Statistical Issues in the Analysis of Correlated Data

by

Rong Xia

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2015

Doctoral Committee:

Professor Thomas M. Braun, Co-Chair
Research Professor Mousumi Banerjee, Co-Chair
Assistant Professor Christopher R. Friese
Associate Professor Brisa N. Sánchez

To my family

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisors Drs. Thomas Braun and Mousumi Banerjee for their immeasurable advice and patience during the course of my dissertation work. Without their generous support, inspiration and encouragement, this dissertation would never have been completed. I am also deeply grateful to Drs. Christopher Friese and Brisa Sánchez for serving on my dissertation committee and providing insightful comments and invaluable suggestions on my research.

Over the course of my graduate study I have been fortunate to work as a graduate student research assistant under the guidance of Dr. Mousumi Banerjee. Her vast knowledge and experience has greatly influenced my statistical skills. I would also like to express my appreciation to our collaborators, Drs. Christopher Friese, Sandra Wong, Michael Sabel, Norah Lynn Henry and Brian Callaghan.

I would like to thank all the faculty members in the Department of Biostatistics at the University of Michigan for their wonderful lectures and brilliant ideas. I would also like to thank all the staffs for their friendly support. My thanks extend to Dr. Kirsten Herold at the School of Public Health Writing Lab for helping me improve my writing skills. My thanks also go to my colleagues who have supported me throughout my graduate study.

Special thanks to my family.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Statistical Issues in the Analysis of Correlated Data

by
Rong Xia

Chair: Thomas M. Braun and Mousumi Banerjee

In the first project, we extend the original classification and regression trees (CART) paradigm [Breiman et al., 1984] to clustered binary outcomes, where individuals within a cluster are correlated. We propose to generate tree models using residuals from a null generalized linear mixed model (with fixed and random intercepts only) as the outcome, which circumvents modeling the correlation structure explicitly while still accounting for the cluster-correlated design, thereby allowing us to adopt the original CART machinery in tree growing, pruning and cross-validation. Based on extensive simulations, we compare our residual-based classification tree to the standard CART that ignores the clustering. We find that our residual-based tree is more appropriate for analyzing clustered binary data, and provides more accurate classification predictions. Our method is also illustrated using data from studies of kidney cancer treatment receipt, surgical mortality after colectomy, and determinant of vaccination coverage. In all studies, residual-based trees identified clinically meaningful subgroups.

The second project is motivated by the analysis of periodontal data. Clinical

attachment level (CAL) is a tooth-level measure that quantifies the severity of periodontal disease. The within-mouth correlation of tooth-level measures of CAL is difficult to model because it must reflect the three-dimensional spatial geography of teeth and their functional similarity. We propose two linear mixed effects (LME) models with random effects that quantify the within-mouth correlation of teeth and their shared functionality. Via simulations, we compare the bias and efficiency of fixed effect estimates computed with our models to corresponding results produced with a t-test and generalized estimating equations. We demonstrate that our mixed models give estimates that are consistent and more efficient than other methods that fail to model the within-mouth correlation of teeth accurately. We also evaluate the performance of the approaches when data are missing under different biologically plausible missing data mechanisms.

Inference for the fixed effects in an LME model is dependent upon the correlation structure implied by the random effects included in the LME model. However, limited methods are available for making inference about the fit of the assumed covariance structure in the LME model. In the third project, we propose three permutation tests, all of which are based on comparing the estimated assumed covariance matrix to the covariance matrix of the marginal residuals. Cholesky residuals, which are exchangeable both within and among subjects, are employed in the permutations. Through simulations, we show that two of our tests have valid size and comparable power in testing different covariance structure assumptions. We also apply our tests to data collected from the periodontal disease study that motivated the methods in our second project.

CHAPTER I

Introduction

Correlated data is abundant in biomedical studies. For example, in longitudinal studies, measurements on a subject are collected repeatedly over time, and thereby are correlated within-subject through shared subject-specific characteristics. Such correlation is commonly known as serial correlation. In multi-level or clustered data, subjects are nested within clusters, and subjects within the same level or cluster are generally more similar to each other than subjects from different clusters. This clustering effect induces intra-cluster correlation among subjects within the same cluster, e.g., in familial segregation studies, family members are usually similar as they share genetic factors. Thus, the assumption of independence, which is common for most standard statistical methods, i.e., ordinary linear models, is violated in correlated data. Therefore, in order to draw valid statistical inferences, special methods are required for analyzing correlated data, as it is necessary to account for the correlation. Ignoring this correlation could inflate the variance estimates.

Mixed effects models are a rich family of models containing both fixed and random effects, which are widely adopted in analyzing correlated data. The fixed effects coefficients play the same role as the coefficients in an ordinary linear model, and are interpreted as estimates of the average population effect. The random effects are

subject-specific, which represent an aggregation of factors that make measurements on the same subject intrinsically similar, and are often assumed to follow certain distributions. The usage of random effects and/or random errors creates a flexible class of covariance structures that allows us to account for and take advantage of the structured patterns in the correlated data. Two popular mixed effects models are the linear mixed effects (LME) models for normally distributed outcomes [Laird and Ware, 1982], and the generalized linear mixed models (GLMMs) for non-normal outcomes (e.g., binary, count etc.) [Breslow and Clayton, 1993]. In both models, parameter estimations involve likelihood based methods implemented with the expectation-maximization (EM) algorithm. These likelihood methods rely on the distribution of the outcomes, and therefore require us to model the covariance structure accurately, in order to obtain consistent and unbiased estimates.

An alternative to mixed effects models is the method of generalized estimating equations (GEEs), where the population-averaged effects are estimated by solving estimating equations [Liang and Zeger, 1986]. The estimating equations are based on the moments of the outcomes, rather than the full distribution. Therefore, when GEEs are used in conjunction with the "sandwich" estimator, the resulting population effects variance estimates are robust to misspecified covariance structures, as long as the mean structure is specified correctly. However, GEEs require stronger assumptions than mixed effects models on missing data and are less efficient than the estimator which uses the correct covariance model.

Tree-based methods have become one of the most flexible, intuitive, and powerful data analytic tools for exploring complex data. The arguably most widely used tree model is the Classification and Regression Trees (CART) introduced by Breiman et al. [1984]. In the CART paradigm, the covariate space is recursively partitioned

into disjoint regions and the corresponding data is split into groups (nodes), with the intent of increasing within-node homogeneity in the response distribution. The final resulting tree can be represented as a binary tree, and its terminal nodes represent subgroups characterized by common covariate values and homogeneous outcomes. However, the standard CART is not designed for handling clustered data, where subjects within a cluster are usually correlated, and accounting for the clustering effect could potentially improve the validity of statistical analysis.

Several researchers have studied extensions of the standard CART to clustered data. One type of approaches is to generalize the univariate impurity function to multivariate outcomes [Segal, 1992, Zhang, 1998]; However, these methods allow splits to only be based upon cluster-level covariates, and splits on subject-level covariates are prohibited. Another type of extension is to combine a regular tree model with cluster-level random effects to grow a mixed effects tree [Hajjem et al., 2011, Sela and Simonoff, 2012]; However, software implementations of such methods to date have been limited.

In sight of these limitations, we propose a new method to extend the standard CART to clustered binary outcomes in Chapter II. Our method is based on using residuals from a null generalized linear mixed model, which only contains fixed intercept and cluster-level random intercepts, as the outcome to partition the covariate space into rectangles. This circumvents modeling the correlation structure explicitly while still accounting for the cluster-correlated design, thereby allowing us to adopt the original CART machinery in tree growing, pruning and cross-validation. We compare our residual-based tree to the standard CART via a series of simulations. We also illustrate our method using data from studies of kidney cancer treatment receipt, surgical mortality after colectomy, and determinants of vaccination coverage

in Uttar Pradesh, India.

Chapter III is motivated by a periodontal disease study, conducted at the the Michigan Center for Oral Health Research [Ramseier et al., 2009, Kinney et al., 2011]. Periodontal disease is a chronic inflammatory disorder that affects the gingiva, the supporting connective tissue and the alveolar bone, all of which anchor the teeth in the jaws. Periodontal disease is the most common cause of tooth loss in adults in the United States, and it has a prevalence around 50% [Eke et al., 2012]. Diagnosis of periodontal disease often involves the evaluation of periodontal outcomes such as clinical attachment loss (CAL), which measures the extent to which the gingiva has lost its attachment to a tooth. The difficulty in modeling periodontal outcomes lies in the fact that teeth from the same subject are usually correlated due to their three-dimensional spatial geography, functional similarity, and the natural symmetry of mouth. Traditionally, researchers have analyzed periodontal outcomes using ordinary linear regression, t-test or generalized estimating equations (GEE). However, these methods have low efficiency because they fail to model the within mouth correlation accurately. Recently, Reich et al. [2007] , Reich and Hodges [2008] and Reich et al. [2013] have proposed to use conditionally auto-regressive (CAR) models for periodontal outcomes, which take account of the spatial correlations by smoothing over neighboring teeth. However, these methods require complicated statistical programming that does not yet exist in standard statistical packages, so that these methods have not been widely adopted in periodontal studies.

Thus in Chapter III, we develop two linear mixed effects (LME) models for periodontal outcomes, where we use random effects and random errors to quantify the complex within-mouth correlation of teeth. Our intention is to create accurate models that are easily accessible to periodontal researchers, therefore can be widely

used in periodontal studies. We also discuss criteria for the selection between these two models. Via simulations, we compare the estimates from our models to the corresponding results of t-tests and GEEs. We further evaluate the performance of our models when data are missing under different biologically plausible missing data mechanisms. Finally we apply our methods to the Michigan periodontal data and estimate the mean difference in CAL values between periodontal diseased and healthy subjects.

When we applied the two LME models to the Michigan periodontal data, one remaining challenge was to assess their goodness of fit. As mentioned earlier, inference for the fixed effects in an LME model depends upon the correlation structure implied by the model. It is important to appropriately model the true covariance structure, otherwise the variance of fixed effects estimates may be biased. However, diagnostic methods for evaluating the fit of the assumed covariance structure in an LME model remain underdeveloped. To the best of our knowledge, the quantile-quantile (Q-Q) plots of Cholesky residuals proposed by Houseman et al. [2004] and Jacqmin-Gadda et al. [2007], and an informal check recommended by Verbeke and Molenberghs [2009] are the only approaches available for evaluating the overall covariance assumption directly. However, these two methods do not provide any formal statistical inferences. Other methods include successively testing for the inclusion or exclusion of each possible random effect, which is difficult because the variance component is on the boundary of the parameter space under the null hypothesis. To overcome these issues, in Chapter IV we propose three permutation tests employing test statistics that quantify the difference between the estimated assumed covariance of the LME model and the smoothed sample covariance of the marginal residuals. The empirical null distributions of our test statistics are generated by permuting

the Cholesky residuals both within and among subjects. Through simulations, we show that two of our tests have valid size and comparable power in testing different covariance structure assumptions. We also apply our permutation tests to Michigan periodontal study and evaluate the fit of the two proposed LME models.

CHAPTER II

Residual-Based Tree for Clustered Binary Data

2.1 Introduction

Tree-based methods have become one of the most flexible, intuitive, and powerful data analytic tools for exploring complex data structures. The applications of these methods are far reaching. The best documented, and arguably most popular uses of these methods are in health sciences research where classification is a central issue. Tree-based methods partition the covariate space into a set of rectangles, leading to a fitted model that is piecewise constant over regions of the covariate space. Some interesting applications of tree-based methods in the health sciences literature are described by Zhang and Singer [1999], Banerjee et al. [2004] and Segal et al. [2004].

Tree-based methods were originally introduced by Morgan and Sonquist [1963], and further advanced by Breiman et al. [1984] in their monograph on Classification and Regression Trees (CART). In the CART paradigm, the covariate space is recursively partitioned into disjoint regions and the corresponding data is split into groups (nodes). The partitions are intended to increase within-node homogeneity in the response distribution. For each node, extent of homogeneity is measured quantitatively using an impurity function, e.g., residual sum of squares for continuous outcomes, and Gini or entropy for binary outcomes. At each stage of the splitting process, a

parent node gives rise to two daughter nodes (binary partitioning). Goodness of a split is assessed by the reduction in impurity going from a parent node to the two daughter nodes. All possible splits for each covariate are evaluated, and the covariate with the corresponding split point that results in the maximum impurity reduction is chosen. This splitting procedure is applied recursively until each node is pure in response or only contains a few subjects. After a large tree is grown, there are rules for pruning and readjusting the size of the tree. The final result can be represented as a binary tree, and its terminal nodes represent subgroups characterized by common covariate values and homogeneous outcomes.

Clustered, or more specifically cluster-correlated data arise when there is nested structure to the data. Data of this sort frequently arise in the social, behavioral, and health sciences since individuals can be grouped in many different ways. For example, in studies of health services and outcomes, assessments of quality of care are often obtained from patients who are nested within physicians and/or hospitals [Miller et al., 2008, Haymart et al., 2011]. Such data are also referred to as hierarchical/multilevel, with patients referred to as level 1 units and physicians/hospitals as level 2 units. The clustering induces correlation among units within the same cluster, and this intra-cluster correlation has to be accounted for in order to obtain valid statistical inferences.

Some authors have studied extensions of original tree-based methods to multilevel data. Segal [1992] developed trees for multilevel continuous outcomes using weighted residual sum of squares as a measure of impurity, where the weights were based on the estimated covariance matrix of some simple variance models. Abdolell et al. [2002] used the likelihood ratio statistic for evaluating splits. Hajjem et al. [2011] and Sela and Simonoff [2012] independently developed mixed effects tree models by

combining a regular tree with cluster-level random effects. For multilevel binary outcome, Zhang [1998] developed classification trees using generalized entropy and Gini criteria for splitting. Keon Lee [2005] proposed building multivariate decision trees that employed generalized estimating equations. However, all of these methods suffer from one or several drawbacks, namely 1) can handle only cluster-level covariates, such that subjects within a cluster always end up in the same node of the tree; 2) require the clusters/groups to be balanced in size; and 3) do not have available software implementations, thereby limiting their applicability.

In this chapter, we develop a methodology for growing trees in the setting of cluster-correlated binary data. As opposed to the conventional CART, our approach uses the residuals from a null generalized linear mixed model as the outcome to partition the covariate space into rectangles. This circumvents modeling the correlation structure explicitly while still accounting for the cluster-correlated design, thereby allowing us to adopt the original CART machinery in tree growing, pruning and cross-validation. Our proposed method is flexible at handling both individual- and cluster- level covariates, does not require balance in cluster sizes, and can be easily implemented in any standard software for binary recursive partitioning. Furthermore, our method lends itself to a natural extension to an ensemble of trees, that can often give more accurate predictions and address instability in a single tree.

This chapter is organized as follows. In Section 2.2, we introduce the methodology for growing trees for clustered binary outcome using residuals from a null generalized linear mixed model. Section 2.3 compares our residual-based trees to the standard classification trees via simulations. We illustrate our methodology in Section 2.4 using data from a health services research study to investigate determinant of kidney cancer treatment receipt. Section 2.5 applies our methodology to study the surgical

mortality after receiving colectomy surgery. Section 2.6 applies our methods to study the vaccination coverage in Uttar Pradesh, India. Finally, Section 2.7 contains some concluding remarks.

2.2 Residual-Based Tree for Clustered Data

For clustered data, individuals within the same cluster are usually correlated, e.g., in familial segregation studies, family members are usually alike as they share the same genetic factors; in clinical studies, patients treated by the same provider are usually more similar in terms of treatment received. Popularized by Breslow and Clayton [1993], generalized linear mixed effects models (GLMMs) have become a standard framework for modeling clustered non-normal data, where the inclusion of cluster-specific random effects induces correlation among individuals within the same cluster. Consider a two-level hierarchical data structure: let y_{ij} be the binary response of the jth individual (level-one unit) in the ith cluster (level-two unit), where 0 stands for 'failure' and 1 stands for 'success', $i = 1, ..., m$, $j = 1, ..., n_i$, $N = \sum_{i=1}^{m} n_i$. The GLMM with logit link can be written as

$$g(\mu_{ij}) = log(\frac{\mu_{ij}}{1-\mu_{ij}}) = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$ are the population level fixed effects coefficients, and $\boldsymbol{b}_i = (b_{i0}, ..., b_{iq})'$ are the random effects for cluster i. The $\boldsymbol{x}_{ij} = (1, x_{ij1}, ..., x_{ijp})'$ and $\boldsymbol{z}_{ij} = (1, z_{ij1}, ..., z_{ijq})'$ are the fixed effects covariates and random effects covariates, respectively, for the jth individual in cluster i. The random effects $\boldsymbol{b}_i$ are assumed to follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The $\mu_{ij} = E(y_{ij}|\boldsymbol{b}_i) = P(y_{ij} = 1|\boldsymbol{b}_i)$ is the conditional expectation of y_{ij} given random effects $\boldsymbol{b}_i$. Given random effects $\boldsymbol{b}_i$, all n_i individuals y_{ij} from cluster i are conditionally independent.

Parameter estimation in GLMMs typically involves maximum likelihood (ML) or variants of ML. In order to obtain the likelihood, integration over the random effects must be evaluated. In general this integration can not be done analytically, instead, numerical algorithm such as penalized quasi-likelihood (PQL) [Breslow and Clayton, 1993] is often employed. The random effects $\boldsymbol{b}_i$ are estimated using empirical Bayes method.

2.2.1 Residuals from Null Model

We propose to fit a null GLMM with only one fixed effect β_0, which is the population level intercept, and one random effect b_{i0}, which is the cluster-level intercept that represents the effect of cluster i. The prediction from this null model is $\hat{\mu}_{ij} = g^{-1}(\hat{\beta}_0 + \hat{b}_{i0})$, where $g^{-1}(.)$ is the inverse of the logit link. It is easily seen from this model that all n_i individuals from cluster i have the same predicted value, which is the estimated cluster-level success probability for cluster i after accounting for the hierarchical structure.

Two types of residuals are commonly used for binary responses: the Pearson residual and the deviance residual. Given $\hat{\mu}_{ij}$, the prediction for individual j in cluster i from the null GLMM, the Pearson residual pr_{ij} can be defined as

$$pr_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - \hat{\mu}_{ij})}}. \tag{2.2}$$

The deviance residual dr_{ij} is defined as

$$dr_{ij} = sign(y_{ij} - \hat{\mu}_{ij})\sqrt{2y_{ij}log(\frac{y_{ij}}{\hat{\mu}_{ij}}) + 2(1 - y_{ij})log(\frac{1 - y_{ij}}{1 - \hat{\mu}_{ij}})}, \tag{2.3}$$

where $sign(y_{ij} - \hat{\mu}_{ij})$ is the sign of $y_{ij} - \hat{\mu}_{ij}$.

2.2.2 Residual-Based Tree

Therneau et al. [1990] had advocated using null martingale residuals from a Cox

proportional hazards model as the responses to grow tree models for survival data. Following their approach, we propose to grow a regression tree using residuals from the null GLMM as our new responses. This circumvents the complexity induced by clustering while still accounting for the correlation structure.

In the CART paradigm, the covariate space is recursively partitioned into disjoint rectangular regions and the data is divided into subgroups (nodes). Each terminal node represents a rectangular region. Here we denote the rectangular region formed by terminal node l as R_l. Suppose tree T has L terminal nodes, then we can envision this tree as an additive model f with L terms, where each term corresponds to a terminal node

$$f(\boldsymbol{x}_{ij}) = \sum_{l=1}^{L} a_l I(\boldsymbol{x}_{ij} \in R_l), \tag{2.4}$$

and a_l is defined as the prediction for all observations fall in terminal node l. Fitting a tree involves greedy searching for the optimized combination of L, R_l's and a_l's.

For clustered binary outcomes, we propose to grow a regression tree using residuals as the outcome variable (i.e. in a transformed scale). At each stage, we search for the best split that maximizes the node impurity, which is measured as the residual sum of squares based on the transformed outcome, i.e., $\sum_{ij \in node}(r_{ij} - \bar{r})^2$. After a large tree is grown, we prune it using cost-complexity pruning. Therefore, L and R_l's are both obtained by optimizing the within-node homogeneity based on the transformed outcome. Once the tree architecture has been selected, class prediction for the $l-th$ terminal node in the original outcome scale is obtained by estimating the proportion of subjects with success or failure in that terminal node, i.e.,

$$\hat{a}_l = \frac{\sum_{i,j} y_{ij} I(\boldsymbol{x}_{ij} \in R_l)}{\sum_{i,j} I(\boldsymbol{x}_{ij} \in R_l)}.$$

The method does not rely on balanced clusters, and can be applied to unbalanced

data with varying cluster sizes. Furthermore, it is possible to choose splits based on both individual- and cluster- level covariates.

2.2.3 Software Implementation

Our residual-based tree algorithm can be easily implemented in the R system. The null GLMM is fitted using the "glmmPQL" function from the "MASS" package, which utilizes the Penalized Quasi-Likelihood algorithm. (As a sensitivity analysis, we have also tried other model fitting algorithms such as Laplace approximation or Gauss-Hermite quadrature.) The regression tree with residuals as the new responses is grown with the standard "rpart" package. We create an R function that extracts the architecture of the regression tree and use it towards terminal node class prediction of the original binary outcomes.

2.3 Simulation Studies

In this section, we compared our residual-based tree to the standard classification tree via simulations. Our comparisons focused on the architectures of the trees as well as prediction performance.

2.3.1 Simulation Design

We generated data from a two-level hierarchical design with 75 clusters and 10 (50 or 100) individuals per cluster. The binary responses y_{ij} were generated from a two-level random intercept model via a latent variable formulation

$$y_{ij}^* = log(\frac{p(\boldsymbol{x}_{ij})}{1-p(\boldsymbol{x}_{ij})}) + b_i + \epsilon_{ij}, \tag{2.5}$$

where the random intercept b_i was generated from normal distribution $N(0, \sigma_b^2)$, and the level-one error ϵ_{ij} was generated from a logistic distribution with mean 0 and variance $\pi^2/3$. We defined $y_{ij} = 1$ if $y_{ij}^* > 0$, and $y_{ij} = 0$ otherwise.

Eight independent covariates (X_1 to X_8) were generated 1) from two distributions, 2) at the cluster- and individual- levels, 3) and divided to signal or noise components. Covariates X_1, X_2, X_5 and X_6 followed standard normal distributions, while X_3, X_4, X_7 and X_8 had Bernoulli distributions with mean 0.5. The X_1, X_3, X_5 and X_7 were individual-level covariates, while the others were at the cluster-level. Covariates X_1 to X_4 contributed to the responses and the rest were noises.

The fixed effect $p(\boldsymbol{x}_{ij})$ depended on covariates X_1 to X_4 via an underlying tree illustrated in Figure 2.1, where $(p_1, p_2, p_3, p_4, p_5)' = (0.9, 0.4, 0.8, 0.3, 0.7)'$ were the marginal probabilities of success in terminal nodes:

Terminal Node I: If $x_{ij3} = 0$ and $x_{ij1} \leq 0$, $p(\boldsymbol{x}_{ij}) = p_1$;

Terminal Node II: If $x_{ij3} = 0$ and $x_{ij1} > 0$ and $x_{ij4} = 0$, $p(\boldsymbol{x}_{ij}) = p_2$;

Terminal Node III: If $x_{ij3} = 0$ and $x_{ij1} > 0$ and $x_{ij4} = 1$, $p(\boldsymbol{x}_{ij}) = p_3$;

Terminal Node IV: If $x_{ij3} = 1$ and $x_{ij2} \leq 0$, $p(\boldsymbol{x}_{ij}) = p_4$;

Terminal Node V: If $x_{ij3} = 1$ and $x_{ij2} > 0$, $p(\boldsymbol{x}_{ij}) = p_5$.

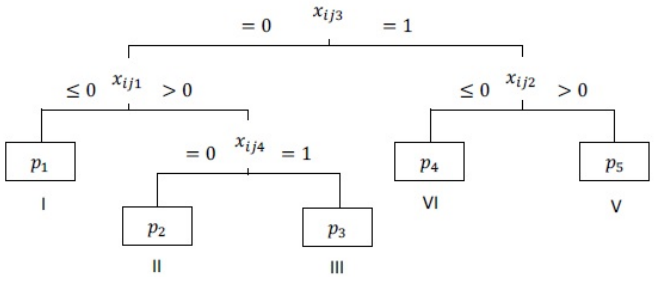

Figure 2.1: Underlying tree structure of the two-level random intercept model used for simulating data, where $(p_1, p_2, p_3, p_4, p_5)'$ are the marginal probabilities of success in terminal nodes.

Under this simulation design, individuals in different clusters were independent while individuals within the same cluster were equally correlated. The strength of correlation between individuals belonging to the same cluster could be expressed by the intra-cluster correlation coefficient (ICC), which was defined as the ratio of

between-cluster variance to total variance, i.e.,

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + (\pi^2/3)}.$$

To study the effect of different within cluster correlations, we varied σ_b^2 between 0, 0.5^2, 1^2 and 2^2, which corresponded to estimated ICC of 0, 0.07, 0.23 and 0.55.

The simulated data were randomly divided into a training set and a testing set. The training set contained 50 clusters and it was used for building the trees. The remaining 25 clusters were employed as the testing set for evaluating predictions. The simulations were repeated for $1,000$ times.

2.3.2 Simulation Results

We compared how similar our residual-based tree or the standard classification tree was to the true underlying tree architecture. Two simple choices of similarity metric were tree size, i.e., the number of terminal nodes, and the number of times each covariate was split in the tree. In addition, we argued that two trees are similar if they place the same individuals together in a terminal node and separate the same individuals in different terminal nodes (i.e., if individuals g and h were placed in two different terminal nodes by Tree A, then these two individuals should also be placed in two different terminal nodes by Tree B for it to be similar to Tree A). As introduced in Banerjee et al. [2012], we employed a metric d to quantify how individuals were clustered in the terminal nodes. For all $\binom{N}{2}$ pairs of individuals, if individual g and h were in the same terminal node by tree T, then $I_T(g,h) = 1$, otherwise $I_T(g,h) = 0$. The difference of terminal nodes clustering between the fitted tree T_1 and the underlying true tree T_0 was then measured as

$$d(T_0, T_1) = \frac{\sum_{g>h} \sum_h |I_{T_0}(g,h) - I_{T_1}(g,h)|}{\binom{N}{2}}, \tag{2.6}$$

where the factor $\binom{N}{2}$ scaled this metric to range from 0 to 1 such that $d = 0$ when the terminal nodes of the fitted tree T_1 were exactly the same as the terminal nodes of the underlying true tree T_0, and $d = 1$ when they were completely different.

Lastly, we compared the residual-based trees and the standard classification tree in terms of prediction accuracy based on the c-statistic obtained using the testing set data.

Table 2.1 contain the averages of the above mentioned statistics over $1,000$ simulations, under different combinations of cluster size (n) and random effect (σ_b^2). To examine the variations over different repeats, we show in Figure 2.2-2.4 the boxplots of fitted tree sizes, terminal nodes clustering metrics d, and C-statistics on the testing set. Within each figure, the top, middle and bottom panel is under cluster size 10, 50 and 100, respectively.

When intra-cluster correlation is none or small, e.g., ($ICC = 0$ or 0.07), the architectures of the fitted standard classification tree (RPART), Pearson residual-based tree (PR) and Deviance residual-based tree (DR) are all similar to the underlying true tree, for cluster sizes 50 and 100. The average fitted tree sizes are all around 5; the four signal covariates X_1 to X_4 are correctly selected for splitting, and each covariate is split once on average; the average terminal nodes clustering metric d are all near 0. The prediction performances of the three fitted trees are also similar based on their c-statistics. The boxplots further indicate that these statistics have little variations over the 1000 simulations.

When intra-cluster correlation is strong, e.g., ($ICC = 0.23$ or 0.55), RPART fits overly complicated trees, with average tree sizes much larger than 5. This is primarily because RPART fails to discriminate between signal and noise variables, and frequently splits on the noise variables X_5, X_6 and X_8. Furthermore, it exhibits a

Table 2.1: Average of statistics over the 1,000 simulations (RPART: standard classification tree; Pearson: Pearson residual-based tree; Deviance: Deviance residual-based tree).

Cluster Size	σ_b^2	ICC	Tree Type	Tree Size	Selection Frequency								d	C-statistic on Testing Set Data
					X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8		
10	0	0.00	RPART	4.0	0.5	1.1	1.0	0.4	0.1	0.0	0.0	0.0	0.156	0.703
			PR	4.1	0.8	0.9	1.0	0.4	0.0	0.0	0.0	0.0	0.093	0.726
			DR	4.2	0.8	0.9	1.0	0.4	0.0	0.0	0.0	0.0	0.088	0.731
	0.5^2	0.07	RPART	4.2	0.5	1.1	1.0	0.4	0.1	0.1	0.0	0.0	0.173	0.686
			PR	3.8	0.8	0.7	1.0	0.3	0.0	0.0	0.0	0.0	0.123	0.701
			DR	4.0	0.8	0.8	1.0	0.4	0.0	0.0	0.0	0.0	0.107	0.710
	1^2	0.23	RPART	4.8	0.6	1.3	1.0	0.3	0.2	0.3	0.0	0.0	0.205	0.657
			PR	3.2	0.7	0.3	0.9	0.2	0.0	0.0	0.0	0.0	0.201	0.654
			DR	3.5	0.8	0.5	1.0	0.2	0.0	0.0	0.0	0.0	0.157	0.671
	2^2	0.55	RPART	10.7	1.3	3.4	0.9	0.5	0.7	2.7	0.1	0.3	0.247	0.574
			PR	2.0	0.4	0.1	0.6	0.0	0.0	0.0	0.0	0.0	0.443	0.566
			DR	2.5	0.5	0.2	0.7	0.1	0.0	0.0	0.0	0.0	0.334	0.588
50	0	0.00	RPART	5.2	1.0	1.0	1.1	1.0	0.0	0.0	0.0	0.0	0.030	0.769
			PR	5.0	1.0	1.0	1.0	0.9	0.0	0.0	0.0	0.0	0.010	0.773
			DR	5.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.006	0.774
	0.5^2	0.07	RPART	5.2	1.0	1.1	1.1	1.0	0.0	0.1	0.0	0.0	0.048	0.752
			PR	4.9	1.0	1.0	1.0	0.9	0.0	0.0	0.0	0.0	0.016	0.759
			DR	5.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.007	0.764
	1^2	0.23	RPART	11.2	1.5	3.4	1.1	1.2	0.4	2.3	0.0	0.2	0.100	0.709
			PR	4.6	1.0	0.7	1.0	0.8	0.0	0.0	0.0	0.0	0.049	0.721
			DR	4.9	1.0	1.0	1.0	0.8	0.0	0.0	0.0	0.0	0.020	0.734
	2^2	0.55	RPART	27.8	2.5	10.7	1.6	1.5	0.9	8.6	0.1	0.9	0.192	0.586
			PR	3.2	0.9	0.1	1.0	0.1	0.0	0.0	0.0	0.0	0.154	0.629
			DR	4.4	1.0	0.8	1.0	0.6	0.0	0.0	0.0	0.0	0.068	0.664
100	0	0.00	RPART	5.2	1.1	1.0	1.1	1.1	0.0	0.0	0.0	0.0	0.015	0.774
			PR	5.3	1.0	1.2	1.0	1.0	0.0	0.0	0.0	0.0	0.008	0.776
			DR	5.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.002	0.776
	0.5^2	0.07	RPART	5.6	1.1	1.2	1.1	1.1	0.0	0.1	0.0	0.0	0.029	0.761
			PR	5.3	1.0	1.3	1.0	1.0	0.0	0.0	0.0	0.0	0.009	0.766
			DR	5.1	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.004	0.766
	1^2	0.23	RPART	16.7	1.7	5.8	1.3	1.5	0.5	4.4	0.0	0.5	0.108	0.713
			PR	5.1	1.0	1.1	1.0	1.0	0.0	0.0	0.0	0.0	0.013	0.740
			DR	5.1	1.0	1.1	1.0	1.0	0.0	0.0	0.0	0.0	0.008	0.742
	2^2	0.55	RPART	33.3	2.6	13.7	2.0	1.7	0.8	10.3	0.1	1.2	0.196	0.585
			PR	3.7	1.0	0.3	1.0	0.4	0.0	0.0	0.0	0.0	0.116	0.645
			DR	5.6	1.0	1.4	1.0	1.0	0.0	0.1	0.0	0.0	0.037	0.676

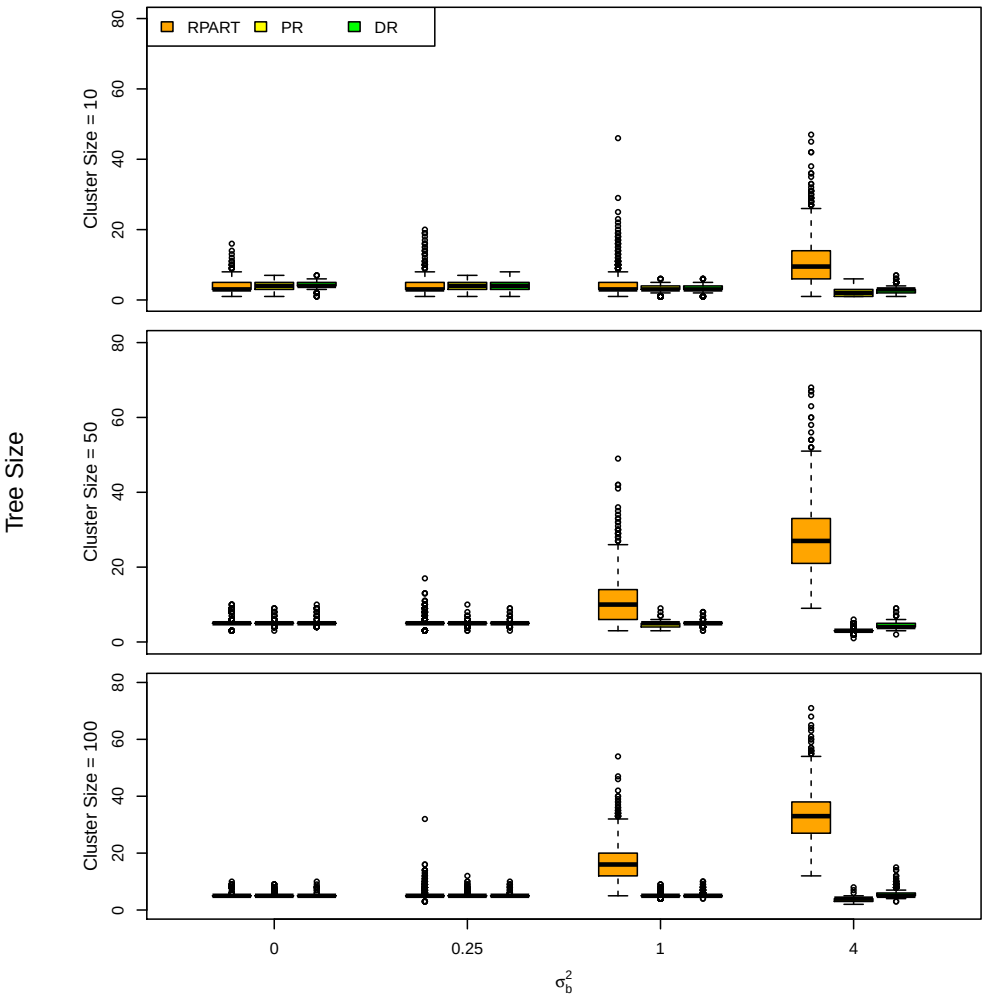

Figure 2.2: Boxplot of tree sizes over different random intercept variance σ_b^2. The top panel is under cluster size 10, the middle panel is under cluster size 50, and the bottom panel is under cluster size 100 (RPART: standard classification tree; PR: Pearson residual-based tree; DR: Deviance residual-based tree).

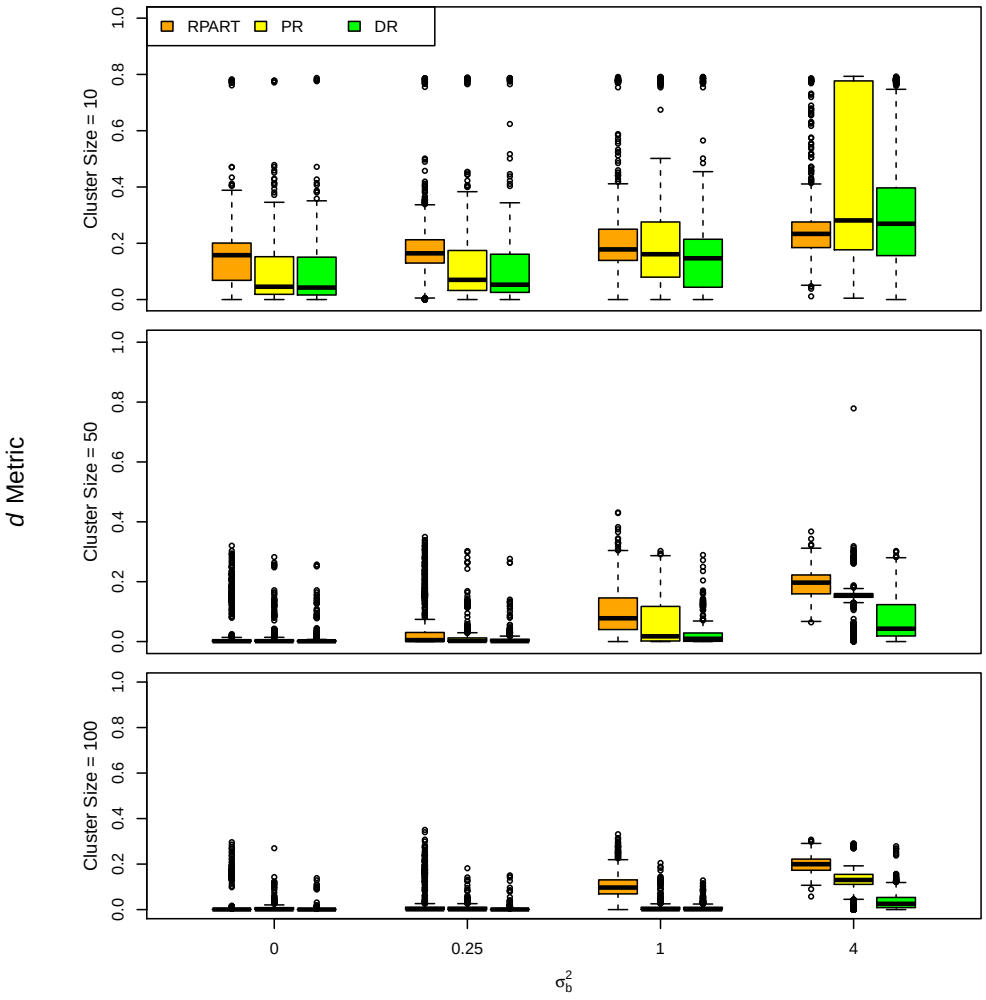

Figure 2.3: Boxplot of terminal nodes clusters metrics d over different random intercept variance σ_b^2. The top panel is under cluster size 10, the middle panel is under cluster size 50, and the bottom panel is under cluster size 100 (RPART: standard classification tree; PR: Pearson residual-based tree; DR: Deviance residual-based tree).

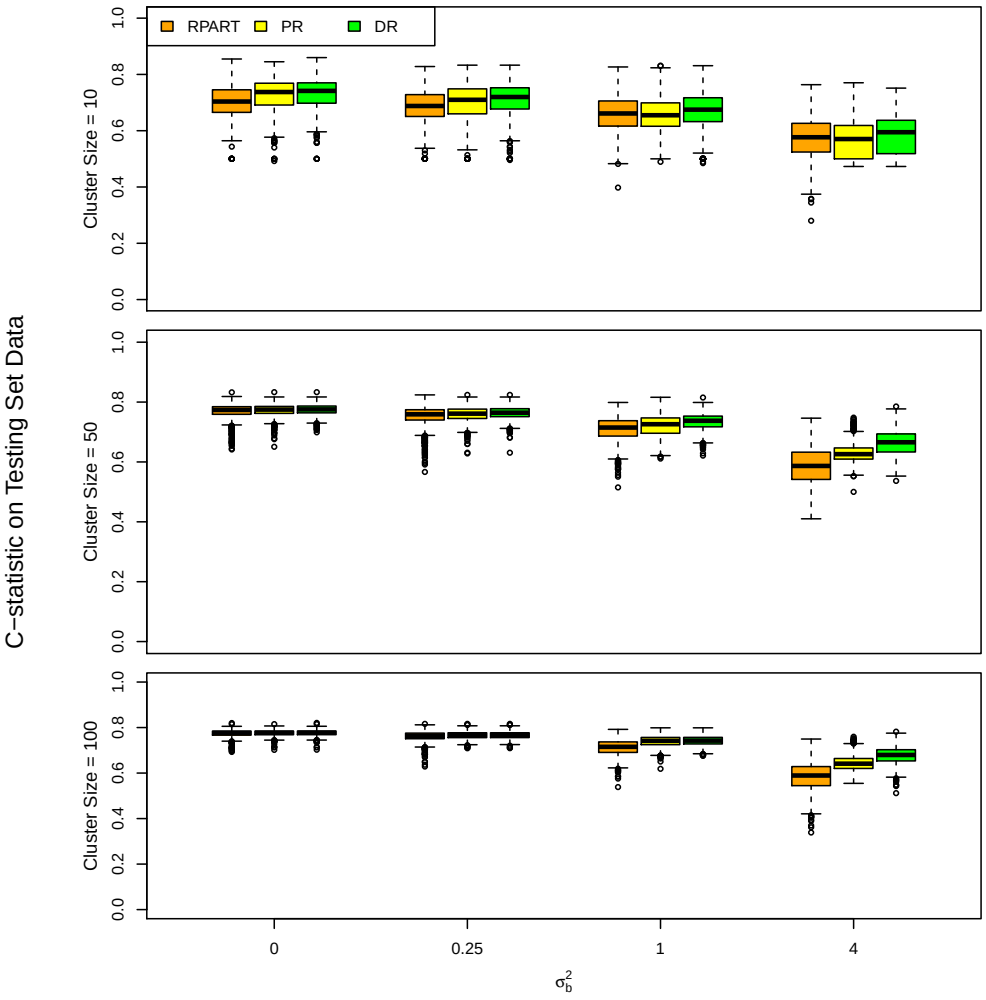

Figure 2.4: Boxplot of C-statistics on the testing set data over different random intercept variance σ_b^2. The top panel is under cluster size 10, the middle panel is under cluster size 50, and the bottom panel is under cluster size 100 (RPART: standard classification tree; PR: Pearson residual-based tree; DR: Deviance residual-based tree).

propensity to over-split individual-level continuous coavariates. This finding agrees with the well-known selection bias issue of RPART [Hastie et al., 2001]. In contrast, PR and DR, particularly DR, fit trees with sizes close to that of the true underlying tree. Both PR and DR are unaffected by noise variables, and in general, make correct splits on the signal covariates. The standard classification tree also exhibits variation in tree sizes over the 1000 simulations. The average of the metric d across the 1000 simulations is consistently smallest for DR, indicating that the DR tree is most similar to the true underlying tree in terms of terminal node clustering. Furthermore, the c-statistic of the DR tree is consistently larger than the PR or RPART tree, demonstrating the former's superior prediction performance. For small clusters (size= 10), however, both PR and DR trees are simpler than the true underlying tree, especially when $ICC = 0.55$. This is possibly due to the biased empirical Bayes estimation of the cluster effect ($\hat{b}_i$) when the cluster size is small [Skrondal and Rabe-Hesketh, 2009].

In summary, based on the simulations, we conclude that for clustered data the residual-based trees are superior to the standard classification tree. In particular, the deviance residual-based tree can better identify the true underlying structure of the data, and provide more accurate predictions. These improvements are substantial when the intra-cluster correlations are strong, given the cluster sizes are moderate to large.

2.4 Application to Kidney Cancer Treatment Receipt Study

To illustrate our method, we present an analysis of data from a population-based study of kidney cancer where the outcome of interest is (binary) receipt of treatment. Radical nephrectomy is the traditional gold standard for treating patients

with organ-confined kidney cancer. During the last two decades, however, the introduction of a nephron-sparing alternative (i.e., partial nephrectomy) to radical excision has appreciably modified the therapeutic options for patients with kidney cancer. Partial nephrectomy yields oncologic outcomes that are indistinguishable from those achieved by radical excision and also preserves long-term renal function while reducing overtreatment of patients with benign tumors. Despite these potential benefits, population-based data suggest that the adoption of partial nephrectomy has been slow, and radical nephrectomy remains the predominant surgical therapy for patients with kidney cancer [Hollenbeck et al., 2006, Banerjee et al., 2014]. The goal of our study was to apply the residual-based trees to understand the pattern of utilization partial nephrectomy in the population.

Our analysis cohort comprised of $11,136$ Medicare beneficiaries treated by $2,031$ urologists for kidney cancer diagnosed between year 1995 and 2006. This data set exhibited a two-level hierarchical structure with patients nested within surgeons. The median number of patients treated by a surgeon was 4. The outcome of interest was receipt of partial versus radical nephrectomy (i.e., binary outcome). Among the $11,136$ patients, $1,667$ underwent partial nephrectomy. A total of sixteen covariates were considered for analysis, which included eight patient characteristics such as socio-demographic variables (age, year of surgery, race/ethnicity, gender, marital status and socioeconomic status), tumor size and the number of preexisting comorbid conditions (using a modification of the Charlson index based on claims submitted during the 12 months before kidney cancer surgery). On the basis of standard clinical guidelines, we categorized tumor size as ≤ 4 cm, $4.1 - 7$ cm and > 7 cm. We also considered eight surgeon-level covariates including a surgeon's age, gender, year of medical school graduation, practice size (solo or two-person, group practice, HMO

or hospital-based, medical school, or other/unclassified), practice location (rural vs. urban), academic affiliation (major, minor, or no academic affiliation), surgeon's association with a National Cancer Institute (NCI)-designated Cancer Center, and surgeon's average annual nephrectomy volume during the study period.

We first implemented our residual-based tree approach on the entire cohort of $11,136$ Medicare beneficiaries. After the full tree was grown, we performed cost-complexity pruning. The final tree was chosen based on 10-fold cross-validation, and the tree with the minimum error on the residualized response scale was selected. The deviance and Pearson residual-based trees were very similar. In Figure 2.5 we present the deviance residual-based tree. At each level of the tree, we show the best split (covariate with cut-point). The numbers in the terminal nodes denote the estimated probability of receiving PAR ($\hat{p}$) and the number of patients (n) in that node. Tumor size and the year of surgery were strong determinants of receipt of PAR. Surgeons affiliated with NCI-designated cancer centers were also more likely to use PAR. Year of medical school graduation and academic affiliation of the surgeon were other important determinants of PAR use.

Single tree model is usually unstable, where a small change in data may largely affect the tree architecture. Another shortcoming of single tree is its modest prediction performance. Ensemble methods such as bagging [Breiman, 1996] and random forest [Breiman, 2001] greatly improve upon these problems. The framework of our residual-based approach can be easily extended to generate residual-based ensembles of trees, where each tree in the ensemble is build on the residualized responses.

We also analyzed this data by growing a deviance residual-based random forests. Individual tree structures were lost in growing the forest, therefore, we evaluated the effect of covariates by examining their permutation variable importance. For

each covariate, its permutation importance was calculated as the average percentage increase in mean squared error (MSE) of the predicted responses (in the residual scale) from the forest, after randomly permuting the values of this variable. The permutation variable importance plot is displayed in Figure 2.6. This variable importance plot again confirms that tumor size is the most important determinant of PAR use. The second and fourth most important factors according to the ranked variable importance, i.e., year of medical school graduation and year of surgery also aligns with our results from the deviance residual-based tree. Surgeon age was also deemed important in the residual-based forest.

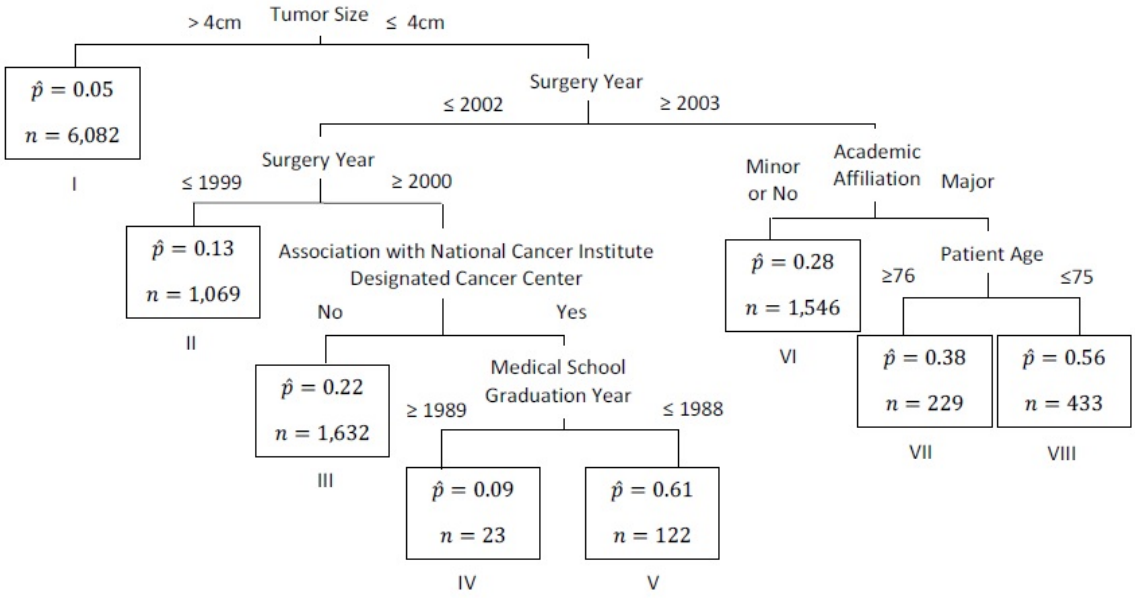

Figure 2.5: Deviance residual-based tree applied to kidney cancer data. In each terminal node, we list the estimated probability of receiving PAR ($\hat{p}$) and the number of patients (n) fall in that node.

2.5 Application to Surgical Mortality after Colectomy Study

Understanding the relationship between hospital/patient characteristics and patient outcomes is important for improving health care quality. In this analysis, we

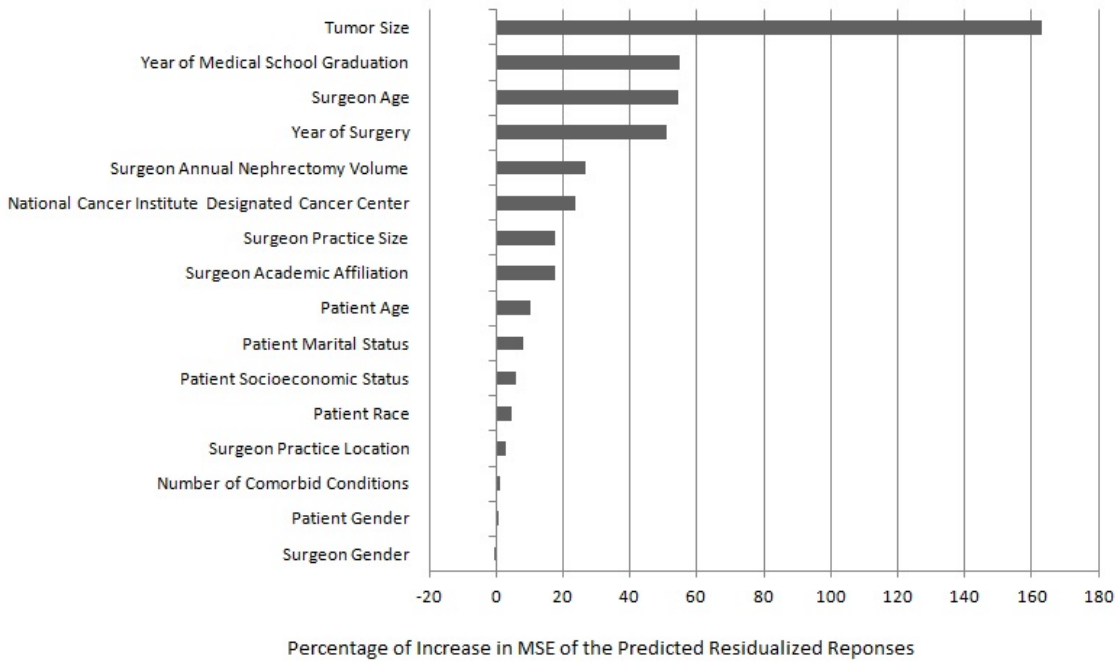

Figure 2.6: Variable importance plot of deviance residual-based random forests applied to kidney cancer data. Variable importance is defined as the percentage of increase in mean squared errors (MSE) of the predicted residualized reponses, after randomly permuting a variable.

were interested in identifying hospital characteristics and patient risk factors that might be associated with patient outcomes after receiving colon resection surgeries [Friese et al., 2015].

We extracted data from nationwide Medicare inpatient claims files between year 2009 and 2010 on patients hospitalized for colon resection. A total of $58,816$ patients 65 years or older, enrolled in fee-for-service Medicare were included in our analysis, and these patients were treated in $3,189$ hospitals. On average, each hospital treated 18 colectomy patients. We measured patient outcomes using failure to rescue (FTR), which is defined as death within 30 days of hospital admission for patients who have experienced a postoperative complication. FTR focuses on a hospital's capability to recognize and address a complication and is less affected by the severity of patients'

illness, therefore, it is considered as a better measure for comparing hospital quality [Ghaferi et al., 2009]. Seven hospital characteristics were considered, including a hospital's recognition of Magnet status by the American Nurses' Credentialing Center, which was a voluntary program reflecting a hospital's nursing care quality; the geographic location (rural vs. urban); whether a hospital had an active organ and/or tissue transplant program; whether a hospital had full-time equivalent medical residents or fellows; the number of staffed beds; a hospital's cost to charge ratio; and a hospital's registered nurse hours per patient day (RNHPPD). Patient risk factors included age (categorized as $65-69$, $70-74$, $75-79$, $80-84$, 85 and older), gender, race/ethnicity, and the number of comorbid conditions reported in their insurance claims.

This data set exhibited a two-level hierarchical structure as patients were nested within hospitals. We accounted for this hierarchical structure by fitting a hospital-specific random effect in the null GLMM model. Deviance residuals from the null model were used as response in growing the tree and random forest. The deviance residual-based tree is presented in Figure 2.7. At each level of the tree, we show the best split covariate along with the cut-point of the best split. For each terminal node, we present the estimated failure to rescue rate ($\hat{p}$) and the number of patients (n) in that node.

The deviance residual-based tree first split by patients' age and divided into three cohorts with age $65-74$, $75-84$, and 85 or older. As expected, FTR rates increased with patients' age. Patients aged $65-74$ were further split by their comorbid conditions: Terminal node I contained the $6,312$ patients with 3 or more comorbid conditions, who had the lowest FTR on average, which was 16%; Terminal node II contained the $15,526$ patients with no more than 2 comorbid conditions, and their

average FTR was 19%. For patients in this age group, hospital characteristics did not show an association with FTR. Among patients aged $75-84$ with 2 or more comorbid conditions, FTR was higher in rural hospitals than in urban hospitals, as we compared terminal node V to terminal nodes III and IV. In addition, terminal nodes I and VII suggested that between patients with no more than 1 comorbid condition, the average FTR was 5% higher for age $80-84$ than $75-79$. Patients older than 85 were further divided by their comorbid conditions, as well as the location and bed size of hospitals they were treated in: Terminal nodes VIII, IX and X indicated that among patients with 2 or more comorbid conditions, the average FTR in rural hospitals was 40%, which was much higher than urban hospitals; For patients with no more than 1 comorbid condition, the average FTR was 42% in hospitals with less than 406 staffed beds, comparing to 34% in hospitals with more than 407 staffed beds, as illustrated by terminal nodes XI and XII.

In summary, through this deviance residual tree, we found that failure to rescue exhibited an increasing trend with patients' age. The effects of hospital characteristics were more evident among older patients, who were commonly considered frailer. Older patients treated in bigger and/or urban hospitals tended to have lower FTR. Our findings on patients' comorbid conditions were confusing, since patients with more comorbid conditions appeared to have lower FTR. The frequency table demonstrated that for patients with 0, 1, 2, and 3 or more comorbid conditions, the crude FTR was 26%, 27%, 24%, and 21%, respectively. Hence this pattern, despite being counterintuitive, actually existed in the raw data, and our residual tree accurately identified this pattern. One possible explanation for this phenomenon is the bias in coding comorbidities, also known as "DRG Creep" [Iezzoni, 2012]. The number of comorbid conditions is collected from a patient's insurance claims, rather than

the medical records. Thus it is not a precise reflection of a patient's illness condition. Furthermore, hospitals with more resources are likely to identify and report more comorbidities in their patients' insurance claims, in order to receive higher reimbursements. These better resourced hospitals usually provide better health care service as well. Therefore, patients' comorbid conditions could be a confounder of hospitals' service quality.

The permutation variable importance plot based on deviance residual-based random forest is shown in Figure 2.8. The two most important variables, patients' age and hospitals' bed size matched with our findings from the single deviance residual-based tree. This confirmed our conclusion that FTR is primarily associated with patients' age, and larger hospitals have lower FTR in general. Interestingly, the importance of hospital location was relatively low, which is possibly due to its confounding with other hospital characteristics such as bed size and teaching program, as urban hospitals are usually bigger and more likely to have teaching program. It is also worth mentioning that patients' comorbidity was not deemed very important in the residual forest.

For this data, we also performed the standard classification tree analysis. However, the standard classification tree was unable to find any splits, and simply returned a root node. Therefore, this surgical mortality example illustrates a real data scenario when our residual-based tree approach served as a helpful alternative to the standard classification tree.

2.6 Application to Determinant of Vaccination Coverage Study

Despite rapid increase in vaccine coverage and substantial reduction in the incidence of many vaccine preventable diseases in India, poor vaccination coverage rates

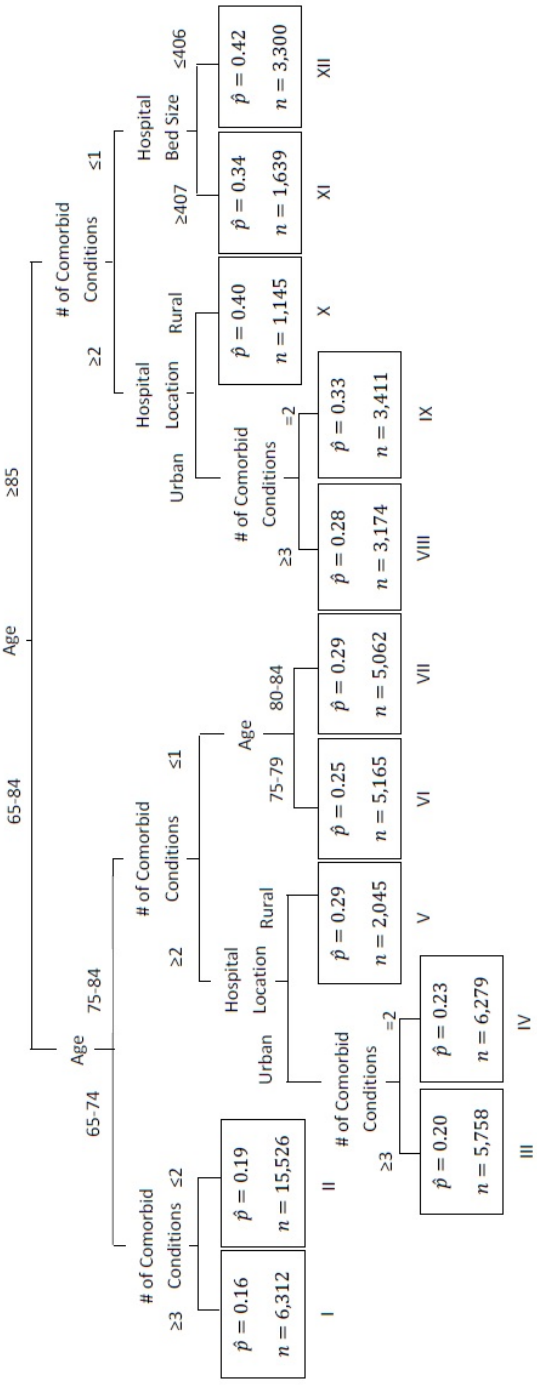

Figure 2.7: Deviance residual-based tree applied to surgical mortality data. In each terminal node, we list the estimated probability of failure to rescue rate ($\hat{p}$) and the number of patients (n) fall in that node.

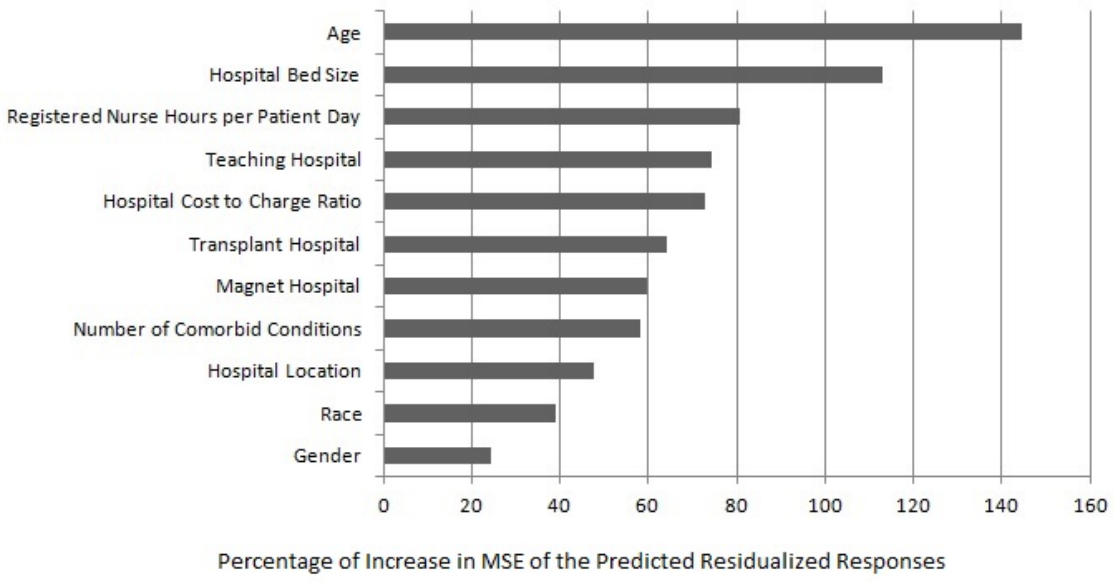

Figure 2.8: Variable importance plot of deviance residual-based random forest applied to surgical mortality data. Variable importance is defined as the percentage of increase in mean squared errors (MSE) of the predicted residualized reponses, after randomly permuting a variable.

still prevail in certain subgroups of children. Coverage rates may be improved upon by targeting these subgroups for interventions. It is therefore crucial to examine characteristics or factors associated with vaccine uptake in order to identify groups with deficient vaccination coverage.

Vaccination status is determined by a range of factors, from supply-side issues of service availability to the more demand-side determinants related to affordability and acceptability. Numerous studies have explored socio-economic and demographic determinants or demand-side factors of vaccination coverage. However, health care system drivers or supply side factors of vaccination coverage have received little attention in the literature and are not well-understood. The complex interplay between the demand and supply-side variables has rarely been examined. Traditional statistical methods, *e.g.*, logistic regression, are limited when analyzing variables that

may interact in complex ways, as interactions must be specified a priori. Therefore, we propose to use residual based tree methodology to study the role of supply-side constraints and demand-side determinants of immunization coverage.

We used data from the third round of the District Level Household Survey (DLHS-3) conducted during 2007-08. It is a large cross-sectional survey covering more than $700,000$ households from 601 districts ($1,000$-$1,500$ households per district) in 28 States and 6 Union Territories in India. DLHS-3 adopted a multistage stratified sampling design and interviewed more than $600,000$ ever married women aged $15-49$ years from the sampled households. As a preliminary analysis, we focused on data from Uttar Pradesh, which is a state located in northern India. A total of $7,704$ children from $2,595$ villages of 70 districts in this state were studied. The outcome of interest is whether a child is fully compliant with recommended vaccines. We examined seven supply side factors at village level, *e.g.*, availability of electricity, availability of anganwadi centre, has health subcentre within 3 kilometers, has primary health centre (PHC) within 5 kilometers, is connected by all-weather road to subcentre or PHC, availability of accredited social health activists and availability of auxiliary nurse midwives. On the demand side, we considered birth order, age and gender of the child, the educational status of the parents, mother's age and health knowledge, household head's caste, religion, household wealth index and location. In addition, we included the proportion of illiterate women and the proportion of households with higher (no less than 4) birth order children in the district.

This data exhibit a three-level structure as children are nested within villages nested within districts. We accounted for this hierarchical structure by assigning random effects to both villages and districts in the NULL GLMM. Residuals from the NULL GLMM were then used for growing the trees, which were further pruned

by minimizing the 10-fold cross-validation error. The Pearson and deviance residual-based trees were identical in this analysis, and we displayed the deviance residual-based tree in Figure 2.9. At each level of the tree, we show the covariate and cut-point of the best split. In each terminal node, we present the number of children (n) and the estimated probability of being fully immunized ($\hat{p}$). We also present the permutation variable importance plot based on deviance residual-based random forest in Figure 2.10.

Figure 2.9 suggests that mother's education is strongly associated with children immunization status. Mothers with less than 6 years of schooling are less likely to have their children fully immunized. Among these children, if their fathers receive less than 1 year of education, then the estimated fully immunization rate is as low as 0.15, and $2,335$ children belongs this group. For mothers with less than 6 years of schooling, if their husbands receive more than 1 year of education, then the immunization status is further associated with household wealth: around 24% children from poor (lower 60% quantile) families are fully immunized, comparing at 37% for children from rich (upper 40% quantile) families. For mothers with 6 to 10 years of schooling, their children have an estimated fully immunization rate of 0.45, and 976 children failing into this group. The immunization rate is as high as 0.61 if mothers have over 10 years of schooling, however, only 667 children belong to this cohort. The residual-based random forests confirmed that parents' education, especially mother's education is the most important variable. It is also meaningful to see that the proportion of illiterate women in a district is highly important.

In summary, parents' education, specially mother's education, are associated with children's immunization status. Increased efforts should be focused on less educated and low income families to improve the vaccination coverage.

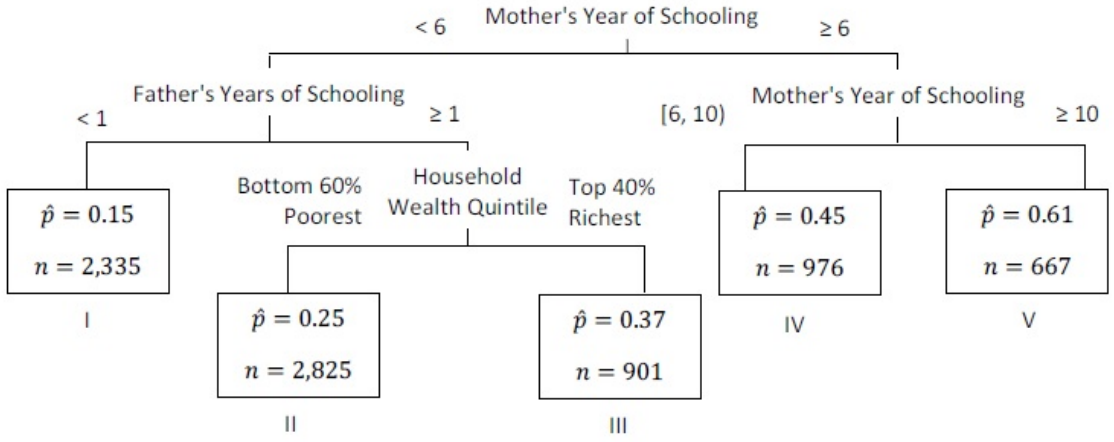

Figure 2.9: Deviance residual-tree applied to vaccination coverage data. In each terminal node, we list the estimated probability of receiving full immunization ($\hat{p}$) and the number of children (n) fall in that node.

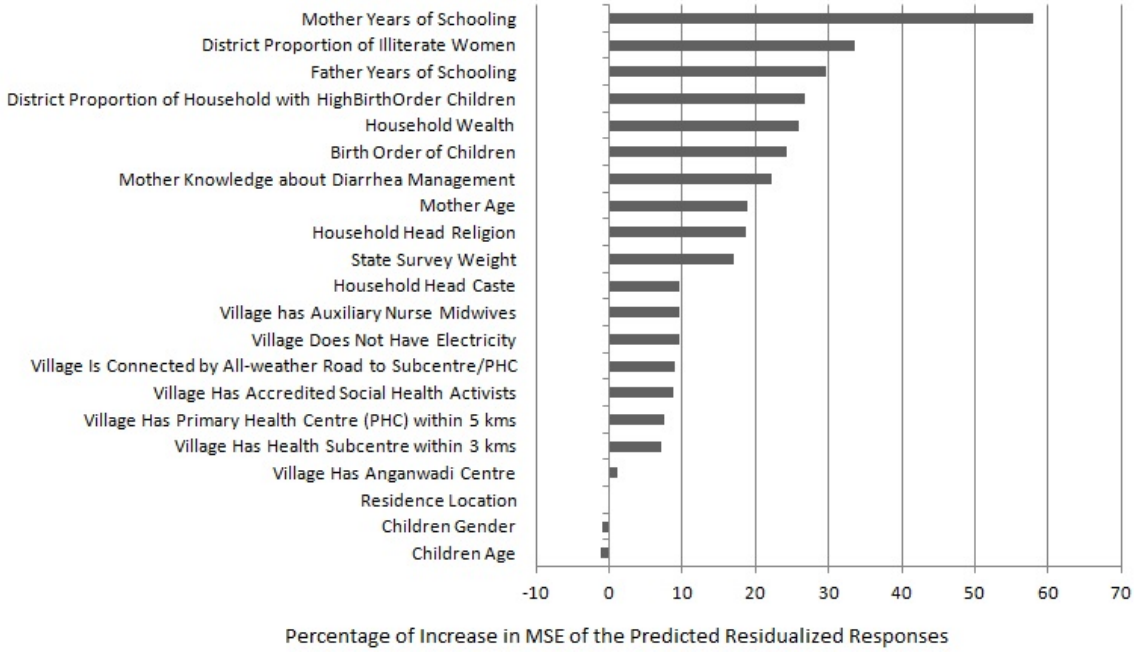

Figure 2.10: Variable importance plot of deviance residual-based random forest applied to vaccination coverage data. Variable importance is defined as the percentage of increase in mean squared errors (MSE) of the predicted residualized reponses, after randomly permuting a variable.

2.7 Discussion

Clustered data are abundant in practice, where observations within a cluster are usually correlated. This intra-cluster correlation needs to be accounted for when performing statistical analyses. Tree-based methods are one of the most flexible, intuitive, powerful data analytic tools for exploring complex data structures, however, the standard classification and regression trees (CART) paradigm is not designed for handling clustered data. In this chapter, we extended CART to handle clustered binary outcomes. Our approach was based on using residuals from a null generalized linear mixed model as the outcome. This circumvents modeling the correlation structure explicitly while still accounting for the cluster-correlated design, thereby allowing us to adopt the original CART machinery in tree growing, pruning and cross-validation. Class predictions for the terminal nodes of our residual-based tree were estimated based on success probabilities within each terminal node. We also provide a natural and direct extension of our residual-based tree to random forest.

Through extensive simulation studies, we have shown that our residual-based trees, especially the deviance residual-based tree, are more appropriate for analyzing clustered binary data than the standard CART. The residual-based trees were better adept in identifying the true structure in the data, and provided more accurate predictions. The improvements over the standard CART are substantial when the intra-cluster correlations are strong, given moderate cluster sizes. We also applied our residual-based approaches to studies of kidney cancer treatment receipt and surgical mortality after colectomy, where the data exhibited cluster-correlated structures. In both studies, residual-based tree and forest identified clinically meaningful subgroups. For the surgical mortality data, standard CART failed to split at all,

further demonstrating the advantage of our residual-based approach. One caveat of our approach is that when fitting the null generalized linear mixed models (GLMMs), at least moderate cluster sizes are needed in order to correctly estimate the cluster-specific effects. When the cluster sizes are small, the estimated random effects might be biased which in turn could affect the performance of our residual-based trees. In a sensitivity analysis, we also tried fititng the null GLMM using other algorithms such as Laplace approximation or Gauss-Hermite quadrature. However, we did not see significant improvements.

An R program implementing the residual-based tree algorithm is available.

CHAPTER III

Reflecting the Orientation of Teeth in Random Effects Models for Periodontal Outcomes

3.1 Introduction

Periodontal disease (PD) is a chronic inflammatory disorder that affects the gingiva, the supporting connective tissue and the alveolar bone, all of which anchor the teeth in the jaws. PD is the most common cause of tooth loss in adults in the United States, and moderate periodontal disease affects about half of the US population [Eke et al., 2012]. Given the increased life expectancy of US adults, the prevalence of periodontal disease may even increase in the future [Williams, 1990].

The diagnosis of periodontal disease involves the evaluation of gingival inflammation and tooth attachment structure destruction. The clinical parameters most commonly used in the diagnosis of PD are radiographically measured alveolar bone level (BL), bleeding on probing (BOP), clinical attachment level (CAL) and pocket depth (PKD). A tooth can be anatomically divided into the crown, which is covered by enamel, and the root, which is covered by the cementum. The border where the enamel meets the cementum is known as the cementoenamel junction (CEJ). Alveolar bone surrounds and supports the root of the tooth. Any detachment of the gingiva from the cementum forms a gap between the gum and the tooth, commonly referred to as a periodontal pocket. PKD quantifies the depth of the pocket, while CAL

quantifies the vertical distance from the CEJ to the bottom of the pocket. Figure 3.1 illustrates the clinical parameters and compares a normal tooth to a periodontally diseased tooth [Arora et al., 2009].

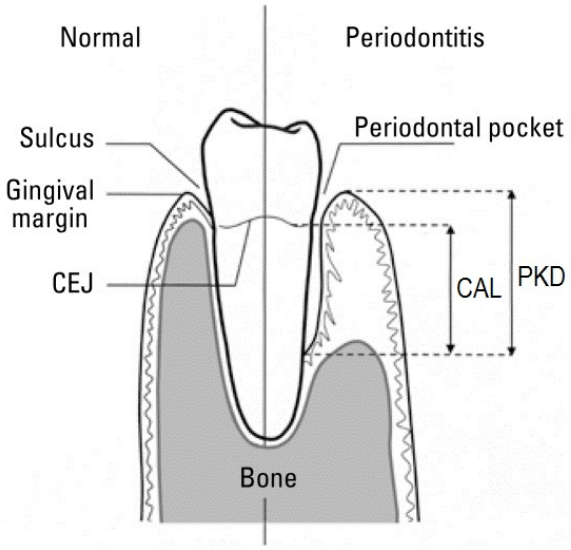

Figure 3.1: Diagram comparing clinical parameters in normal (left) and periodontally diseased (right) tooth [Arora et al., 2009].

BOP, CAL and PKD are typically measured manually via a periodontal probe; an examiner gently inserts the probe between the tooth and gingiva until slight resistance is felt. BOP is the indicator of bleeding resulting from the probe, while CAL and PKD are the corresponding distance read from the probe calibration, rounded to the nearest whole millimeter. CAL reflects both destruction of periodontal ligament and resorption of alveolar bone, and is considered as the "gold standard" for identifying periodontitis. According to the American Academy of Periodontology, severity of periodontitis has a site-specific, three-category definition based on the amount of CAL and is designated as slight ($1-2$ mm), moderate ($3-4$ mm) or severe (≥ 5 mm) [Wiebe and Putnins, 2000].

During a full periodontal exam, BOP, CAL and PKD are measured around each tooth at six sites: mesial-buccal (MB), buccal (B), distal-buccal (DB), distal-lingual

(DL), lingual (L) and mesial-lingual (ML); see Figure 3.2. According to the American Dental Association, we number the teeth from 1 to 32 using the Universal Numbering System, with numbers 1-16 referring to the sixteen teeth in the upper jaw (maxillary) and numbers 17-32 referring to the sixteen teeth in the lower jaw (mandibular). Due to the natural symmetry of a mouth, we further divide teeth equally into four quadrants of eight teeth each as shown in Figure 3.3. Based on their different functionality, teeth are classified as incisors, cuspids, bicuspids and molars. Since the third molars (also known as the "wisdom teeth", teeth 1, 16, 17 and 32) are often removed in most adults, even when healthy, these teeth are usually omitted from periodontal studies. Thus, BOP, CAL, and PKD can be measured at a maximum of 168 sites, six from each of the 28 teeth.

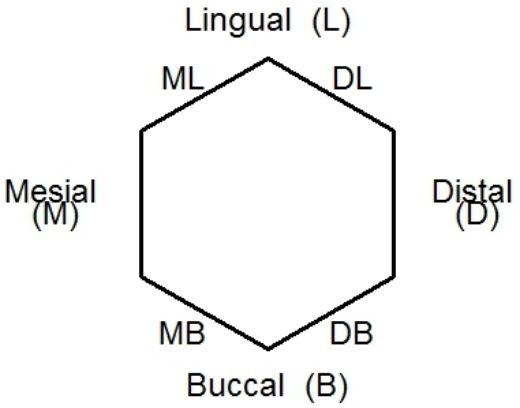

Figure 3.2: Schematic of a single tooth and sites of clinical examination.

Numerous studies have been conducted to identify risk factors of periodontal disease and to assess the effectiveness of treatments [Genco and Borgnakke, 2013]. Historically, statistical analysis in periodontal studies has been performed at the site-level via standard methods such as t-tests or regression models. However, these analyses assumed independence of sites and completely ignored the potential corre-

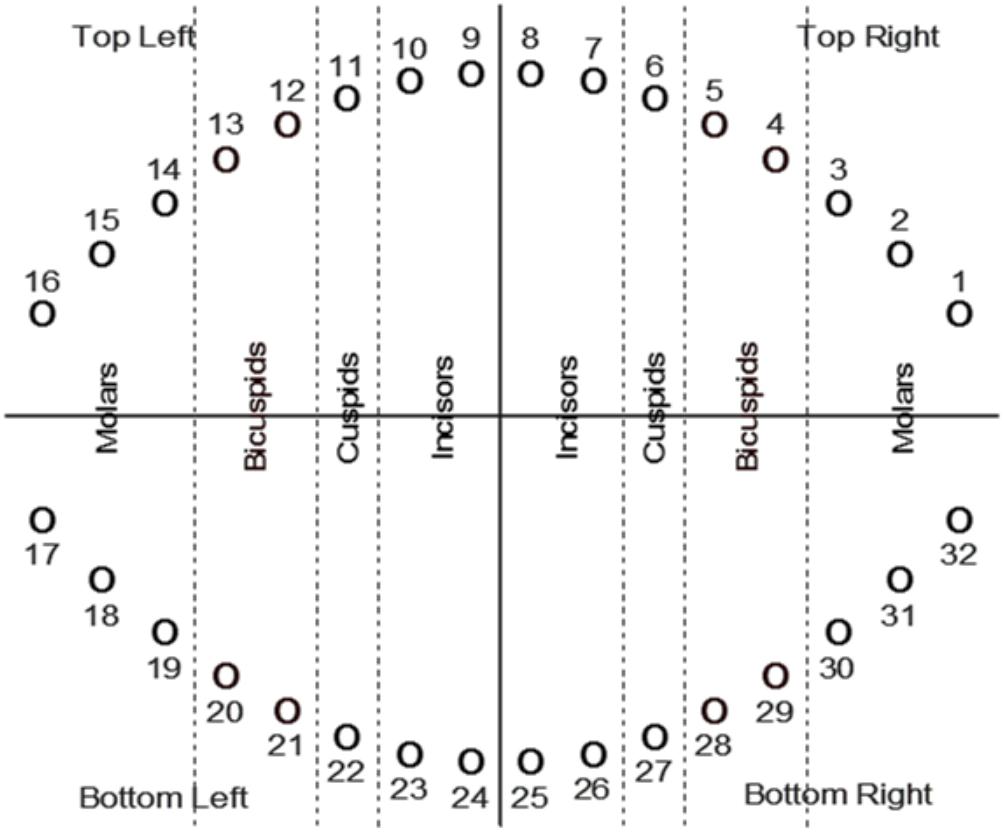

Figure 3.3: Diagram of teeth including numbering and functional groupings.

lations between sites. Alternatively, other analyses summarized the site-level measurements to mouth-level averages, leading to inefficient results [Emrich, 1990]. Due to the bias and inefficiency of these traditional approaches, conflicting conclusions have been made between periodontal studies [Harrel and Nunn, 2001].

In the last decade, several advanced statistical methods have been introduced to improve upon the weakness of traditional approaches. By treating periodontal data with a three-level hierarchical structure for mouth, tooth and site, multilevel models were introduced to model site-level measurements [Axtelius et al., 1999, Tu et al., 2004, Müller, 2009, Wan et al., 2009]. In these multilevel models, random effects were incorporated to account for the correlations within each level of the data. However, functionality was not incorporated in these models. Generalized estimating equations (GEE) with exchangeable working correlation structure were employed to study tooth-level PKD and CAL, which were obtained by averaging over all sites of each tooth [Harrel and Nunn, 2001]. Maitra [2012] first identified regions of the mouth that were most susceptible to periodontal disease via GEE by assuming the directions of the diseased teeth to follow a generalized von Mises distribution. To account for the within-mouth spatial correlation of teeth and sites, Reich et al. [2007] analyzed baseline site-level CAL data with a conditionally autoregressive (CAR) model. Reich and Hodges [2008] then extended the spatial model to a nonstationary spatiotemporal model to study longitudinal CAL data. Most recently, Reich et al. [2013] proposed a semi-parametric model to jointly model CAL and the location of missing teeth via kernel convolution methods.

Multilevel models assume that clinical parameters at the same level are equally correlated, i.e., all teeth within the same mouth are equally correlated and all sites within the same tooth are equally correlated, and ignores the spatial proximity be-

tween measurements. CAR models smooth only over adjacent neighbors, which are based solely on their spatially-defined distance. However, several studies have suggested that presence of periodontal disease is usually symmetric between the left side and the right side of a mouth [Mombelli and Meier, 2001, Minaya-Sánchez et al., 2010]. These authors also found that different functional types of teeth contributed differently to periodontal outcomes. Dowsett et al. [2002] stressed that the mouth exhibited symmetry among quadrants. Based on these findings, we have chosen to model the complex within-mouth correlation of periodontal outcomes by exploring the contributions of spatial proximity, the biological function of the teeth and the natural symmetry of the mouth.

Furthermore, although the semi-parametric spatial model proposed by Reich et al. [2013] could account for spatial proximity, biological function of the teeth as well as the symmetry of the mouth, it requires complex computational effort that does not exist in standard statistical packages. In this study, we propose to model periodontal outcomes with linear mixed models that can be implemented in standard statistical software packages. We will adjust for the complex within-mouth correlation by incorporating various random effects, and we will also compare our mixed models with GEEs and t-test via simulations and an application on actual data. Finally we will evaluate the performance of these approaches when data are missing under different biologically plausible mechanisms.

3.2 Methods

To explain our methods, we will focus on tooth-level CAL, although our concepts are applicable to PKD, BOP and BL as well. Tooth-level CAL is calculated as the average of measured CAL on the six sites of a tooth. Although each measurement

is a non-negative integer, we will treat the tooth-level average CAL as a continuous variable.

A total of m subjects are enrolled in the study, and for subject i, we observe a vector of $n_i \times 1$ outcomes $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{in_i})'$, where Y_{ij} is the CAL of tooth j. In periodontal studies, a healthy person has a maximum of 28 teeth, thus $n_i \leq 28$. The teeth are numbered as in Figure 3.3. We assume that $\boldsymbol{Y}_i$ has a multivariate normal distribution, i.e., $\boldsymbol{Y}_i \mid \omega_i, \boldsymbol{\beta} \sim \mathcal{N}[\boldsymbol{X}_i\boldsymbol{\beta}, \Sigma(\omega_i)]$, where $\boldsymbol{X}_i$ is a $n_i \times p$ matrix of covariates and $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients. Σ is a $n_i \times n_i$ covariance matrix depending on parameters ω_i, which reflects the within-mouth correlation structure. We further assume that teeth from different subjects are independent, i.e., $\boldsymbol{Y}_i \perp \boldsymbol{Y}_k$.

3.2.1 Introduction to Generalized Estimating Equations

Introduced by Liang and Zeger [1986], generalized estimating equations (GEE) is used to model correlated data and produces a moment based estimator. Unlike linear mixed models (described next), GEE does not require explicit assumptions on the joint distribution of $\boldsymbol{Y}_i$ and the correlation structures $\Sigma(\omega_i)$. Instead, GEE assumes that the marginal mean and variance of the outcomes are $E(Y_{ij}) = \mu_{ij}$, $\text{Var}(Y_{ij}) = \phi a_{ij}^{-1} v(\mu_{ij})$, and the mean model is $g(\mu_{ij}) = \boldsymbol{X}_{ij}^{'}\boldsymbol{\beta}$ ($g(.)$ is the link function and $g(\mu_{ij}) = \mu_{ij}$ for normal outcomes). Estimation of $\boldsymbol{\beta}$ is obtained via numerically solving the equation

$$\sum_{i=1}^{m} \boldsymbol{D}_i^{'} \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0,$$

where $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i / \partial\boldsymbol{\beta}^{'}$, and $\boldsymbol{V}_i$ is the working covariance matrix. The working covariance matrix $\boldsymbol{V}_i = \boldsymbol{V}_{M_i}^{1/2} \boldsymbol{R}_i(\alpha) \boldsymbol{V}_{M_i}^{1/2}$, where $\boldsymbol{V}_{M_i} = diag\{\phi a_{ij}^{-1} v(\mu_{ij})\}$ is the marginal variance and $\boldsymbol{R}_i(\alpha)$ is a working correlation matrix, where α is the correlation parameter. Common choices of $\boldsymbol{R}_i(\alpha)$ include independence, which assumes that all

teeth are independent; exchangeable, which assumes that all pairs of teeth have the same correlation; and autoregressive(1), which assumes that the correlation between two teeth decreases as their distance (measured by teeth number) increases.

The advantage of GEE is that the estimated $\hat{\boldsymbol{\beta}}$ is consistent given that the mean model is correctly specified, even if the correlation matrix $\boldsymbol{R}_i(\alpha)$ is misspecified. However, if $\boldsymbol{R}_i(\alpha)$ is correctly specified, the estimation $\hat{\boldsymbol{\beta}}$ is efficient within the linear estimating function family [Lipsitz et al., 1994]. Due to the complex within-mouth correlation, the assumption of any of the standard correlation structures seems unreasonable for periodontal data, while use of an unstructured form for $\boldsymbol{R}_i(\alpha)$ will require estimation of too many parameters, motivating the use of a linear mixed effects model.

3.2.2 Linear Mixed Effects Models

A linear mixed effect (LME) model is a linear model that contains both fixed and random effects, which provides a flexible framework for modeling correlated data [Henderson, 1950, Laird and Ware, 1982]. Following the notation introduced in Section 3.2.1, an LME model is written as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \epsilon_i,$$

where $\boldsymbol{b}_i \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{D})$ and $\epsilon_i \sim \mathcal{N}_{n_i}(\boldsymbol{0}, \boldsymbol{R}_i)$. Here $\boldsymbol{X}_i$ is the $n_i \times p$ matrix of fixed effects covariates, $\boldsymbol{Z}_i$ is the $n_i \times q$ matrix of random effects covariates, $\boldsymbol{b}_i = (b_{i1}, ..., b_{iq})'$ is the $q \times 1$ unknown vector of random effects for subject i, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{in_i})'$ is the $n_i \times 1$ vector of errors. The covariance matrix of $\boldsymbol{b}_i$ is $\boldsymbol{D}$ while $\boldsymbol{R}_i$ is the covariance matrix of $\boldsymbol{\epsilon}_i$. An LME model usually assumes that the random effects $\boldsymbol{b}_i$ are independent of the errors $\boldsymbol{\epsilon}_i$. Therefore, the covariance matrix of the responses $\boldsymbol{Y}_i$ of subject i is $\Sigma(\omega_i) = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{R}_i$.

The parameters in a LME model can be estimated by the maximum likelihood (ML) method implemented with the expectation-maximization (EM) algorithm [Laird and Ware, 1982]. The EM algorithm treats the maximization of the likelihood as a missing data problem, where the $\boldsymbol{Y}_i$ are the observed data and $\boldsymbol{b}_i$ are the missing data. Therefore, the full data are $(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{b}_i)$ with parameters $\boldsymbol{\beta}$, $\boldsymbol{D}$ and $\boldsymbol{R}_i$. This algorithm calculates the expected values of the missing values $\boldsymbol{b}_i$, given the current observed data and estimated parameter values (the expectation step), and then uses the expected values to update the estimates of the parameters $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{D}}$ and $\hat{\boldsymbol{R}}_i$ (the maximization step). These two steps are repeated until convergence to achieve valid estimates.

3.2.3 Functional and Spatial Modeling

The correlations among multiple teeth of a subject are not only related to the biological proximity of the teeth, but also to their functionalities. In order to properly account for this complex within-mouth correlation, we propose two linear mixed effect (LME) models to model periodontal data.

In the first model, we model the within-mouth variation between the maxillary and mandibular arches with random effects. The functional variation between different types of teeth (molar, bicuspid, cuspid and incisor) are also represented by random effects. In addition, we constrain the 28 teeth to be uniformly distributed around a unit-radius circle and model the spatial correlation as a circular effect. Thus, we can write our LME model as:

$$Y_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \sum_{k=1}^{2} U_{kij} a_{ki} + \sum_{l=1}^{4} Z_{lij} b_{li} + \epsilon_{ij}, \tag{3.1}$$

where $U_{1ij} = I(j \leq 15)$, $U_{2ij} = I(j \geq 18)$, Z_{1ij} = I(tooth j is a molar), Z_{2ij} = I(tooth j is a bicuspid), Z_{3ij} = I(tooth j is a cuspid), Z_{4ij} = I(tooth j is an

incisor). The random effects are mutually independent and are marginally distributed as $a_{ki} \sim \mathcal{N}(0, \gamma_k)$ and $b_{li} \sim \mathcal{N}(0, \tau_l)$. The variance parameter γ_1 represents the maxillar variation while γ_2 represents the mandibular variation. Functional variation is reflected by $\tau_l, l = 1, ..., 4$, where τ_1 represents molars, τ_2 represents bicuspids, τ_3 represents cuspids and τ_4 represents incisors. The vector of errors $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{in_i})'$ follows a multivariate normal distribution with mean $\mathbf{0}$, variance $\text{Var}(\epsilon_{ij}) = \sigma^2$ and correlations $\text{Corr}(\epsilon_{ij}, \epsilon_{ij'}) = \exp\{-d^2_{jj'}/\theta\}$. Here $d^2_{jj'}$ is the distance between tooth j and j' on the unit-radius circle and is calculated as

$$d^2_{jj'} = \{[\cos(A_j) - \cos(A_{j'})]^2 + [\sin(A_j) - \sin(A_{j'})]^2\},$$

where A_j is the angle of tooth j in the polar coordinate system (i.e., $A_2 = 0.295$ and $A_{31} = -0.295$). The variance parameter θ describes the spatial correlation and σ^2 is the residual variation.

3.2.4 Quadrant and Spatial Modeling

In the second model, we consider the natural symmetry of a mouth and divide it into four correlated quadrants, as is often done in clinical practice and research. The correlated quadrant effects are modeled with random effects. Similar to the model presented in Section 3.2.3, we utilize the polar coordinate distances to measure the spatial correlations. In addition, we introduce greater heterogeneity among teeth by allowing the residual variation to differ among the different types of teeth, rather than by function. Therefore, the LME can be written as:

$$Y_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \sum_{k=1}^{4} U_{kij} a_{ki} + \epsilon_{ij}, \tag{3.2}$$

where $U_{1ij} = I(j \leq 15)$, $U_{2ij} = I(j \geq 18)$, $U_{3ij} = I(9 \leq j \leq 24)$, $U_{4ij} = I(j \leq 8 \text{ or } j \geq 25)$. The random effects $(a_{1ij}, ..., a_{4ij})'$ follow a multivariate normal distribution centered at zero with a compound symmetry covariance matrix such that

$\text{Var}(a_{ki}) = \tau, k = 1, ..., 4$ and $\text{Cov}(a_{ki}, a_{k'i}) = \phi$ for $k \neq k'$. The variance parameters τ and ϕ describe the correlation among quadrants.

The vector of errors $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{in_i})'$ follows a multivariate normal distribution with mean $\mathbf{0}$, and variance $\text{Var}(\epsilon_{ij}) = \sigma^2(\sum_{l=1}^{7} Z_{lij}\rho_l)$, where $Z_{1ij} = I(j \in \{8, 9, 24, 25\})$, $Z_{2ij} = I(j \in \{7, 10, 23, 26\})$, $Z_{3ij} = I(j \in \{6, 11, 22, 27\})$, $Z_{4ij} = I(j \in \{5, 12, 21, 28\})$, $Z_{5ij} = I(j \in \{4, 13, 20, 29\})$, $Z_{6ij} = I(j \in \{3, 14, 19, 30\})$, $Z_{7ij} = I(j \in \{2, 15, 18, 31\})$. By doing so, we place the teeth into seven categories. Type 7 teeth (Teeth 2, 15, 18 and 31) are chosen as the reference and hence ρ_7 is constrained to be 1. Parameters $\rho_1, .., \rho_6$ reflect the variation of other types of teeth relative to Type 7 teeth. The variance parameter σ^2 is the residual variation of Type 7 teeth. The correlations $\text{Corr}(\epsilon_{ij}, \epsilon_{ij'}) = \exp\{-d_{jj'}^2/\theta\}$ are defined in the same way as they were in Section 3.2.3.

To better understand the difference between the functional and spatial LME (3.1) and the quadrant and spatial (3.2), in Table 3.1 we present a direct comparison of the covariance parameters of these two models.

Table 3.1: Comparison of covariance parameters between the functional and spatial LME (3.1) and the quadrant and spatial LME (3.2).

<table>
<tr><td colspan="5">Functional and Spatial LME (3.1)</td><td colspan="5">Quadrant and Spatial LME (3.2)</td></tr>
<tr><td colspan="2">Maxillar</td><td colspan="2">Mandibular</td><td></td><td colspan="2">Maxillar</td><td colspan="2">Mandibular</td><td></td></tr>
<tr><td>Left</td><td>Right</td><td>Left</td><td>Right</td><td></td><td>Left</td><td>Right</td><td>Left</td><td>Right</td><td></td></tr>
<tr><td colspan="4">Tooth Number</td><td></td><td colspan="4">Tooth Number</td><td></td></tr>
<tr><td>9</td><td>8</td><td>24</td><td>25</td><td rowspan="2">τ_4</td><td>9</td><td>8</td><td>24</td><td>25</td><td>ρ_1</td></tr>
<tr><td>10</td><td>7</td><td>23</td><td>26</td><td>10</td><td>7</td><td>23</td><td>26</td><td>ρ_2</td></tr>
<tr><td>11</td><td>6</td><td>22</td><td>27</td><td>τ_3</td><td>11</td><td>6</td><td>22</td><td>27</td><td>ρ_3</td></tr>
<tr><td>12</td><td>5</td><td>21</td><td>28</td><td rowspan="2">τ_2</td><td>12</td><td>5</td><td>21</td><td>28</td><td>ρ_4</td></tr>
<tr><td>13</td><td>4</td><td>20</td><td>29</td><td>13</td><td>4</td><td>20</td><td>29</td><td>ρ_5</td></tr>
<tr><td>14</td><td>3</td><td>19</td><td>30</td><td rowspan="2">τ_1</td><td>14</td><td>3</td><td>19</td><td>30</td><td>ρ_6</td></tr>
<tr><td>15</td><td>2</td><td>18</td><td>31</td><td>15</td><td>2</td><td>18</td><td>31</td><td>1</td></tr>
<tr><td colspan="2">γ_1</td><td colspan="2">γ_2</td><td></td><td colspan="2">a_1</td><td colspan="2">a_2</td><td></td></tr>
<tr><td colspan="4"></td><td></td><td>a_3</td><td>a_4</td><td>a_3</td><td>a_4</td><td></td></tr>
<tr><td colspan="4" rowspan="2">These six random effects
are mutually independent</td><td></td><td colspan="4">$Var(a_k) = \tau$</td><td></td></tr>
<tr><td></td><td colspan="4">$Cov(a_k, a_{k'}) = \phi$</td><td></td></tr>
</table>

3.2.5 Selecting Between Linear Mixed Effects Models

Given the two linear mixed effects models in Section 3.2.3 and 3.2.4, we want to choose which model best fits a given set of data. We are especially interested in choosing a parsimonious covariance matrix that produces the LME model with meaningful interpretations and efficient estimations. Many methods exist for model selection with LME models, and we will focus on two types: (1) information based on likelihood, specifically Akaike's information criterion (AIC) [Akaike, 1998] and Bayesian information criterion (BIC) [Schwarz, 1978]; and (2) the geodesic distance proposed by Carey and Wang [2011].

AIC attempts to prevent overparameterization of a model by penalizing the log-likelihood for the number of parameters used in the model. Although two types of maximum likelihood exist for LME models (maximum likelihood (ML) and restricted maximum likelihood (REML)), we will use REML to derive the AIC since our estimates are based on REML and our candidate models have the same mean structure. If we let l_R be the natural logarithm of the restricted likelihood of a LME model, its AIC is defined as AIC $= -2l_R + 2(p+q)$, where p denotes the number of fixed effect parameters and q denotes the number of variance parameters. Several methods have been proposed for treating fixed and random effects parameters differently [Müller et al., 2013], but we will avoid these differences by treating the variance parameters in the same way as the fixed effect parameters.

The simplest and most widely used BIC for LME models is BIC $= -2l_R + \log(n)(p+q)$, where n is the sample size. Here n is chosen as the total number of teeth over all subjects, i.e., $n = \sum_{i=1}^{m} n_i$. Models with smaller values of AIC and BIC are preferred to larger values. When the sample size is large ($log(n) > 2$), BIC tends to select more parsimonious models by putting higher penalty on the number

of parameters.

Carey and Wang [2011] derived a geodesic distance that quantifies the discrepancy between the working covariance and the empirical covariance, and allows for selection of the working covariance models for GEE. Inspired by their approach, we propose to choose the LME model whose model-based covariance is closest to the true covariance. We evaluate the discrepancy between the model-based covariance and the empirical covariance using the following statistics:

$$\begin{aligned} Q_0 &= \sum_{i=1}^{m} \boldsymbol{X}_i' \Sigma^{-1}(\hat{\omega}_i) \boldsymbol{X}_i, \\ Q_1 &= \sum_{i=1}^{m} \boldsymbol{X}_i' \Sigma^{-1}(\hat{\omega}_i) \boldsymbol{e}_i \boldsymbol{e}_i' \Sigma^{-1}(\hat{\omega}_i) \boldsymbol{X}_i, \end{aligned}$$

where $\Sigma(\hat{\omega}_i) = \boldsymbol{Z}_i \hat{\boldsymbol{D}} \boldsymbol{Z}_i' + \hat{\boldsymbol{R}}_i$ is the estimated covariance matrix from the LME model, and $\boldsymbol{e}_i = \boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}$ is the vector of residuals.

If we let c_k be the eigenvalues of $Q_0^{-1} Q_1$, Rotnitzky and Jewell [1990] proved that all elements of $c_k = 1$ whenever the model-based covariance and the true covariance coincide. Following this fact, we consider the following criteria: (i) $\Delta_1 = \sum_{k=1}^{p} (c_k - 1)^2 / p$; and (ii) $\Delta_2 = \sum_{k=1}^{p} (log(c_k))^2$. These two criteria should be close to zero when model-based covariance approximates the true covariance, which indicates that the LME model has properly accounted for the within-mouth correlation.

3.3 Simulation Studies

Our simulations are motivated by a clinical trial described in Ramseier et al. [2009] and Kinney et al. [2011], and a subset of data from this clinical trial are analyzed in Section 3.4. We apply the two proposed LME models, GEE with an exchangeable working correlation structure, GEE with AR-1 structure and a t-test using mouth-

level averages to the simulated data. We are primarily interested in comparing the efficiency of the fixed effect estimators from these different approaches. We also study the performance of the model selection criteria described in Section 3.2.5. Both complete data and data with missing outcomes are considered.

3.3.1 Complete Cases

We have 100 independent subjects, and every subject has 28 teeth. The only covariate considered is a subject-level indicator of periodontal disease, and we assume that diseased subjects have 1mm more CAL, on average, than healthy subjects. There are 50 subjects each in the diseased group and the healthy group. For all scenarios, we generate tooth-level CAL from a multivariate normal distribution, with mean equal to 2mm for the healthy group and 3mm for the diseased group, i.e., $\beta_0 = 2$ and $\beta_1 = 1$. In scenario 1, the covariance matrix, Σ, is chosen as the (unstructured) empirical within-mouth covariance of the baseline CAL in the motivating data set. For scenario 2, Σ is based upon the functional and spatial mixed model (LME (3.1)) described in Section 3.2.3. The variance parameters are chosen as $\gamma_1 = 0.7907$, $\gamma_2 = 0.4702$, $\tau_1 = 0.4830$, $\tau_2 = 0.0025$, $\tau_3 = 0.0991$, $\tau_4 = 0.1288$, $\theta = 0.3473$ and $\sigma^2 = 0.6725$. For scenario 3, Σ is based upon the quadrant and spatial mixed model (LME (3.2)) described in Section 3.2.4, with $\tau = 0.1777$, $\phi = 0.1600$, $\rho_1 = 0.8128$, $\rho_2 = 0.7570$, $\rho_3 = 0.8214$, $\rho_4 = 0.7233$, $\rho_5 = 0.7680$, $\rho_6 = 0.9836$, $\sigma^2 = 1.1569$ and $\theta = 0.3536$. The chosen variance parameter values in scenarios 2 and 3 are based on the motivating data set as well. One thousand simulations are performed for each scenario.

We are primarily interested in comparing the efficiency of the fixed effect estimator $\hat{\beta}_1$ among the different modeling approaches. We summarize the $1,000$ simulations by using (a) the empirical mean of $\hat{\beta}_1$ and (b) the empirical standard deviation (S.D.)

of $\hat{\beta}_1$. We then compare (a) to the true value $\beta_1 = 1$ and (b) to the mean of the model-based standard error (S.E.) or the robust S.E. To evaluate the performance of the model selection criteria, for each LME model, we compute the median values of AIC, BIC, Δ_1 and Δ_2, as well as the frequency each model is selected by each criterion. The results from the three simulation scenarios are presented in Table 3.2.

Across all three simulation scenarios, the empirical mean of $\hat{\beta}_1$ from all five methods is very close to the true value 1, which suggests that these methods are all unbiased when we have complete data. GEE with exchangeable correlation structure is identical to the t-test, which is due to the balanced design. The empirical S.D.'s resulting from both LME models are always smaller than the empirical S.D.'s from both GEEs and the t-test, which indicates that our LME models have improved the efficiency of estimates by properly modeling the complex within-mouth correlation. Under scenarios 2 and 3, the corresponding LME model used for generating the data gives the smallest empirical S.D. For GEE with exchangeable correlation structure, the empirical S.D., the model-based S.E., and the robust S.E. are similar, which suggests that the empirical correlation of the simulated data is close to exchangeable.

As to the model selection criteria, we notice that both AIC and BIC are able to identify the true LME model used for generating data under scenarios 2 and 3. However, information criteria do not guarantee selecting the model with higher efficiency, as illustrated in scenario 1, where LME (3.2) is selected while LME (3.1) has a smaller empirical S.D.

It is worth noticing that for LME (3.1), the model-based S.E. underestimates the empirical S.D. under scenarios 1 and 3, which indicates that LME (3.1) does not model the true covariance structure correctly in these situations. In the contrast, the

Table 3.2:

Summary of estimated $\hat{\beta}_1$ over $1,000$ simulations when data are generated from the empirical distribution under scenario 1, the functional and spatial mixed model (LME (3.1)) under scenario 2 and the quadrant and spatial mixed model (LME (3.2)) under scenario 3 (S.F.=Selection Frequency, * are based on 100 simulations.)

Scenario		GEE:Exch	GEE:AR-1	t-test	LME (3.1)	LME (3.2)
1	Mean $\hat{\beta}_1$	1.0091	1.0063	1.0091	1.0083	1.0092
	Empirical S.D.	0.1782	0.1802	0.1782	0.1725	0.1761
	Model-based S.E.	0.1673	0.1654		0.1182	0.1670
	Robust S.E.	0.1673	0.1686			
	AIC (S.F.)				6521.23 (0)	6366.81 (1000)
	BIC (S.F.)				6580.60 (0)	6438.05 (1000)
	Δ_1 (S.F.)	0.25*	0.16*		1.55 (0)	0.02 (1000)
	Δ_2 (S.F.)	0.17*	0.23*		0.78 (0)	0.02 (1000)
2	Mean $\hat{\beta}_1$	1.0070	1.0081	1.0070	1.0047	1.0065
	Empirical S.D.	0.1405	0.1515	0.1405	0.1331	0.1390
	Model-based S.E.	0.1325	0.1269		0.1263	0.1317
	Robust S.E.	0.1325	0.1428			
	AIC (S.F.)				6449.42 (1000)	6761.07 (0)
	BIC (S.F.)				6508.78 (1000)	6832.31 (0)
	Δ_1 (S.F.)	0.53*	0.46*		0.04 (268)	0.02 (732)
	Δ_2 (S.F.)	0.32*	0.87*		0.04 (279)	0.02 (721)
3	Mean $\hat{\beta}_1$	1.0099	1.0082	1.0099	1.0096	1.0098
	Empirical S.D.	0.1809	0.1925	0.1809	0.1807	0.1799
	Model-based S.E.	0.1722	0.1680		0.1287	0.1732
	Robust S.E.	0.1722	0.1864			
	AIC (S.F.)				6690.01 (0)	6518.59 (1000)
	BIC (S.F.)				6749.37 (0)	6589.83 (1000)
	Δ_1 (S.F.)	0.53*	0.11*		1.33 (0)	0.02 (1000)
	Δ_2 (S.F.)	0.34*	0.15*		0.69 (0)	0.02 (1000)

model-based S.E. is close to the empirical S.D. for LME (3.2) across all three scenarios. Through additional simulations (not shown), we find that this is because LME (3.2) allows for quadrant-level correlations, while LME (3.1) restricts the maxillary and the mandibular arches to be uncorrelated. While simulating from LME (3.1), even though the true quadrant-level correlation is zero between Quadrants I and III, I and VI, II and III, II and IV, the fitted LME (3.2) is able to approximate the true correlations well through the covariance parameters τ and ϕ. On the other hand, while simulating from LME (3.2), in which the true correlations between all pairs of quadrants are non-zero, the fitted LME (3.1) forces a zero correlation between Quadrants I and III, I and VI, II and III, II and IV, which leads to a model-based covariance that deviates from the truth.

Unlike the GEEs, the LME models require correct modeling of the covariance structure to obtain valid estimation; otherwise, the model-based S.E. might underestimate the true variability [Gałecki and Burzykowski, 2013]. Through our simulations, we find that the geodesic distance statistics Δ_1 and Δ_2 serve as good criteria for identifying the LME model that has modeled the true covariance appropriately; the values of Δ_1 and Δ_2 are near zero when the LME model-based S.E. is similar to the empirical S.D. However, smaller Δ_1 and Δ_2 are not necessarily associated with a more efficient LME model, as illustrated by scenario 2, where LME (3.1) has a smaller empirical S.D. and LME (3.2) is more frequently preferred by these geodesic distance statistics.

We also calculated the median geodesic distance of the two GEEs over the first 100 simulations. It is interesting to notice that under scenarios 1 and 3, the Δs of GEEs are between the Δs of LME (3.2) and LME (3.1); Under scenario 2, GEEs have larger Δs than the two LME models. This matches our previous finding that

higher Δs are related to more bias in the model-based S.E.

Based on these simulations, we conclude that when data are complete, estimates from the proposed LME models, the GEEs and the t-test are all unbiased. The proposed LME models could moderately improve the efficiency of the estimates. The LME model with smaller values in information criteria, Δ_1 and Δ_2 should be selected for making inference.

3.3.2 With Missing Data

We simulate data under different missing mechanisms and examine the effect of missingness on the above mentioned methods. The tooth-level CALs are generated from the empirical multivariate normal distribution, as described in scenario 1 of Section 3.3.1. We then impose missing CAL according to three different biologically plausible mechanisms: (1) when missingness depends on the covariates but not on the outcomes, i.e., covariate-dependent missingness (CDM); (2) when missingness only depends on the observed outcomes, i.e., missing at random (MAR); (3) when missingness depends on both observed and unobserved outcomes, i.e., missing not at random (MNAR).

We assume that around 5.5% of teeth are missing, which is the percentage of missing teeth at baseline in the motivating data set. When simulating under CDM, we assume that diseased subjects have more missing teeth than healthy subjects. We generate binary missing indicators with probability 0.093 for the diseased group, and with probability 0.017 for the healthy group, thus the missingness depends on the observed covariate (disease group) only. When simulating under MAR, we assume that for each subject, the first tooth (Tooth 1) is always observed. A missing indicator for Tooth 2 depends on the CAL value of Tooth 1 via a logistic model, i.e., $logit(\text{P}(\text{Tooth 2 is missing}|\text{Tooth 1 CAL})) = -8.935 + 1.721 \times (\text{Tooth 1 CAL})$. Sim-

ilarly, for every following tooth, the probability of missingness depends on the CAL value of the closest existing tooth via the same logistic model. The coefficient values in the logistic model are chosen such that the missing probabilities match with CDM. When simulating under MNAR, we set any teeth with CAL values larger than 4.5mm to be missing. As a result, the missingness depends on both observed and unobserved outcomes. Here we have chosen 4.5mm as the cutoff so that around 5.5% of teeth are removed, which is consistent with the two previous situations. Furthermore, as mentioned in Section 3.1, 5mm is the cutoff for classifying severe periodontitis, and such teeth are usually extracted by the dentist.

The simulations are repeated $1,000$ times under each missing mechanism; we display the summary statistics in Table 3.3.

Based on the simulations, we find that with a moderate missing percentage, CDM does not impact the performances of the GEEs, the t-test and the LME models, as their estimates are still unbiased and the empirical S.D.'s remain roughly the same as with complete data. The LME models are still more efficient than their competitors, and the model selection criteria work as before. When data are MAR, the estimates from the LME models are unbiased while the estimates from the GEEs and the t-test are biased. LME (3.1) is the most efficient method as it has the smallest empirical S.D. When data are MNAR, all five methods suffer from biased estimates and loss of efficiency, although the bias in the LME models is less severe than the GEEs and the t-test. LME (3.1) is still the most efficient among these methods.

3.4 Application to Michigan Periodontal Study

We applied our LME models and the competing approaches to the data motivating our simulations. This non-randomized longitudinal observational study, conducted

Table 3.3:
Summary of estimated $\hat{\beta}_1$ over $1,000$ simulations when data are generated from the empirical distribution under CDM, MAR or MNAR. (S.F.=Selection Frequency.)

		GEE:Exch	GEE:AR-1	t-test	LME (3.1)	LME (3.2)
CDM	Mean $\hat{\beta}_1$	1.0049	1.0062	1.0049	1.0056	1.0051
	Empirical S.D.	0.1792	0.1831	0.1792	0.1731	0.1767
	Model-based S.E.	0.1676	0.1657		0.1187	0.1674
	Robust S.E.	0.1676	0.1691			
	AIC				6214.56	6057.78
	(S.F.)				(2)	(998)
	BIC				6273.31	6128.38
	(S.F.)				(3)	(997)
	Δ_1				1.53	0.02
	(S.F.)				(0)	(1000)
	Δ_2				0.77	0.02
	(S.F.)				(0)	(1000)
MAR	Mean $\hat{\beta}_1$	0.9790	1.0118	0.9847	0.9916	0.9974
	Empirical S.D.	0.1671	0.1762	0.1681	0.1649	0.1689
	Model-based S.E.	0.1570	0.1557		0.1180	0.1669
	Robust S.E.	0.1638	0.1696			
	AIC				6162.32	6013.86
	(S.F.)				(0)	(1000)
	BIC				6221.03	6084.33
	(S.F.)				(0)	(1000)
	Δ_1				1.50	0.02
	(S.F.)				(0)	(1000)
	Δ_2				0.76	0.02
	(S.F.)				(0)	(1000)
MNAR	Mean $\hat{\beta}_1$	0.8843	0.8930	0.8889	0.9005	0.9042
	Empirical S.D.	0.1536	0.1569	0.1540	0.1516	0.1550
	Model-based S.E.	0.1455	0.1431		0.1073	0.1509
	Robust S.E.	0.1480	0.1502			
	AIC				5914.97	5774.89
	(S.F.)				(0)	(1000)
	BIC				5973.80	5845.42
	(S.F.)				(0)	(1000)
	Δ_1				1.59	0.04
	(S.F.)				(0)	(1000)
	Δ_2				0.78	0.04
	(S.F.)				(0)	(1000)

at the Michigan Center for Oral Health Research, involved 50 periodontally healthy and 50 periodontally diseased subjects. Periodontal exams were given to the subjects periodically at the baseline, six and twelve months after enrollment. We were interested in estimating the difference in CAL between the periodontally diseased and healthy group at the baseline, and comparing the estimates from our LME models to other methods.

A total of $2,646$ teeth were observed at the baseline, which suggested a 5.5% rate of missing. On average, healthy subjects lost 1.7% of their teeth while periodontally diseased subjects lost 9.3%. The histogram plot suggested that tooth-level CAL was not normally distributed, but positively skewed. Although a transformation could fix the skewness, it would hamper the interpretation of the estimates. In addition, Jacqmin-Gadda et al. [2007] has shown that linear mixed model is robust to a non-normal error distribution. Therefore, we decided to analyze the tooth-level CAL data without transformation.

The results were presented in Table 3.4. The estimated mean CAL of periodontally diseased subjects was 1.6439mm larger than the healthy subjects by the functional and spatial LME (3.1), and it was 1.7546mm by the quadrant and spatial LME (3.2). Both LME estimates, which were more robust to missing data, were smaller than the estimates of the GEEs and the t-test. Comparing the two LME models, both AIC and BIC preferred LME (3.1), which had the smallest model-based S.E. across all five methods. However, since the geodesic distance statistics were larger for LME (3.1), which indicated that its model-based S.E. might have underestimated the truth, it was more proper to make inference using LME (3.2), which also had a smaller model-based S.E. than the robust S.E. of the GEEs and the t-test. Therefore, our LME models have improved the efficiency of the estimate, although the improvement

was minor in this application. It is worth noticing that for GEEs, the model-based S.E. was different from the robust S.E., which indicated that the true within-mouth correlation in this actual data set was more complex than exchangeable or AR-1.

Table 3.4: Analysis results of Michigan Center for Oral Health Research Data.

	GEE: Exch	GEE: AR-1	t-test	LME (3.1)		LME (3.2)	
Estimated Difference	1.8252	2.1135	1.8273	1.6439		1.7546	
Model-based S.E.	0.1674	0.1751	0.1828	0.1271		0.1798	
Robust S.E.	0.1805	0.1869		0.1726		0.1781	
AIC				6154.86		6180.49	
BIC				6213.66		6251.05	
Δ_1				4.23		0.81	
Δ_2				1.41		1.37	
Covariance Parameters				$\hat{\gamma}_1$	0.7907	$\hat{\tau}$	0.1777
				$\hat{\gamma}_2$	0.4702	$\hat{\phi}$	0.1777
				$\hat{\tau}_1$	0.4830	$\hat{\rho}_1$	0.8128
				$\hat{\tau}_2$	0.0000	$\hat{\rho}_2$	0.7570
				$\hat{\tau}_3$	0.0991	$\hat{\rho}_3$	0.8214
				$\hat{\tau}_4$	0.1288	$\hat{\rho}_4$	0.7233
				$\hat{\theta}$	0.3473	$\hat{\rho}_5$	0.7680
				$\hat{\sigma^2}$	0.6725	$\hat{\rho}_6$	0.9836
						$\hat{\theta}$	0.3536
						$\hat{\sigma^2}$	1.1569

3.5 Discussion

Periodontal disease is prevalent in the United States. The relatively small signal-to-noise ratio in periodontal outcomes has raised the request for proper statistical methods. In this chapter, we have proposed to model tooth-level periodontal outcomes using two linear mixed effects models, which could account for the complex within-mouth correlation and provide better estimates. Via simulations, we have shown that our mixed models are more robust to missing data (unbiased provided that data is not MNAR), and more efficient than traditional methods such as GEE and t-test in periodontal analysis. We have also suggested model selection criteria for choosing the LME model that better fits the data. The proposed LME models and the selection criteria can be conveniently implemented in standard software

packages, which makes them readily accessible to dentists.

One disadvantage of the geodesic distance based selection criteria is that they do not formally test how well the LME models approximate the true data-based covariance. Incorporating more random effects into a LME model could better approximate the data-based within-mouth correlation. However, over-parameterized models face the risk of overfitting, and the complex null distributions in hypothesis testing random effects make it impractical to extensively test all random effects one by one. It is therefore beneficial to derive a statistical test based on the geodesic distance that can identify whether the LME model has adequately approximated the true within-mouth correlation. We will explore this idea in the next chapter.

Longitudinal data are common in periodontal studies, where the repeated measurements over time are usually correlated. In order to efficiently analyze longitudinal periodontal outcomes, it is crucial to generalize our LME models to account for the temporal correlation as well.

Finally, periodontal disease is a leading cause of tooth loss, and teeth with larger periodontal outcomes have a higher chance of being removed. Therefore, informative missing, i.e., MNAR is inevitable in periodontal studies. Through the simulations, we have seen that our LME models are biased and less efficient when data are MNAR. Joint modeling of missing teeth and periodontal outcomes rises as an interesting and rewarding direction for future studies.

CHAPTER IV

Permutation Tests for Covariance Structure Assumption in Linear Mixed Effects Models

4.1 Introduction

Correlated data is abundant in biomedical studies. For example, in longitudinal or repeated measures data, outcomes for subjects are collected repeatedly over time, and thereby are typically correlated within-subject through sharing subject-specific characteristics. In multilevel or clustered data, observations within the same level or cluster are generally more similar to each other than observations from different clusters, which induces within-level or within-cluster correlations. Linear mixed effects (LME) models are a rich family of models containing both fixed and random effects, which are widely adopted in modeling correlated data [Laird and Ware, 1982]. The random effects and residual errors in LME models create a flexible class of covariance structures that allows us to account for and take advantage of the structured patterns in the correlated data.

In applying LME models, it is important to appropriately model the true covariance structure in order to obtain efficient standard errors and valid statistical inference for fixed effect parameters. Lange and Laird [1989] demonstrated that, in general, variance of fixed effects estimates and random effects may be biased when the covariance structure is not correct. Taylor and Law [1998] showed that individual

predictions are affected by misspecified covariance structures. Although the "sandwich" estimator recommended by Liang and Zeger [1986] brought in robustness to misspecified covariance, the sandwich estimator is less efficient than the estimator using the correct covariance model. Valid inference with the sandwich estimator also requires additional assumptions about the missing data, and the sandwich estimator has not been fully evaluated in small samples [Verbeke and Molenberghs, 2009]. Therefore, it is still desirable to accurately model the true covariance structure when fitting LME models.

Despite the importance of appropriately modeling the true covariance structure in LME models, the diagnostic methodology for evaluating the covariance structure assumption remains relatively underdeveloped. Houseman et al. [2004] and Jacqmin-Gadda et al. [2007] proposed drawing quantile-quantile (Q-Q) plots of Cholesky residuals to graphically examine the goodness of fit in LME models. The former paper also established the asymptotic properties of the Cholesky residuals. Verbeke and Molenberghs [2009] suggested an informal check for the appropriateness of the selected random effects by comparing the fitted covariance function based on an LME model to the smoothed sample covariance function of the marginal residuals. However, these two approaches do not provide any formal statistical inferences.

An alternative solution is to successively test for the inclusion or exclusion of all possible random effects; i.e., testing variance components against 0. However, it is challenging to test for random effects because the variance component is equal to 0 under the null hypothesis, which is on the boundary of the parameter space. As a result, the asymptotic null distribution of the Wald, score, and likelihood ratio tests no longer follow the typical χ^2 distributions, but often follow mixture of χ^2 distributions [Stram and Lee, 1994, Verbeke and Molenberghs, 2003, Silvapulle,

1992]. In addition, when testing for multiple variance components simultaneously, it is especially difficult to determine the mixture weights. Other approaches include the random effect selection methods proposed by Chen and Dunson [2003] and Kinney and Dunson [2007]. These authors proposed the use of a Bayesian stochastic search to identify nonzero random effect variances in LME models. However, these approaches are computationally expensive and do not currently exist in standard statistical packages.

Permutation tests provide a viable alternative to the aforementioned methods. A permutation test determines the null distribution of the test statistic through permutations of the data and circumvents the difficulties with explicitly deriving an asymptotic distribution. For LME models, Fitzmaurice et al. [2007] first introduced using permutation tests for the inclusion of a single random effect in multilevel models. Lee and Braun [2012] proposed two permutation tests, one based on the best linear unbiased predictors and one based on the restricted likelihood ratios test statistic, for testing single or multiple random effects. Drikvandi et al. [2013] proposed testing for multiple random effects, by defining a test statistic based on the variance least square estimator of variance components, and applied a permutation procedure to approximate its null distribution. However, these permutation tests were limited to the testing of inclusion or exclusion of specific random effects, rather than the overall fit of the assumed covariance structure. Finally, it is worth mentioning that Schmoyer [1994] proposed using permutations of the regression residuals to test for correlation in errors of ordinary linear models.

Our method integrates the informal check suggested in Verbeke and Molenberghs [2009] and the permutation procedures introduced in Lee and Braun [2012], and it leads to three permutation tests that allow for inference on the overall covariance

structure assumption of LME models. All three test statistics are defined as different metrics that quantify the discrepancy between the fitted covariance matrix based on the LME model and the smoothed sample covariance matrix of the marginal residuals; the empirical null distributions are generated by permutations of the Cholesky residuals. Via simulations, we demonstrate that two of our tests have valid size and sufficient power under different covariance structure assumptions.

The rest of this chapter is organized as follows. In Section 4.2, we review some background of LME models and introduce our permutation tests on covariance structures. Section 4.3 presents simulation studies designed to evaluate the validity and powers of our proposed tests for different components of covariance structures. We illustrate our methods in Section 4.4 using data from a periodontal disease study. Section 4.5 contains some concluding remarks.

4.2 Methods

4.2.1 Linear Mixed Effects Models

Consider a repeated measures scenario in which Y_{ij} is the jth measurement of subject i for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n_i$. Then vector $\boldsymbol{Y}_i = \{Y_{i1}, ..., Y_{in_i}\}'$ represents all n_i outcomes of subject i. An LME model can be expressed as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{4.1}$$

where $\boldsymbol{\beta} = \{\beta_1, ..., \beta_p\}'$ is a $p \times 1$ vector of population level fixed effect coefficients, $\boldsymbol{b}_i = \{b_{i1}, ..., b_{iq}\}'$ is the $q \times 1$ vector of random effect coefficients for subject i, and $\boldsymbol{\epsilon}_i = \{\epsilon_{i1}, ..., \epsilon_{in_i}\}'$ is the $n_i \times 1$ vector of random errors of subject i. The $n_i \times p$ matrix $\boldsymbol{X}_i$ contains fixed effect covariates, and $\boldsymbol{Z}_i$ is the $n_i \times q$ matrix of random effect covariates, respectively, for the ith subject. Generally, all elements of the first column of $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are equal to 1 to represent the fixed and random intercept,

respectively.

For an LME model, it is commonly assumed that the random effects $\boldsymbol{b}_i$ follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{D}$; i.e., $\boldsymbol{b}_i \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{D})$; and the random errors $\boldsymbol{\epsilon}_i$ follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{R}_i$; i.e., $\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \boldsymbol{R}_i)$. The covariance matrix $\boldsymbol{R}_i$ is usually assumed to be diagonal. However, in cases when the variability in observations cannot be completely modeled by the random effects, we also introduce correlated random errors via non-diagonal $\boldsymbol{R}_i$ to allow for more flexible covariance structures. Finally, random effects $\boldsymbol{b}_i$ and random errors $\boldsymbol{\epsilon}_i$ are usually assumed to be independent.

To simplify our notation, we combine data over all m subjects by stacking vectors $\boldsymbol{Y}_i$, $\boldsymbol{b}_i$, $\boldsymbol{\epsilon}_i$, and matrices $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$, $\boldsymbol{R}_i$ respectively, and re-write our LME model as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$. The formulation of this LME model implies an assumption that the covariance matrix of the outcomes $\boldsymbol{Y}$, $var(\boldsymbol{Y})$, is identical to the model-based covariance structure $\boldsymbol{W} = \boldsymbol{Z}^T\boldsymbol{D}\boldsymbol{Z} + \boldsymbol{R}$.

Estimation of parameters $\boldsymbol{\beta}$, $\boldsymbol{D}$ and $\boldsymbol{R}$ is typically done through maximum likelihood (ML) or restricted maximum likelihood (REML), and the subject-specific random effects $\boldsymbol{b}$ can be predicted using the best linear unbiased predictions (BLUPs). Verbeke and Molenberghs [2009] provides a comprehensive discussion of these topics. According to Henderson [1950], the estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$ and the prediction of $\boldsymbol{b}$, $\tilde{\boldsymbol{b}}$, can be obtained analytically as solutions to the following mixed model equations:

$$\begin{aligned} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\tilde{\boldsymbol{b}} &= \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Y}, \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} + (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{D}^{-1})\tilde{\boldsymbol{b}} &= \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Y}, \end{aligned} \tag{4.2}$$

which lead to

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T \hat{\boldsymbol{W}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \hat{\boldsymbol{W}}^{-1} \boldsymbol{Y}, \\ \tilde{\boldsymbol{b}} &= \hat{\boldsymbol{D}} \boldsymbol{Z} \hat{\boldsymbol{W}}^{-1} \hat{\boldsymbol{e}}, \end{aligned} \tag{4.3}$$

where $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ are the marginal residuals from the fitted LME model, and $\hat{\boldsymbol{W}} = \boldsymbol{Z}^T \hat{\boldsymbol{D}} \boldsymbol{Z} + \hat{\boldsymbol{R}}$ is the estimated model-based covariance matrix for $\boldsymbol{Y}$.

Equation 4.3 induces the so called robust "sandwich" estimator for the variance of the estimated fixed effects $\hat{\boldsymbol{\beta}}$:

$$\hat{var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \hat{\boldsymbol{W}}^{-1} \boldsymbol{X})^{-1} (\boldsymbol{X}^T \hat{\boldsymbol{W}}^{-1} \hat{var}(\boldsymbol{Y}) \hat{\boldsymbol{W}}^{-1} \boldsymbol{X}) (\boldsymbol{X}^T \hat{\boldsymbol{W}}^{-1} \boldsymbol{X})^{-1}, \tag{4.4}$$

where $\hat{var}(\boldsymbol{Y})$ is the smoothed empirical estimator of $var(\boldsymbol{Y})$ that is based on the marginal residuals $\hat{\boldsymbol{e}}$ [Liang and Zeger, 1986]. Although the sandwich estimator is robust to misspecified covariance structure, Verbeke and Molenberghs [2009] noted that, (i) the sandwich estimator is less efficient than the one using the correct covariance model; (ii) valid inference requires additional assumptions about missing data; and (iii) the sandwich estimator has not been fully evaluated in small samples. Therefore, in practice, we usually use the reduced estimator

$$\hat{var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \hat{\boldsymbol{W}}^{-1} \boldsymbol{X})^{-1} \tag{4.5}$$

instead. The validity of the model-based estimator in Equation (4.5) depends on the assumption that the model-based covariance matrix $\boldsymbol{W}$ is identical to the true covariance matrix of $\boldsymbol{Y}$, $var(\boldsymbol{Y})$, i.e., the covariance structure of the LME model is correctly specified. When this assumption does not hold, the model-based estimator will result in a biased assessment of the variability of $\hat{\boldsymbol{\beta}}$, which might impact the validity of statistical inferences. It is therefore essential to examine the covariance structure assumption of an LME model before applying it for inference.

4.2.2 Permutation Tests for Covariance Structure Assumption

We wish to test the null hypothesis

$$\boldsymbol{H}_0 : \boldsymbol{W} = var(\boldsymbol{Y}) \tag{4.6}$$

versus the alternative hypothesis $\boldsymbol{H}_1 : \boldsymbol{W} \neq var(\boldsymbol{Y})$.

The model-based covariance matrix can be estimated by $\hat{\boldsymbol{W}} = \boldsymbol{Z}^T \hat{\boldsymbol{D}} \boldsymbol{Z} + \hat{\boldsymbol{R}}$, where $\hat{\boldsymbol{D}}$ and $\hat{\boldsymbol{R}}$ are the variance parameter estimates from the LME model. Conditioning on a correctly specified mean structure; i.e., $\boldsymbol{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$, an unbiased estimator of $var(\boldsymbol{Y})$ is the smoothed sample covariance of the marginal residuals; i.e., $\hat{\boldsymbol{V}} = \hat{\boldsymbol{e}}\hat{\boldsymbol{e}}^T/(m-1)$, where $\hat{\boldsymbol{e}} = \{\hat{\boldsymbol{e}}_1, ..., \hat{\boldsymbol{e}}_m\}$ is a $n \times m$ matrix combining all m subjects' marginal residuals, and $\hat{\boldsymbol{e}}_i = \{\hat{e}_{i1}, ..., \hat{e}_{in}\}'$ is a vector containing all n marginal residuals of subject i (assuming all subjects have the same number of measurements). Inspired by the informal check suggested by Verbeke and Molenberghs [2009], we propose to define our test statistics as metrics that quantify the discrepancy between the estimated model-based covariance $\hat{\boldsymbol{W}}$ and the empirical covariance $\hat{\boldsymbol{V}}$.

We first compare the two estimated covariance matrices by examining the multiplication $\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{V}}$, and define our first two test statistics as:

$$T_1 = \sum_{k=1}^{n} (c_k - 1)^2, \tag{4.7}$$

$$T_2 = \sum_{k=1}^{n} (log(c_k))^2, \tag{4.8}$$

where the c_k are the eigenvalues of matrix $\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{V}}$. Rotnitzky and Jewell [1990] proved that all elements of c_k should equal to 1 when $\boldsymbol{W} = var(\boldsymbol{Y})$. Thus, we will reject our null hypothesis H_0 if T_1 and T_2 deviate much from 0.

Covariance structure analysis is also used in structural equations modeling (SEM), and one intuitive index is the standardized element-wise difference between the estimated model-based covariance and the sample residual covariance [Hu and Bentler, 1999]. This index motivates us to define our third test statistic as:

$$T_3 = \sum_{i=1}^{n}\sum_{j=1}^{n} |\hat{w}_{ij} - \hat{v}_{ij}|, \tag{4.9}$$

which is the L^1-norm of the difference matrix $\hat{\boldsymbol{W}} - \hat{\boldsymbol{V}}$. It is easy to see that $T_3 = 0$ when $\boldsymbol{W} = var(\boldsymbol{Y})$. Thus, we will reject our null hypothesis H_0 if T_3 deviates much from 0.

Despite the intuitive nature of our three test statistics, it is not straightforward to derive their exact or asymptotic distributions. Instead of seeking analytical solutions, we propose to estimate their distributions numerically using permutations. A permutation test is one that approximates the null distribution of the test statistic via permutations of the data. The test will have a nominal size as long as we have exchangeability of the data under the null hypothesis. For a vector $\boldsymbol{Y}$, it is exchangeable if, for any permutation of $\boldsymbol{Y}$, denoted as $\boldsymbol{Y}^*$, the distribution of $\boldsymbol{Y}^*$ is the same as that of $\boldsymbol{Y}$. Good [2006] provides a comprehensive explanation of permutations tests.

We now give a detailed explanation of how to perform our permutation tests for the covariance structure assumption in LME models. After a LME model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\epsilon}$ is fitted, we first obtain the estimated model-based covariance matrix as $\hat{\boldsymbol{W}} = \boldsymbol{Z}^T\hat{\boldsymbol{D}}\boldsymbol{Z} + \hat{\boldsymbol{R}}$, where $\hat{\boldsymbol{D}}$ and $\hat{\boldsymbol{R}}$ are the variance component estimates from the fitted LME model. We also estimate the marginal residuals from $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and calculate their smoothed sample covariance matrix as $\hat{\boldsymbol{V}} = \hat{\boldsymbol{e}}\hat{\boldsymbol{e}}^T/(m-1)$. We then calculate our three test statistics $T_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{V}})$, $T_2(\hat{\boldsymbol{W}}, \hat{\boldsymbol{V}})$, and $T_3(\hat{\boldsymbol{W}}, \hat{\boldsymbol{V}})$, based

on the estimated $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{V}}$.

In order to construct the permutation distributions, we would like to permute the marginal errors, i.e., $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$. However, as indicated by Equation 4.1, conditioning on a correctly specified mean structure, under the null hypothesis H_0, the marginal errors are normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{W} = \boldsymbol{Z}^T\boldsymbol{D}\boldsymbol{Z} + \boldsymbol{R}$. Therefore the marginal errors are not immediately exchangeable. To solve this issue, we propose to weight the errors by the matrix $\boldsymbol{L}^{-1}$, where $\boldsymbol{L}$ is the Cholesky decomposition of $\boldsymbol{W}$, i.e., $\boldsymbol{W} = \boldsymbol{L}\boldsymbol{L}^T$. As a result, the Cholesky marginal errors, $\boldsymbol{L}^{-1}\boldsymbol{\epsilon}$ become normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{I}$, and thereby are exchangeable, allowing for permutations both within and among subjects.

In practice, we work with the marginal residuals, $\hat{\boldsymbol{e}}$ and the estimated model-based covariance matrix, $\hat{\boldsymbol{W}}$, and permute the Cholesky marginal residuals, i.e., $\hat{\boldsymbol{L}}^{-1}\hat{\boldsymbol{e}}$, where $\hat{\boldsymbol{W}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T$, both within and among subjects. Let $(\hat{\boldsymbol{L}}^{-1}\hat{\boldsymbol{e}})^*$ denote the permuted Cholesky marginal residuals; we further re-weight them with $\hat{\boldsymbol{L}}$, and obtain $\hat{\boldsymbol{L}}(\hat{\boldsymbol{L}}^{-1}\hat{\boldsymbol{e}})^*$. Under the null hypothesis H_0, the smoothed sample covariance of $\hat{\boldsymbol{L}}(\hat{\boldsymbol{L}}^{-1}\hat{\boldsymbol{e}})^*$, i.e., $\hat{\boldsymbol{V}}^* = [\hat{\boldsymbol{L}}(\hat{\boldsymbol{L}}^{-1}\hat{\boldsymbol{e}})^*][\hat{\boldsymbol{L}}(\hat{\boldsymbol{L}}^{-1}\hat{\boldsymbol{e}})^*]^T/(m-1)$ should be identical to the estimated model-based covariance matrix $\hat{\boldsymbol{W}}$. Therefore, we calculate our permuted test statistics based on $\hat{\boldsymbol{V}}^*$ and $\hat{\boldsymbol{W}}$, and obtain $T_1(\hat{\boldsymbol{W}}, \hat{\boldsymbol{V}}^*)$, $T_2(\hat{\boldsymbol{W}}, \hat{\boldsymbol{V}}^*)$, and $T_3(\hat{\boldsymbol{W}}, \hat{\boldsymbol{V}}^*)$, respectively.

As recommended by Good [2006], we perform the permutation procedure $1,000$ times, and obtain $1,000$ permuted values for each test statistic, i.e., T^*_{1l}, T^*_{2l}, T^*_{3l}, $l = 1, ..., 1000$. These $1,000$ permuted values provide an approximate empirical null distribution for each test statistic. Then for each test, we generate its p-value by counting the percentage of permutations with permuted values T^* greater than T,

e.g., $p_1 = \sum_{l=1}^{1000} I(T_1 < T_{1l}^*)/1000$ for Test 1, where $I(T_1 < T_{1l}^*) = 1$ if $T_1 < T_{1l}^*$.

4.3 Simulation Studies

We performed a series of simulations to study the performance of our permutation tests in examining different components of the covariance structure assumption. The first simulation tested the inclusion or exclusion of random effects; and the following two simulations tested the appropriateness of assumptions on random errors. In each simulation, we generated data with a known covariance, and fitted LME models under correct or incorrect covariance structure assumptions. We verified the validity of our permutation tests when the covariance structure was correctly specified; and evaluated the power of our permutation tests in detecting incorrectly specified covariance structures. Each simulation was repeated $1,000$ times. All simulations were performed in the *R* system using the *lme()* function from the *nlme* package [Pinheiro et al., 2015].

4.3.1 Testing for a Random Slope

Testing for the inclusion or exclusion of random effects is arguably the most commonly encountered situation in fitting LME models. This can also be expressed as examining the appropriateness of covariance matrix $\boldsymbol{D}$ in LME model (4.1). As a special case, here we present simulations testing for the inclusion of a random slope given an independent random intercept. We considered situations when the random effect covariates were the same for all subjects or varied among subjects.

Measurements Occur at the Same Time Points

The data set was generated from an LME model with a random intercept and a possible random slope $Y_{ij} = \beta_0 + x_{ij1}\beta_1 + b_{i0} + z_{ij1}b_{i1} + \epsilon_{ij}$, where the fixed effect coefficients $\beta_0 = 3$, $\beta_1 = 2.75$, fixed effect covariate $x_{ij1} \sim \mathcal{N}(0, 1)$, random effects

$b_{i0} \sim \mathcal{N}(0, \sigma_0^2)$, $b_{i1} \sim \mathcal{N}(0, \sigma_1^2)$, and error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. The random effect covariate $z_{ij1} = z_{j1}$ for all subjects, and $z_{j1} \sim \mathcal{N}(0, 1)$. Thus, this design could be viewed as a longitudinal study where observations were measured at the same time points for all subjects. Both x_{ij1} and z_{ij1} were centered at 0 and scaled by their standard errors. Here we set $\sigma_0^2 = 1$, $\sigma^2 = 1$, and let $\sigma_1^2 = 0, 0.15^2, 0.2^2, 0.3^2$, respectively. We varied the number of subjects to be either $n = 50$ or $n = 10$, and the number of measurements per subject to be $m = 10$ or $m = 5$.

We fit an LME model with a random intercept only. When $\sigma_1^2 = 0$, the fitted model had correctly specified the covariance structure, and we verified the validity of our permutation tests; when $\sigma_1^2 = 0.15^2, 0.2^2, 0.3^2$, the covariance structure implied by the fitted model was different from the truth, and we evaluated the power of our tests. The rejection rates of our permutation tests over the $1,000$ simulations are presented in Table 4.1.

Based on this simulation, we find that our permutation Test 1 (4.7) and Test 2 (4.8) are valid in testing the random intercept, as they have nominal size when H_0 holds. However, these two tests are more conservative when $m = 5$ than when $m = 10$. When σ_1^2 increases to 0.15^2, 0.2^2, and 0.3^2, the covariance structure of the fitted model gradually deviates from the truth, and the power of Tests 1 and 2 increase as well. Test 1 seems to be more powerful than Test 2 over all scenarios. As expected, the power of these two permutation tests decrease as fewer subjects are included in the study. When the number of subjects equals 10 or 5, Test 2 has very limited power. The performance of permutation Test 3 (4.9) is disappointing, as it has neither a valid size nor any power.

Table 4.1: Testing for a random slope when measurements occur at the same time points for all subjects. Rejection rates (expressed as percentages) of our permutation tests (at 5% level) over $1,000$ simulations.

n	m	σ_1^2	Test 1	Test 2	Test 3
50	10	0	4.3	3.8	0.0
		0.15^2	73.4	37.5	0.1
		0.2^2	91.0	59.9	0.4
		0.3^2	99.1	86.3	5.2
50	5	0	2.3	2.6	0.0
		0.15^2	28.8	19.9	0.4
		0.2^2	47.3	32.9	0.8
		0.3	77.6	62.0	4.2
10	10	0	5.5	5.7	0.2
		0.15^2	16.0	5.1	0.6
		0.2^2	23.3	4.9	1.0
		0.3^2	42.9	5.7	2.5
10	5	0	3.6	4.0	0.1
		0.15^2	6.3	5.9	0.9
		0.2^2	8.0	5.9	0.9
		0.3^2	14.2	7.4	2.2

Measurements Occur at Different Time Points

In real longitudinal studies, observations are rarely measured at the exactly same time points for different subjects. Instead, measurements may occur at slightly different time points that vary among subjects. To mimic this situation, we designed a simulation study to test for a random slope, when the random effect covariate varied by subjects.

Similar to Section 4.3.1, the data set was also generated from an LME model with a random intercept and a possible random slope $Y_{ij} = \beta_0 + x_{ij1}\beta_1 + b_{i0} + z_{ij1}b_{i1} + \epsilon_{ij}$, where the fixed effect coefficients $\beta_0 = 3$, $\beta_1 = 2.75$, fixed effect covariate $x_{ij1} \sim \mathcal{N}(0,1)$, random effects $b_{i0} \sim \mathcal{N}(0, \sigma_0^2)$, $b_{i1} \sim \mathcal{N}(0, \sigma_1^2)$, and error $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. For each subject i, random effect covariate $z_{ij1} \sim \mathcal{N}(\mu_j, \sigma_z^2)$, where $(\mu_1, ..., \mu_j, ..., \mu_m)$ is a sequence of numbers increasing from $-(m-1)/2$ to $(m-1)/2$ by 1. Both x_{ij1} and z_{ij1} were centered at 0 and scaled by their standard errors. We considered three scenarios with $\sigma_z^2 = 0.2^2$, 1 and 5^2, respectively. Under each scenario, we set $\sigma_0^2 = 1$,

$\sigma^2 = 1$, and let $\sigma_1^2 = 0, 0.15^2$. We also varied the number of subjects to be either $n = 50$ or $n = 10$, and the number of measurements per subject to be $m = 10$ or $m = 5$.

We again fit an LME model with a random intercept only. When $\sigma_1^2 = 0$, the fitted model had correctly specified the covariance structure, and we verified the validity of our permutation tests; when $\sigma_1^2 = 0.15^2$, the covariance structure implied by the fitted model were different from the truth, and we evaluated the power of our permutation tests. The rejection rates of our permutation tests over the $1,000$ simulations are presented in Table 4.2.

Across all situations, the rejection rates of permutation Test 1 and Test 2 are around 5% when $\sigma_1^2 = 0$. Thus we confirm that these two tests are valid in testing for the random intercept, even when measurements occur at different time points. In addition, these tests are more conservative when $m = 5$ than when $m = 10$. When $\sigma_z^2 = 0.2^2$, the time points of measurements only vary slightly for different subjects, and the power of Tests 1 and 2 is still prominent. As σ_z^2 enlarges, measurements time points are more significantly different over subjects. As a result, the power of Tests 1 and 2 diminishes. We hardly see any power when $\sigma_z^2 = 5^2$. We also notice decreased power with fewer subjects, which is expected. In general, Test 1 is has more power than Test 2, and permutation Test 3 is not valid.

4.3.2 Testing for Serial Correlations among Random Errors

In longitudinal studies, measurements on the same subject are usually correlated via serial correlations, e.g., autoregressive (AR). An LME model allows accounting for the serial correlation by introducing correlated random errors through the $\boldsymbol{R}$ matrix in LME model (4.1). Here we designed simulations to evaluate the performance of our permutation tests in examining serial correlations.

Table 4.2: Testing for a random slope when measurements occur at different time points. Rejection rates (expressed as percentages) of our permutation tests (at 5% level) over $1,000$ simulations.

σ_z^2	n	m	σ_1^2	Test 1	Test 2	Test 3
0.2^2	50	10	0	4.8	5.4	0.0
			0.15^2	75.0	39.1	50.0
	50	5	0	3.1	4.2	0.0
			0.15^2	27.6	18.8	17.6
	10	10	0	4.2	4.9	0.0
			0.15^2	13.3	4.9	8.5
	10	5	0	3.6	5.1	0.3
			0.15^2	6.8	5.7	4.5
1	50	10	0	4.0	4.1	0.0
			0.15^2	64.8	28.7	37.7
	50	5	0	4.1	4.0	0.0
			0.15^2	12.9	11.2	7.5
	10	10	0	4.1	4.8	0.2
			0.15^2	13.0	5.6	7.5
	10	5	0	3.4	3.7	0.2
			0.15^2	5.0	3.2	3.3
5^2	50	10	0	4.6	3.7	0.0
			0.15^2	8.9	6.0	3.8
	50	5	0	3.2	3.2	0.0
			0.15^2	4.3	4.3	2.9
	10	10	0	4.7	4.2	0.0
			0.15^2	7.0	5.4	3.6
	10	5	0	3.4	3.2	0.1
			0.15^2	3.4	4.2	3.5

The data set was generated from an LME model with a random intercept and possibly correlated random errors $Y_{ij} = \beta_0 + x_{ij1}\beta_1 + b_{i0} + \epsilon_{ij}$, where the fixed effect coefficients $\beta_0 = 3$, $\beta_1 = 2.75$, fixed effect covariate $x_{ij1} \sim \mathcal{N}(0, 1)$, and it was centered at 0 and scaled by its standard error, random effect $b_{i0} \sim \mathcal{N}(0, \sigma_0^2)$. For each subject i, its random errors $(\epsilon_{i1}, ..., \epsilon_{ij}, ...\epsilon_{ik}, ..., \epsilon_{im})'$ followed a multivariate normal distribution with mean 0 and AR-1 covariance; i.e., $var(\epsilon_{ij}) = \sigma^2$, $corr(\epsilon_{ij}, \epsilon_{ik}) = \rho^{|j-k|}$. We set $\sigma_0^2 = 1$, $\sigma^2 = 1$, and let $\rho = 0, 0.2, 0.3, 0.4, 0.6$, respectively. We again varied the number of subjects to be either $n = 50$ or $n = 10$, and the number of measurements per subject to be $m = 10$ or $m = 5$.

We fit an LME model with a random intercept and independent errors. When $\rho = 0$, the fitted model had correctly specified the covariance structure, and we ver-

ified the validity of our permutation tests; when $\rho = 0.2, 0.3, 0.4, 0.6$, the covariance structure implied by the fitted model were different from the truth, and we evaluated the power of our permutation tests. The rejection rates of our permutation tests over the $1,000$ simulations are presented in Table 4.3.

In this simulation, our permutation Test 1 and Test 2 have nominal size when $\rho = 0$, i.e., H_0 holds. Thus these two tests are valid for testing serial correlations among random errors. However, these two tests are generally more conservative when $m = 5$ than when $m = 10$. As ρ increases, the serial correlation becomes stronger in the data, and Tests 1 and 2 have a higher chance to detect it. The power of these two tests is affected by the sample size: having fewer subjects would reduce the power. Test 1 has slightly higher power than Test 2 in general. Permutation Test 3 is not valid again.

Table 4.3: Testing for serial correlations among random errors. Rejection rates (expressed as percentages) of our permutation tests (at 5% level) over $1,000$ simulations.

n	m	ρ	Test 1	Test 2	Test 3
50	10	0	4.3	3.8	0.0
		0.2	32.6	25.1	0.0
		0.3	76.6	66.1	0.0
		0.4	97.3	95.3	0.0
		0.6	100.0	100.0	0.2
50	5	0	3.7	3.4	0.0
		0.2	15.4	15.0	0.0
		0.3	37.8	32.0	0.0
		0.4	63.6	57.4	0.0
		0.6	97.2	96.4	0.0
10	10	0	2.8	6.3	0.0
		0.2	9.5	7.1	0.2
		0.3	16.5	5.6	0.2
		0.4	24.5	6.0	0.2
		0.6	61.4	6.8	0.3
10	5	0	4.0	3.7	0.1
		0.2	5.3	4.1	0.5
		0.3	8.7	5.2	0.5
		0.4	11.4	7.9	0.6
		0.6	24.9	15.0	0.3

4.3.3 Testing for Heterogeneous Random Errors

Another commonly encountered situation in longitudinal studies is that measurements observed at different time points may have different variances. For example, the machine used for collecting data is gradually worn over time, and the variability of the measurements increases. An LME model can take such variability into account by allowing for heterogeneous random errors; i.e., the diagonal elements of matrix $\boldsymbol{R}$ in LME model (4.1) may differ among measurements. Here we designed simulations to evaluate the performance of our permutation tests in diagnosing heterogeneous errors.

The data set was generated from an LME model with a random intercept and possibly heterogeneous random errors $Y_{ij} = \beta_0 + x_{ij1}\beta_1 + b_{i0} + \epsilon_{ij}$, where the fixed effect coefficients $\beta_0 = 3$, $\beta_1 = 2.75$, fixed effect covariate $x_{ij1} \sim \mathcal{N}(0, 1)$, and it was centered at 0 and scaled by its standard error, random effect $b_{i0} \sim \mathcal{N}(0, \sigma_0^2)$. For each subject i, its random errors $(\epsilon_{i1}, ..., \epsilon_{ij}, ..., \epsilon_{im})'$ followed a multivariate normal distribution with mean 0 and diagonal covariance matrix; i.e., $var(\epsilon_{ij}) = \sigma^2 h_j$, where $(h_1, ..., h_j, ..., h_m)$ is a sequence of numbers increasing from h to 1 by $(1-h)/(m-1)$. We set $\sigma_0^2 = 1$, $\sigma^2 = 1$, and let $h = 1, 0.7, 0.5, 0.3, 0.1$, respectively. We varied the number of subjects to be either $n = 50$ or $n = 10$, and the number of measurements per subject to be $m = 10$ or $m = 5$ as before.

We fit an LME model with a random intercept and homogeneous errors. When $h = 1$, the fitted model had correctly specified the covariance structure, and we verified the validity of our permutation tests; when $h = 0.7, 0.5, 0.3, 0.1$, the covariance structure implied by the fitted model were different from the truth, and we evaluated the power of our permutation tests. The rejection rates of our permutation tests over the $1,000$ simulations are presented in Table 4.4.

In this simulation, permutation Test 1 and Test 2 have nominal size when the null hypothesis H_0 holds. Thus these two tests are valid for testing heterogeneous random errors. However, these two tests are more conservative when $m = 5$ than when $m = 10$. As the heterogeneity parameter range enlarges, random errors have more variability, and Tests 1 and 2 have higher chance in identifying this heterogeneous structure. The power of these two tests decreases with sample size. Unlike previous simulations, neither Test 1 nor Test 2 has universally higher power. Permutation Test 3 is not valid as before.

Table 4.4: Testing for heterogeneous random errors. Rejection rates (expressed as percentages) of our permutation tests (at 5% level) over $1,000$ simulations.

n	m	Heterogeneity Parameter Range	Test 1	Test 2	Test 3
50	10	[1, 1]	5.7	4.3	0.0
		[0.7, 1]	6.8	6.9	0.0
		[0.5, 1]	18.9	16.6	0.0
		[0.3, 1]	58.2	55.2	0.0
		[0.1, 1]	98.7	100.0	0.0
50	5	[1, 1]	3.3	3.0	0.0
		[0.7, 1]	6.1	5.9	0.0
		[0.5, 1]	15.4	15.8	0.0
		[0.3, 1]	44.4	51.5	0.0
		[0.1, 1]	95.1	98.9	0.0
10	10	[1, 1]	4.3	4.2	0.0
		[0.7, 1]	3.6	5.7	0.0
		[0.5, 1]	5.1	6.1	0.0
		[0.3, 1]	8.5	5.0	0.0
		[0.1, 1]	16.2	7.3	0.0
10	5	[1, 1]	3.4	3.1	0.2
		[0.7, 1]	3.6	4.2	0.4
		[0.5, 1]	4.5	5.0	0.4
		[0.3, 1]	8.2	6.4	0.2
		[0.1, 1]	12.6	13.8	0.2

In summary, through these simulation studies, we show that our permutation Test 1 and Test 2 have nominal size in testing different components, i.e., random effects and/or random errors, of the covariance structure assumption in LME models. However, these two tests are generally more conservative when $m = 5$ than when

$m = 10$. The power of Test 1 and 2 are sufficient. Permutation Test 3 seems to be invalid.

4.4 Application to Michigan Periodontal Study

Here we apply our permutation tests to evaluate the two LME models developed for analyzing periodontal data presented in Chapter III. It is challenging to model periodontal outcomes due to the complex within-mouth correlation induced by the three-dimensional spatial geography of teeth and their functional similarity. Thus we have proposed two LME models with random effects and correlated random errors that quantify the within-mouth correlation of teeth. However, we were not able to find proper statistical tests to evaluate the fit of our LME models. In this application, we fit the two LME models to the Michigan periodontal data, and assessed their covariance structure assumptions using our permutation tests.

The data set contained clinical attachment level (CAL), a tooth-level measure that quantifies the severity of periodontal disease, of $2,646$ teeth collected from 50 periodontally healthy and 50 periodontally diseased subjects. The goal of the study was to compare the difference in CALs between periodontally healthy and diseased subjects. As an illustration, we focused on the 38 periodontally healthy and 9 diseased subjects who had complete CALs on all 28 teeth.

The first proposed LME model used random effects to account for the within-mouth variation between the maxillary and mandibular arches, and the functional variation between different types of teeth (molar, bicuspid, cuspid and incisor); it also modeled the spatial proximity of teeth via circularly correlated random errors. We called this model the functional and spatial LME 1 (3.1). The second proposed LME model used random effects to model the natural symmetry between the four

quadrants; and it employed circularly correlated heterogeneous random errors to account for the extra variability. This model was called the quadrant and spatial LME 2 (3.2). Please refer to Chapter III for details about these two models.

We fit the functional and spatial LME 1 and the quadrant and spatial LME 2, with a fixed effect intercept and a fixed effect binary indicator for periodontal disease status. Then we applied our permutation tests to examine the covariance assumptions of the two fitted models. The null hypothesis was H_0: "The covariance matrix implied by LME 1 (or 2) is identical to the true covariance of CALs". The p-values of our permutation tests are presented in Table 4.5. For both LME 1 and LME 2, the p-values of our permutation Test 1 and 2 are < 0.01, which indicate strong rejections of the null. Thus LME 1 and LME 2 did not correctly model the within-mouth correlation for subjects in the Michigan periodontal study, and the standard errors of the fixed effects estimates from these LME models might be biased. However, we should not over interpret these p-values, as they only reflect the appropriateness of LME 1 and LME 2 in modeling these 47 subjects, rather than overall evaluations of the merits of the two LME models.

Table 4.5: p-values from applying our permutation tests to evaluate the covariance structure assumption of the functional and spatial LME 1, and the quadrant and spatial LME 2 fitted to Michigan data.

Fitted Model	Test 1	Test 2
Functional and Spatial LME 1 (3.1)	< 0.01	< 0.01
Quadrant and Spatial LME 2 (3.2)	< 0.01	< 0.01

One limitation of this application is that we are unclear whether the mean structure of the LME models are specified correctly. It is very likely that there are other influential fixed effects, e.g., gender and age, that have not been included in the analysis. Therefore, we proposed a small simulation study to compare the functional and spatial LME 1 to the quadrant and spatial LME 2. We generated 50 periodontally

healthy and 50 diseased subjects with known mean profile and known covariance matrices. Two covariance matrices were considered, one was based on the functional and spatial LME 1; and the other one was based on the quadrant and spatial LME 2. For each simulated data set, we fitted the two LME models, and applied our permutation tests to evaluate the covariance assumption of the fitted models. The simulations were repeated 200 times. The rejection rates of our permutation tests over the 200 simulations are presented in Table 4.6.

When the correct model is fitted to the simulated data, its covariance structure assumption is satisfied, and our permutation Test 1 and Test 2 are unlikely to reject the null hypothesis. In the contrast, if the fitted model is different from the model used for generating data, Test 1 and 2 will almost always reject the fitted model. This simulation convinces that our permutation Test 1 and 2 are valid for evaluating rather complex covariance structures in LME models.

Table 4.6: Applying permutation tests to evaluate the covariance structure assumption of the fitted functional and spatial LME 1, and the quadrant and spatial LME 2 when data is generated from a known model. Rejection rates (expressed as percentages) of our permutation tests (at 5% level) over 200 simulations.

True Model	Fitted Model	Test 1	Test 2
Functional and Spatial LME 1 (3.1)	Functional and Spatial LME 1 (3.1)	2.0	2.5
	Quadrant and Spatial LME 2 (3.2)	100.0	100.0
Quadrant and Spatial LME 2 (3.2)	Functional and Spatial LME 1 (3.1)	99.5	100.0
	Quadrant and Spatial LME 2 (3.2)	3.0	4.5

4.5 Discussion

In this chapter, we have proposed three permutation tests for examining the covariance structure assumption in linear mixed effects models. Our methods are eligible for testing different components of the covariance structure in an LME model by comparing the estimated model-based covariance matrix to the smoothed sample covariance matrix of the marginal residuals. To the best of our knowledge, our

permutation tests are the first methods that provide formal statistical inference on the overall appropriateness of covariance structure in LME models. Through simulations, we have seen that our permutation Test 1 and Test 2 have valid size and sufficient power. Our methods can be easily implemented in standard statistical software and it has an immediate extension to other models such as structural equation modeling.

Through simulations, we have seen that Permutation Test 3 has neither valid size nor sufficient power to be useful. One possible explanation is that test statistic T_3 in Equation (4.9) is an element-wise comparison of the two estimated covariance matrices, and the true difference could be overwhelmed by the level of noise associated with the estimates of all the individual elements of the matrices. In addition, Permutation Test 1 appears to be more powerful than Test 2 in most settings, even though test statistics T_1 and T_2, in Equations (4.7) and (4.8), respectively, are computed from the same set of eigenvalues. We think the difference in power is possibly due to computational reasons, as the value of $(\log(c_k))2$ will be extremely large for any eigenvalue c_k close to 0. Thus, relative to a permutation test based upon T_1, small eigenvalues could overwhelm the computation of test statistic T_2 across permutations, leading to a permutation distribution that is less variable than desired, thereby reducing the power of the permutation test. Future research is needed to explore the differences among these three test statistics, which might help us to identify the differences in their operating characteristics, as well as propose additional permutation tests with greater power.

One limitation of our methods is that we have assumed the same number of measurements for each subject, which is required by our estimation of the empirical covariance $\hat{\boldsymbol{V}}$. In addition, our methods assume a common estimated model-based

covariance matrix for all subjects, which has restricted their application to situations when random effect covariates vary largely by subjects. One possible approach to eliminate this limitation is to average the estimated subject-specific model-based covariance matrices over all subjects; and use the resulting smoothed model-based covariance matrix in calculating our test statistics.

We have assumed that the mean structure of the fitted LME model is correct. It will provide more insights on our permutation tests by evaluating their performance under situations when the mean profile is not modeled correctly. In addition, we have assumed normal distributed random effects and random errors when we performed our simulations. However, our test statistics only rely on the moments, rather than the full distributions. Therefore, it will be beneficial to perform sensitivity analysis of our permutation tests to nonnormal random effects and/or errors. Finally, extending our permutation tests to generalized linear mixed models may be rewarding.

CHAPTER V

Conclusion and Future Work

In this dissertation we have developed three methods for handling correlated data.

In Chapter II, we extended the standard classification and regression trees (CART) method to clustered binary outcomes. As opposed to the conventional CART, we propose to build tree models using the residuals from a null generalized linear mixed model (GLMM) as the outcome. This circumvents modeling the correlation structure explicitly while still accounting for the cluster-correlated design, thereby allowing us to adopt the original CART machinery in tree growing, pruning and cross-validation. Class predictions for the terminal nodes of our residual-based tree are estimated based on success probabilities within each terminal node. We also provide a natural and direct extension of our residual-based tree to random forest.

Through extensive simulation studies, we have shown that our residual-based trees, especially the deviance residual-based tree, are more appropriate for analyzing clustered binary data than the standard CART. The residual-based trees are better adept in identifying the true relationship in the data, and provide more accurate predictions. The improvements over the standard CART are substantial when the intra-cluster correlations are strong, given moderate cluster sizes. We also applied our residual-based approaches to studies of kidney cancer treatment receipt, surgical

mortality after colectomy and determinants of vaccination coverage, where the data exhibited cluster-correlated structures. In all studies, residual-based tree and forest identified clinically meaningful subgroups.

One caveat of our approach is that when fitting the null GLMMs, at least moderate cluster sizes are needed in order to correctly estimate the cluster-specific random effects. When the cluster sizes are small, the estimated random effects might be biased, which in turn could affect the performance of our residual-based trees. It will be beneficial to find other algorithms that could reduce the bias in estimating the random effects of GLMMs under small cluster sizes.

In Chapter III, we have proposed two linear mixed effects (LME) models for tooth-level periodontal outcomes, which can account for the complex within-mouth correlation via the usage of random effects and random errors. Through simulations, we have shown that our LME models are more robust to "missing at random", and more efficient than traditional methods such as GEE and t-tests in periodontal analysis. We have also suggested model selection criteria for choosing the LME model that better fits the data. The proposed LME models and the selection criteria can be conveniently implemented in standard software packages, which makes them readily accessible to periodontal researchers.

Longitudinal data are common in periodontal studies, where each tooth are monitored repeatedly over time. This temporal effect, along with the within mouth correlation, will induce even more complex correlation structures. Further research could be conducted to generalize our LME models to longitudinal periodontal outcomes. In addition, periodontal disease is a leading cause of loss tooth, and teeth with larger periodontal outcomes have a higher chance of being removed. Therefore, informative missing, i.e., MNAR is inevitable in periodontal studies. Through the

simulations, we have seen that our LME models are biased and less efficient when data are MNAR. Joint modeling of missing teeth and periodontal outcomes rises as an interesting and rewarding direction for future studies.

In Chapter IV, we have proposed three permutation tests for evaluating the covariance structure in linear mixed effects models. Our methods are among the first few efforts to provide formal statistical inferences on the appropriateness of the covariance structure implied by an LME model. Through simulations, we have shown that our permutation Test 1 and Test 2 have valid size and comparable power in testing different covariance structure assumptions. We also applied our tests to the Michigan periodontal study and evaluated the two LME models proposed in Chapter III. We confirmed that our permutation tests can identify the LME model that has accurately modeled the within mouth correlation of periodontal outcomes.

Our permutation tests assume a common covariance structure for all subjects. Thus our methods are restricted to balanced data or situations when only slight variations are allowed in random effect covariates. One possible approach to eliminate these limitations is to average the estimated subject-specific model-based covariance over all subjects, and use the resulting smoothed covariance in calculating our test statistics. Future research could be conducted to examine this possible solution.

In addition, sensitivity analysis under misspecified mean structure or nonnormal random effects and/or random errors would provide a more comprehensive assessment of our permutation tests. Finally, it will be rewarding to extend our permutation tests to generalized linear mixed models.

BIBLIOGRAPHY

Bibliography

M. Abdolell, M. LeBlanc, D. Stephens, and R. V. Harrison. Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Statistics in Medicine*, 21(22):3395–3409, 2002.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

M. Arora, J. L. Weuve, J. D. Schwartz, and R. O. Wright. Association of environmental cadmium exposure with periodontal disease in us adults. *Environmental Health Perspectives*, 117(5):739–744, 2009.

B. Axtelius, B. Söderfeldt, and R. Attström. A multilevel analysis of factors affecting pocket probing depth in patients responding differently to periodontal treatment. *Journal of Clinical Periodontology*, 26(2):67–76, 1999.

M. Banerjee, J. George, E. Y. Song, A. Roy, and W. Hryniuk. Tree-based model for breast cancer prognostication. *Journal of Clinical Oncology*, 22(13):2567–2575, 2004.

M. Banerjee, Y. Ding, and A.-M. Noone. Identifying representative trees from ensembles. *Statistics in Medicine*, 31(15):1601–1616, 2012.

M. Banerjee, C. Filson, R. Xia, and D. C. Miller. Logic regression for provider effects on kidney cancer treatment delivery. *Computational and Mathematical Methods in Medicine 2014*, 2014.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

V. J. Carey and Y.-G. Wang. Working covariance model selection for generalized

estimating equations. *Statistics in Medicine*, 30(26):3117–3124, 2011.

Z. Chen and D. B. Dunson. Random effects selection in linear mixed models. *Biometrics*, 59(4):762–769, 2003.

S. A. Dowsett, G. J. Eckert, and M. J. Kowolik. The applicability of half-mouth examination to periodontal disease assessment in untreated adult populations. *Journal of Periodontology*, 73(9):975–981, 2002.

R. Drikvandi, G. Verbeke, A. Khodadadi, and V. P. Nia. Testing multiple variance components in linear mixed-effects models. *Biostatistics*, 14(1):144–159, 2013.

P. I. Eke, B. A. Dye, L. Wei, G. O. Thornton-Evans, and R. J. Genco. Prevalence of periodontitis in adults in the united states: 2009 and 2010. *Journal of Dental Research*, 91(10):914–920, 2012.

L. J. Emrich. Common problems with statistical aspects of periodontal research papers. *Journal of Periodontology*, 61(4):206–208, 1990.

G. M. Fitzmaurice, S. R. Lipsitz, and J. G. Ibrahim. A note on permutation tests for variance components in multilevel generalized linear mixed models. *Biometrics*, 63 (3):942–946, 2007.

C. R. Friese, R. Xia, A. Ghaferi, J. D. Birkmeyer, and M. Banerjee. Hospitals in 'magnet' program show better patient outcomes on mortality measures compared to non-'magnet' hospitals. *Health Affairs*, 34(6):986–992, 2015.

A. Gałecki and T. Burzykowski. *Linear mixed-effects models using R: A step-by-step approach*. Springer, 2013.

R. J. Genco and W. S. Borgnakke. Risk factors for periodontal disease. *Periodontology 2000*, 62(1):59–94, 2013.

A. A. Ghaferi, J. D. Birkmeyer, and J. B. Dimick. Variation in hospital mortality associated with inpatient surgery. *New England Journal of Medicine*, 361(14): 1368–1375, 2009.

P. I. Good. *Permutation, parametric, and bootstrap tests of hypotheses*. Springer Science & Business Media, 2006.

A. Hajjem, F. Bellavance, and D. Larocque. Mixed effects regression trees for clustered data. *Statistics & Probability Letters*, 81(4):451–459, 2011.

S. K. Harrel and M. E. Nunn. Longitudinal comparisons of the periodontal status of patients with moderate to severe periodontal disease receiving no treatment, nonsurgical treatment, and surgical treatment utilizing individual sites for analysis.

Journal of Periodontology, 72(11):1509–1519, 2001.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

M. R. Haymart, M. Banerjee, A. K. Stewart, R. J. Koenig, J. D. Birkmeyer, and J. J. Griggs. Use of radioactive iodine for thyroid cancer. *JAMA*, 306(7):721–728, 2011.

C. R. Henderson. Estimation of genetic parameters. *Biometrics*, 6(2):186–187, 1950.

B. K. Hollenbeck, D. A. Taub, D. C. Miller, R. L. Dunn, and J. T. Wei. National utilization trends of partial nephrectomy for renal cell carcinoma: a case of underutilization? *Urology*, 67(2):254–259, 2006.

E. A. Houseman, L. M. Ryan, and B. A. Coull. Cholesky residuals for assessing normal errors in a linear model with correlated outcomes. *Journal of the American Statistical Association*, 99(466):383–394, 2004.

L.-T. Hu and P. M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55, 1999.

L. I. Iezzoni. *Risk adjustment for measuring healthcare outcomes*. Health Administration Press, 2012.

H. Jacqmin-Gadda, S. Sibillot, C. Proust, J.-M. Molina, and R. Thiébaut. Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics & Data Analysis*, 51(10):5142–5154, 2007.

S. Keon Lee. On generalized multivariate decision tree by using gee. *Computational Statistics & Data Analysis*, 49(4):1105–1119, 2005.

J. S. Kinney, T. Morelli, T. Braun, C. A. Ramseier, A. E. Herr, J. V. Sugai, C. E. Shelburne, L. A. Rayburn, A. K. Singh, and W. V. Giannobile. Saliva/pathogen biomarker signatures and periodontal disease progression. *Journal of Dental Research*, 90(6):752–758, 2011.

S. K. Kinney and D. B. Dunson. Fixed and random effects selection in linear and logistic models. *Biometrics*, 63(3):690–698, 2007.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

N. Lange and N. M. Laird. The effect of covariance structure on variance estimation in balanced growth-curve models with random parameters. *Journal of the American*

Statistical Association, 84(405):241–247, 1989.

O. E. Lee and T. M. Braun. Permutation tests for random effects in linear mixed models. *Biometrics*, 68(2):486–493, 2012.

K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

S. R. Lipsitz, G. M. Fitzmaurice, E. J. Orav, and N. M. Laird. Performance of generalized estimating equations in practical situations. *Biometrics*, pages 270–278, 1994.

S. Maitra. *Applications of circular distributions and spatial point processes to the analysis of periodontal data*. PhD thesis, The University of Michigan, 2012.

D. C. Miller, C. S. Saigal, M. Banerjee, J. Hanley, and M. S. Litwin. Diffusion of surgical innovation among patients with kidney cancer. *Cancer*, 112(8):1708–1717, 2008.

M. Minaya-Sánchez, A. A. Vallejos-Sánchez, A. J. Casanova-Rosado, J. F. Casanova-Rosado, C. E. Medina-Solís, G. Maupomé, M. d. L. Márquez-Corona, and H. Islas-Granillo. Confirmation of symmetrical distributions of clinical attachment loss and tooth loss in a homogeneous mexican adult male population. *Journal of Dental Sciences*, 5(3):126–130, 2010.

A. Mombelli and C. Meier. On the symmetry of periodontal disease. *Journal of Clinical Periodontology*, 28(8):741–745, 2001.

J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302):415–434, 1963.

H.-P. Müller. Dealing with hierarchical data in periodontal research. *Clinical Oral Investigations*, 13(3):273–278, 2009.

S. Müller, J. L. Scealy, and A. H. Welsh. Model selection in linear mixed models. *Statistical Science*, 28(2):135–167, 2013.

J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2015. URL `http://CRAN.R-project.org/package=nlme`. R package version 3.1-121.

C. A. Ramseier, J. S. Kinney, A. E. Herr, T. Braun, J. V. Sugai, C. A. Shelburne, L. A. Rayburn, H. M. Tran, A. K. Singh, and W. V. Giannobile. Identification of pathogen and host-response markers correlated with periodontal disease. *Journal of Periodontology*, 80(3):436–446, 2009.

B. J. Reich and J. S. Hodges. Modeling longitudinal spatial periodontal data: A spatially adaptive model with tools for specifying priors and checking fit. *Biometrics*, 64(3):790–799, 2008.

B. J. Reich, J. S. Hodges, and B. P. Carlin. 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, 2007.

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

A. Rotnitzky and N. P. Jewell. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77 (3):485–497, 1990.

R. L. Schmoyer. Permutation tests for correlation in regression errors. *Journal of the American Statistical Association*, 89(428):1507–1516, 1994.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

M. R. Segal. Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, 87(418):407–418, 1992.

M. R. Segal, J. D. Barbour, and R. M. Grant. Relating hiv-1 sequence variation to replication capacity via trees and forests. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

R. J. Sela and J. S. Simonoff. Re-em trees: a data mining approach for longitudinal and clustered data. *Machine Learning*, 86(2):169–207, 2012.

M. J. Silvapulle. Robust wald-type tests of one-sided hypotheses in the linear model. *Journal of the American Statistical Association*, 87(417):156–161, 1992.

A. Skrondal and S. Rabe-Hesketh. Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3): 659–687, 2009.

D. O. Stram and J. W. Lee. Variance components testing in the longitudinal mixed effects model. *Biometrics*, pages 1171–1177, 1994.

J. M. Taylor and N. Law. Does the covariance structure matter in longitudinal modelling for the prediction of future cd4 counts? *Statistics in Medicine*, 17(20): 2381–2394, 1998.

T. M. Therneau, P. M. Grambsch, and T. R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

Y.-K. Tu, M. S. Gilthorpe, G. S. Griffiths, I. H. Maddick, K. A. Eaton, and N. W. Johnson. The application of multilevel modeling in the analysis of longitudinal periodontal data-part i: absolute levels of disease. *Journal of Periodontology*, 75 (1):127–136, 2004.

G. Verbeke and G. Molenberghs. The use of score tests for inference on variance components. *Biometrics*, 59(2):254–262, 2003.

G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*. Springer Science & Business Media, 2009.

C. P. Wan, W. K. Leung, M. Wong, R. Wong, P. Wan, E. Lo, and E. F. Corbet. Effects of smoking on healing response to non-surgical periodontal therapy: a multilevel modelling analysis. *Journal of Clinical Periodontology*, 36(3):229–239, 2009.

C. B. Wiebe and E. E. Putnins. The periodontal disease classification system of the american academy of periodontology-an update. *Journal of the Canadian Dental Association*, 66(11):594–599, 2000.

R. C. Williams. Periodontal disease. *New England Journal of Medicine*, 322(6): 373–382, 1990.

H. Zhang. Classification trees for multiple binary responses. *Journal of the American Statistical Association*, 93(441):180–193, 1998.

H. Zhang and B. Singer. *Recursive partitioning in the health sciences*. Springer, 1999.