-mixture models run on the datasets with MAR and MCAR missing data. Ultimately, the two pattern-mixture models appear to not be nuanced enough to compensate for their increase in parameter estimate variance. This was also confirmed in the posterior predictive checks, as the selection model performed better than the others. More nuanced patterns will be controlled for in the methamphetamine dataset and compared to the two pattern-mixture models.

Additionally, a sensitivity analysis was conducted, and the parameter estimates on the models that did not control for the missing-data mechanism were not significantly changed. However,

the selection and pattern-mixture models were sensitive to changes in prior distributions. Non-informative and high-variance prior distributions resulted in identifiability and convergence issues. Regression parameter inferences did not change substantially for normal priors centered at 0 with standard deviation below 10. The over-dispersion parameters on the Negative-Binomial and Beta-Binomial Models were sensitive to prior distribution variance and required stronger prior distributions to identify the models. A Gamma(1,1) prior was not restrictive enough to achieve an identified model, but Gamma(1, 5) identified the model, which imposed strict smoothing on the over-dispersion parameter. However, use of 'rjags' and the nbetabinom function, which is a reparameterization of the Beta-Binomial model, allowed the Beta-Binomial to converge with the uniform(0, 10) prior on the over-dispersion parameter. A similar sensitivity analysis will be performed and discussed on the methamphetamine data, and highlighting the importance of three-pattern Pattern-Mixture Models to identify the parameters.

4.4 Data Analysis and Discussion

4.4.1 Methamphetamine Data: MAR

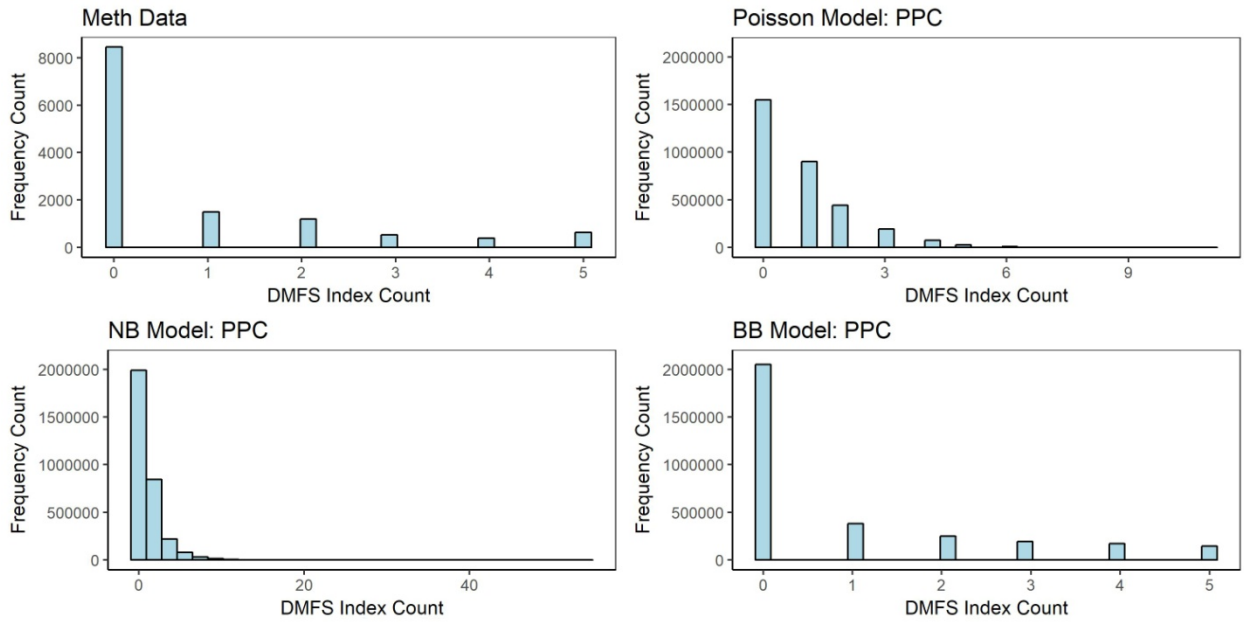
Three models were run on the methamphetamine using data set that did not control for the missing-data mechanism. The likelihood distributions of these three models included Poisson, Negative-Binomial, and Beta-Binomial distributions and the covariates sex, race, age, and the aforementioned tooth location variables described in section 4.2.2. The parameter estimates are in Table 14 below:

Table 14: Coefficient Estimates for Base Models (Real Data)

	Poisson	Negative-Binomial	Beta-Binomial
intercept	0.45 (0.37, 0.53)	0.41 (0.26, 0.55)	-0.64 (-0.77, -0.48)
Age	0.41 (0.36, 0.46)	0.44 (0.37, 0.51)	0.49 (0.41, 0.57)
Sex	-0.3 (-0.34, -0.25)	-0.3 (-0.37, -0.22)	-0.44 (-0.52, -0.35)
Latino	-0.03 (-0.05, 0.03)	-0.02 (-0.11, 0.07)	0.001 (-0.09, 0.09)
Black	-0.05 (-0.1, 0.01)	-0.06 (-0.14, 0.03)	-0.02 (-0.11, 0.07)
Other	0.21 (0.13, 0.3)	0.34 (0.2, 0.47)	0.25 (0.11, 0.37)
Molars 2	-0.06 (-0.14, 0.02)	-0.04 (-0.19, 0.11)	-0.05 (-0.2, 0.1)
Molars 3	0.11 (0.04, 0.19)	0.15 (-0.001, 0.3)	0.2 (0.05, 0.35)
Molars 4	0.01 (-0.07, 0.09)	0.03 (-0.11, 0.17)	0.08 (-0.06, 0.23)
Premolars 1	-0.44 (-0.52, -0.35)	-0.41 (-0.56, -0.27)	-0.68 (-0.84, -0.53)
Premolars 2	-0.41 (-0.5, -0.32)	-0.4 (-0.54, -0.25)	-0.67 (-0.82, -0.53)
Premolars 3	-0.92 (-1.01, -0.82)	-0.94 (-1.09, -0.79)	-1.14 (-1.3, -0.98)
Premolars 4	-0.85 (-0.95, -0.76)	-0.86 (-1, -0.7)	-1.12 (-1.27, -0.97)
Canine 1	-1.5 (-1.67, -1.31)	-1.52 (-1.73, -1.3)	-1.9 (-2.1, -1.72)
Canine 2	-1.5 (-1.69, -1.34)	-1.52 (-1.7, -1.3)	-2.03 (-2.21, -1.83)
Canine 3	-2.23 (-2.46, -1.96)	-2.26 (-2.5, -2)	-2.35 (-2.53, -2.17)
Canine 4	-2.08 (-2.29, -1.81)	-2.08 (-2.3, -1.85)	-2.28 (-2.47, -2.09)
Incisors 1	-0.86 (-0.94, -0.78)	-0.87 (-1, -0.75)	-1.16 (-1.3, -1.02)
Incisors 2	-2.4 (-2.6, -2.4)	-2.5 (-2.7, -2.37)	-2.53 (-2.67, -2.36)

There were no problems with convergence of parameters or identifiability with the base prior distributions discussed in the simulated data section. As seen in Table 14, parameter estimates across the three models were all similar and the intercept being the only estimate where there is a difference in parity of sign. The dispersion parameter of the Negative-Binomial, r = 0.66 (0.62, 0.71), and the θ parameter in the Beta-Binomial model, θ = 1.2 (1.14, 1.26), indicate that over-dispersion is present in the data relative to Poisson variation. Figure 11 displays the distribution of DMFS index frequency from 200 generated data sets using the trained models where the parameters were selected from every 25th iteration of the Gibbs sampling process (autocorrelation function did not report significant correlation between these iterations):

Figure 29: DMFS Index Distributions for Base Models



In Figure 29 it is clear that the Beta-Binomial model most closely resembles the actual data. To better assess the model fit, each of the 200 simulated data sets will have the proportion of their observations at 0, (1, 2, or 3), (4 or 5), and 5+ calculated and plotted in Figure 30:

Figure 30: Proportion of Posterior Predictive Checks in Each Category (Base Models)

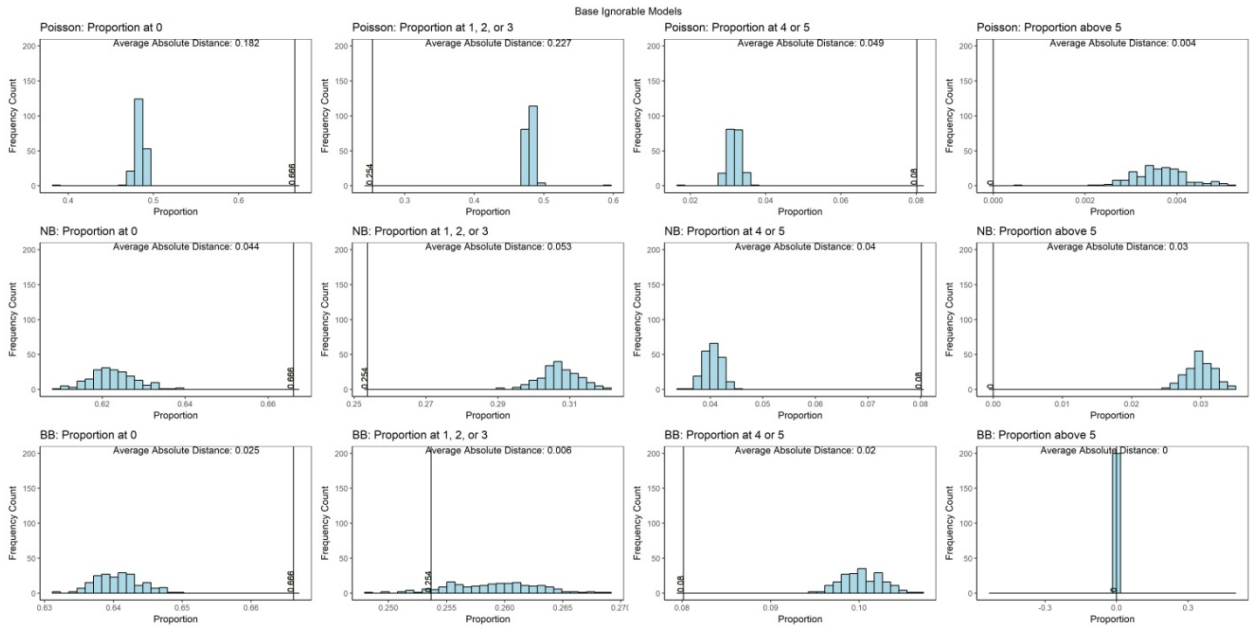
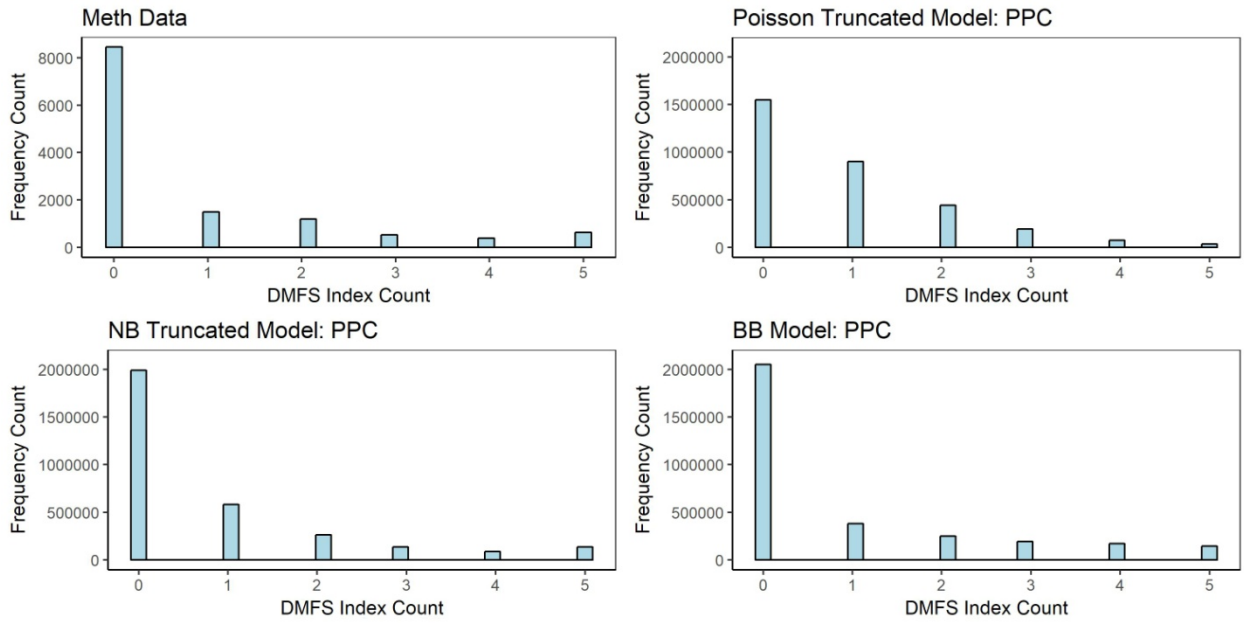


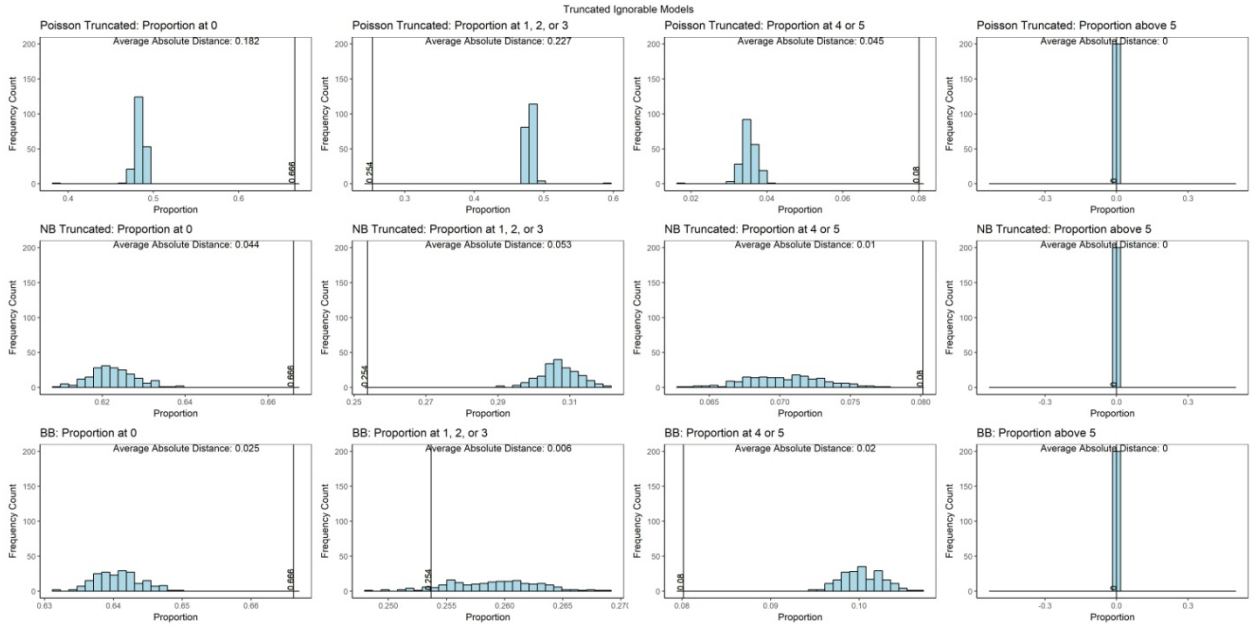
Figure 30 shows that the Beta-Binomial model and Negative-Binomial have closer proportions to the data than the Poisson model. The Beta-Binomial model outperforms the Negative-Binomial in all categories and especially in the (1, 2, or 3) category. As expressed before, the Beta-Binomial has a natural truncation at the tooth's maximal count and this can give the model an advantage over the Negative-Binomial and Poisson models as seen in the 5+ category of Figure 30. Figure 31 displays the posterior predictive checks for the models again, but with the Poisson and Negative-Binomial data truncated at the tooth's maximal DMFS index:

Figure 31: Distributions for Base Models (Truncated)



The Negative-Binomial model, unlike the Beta-Binomial model, captures that there are more teeth with a count of 5 than a count of 4. However, the truncation still does not remedy the excess density it places at 1. Figure 32 displays the distribution of the proportions for each model in the four categories:

Figure 32: Proportion of Posterior Predictive Checks in Each Category (Truncated Base Models)



The Negative-Binomial demonstrates its improved (4 or 5) category where the observed data's proportion was close to the generated data's proportion distribution. Overall, the truncated Negative-Binomial model outperforms the Beta-Binomial model in the (4 or 5) category by 1% in mean absolute error, but underperforms in the 0 and (1, 2, or 3) category by wider margins. The increased performance of the Negative-Binomial when truncated is still insufficient to choose it over the Beta-Binomial model. To be thorough, the missing teeth in the data were identified and their analogues teeth in the generated data from the models were removed and the resulting distributions are displayed in Figure 33:

Figure 33: Distributions for Base Models (Missing Removed)

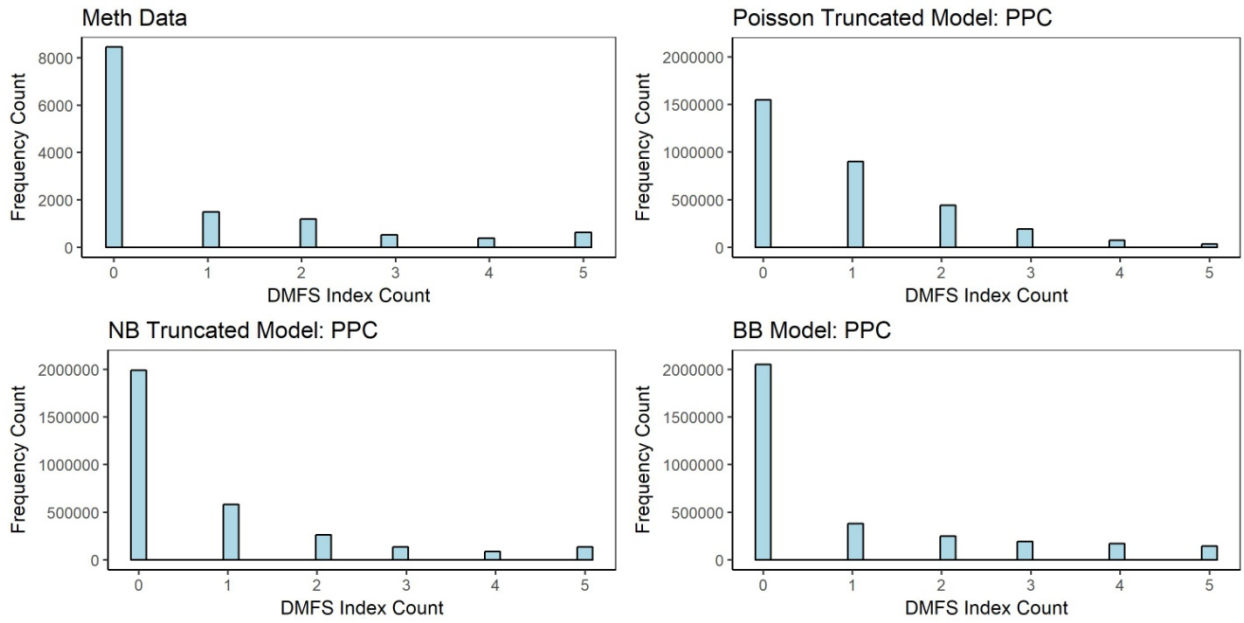
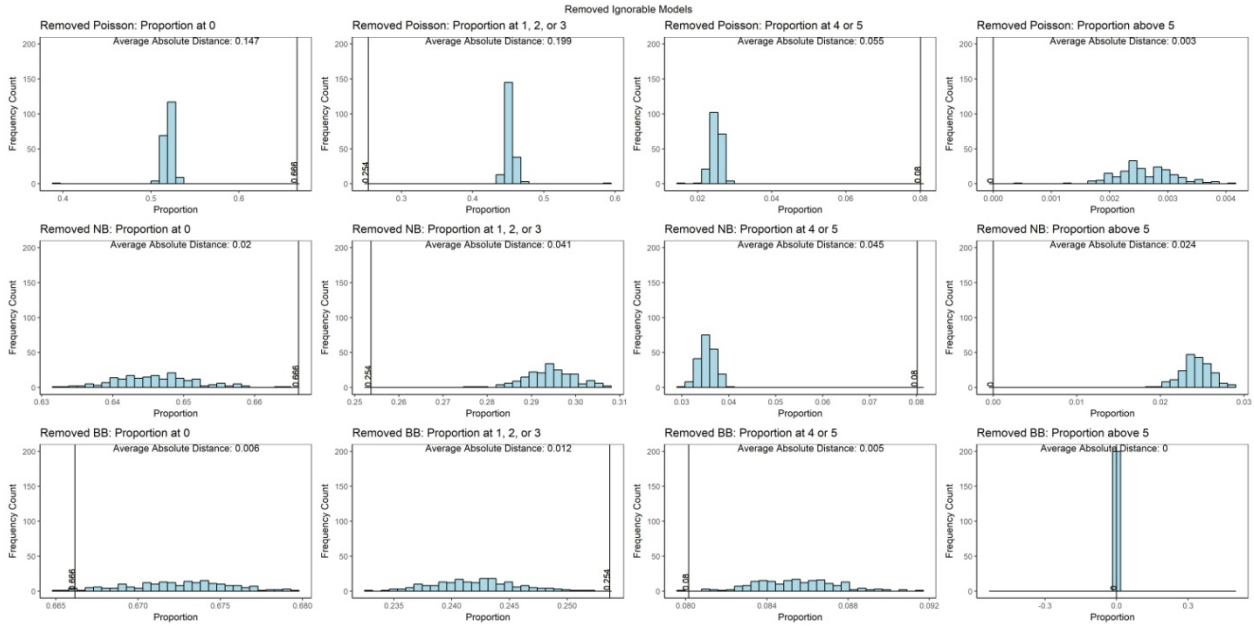


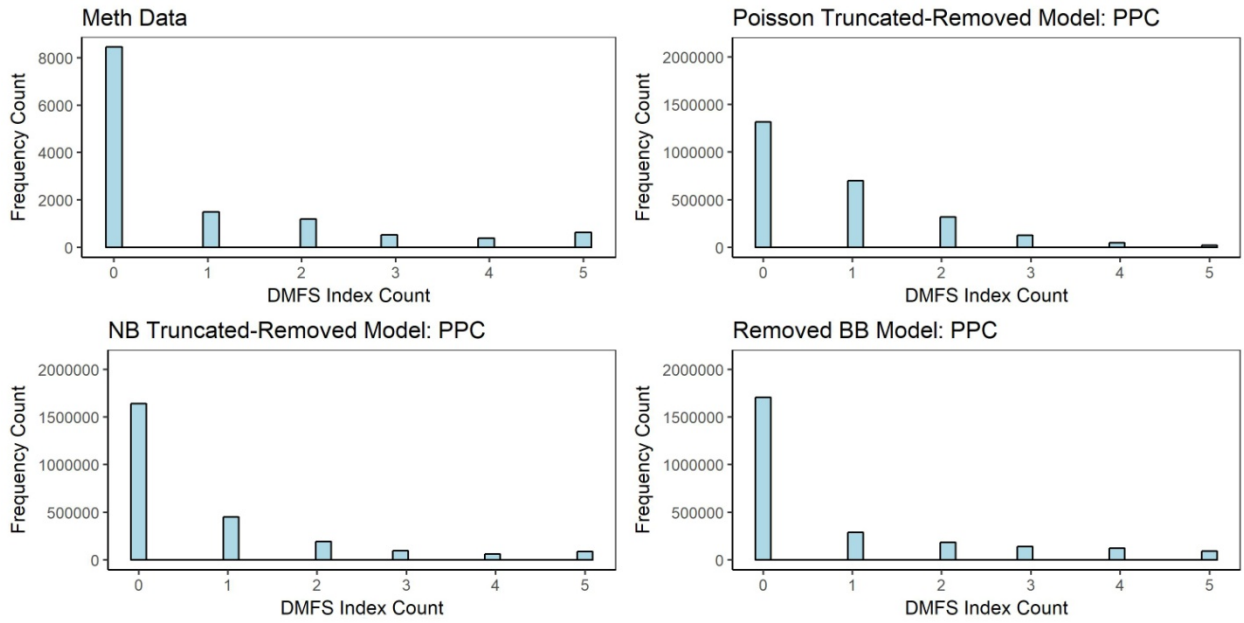
Figure 33 displays no clear differences from previous posterior predictive checks. Figure 34 displays the distribution of proportions of the four categories previously defined:

Figure 34: Proportion of Posterior Predictive Checks in Each Category (Missing Removed Base Models)



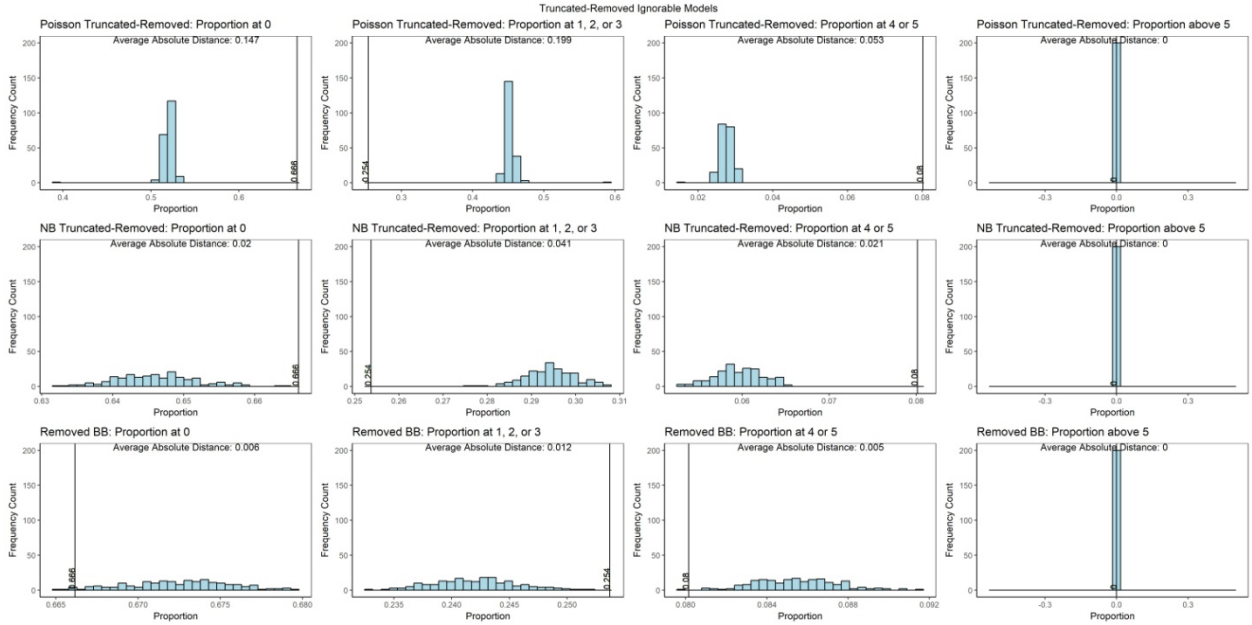
All models improve slightly in all categories with the Beta-Binomial model having 1.2% as its largest mean absolute error and the true data proportion more or less contained or touching the distribution of the posterior predictive check distribution. However, the difference between the Negative-Binomial and Beta-Binomial models remains roughly the same magnitude. The Poisson model improves in all categories with the removal of missing teeth. The Poisson and Negative-Binomial models, with the missing teeth removed, will now be truncated to make them more comparable to the Beta-Binomial model in Figure 35:

Figure 35: Distributions for Base Models (Truncated and Missing Removed)



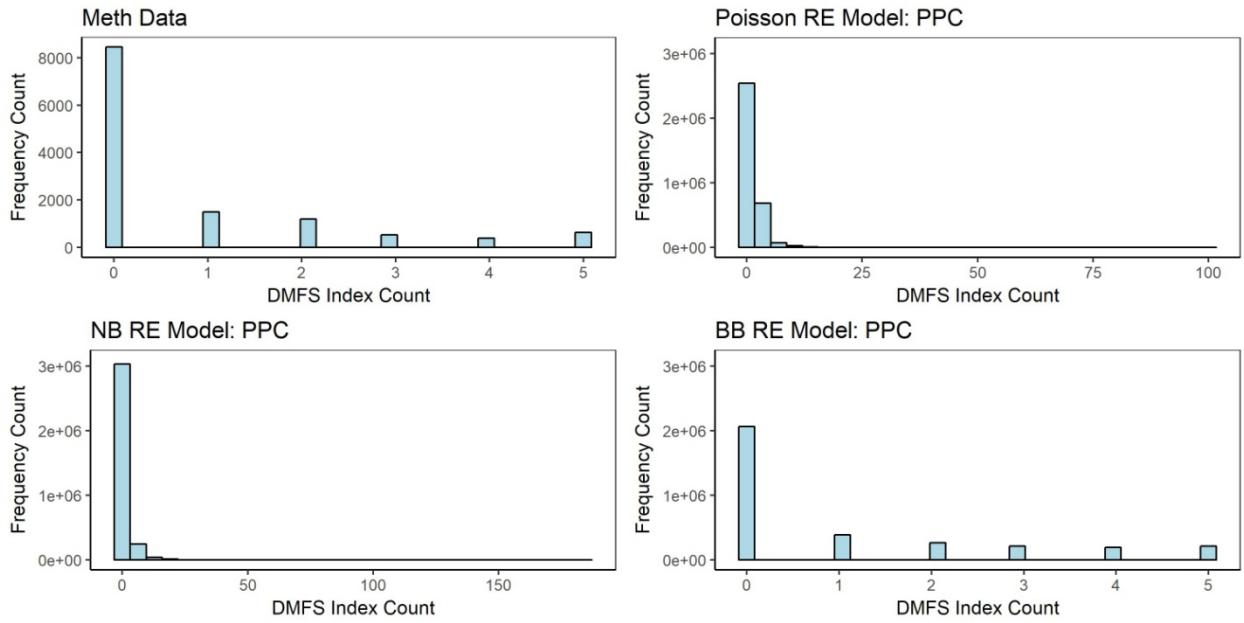
It is evident that the removal of missing teeth has decreased the amount of density that the Negative-Binomial model places at 5. Figure 36 shows the distributions of the proportions in the four categories for these data:

Figure 36: Proportion of Posterior Predictive Checks in Each Category (Truncated and Missing Removed Base Models)



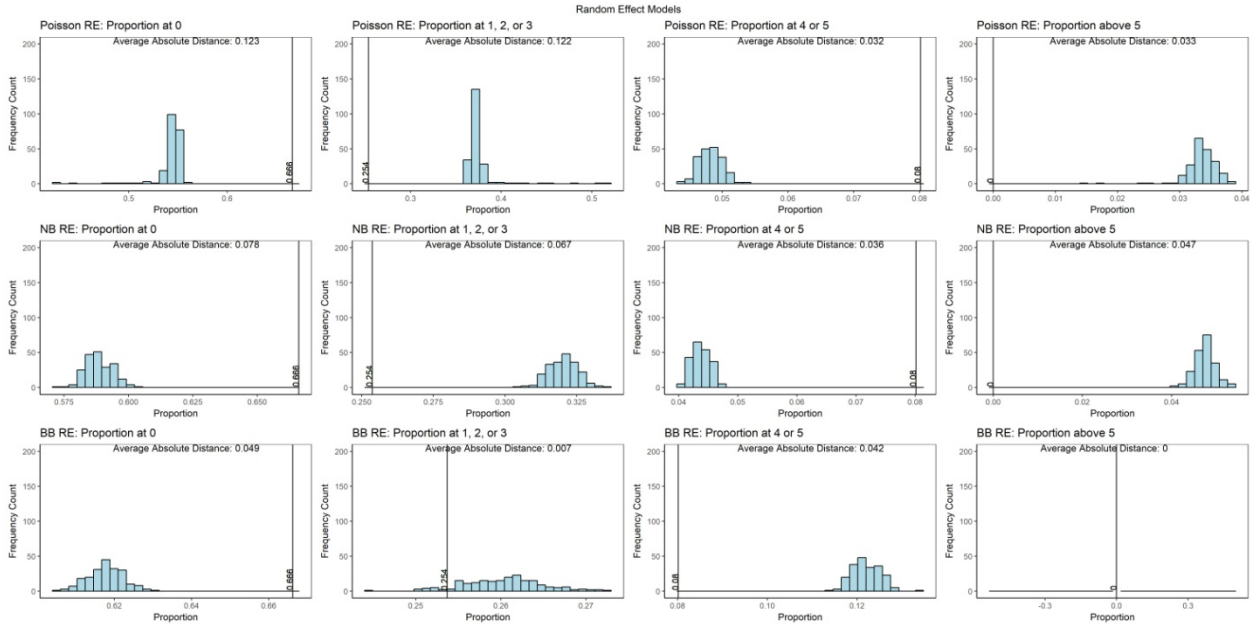
The average absolute difference measure for the Beta-Binomial and Negative-Binomial models maintain their differences with the missing data removed and truncation applied. Based on previous work where missing teeth were assigned maximum DMFS index, inclusions of a subject level random effect was shown to aid model fit and performance. For continuity, a subject level random effect was added to the base models and run on the data with posterior predictive checks in Figure 37:

Figure 37: Distributions for Random Effect Models



From Figure 37, it is already clear based on the scale of the x-axis that several higher valued DMFS index values were generated in the Poisson and Negative-Binomial models. Figure 38 uses these generated data to make the four categories of proportions:

Figure 38: Proportion of Posterior Predictive Checks in Each Category (Random Effect Models)



The mean absolute error increases in the 0 and (4 or 5) categories for the Beta-Binomial model, which shows that the random effect model may perform better for the situation where missing teeth have maximal tooth scores, but not in a data imputation setting. The remainder of the models, Selection and Pattern-Mixture models, report similar parameter estimates for the additional posterior predictive checks displayed (truncated, removed NA, and truncated-removed NA) and therefore, for conservation of space, their posterior predictive checks and proportion distribution figures are omitted because they are qualitatively similar to those already presented.

4.4.2 Methamphetamine Data: Selection Model

When the selection model was run on the methamphetamine data set, errors occurred relating to the Beta distribution where it was placing infinite density on a certain parameter. This occurs when both α and β both approach 0 simultaneously in the sampling space. A remedy for this

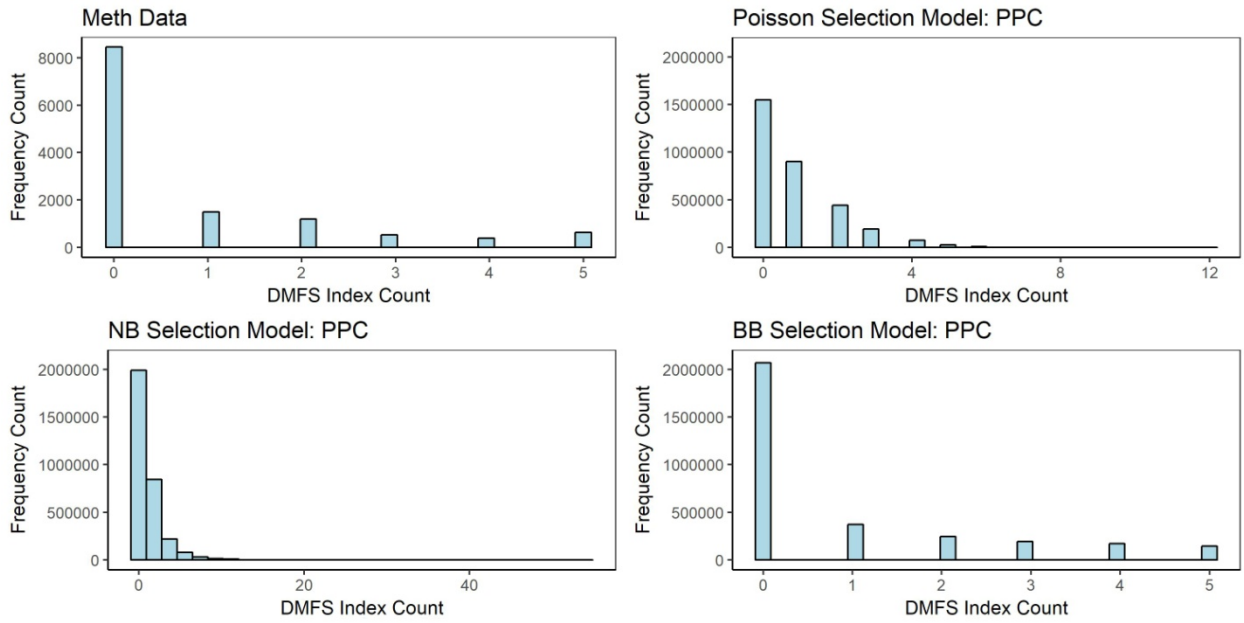
error is to add a constant (0.1 or some other small value) to the parameters, effectively placing a floor in for lowest achievable value. This corrective measure allows for the model to avoid this error, but does change parameter estimates slightly.

Table 15: Coefficient Estimates for Selection Models (Real Data)

	Poisson	Negative-Binomial	Beta-Binomial
intercept	0.45 (0.36, 0.54)	0.4 (0.24, 0.55)	-0.76 (-0.98, -0.56)
Age	0.41 (0.36, 0.46)	0.44 (0.37, 0.52)	0.81 (0.68, 0.93)
Sex	-0.3 (-0.35, -0.25)	-0.29 (-0.37, -0.21)	-0.7 (-0.83, -0.57)
Latino	-0.03 (-0.08, 0.03)	-0.02 (-0.11, 0.08)	0.03 (-0.12, 0.2)
Black	-0.05 (-0.1, 0.01)	-0.05 (-0.15, 0.04)	-0.02 (-0.16, 0.13)
Other	0.21 (0.14, 0.3)	0.34 (0.21, 0.48)	0.21 (-0.03, 0.44)
Molars 2	-0.06 (-0.14, 0.02)	-0.04 (-0.18, 0.11)	-0.12 (-0.31, 0.06)
Molars 3	0.11 (0.03, 0.19)	0.15 (-0.004, 0.3)	0.2 (0.01, 0.38)
Molars 4	0.01 (-0.06, 0.09)	0.03 (-0.12, 0.18)	0.05 (-0.13, 0.24)
Premolars 1	-0.43 (-0.52, -0.34)	-0.41 (-0.56, -0.26)	-1 (-1.22, -0.79)
Premolars 2	-0.4 (-0.48, -0.31)	-0.39 (-0.54, -0.25)	-0.98 (-1.94, -0.77)
Premolars 3	-0.92 (-1.01, -0.83)	-0.93 (-1.08, -0.78)	-1.79 (-2.08, -1.51)
Premolars 4	-0.86 (-0.95, -0.76)	-0.86 (-1, -0.71)	-1.71 (-1.98, -1.46)
Canine 1	-1.49 (-1.65, -1.3)	-1.51 (-1.72, -1.3)	-11.3 (-23.8, -4.69)
Canine 2	-1.51 (-1.67, -1.31)	-1.52 (-1.73, -1.3)	-12.06 (-25.06, -5.28)
Canine 3	-2.22 (-2.45, -1.96)	-2.26 (-2.52, -2.01)	-12.71 (-25.22, -6.22)
Canine 4	-2.08 (-2.29, -1.83)	-2.08 (-2.32, -1.84)	-12.5 (-24.7, -6.01)
Incisors 1	-0.87 (-0.95, -0.78)	-0.87 (-1, -0.74)	-1.86 (-2.08, -1.63)
Incisors 2	-2.44 (-2.58, -2.12)	-2.53 (-2.71, -2.36)	-13.79 (-25.4, -7.63)

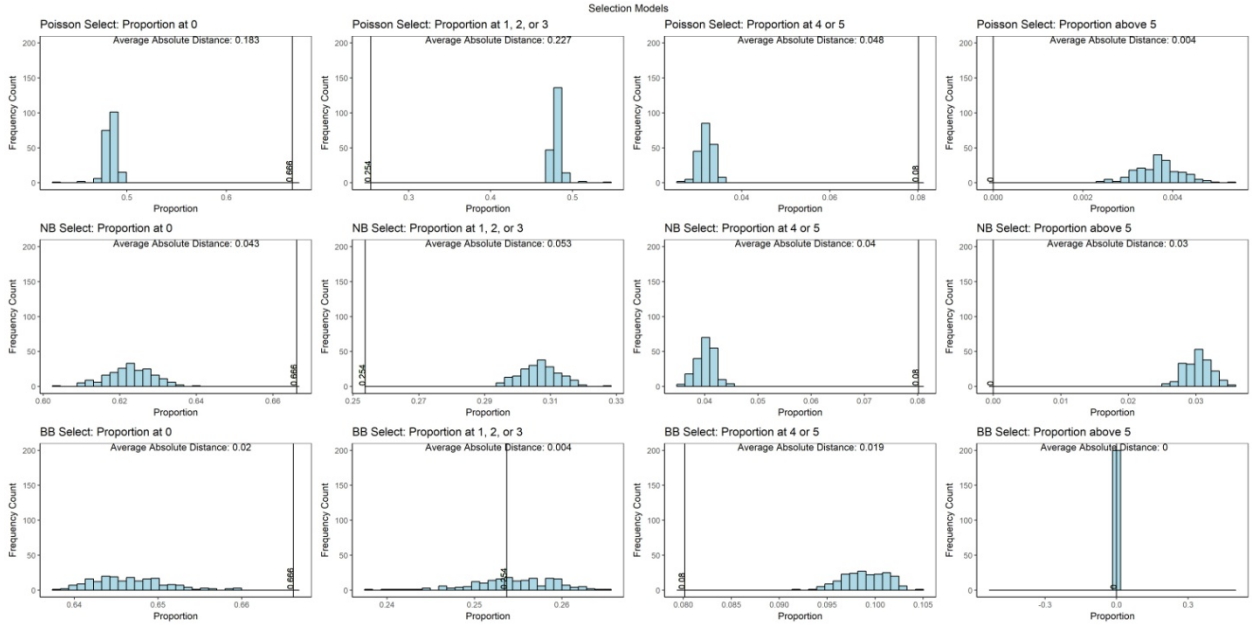
The parameter estimates in Table 15 for the incisors and canines in the Beta-Binomial model exhibit small, negative values that are on another level of magnitude when compared to the other models. These parameters may require more informative priors to compensate for the correlation in the features and lack of information in the data. Parameter estimates between the selection model and the models not controlling for the missing-data mechanism do not differ considerably and the posterior distributions for the DMFS index are similar to the ignorable models Figure 39:

Figure 39: Distributions for Selection Models



The distribution of proportions in the four categories also does not differ considerably from the previous base models where the Negative-Binomial and Beta-Binomial models outperforming the Poisson model. The Negative-Binomial and Beta-Binomial outperform the Poisson model, but the Beta-Binomial continues to be the best fit. The distribution of proportions for the four categories can be seen in Figure 40:

Figure 40: Proportion of Posterior Predictive Checks in Each Category (Selection Models)



Compared to the ignorable models, the selection model performs similarly. However, the selection model requires significantly more computation.

4.4.3 Methamphetamine Data: Pattern-Mixture Model

The pattern-mixture models were initially run so that the only two patterns of missing data were missing and not missing (indicator variable):

Table 16: Coefficient Estimates for Pattern-Mixture Models (Real Data)

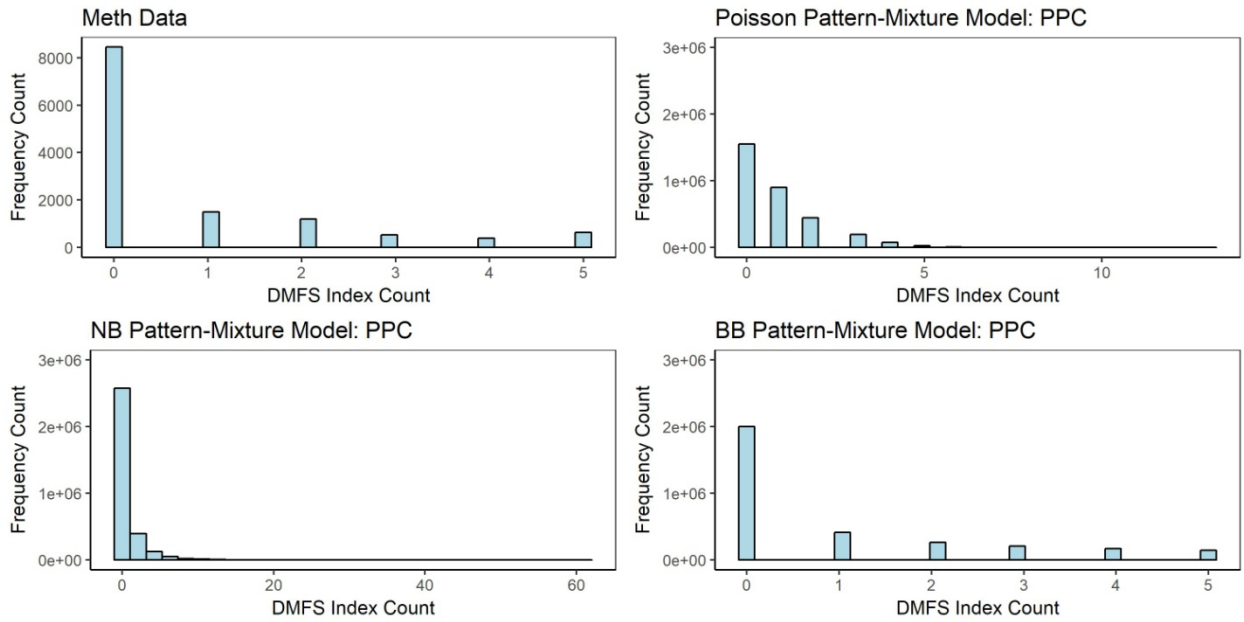
	Poisson	Negative-Binomial	Beta-Binomial
intercept	0.45 (0.21)	0.41 (0.22)	-0.84 (0.23)
Age	0.41 (0.21)	0.44 (0.21)	0.79 (0.22)
Sex	-0.3 (0.2)	-0.29 (0.21)	-0.7 (0.21)
Latino	-0.02 (0.21)	-0.02 (0.21)	0 (0.22)
Black	-0.04 (0.21)	-0.06 (0.21)	-0.04 (0.22)
Other	0.22 (0.21)	0.34 (0.22)	0.23 (0.23)
Molars 2	-0.06 (0.21)	-0.04 (0.22)	-0.01 (0.23)
Molars 3	0.11 (0.21)	0.15 (0.22)	0.31 (0.22)
Molars 4	0.01 (0.21)	0.03 (0.22)	0.17 (0.22)
Premolars 1	-0.43 (0.21)	-0.41 (0.22)	-0.87 (0.23)
Premolars 2	-0.4 (0.21)	-0.39 (0.22)	-0.83 (0.24)
Premolars 3	-0.91 (0.22)	-0.94 (0.22)	-1.63 (0.24)
Premolars 4	-0.85 (0.21)	-0.87 (0.22)	-1.58 (0.24)
Canine 1	-1.49 (0.23)	-1.52 (0.23)	-3.65 (0.44)
Canine 2	-1.53 (0.24)	-1.52 (0.23)	-3.97 (0.49)
Canine 3	-2.23 (0.27)	-2.26 (0.24)	-4.72 (0.47)
Canine 4	-2.08 (0.26)	-2.08 (0.23)	-4.53 (0.48)
Incisors 1	-0.86 (0.21)	-0.87 (0.21)	-1.71 (0.25)
Incisors 2	-2.45 (0.27)	-2.53 (0.22)	-6 (0.44)

The parameter estimates again do not differ from the models that do not control for the missing-data mechanism except for the Beta-Binomial model, which exhibited larger negative coefficients, but were not as extreme as the Selection model. The Pattern-Mixture models were run again with three patterns, which were detailed in a previous section.

These parameter estimates displayed below in Figure 43, 44, and 45 differ slightly from what the other models provided, but were more stable in their estimation compared to the previous Pattern-Mixture model. Additionally, as displayed, the three patterns differ in coefficient values and this kind of model can give insight into the associations and correlations between

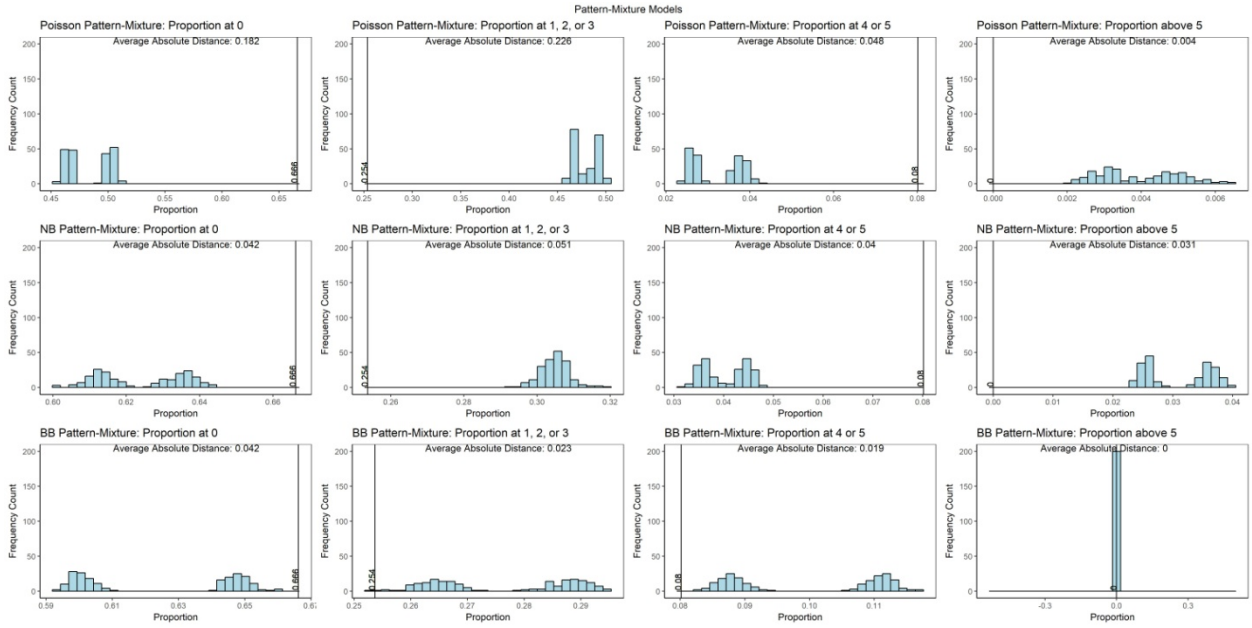
coefficients and dental decay for each pattern type allowing for more nuanced conclusions to be drawn from the model. Posterior distributions for DMFS index exhibited similar characteristics to the real data Figure 41:

Figure 41: Distributions for Pattern-Mixture Models



The distribution of the proportions for the models demonstrated the mixture of the differing patterns Figure 42:

Figure 42: Proportion of Posterior Predictive Checks in Each Category (Pattern-Mixture Models)



Relative to all other models, the three pattern Pattern-Mixture models display differing results. The distributions of the proportions yield a multimodal distribution. Parameters for the mixed models using the delta-method formula presented early were similar to other models. Below in Figure 43 are the posterior distributions for the three patterns in the Poisson pattern-mixture model:

Figure 43: Poisson Pattern-Mixture Model Coefficient Posterior Distributions

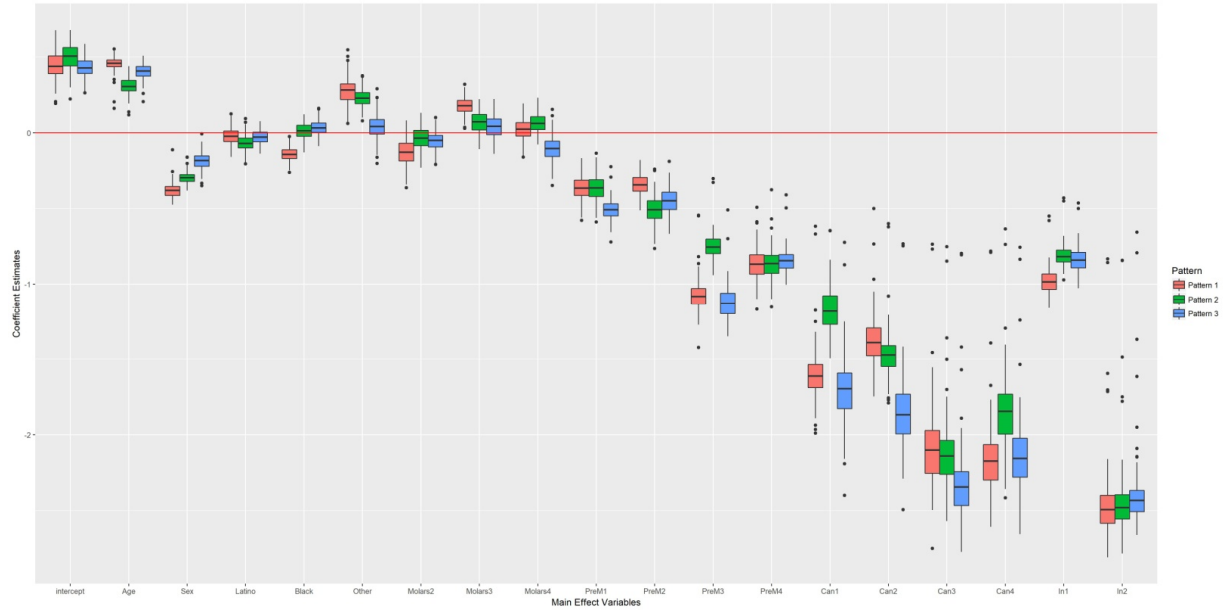
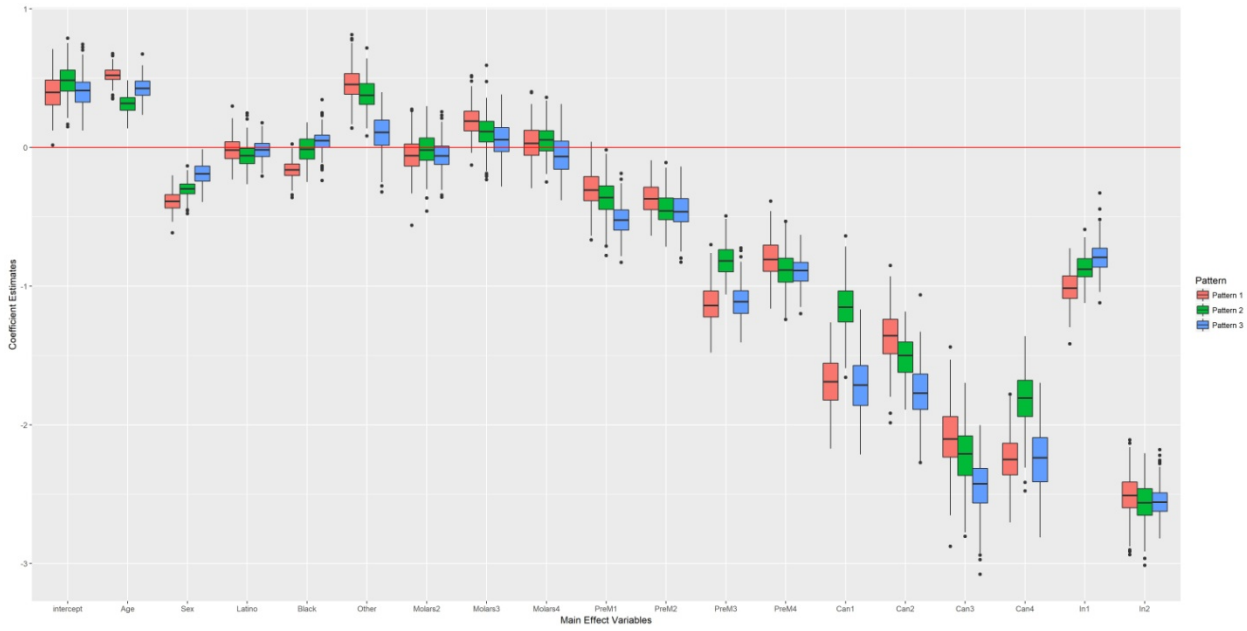


Figure 44 below represents the same, but for the Negative-Binomial model:

Figure 44: Negative-Binomial Pattern-Mixture Model Coefficient Posterior Distributions



And lastly the Beta-Binomial pattern-mixture model in Figure 45:

Figure 45: Beta-Binomial Pattern-Mixture Model Coefficient Posterior Distributions

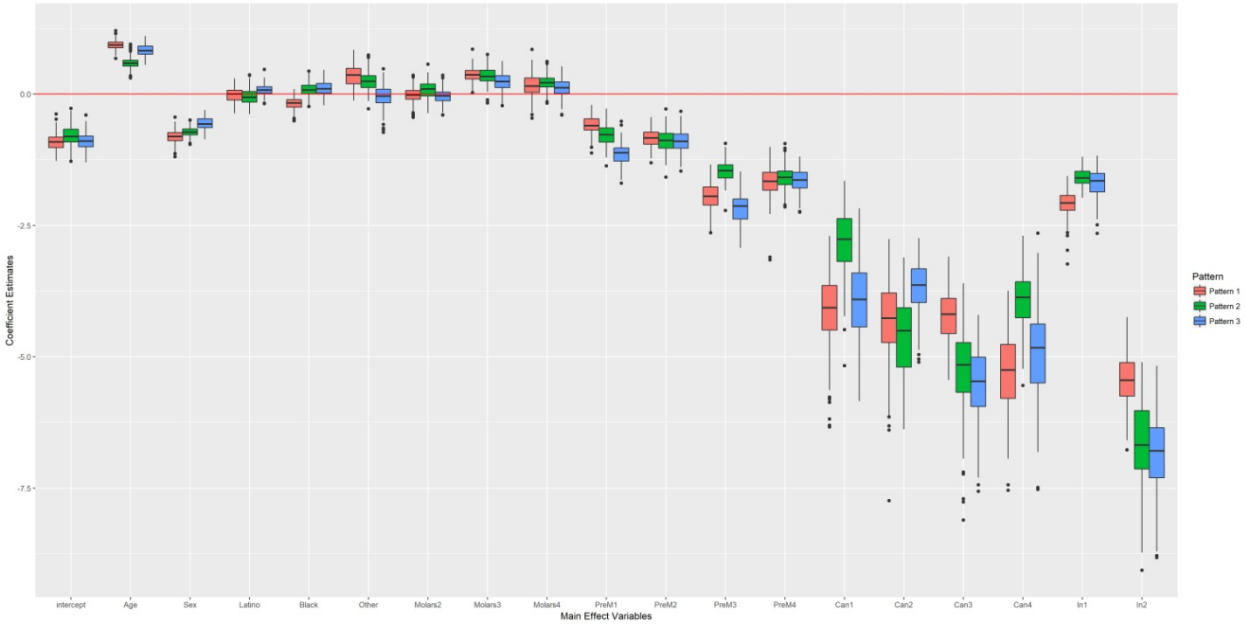


Figure 43, 44, and 45 show that the regression parameters between models all have similar posterior distributions with Pattern 2 differentiating itself in age, premolar 3 (lower left premolar), canine 1 (upper right canine) and canine 4 (lower right canine). The other two pattern groups display differentiation in race, premolar 1, canine 2, and incisors 1 and 2. Additionally, the range of the posterior distributions for the coefficient estimates increases significantly for the canines and incisors 2. The difference between the exchangeable pattern groups is likely the reason for the bimodality in Figure 24 and provides a more nuanced understanding of the patterns of decay and missing data in the mouth. It is this level of granular detail that is unique to pattern-mixture models and allows for the researcher to better understand the data being analyzed.

4.4.4 Discussion and Conclusion

Overall, there was very little difference in parameter estimation and mean absolute error of distribution of proportions between the ignorable models, selection models, and two patterns Pattern-Mixture model. However, within this set of models the Negative-Binomial and Beta-Binomial models outperformed the Poisson Model in absolute mean error. In particular, the truncated Negative-Binomial had the best performance of the Negative-Binomial models. Overall, the Beta-Binomial model outperformed the Negative-Binomial and Poisson models in every category. The Selection-Model evidenced a similar performance to the ignorable and the question of whether to utilize it depends on computational complexity concerns. The three pattern Pattern-Mixture model exhibited improved mean absolute error with the Negative-Binomial and Beta-Binomial outperforming the Poisson model. This Pattern-Mixture model evidenced a mutli-model distribution in the proportions in the four categories, which is what might be expected for a mixture model. Even though the model had a larger absolute mean difference, the Pattern-Mixture model allows for parameters estimates in each decay pattern. This level of granularity can be advantageous from a modeling perspective to see if parameter estimates make sense given their pattern and from a scientific basis because we are now able to understand the influence of variables on decay status of the teeth for different levels of severity. Model fit could be improved given different patterns. The use of pattern-mixture models bring up another idea: that missing data should not be viewed simply as something to be imputed, but rather a data category in itself (Little and Rubin, 2002). The missing data provide information by differentiating teeth from certain generative patterns of decay/missingness based on the patterns we assign. Allowing for separate models to be fit on these data, outside of imputation, give nuanced parameter estimates that can assist with scientific understanding and inferences drawn.

From a modeling perspective, Beta-Binomial models consistently outperformed the Negative-Binomial and Poisson models. From the simulated data and the methamphetamine data, it is clear that the truncation of the model at the maximal value of the tooth improves fit and results in posterior predictive checks that resemble the data more accurately. The Beta-Binomial model has a natural truncation due to the Binomial portion of the model where the number of trials must be specified, which is the number of surfaces on the tooth. The Negative-Binomial model requires truncation after running the model, but requires little post processing effort. So, for truncated count data, these two models should be considered with knowledge that the Beta-Binomial is likely to have a better fit.

Modeling truncated count data where a majority of the density is placed at one extreme, but with a 'U' shape where the other extreme has some density presents several challenges. The flexibility of the likelihood selected for modeling must be able to capture this shape for accurate inference. The Beta-Binomial model represents a natural solution that contains truncation and the flexibility to capture this shape. The only shortcoming of the Beta-Binomial model was its inability to model the drop in observations at the DMFS index of 3 and 4. In the observed data, the DMFS index distribution represents a 'U' shape except for where there is a sharp drop in the 3's. It appears that the Beta-Binomial model is able to form the 'U' shape, but does not model the sharp drop accurately. The inaccuracy in modeling the drop likely leads to the 2% error in the (4 or 5) category, but this level of error is tolerable in most situations and does not detract from the overall inference and conclusions drawn. An alternative model is the Negative-Binomial, which requires truncation in post processing, but does not perform as well as the Beta-Binomial on these DMFS index dental data. Both of these models should be considered when analyzing

dental data as they outperform and generate posterior predictive checks more similar to the true data than commonly used models for DMFS index like Linear and Poisson regression. Additionally, when missing data are present, the Pattern-Mixture model should be considered as a possible solution. It allows for nuanced imputation and model parameter estimation that account for decay/missing-data patterns in the mouth. Controlling for specified patterns is of both modeling and scientific interest as it can improve model fit and allow for more nuanced scientific conclusions to be drawn. Therefore, when analyzing dental data with missing teeth, Pattern-Mixture models using a Beta-Binomial or Truncated Negative-Binomial likelihood should be considered as possible modeling options.

5 Missing Data Methods for Ecological Momentary Assessments in Mobile Health Data

Information concerning oral health behavioral patterns is vital for informing clinical decisions to prevent and understand dental decay (Lie et. al., 2014). In a dental setting, these data are typically collected in person through exchanges at a dental office or clinic although it is also possible to collect data on oral health behaviors through a survey. The patient's answers are based on recollection and subject to recall bias, which can influence model parameter estimates and the conclusions ultimately drawn. Mobile health technology offers a possible way to mitigate recall bias through utilizing Ecological Momentary Assessments (EMAs) in real time using small surveys administered on the patient's smart phone. By collecting the data closer in time to the behavior in question, recall bias may be reduced (Shiffman, Stone, and Hufford, 2008). However, this technology presents its own challenges, as non-response and missing data can be expected. In the context of a pilot investigation of EMAs of measures related to oral health, attempts were made to minimize sample size while controlling for design factors such as length and frequency of EMAs by using an Orthogonal Array study design (Taguchi et. al., 2004). The efficiency of the orthogonal array study design limits the degrees of freedom available at the analysis stage and makes modeling of high-order interaction terms difficult, but by intentionally balancing the number of times each level of each treatment appears with the levels of other treatment factors, orthogonal array designs are not motivated by interest in estimation of two-

way interactions. Implementations of variations of the one degree of freedom test for non-additivity (Tukey, 1949) and a hierarchical Bayesian model with smoothing of interaction terms (Rubin, Schafer, and Schenker, 1988) were used to conserve degrees of freedom while controlling for interaction terms. Even with the efficiency of the orthogonal array, missing data can still occur. It is possible that the missing data will exhibit distinct patterns in the survey data. Missing data methods were applied assuming the missing-data mechanism to be ignorable. However, with the distinct, observed patterns of missingness possibly being informative, pattern-mixture models were also implemented to model the data and the missing-data mechanism. After all missing data models were fit, a sensitivity analyses was carried out to assess robustness of study findings. Ultimately, the one degree of freedom test for non-additivity was able to model lower order interaction terms, and the pattern-mixture models, specifically adapted to the unique patterns of missing data observed in the EMAs, provided the most accurate inferences when missing data are non-ignorable and should be considered in any future EMA survey analysis.

5.1 Introduction

Collection of oral health behavior (OHB) data are vital to understand dental decay and assist patients in controlling it in a clinical setting (Lie et. al., 2014). Previous work by our group (Shetty et al., 2016) suggested that the dental decay observed in methamphetamine users was associated with oral health behaviors, such as frequency of brushing, dental visits, and diet, opposed to frequency and method of methamphetamine use (smoking, injecting, snorting, and other). This researcg suggests that dental decay is strongly associated with patterns of oral health behavior sustained over long periods of time and illustrates the importance of these data. However, when such data are collected, most often in a clinical setting, the patient is required to

recall very specific behaviors, which ultimately is dependent on the accuracy of the patient's memory and is subject to recall bias (Hassan, 2005). Recall bias can have a meaningful influence on the data collected and can consequently influence parameter estimates in a statistical model, carrying the potential to yield inaccurate conclusions.

Mobile health technology, which has demonstrated clinical effectiveness in addiction studies for influencing behaviors through interventions administered on the patient's phone, offers a possible solution to recall bias by delivering surveys in the form of EMAs in real time on a patient's smart phone (Gustafson et al., 2014; Alessi and Petry, 2013; Quanbeck et al., 2014). By ascertaining answers to questions of interest closer to the actual behavior, the accuracy of the answer can be expected to be less influenced by recall bias. With data more representative of true oral health behaviors, more accurate parameter estimates and conclusions can be drawn from the statistical model (Shiffman, Stone, and Hufford, 2008). In some contexts, it is realistic to administer EMAs multiple times on a weekly basis.

However, this new method for collecting data creates new obstacles that must be addressed. Initially, the frequency and length of the EMAs must be considered to facilitate compliance and minimize the amount of missing data. After the EMA data are collected, even with survey frequency and length optimized, missing data are likely to be present. Due to the nature of the surveys, distinct patterns of missing data are likely to arise and may be informative in data imputation and modeling. Therefore, the rest of the chapter will be organized into the following sections: 1) Study designs that optimize efficiency, Latin Squares and Orthogonal Arrays, will be discussed and applied to the EMA framework, 2) Missing data and anticipated EMA missing-

data patterns unique to the Orthogonal Array study design will be discussed 4) Missing data methods for modeling these data presented, and strategies for modeling interaction terms in high-dimensional settings will be proposed, 5) Simulation data from several possible underlying models and missing-data mechanisms will be utilized to assess accuracy of the proposed models for EMA data and a sensitivity analysis conducted to determine their robustness 6) Results of the model fit and sensitivity analysis will be discussed and final conclusions drawn.

5.1.1 Study Design

When considering the experimental design for EMA data, it is important to examine possible biases and sources of variability present as well as trying to be as efficient as possible to minimize required sample size. There are a host of possible biases (systematic, selection, and accidental bias, for example) that must be accounted for in the design of the study. R.A. Fisher proposed the use of randomized clinical trials to control for such biases through balancing the distributions of potential confounding variables, across treatment groups (Fisher 1926 and 1935). When multiple sources of variance need to be controlled, the full factorial design is an efficient way of doing so and allows for full study of each factor and interactions between factors (Fisher, 1935; Cochran and Cox, 1957; Box et al., 1978). However, as the number of factors increases, sample size requirements on become untenable. Randomized block designs represent a possible solution to avoid certain types of confounding. A more efficient solution that requires fewer observations are carefully planned fractional factorial designs, including Latin Squares, and related approaches such as Orthogonal Array designs.

Latin squares allow for a single treatment factor of interest, several nuisance factors, and a need for efficiency (Cox, 1958). The classic Latin Square design uses two nuisance factors as blocking variables, where a tabular grid is constructed and each row and column receive each treatment combination exactly once (Box, Hunter, and Hunter, 1978). When more than two nuisance factors are present, the Graeco-Latin Square design permits three nuisance factors, and the hyper-Graeco-Latin Square design permits four. The Latin Square design requires that the number of levels of the blocking variables must equal the number of levels in the treatment, and the design assumes no interaction between blocking factors or between the treatment and individual blocking factors.

Orthogonal Arrays represent a reformulation and generalixation of Latin Squares, and allowing for multifactor experiments with treatment factors that have varying numbers of levels (Kacker, Lagergen, and Filliben, 1991; Taguchi, 1987). Taguchi denotes the arrays $L_N(s^m)$ with s elements in an N x m matrix where the L stands for Latin square. The columns of the experimental matrix can be viewed as factors, the entries in the columns being the test levels of those factors, and the rows are the runs or study participants. Orthogonal arrays have the property that every pair of columns have the ordered pairs of elements appearing the same number of times, as seen in Figure 46:

Figure 46: Orthogonal Array

1	1	1
0	0	1

1	0	0
0	1	0

This results in the design being balanced and allowing for all levels of the factors to be equally considered. Additionally, it permits the factors to be evaluated independently, uncorrelated, of each other despite the fractionality of the design (Bose, 1947; Bose and Bush, 1952). The main effects and first-order interactions are considered while higher-order interactions are assumed to be nonexistent. The flexibility of the design allows for any column to be removed and the table remain an orthogonal array (Taguchi, 1987). Orthogonal Arrays remain Orthogonal Arrays under row permutation, column permutation, and elements within a column being permuted as well. With its efficiency, flexibility, and generalizibility, Orthogonal Arrays offer an efficient study design option.

In the simulated data, the flexibility and efficiency of the Orthogonal Array will be utilized to represent possible EMA data. To be consistent with our study, an orthogonal array with 36 subjects, or rows, was selected with 8 binary variables and 3 variables with 6 levels (Zhang et al., 2001). The orthogonal array is presented below in Figure 47:

Figure 47: Study Orthogonal Array

1	0	0	0	0	0	0	0	0	0	0
4	1	1	0	0	0	1	0	1	0	1
2	0	1	1	1	1	1	0	0	0	1

3	1	0	1	1	0	1	0	0	1	0
1	2	4	1	0	0	1	1	1	1	1
2	3	5	0	1	0	0	1	0	1	1
4	2	5	0	0	1	1	1	0	1	0
3	3	4	1	0	1	0	1	0	0	1
3	4	2	0	1	1	0	0	1	1	1
2	5	3	1	0	1	0	0	1	1	0
4	4	3	1	1	0	0	1	1	0	0
1	5	2	0	1	1	1	1	1	0	0
2	2	2	0	0	0	0	0	0	0	0
5	3	3	0	0	0	1	0	1	0	1
0	2	3	1	1	1	1	0	0	0	1
4	3	2	1	1	0	1	0	0	1	0
2	4	0	1	0	0	1	1	1	1	1
0	5	1	0	1	0	0	1	0	1	1
5	4	1	0	0	1	1	1	0	1	0
4	5	0	1	0	1	0	1	0	0	1
4	0	4	0	1	1	0	0	1	1	1
0	1	5	1	0	1	0	0	1	1	0
5	0	5	1	1	0	0	1	1	0	0
2	1	4	0	1	1	1	1	1	0	0
0	4	4	0	0	0	0	0	0	0	0
3	5	5	0	0	0	1	0	1	0	1

1	4	5	1	1	1	1	0	0	0	1
5	5	4	1	1	0	1	0	0	1	0
0	0	2	1	0	0	1	1	1	1	1
1	1	3	0	1	0	0	1	0	1	1
3	0	3	0	0	1	1	1	0	1	0
5	1	2	1	0	1	0	1	0	0	1
5	2	0	0	1	1	0	0	1	1	1
1	3	1	1	0	1	0	0	1	1	0
3	2	1	1	1	0	0	1	1	0	0
0	3	0	0	1	1	1	1	1	0	0

It should be noted, concerning the interaction terms and a statistical model's ability to estimate corresponding regression coefficients, an interaction term in this context constructed by multiplying two binary variables will have 9 non-zero elements, which makes it more likely that this term's parameter may be estimated. However, and interaction term between a binary and 6-level column and two 6-evel columns only have 3 and 1 non-zero term for each level of the 6-level column. With so little information at each level, estimation of interaction parameters may be difficult or, in some cases, impossible without relying on assumptions that go beyond the usual foundations of parameter estimation in regression models.

5.1.2 Missing Data

Missing data and nonresponse are likely to occur in EMA data, and the structure of the orthogonal array study design present new challenges to imputation of these data. Given the

survey nature of the data, distinct patterns of missing data are more probable than others. For instance, a monotone pattern of missing data, where the rate of missing data increases as the number of questions or surveys increase, may be expected. During the actual study, EMAs will be delivered at a set time and remain available for several hours after. The variability in the day that the EMA is given and the week of the study are apt to be contributors to the probability of EMA non-response, which could lead to data where subjects have missing data on certain days of the week over the study. Additionally, a monotone missing data pattern would not be surprising, where the probability of non-response increases as time in the study increases.

When handling a missing data problem, it is important to consider the mechanism that induces the missing data for both modeling and imputation considerations (Little and Rubin, 2002; Rubin, 2004). To better understand the mechanism that leads to missing data, we consider the joint distribution of the outcome variable $Y = (Y_1, Y_2, \ldots, Y_n)$ and a missing/observed indicator variable $M = (M_1, M_2, \ldots, M_n)$ where $M_i = 1$ if tooth i is missing and 0 otherwise:

$$P(Y, M|X, \beta, \theta) = P(M|\theta, X, Y)P(Y|X, \beta)$$

In this formula, $Y = \{Y_{obs}, Y_{mis}\}$ is composed of the observed and the missing data, β are the parameters of the model of interest, θ are the parameters for this missing-data mechanism function, and X are the observed covariates. Note that this uses a selection model factoring of the joint distribution and assumes conditional independence between θ and β. The factored joint distribution has two parts: the missing-data mechanism $P(M|\theta, X, Y)$ and the model of interest $P(Y|X, \beta)$.

Under Rubin's classification system (Rubin, 1976), the missing-data mechanism can be either missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). This assumption made about the missing-data mechanism directly affects how the joint distribution is modeled. In a missing data situation where outcome values, Y, are MCAR, the probability of missing data in Y is unrelated to other measured or unmeasured variables, parameters of interest, and underlying values of the incomplete data in Y. This implies that a parameter, independent of other parameters in the model and independent of the observed and unobserved data, generated the missing data and the missing-data mechanism can be modeled as shown below:

$$P(Y, M|X, \beta, \theta) = P(Y_{obs}, Y_{mis}|X, \beta)P(M|\theta, X, Y_{obs}, Y_{mis}) = P(Y_{obs}, Y_{mis}|X, \beta)P(M|\theta)$$

$$\int P(Y_{obs}, Y_{mis}|X, \beta)P(M|\theta)dY_{mis} = P(Y_{obs}, |X, \beta)P(M|\theta)$$

Missing data in the MCAR case are unsystematic, and the observed data can be viewed as a random subsample of the hypothetically complete data. MCAR is considered a strict assumption because it requires that missingness be unrelated to any of the data collected in the study, a scenario that is unlikely to be satisfied unless it is imposed by design. A less strict assumption for the missing-data mechanism, often used in practice, is the MAR assumption. Under the MAR assumption the probability of missing data in Y is fully explained by the observed data and not dependent on the underlying values of the incomplete data in Y:

$$P(Y_{obs}, Y_{mis}|X, \beta)P(M|\theta, X, Y_{obs}, Y_{mis}) = P(Y_{obs}, Y_{mis}|X, \beta)P(M|\theta, X, Y_{obs})$$

$$\int P(Y_{obs}, Y_{mis}|X,\beta)P(M|\theta,X,Y_{obs})dY_{mis} = P(Y_{obs},|X,\beta)P(M|\theta,X,Y_{obs})$$

MCAR and MAR assumptions are related to the concept of ignorable missing data (Rubin, 1976 and 1978; Mealli and Rubin, 2015). From a Bayesian perspective, if the missing data are ignorable, MCAR or MAR, it means that the posterior distribution is conditionally independent, given the observed data, of the indicators for missingness status (Gelman et al., 2013). The conditional independence implies that there is no need to model the missing-data mechanism represented by $P(M|\theta)$ under MCAR and $P(M|\theta,X,Y_{obs})$ under MAR to achieve valid Bayesian or likelihood-based inference for θ.

However, when the target posterior distribution depends on the missingness indicators and joint modeling for the data and the indicators is necessary for valid inference, this is known as non-ignorable missing data. The concept of non-ignorable missing data, where the missing-data mechanism must be jointly modeled with the data for valid inference, occurs when the missing data are missing not at random (MNAR). Formally, missing data are MNAR when the probability of missingness is systematically related to the hypothetical values that are missing. Therefore, there is no way to simplify the original factorization of the joint distribution:

$$P(Y_{obs}, Y_{mis}|X,\beta)P(M|\theta,X,Y_{obs},Y_{mis})$$

For analysis of dental caries data, we start by assuming the missingness is ignorable and model the data, including the imputation process, without jointly modeling the missing-data mechanism. In Bayesian analysis this is done by replacing the missing data with parameters and then running the model as would be done if the data were fully observed (Gelman et al., 2013).

This assumes that the imputation model and the model run on the data are the same. In our data, the outcome variable, the EMA data, is the only data experiencing missing values. The predictor variables, which are fully observed, are binary and categorical, which are shown in Figure 2. These same covariates will be used in all models in this chapter. The next section will discuss the modeling techniques required when the missing data are assumed to be MNAR.

5.1.3 Pattern-Mixture Model

Pattern-mixture models (Rubin, 1974; Little, 1992) are an alternative way of jointly modeling (R, Y) where the complete data distribution is a mixture of observed and missing data distributions:

$$P(Y_{obs}, Y_{mis}, M|X, \beta, \theta) = P(Y_{obs}, Y_{mis}|X, \beta, M)P(M|\theta, X)$$

where $P(M|\theta, X)$ is usually a logistic or probit model (we will be using a Logistic model for our analysis) and $P(Y_{obs}, Y_{mis}|X, \beta, M)$ is a model that fits the observed data.

5.1.4 Pattern-Mixture Models with Multiple Patterns

Pattern-mixture models were presented in the previous section using the observed/missing indicator variable M, which was unique for each value for Y. However, M can also be a categorical variable where the different values represent a separate pattern of missing data. This implies that pattern of the missing data is informative and should be accounted for opposed to just the observed/missing status of the Y observation. The joint model is similar to the previous formulation, but specific to the pattern of missingness M = m:

$$P(Y_{obs}, Y_{mis}, M|X, \beta, \theta) = P(Y_{obs}, Y_{mis}|X, \beta_m, M = m)P(M = m|\theta, X)$$

This makes the parameters pattern dependent. A drawback of this framework is that the marginal density of the full data is not explicitly represented. Rather, a mixture of the missingnss pattern distributions is required to get the full data distribution:

$$P(Y_{obs}, Y_{mis}|X, \beta, \theta) = \sum_{m} P(Y_{obs}, Y_{mis}|X, \beta_m, M = m) P(M = m|\theta, X)$$

This is similar to the case when R is the observed/missing indicator variable, but requires more summed values to account for all missingness patterns. In EMA data, the patterns of missigness are informative as non-response occurs in certain patterns. The pattern of non-response observed in the EMA data can inform the modeling process. Thus, pattern-mixture models will be one of the models run on the EMA data.

5.2 Modeling Strategies for Orthogonal Array Study Design with Missing Data

The orthogonal array study design is efficient and allows for analysis of the main effects to be done with reduced sample size. Reduction in sample size is advantageous for reduction of cost and other organizational issues in studies, but present new challenges for statistical analysis. Small sample sizes result in fewer degrees of freedom and the orthogonal array used in this chapter has 11 covariates of interest (23 if dummy variables are created for the three categorical variables), which makes each modeling decision important as degrees of freedom are at a premium. In particular, this creates a challenge when controlling of interaction terms. With 23

covariates (including indicator variables for levels of the categorical variables), having all two-way interaction terms in a model without regularization is not feasible due to there being more covariates than observations in the data. Thus, it must be assumed that there are no interaction terms present and that only inference on the main effects is possible. This section presents alternative strategies to modeling interaction terms in the orthogonal array study design and later will be implemented in three simulated data sets, which have varying orders of interaction terms present (none, two-way only, and two and three-way interaction terms).

5.2.1 Hierarchical Bayesian Model

Rubin, Schafer, and Schenker, working with the U.S. Census Bureau's post-enumeration survey, developed a pattern-mixture model to impute missing categorical data (Rubin, Schafer, and Schenker, 1988). Their approach to the problem of missing data in the census will serve as the basis for our pattern-mixture model and modeling strategy for interaction terms in other models. In their model, to permit high-order interaction terms, they utilized a hierarchical Bayesian imputation model:

$$\log(\theta_{ijk..p}) = \mu + \mu_{1(i)} + \mu_{2(j)} + \cdots + \mu_{8(p)} + \mu_{12(ij)} + \mu_{13(ik)} + \cdots + \mu_{123..8(ijk..p)}$$

Where $\theta_{ijk..p}$ is the probability an observation falls in cell ijk…p and the $\mu's$ represent the main effects, one way, two way, three way, and high-order interaction terms. The prior distributions placed on the interaction terms allow for increased smoothing the higher the order the interaction term:

$$\mu_i \sim N(0, \sigma^2)$$

$$\mu_{ij} \sim N\left(0, \frac{\sigma^2}{\tau}\right)$$

$$\mu_{ijk} \sim N\left(0, \frac{\sigma^2}{\tau^2}\right)$$

.

.

$$\mu_{ijk..p} \sim N\left(0, \frac{\sigma^2}{\tau^7}\right)$$

For $\sigma > 0$ and $\tau > 1$. This hierarchical Bayesian model permits high-order interaction terms to be in the model, but smoothes them heavily toward zero as order of the interaction term increases. Additionally, minimizing the number of parameters present in the model is possible by having a common σ^2 and setting τ as a constant. It is important to remember that with Latin Squares and Orthogonal Arrays, the first order interaction terms can be controlled for, but higher-order terms are assumed to be non-existent. Therefore, this model will be used for the first order interaction terms and experimentation will be done with higher order interaction terms using.

5.2.2 Tukey's Test of Additivity

The hierarchical Bayesian model presented previously attempts to control for interaction terms and conserve degrees of freedom utilizing a smoothing framework. Even with the added regularization, the model can struggle in high-dimensional settings and require large τ values to achieve convergence. Additionally, the orthogonal array design has little data or replicate values in the data set for two-way and three-way interaction terms and with missing data present there is likely to be an even further reduction in the number of replicates present for analysis. Using

Figure 2, it can be seen, on average, that two-way interactions between binary variables have 9 replicates. Combinations of a 6 level categorical variable and a binary variable have 3 and two 6 categorical variables only 1 replicate for each level. This makes controlling for these effects difficult in addition to having few degrees of freedom. Tukey's test of additivity allows for assessment of whether factor variables are additively related to the expected value of the response variable (Tukey, 1949). In particular, it reduces the degrees of freedom used for interaction terms with categorical variables with more than 2 levels and permits interaction terms where there are no replicates present. In a two-way ANOVA, where i are the levels of the first main effect and j the levels for the second main effect, the Tukey test of additivity is written:

$$E[Y_{ij}] = \mu + \alpha_i + \beta_i + \lambda\alpha_i\beta_i$$

This strategy will be extended to accommodate our 11 main effects, used for controlling for interaction terms in all models as well, and compared to the previous hierarchical Bayesian model.

5.3 Simulation Studies

5.3.1 Explanation of Simulated Data Sets

The efficiency of the orthogonal array study design is desirable, but can result in few degrees of freedom and make estimation of interaction term parameters difficult or impossible. To assess the ability of the proposed models (main effects only, hierarchical Bayesian, Tukey test of additivity, and pattern-mixture model with only main effects) to capture these interaction terms, three simulated data sets were generated from: 1) model with only main effects and no interaction terms, 2) model with main effects and two-way interaction terms, and 3) model with main effects, two-way and three-way interaction terms. The parameters for the generative models

were determined by random draws from a uniform distribution and their significance was further determined by letting each coefficient have a 50% probability of being reassigned to 0. These parameters and mean structures were then used to calculate probabilities and ultimately to generate binary and categorical variables. The simulated data sets were in the form of an EMA, where there were 5 questions (2 binary and 3 categorical).

Three missing data types were then generated on the 3 previously described simulated data sets resulting in a total of 9 data sets. The first kind of missingness pattern was monotone missing data. The probability of missing data for each question was determined by a random sample from a uniform distribution with increasing maximum and minimum for each question [(0, 0.1), (0.05, 0.2), (0.15, 0.25), (0.2, 0.35), and (0.3, 0.4)] to increase the likelihood of missing data for the later questions in the EMA. Next, MAR data were imposed on the three simulated data sets where the same uniform distribution/monotone missing data scheme was used, but probability of missingness increased by a value randomly drawn from a Uniform(0, 0.1) if the first binary questions was answered as 1 in the orthogonal array/covariate matrix. The final pattern of missing data included the monotone missing data scheme, but added a pattern of missing data where the last three questions in the survey had increased probability of missingness, based on a random draw from Uniform(0, 0.1), if the fifth questions was answered as 1. The count distribution of missing data in the three resulting data sets for the base data set with just main effects can be seen in Figure 48:

Figure 48: Data for No Interaction Terms Surveys

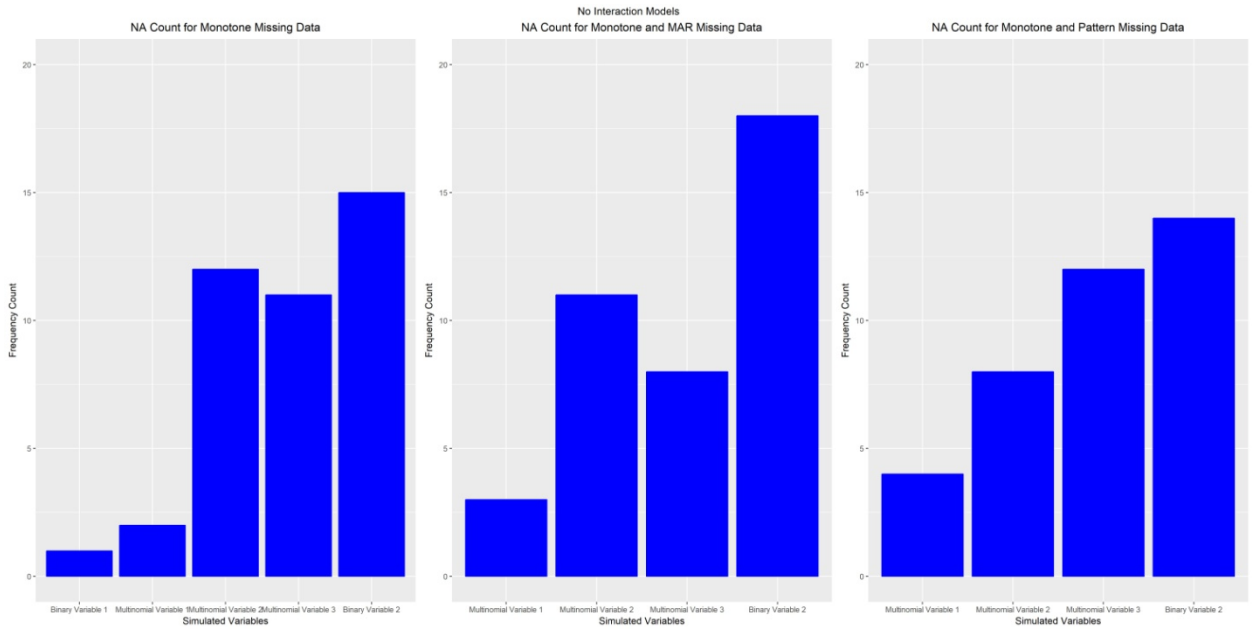
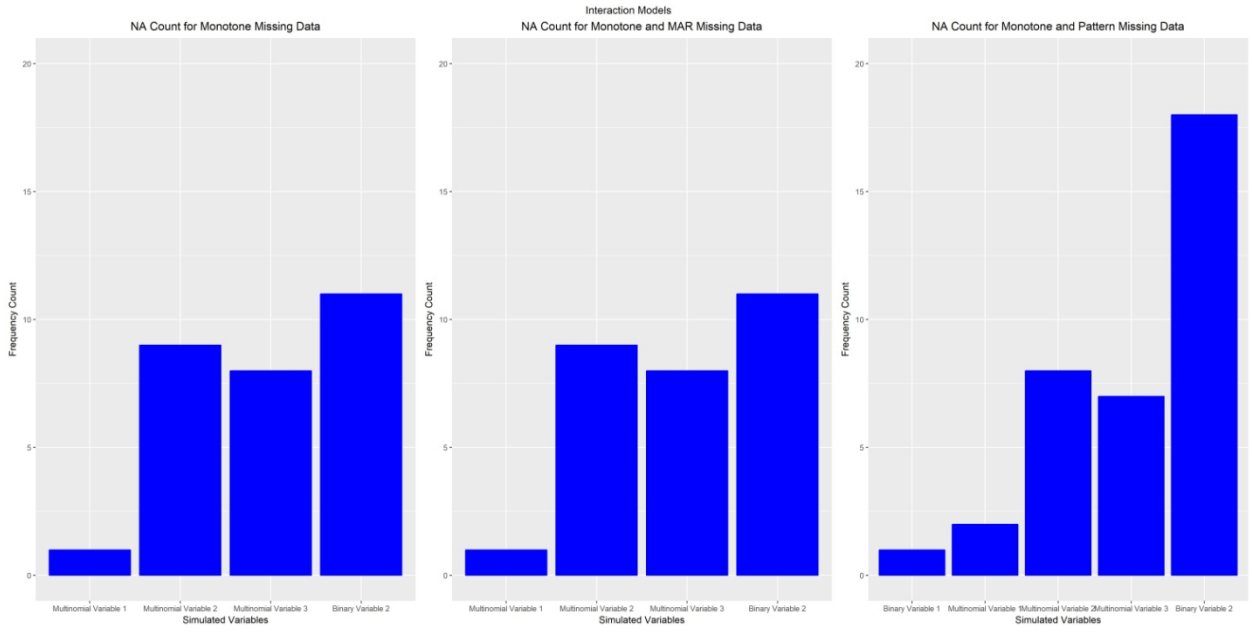


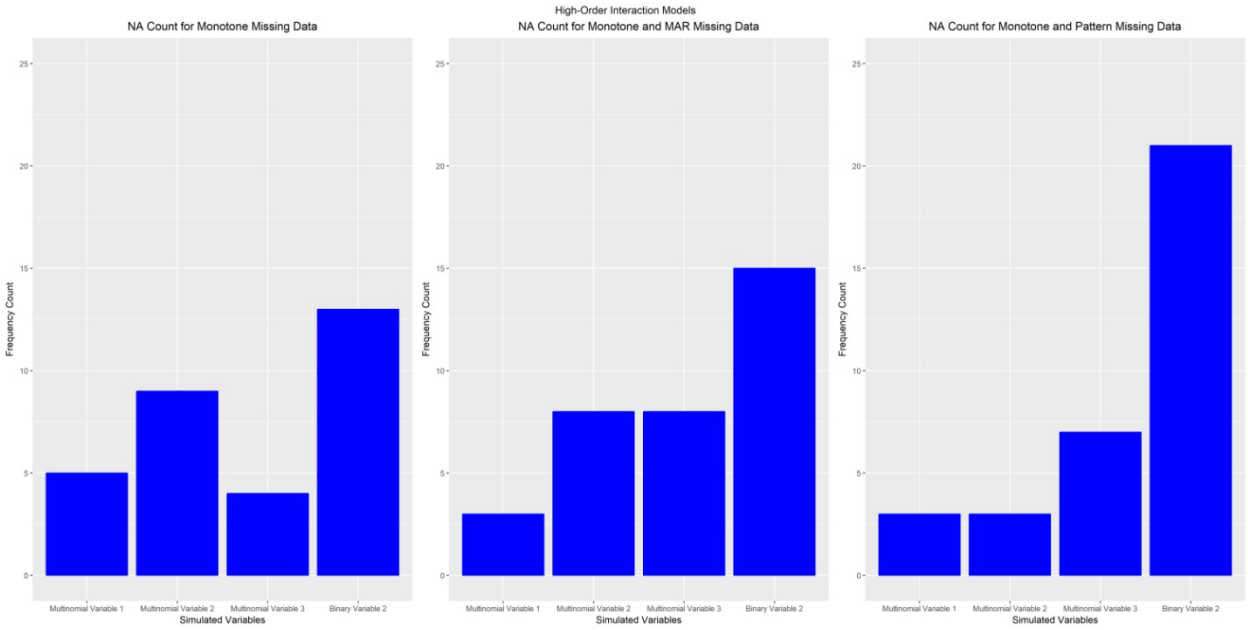
Figure 49 shows the same, but for the base data set with main effects and two-way interaction terms:

Figure 49: Data for Two-Way Interaction Terms Surveys



Lastly, Figure 50 displays the count distribution of missing data for the base data set with main effects, two-way and three-way interaction terms:

Figure 50: Data for High-Order Interaction Terms Surveys



Analysis of these data were restricted to question 5 or Binary_2 in Figure 48, 49, and 50 for simplicity. This analysis assumed independence between questions with the pattern-mixture model being the only exception for the missing data pattern. Thus, answers given in questions other than question 5 were not included as covariates in any of the models, but may be an area to further explore in future work. Given that only question 5 was used in the analysis, examination of its distribution and missing data are necessary. Figure 51 shows these distributions for the base model with only main effects:

Figure 51: Question 5 Main Effects Only Datasets

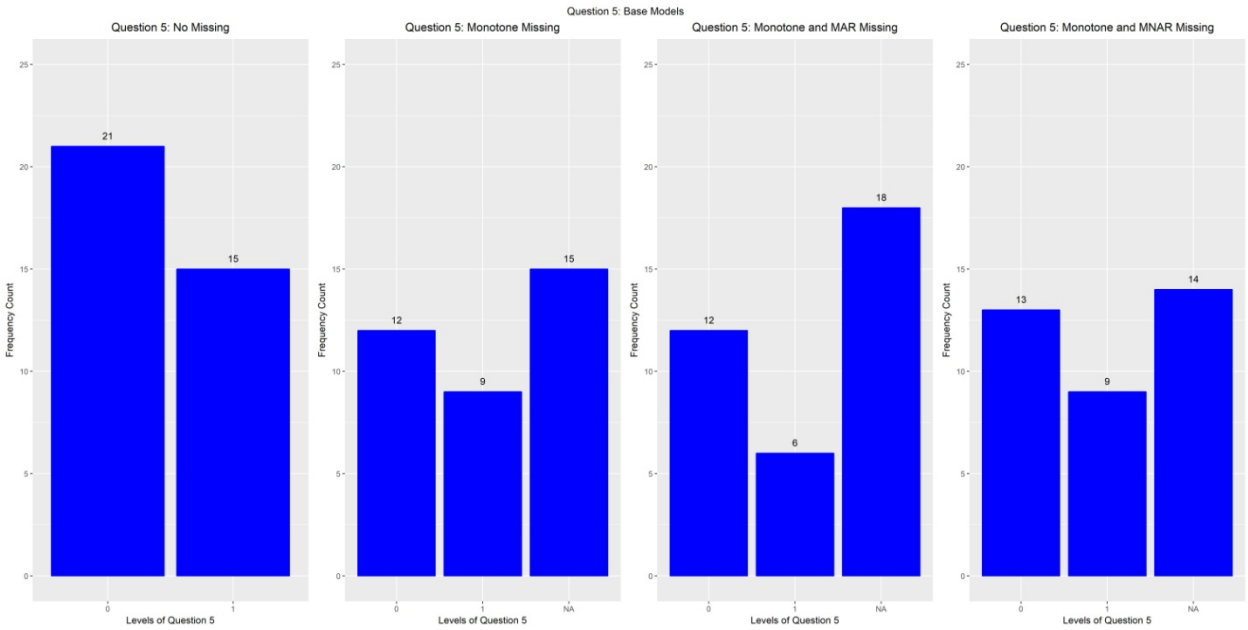
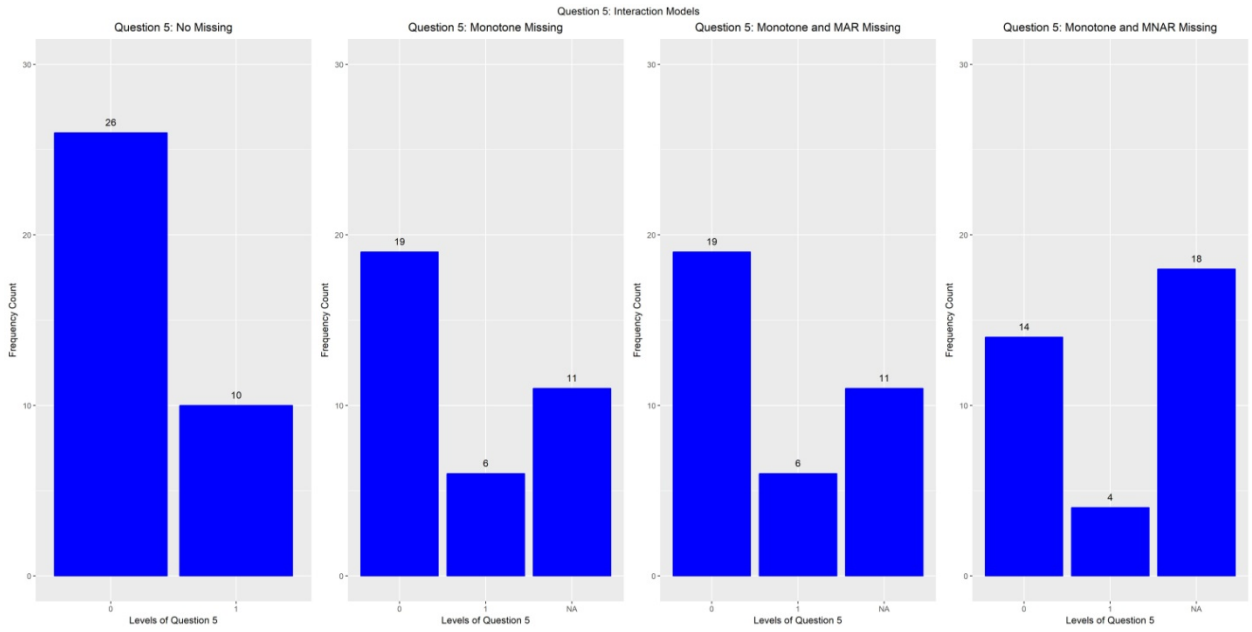


Figure 51 exhibits a distribution where the answer of 0 experiences less missing data than 1 (especially in the MAR missing data). The imbalance between will be monitored during the analysis. Figure 52 shows these same data for question 5 in the simulated data set with main effects and two-way interaction terms:

Figure 52: Question 5 Two-Way Interaction Terms Data Sets



In figure 52, we see that there were more that answered 0 than 1, as noted in Figure 51, but even more so. Figure 53, displays these distributions for the data set generated from the model with main effects, two-way and three-way interaction terms:

Figure 53: Question 5 High-Order Interaction Terms

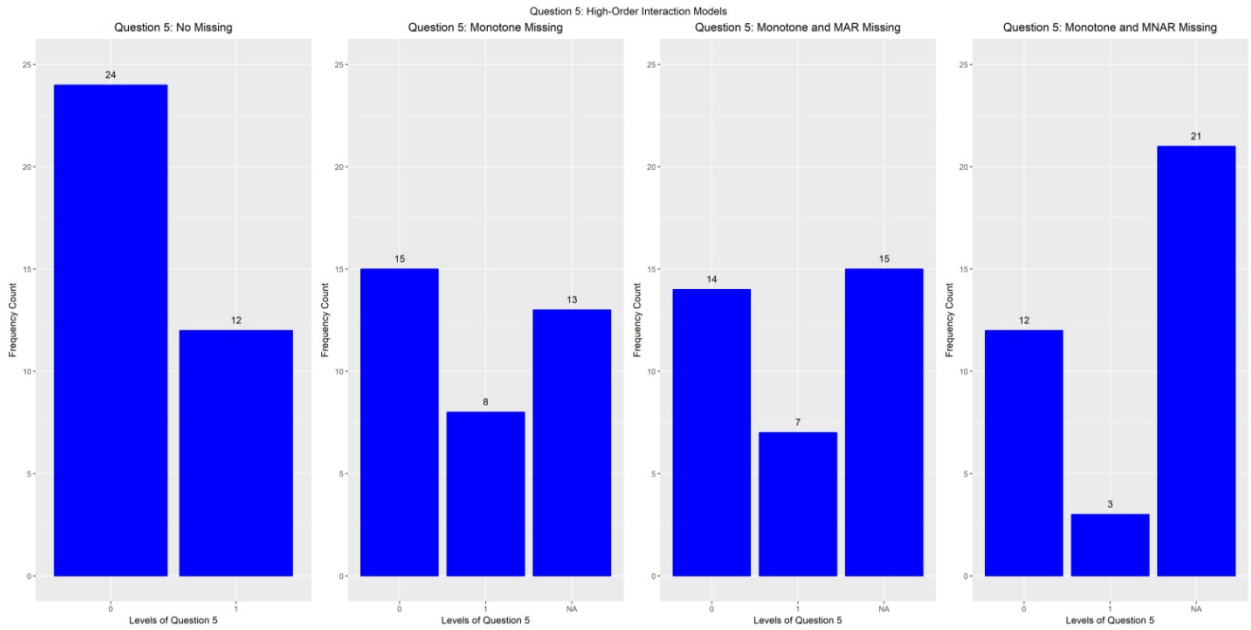


Figure 53 shows that the MNAR case has very few 1 cases that are not missing. Imputation and posterior predictive checks of these data will be important as the pattern of missingness in MNAR was constructed to increase probability of missingness if the answer was 1. This again is something that is noted and will be examined in the analysis results.

5.3.2 Prior Distributions

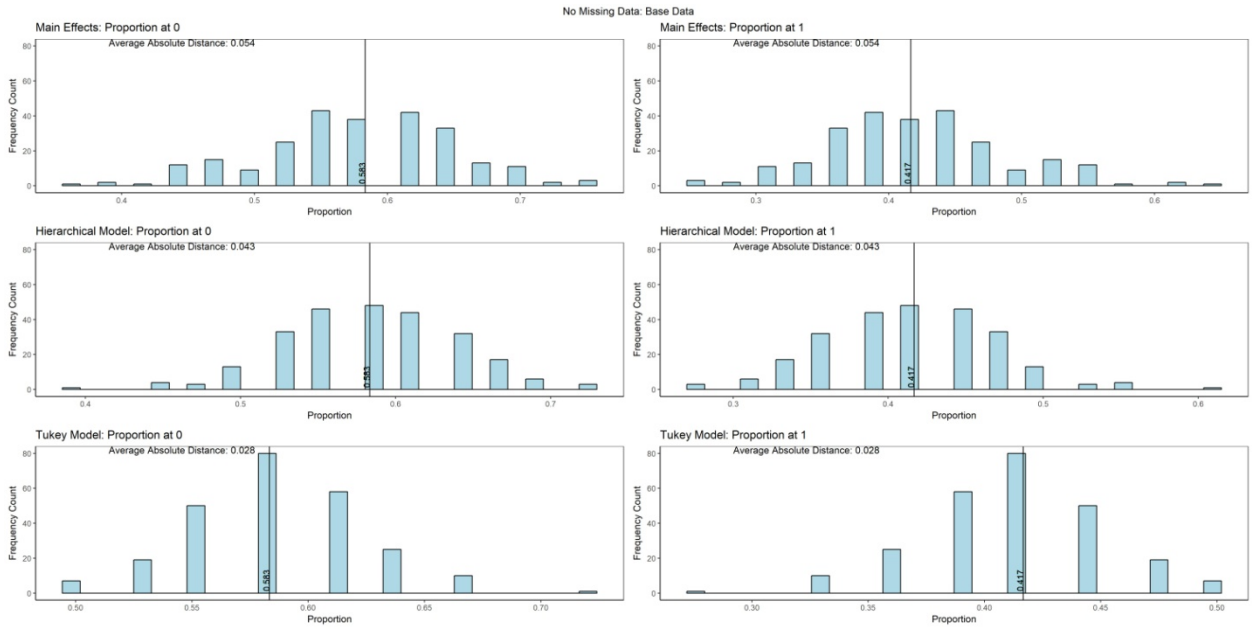
With question 5 being the outcome variable, a logistic regression model was constructed for a main effects only model (Model 23), a Tukey model (Model 24), hierarchical model (Model 25), and pattern-mixture model (Model 26). The covariate parameters were assigned a Normal prior distribution where the precision was determined by running a grid of values where one was selected for the final model based on the fit of the posterior predictive checks (PPC). In the Rubin, Schafer, and Schenker hierarchical model, a grid was also run on the smoothing parameter for interaction terms. Ultimately, the optimal precision parameter for the regression

coefficients was 1.1 for the smoothing parameter. In this case it would be multiplying rather than dividing by 1.1 to the precision in order to increase the precision of the prior or its strength of smoothing. Model fit was assessed using PPC compared to the true, simulated data. All computation was done with open source software R and JAGS, developed by Martyn Plummer and publically released in 2003, where there were 4 chains having arbitrary starting values with 5,000 total sampling iteration and a 2,500 burn in. Autocorrelation plots were examined and the Gelman-Rubin diagnostic (1992) was used to assess convergence of the chains to the same stationary distribution.

5.3.3 Simulated Data Set 1 and Results

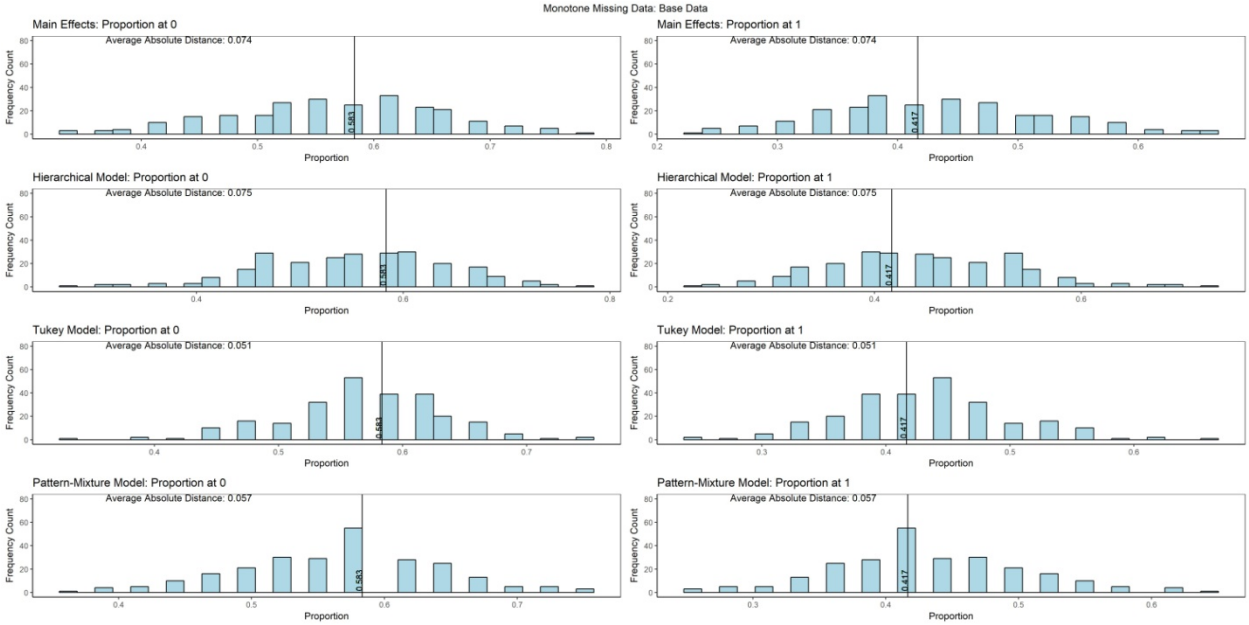
The first simulated data set was constructed using only main effects and uniformly generated regression coefficients. Additionally, each main effect coefficient had a 50% chance of being reassigned to 0. Initially, the models were run on this dataset where there were no missing values to assess how well the models fit the data. This was accomplished by generating PPC from every 20^{th} iteration of the parameter sampling (5000 iterations with 2500 burn implies 125 samples of PPC). The proportion of observations of 1's and 0's for each of the 125 generated datasets were then calculated, plotted, and compared to the true proportion values in each category for question 5 from our base (main effects only) simulated data, which is displayed below in Figure 54:

Figure 54: No Missing Data or Interaction Terms



The pattern-mixture model was not run on these data because there was no missing data present. All models perform similarly, but the Tukey model had the best performance. Additionally, the hierarchical model, with its large number of parameters, required significantly more computation time (Tukey model = 30 seconds, Hierarchical model = 300 seconds). The introduction of monotone missing data, as described previously and shown below in Figure 55, increases each model's error rate, but the Tukey model continues to distinguish itself:

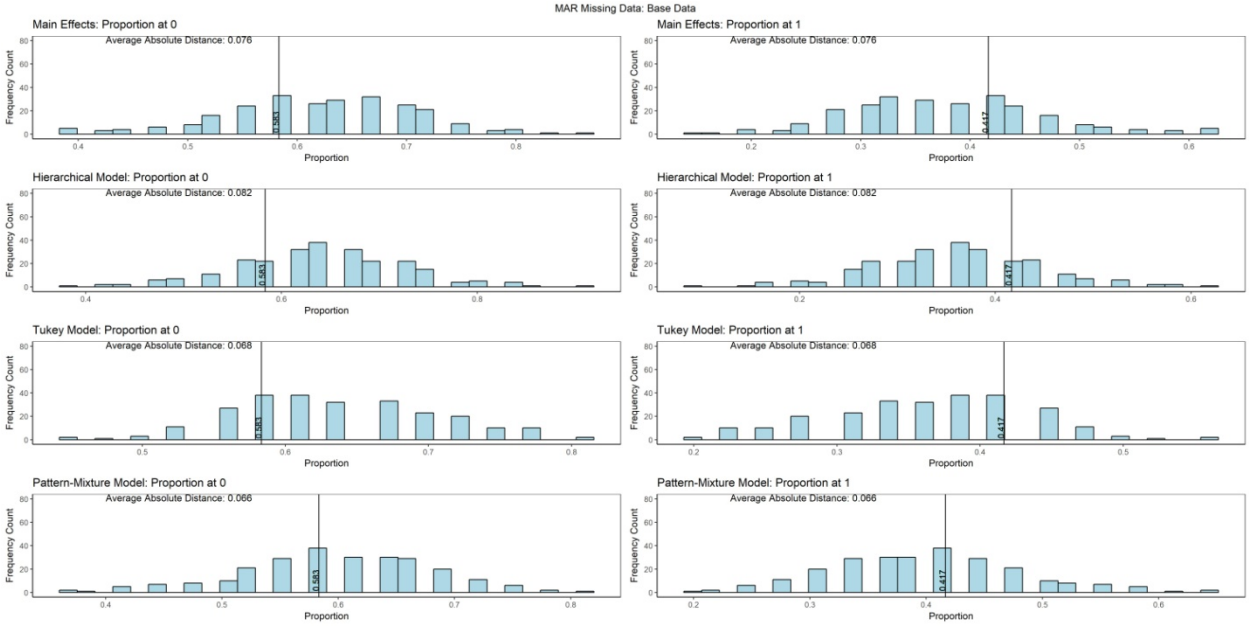
Figure 55: Monotone Missing Data and No Interaction Terms



The Tukey and Pattern-Mixture models perform best with a mean absolute error near 0.06.

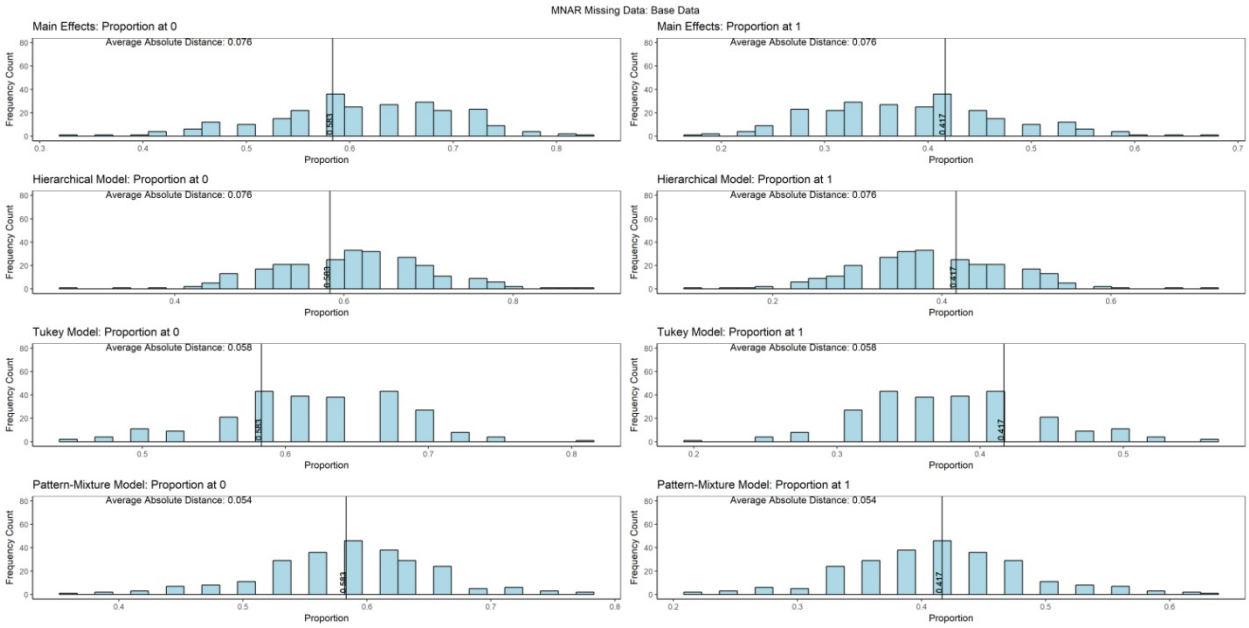
However, in the MAR missing data, all the non-base models perform similarly Figure 56:

Figure 56: MAR Missing Data and No Interaction Terms



When MNAR were introduced (Figure 57) The Tukey and Pattern-Mixture model again exhibit similar performance:

Figure 57: MNAR Missing Data and No Interaction Terms



In Figure 12, however, the pattern-mixture models mode is located near the true value of the data whereas the Tukey mode is not. Assessment of the mean posterior values for the main effects parameters of the Tukey and Pattern-Mixture model model reveled that they were similar to the true values of the model that generated the data.

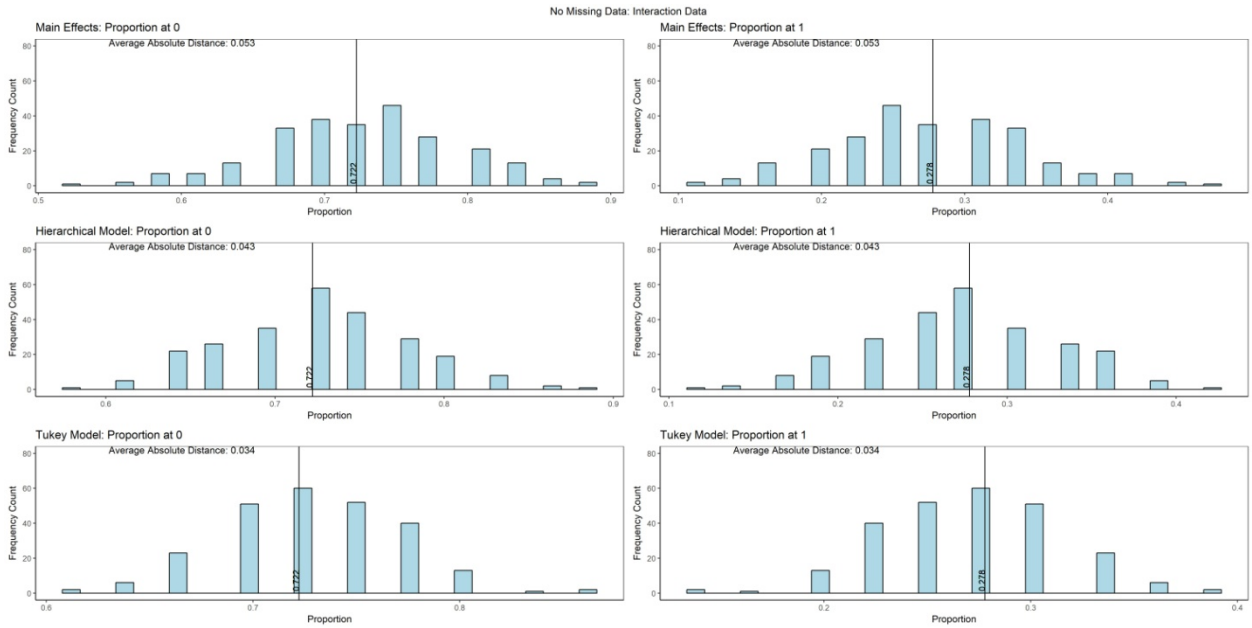
5.3.4 Simulated Data Set 2 and Results

The second data set was generated using a model that had main effects and two two-way interaction terms. The coefficient parameters were generated using samples from a uniform distribution and the main effects had a 50% chance of being reassigned to 0. The first interaction term constructed was between the first and 6 level categorical variables and the second between the last and second to last binary variables. The reason for choosing these interactions was to

illuminate a challenge in controlling for interaction terms in an orthogonal array study design: the 6 level categorical variables have limited overlap, which creates a situation, where, depending on the variables chosen in the orthogonal array, there is at most 1 non-zero interaction between each level of the two categorical variables observed when an interaction term is constructed between two 6-level categorical variable. Even without missing data being introduced to these data, this creates an estimation challenge, making estimation of the regression coefficient for the interaction term between the two 6 level categorical variables untenable. Estimation of the regression coefficient for the second interaction term between two binary variables is more promising as there significantly more non-zero instances of the variable. Interaction terms between binary and categorical variables may still prove difficult for estimation with non-zero terms having 3 total samples, but better than two categorical variables. Coefficient estimation compared to the true value of the interaction terms will be monitored and reported.

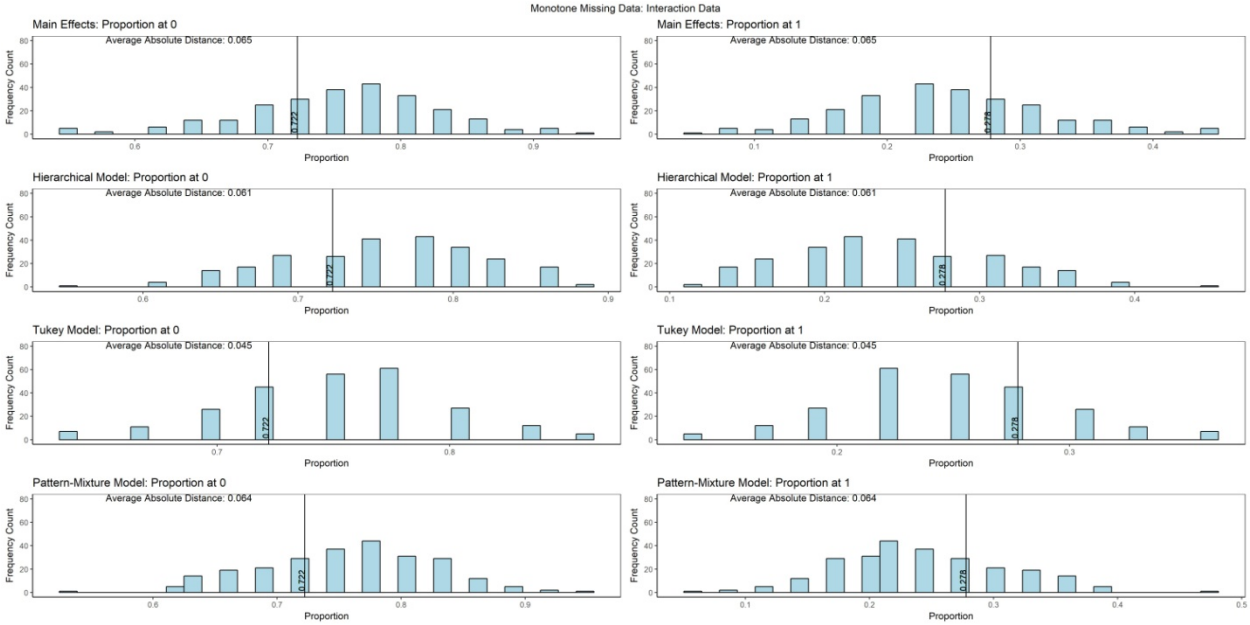
The models were initially run on the simulated data with interaction terms with no missing data present and the same metric as shown in the previous section was used to assess performance, which is displayed below in Figure 58:

Figure 58: No Missing Data and Two-Way Interaction Terms



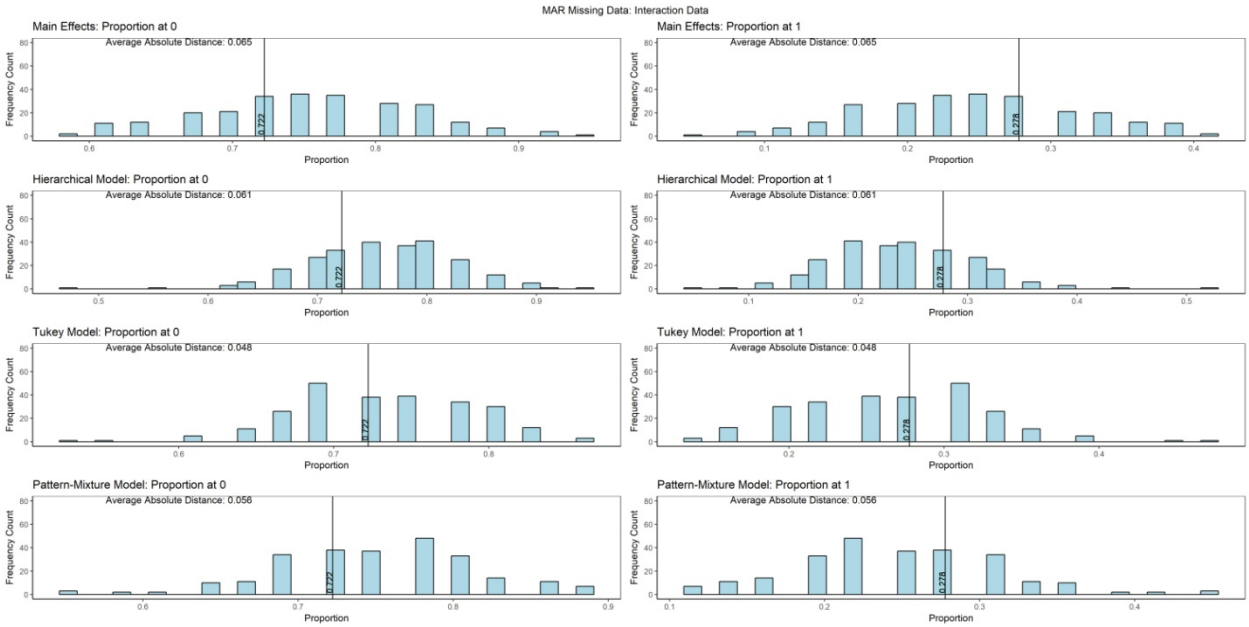
The Tukey and hierarchical model have improved performance over the main effects only, which is what is expected of these models given that they control for interaction effects. The Tukey model did surprisingly well with only 0.03 error. This error rate increased for all models with the addition of monotone missing data to the simulated data set in Figure 59:

Figure 59: Monotone Missing Data and Two-Way Interaction Terms



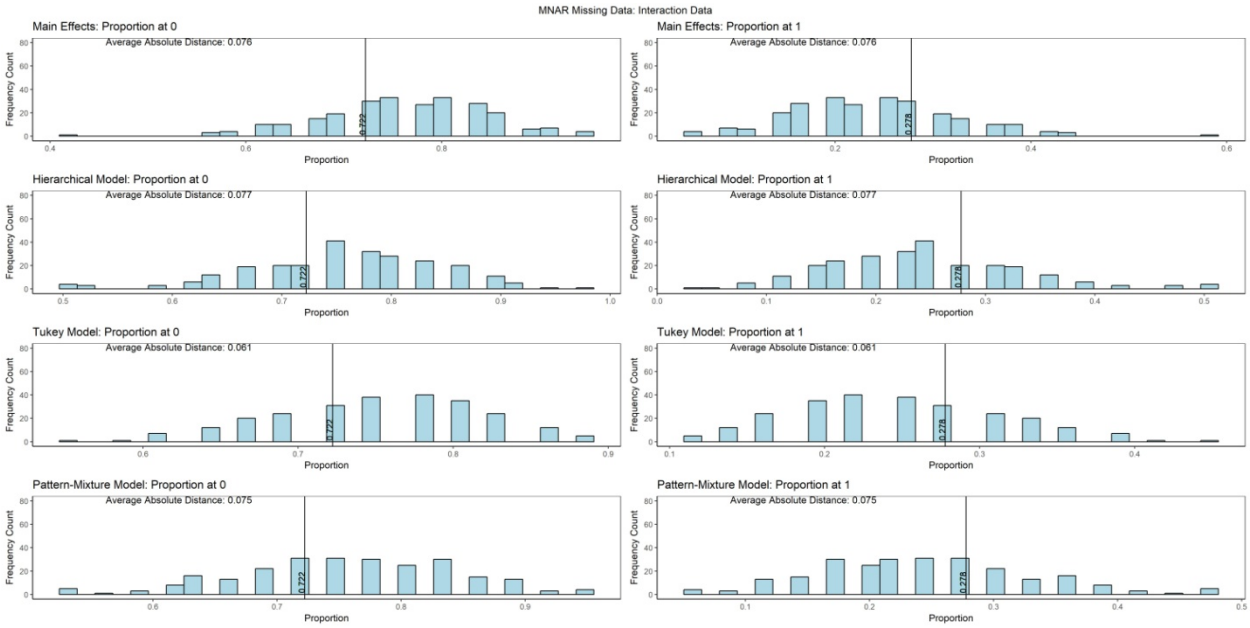
The Tukey model continues to exhibit the best performance where the other three models are all similar. There is a visible skewness to the generated distributions that is interesting. Introduction of MAR missing data sustain this trend and has slightly increased error as seen in Figure 60:

Figure 60: MAR Missing Data and Two-Way Interaction Terms



Addition of MNAR missing data again increases mean absolute error. However, the Tukey and Pattern-Mixture model have the best performance as seen in Figure 61:

Figure 61: MNAR Missing Data and Two-Way Interaction Terms



The other models all perform similarly. The introduction of interaction terms to the generative model again showed that the Tukey model outperformed the other competing models in the no missing data, monotone, and MAR cases. The Tukey model did perform well in the MNAR case and the pattern-mixture model showed promise. It should be noted that the pattern-mixture model did not contain interaction terms, but it is possible that the addition of such terms might further improve its ability to fit these data.

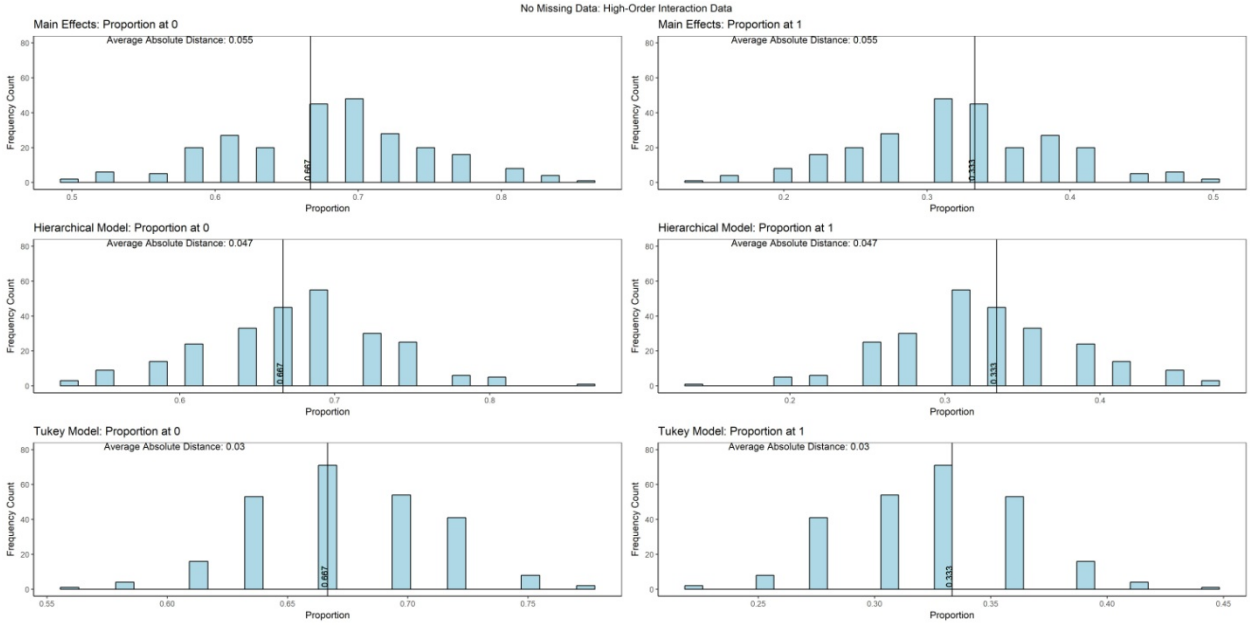
5.3.5 Simulated Data Set 3 and Results

The last data set generated contained main effects, two two-way and one three-way interaction term. The two-way interaction terms are the same as described in the previous section. However,

the one three-way interaction is composed of two binary variables and one 6 level categorical variable. Similar concerns are present for the three-way interaction as there were for the two-way interaction term between two 6 level categorical variables: the instances of non-zero combinations of the three variables involved are few. For instance, the combination of variables used for the simulation data had one non-zero instance for each level of the categorical variable. This is likely to make estimation of these effects challenging.

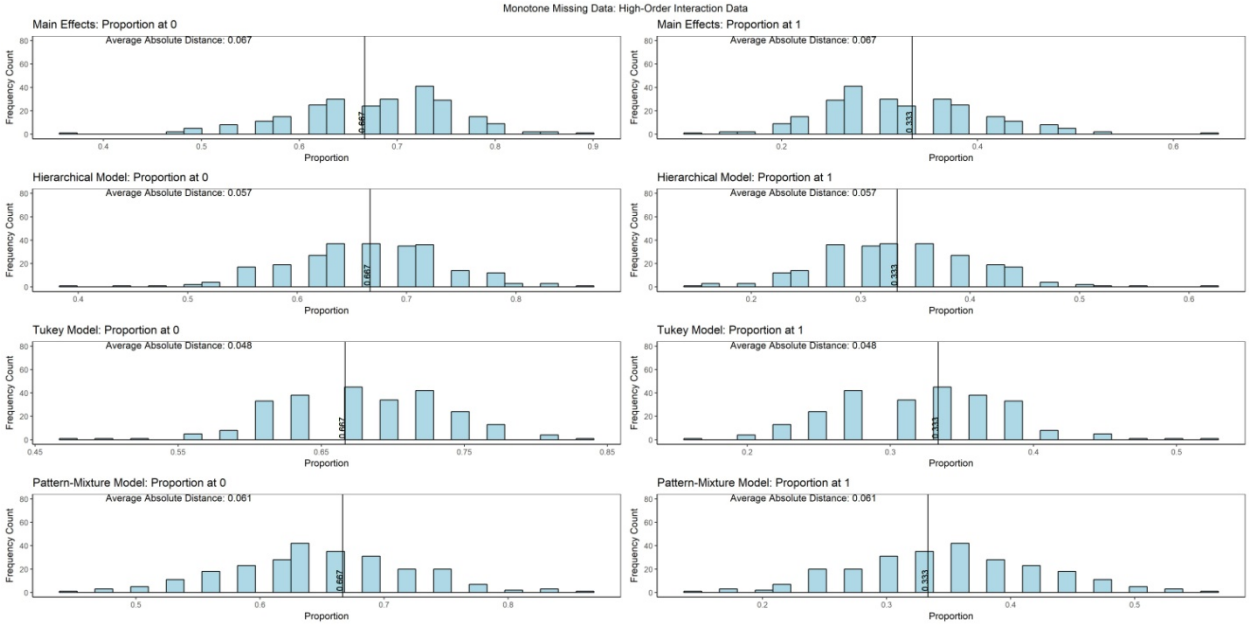
Performance of the models in the complete data case was similar to that of the two-way interaction only simulated data Figure 62:

Figure 62: No Missing Data and High-Order Interaction Terms



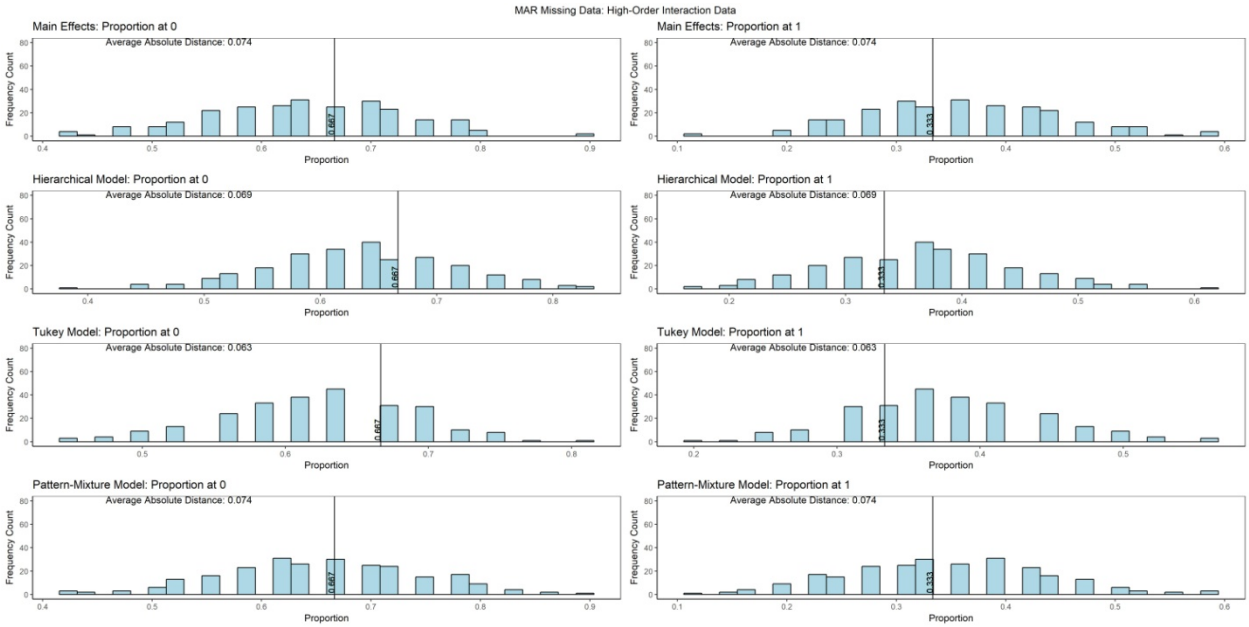
The Tukey model has the best performance out of the three models achieving the smallest error rate. This is sustained when monotone missing data are introduced which increases in error rate in all models as seen in Figure 63:

Figure 63: Monotone Missing Data and High-Order Interaction Terms



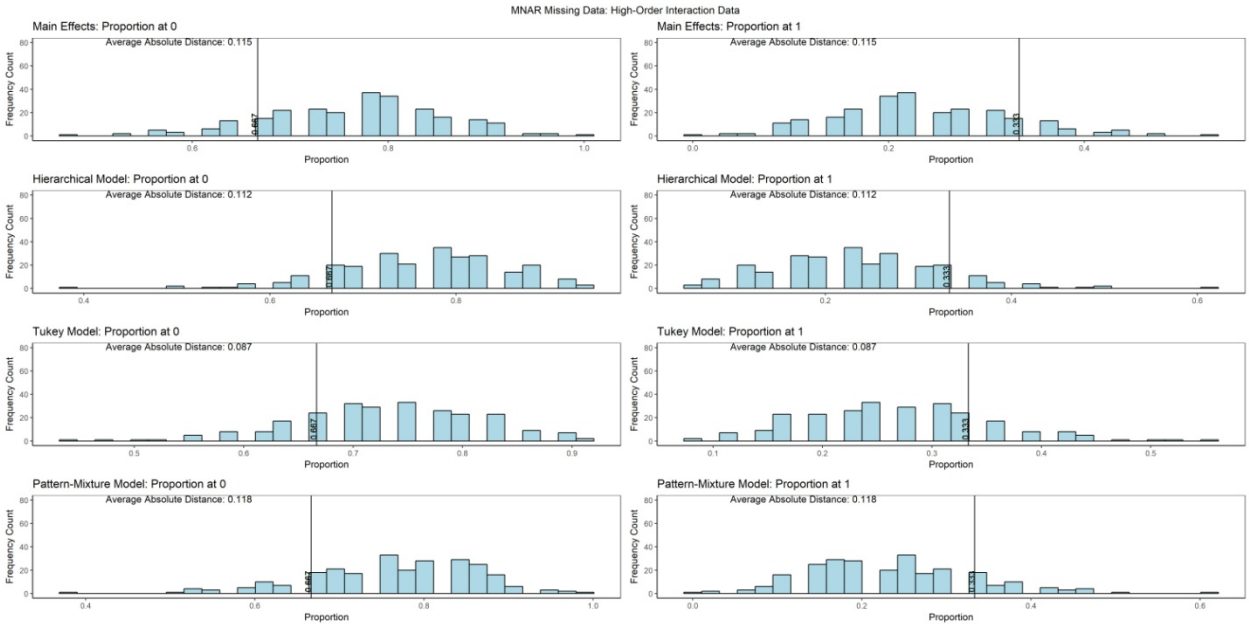
The Tukey and Patter-Mixture model perform best when MAR missing data are added to the simulated data as shown in Figure 64:

Figure 64: MAR Missing Data and High-Order Interaction Terms



The introduction of MNAR data occur, the error rates increase significantly. The Tukey model continues to be the best performer, but all models experience and doubling in their error rate. This is likely due to the percentage of missing data (58.3%) making estimation of parameters difficult:

Figure 65: MNAR Missing Data and High-Order Interaction Terms



5.3.6 Conclusion

The three simulated data sets, main effects only, main effects and two-way interactions, and main effects, two-way, and three-way interaction terms, allow for testing of the proposed models on likely situations given an EMAs orthogonal array study design. Implementation of these models has demonstrated the importance of controlling for interaction effects in the data and the challenge of high-dimensionality. The orthogonal array study design is efficient and limits the amount of data available, which creates challenges for estimating interaction effects if they are present. The Tukey model, also known as one degree of freedom for non-additivity or Tukey's

test of additivity, allows for the modeling of interaction effects at the cost of a single degree of freedom for each combination of main effect variable. In the orthogonal array design chosen for this chapter, the 6 level categorical variables benefit significantly from this efficiency. For instance, for two-way interaction effects, the Tukey stratedgy yield 55 interaction terms whereas a straight forward construction of all two-way interaction terms results in 112. This conservation of degrees of freedom is key for parameter estimation in the models and likely the reason for the Tukey model's improved performance. In the hierarchical model, which contained all 112 interactions with regularization, many of the coefficient parameters had posterior distributions centered at 0. Even with regularization, the number of terms, given that there were only 36 observations present, proved challenging. The Tukey model, with fewer degrees of freedom spent and regularization included consistently outperformed the hierarchical model.

With the improved performance of the Tukey model and its ability to capture or show that there is a non-additive relationship between binary variables, it may be a useful tool for identifying which interactions to control for. Ultimately, in the orthogonal array study design, degrees of freedom and information concerning interaction terms are at a premium. Methods that conserve degrees of freedom are necessary for determining which interaction terms are relevant to the model. The Tukey model offers a possible method for doing so. Additionally, the Tukey model performed well in complete, monotone, and MAR missing data, but struggled when MNAR data were present. Therefore, to further conserve degrees of freedom, using the Tukey model, interactions that are significant can be used in a hierarchical model and non-significant terms dropped. This will allow for more degrees of freedom and possible inclusion of the pattern-mixture model when MNAR data are suspected.

Models

Model 1: Binomial Regression in Stan

```
data {
 int N;
 int K;
 int y[N];
 int W[N];
 matrix[N, K] X;
}
parameters {
 vector[K] betas;
}
transformed parameters{
 vector[N] probs;
 probs = inv_logit(X*betas);
}
model {
 for(j in 1:K)
  // betas[j] ~ cauchy(0, 2.5);
  betas[j] ~ normal(0, tau);
 y ~ binomial(W, probs);
}
generated quantities {
  vector[N] log_lik;
  for (n in 1:N){ log_lik[n] =  binomial_lpmf(y[n] | W[n], inv_logit(X[n]*betas));}
}
```

Model 2: Poisson Regression in Stan

```
data {
        int N;
        int K;
        int y[N];
        matrix[N,K] X;
}
parameters {
        vector[K] beta;
}
transformed parameters {
        vector[N] mu;
        mu = exp(X*beta);
}
model {
        //Priors
```

```
        for(i in 1:K)
         //beta[i] ~ cauchy(0, 2.5); //prior for coefficients
         beta[i] ~ normal(0, tau);
        //Likelihood
        y ~ poisson(mu);
}
generated quantities {
  vector[N] log_lik;
  vector[N] y_pred;
  for(n in 1:N){log_lik[n] = poisson_lpmf(y[n]|exp(X[n]*beta));
                y_pred[n] = poisson_rng(mu[n]);}
}
```

Model 3: Negative-Binomial Regression in Stan

```
data {
  int N;
  int K;
  int y[N]; //
  matrix[N,K] X;
}
parameters {
  vector[K] beta;
  real<lower = 0> phi;
}
transformed parameters {
  vector[N] mu;
  mu = exp(X*beta);
}

model {
  phi ~ uniform(0, upper_bound_uniform);
  for(i in 1:K){ beta[i] ~ normal(0, tau); }
  y ~ neg_binomial_2(mu,phi);
}
generated quantities {
  vector[N] log_lik;
  for(n in 1:N){log_lik[n] = neg_binomial_2_lpmf(y[n]| mu[n], phi);}
}
```

Model 4: Beta-Binomial Regression in Stan

```
data {
  int N;
  int K; //
  int y[N];
  matrix[N,K] X;
```

```
  int W[N]; //
  int teeth;
  int<lower = 1, upper = teeth> tooth[N];
}
parameters {
  vector[K] betas;
  vector<lower = 0.001>[teeth] phi;
}
transformed parameters {
  vector[N] mu;
  vector[N] alpha;
  vector[N] beta;
  vector[N] phi_new;

  for(i in 1:N){phi_new[i] = phi[tooth[i]];}
  for(n in 1:N){
   mu[n] = inv_logit(X[n,]*betas);
   alpha[n] = mu[n] * phi_new[n];
   beta[n] = (1-mu[n]) * phi_new[n];
  }
}
model {
  phi ~ gamma(a_gamma, b_gamma);
  for(i in 1:K)
   betas[i] ~ normal(0, tau);
  y ~ beta_binomial(W,alpha,beta);
}
generated quantities {
  vector[N] log_lik;
  vector[N] y_pred;
  for(n in 1:N){log_lik[n] = beta_binomial_lpmf(y[n]| W[n], alpha[n], beta[n]);
y_pred[n] = beta_binomial_rng(W[n],alpha[n],beta[n]);}
}
```

Model 5: Over-dispersed Poisson Regression in Stan

```
data {
        int<lower = 1> N;
        int<lower = 1> K; //
        int y[N];
        matrix[N,K] X;
}
parameters {
        vector[K] beta;
        vector[N] OD_RE;
        real<lower = 0> sigma_OD;
```

```
}
transformed parameters {
        vector[N] mu;
        mu = exp(X*beta + OD_RE);
}
model {
        //Priors
        OD_RE ~ normal(0, sigma_OD);
        for(i in 1:K){
         beta[i] ~ normal(0, tau); //prior for coefficients cauchy(0, 2.5)
        }
        //Likelihood
        y ~ poisson(mu);
}
generated quantities {
  vector[N] log_lik;
  for(n in 1:N){log_lik[n] = poisson_lpmf(y[n]|mu[n]);}
}
```

Model 6: Poisson Regression with Random Effects in Stan

```
data {
        int<lower = 1> N;
        int<lower = 1> K;
        int y[N];
        matrix[N,K] X;
        int<lower = 0> teeth;
        int<lower = 0> people;
        int<lower = 1, upper = teeth> tooth[N];
        int<lower = 1, upper = people> subject[N];
}

parameters {
        vector[K] beta;
        vector[teeth] teeth_RE;
        vector[people] subject_RE;
        real<lower = 0> sigma_sub;
        real<lower = 0> sigma_tooth;
}
transformed parameters{
        vector[N] tooth_level_RE;
        vector[N] subject_level_RE;
        vector[N] mu;
        for (i in 1:N){
          tooth_level_RE[i] = teeth_RE[tooth[i]];
          subject_level_RE[i] = subject_RE[subject[i]];
```

```
        }
        mu = exp(X*beta  + subject_level_RE + tooth_level_RE);
}

model {

        //Priors
        for(i in 1:K){
         beta[i] ~ normal(0, tau); // cauchy(0, 2.5)
        }
        teeth_RE ~ normal(0, sigma_tooth);
        subject_RE ~ normal(0, sigma_sub);
        //Likelihood
        y ~ poisson(mu);
}
generated quantities{
        vector[N] log_lik;
        for(n in 1:N){log_lik[n] = poisson_lpmf(y[n]|mu[n]);}
}
```

Model 7: Over-Dispersed Poisson with Random Effects in Stan

```
data {
        int<lower = 1> N;
        int<lower = 1> K;
        int y[N];
        matrix[N,K] X;
        int<lower = 0> teeth;
        int<lower = 0> people;
        int<lower = 1, upper = teeth> tooth[N];
        int<lower = 1, upper = people> subject[N];
}

parameters {
        vector[K] beta;
        vector[teeth] teeth_RE;
        vector[people] subject_RE;
        real<lower = 0> sigma_sub;
        real<lower = 0> sigma_tooth;
        vector[N] OD_RE;
        real<lower = 0> sigma_OD;
}
transformed parameters{
        vector[N] tooth_level_RE;
        vector[N] subject_level_RE;
        vector[N] mu;
```

```
        for (i in 1:N){
          tooth_level_RE[i] = teeth_RE[tooth[i]];
          subject_level_RE[i] = subject_RE[subject[i]];
        }
        mu = exp(X*beta  + subject_level_RE + OD_RE + tooth_level_RE);
}
model {
        //Priors
        OD_RE ~ normal(0, sigma_OD);
        for(i in 1:K){
         beta[i] ~ normal(0, tau); // cauchy(0, 2.5)
        }
        teeth_RE ~ normal(0, sigma_tooth);
        subject_RE ~ normal(0, sigma_sub);

        //Likelihood
        y ~ poisson(mu);
}
generated quantities{
        vector[N] log_lik;
        for(n in 1:N){log_lik[n] = poisson_lpmf(y[n]|mu[n]);}
}
```

Model 8: Negative-Binomial Regression with Random Effects in Stan

```
data {
  int<lower = 1> N;
  int<lower = 1> K;
  int y[N];
  matrix[N,K] X;
  int<lower = 0> people;
  int<lower = 0> teeth;
  int<lower = 1, upper = people> subject[N];
  int<lower = 1, upper = teeth> tooth[N];
}

parameters {
  vector[K] beta;
  real<lower = 0> phi; //
  vector[people] subject_RE;
  vector[teeth] teeth_RE;
  real<lower = 0> sigma_tooth;
  real<lower = 0> sigma_sub;
}
transformed parameters{
          vector[N] tooth_level_RE;
```

```
            vector[N] mu;
          vector[N] subject_level_RE;
            for (i in 1:N){
            tooth_level_RE[i] = teeth_RE[tooth[i]];
            subject_level_RE[i] = subject_RE[subject[i]];
            }
            mu = exp(X*beta + subject_level_RE + tooth_level_RE);
}
model {
  // Priors
  phi ~ uniform(0, upper_bound_uniform);
  subject_RE ~ normal(0, sigma_sub);
  teeth_RE ~ normal(0, sigma_tooth);
  for(i in 1:K){
   beta[i] ~ normal(0, tau);
   }
  y ~ neg_binomial_2(mu,phi);
}
generated quantities{
          vector[N] log_lik;
          for(n in 1:N){log_lik[n] = neg_binomial_2_lpmf(y[n]|mu[n], phi);}
}
```

Model 9: Beta-Binomial Regression with Random Effects in Stan

```
data {
  int N;
  int K;
  int y[N];
  matrix[N,K] X;
  int W[N];
  int<lower = 0> people;
  int<lower = 0> teeth;
  int<lower = 1, upper = people> subject[N];
  int<lower = 1, upper = teeth> tooth[N];

}
parameters {
  vector[K] betas;
  vector<lower = 0>[teeth] phi;
  vector[teeth] teeth_RE;
  vector[people] subject_RE;
  real<lower = 0> sigma_sub;
  real<lower = 0> sigma_tooth;
}
transformed parameters{
```

```
  vector[N] subject_level_RE;
  vector[N] tooth_level_RE;
  vector[N] phi_new;
  vector[N] mu;
  vector[N] alpha;
  vector[N] beta; //
  for (i in 1:N){
            phi_new[i] = phi[tooth[i]];
            subject_level_RE[i] = subject_RE[subject[i]];
            tooth_level_RE[i] = teeth_RE[tooth[i]];
   }
  for(n in 1:N){
  mu[n] = inv_logit(X[n, ]*betas + subject_level_RE[n] + tooth_level_RE[n]);
  alpha[n] = mu[n] * phi_new[n];
  beta[n] = (1-mu[n]) * phi_new[n];
  }
}

model {
  // Priors
  phi ~ gamma(a_gamma,  b_gamma);
  subject_RE ~ normal(0, sigma_sub);
  teeth_RE ~ normal(0, sigma_tooth);
  for(i in 1:K){betas[i] ~ normal(0, tau);}
  // Likelihood
  y ~ beta_binomial(W,alpha,beta);
}

generated quantities {
  vector[N] log_lik;
  vector[N] y_pred;
  for(n in 1:N){log_lik[n] = beta_binomial_lpmf(y[n]| W[n], alpha[n], beta[n]);
              y_pred[n] = beta_binomial_rng(W[n],alpha[n],beta[n]);}
}
```

Model 10: Beta-Binomial Regression with Only Subject Level Random Effects in Stan

```
data {
  int N;
  int K;
  int y[N];
  matrix[N,K] X;
  int W[N];
  int<lower = 0> people;
  int<lower = 0> teeth;
  int<lower = 1, upper = people> subject[N];
```

```
  int<lower = 1, upper = teeth> tooth[N];

}
parameters {
  vector[K] betas;
  vector<lower = 0>[teeth] phi;
  vector[teeth] teeth_RE;
  vector[people] subject_RE;
  real<lower = 0> sigma_sub;
  real<lower = 0> sigma_tooth;
}
transformed parameters{
  vector[N] subject_level_RE;
  //vector[N] tooth_level_RE;
  vector[N] phi_new;
  vector[N] mu;
  vector[N] alpha;
  vector[N] beta;
  for (i in 1:N){
            phi_new[i] = phi[tooth[i]];
            subject_level_RE[i] = subject_RE[subject[i]];
   }
  for(n in 1:N){
  mu[n] = inv_logit(X[n, ]*betas + subject_level_RE[n]);
  alpha[n] = mu[n] * phi_new[n];
  beta[n] = (1-mu[n]) * phi_new[n];
  }
}

model {
  // Priors
  phi ~ gamma(a_gamma, b_gamma);
  subject_RE ~ normal(0, sigma_sub);

  for(i in 1:K){betas[i] ~ normal(0, tau);}
  // Likelihood
  y ~ beta_binomial(W,alpha,beta);
}

generated quantities {
  vector[N] log_lik;
  vector[N] y_pred;
  for(n in 1:N){log_lik[n] = beta_binomial_lpmf(y[n]| W[n], alpha[n], beta[n]);
              y_pred[n] = beta_binomial_rng(W[n],alpha[n],beta[n]);}
}
```

Model 11: Beta-Binomial Regression with Spatial Random Effects and CAR Prior Distribution in Stan

```
functions{
real sparse_iar_lpdf(vector phi, real tau, int[,] W_sparse, vector D_sparse, vector lambda, int N,
int W_n, int people){
        row_vector[N] phit_D;
        row_vector[N] phit_W;
        vector[N] ldet_terms;

        phit_D = (phi .* D_sparse)';
        phit_W = rep_row_vector(0, N);
        for(i in 1:W_n){
                phit_W[W_sparse[i, 1]] = phit_W[W_sparse[i, 1]] + phi[W_sparse[i, 2]];
                phit_W[W_sparse[i, 2]] = phit_W[W_sparse[i, 2]] + phi[W_sparse[i, 1]];
        }
        return 0.5 * ((N - people) * log(tau) - tau * (phit_D * phi - (phit_W * phi)));
  }
matrix Kronecker(int people, int N_adj, matrix adj, int N){
        vector[people] peeps;
        matrix[N, N] adj_mat;
        matrix[people, people] ID;
        peeps = rep_vector(1, people);
        ID = diag_matrix(peeps);
        for(i in 1:people)
          for(j in 1:people)
            for(k in 1:N_adj)
              for(l in 1:N_adj)
                  adj_mat[N_adj*(i-1)+k, N_adj*(j-1)+l] = ID[i, j] * adj[k, l];
        return adj_mat;
 }
int[,] Sparse_W(int W_n, int N, matrix adj, int people, int N_adj){
        matrix[N, N] ADJ_New;
        int W_spar[W_n, 2]; //adjacency pairs
        int counter;
        ADJ_New = Kronecker(people, N_adj, adj, N);
        counter = 1;
        for(i in 1:(N - 1)){
         for(j in (i + 1):N){
          if(ADJ_New[i, j] == 1){
           W_spar[counter, 1] = i;
           W_spar[counter, 2] = j;
           counter = counter + 1;
           }
          }
         }
```

```
        return W_spar;
 }
vector Sparse_D(int N, matrix adj, int people, int N_adj){
        matrix[N, N] ADJ_New;
        vector[N] D_spar;
        ADJ_New = Kronecker(people, N_adj, adj, N);
        for(i in 1:N) D_spar[i] = sum(ADJ_New[i]);
        return D_spar;
 }
vector lamby(vector D_sparse, int N, matrix adj, int people, int N_adj){
        vector[N] invsqrtD;
        vector[N] lambs;
        matrix[N, N] ADJ_New;
        ADJ_New = Kronecker(people, N_adj, adj, N);
        for(i in 1:N){
                invsqrtD[i] = 1 / sqrt(D_sparse[i]);
        }
        lambs = eigenvalues_sym(quad_form(ADJ_New, diag_matrix(invsqrtD)));
        return lambs;
 }
}

data {
        int N;
        int K;
        int y[N];
        int W[N];
        matrix[N, K] X;
        int N_adj;
        matrix<lower = 0, upper = 1>[N_adj, N_adj] adj;
        int W_n;
        int<lower = 0> people;

}
transformed data{
        vector[N] D_sparse;
        vector[N] lambda;
        int W_sparse[W_n, 2];

        W_sparse = Sparse_W(W_n, N, adj, people, N_adj);
        D_sparse = Sparse_D(N, adj, people, N_adj);
        lambda = lamby(D_sparse, N, adj, people, N_adj);
}
parameters {
  vector[K] betas;
  real<lower = 0> phi;
```

```
  vector[N] phi_unscaled_CAR;
  real<lower = 0> tau;

}
transformed parameters {

  vector[N] mu;
  vector[N] alpha;
  vector[N] beta;
  vector[N] phi_CAR;
  phi_CAR = phi_unscaled_CAR - mean(phi_unscaled_CAR);

  for(n in 1:N)
  mu[n] = inv_logit(X[n,]*betas + phi_CAR[n] );
  alpha = mu * phi;
  beta = (1-mu) * phi;
}
model {
  phi_unscaled_CAR ~ sparse_iar_lpdf(tau, W_sparse, D_sparse, lambda, N, W_n, people);
  tau ~ gamma(a_gamma,  b_gamma);

    phi ~ uniform(0, upper_bound_uniform);
  for(i in 1:K)
   betas[i] ~ normal(0, sigma_betas);

  y ~ beta_binomial(W,alpha,beta);
}
generated quantities {
  vector[N] log_lik;
  for(n in 1:N){log_lik[n] = beta_binomial_lpmf(y[n]| W[n], alpha[n], beta[n]);}
}
```

Model 12: Beta-Binomial SANOVA Model in Stan

```
data{
 int N;
 int K1;
 int K2;
 int N_I;
 int W[N];
 int smoothy[K2];
 int y[N];
 vector[K2] zeros;
 matrix[K2, K2] ident;
 matrix[N, K2] XINT;
 matrix[N, K1] X1;
```

```
 int under;
 int<lower = 0> people;
 int<lower = 1, upper = people> subject[N];
 int teeth;
 int<lower = 1, upper = teeth> tooth[N];
}

parameters{
 vector[K1] betas;
 vector[K2] alphas;
 vector<lower = 0.0001>[under] tau;
 real<lower = 0.0001> tauM;
 vector[people] subject_RE;
 vector<lower = 0.001>[teeth] phi;
}

transformed parameters{
 vector[N] subject_level_RE;
 vector[N] muM;
 vector[K2] mu;
 vector[N] alpha;
 vector[N] beta;
 vector[N] phi_new;

 for(i in 1:N){
          subject_level_RE[i] = subject_RE[subject[i]];
          phi_new[i] = phi[tooth[i]];}


 muM = inv_logit(X1*betas + XINT*alphas + subject_level_RE);
 mu = ident*alphas;

          for(i in 1:N){
           alpha[i] = muM[i] * phi_new[i];
           beta[i] = (1-muM[i]) * phi_new[i];}
}

model{
 tau ~ gamma(a_gamma1,  b_gamma1);
 tauM ~ gamma(a_gamma2, b_gamma2);

 for(j in 1:K1){
 betas[j] ~ normal(0, sigma_betas);
 }
 for(k in 1:K2){
 alphas[k] ~ normal(0,  sigma_alphas);
```

```
 }
 y ~ beta_binomial(W,alpha,beta);
 for(q in 1:K2){
 zeros[q] ~ normal(mu[q],tau[smoothy[q]]);
 }
}

generated quantities{
 vector[N] log_lik;
 vector[N] y_pred;
 for(n in 1:N){
        log_lik[n] = beta_binomial_lpmf(y[n]| W[n], alpha[n], beta[n]);
                        y_pred[n] = beta_binomial_rng(W[n],alpha[n],beta[n]);
 }
}
```

Model 13: Beta-Binomial Hierarchical Model in Stan

```
data {
 int N;
 int N1;
 int N2;
 int N3;
 int N4;
 int K1;
 int K2;
 int K3;
 int K4;
 int teeth;
 int<lower = 1, upper = teeth> tooth[N];

 real SM;
 int y[N];
 int W[N];

 matrix[N1, K1] X1;
 matrix[N2, K2] X2;
 matrix[N3, K3] X3;
 matrix[N4, K4] X4;
}

parameters {
 vector[K1] betas;
 vector[K2] alphas;
 vector[K3] omegas;
 vector[K4] gammas;
```

```
 real<lower = 0.0001> tau;
 vector<lower = 0.001>[teeth] phi;
}

transformed parameters {
  vector[N] mu;
  vector[N] alpha;
  vector[N] beta;
  vector[N] phi_new;

  for(i in 1:N){phi_new[i] = phi[tooth[i]];}
  mu = inv_logit(X1*betas + X2*alphas + X3*omegas + X4*gammas);
  for(n in 1:N){
   alpha[n] = mu[n] * phi_new[n];
   beta[n] = (1-mu[n]) * phi_new[n];
  }
}

model {

 phi ~ gamma(a_gamma, b_gamma);

 for(j in 1:K1)
  betas[j] ~ normal(0, tau);
 for(l in 1:K2)
  alphas[l] ~ normal(0, tau/(SM)^2);
 for(m in 1:K3)
  omegas[m] ~ normal(0, tau/(SM)^3);
 for(b in 1:K4)
  gammas[b] ~ normal(0, tau/(SM)^4);

 y ~ beta_binomial(W,alpha,beta);
}


generated quantities {
  vector[N] y_pred;
  vector[N] log_lik;
  for(n in 1:N){log_lik[n] = beta_binomial_lpmf(y[n]| W[n], alpha[n], beta[n]);
              y_pred[n] = beta_binomial_rng(W[n],alpha[n],beta[n]);}
}
```

Model 14: Poisson Model in JAGS

```
model{
   ## Likelihood
```

```
   for(i in 1:N){
     y[i] ~ dpois(lambda[i])
     log(lambda[i]) <- mu[i]
     mu[i] <- inprod(beta[],X[i,])
     }
   ## Priors
   for(j in 1:K){
   beta[j] ~ dnorm(0,tau)}  # multivariate Normal prior

   for (i in 1:N) {
      y_pred[i]~dpois(lambda[i])
      }
}
```

Model 15: Negative-Binomial Model in JAGS

```
model{
   ## Likelihood
   for(i in 1:N){
     y[i] ~ dnegbin(p[i],r)
     p[i] <- r/(r+lambda[i])
     log(lambda[i]) <- mu[i]
     mu[i] <- inprod(beta[],X[i,])
   }
   ## Priors
   for(i in 1:K){
   beta[i] ~ dnorm(0,tau)}
   r ~ dunif(0,uniform_upper)
  for (i in 1:N) {
      y_pred[i]~dnegbin(p[i],r)
      }
}
```

Model 16: Beta-Binomial Model in JAGS

```
model{
  for(i in 1:N){
            #likelihood
            y[i] ~ dbinom(p[i],W[i])
            p[i] ~ dbeta(a[i], b[i])
            a[i] <- mu[i] * phi
            b[i] <- (1 - mu[i]) * phi
            logit(mu[i]) <- inprod(beta[],X[i,])
            }
            ##priors
```

```
   for(i in 1:K){
   beta[i] ~ dnorm(0,tau)}
   phi ~ dunif(0,uniform_upper)
  for (i in 1:N) {
      y_pred[i]~dbinom(p[i],W[i])
      }
}
```

Model 17: Poisson Selection Model in JAGS

```
model{
   ## Likelihood
   for(i in 1:N){
     y[i] ~ dpois(lambda[i])
     log(lambda[i]) <- mu[i]
     mu[i] <- inprod(beta[],X[i,])

   ## Missing Data Mechanism
     miss[i] ~ dbern(p[i])
     logit(p[i]) <- inprod(alpha[],X[i,]) + (xi * y[i])
     }
   ## Priors
   for(j in 1:K){
   beta[j] ~ dnorm(0,tau1)
   alpha[j] ~ dnorm(0,tau2)}
   xi ~ dnorm(0,tau3)
   for (i in 1:N) {
      y_pred[i]~dpois(lambda[i])
      }
}
```

Model 18: Negative-Binomial Selection Model in JAGS

```
model{
   ## Likelihood
   for(i in 1:N){
     y[i] ~ dnegbin(p[i],r)
     p[i] <- r/(r+lambda[i])
     log(lambda[i]) <- mu[i]
     mu[i] <- inprod(beta[],X[i,])

    ## Missingnes Mechanism
     miss[i] ~ dbern(p.miss[i])
     logit(p.miss[i]) <- inprod(alpha[],X[i,]) + (xi * y[i])

      }
```

```
   ## Priors
   for(i in 1:K){
   beta[i] ~ dnorm(0,tau1)
   alpha[i] ~ dnorm(0,tau2)}
   r ~ dgamma(a_gam,b_gam)
   xi ~ dnorm(0,tau3)
  for (i in 1:N) {
      y_pred[i]~dnegbin(p[i],r)
      }
}
```

Model 19: Beta-Binomial Selection Model in JAGS

```
model{
  for(i in 1:N){
            ##likelihood
            y[i] ~ dbinom(p[i],W[i])
            p[i] ~ dbeta(a[i] + 0.1, b[i] + 0.1)
            a[i] <- (mu[i] * phi)
            b[i] <- ((1 - mu[i]) * phi)
            logit(mu[i]) <- inprod(beta[],X[i,])


            ## missing data mechanism
            miss[i] ~ dbern(p.miss[i])
            logit(p.miss[i]) <- inprod(alpha[],X[i,]) + (xi * y[i])
}
            ##priors
   for(i in 1:K){
   beta[i] ~ dnorm(0,tau1)
   alpha[i] ~ dnorm(0,tau2)}
   phi ~ dgamma(a_gam,b_gam)
   xi ~ dnorm(0,tau3)
   for (i in 1:N) {
      y_pred[i] ~ dbinom(p[i],W[i])
      }

}
```

Model 20: Beta-Binomial Pattern-Mixture Model in JAGS

```
model{
   ## Likelihood
   for(i in 1:N){
     y[i] ~ dpois(lambda[i])
     log(lambda[i]) <- mu[i]
```

```
    mu[i] <- inprod(beta[],X[i,]) + inprod(alpha2[],X_miss2[i,]) + inprod(alpha3[],X_miss3[i,])}

  ## Priors
  for(j in 1:K){
  beta[j] ~ dnorm(0,tau1)
  alpha2[j] ~ dnorm(0,tau2)
  alpha3[j] ~ dnorm(0,tau3)}

}
```

Model 21: Beta-Binomial Pattern-Mixture Model in JAGS

```
model{
  ## Likelihood
  for(i in 1:N){
    y[i] ~ dnegbin(p[i],r)
    p[i] <- r/(r+lambda[i])
    log(lambda[i]) <- mu[i]
    mu[i] <- inprod(beta[],X[i,]) + inprod(alpha2[],X_miss2[i,]) + inprod(alpha3[],X_miss3[i,])}

  ## Priors
  for(i in 1:K){
  beta[i] ~ dnorm(0,tau1)
  alpha2[i] ~ dnorm(0,tau2)
  alpha3[i] ~ dnorm(0,tau3)}
  r ~ dgamma(a_gam,b_gam)
}
```

Model 22: Beta-Binomial Pattern-Mixture Model in JAGS

```
model{
  for(i in 1:N){
            ##likelihood
            y[i] ~ dbinom(p[i],W[i])
            p[i] ~ dbeta(a[i] + 0.1, b[i] + 0.1)
            a[i] <- mu[i] * phi
            b[i] <- (1 - mu[i]) * phi
            logit(mu[i]) <- inprod(beta[],X[i,]) + inprod(alpha2[],X_miss2[i,]) +
inprod(alpha3[],X_miss3[i,])
            }
            ##priors
  for(i in 1:K){
  beta[i] ~ dnorm(0,tau1)
  alpha2[i] ~ dnorm(0,tau2)
  alpha3[i] ~ dnorm(0,tau3)}
  phi ~ dgamma(a_gam, b_gam)
```

```
}
```

Model 23: Base Model in JAGS

```
model{
  ## Likelihood
  for(i in 1:N){
        y[i] ~ dbern(p[i])
        p[i] <- ilogit(inprod(beta1[],X[i,]))

        }

   ## Priors
   for(j in 1:K){
        beta1[j] ~ dnorm(0, tau)
        }

   for(i in 1:N){
        y_pred[i] ~ dbern(p[i])
        }
}
```

Model 24: Tukey Model in JAGS

```
model{
  ## Likelihood
  for(i in 1:N){
        y[i] ~ dbern(p[i])
        p[i] <- ilogit(inprod(beta1[],X[i,]) + inprod(lambda[], tuk[i,]))

        }

   ## Priors
   for(l in 1:P){
   lambda[l] ~ dnorm(0, tau1)
        }

   for(j in 1:K){
        beta1[j] ~ dnorm(0, tau2)
        }

   for(i in 1:N){
        y_pred[i] ~ dbern(p[i])
        }
}
```

Model 25: Hierarchical Model in JAGS

```
model{
  ## Likelihood
  for(i in 1:N){
            y[i] ~ dbern(p[i])
            p[i] <- ilogit(inprod(beta1[],X[i,]) + inprod(beta2[], int_2[i,]))
            }

   ## Priors
   for(j in 1:N_base){
            beta1[j] ~ dnorm(0, tau1)
            }
   for(k in 1:N_int1){
            beta2[k] ~ dnorm(0, tau2)
            }

   for(i in 1:N){
            y_pred[i] ~ dbern(p[i])
            }
}
```

Model 26: Pattern-Mixture Model in JAGS

```
model{
  ## Likelihood
  for(i in 1:N){
            y[i] ~ dbern(p[i])
            p[i] <- ilogit(inprod(beta1[],X[i,]) + inprod(alpha1[],X_int[i,]))
            }

   ## Priors
   for(j in 1:K){
            beta1[j] ~ dnorm(0, tau1)
            alpha1[j] ~ dnorm(0, tau2)
            }

   for(i in 1:N){
            y_pred[i] ~ dbern(p[i])
            }
}
```

References

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*,*60*(2), 255-265.

Alessi, S.M, Petry, N.M. (2013). A randomized study of cellphone to reinforce alcohol abstinence in the natural environment. Addiction, 108(5): 900-909.

Allan, F. E., & Wishart, J. (1930). A method of estimating the yield of a missing plot in field experimental work. *The Journal of Agricultural Science*, *20*(3), 399-406.

Altinisik, Y. (2013). Intrablock, Interblock and Combined Estimates in Incomplete Block Designs: A Numerical Study.

Arnold, B. C., & Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhyā: The Indian Journal of Statistics, Series B*, 233-243.

Bandyopadhyay, D., Reich, B. J., & Slate, E. H. (2011). A spatial beta-binomial model for clustered count data on dental caries. *Statistical methods in medical research*, *20*(2), 85-102.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Crc Press.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 192-236.

Best, N. and Mason, A. (January 30, 2012). Bayesian approaches to handling missing data. BIAS short course.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, *43*(1), 1-20.

Besag, J., & Higdon, D. (1999). Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(4), 691-746.

Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Bodecker, C. F. (1939). The modified dental caries index. *The Journal of the American Dental Association*, *26*(9), 1453-1460.

Bose, R. C. (1947). Mathematical theory of the symmetrical factorial design. *Sankhyā: The Indian Journal of Statistics*, 107-166.

Bose, R. C., & Bush, K. A. (1952). Orthogonal arrays of strength two and three. *The Annals of Mathematical Statistics*, 508-524.

Box, G. E., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters: an introduction to design, data analysis, and model building* (Vol. 1). New York: Wiley.

Broadbent JM, Thomson WM. For debate: problems with the DMF index pertinent to dental caries data analysis. Community Dent Oral Epidemiol. 2005;33(6):400-9.

Brook, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*,*51*(3/4), 481-483.

Carlevaro, F., Croissant, Y., & Hoareau, S. (2009). Multiple Hurdle Tobit Models in R: The mhurdle Package.

Carlin, B. P., & Banerjee, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data.*Bayesian statistics*, *7*, 45-63.

Carpenter, J., Kenward, M., Evans, S., & White, I. (2004). Last observation carry-forward and last observation analysis.*Statistics in medicine*, *23*(20), 3241-3242.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167-174.

Clague, J., Belin, T. R., & Shetty, V. (2017). Mechanisms underlying methamphetamine-related dental disease. *The Journal of the American Dental Association*, *148*(6), 377-386.

Clifford, P. (1990). Markov random fields in statistics. *Disorder in physical systems: A volume in honour of John M. Hammersley*, 19-32.

Cochran, W. G. & Cox, G. M. (1957). *Experimental designs*. Wiley.

Cook, R. J., Zeng, L., & Yi, G. Y. (2004). Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics*, *60*(3), 820-828.

Cox, D. R. (1958). Planning of experiments.

Cressie, N. (1993). Statistics for spatial data. New York: John Wiley.

Darby M.L., Walsh M.M. Dental hygiene: theory and practice (2nd edition). W. B. Saunders Company, USA; 2003.

Demirci M, Tuncer S, Yuceokur AA. Prevalence of caries on individual tooth surfaces and its distribution by age and gender in university clinic patients. Eur J Dent. 2010;4(3):270-9.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

Donaldson, M., & Goodchild, J. H. (2006). Oral health of the methamphetamine abuser. *American Journal of Health-System Pharmacy*, *63*(21).

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, *195*(2), 216-222.

Earnest, A., Morgan, G., Mengersen, K., Ryan, L., Summerhayes, R., & Beard, J. (2007). Evaluating the effect of neighbourhood weight matrices on smoothing properties of Conditional Autoregressive (CAR) models. *International journal of health geographics*, *6*(1), 54.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. CRC press.

Fichman, M., & Cummings, J. N. (2003). Multiple imputation for missing data: Making the most of what you know.*Organizational Research Methods*, *6*(3), 282-308.

Fisher, R.A. (1926). The Arrangement of Field Experiments. Journal of the Ministry of Agriculture of Great Britain, 33: 504-513.

Fisher, R. A. (1935). The Design of Experiments (Hafner, New York).

García-Zattera, M. J., Jara, A., Lesaffre, E., & Declerck, D. (2007). Conditional independence of multivariate binary data with an application in caries research. *Computational statistics & data analysis*, *51*(6), 3223-3234.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, *85*(410), 398-409.

Gelfand, A. E., & Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis.*Biostatistics*, *4*(1), 11-15.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457-472.

Gelman, A. (2004). Parameterization and Bayesian modeling.*Journal of the American Statistical Association*, *99*(466), 537-545.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (Vol. 3). Boca Raton, FL: CRC press.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997-1016.

Geman, S., Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6), 721-741.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association*, *88*(423), 984-993.

Golub, G. H., & Van Loan, C. F. (2012). *Matrix computations*(Vol. 3). JHU Press.

Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. *NIDA research monograph*, *142*, 13-13.

Gustafson DH, McTavish FM, Chih MY, et al. A smartphone application to support recovery from alcoholism: a randomized clinical trial. *JAMA Psychiatry*. 2014;71(5):566-72.

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

Hammersley, J. M., & Clifford, P. (1971). Markov fields on finite graphs and lattices.

Hassan, E. (2005). Recall Bias can be a threat to retrospective and prospective research designs. The internet journal of epidemiology, 3(2).

He, Y., Hodges, J. S., & Carlin, B. P. (2007). Re-considering the variance parameterization in multiple precision models.*Bayesian Analysis*, *2*(3), 529-556.

Hinde, J., Demétrio, C. G. (1998). Overdispersion: models and estimation. *Computational Statistics & Data Analysis*,*27*(2), 151-170.

Hodges J.S., Cui Y., Sargent D.J., Carlin B.P (2007). Smoothing Balanced Single-Error-Term Analysis of Variance. Technometrics; 49(1): 12-25

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.*Journal of Machine Learning Research*, *15*(1), 1593-1623.

Hogan, H.W., Laird, N.M. (1998) Mixture Models for the Joint Distribution of Repeated Measures and Event Times. Statistics in Medicine, 16(3): 239-257.

Horn, R. A., & Johnson, C. R. (1990). *Matrix analysis*. Cambridge university press.

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G., & Mallinckrodt, C. (2006). Analyzing incomplete discrete longitudinal clinical trial data. *Statistical Science*, *21*(1), 52-69.

Joseph M. Exact Sparse CAR Models in Stan. Stan Official Website. August 20, 2016. Available at: http://mc-stan.org/documentation/case-studies/mbjoseph-CARStan.html. Accessed on March 20, 2017.

Kacker, R. N., Lagergren, E. S., & Filliben, J. J. (1991). Taguchi's orthogonal arrays are classical designs of experiments. *Journal of research of the National Institute of Standards and Technology*, *96*(5), 577.

Khodai, N. The universal numbering system assigns and individual numbers to each tooth in your mouth. Barranca Dental, 8 Jan. 2012, http://www.barrancadental.com/tooth-numbers/

Klein, H., Palmer, C. E., & Knutson, J. W. (1938). Studies on dental caries: I. Dental status and dental needs of elementary school children. *Public Health Reports (1896-1970)*, 751-765.

Kishen, K. (1942), "On latin and hyper-graeco cubes and hypercubes", Current Science, 11: 98–99.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kumar, S., Abowd, G.D., Abraham, W.T., al'Absi, M., Beck, J.G, Chau, D.H., Condie, T., Conroy, D.E., Ertin, E., Estrin, D., Ganesan, D., Lam, C., Marlin, B., Marsh, C.B., Murphylnbal, S.A., Patrick, J., Rehg, J.M., Sharmin, M., Shetty, V., Sim, I., Spring, B., Srivastava, M., Wetter, D.W. (2015). Center of Excellence for Mobile Sensor Data-To-Knowledge (MD2K). Journal of the American Medical Informatics Association, 22(6), 1137 – 1142.

Lawson, J. (2010). *Design and Analysis of Experiments with SAS*. CRC Press.

LeSage, J. P., & Pace, R. K. (2004). Models for spatially dependent missing data. *The Journal of Real Estate Finance and Economics*, *29*(2), 233-254.

Linde, A. (2005). DIC in variable selection. *Statistica Neerlandica*, *59*(1), 45-56.

Little, R.J. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. Journal of the American Statistical Association, 83(404): 1198-1202.

Little, R. J., Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, *18*(2-3), 292-326.

Little, R. J. (1992). Regression with missing X's: a review.*Journal of the American Statistical Association*, *87*(420), 1227-1237.

Little, R.J. (1993). Pattern-Mixture Models for Multivariate Incomplete Data. Journal of the American Statistical Association, 88(421), 125 – 134.

Little, R.J.A. & Schenker, N. Missing data. (1995). In G. Arminger, C.C. Clogg & M.E. Sobel [Eds.] *Handbook of statistical modeling for the social and behavioral sciences.* New York: Plenum.

Little, R. J., Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons.

Liu Z, Yu D, Luo W, et al. Impact of oral health behaviors on dental caries in children with intellectual disabilities in Guangzhou, China. Int J Environ Res Public Health. 2014;11(10):11015-27.
Published 2014 Oct 22. doi:10.3390/ijerph111011015

Mason, A.J. (2009). Bayesian Methods for Modelling Non-Random Missing Data Mechanisms in
Logitudinal Studies. Retrieved from http://www.bias-project.org.uk/ Imperial College, London.

Mealli, F., Rubin, D.B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. Biometrika, 102(4): 995-1000.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*,*21*(6), 1087-1092.

Miller, R. G. (1974). The jackknife-a review. *Biometrika*, *61*(1), 1-15.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Neal, R.M. (1993). Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics.*Handbook of Markov Chain Monte Carlo*, *2*(11).

Nesterov, Y. (2009). Primal-dual subgradient methods for convex problems. *Mathematical programming*, *120*(1), 221-259.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*(4), 558-625.

Newcomb, R. W. (1961). On the simultaneous diagonalization of two semi-definite matrices. *Quarterly of Applied Mathematics*, *19*(2), 144-146.

Nie, N. H., Bent, D. H., & Hull, C. H. (1970). *SPSS: Statistical package for the social sciences* (No. HA29 S6). New York: McGraw-Hill.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. DSC Working Papers.

Ott, R. L., Longnecker, M. (2010). Using surveys and experimental studies to gather data. *An introduction to statistical methods and data analysis*, 16-48.

Quanbeck A, Chih MY, Isham A, Gustafson D. Mobile Delivery of Treatment for Alcohol Use Disorders: A Review of the Literature. *Alcohol Res*. 2014;36(1):111-22.

Rao, C.R. (1946), "Hypercubes of strength "d" leading to confounded designs in factorial experiments", Bulletin of the Calcutta Mathematical Society, 38: 67–78.

Reich, B. J., Hodges, J. S., & Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, *62*(4), 1197-1206.

Reich, B. J., Hodges, J. S., & Carlin, B. P. (2007). Spatial analyses of periodontal data using conditionally autoregressive priors having two classes of neighbor relations. *Journal of the American Statistical Association*, *102*(477), 44-55.

Reich, B. J., Bandyopadhyay, D. (2010). A latent factor model for spatial data with informative missingness. *The annals of applied statistics*, *4*(1), 439.

Reich, B. J., Bandyopadhyay, D., & Bondell, H. D. (2013). A nonparametric spatial model for periodontal data with nonrandom missingness. *Journal of the American Statistical Association*, *108*(503), 820-831.

Rosenbaum, P. R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688-701.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*,*63*(3), 581-592.

Rubin, D. B. (1977a). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, *72*(359), 538-543.

Rubin, D. B. (1977b). Assignment to Treatment Group on the Basis of a Covariate. *Journal of educational Statistics*, *2*(1), 1-26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 34-58.

Rubin, D. B. (1996). Multiple imputation after 18+ years.*Journal of the American statistical Association*, *91*(434), 473-489.

Rubin, D. B., Schafer, J. L., & Schenker, N. (1988). Imputation strategies for missing values in post-enumeration surveys.*Survey Methodology*, *14*(1), 2.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Rubin, D.B. (2004). Direct and Indirect Causal Effects via Potential Outcomes. Scandinavian Journal of Statistics, 31(2), 161-170.

Rue, H., & Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

Shetty, V., Mooney, L. J., Zigler, C. M., Belin, T. R., Murphy, D., & Rawson, R. (2010). The relationship between methamphetamine use and increased dental disease. *The Journal of the American Dental Association*, *141*(3), 307-318.

Shetty, V., Harrell, L., Murphy, D. A., Vitero, S., Gutierrez, A., Belin, T. R., ... & Spolsky, V. W. (2015). Dental disease patterns in methamphetamine users: Findings in a large urban sample. *The Journal of the American Dental Association*,*146*(12), 875-885.

Shetty, V., Harrell, L., Clague, J., Murphy, D. A., Dye, B. A., & Belin, T. R. (2016). Methamphetamine users have increased dental disease: a propensity score analysis. *Journal of dental research*, *95*(7), 814-821.

Shievitz, P. The Effect of a Non-Steroidal Anti-Inflammatory Drug on Periodontal Clinical Parameters After Scaling. Unpublished Master's Thesis, University of Minnesota, School of Dentistry. 1997.

Shiffman, S., Stone, A.A., Hufford, M.R. (2008). Ecological Momentary Assessment. Annual Review of Clinical Psychology, 4: 1-32.

Smith, A.F., & Gelfand, A.E. (1990). Sampling-Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association, 85(410): 398-409.

Smith, A. F., Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, *46*(2), 84-88.

Son, H., Friedmann, E., Thomas, SA. (2012). Application of pattern mixture models to address missing data in longitudinal data analysis using SPSS. Nurs Res. 61(3): 195-203.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583-639.

Sun, D., Tsutakawa, R. K., Kim, H., & He, Z. (2000). Bayesian analysis of mortality rates with disease maps. *Statistics in Medicine*, *19*, 2015-2035.

Taguchi, G. (1986). *Introduction to quality engineering: designing quality into products and processes*.

Taguchi, G., Jugulum, R., Taguchi, S. (2004). Computer-Based Robust Engineering: Essemtials for DFSS. ASQ Quality Press: Milwoukee, WI.

Taguchi, G., Taguchi, G. (1987). *System of experimental design; engineering methods to optimize quality and minimize costs* (No. 04; QA279, T3.).

Tanner, M. A., Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528-540.

Thomas, S. A., Friedmann, E., Lee, H. J., Son, H., Morton, P. G., & HAT Investigators. (2011). *Changes in anxiety and depression over 2 years in medically stable patients after myocardial infarction and their spouses in the Home Automatic External Defibrillator Trial (HAT): A longitudinal observational study*. Heart, 97, 371Y381. doi:10.1136/ hrt.2009.184119.

Todem D. Oral Health. In Boslaugh S (ed.), Encyclopedia of epidemiology, vol. 2. Los Angeles: Sage, 762 – 64; 2008.

Todem, D. (2012). Statistical models for dental caries data. In*Contemporary Approach to Dental Caries*. InTech.

Tukey, John (1949). "One degree of freedom for non-additivity". Biometrics. 5 (3): 232-242.

Van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, *18*(6), 681-694.

Van Der Linde, A. (2005). DIC in Variable Selection. Statistica Neerlandica, 59(1), 45-56.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413-1432.

Vehtari A, Gelman A, and Gabry J (2016). "loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models." R package version 1.1.0, https://CRAN.R-project.org/package=loo.

Verbeke, G., Molenberghs, G. (2000). A model for Longitudinal Data. *Linear mixed models for longitudinal data*, 19-29.

Wall, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of statistical planning and inference*, *121*(2), 311-324.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

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

West, M., and Harrison, J. (1997). Bayesian Forecasting and Dynamic Models. Springer.

Yates, F. (1933). The principles of orthogonality and confounding in replicated experiments.(With Seven Text-figures.). *The Journal of Agricultural Science*, *23*(1), 108-145.

Yates, F. (1936). Incomplete randomized blocks. *Annals of Human Genetics*, *7*(2), 121-140.

Yates, F., Hale, R. W. (1939). The analysis of Latin squares when two or more rows, columns, or treatments are missing.*Supplement to the Journal of the Royal Statistical Society*,*6*(1), 67-79.

Zhang Y., Todem D., Kim K., Lesaffre E. Bayesian latent variable models for spatially correlated tooth-level binary data in caries research. Statistical Modelling. 2011; 11(1): 25-47.

Zhang, Y., Pang, Y., Wang, Y. (2001). Orthogonal arrays obtained by generalized hadamard product. Discrete Mathematics, 238: 151-170.

Zorn, C. J. (1998). An analytic and empirical examination of zero-inflated and hurdle Poisson specifications. *Sociological Methods & Research*, *26*(3), 368-400.