INNOVATIVE METHODS FOR SOME STATISTICAL ISSUES IN CLINICAL TRIALS

Diana Fong-Hor Lam

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2015

Approved by:

Gary G. Koch

Joseph G. Ibrahim

John S. Preisser

Amy Herring

Hongtu Zhu

Geraldo Heiss

ABSTRACT

Diana Fong-Hor Lam: Innovative Methods for Some Statistical Issues in Clinical Trials
(Under the direction of Gary Koch)

Clinical trials have many different aspects to them, and here three topics will be explored: Bayesian sensitivity analysis in survival models, ROC methods in the presence of verification bias, and non-parametric adjustment of covariates in randomized clinical trial with and without missing data.

The aim of the first paper is to develop several Bayesian influence measures to assess the influence of the prior, the sampling distribution, and individual observations in survival analysis with the presence of missing covariate data and describe these changes in a modified likelihood model (the perturbation model). We construct a Bayesian perturbation manifold to the perturbation model and calculate its associated geometric quantities and influence measures based on several objective functions to quantify the degree of various perturbations to statistical models. We carry out several simulation studies and analyze a real data set to illustrate the finite sample performance of our Bayesian influence method.

While clinical trials concerned with survival track how long people will live given that they have the disease, some trials are concerned with using screeners to predict disease in the first place. Chronic obstructive pulmonary disease (COPD) affects 5% of the adult population in the United States, but a general screener has not been evaluated for the disease at a large scale. In this paper a COPD screener is evaluated using an innovative application of sampling weights. With these sampling weights, which help us adjust for verification bias, we explore the different variables to use in the screener. The optimal variable on the pocket screener was forced expiratory volume in 1 second as compared to peak expiratory flow.

While screeners can help predict disease, mainly clinical trials are designed to evaluate treatments that cure disease or improve health outcomes. Non-parametric adjustment of covariates is an attractive methodology in the regulatory setting as it requires few assumptions. We develop methodology to estimate the treatment effect in a longitudinal logistic randomized trial after non-parametrically adjusting for covariates. We also develop methodology to estimate treatment effect when the outcome is ordinal (with more than two groups) instead of binary, when there is missing data among the baseline covariates, and when there

are multiple treatment groups to be evaluated. This methodology and its extensions are applied to a data set evaluating Cushing's disease and another data set evaluating multiple doses of a neurological disorder medication.

I would like to dedicate my dissertation to my family for their unwavering support during my graduate school years.

ACKNOWLEDGEMENTS

I would like to thank Dr. Gary G. Koch for his patience, kindness, and guidance over many a breakfast meeting at Bob Evans and Whole Foods.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In today's world of clinical trials, there are many statistical analyses that we perform to adjust for the fact that trials occur in real life and are not carried out as perfectly as we would wish. There are methods to adjust for missing data, sensitivity analyses to test model assumptions, and weighting schemes to adjust the data to a more general population. Furthermore, clinical trials have many outcome variables that can be addressed: length of survival in the study, dichotomous outcomes such as diseased or non-diseased, or ordinal outcomes such as diseased, partially diseased, and non-diseased.

1.1 SURVIVAL ANALYSIS

In one type of clinical trial, the objective is to assess whether a treatment can improve the survival rates. There are many types of survival models. Survival models are inherently different from general linear models because they contain a censoring variable. Since it is unlikely that all patients in the study will experience the event or events during the time frame of the study, it is necessary to include a variable indicating whether or not the patient was event-free at the time they left the study or the time that the study ended.

A subject with an uncensored event is a subject who experienced the event of interest during the observed portion of the study. The event could be death, recurrence of disease, organ transplant, to name a few. A patient who is censored is a patient who does not experience the event during his time in the study. Patients may be censored for one of many reasons: they completed the study period without incident, they withdrew from the study, they were loss to follow-up, etc. For the purpose of this paper, it is assumed that all patients have a known starting time in the study. Therefore any censoring that occurs is right-sided censoring.

While many models use maximum likelihood methods by estimating parameters in the density function, survival models want to maximize survival, which is measured as a function of the cumulative distribution. Let $F(t, \beta, x)$ be the cumulative distribution function for a subject at time t, and covariates in the vector x, with parameters in the vector β, and let $f(t, \beta, x)$ be the density function. The survivorship function is then $S(t, \beta, x) = 1 - F(t, \beta, x)$. If an event occurred, then that helps to estimate the density function, but if the observation is censored, that means the survivorship function is needed. The likelihood that needs to be

maximized must incorporate both the density of events and survivorship function. Therefore, a censorship indicator variable, C, is created, where C=1 if the subject experienced the event and C=0 otherwise.

The likelihood function we want to maximize for survival models for all $i = 1, \ldots, N$ subjects is

$$\prod_{i=1}^{N} [f(t, \beta, x)]^c [S(t, \beta, x)]^{1-c}$$

In order to estimate the survivorship function ($S(t, \beta, x)$), we use the cumulative hazard function and the density of the hazard function. The hazard function represents the risk of experiencing an event in an interval after time t, given that the subject has survived to time t (Hosmer and Lemeshow, 1999). If the hazard function can be estimated parametrically, it is usually modeled using traditional linear modeling methods. The cumulative hazard function is the sum of the hazards over all intervals.

If the hazard function has a parametric form, the hazard, cumulative hazard, and survivorship function are related with the following equations:

$$H(t) = \int_0^t h(u)\,\mathrm{d}u$$
$$S(t) = \exp(-H(t))$$

A popular method for analyzing such data is the Cox proportional hazards model (Cox, 1972), which estimates the effects of the covariates on the hazard ratio with respect to an unspecified baseline hazard. It assumes that hazards are proportional between groups with respect to time. The model for the Cox proportional hazards model is

$$\lambda(t|\boldsymbol{z}_j) = \lambda_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{z}_j) \tag{1.1}$$

where $\lambda_0(t)$ is the baseline hazard given the subject had survived until time t, $\boldsymbol{z}_j$ is the vector of covariate values for patient j, and $\boldsymbol{\beta}$ is the vector of treatment effect parameters.

1.2 MISSING DATA AND SENSITIVITY ANALYSIS

Survival models, with their inherent censoring issues and numerous covariates, have posed a problem for statisticians wishing to perform sensitivity analysis on them. Any measure of sensitivity must be able to

accommodate the censored structure of the data and account for binary, categorical, and continuous covariates. Ad hoc methods are sometimes used for treating missing data, such as carrying the last observation forward or imputing the missing value using the mean of the covariate (Little and Rubin, 1987). One of the earlier papers to address missing data in survival models was Schluchter and Jackson (1989). They specified the joint distribution of failure time and covariates as the conditional distribution of failure time given covariates, and a marginal distribution on covariates, which was treated as a nuisance parameter. Lin and Ying (1993) came up with an estimator using the approximate partial likelihood estimator. Zhou and Pepe (1995) expanded on this work and applied this EM-type algorithm to data that were missing at random and monotone.

Most methods assume that missing covariates are either missing completely at random (MCAR) or missing at random (MAR). Covariates that are MCAR means the missing data mechanism does not depend on the observed or missing observations (Little and Rubin, 1987). In other words the missing observations are themselves a simple random sample. Covariates that are MAR means the data missing mechanism depends on the observed values but not the missing values (Little and Rubin, 1987). The missing observations are a simple random sample given the observed values. Paik and Tsai (1997) proposes using multiple imputation to address the missing covariates under MAR conditions. Chen and Little (1999) introduce a non-parametric likelihood which uses a discretized version of the likelihood where the baseline hazard is cast as a step function with jumps at the observed values. This method requires the specification of the missing data distribution and assumes missingness is MAR. Furthermore the method is only valid for covariates that are all discrete or all normally distributed (Herring and Ibrahim, 2001). Another method of handling missing covariates is to use estimating equations to produce parameter estimates. These methods generally require the missing data mechanism to be specified. The likelihood often has an expectation of the complete data given the observed data. Lipsitz and Ibrahim (1998) have estimating equations that are applicable only to categorical covariates. Herring and Ibrahim (2001) expand on the methodology of Lipsitz and Ibrahim (1998) to the situation looking at both continuous and categorical covariates. They provide weighted estimating equations in the log likelihood where the weights are the probabilities of a particular missing data pattern for a subject. These methods (Lipsitz and Ibrahim, 1998; Chen and Little, 1999; Herring and Ibrahim, 2001) all use an EM algorithm (and possibly adaptive rejection metropolis sampling) to get estimates.

Semiparametric methods required fast computing capabilities. With faster capabilities, more nuanced ways of handling missing data in survival models became available. Herring et al. (2002), suggested using Gibbs sampler along with Monte Carlo EM algorithm to obtain parameter estimates. Herring et al. (2004) later

expanded on the methods of Leong et al. (2001), who had done work on non-ignorable data mechanisms for binary covariates, to the case when the non-ignorable mechanism was applied to both discrete and continuous covariates. More recently, Ibrahim et al. (2008) developed semi-conjugate priors for missing at random covariate data and developed a variation on DIC for survival models. As the methods become more complex, other papers such as Cho et al. (2009) have found creative ways of getting around computational issues for survival models.

Bayesian work in survival models has focused on two main areas: extending frequentist hypothesis testing methods into a Bayesian setting and developing more complex priors to better suit the nature of survival parameters. Sinha et al. (2003) extended Cox's partial likelihood into a Bayesian context. Dunson and Herring (2003) developed Bayesian methods for testing null hypothesis versus order-restricted alternatives. In addition to exploring survival model testing, Dunson and Park (2008) developed a kernel stick breaking prior to model an uncountable number of random events. Other types of priors used in Bayesian survival analysis include beta process and Dirichlet process. The beta process treats discrete intervals of the hazard function as having a beta distribution. The Dirichlet process is a non-parametric prior that involves putting a Dirichlet distribution on the cumulative hazard over disjoint intervals. A more detailed summary of these priors can be found in Ibrahim et al. (2001).

Bayesian diagnostics on survival models and other generalized models has grown out of frequentist diagnostics. The most commonly used frequentist diagnostic is Cook's distance (Cook, 1986). One extension of Cook's distance is the conditional posterior ordinate (CPO), which can be thought of as the marginal posterior predictive density of the i^{th} case if it were deleted from the data set (Gelfand et al., 1992; Gelfand, 1999; Wei and Su, 1999). In addition to case deletion diagnostics, there are also systemic and isolated departures to examine (Davison and Tsai, 1992). Systemic departures come about when features of the data set have not been captured or a large group of observations does not fit well. Isolated departures, such as outliers, arise when a few observations do not fit the model.

Detecting systemic departures is part of doing robust Bayesian analysis, which came to light in the early 90's (Berger, 1990, 1994). The idea behind robust Bayesian analysis was that the outcome of the analysis was an imprecise judgment quantifiable only on certain probabilistic intervals. Berger and Berliner (1986) tested the robustness of priors by developing an $\epsilon-$contamination class of priors, that allowed for a related family of priors to be considered at once; however, they did not quantify the change caused by going from different priors. McCulloch (1989) developed a method of quantifying the change caused by changes in the

prior, and this method is called local sensitivity analysis. Although measuring the many different changes at once is possible, Gustafson and Wasserman (1995) suggested that examining individual elements of change is preferred. Hyperparameters or clusters that had high local sensitivity were a cause for concern. Zhu et al. (2007) developed advanced techniques for exploring clusters that did not fit well.

1.3 RECEIVER OPERATOR CHARACTERISTIC CURVES IN THE PRESENCE OF VERIFICATION BIAS

Another aspect of clinical trials besides survival is determining whether continuous or categorical measures can be used to discriminate between diseased and non-diseased patients. Initially the methods developed were used to detect signals in the presence of noise(Green and Swets, 1966). Early evaluations focused on categorical measures. The estimates used to evaluate the measures were sensitivity, the probability of the test being positive given the subject was diseased, and specificity, the probability of the test being negative given the subject was not diseased. Eventually, methods were created to handle continuous measures. These methods created a Receiver Operating Characteristic curve (ROC curve), which plots the sensitivity versus 1-specificity at a variety of cutoff points, where the cutoff point differentiates between diseased and non-diseased subjects. The Area Under the Curve (AUC) gives a sense of the measures accuracy for differentiating between diseased and non-diseased states and ranges from 0.5 to 1. An AUC of 0.5 represents a measure whose accuracy is no better than a random guess, while an AUC of 1 represents a measure that predicts diseased or non-diseased status perfectly. The AUC corresponds to the Mann-Whitney statistic (Bamber, 1975).

In addition, statisticians can compare the AUC's of multiple measures that were conducted on the same people to see if one measure is better able to discriminate than others (DeLong et al., 1988). Multiple tests or screeners can be tested on the same subjects in an effort to compare different screeners. The AUC's for the different screeners are then correlated. Statistical tests that wish to determine whether one screener is significantly different than the other needs to take into account this correlation when calculating the variance of the difference between the AUC's.

One of the first papers to address the issue of the variance of the difference between two correlated AUC's was Hanley and McNeil (1983). The proposed method for estimating the standard error of the difference of two correlated AUC's involved calculating r, the correlation introduced between the two areas by studying the same sample of patients. The correlation, r, is a function of the average of the correlation between the two

tests in the non-diseased group and the correlation between the two tests in the diseased group and the average of the AUC's for each test was calculated. A derived table with average correlation along the rows and average area under the curve along the columns was then used to calculate r . The standard error formula is the same as normal: $\text{SE}(\text{AUC}_1 - \text{AUC}_2) = \sqrt{\text{SE}^2(\text{AUC}_1) + \text{SE}^2(\text{AUC}_2) - 2 \times r \times \text{SE}(\text{AUC}_1) \times \text{SE}(\text{AUC}_2)}$.

In addition to calculating the correlation between the measures, it is also important to consider verification bias when calculating both the estimate of the difference between correlated AUC's and the variance of the difference. There have been 2 main ways of dealing with verification bias: imputation and reweighting. Some of the imputation methods are full imputation (Rotnitzky et al., 2006; Alonzo and Pepe, 2005) and mean score imputation (Alonzo and Pepe, 2005). These imputation methods extend the ideas initially proposed by Begg and Greenes (1983) to screener tests with continuous outcomes. Unlike imputation methods that treat the disease status as missing, reweighting methods use only the subjects that has both test and disease information. But like imputation methods, reweighting methods assumes that the population of subjects with the test results is the population that we wish to make inference on.

For all methods described, let T_i be the test result for the i^{th} person, V_i be the indicator that the i^{th} patient went on for disease verification, D_i be the true disease status, and X_i be the vector of covariates of interest. Thus in the presence of verification bias, T_i, V_i, and X_i are observed for all $i = 1, \ldots, n$ patients where $n =$total number of patients in the study, and D_i is missing for a subset of patients.

In the full imputation method the missing disease status is predicted by modeling $P(D = 1|T, X)$. The disease status is replaced by the predicted probability of disease even for subjects whose disease status is known. The estimated probability of disease for the i^{th} patient is $\hat{\rho}_i = P(D_i = 1|T_i, X_i)$. Therefore for the screener cutoff of c the senstivity with full imputation is

$$\text{sens(c)}_{\text{FI}} = \frac{\sum_{i=1}^{n} I(T_i \geq c)\hat{\rho}_i}{\sum_{i=1}^{n} \hat{\rho}}.$$

It seems the method of Rotnitzky et al. (2006) is the most popular full imputation method due to the flexibility, development of statistical estimators, and double robustness property. Details follow, but the main steps that Rotnitzky et al. (2006) use to deal with verification bias are to first model the probability of being sent on for disease verification given the test result and other measured covariates (1.2), and then model the probability of disease given that the disease was verified (1.3). Using the estimates from these models, they

create an estimator for the disease status based on the model parameters, test result, and other measured covariates.

The model of log odds of no disease verification is expressed as

$$log\left\{\frac{Pr(V=0|T,X,D)}{Pr(V=1|T,X,D)}\right\} = h(T,X) + q(T,X)D. \tag{1.2}$$

where $h(T,X)$ is a function of the test result and covariates associated with the log odds of not having disease verification depend on disease status that do not have any interaction with the disease status. Let $q(T,X)$ be a function of how test result and covariates interact with disease status to influence the log odds of no disease verification. For example, if we had non-ignorable verification bias, then we expect $q(T,X) \neq 0$. Then $h(T,X)$ is assumed to have a parametric form such that $h(T,X) = h(T,X|\gamma)$.

The next step is to model the probability of being diseased given that the subject went on for verification

$$log\left\{\frac{Pr(D=1|V=1,T,X)}{Pr(D=0|V=1,T,X)}\right\} = m(T,X). \tag{1.3}$$

Let $m(T,X)$ be a function of the test result and covariates that models log odds of disease status among those who have their disease status verified. $m(T,X)$ is also assumed to have a parametric form $m(T,X) = m(T,X|\mu)$.

The double robust estimator, $D_{DR}(\gamma,\mu)_i$ or D_{DRi} for short, for the i^{th} patient's disease status is

$$D_{DR}(\gamma,\mu)_i = P(V_i,T_i,X_i|\mu) + U_i(\gamma|\mu)$$

where

$$\begin{aligned} U_i(\gamma,\mu) = V_i[\{D_i - P(1,T_i,X_i|\mu)\} + \exp\{h(T_i,X_i|\gamma) \\ + q(T_i,X_i)D_i\} \times \{D_i - P(0,T_i,X_i|\mu)\}] \end{aligned}$$

and

$$P(V_i, T_i, X_i) = Pr(D_i = 1|V_i = 1, T_i, X_i) \times \{V_i + \frac{(1 - V_i)\exp\{q(T_i, X_i)\}}{1 - Pr(X_i = 1|V_i = 1, T_i, X_i) \times \{1 - \exp\{q(T_i, X_i)\}\}}\}.$$

The process gives us 2 chances to get the model right. The estimator's double robustness property means that if either the verification or disease model is mis-specified the estimator will still be consistent. D_{DRi} is then calculated for all subjects and depends only on observed variables, therefore it can be calculated regardless of verification status. The estimator of the AUC is then calculated using the normal formula with D_{DRi} substituted for D_i for the entire data set. The formula for the AUC and its double robust estimator is

$$\nu = Pr(T_2 > T_1|D_1 = 1, D_2 = 0) + 0.5 \times Pr(T_2 > T_1|D_1 = 1, D_2 = 0)$$

$$\hat{\nu} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\{D_{DRi}(1 - D_{DRj}) \times [I(T_i > T_j) + I(T_i = T_j)]\}}{\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\{D_{DRi}(1 - D_{DRj})\}}.$$

Instead of treating the disease status as missing as in imputation methods, the inverse probability weighting (IPW) estimator uses weights to calculate the probability of verification. It weights each observation in the verified sample by the inverse of the probability of being selected for verification (Alonzo and Pepe, 2005). Let $\hat{\pi}_i = P(V_i|T_i, X_i)$, which is estimated from disease-verified patients only. The estimate for sensitivity using IPW for the cutoff c is

$$\text{sens(c)}_{\text{IPW}} = \frac{\sum_{i=1}^{n} I(T_i \geq c)\{V_i * D_i/\hat{\pi}_i\}}{\sum_{i=1}^{n}\{V_i * D_i/\hat{\pi}_i\}}.$$

This approach yields bias estimates of AUC when verification model is mis-specified (Alonzo and Pepe, 2005). However, in randomized studies, often the verification model is under investigators control. In these situations, investigators send a random sample of subjects on for verifiation based on pre-specified criteria.

Another method derived by Alonzo et al. (2003) is the semiparametric efficient estimator (SPE). This is a combination of both the IPW and full imputation method. The estimator is

$$\text{sens(c)}_{\text{SPE}} = \frac{\sum_{i=1}^{n} I(T_i \geq c)\{V_i * D_i/\hat{\rho} - (V_i - \hat{\pi}_i) * \hat{\rho}_i\hat{\pi}_i\}}{\sum_{i=1}^{n}\{V_i * D_i/\hat{\pi}_i - (V_i - \hat{\pi}_i) * \hat{\rho}_i/\hat{\pi}_i\}}.$$

This method is semiparametric since it parametrically models $P(D|T, X)$ and $P(V|T, X)$ but does not specify a distribution for $P(D, T, X)$. This method is also doubly robust. Furthermore, this method combines both the efficiency of imputation methods with the robustness of re-weighting methods. Finally, another concern when evaluating ROC's is the presence of verification bias, sometimes referred to as work up bias.

Selection bias, sometimes referred to as verification bias or work-up bias, occurs when the measures used to evaluate a test are based only on the subset of patients who had the disease verified (Begg, 1987). The stronger the association between the test result and disease verification, the larger the bias (Begg, 1987). The bias is important to adjust for since positive tests will be over-represented in the disease-verified population, leading to artificially inflated sensitivity and deflated specificity estimates.

Another bias that occurs is the bias of uninterpretable results. Uninterpretable results are those that are *technically* unacceptable. For example, if the test is an ultrasound of the pancreas and the pancreas cannot be viewed, the test is uninterpretable (Begg et al., 1986). This is different from indeterminate results in which the meaning of the test is unclear. In situations in which the test result is uninterpretable, the uninterpretable subjects may be biasing the results if the uninterpretability is associated with test results (Begg et al., 1986). Begg et al. (1986) suggests adjusting the posterior odds by the probability of the odds of uninterpretable results of diseased groups to non diseased groups. Another suggestion by Begg et al. (1986) was to consider the uninterpretable readings as non-diseased, and adjust the numerator and denominator of specificity by that number.

1.4 NON-PARAMETRIC COVARIATE ADJUSTMENT IN LONGITUDINAL BINARY AND ORDINAL SETTINGS

For many clinical trials, the main purpose of the trial is to evaluate the treatment effect. Analysis plans need to state *a priori* the statistical methods they plan on using to determine if the treatment effect is significant. This can be difficult when some methods require checking the data to determine if certain assumptions are met. Therefore, the fewer assumptions a method requires the more attractive it is in a regulatory setting. One of the benefits of operating in a regulatory setting is that oftentimes, the treatment groups are randomized,

meaning that covariate imbalances between the treatment group(s) and control(s) are due to random chance. Non-parametric adjustment of covariates exploits this assumption of random covariate imbalance.

The advantages and limits of covariate modeling versus non parametric modeling have been outlined extensively by Tangen and Koch (1999). Some of the advantages of traditional covariate modeling are that the odds ratio for the treatment applies to patients in the same subpopulation according to the covariates in the model (Tangen and Koch, 1999). That then assumes that the patients at each cross classification of the covariates represent a stratified simple random sample of subjects based on explanatory variables. Also, traditional methods produce estimates of the effect of the covariates. Furthermore, non-parametric adjustment treats covariates and stratification variables differently while modeling treats them the same (Tangen and Koch, 1999). The main advantage of non-parametric adjustment is that it generally reduces the variance of the treatment effect without impacting the parameter estimate itself, and the treatment effect is still applicable to the general population the subjects come from (Koch et al., 1998). One of the advantages of adjusting non-parametrically is that the sample size only needs to take into account the number of parameters that pertain to treatment, without regard for the number of covariates you wish to adjust by.

After adjusting for covariates, we still have multiple ways of modeling the treatment effect, depending on the outcome variable. Oftentimes, the outcome variable in clinical settings is a dichotomous variable, however, there has been some work to extend methodology to an ordinal outcome. A few different methods deal with ordinal outcome variables. The generalized logistic regression model developed by McCullagh (1980) models the log of the probability of one level with respect to the probability of a reference level of the outcome variable. For simplicity's sake, we will assume the reference level is $j = 1$, although in reality, the reference could be any level of Y. Let Y_i be the outcome for the i^{th} subject, where $Y_i \in 1, \ldots, k$, $\gamma_j = Pr(Y = j|\boldsymbol{x})$, and $\boldsymbol{x}_i$ be a vector of covariates for the i^{th} subject. The generalized logistic regression models

$$log\left(\frac{\gamma_j}{\gamma_1}\right) = \alpha_j + \boldsymbol{x}'\boldsymbol{\beta}_j, j = 2, \ldots, k,$$

where α_j represents the j^{th} log odds ratio of the j^{th} outcome to the first outcome, holding all other variables constant. $\boldsymbol{\beta}_j$ represents the effect of the covariates on the log odds ratio of the j^{th} outcome with respect to the first outcome.

Unfortunately, the generalized logistic regression has two disadvantages when the outcome is ordinal. As each outcome level except the reference has its own parameter effects and intercept, the sample size required to produce stable parameter estimates is large. It does not make use of the ordinal nature of the outcome variable, as odds are calculated with respect to a reference level. A model that does make use of the ordinal nature of the outcome variable is the proportional odds model developed by McCullagh (1980). Let $\pi_j = Pr(Y \leq j|\boldsymbol{x})$, so the linear model is

$$logit(\pi_j) = \alpha_j + \boldsymbol{x}'\boldsymbol{\beta}, j = 2, \ldots, k.$$

In this setting, α_j represents the log odds ratio of $Y \leq j$ if all covariates are equal to 0. $\boldsymbol{\beta}$ represents the effect of the covariates on the log odds of more favorable values to less favorable values. This model requires that the effect of the covariates be the same regardless of which value differentiates between a success and failure for the odds, as there is only one parameter of covariate effect regardless of the π_j being modeled. This assumption is called the proportional odds assumption and requires the data to check.

A model that combines features of the 2 is the partial proportional odds model, developed by Peterson and Harrell (1990). It can be modeled as

$$logit(\pi_j) = \alpha_j + \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{t}\boldsymbol{\phi}_j$$

where $\boldsymbol{t}$ is a vector of a subset of the original covariates which do not satisfy the proportional odds assumption. The partial proportional odds model uses both the ordinal nature of the outcome, and relaxes the proportional odds assumption for the covariates in the model. The advantage of this model is that for covariates for which proportional odds holds, we only need one parameter (expressed in $\boldsymbol{\beta}$, but we still have flexibility for covariates which do not satisfy the proportional odds assumption.

Today's world of clinical trials is moving more towards longitudinal studies where the effects of time can be teased out. While longitudinal studies can provide a clearer picture of treatment effect, they also present a challenge for analysis due to the correlation between repeated measures. The same variable measured multiple times within the same person will have some correlation between measurement. There are two methods for dealing with longitudinal data. In the generalized linear model (GLM) method, the correlation is treated like a nuisance parameter, leading to inference on the marginal model only. In the generalized linear

mixed model (GLMM) method the correlation is treated like an additional parameter to be estimated. The GLMM method requires large sample sizes, so if the correlation of the outcome variable between repeated visits is not of interest to inference, then the GLM method is more practical to use.

Within the GLM method, there are 2 methods for adjusting for the correlation between repeated measures. Maximum likelihood methods require specifying the correlation structure between measurements. However this can lead to bias if the correlation structure is mis-specified. An alternative method for parameter covariance estimation is to use generalized estimating equations (Liang and Zeger, 1986). The advantage of this method is that it is robust even if the correlation structure is mis-specified. Generalized estimating equations use a generalized linear model to estimate the marginal distribution of the outcome variable, and then uses a combination of sandwhich estimators and working correlation matrices to estimate the parameter and its covariance (Liang and Zeger, 1986).

Zeger et al. (1985) formed a model for the analysis of binary longitudinal data with time-independent covariates. If we let Y_{it} be the binary outcome variable for the i^{th} subject for the t^{th} time point and $\boldsymbol{X}_i$ be the covariate vector for the i^{th} subject, then Zeger et al. (1985) models $\pi_i = Pr(Y_{it} = 1)$ and the correlation between consecutive time points ρ. The models used is $logit\pi_i = \boldsymbol{X}_i'\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is the effect of the covariates on the log odds ratio and $corr(Y_{it}, Y_{i,t-1}) = \rho$. This model treats the binary series as the realization of a stationary Markov chain. One of the issues with this method was that there might also be correlation between outcomes that were more than 1 repeated measure apart. Furthermore, all time dependence is modeled in one parameter: ρ.

Stram et al. (1988) formed a more flexible model than Zeger et al. (1985) that could also work for repeated ordered categorical outcomes. The model relies on the proportional odds assumption for marginal probabilities, and each time point has its own model. By modeling each time point separately, the dependence between time points does not need to be modeled and instead is calculated empirically. Furthermore, the paper suggests a multi-stage testing procedure to check whether the time coefficients are significantly different from each other. The procedure relies on the fact that all hypotheses need to be rejected to move on to a subset of the hypothesis.

Schacht et al. (2008) develops a different type of testing frame work. He uses non-parametric baseline covariate adjustment for two groups, but not in a longitudinal setting. Instead of finding a parameter estimate, he built a testing framework. Let Y_i be the distribution of the outcome for treatment group i where $i = 1$ if the patient is on treatment and 0 otherwise. He tested $p = Pr(Y_0 < Y_1) + 0.5Pr(Y_0 = Y_1)$. Under

the null hypothesis $p = 0.5$. The estimate of $p, \hat{p}$ is then adjusted by $\sum_{c=1}^{d} \hat{\gamma}_c(\hat{q}_c - 0.5)$ where $\hat{q}_c$ is $Pr(Q_{c0} < Q_{c1}) + 0.5P(Q_{c0} = Q_{c1})$. Q_{ci} is the value of the c^{th} covariate for the i^{th} treatment group. $\hat{\gamma}_c$ is then estimated from the covariance matrix of the covariates with each other and the covariates with the outcome. This method is similar to a simplified version of the method we propose. The advantages of our method is that we calculate the estimate of the difference between the two groups and the calculation of $\hat{\gamma}$ is unnecessary.

This literature review covers many different aspects of clinical trials. We have reviewed survival models and aspects of them that present challenges for analysis. We have talked about more preventative-oriented clinical trials that aim to determine disease status using some other measure which is usually either easier or cheaper than a gold standard, and the estimates that go with those studies. Finally, we addressed clinical trials in a regulatory setting which measure whether the treatment effect can provide some benefit.

CHAPTER 2: BAYESIAN SENSITIVITY ANALYSIS IN SURVIVAL MODELS

2.1 INTRODUCTION OF SURVIVAL DATA SET IN ONCOLOGY STUDY

In this paper, we extend the methods of Zhu et al. (2007, 2010) to data from the Eastern Cooperative Oncology Group (ECOG), who carried out a phase III clinical trial for high dose interferon on multiple melanoma patients. In this data set, we have $n = 285$ subjects and either their time to relapse or time to censoring. There were 196 relapses and 89 censored observations. The four covariates of interest are treatment, age, Breslow score, and size. Breslow score and size had missing values.

Briefly, Zhu et al. (2007) outlines 4 steps to calculating Bayesian local sensitivity measures for survival models.

1. Define what aspects of the model are to be tested, ie, individual data points, prior assumptions, etc, and create a vector, $\boldsymbol{\omega}$ that is associated with these aspects

2. Adjust the likelihood model such that $\boldsymbol{\omega}$ is incorporated

3. Choose how sensitivity will be measured (posterior mean distance, ϕ-divergence,. . .)

4. Calculate sensitivity measures that are generally derivatives of the adjusted likelihood in step 2 with respect to $\boldsymbol{\omega}$

We will apply this method to both a simulation, a real data set, and then we will explore some of its properties using a bootstrap simulation.

Mahabadi and Ganhali (2013) uses an index of sensitivity to non-ignorability (ISNI) to meausre how sensitive model parameters are to departures in the ignorability assumptions. It is calculated in a similar manner to the measures we develop, however, it is not as flexible as the model we develop, as it can only test the ignorability assumption. Furthermore, it is a frequentist measure as opposed to ours which is a Bayesian one.

This chapter is organized as follows, in Section 2.2, we develop measures for calculating first and second order influence measures. Since the measures are based on geometric tensors and covariant derivatives as described in differential geometry and require only a proper sampling distribution, the censoring in survival models does not pose a problem. In Section 2.3, we apply the influence measures to a variety of simulated data sets and the ECOG data. In Section 2.4, explore the properties of the Bayesian influence measures via bootstrap samples of the ECOG data set. Section 2.5 contains a discussion of the measures.

2.2 BAYESIAN SURVIVAL MODELS WITH MISSING COVARIATES

2.2.1 STATISTICAL SURVIVAL MODELS WITH MISSING DATA

When covariates are missing, we need to think about whether or not the data are missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). To access aspects of the missingness pattern, we need to first establish some notation. For our data, we let $D_o = (x_{1,o}, ..., x_{n,o})$, $D_m = (x_{1,m}, ..., x_{n,m})$, and $D_c = (D_o, D_m)$ represent the observed, missing, and complete data respectively. Let $x_{i,m}$ be the vector of missing values for the i^{th} subject and $x_{i,o}$ be the vector of observed values. The vector of covariates for the i^{th} subject is $x_i = (x_{i,m}, x_{i,o})$ for $i = 1, \ldots, n$. For missing data problems, we want to model $p(D_c|\theta)$ as the product of $p(D_o|\theta)$ and a model of the missing data given the observed data ($p(D_m|D_o, \theta)$), where θ is a vector of necessary parameters. We usually use Markov chain Monte Carlo (MCMC) methods to simulate $p(\theta|D_o) \propto \int p(D_c|\theta) * p(\theta) dD_m$. With missing data, we also need a missingness indicator, r_{ij}, to indicate missingness of the j^{th} covariate of the i^{th} subject. Therefore, $r_{ij} = 1$ if x_{ij} is missing and 0 otherwise, and each subject's missingness vector is r_i. While $i = 1, \ldots, n$ and the number of covariates is p, j goes from $1, \ldots, q$ where q is the number of possible missing covariates, $q \leq p$, and the difference $(p - q)$ is the number of completely observed covariates. Finally, let y_i represent the survival data for the i^{th} observation, with covariates x_i for $i = 1, \ldots, n$.

Each of the three missingness mechanisms has its own strategies for analysis. When the data are MCAR, then the probability of observing x_i is independent of y_i and the observed and unobserved values of x_i. In this case, the x_i represents a random sample of the data. While doing a complete case analysis where only observations that are completely observed are used would be inefficient, it would at least be unbiased. When the data are MAR, the probability of observing x_i conditional on the observed data does not depend on the unobserved data. If the data are MAR and missingness depends only on the covariates and not the

response, then a complete case analysis will lead to unbiased results. In both the MCAR and MAR case, the data missing mechanism can be ignored. When the data are MNAR then the probability of observing x_i conditional on the observed data is dependent on the missing data itself. Again, if missingness does not depend on the response, a complete case analysis will be unbiased (Ibrahim et al., 2005).

We can model the probability of the i^{th} observation having a particular survival time, covariate values, and missingness indicator as

$$p(y_i, x_i, r_i|\theta) = p(y_i|x_i, \theta) * p(x_i|\theta) * p(r_i|x_i, y_i, \theta)$$

$p(y_i|x_i, \theta)$ is a general survival model, which depends on the covariates being known. $p(x_1|\theta)$ is a model for the covariates, both missing and observed, and $p(r_i|x_i, y_i, \theta)$ is the model of the missingness indicator of the covariates. One way to perturb the missing data mechanism from MAR to MNAR is to use a Bayesian Perturbation Maniforld.

2.2.2 BAYESIAN PERTURBATION MANIFOLD

The mechanics of measuring perturbations has its roots in differential geometry. We first choose a function that we would like to evaluate our diagnostic measures with respect to; measures such as the Bayes factor, ϕ-divergence, Kullback-Leibler distance (KL distance) to name a few. Then, we envision the function as a manifold, or surface in more than 3 dimensions, over all changes of interest. The change in the manifold curvature going in one direction of change represents the amount of change that perturbation affects. There are certain guidelines as to what makes a viable perturbation scheme.

We represent pertubations to the complete-data model, $p(D_c, \theta)$ using the vector $\omega = \omega(D_c, \theta)$ in a set Ω. For example, if we are interested in investigating if our model has heteroscedastic variance in a general linear model $Y_i = x_i\beta + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$ and we wish to use the Bayes factor to measure how the pertrubed model changed from the unperturbed model, we would go from our unpertubed model, $Y_i \sim N\left(x_i'\beta, \sigma^2\right)$ to the pertrubed model $Y_i|\omega \sim N\left(x_i'\beta, \frac{\sigma^2}{\omega_i}\right)$. By dividing each observation's variance by an element of ω, we can then examine how our manifold of the Bayes factor, curves as we let each of the observations have its own variance. Our perturbation vector would be $\omega = (\omega_1, \ldots, \omega_n)$ where n is the number of observations. The value of ω that represents no perburation is $\omega_0 = (1, \ldots, 1)_{n\times 1}$. All perturbations must have some $\omega_0 \in \Omega$ that represents no change in the model. Furthermore, the distribution with the perturbation must still

follow the rules of probability. For example, if we perturb the sample variance, then $\int p(y|\omega)dy = 1$ and $p(y|\omega) > 0$. Ideally each element of ω needs to be independent, and the vector elements need to be scaled so that the effect sizes are comparable. For example, when using Cook's distance to examine clustered data, larger clusters have more influence due to their size (Zhu et al., 2007). Perturbations should not have this problem.

We put the perturbations in a geometric framework called a Bayesian perturbation manifold. We can consider all possible perturbations of interest as a Riemannian Hilbert manifold under some conditions, $M = p(D_c, \theta) : \omega \in \Omega$. We can parameterize tangent curves from ω as

$$C(t) = p(D_c, \theta|\omega(t)) : [-\epsilon, \epsilon] \to M, C(0) = p(D_c, \theta|\omega)$$

and
$\int \dot{l}(D_c, \theta|\omega(t))^2 p(D_c, \theta|\omega(t)) dD_c d\theta < \infty$. On M we define the tangent space of all possible curves M at ω as those curves that take form of $C(t)$ as $T_\omega M$. We can define the inner product of two tangent vectors $v_1(\omega)$ and $v_2(\omega)$ on $T_\omega M$ as

$$< v_1, v_2 > (\omega) = \int \{v_1(\omega)v_2(\omega)\}p(D_c, \theta|\omega)dD_c d\theta \tag{2.1}$$

In $T_\omega M$ we can measure pertrubations. We look at several measusres: $G(\omega_0), \text{FI}_{RI}[\mathbf{v}](\boldsymbol{\omega}(0))$ and $\text{SI}_{RI}[\mathbf{v}](\boldsymbol{\omega}(0))$. These terms approximately represent distance, first-order curvature of the manifold, and second-order curvature of the manifold. The geometric tensor, $G(\omega)$ is defined as

$$G(\omega(0)) = \int [\partial_\omega l(D_c, \theta|\omega(0))]^{\otimes 2} p(D_c, \theta|\omega) dD_c d\theta \tag{2.2}$$

. If the covariant derivative, $\nabla_{\boldsymbol{v}}\boldsymbol{u}(\boldsymbol{\omega}) \neq 0$ then we use first-order influence measure ($\text{FI}_{RI}[\mathbf{v}](\boldsymbol{\omega}(0))$) as the diagnostic of interest. If $\nabla_{\boldsymbol{v}}\boldsymbol{u}(\boldsymbol{\omega}) = 0$ then use we use the second-order influence measure ($\text{SI}_{RI}[\mathbf{v}](\boldsymbol{\omega}(0))$) as the diagnostic. The covariant derivative $\nabla_{\boldsymbol{v}}\boldsymbol{u}(\boldsymbol{\omega})$ measures the initial rate of change of $IF(\omega_0)$ as we move from ω_0 in the direction of ω and is defined as $\nabla_{\boldsymbol{v}}\boldsymbol{u}(\boldsymbol{\omega})$,

$$d\boldsymbol{u}[\boldsymbol{v}](\boldsymbol{\omega}) - 0.5\{\boldsymbol{u}(\boldsymbol{\omega})\boldsymbol{v}(\boldsymbol{\omega})p(\mathbf{z}_{com}, \boldsymbol{\theta} \mid \boldsymbol{\omega}) - \int \boldsymbol{u}(\boldsymbol{\omega})\boldsymbol{v}(\boldsymbol{\omega})p(\mathbf{z}_{com}, \boldsymbol{\theta} \mid \boldsymbol{\omega})d\mathbf{z}_{com}d\boldsymbol{\theta}\}. \tag{2.3}$$

The first and second order influence measure are written with respect to an intrinsic influence measure, $\text{IF}(\boldsymbol{\omega}) = \text{IF}(p(\boldsymbol{\theta} \mid \boldsymbol{z}_{obs}, \boldsymbol{\omega}))$. We are usually interested in letting $IF(\boldsymbol{\omega})$ represent ϕ-divergence, the Bayes Factor, or the posterior mean.

We must then define a relative intrinsic influence measure (RIFM, $RI(\omega, \omega_0)$) as a function of both $p(\theta|D_o, \omega)$ and $p(\theta|D_o, \omega_0)$ so that the difference in the intrinsic influence measure between the perturbed and unperturbed situation can be measured on the Bayesian Perturbation Manifold. The simplest example is to let $RI(\omega, \omega_0) = IF(\omega) - IF(\omega_0)$. In addition, we want to scale the $RI(\omega, \omega_0)$ by the minimal geodesic distance between $p(D_c, \theta|\omega)$ and $p(D_c, \theta|\omega_0)$. We will define this as the intrinsic influence measure.

$$\text{IGI}_{RI}(\boldsymbol{\omega}, \boldsymbol{\omega}_0) = \frac{\text{RI}(\boldsymbol{\omega}, \boldsymbol{\omega}_0)^2}{g(\boldsymbol{\omega}, \boldsymbol{\omega}_0)^2}. \tag{2.4}$$

Since we are in the space $T\omega M$, we can describe the local behavior of $RI(\omega, \omega_0)$ as $RI(\omega(t), \omega_0)$ as $t \to 0$ along all possible curves $p(D_c, \theta|\omega(t))$ passing through $\omega(0) = \omega_0$. The first-order influence measure is defined as

$$\text{FI}_{RI}[\mathbf{v}](\boldsymbol{\omega}(0)) = \lim_{t \to 0} \text{IGI}_{RI}(\boldsymbol{\omega}(0), \boldsymbol{\omega}(t)) = \frac{\{d(\text{RI})[\mathbf{v}](\boldsymbol{\omega}(0))\}^2}{< \mathbf{v}, \mathbf{v} > (\boldsymbol{\omega}(0))} \tag{2.5}$$

When $\partial RI(\omega(0)) = 0$, the first-order influence measure is 0, so we use $\partial^2 RI(\omega(0))$ to calculate the second-order influence measure, defined as

$$\text{SI}_{RI}[\mathbf{v}](\boldsymbol{\omega}(0)) = \frac{\partial^2 \text{RI}(\boldsymbol{\omega}(0))}{< \mathbf{v}, \mathbf{v} > (\boldsymbol{\omega}(0))} \tag{2.6}$$

We can structure the missing data as a sequence of one-dimensional conditional distributions as in (Ibrahim and Lipsitz, 1999). In order to deal with missing data, we cast the influence measures, $FI_{RI}(\mathbf{v})$ or $SI_{RI}(\mathbf{v})$ as an expectation with respect to the joint distribution of the parameters and missing data. We use Gibbs sampling in adaptive rejection Metropolis sampling (ARMS) to get an empirical missing data distribution. We treat the missing values for each observation as a parameter to be estimated in Gibbs sampling, meaning that each observation that was missing was modeled with a prior distribution. The parameters for the missing variable distribution can be thought of as nuisance parameters.

In order to deal with missing data, we cast the influence measures, $FI_{RI}[v]$ or $SI_{RI}[v]$ as an expectation with respect to the joint distribution of the parameters and missing data. We use Gibbs sampling in adaptive

rejection metropolis sampling to get an empirical missing data distribution. We treat the missing values for each observations as a parameter to be estimated in Gibbs sampling meaning that each observation htat was missing was modeled with a prior distribution.

When we want to perturb the missing data mechanism from MAR to MNAR, then we perturb $p(r_i|y_i, x_i, \theta)$, where $x_i = (x_{i,o}, x_{i,m})$. Therefore the missingness will depend on both the observed ($x_{i,o}$) and unobserved ($x_{i,m}$) covariates. Since r_i is a binary variable vector, we perform a logistic regression on r_i with the observed covariates and failure time as the predictors. To model the missingness indicator, we have $logit[P(r_i = 1|\omega)] = \phi_0 + \phi_1 y_i + \omega x_i$, where ϕ_0 is the intercept, ϕ_1 is the coefficent of the failure time, and ω is a perturbation of the affect of the observed covariate. Let ϕ_0 and ϕ_1 be hyperparameters, used in the following logistic regression:

$$P(r_i = 1|x_i, y_i) = \left(\frac{\exp\{\phi_0 + \phi_1 y_i + \omega x_i\}}{1 + \exp\{\phi_0 + \phi_1 y_i + \omega x_i +\}}\right)^{r_i} \left(\frac{1}{1 + \exp\{\phi_0 + \phi_1 y_i + \omega x_i\}}\right)^{1-r_i}. \tag{2.7}$$

When we have more than one covariate missing, we can use the following one-dimensional conditional distribution of the missing variables to simplify the distrubtion of the missing covariate:

$$logit[P(r_{iq} = 1|r_{i1}, \ldots, r_{iq-1}, x_{i,o}, y_i)] = \phi_0 + \phi_1 y_i + \tilde{\omega}_q' \tilde{x}_{iq} + \phi' x_{i,o} \tag{2.8}$$

$$logit[P(r_{iq-1} = 1|r_{i1}, \ldots, r_{iq-2}, x_{i,o}, y_i)] = \eta_0 + \eta_1 y_i + \tilde{\omega}_{q-1}' \tilde{x}_{iq-1} + \eta' x_{i,o}$$

$$\vdots \tag{2.9}$$

$$logit[P(r_{i1} = 1|x_{i,o}, y_i)] = \psi_0 + \psi_1 y_i + \omega_1 x_{i1} + \psi' x_{i,o}.$$

Let $\tilde{\omega_q}$ be a $q \times 1$ vector of ω elements and $\tilde{x}_{iw}$ is the vector of missing elements $1, \ldots, w \leq q$. The elements of each $\tilde{\omega}$ are modeled independently, i.e., the model for the first element of $\tilde{\omega}_q$ does not depend on the first element of $\tilde{\omega}_{q-1}$

2.3 EXAMPLES

2.3.1 SIMULATED WEIBULL

We performed a simulation study for the Weibull model to determine whether our method was accurately detecting perturbations. We wanted the final data set to have $n = 250$ observations. The failure times were chosen from

$$Y_i|x_i \sim Weibull(\alpha, \exp(\tilde{x}_i'\beta)) \tag{2.10}$$

for $i = 1, \ldots, n$. The vector $\tilde{x}_i = (1, x_i)$ where x_i is a simulated continuous covariate distributed $X_i \sim N(0, 1)$. β is the parameter of the effect of the intercept and covariate $\beta = (\beta_0, \beta_1)'$. Let α affect the rate of the hazard, and $\theta = (\beta', \alpha)'$. Censoring times were created from an Exponential(0.5) distribution. The observed time was the minimum of the failure and censoring time. Overall 32% of the observations were censored.

Next, we added missingness to the covariates. We let the covariate be missing at random. The probability of missing depended on a random Bernoulli(logit$^{-1}(1.5 + 0.2 * y_i)$) variable, where the probability of missing depended on failure time. Overall we had 21.2% missing covariates.

In addition to creating the data set, we needed to have 5 observations for the influence measure to detect. We let the first 245 failure times come from the unperturbed Weibull distribution (2.10), but observations 246-250 came from a perturbed model. Instead of the normal hazard: $h(t|x_i) = \alpha t^{(\alpha-1)} \exp(\beta_0 + \beta_1 x_i)$, we perturb the hazard for the last 5 observations to $h(t|x_i) = \alpha t^{(\alpha-1)} \exp(\beta_0 + \beta_1 * x_i - 2x_i^2)$. The difference of $2x_i$ can be represented in the perturbed hazard as:

$$h(t|x_i, \omega_i) = \alpha t^{(\alpha-1)} \exp((\beta_0 + \beta_1 x_i) - \omega_i), \omega(0) = (0, \ldots, 0). \tag{2.11}$$

The perturbation to the hazard implies we have also pertrubed the log-likelihood:

$$l(p(y|\alpha, \beta, \omega)) = d \log(\alpha) + \sum_{i=1}^{n} [v_i(\alpha - 1) \log y_i + v_i \tilde{x}_i'\beta + v_i\omega - y_i^{\alpha} \exp(\tilde{x}_i'\beta + \omega_i)] \tag{2.12}$$

where d is the number of subjects who died in the study and v_i is an indicator of whether or not the subject was censored.

Having simulated the data, we also needed to come up with priors for β and α. We chose flat priors for both parameters, $\beta \sim N_2\left((0,0)', 10^{-6}I_2\right)$ where I_2 is a 2×2 identity matrix, and $\alpha \sim \Gamma(0.001, 0.001)$. The true values are $\beta = (0.5, -1)'$, and $\alpha = 2$. The prior on the missing covariates was a normal distribution with the mean and variation of the observed covariates.

The geometric tensor for our perturbation to the prior is the $n \times n$ identity matrix. Therefore, the perturbation does not need to be shifted or scaled. The first-order influence measure for the Bayes factor for the i^{th} observation is

$$FI_{RI}[\boldsymbol{v_i}](\omega(0)) = \left(I(v_i = 1) - \int y_i^{\alpha} \exp(\tilde{x}_i'\beta + \omega_i) p(\beta, \alpha | D_o) d\Lambda(\alpha, \beta)\right)^2. \tag{2.13}$$

Table 2.1: FI for unperturbed and perturbed probability densities

Obs	Time	x	P(time$\|x, \theta$)	P(time$\|x, \theta, \omega$)	FI_{Bayes}
246	0.926	-1.89	0.00167	0.0153	17.0
247	0.442	-0.263	1.25	1.15	0.301
248	1.01	-0.208	0.515	0.560	0.701
249	0.451	-1.04	1.63	0.431	0.0756
250	1.19	-1.54	0.000361	0.143	26.8

As Table (2.1) shows, there are two points that have a large FI. Given our perturbation scheme, we suspect that these two points may have a different hazard function that the rest of the data, as the Bayes factor has a large change for observations 246 and 250. As we can see from Table 2.1, observations 247-249 have very small influence measures. Since we know the hazard function that generated these points, we can look at the presumed probability density (p(time$|x, \theta$)), and the true probability density (p(time$|x, \theta, \omega$)). For the observations with high FI values, we see that the observations are much more likely under the perturbed density than the unperturbed density. For the observations with small FI values, it can be seen that the observations do not have a large difference in densities.

Our influence measure is able to detect the perturbation despite missing covariates in the data. Furthermore, the measures are robust, as only large perturbations were detected. From Figure (2.1), we would suspect observations 246 and 250 came from a different hazard since their FI are so much greater than the other observations'. Unfortunately, there is no way to test whether observations are significant or not and the scale is meaningless. "Large" observations are relative to other points.

2.3.2 PIECEWISE EXPONENTIAL

We applied the piecewise exponential model to data from the Eastern Cooperative Oncology Group, who carried out a phase III clinical trial for high dose interferon on multiple melanoma patients. Survival time was defined as the time between enrollment in the study and progression of the tumor or death, whichever came first. In this data set, we have $n = 285$ subjects and either their time to relapse or time to censoring. There were 196 relapses and 89 censored observations. The 4 covariates of interest are treatment (binary), age (17-78 years), Breslow score (0.2-35), and size (0-476). Only the Breslow score which measures the thickness of the tumor and size (area) of the tumor have missing observations (30 and 55 respectively, and 11 with both covariates missing). Breslow and size were log transformed in order to use a normal distribution, and their log values were standardized. Observations with a fail-time of 0 were recorded as 0.01 since survival times are always positive.

The piecewise exponential model is constructed by partitioning the time axis at times $0 < s_1 < s_2 < \ldots < s_J$, where s_J is large enough to ensure that no survival times fall outside of it. Within each of the J intervals $[0, s_1), [s_1, s_2), \ldots, [s_{J-1}, s_J)$ we assume a constant hazard, i.e. an exponential survival function for the interval. We also chose to model three windows. They went from [0,0.378), [0.378, 1.16), [1.16, 10). There were approximately equal number of failures in each window.

The hazard for the i^{th} subject with covariates x_i over the j^{th} time interval is $h(t) = \exp(\phi_j + x_i'\beta)$ for $s_{j-1} \leq t < s_j$. For the complete data distribution, we have two parameters of interest: β, and ϕ. Let β represent the effect of the covariates of interest, which we assume remains the same for all windows and ϕ a vector of the log of baseline hazards for the J time windows. As the range for both parameters is $(-\infty, \infty)$, we use multivariate normal (MVN) priors on both. We model $\beta \sim MVN_4(\mu_0, \Sigma_0)$ where Σ_0 is a large multiple of the 4x4 identity matrix. Let ϕ be modeled as $\phi \sim MVN_{J=3}\left(\gamma_0, diag(a_1, a_2, a_3)\right)$, where $diag(a_1, a_2, a_3)$ is a diagonal matrix whose diagonal is formed by the enclosed vector.

In addition to the complete data distribution, we also need to define the missing variable distribution. This data set has two possible missing covariates. We divide the missing observations into 3 groups: (1) those with only Breslow score missing, (2) only size missing, and (3) both missing. We express the vector of missingness for the i^{th} observation as $x_{i,m}$. When the i^{th} observation is missing only Breslow score, $x_{i,m} \sim N(\eta_1, \sigma_{11})$. When the i^{th} observation is missing only the size, $x_{i,m} \sim N(\eta_2, \sigma_{22})$. When the i^{th} observation is missing both covariates, $x_{i,m} \sim MVN_2\left(\eta = (\eta_1, \eta_2)', \Sigma\right)$, where $\Sigma = \begin{smallmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{smallmatrix}$. The mean

and variance of the observation missing only Breslow or only size is the respective marginal distribution of the observation missing both covariates. We can express all three situations as $x_{i,m} \sim MVN_{d_i}(\eta_{e_i}, \Sigma_{s_i})$, where in the case of only Breslow score missing, $d_i = 1, \eta_{e_i} = \eta_1, \Sigma_{s_i} = \sigma_{11}$. In the case of only size missing, $d_i = 1, \eta_{e_i} = \eta_2, \Sigma_{s_i} = \sigma_{22}$. In the case of both covariates missing, $d_i = 2, \eta_{e_i} = \eta, \Sigma_{s_i} = \Sigma$.

The missing variable distribution has two parameters of interest: η and Σ. η is a 2-dimensional vector of the mean of the missing Breslow scores and sizes while Σ is the covariance matrix of the bivariate distribution. We let $\eta \sim MVN_2(\zeta_0, \Psi_0)$ and model $\Sigma \sim Wishart_2(c_0, d_0 * I_2)$, where c_0 and d_0 are scalars, and I_2 is the 2-dimensional identity matrix, as Σ is a semi-positive definite matrix. All multivariate normal prior distributions have means of 0 and variances of 10^6. We let $c_0 = 2$ and $d_0 = 0.5$ which corresponds to an uninformative prior for the Wishart distribution.

The priors for the parameters may not reflect the true situation. If we have previous data on the covariates of interest, we may be interested in modeling a less flat prior on β. The prior on ϕ assumes that the baseline hazards are independent for each interval. We may also be interested in allowing the mean for each missing observation to vary, so that instead of all the missing breslow and size covariates having the same mean, they would depend on which observation was missing. These changes to our model can be represented as perturbations.

When performing model diagnostics, it is important to assess whether the model is sensitive to these perturbations. We could explore how reducing the variance would effect the model by perturbing the variance terms of β. Our perturbed prior would be

$$\beta|\omega \sim N_4\left(\mu_0, diag(\frac{\sigma_1^2}{\omega_{\beta 1}}, \frac{\sigma_2^2}{\omega_{\beta 2}}, \frac{\sigma_3^2}{\omega_{\beta 3}}, \frac{\sigma_4^2}{\omega_{\beta 4}})\right). \tag{2.14}$$

If we want to explore whether or not the baseline hazards are independent, we can perturb or change the prior on ϕ. We perturb ϕ such that

$$\phi|\omega \sim MVN_3(\gamma_0, AR(\omega)), AR(\omega) = \begin{bmatrix} a_1 & \omega_{\phi 1} & 0 \\ \omega_{\phi 1} & a_2 & \omega_{\phi 2} \\ 0 & \omega_{\phi 2} & a_3 \end{bmatrix}. \tag{2.15}$$

When $\omega_{\phi i} = 0, i = 1, 2$ this corresponds to independence. When $0 \neq \omega_{\phi i}$ this corresponds to some correlation between the baseline hazards.

If we were interested in modeling a separate mean for each missing observation, we could change the missing variable distribution for $x_{i,m}$. Instead of $x_{i,m} \sim MVN_{d_i}(\eta_{e_i}, \Sigma_{s_i})$ we can explore,

$$x_{i,m}|\omega_{imis} \sim MVN_{d_i}\left(\omega_{imis}1'_{d_i} + \eta_{e_i}, \Sigma_{s_i}\right) \tag{2.16}$$

where ω_{imis} has the same dimension as $x_{i,m}$. Written out for each of the three scenarios, we have

$$x_{i,m}|\omega_{imis} \sim N\left(\omega_{imis} + \eta_1, \sigma_{11}\right) \text{ if only Breslow missing} \tag{2.17}$$

$$x_{i,m}|\omega_{imis} \sim N\left(\omega_{imis} + \eta_2, \sigma_{22}\right) \text{ if only size missing} \tag{2.18}$$

$$x_{i,m}|\omega_{imis} \sim MVN_2\left((\omega_{imis1}, \omega_{imis2}) * (1,1)' + \eta, \Sigma\right) \text{ if both missing.} \tag{2.19}$$

Let $\omega_{imis} = (\omega_{imis,1}, \omega_{imis,2})$ if x_i is missing 2 variables and ω_{imis} is scalar if x_i is missing only 1 variable. As we have 72 observations with one or two covariates missing, and 83 values of either Breslow score or size missing. Therefore, ω_{mis} is a vector with 83 elements, one for each missing value. All together, $\omega = (\omega_\beta, \omega_\phi, \omega_{mis})$ where ω_β corresponds to the scaling perturbations on β, ω_ϕ to covariance between the log baseline hazards (ϕ), and ω_{mis} to the missing means of $x_{i,m}$ and $\omega_0 = (1,1,1,1,0,0,0,\ldots,0)_{89\times 1}$.

Using these perturbations, we calculate the complete log posterior to be

$$l(D_m, D_o, \beta, \phi, \eta, \Sigma|\omega) \propto l(D_o|D_m, \beta, \phi) + l(D_m|\eta, \Sigma, \omega_{mis}) + l(\beta|\omega_\beta) + l(\phi|\omega_\phi) + l(\eta) + l(\Sigma^{-1}) \tag{2.20}$$

which we will use to calculate $G(\omega)$ and $FI_{RI}[v]$.

The geometric tensor is a block diagonal matrix composed of three block matrices along the diagonal. The first block matrix in the diagonal corresponds to $G(\omega_\beta)_{4\times 4} = diag(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. The second block matrix corresponds to $G(\omega_\phi)_{2\times 2} = \begin{bmatrix} \frac{1}{a_1 a_2} & 0 \\ 0 & \frac{1}{a_2 a_3} \end{bmatrix} + E_\omega \begin{bmatrix} \frac{\phi_2^2}{a_1 a_2^2} + \frac{\phi_1^2}{a_1^2 a_2} & \frac{\phi_1 \phi_3}{a_1 a_2 a_3} \\ \frac{\phi_1 \phi_3}{a_1 a_2 a_3} & \frac{\phi_2^2}{a_2^2 a_3} + \frac{\phi_3^2}{a_2 a_3^2} \end{bmatrix}$, where a_i is the i^{th} diagonal element of $AR(\omega)$ and E_ω is the expected value with respect to $p(D_c, \theta)$. The third and final block matrix corresponds to $G(\omega_{mis})_{72\times 72} = diag(g(\Sigma))$. $g(\Sigma)$ is a vector with 72 elements, one for each patient with missing observations. Two elements of ω_{mis} are used when a patient has two missing covariates. The i^{th}

element of $g(\Sigma)$ is

$$g_i(\Sigma) = \begin{cases} E_\omega\left(\frac{1}{\sigma_{11}}\right) \\ \text{if the } i^{th} \text{ missing observation only has the Breslow score missing} \\ E_\omega\left(\frac{1}{\sigma_{22}}\right) \\ \text{if the } i^{th} \text{ missing observation only has the size missing} \\ E_\omega\left(\frac{\sigma_{22}}{det(\Sigma)} + \frac{\sigma_{11}}{det(\Sigma)} - \frac{2\sigma_{12}}{det(\Sigma)}\right) \\ \text{if the } i^{th} \text{ missing observation has both missing} \end{cases} \tag{2.21}$$

Since both $G(\omega_\phi)$ and $G(\omega_{mis})$ contain parameters in the expression, we rewrite

$$g_{ij}(\omega) = -\int \partial^2_{\omega_i\omega_j} l_c(\omega) p(D_c, \theta) d\Lambda(D_{com}, \theta)$$

as

$$g_{ij}(\omega) = -\int \partial^2_{\omega_i\omega_j} l_c(\omega) p(D_m, \theta | D_o) p(D_o) d\Lambda(D_m, \theta, D_o).$$

By reformulating the geometric tensor as the expected value of $p(D_m, \beta, \alpha | D_o)$ we can calculate this distribution from ARMS, which is also used to calculate $FI_{RI}[v]$, and $SI_{RI}[v]$. We also use the fact that $G(\omega_\phi)$ and $G(\omega_{mis})$ do not contain any terms involving the observed data so the integral with respect to $p(D_o)$ is trivial.

Since $G(\omega)$ is not proportional to the identity matrix, we need to adjust our calculations for $FI_{RI}[v]$, and $SI_{RI}[v]$ so that measures are on the same scale for comparison reasons. We scale by $G^{-1/2}(\omega_0)$. We can easily calculate the diagonal matrices of $G(\omega_\beta)$ and $G(\omega_{mis})$, and for non-diagonal matrix $G(\omega_\phi)$ we can use spectral decomposition. Details are provided in Appendix A1

To calculate $p(D_m, \theta | D_o)$, we use Gibbs sampling in ARMS. We have 4 elements of β, 3 elements of ϕ, 2 elements of η, 3 elements of Σ, and 83 parameters that correspond to the missing covariates, which means that we have 95 parameters to sample. We had a burn-in of 5,000 and trimmed every 5th observation. In the end, we had 105,000 Gibbs samples. Convergence was monitored using traceplots, auto-correlation, and Geweke statistics. The parameters associated with missing variables did not have great convergence.

Let v_i be the elementary vector with 1 in the i^{th} position. Therefore, $v_1 - v_4$ correspond to the perturbations to β, $v_5 - v_6$ correspond to the perturbations to ϕ, and $v_7 - v_{78}$ correspond to the perturbations to the means of the missing observations. There are 72 elements since these are calculated by subject.

For $i \in \{1, 2, 3, 4\}$ which corresponds to scaling the variance of the $i^{th}\beta$ parameter

$$FI_{RI}[v_i] = \left(\int \left(\frac{\sqrt{2}}{2} - \frac{\sqrt{2}(\beta_i - \mu_{0i})^2}{2\sigma_{0i}^2} \right) * p(\beta|D_o)\Lambda(\beta) \right)^2. \tag{2.22}$$

Since $G(\omega_\phi)$ is not diagonal, calculating the $FI_{RI}[v]$ is more difficult. For ω_ϕ, when $i \in \{1, 2\}$ which corresponds to exploring covariance between ϕ_1, and ϕ_2, and ϕ_2 and ϕ_3, respectively

$$d(\text{RI})[\mathbf{v}_{i+4}](\boldsymbol{\omega}(0)) = \int \frac{-(\phi_i - \gamma_{0i})(\phi_{i+1} - \gamma_{0i+1})}{a_i * a_{i+1}} p(\phi|D_o) d\Lambda(\phi) \tag{2.23}$$

For further details regarding the calculation of $FI_{RI}[v_5], FI_{RI}[v_6]$, see Appendix A1.

For ω_{miss}, when $i \in \{1, 2, \ldots, 72\}$, which corresponds to the observations with one or two missing covariates, then

$$FI_{RI}[v_{i+6}] = \begin{cases} \left(\int \left(\frac{(x_{imis1} - \eta 01)}{\sigma_{11}} \right) p(x_{i,m}|D_o)\Lambda(x_{i,m}) \times E_\omega(\frac{1}{\omega_{11}})^{-1/2} \right)^2 \\ \text{if } i^{th} \text{ missing observation has breslow score missing} \\ \left(\int \left(\frac{(x_{imis2} - \eta 02)}{\sigma_{22}} \right) p(x_{i,m}|D_o)\Lambda(x_{i,m}) \times E_\omega(\frac{1}{\omega_{22}})^{-1/2} \right)^2 \\ \text{if } i^{th} \text{ missing observation has size missing} \\ \Big[\int \Big(\frac{\sigma_{22}}{det(\Sigma)}(x_{imis1} - \eta_{01}) + \frac{\sigma_{22}}{det(\Sigma)}(x_{imis2} - \eta_{02}) \\ -\frac{\sigma_{12}}{det(\Sigma)}(x_{imis1} + x_{imis2} - \eta_{01} - \eta_{02}) \Big) p(x_{i,m}|D_o) d\Lambda(x_{i,m}) \\ \times E_\omega(\frac{\sigma_{22}}{det(\Sigma)} + \frac{\sigma_{11}}{det(\Sigma)} - \frac{2\sigma_{12}}{det(\Sigma)})^{-1/2} \Big]^2 \\ \text{if } i^{th} \text{ missing observation has both missing} \end{cases} \tag{2.24}$$

Recall that $E_\omega()$ is the expected value with respect to $p(D_c, \omega)$, and the power is taken after the expected value. The FI for β and the missing observations have already been scaled in their expression.

Table 2.2: First order influence measures for E1684 data

FI in direction of	β_1	β_2	β_3	β_4	ϕ_1	ϕ_2
	0.125	0.125	0.125	0.125	2.3e-4	5.4e9

As we can see from the Table 2.2, there is only a small change in the curvature of the Bayes Factor as we scale the variance due to FI's small value. Because we have put the the $\beta's$ on the same scale, the equivalent FI values show that there is no β that is more influential than the others. This is mostly due to the fact that the hyperparameter for the variance of the $\beta's$ is on the orders of magnitude larger than the mean of the $\beta's$.

Table 2.2 also suggests that there is only a minor change if we decided to take into account covariance between the log baseline hazards in the first and second window, however the covariance between the log baseline hazards in the second and third window has a large effect on the Bayes factor.

The missing data perturbation, as shown in Figure (2.2), shows a similar story. The FI values vary from 1.7e-4 to 85.1, which indicates not much influence over the Bayes factor. The 3 highest FI's come from observations that are missing both covariates, but these values are not much higher than the other FI's. The lack of high FI's suggests that none of the observations is particuarly influential on the Bayes factor.

2.4 SIMULATIONS

To access the variability of the FI measures, we created 100 bootstrap iterations with sample size 285 (the original sample size) with replacement from the ECOG data and linked the FI measures back to the original observations. Figure 2.3 shows that the measure has a lot of variability and is dependent on the sample. An observation might have high FI in one bootstrap sample and low FI in another. Since this is the case, it is prudent to consider not only the original data's FI, but also how high a percentage of the observations might be. To this end, we consider both the 75^{th} and 90^{th} percentile.

We could consider an observation as having high influence if the observation had a large value in the original data set and the 75^{th} percentile was above a certain cutoff. By forcing the 75^{th} percentile to be above a certain cutoff implies that 25% of the bootstrap sample is above the cutoff. The FI graph of the original values, Figure 2.2, suggests that 50 might be a reasonable cutoff. Figure 2.4, which shows the inter-quartile range versus the observation number overlayed with the original FI values, suggests that 30 might be a good cutoff for the 75^{th} percentile in this data set. Using this cutoff suggests that there 6 potential influential observations listed in Table 2.3, only one of which has an original FI above 50 (observation number 75).

Another measure to consider are the observations whose 90^{th} percentile is above a certain cutoff. Figure 2.5 shows the interdecile range versus the observation number overlayed with the original FI values and suggests that 60 might be a good cutoff in this situation. Using this cutoff we see that there are 9 possible

outliers (Table 2.4), all of which have both variables missing, while in the cutoff using the 75^{th} percentile, there were some observations that only have Breslow missing. This suggests that the tails of the distribution of the FI measure could be related to the number of variables that are missing for a particularly observation. Using this cutoff, we see there are 5 observations whose original FI values are greater than 50 (observation numbers 75, 81, 156, 184, and 189) and whose 90^{th} percentile is high.

Table 2.3: Observations with high 75^{th} percentiles

RFS (yrs)	Relapse	Trt	Std. log(Bres.)	Std. log(size)	Std. Age (yrs)	count	FI	75^{th} Pctl
0.39	1	0	9999	9999	-2.30	75	63.07	35.07
1.84	0	1	9999	9999	-0.03	161	4.71	31.21
2.15	0	1	9999	-0.37	0.12	171	35.87	35.55
6.31	0	0	9999	0.35	-1.33	232	38.81	34.35
7.97	0	1	9999	-0.84	-0.93	264	39.68	33.37
8.29	0	1	9999	9999	1.41	272	8.78	41.16

Table 2.4: Observations with high 90^{th} percentile values

RFS (yrs)	Relapse	Trt	Std. log(Bres.)	Std. log(size)	Std. Age (yrs)	count	FI	90^{th} Pctl
0.26	1	1	9999	9999	-0.33	53	0.01	80.66
0.39	1	0	9999	9999	-2.30	75	63.07	68.07
0.44	1	1	9999	9999	1.10	81	63.17	63.58
1.70	1	1	9999	9999	-0.99	156	65.69	85.39
1.84	0	1	9999	9999	-0.03	161	4.71	95.77
3.02	1	1	9999	9999	0.32	184	4.93	97.37
3.28	1	0	9999	9999	-0.70	189	82.34	91.21
4.89	0	1	9999	9999	-0.52	210	85.13	71.85
8.29	0	1	9999	9999	1.41	272	8.78	151.09

[1]

The Tables 2.3 and 2.4 suggest that we have perhaps 1 observation whose 75^{th} and 90^{th} percentile are different from the rest of the data, and whose original FI was high. However, this observation does not seem a cause for concern given the high variability of the FI measure.

[1]RFS: Relapse Free Survival time
Relapse is 1 if relapse occured
Trt is 1 if high interferon dose used, 0 if observation
A value of 9999 indicates missing observation

2.5 DISCUSSION

The purpose of performing model diagnostics is to ensure that the chosen model is not sensitive to small changes, and if the model is sensitive to small changes, we need to be forthwright with investigators about the model's sensitivity. As investigators, we may be interested in changes that that are a result of our modeling or prior assumptions.

In this paper, we used both simulated data and a real data set to explore how our modeling and prior assumptions effected the Bayes factor. Our first simulation example explored perturbations to the model by perturbing the hazard function. By knowing the truth behind the simulations, we can verify that the diagnostics are accurately picking up perturbed observations. We see that the influence measures pick up big changes, and is not sensitive to smaller changes. For diagnostics, that is what we want as we are only interested in knowing about potential large changes to our model.

For our ECOG data set, we explored perturbations to the variances of β, ϕ, and the influence of individual observations. Our perturbation to the variances of β and the covariances of ϕ represent perturbations to the prior assumptions. The FI measure for β were largely dominated by the original hyperparameter variance for β. In non-informative cases, the variance for β is chosen to be many orders of magnitude larger than the hyperparameter mean of β. Since the original hyperparameter for the variance of β appears in the denominator of the FI measure, it ends up dominating the term. No one variable is more influential in the model unless its estimate is large compared to the hyperparameter variance.

We see from our bootstrap sample of the ECOG data set, that the FI measure has much variability. It gives a sense that "large" FI should only be those whose order of magnitude are 2 to 3 times the majority of the data. Knowing this, we suspect that there are not any unduely influential observations in the ECOG data set.

There are many ways of measuring sensitivity. With Bayesian influence measures, we measure how small perturbations will effect a measure of interest. Using graphical methods, we determine which changes to the model will have a large effect on our measure of interest.

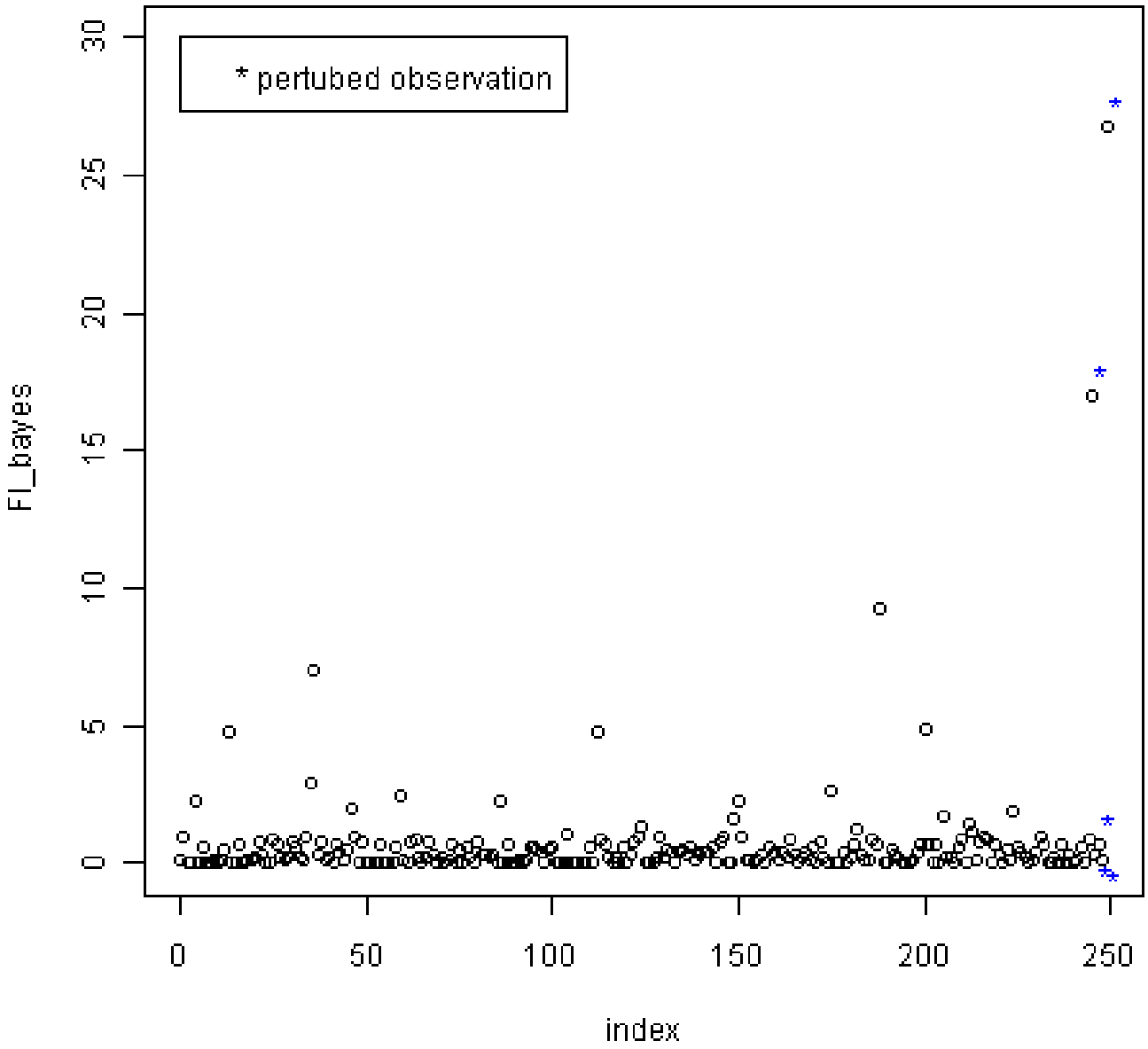

Figure 2.1: First-order influence measure for Bayes factor

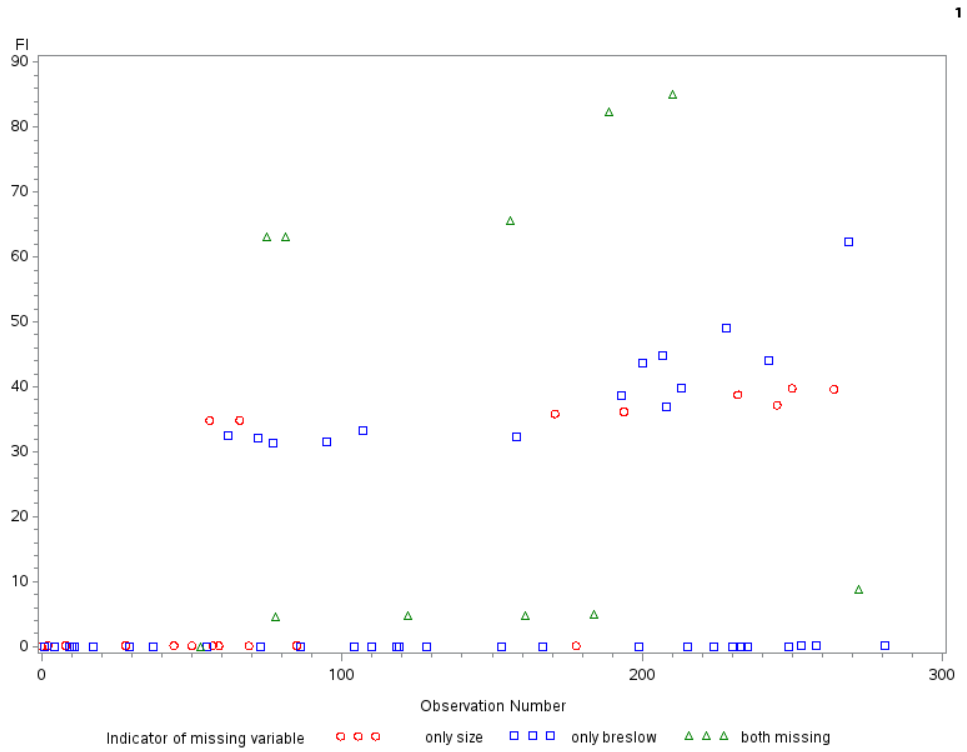

Figure 2.2: First order influence measures for missing observations

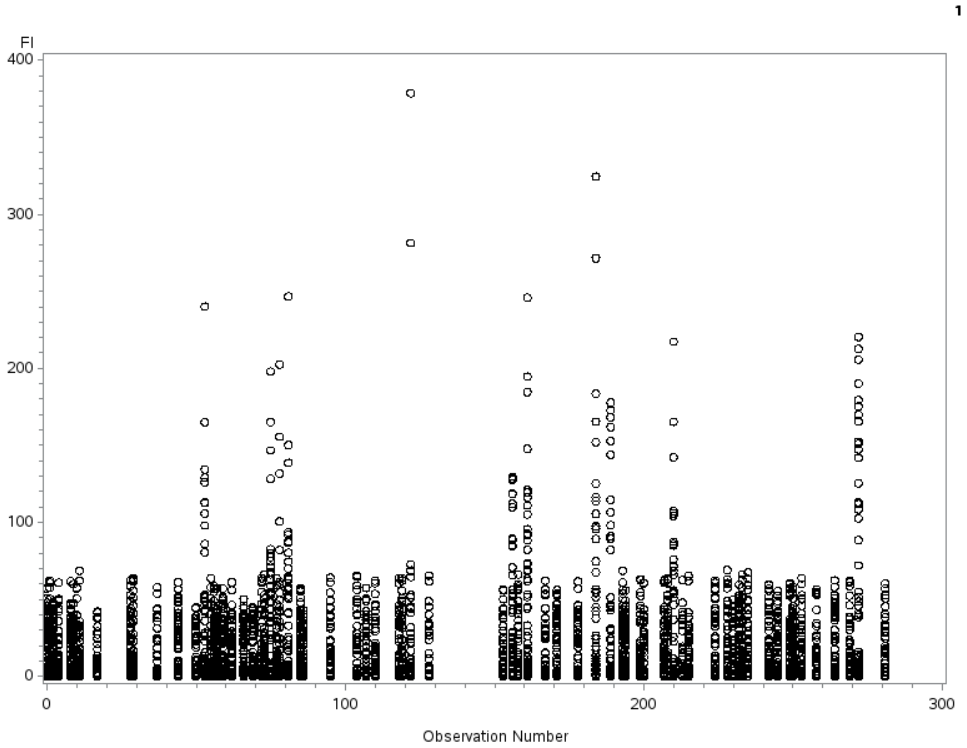

Figure 2.3: Graphic of FI measures across all 100 iterations

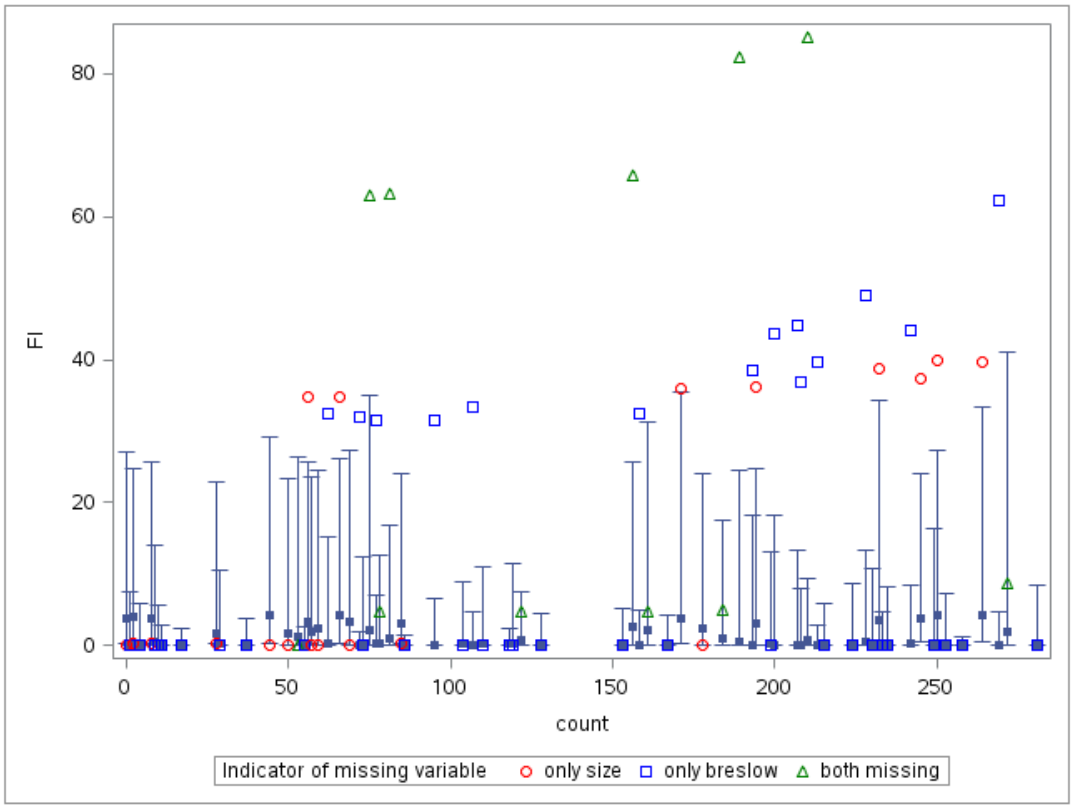

Figure 2.4: Original FI values versus interquartile range (IQR) of 100 bootstraps

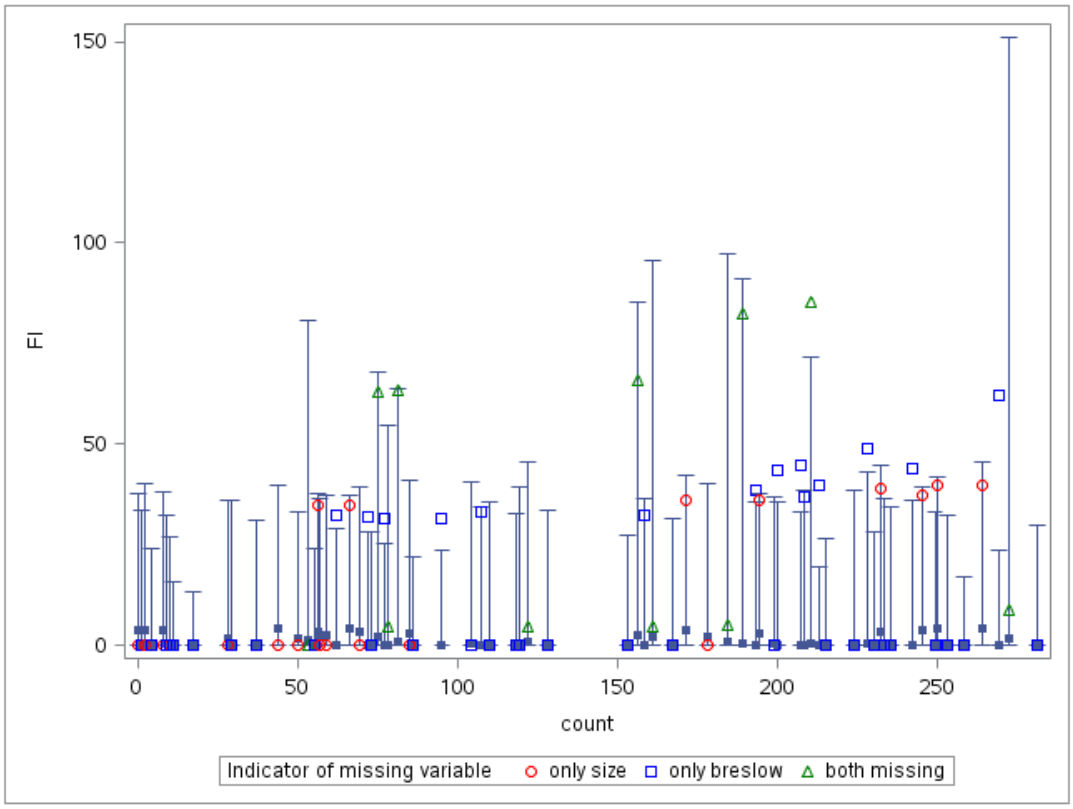

Figure 2.5: Original FI values versus interdecial range (IDR) of 100 bootstraps

CHAPTER 3: ROC ANALYSIS IN THE PRESENCE OF VERIFICATION BIAS

3.1 INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is the fourth leading cause of death in the United States, affecting more than 5% of the adult population (Lin et al., 2008). Despite its high prevalence, COPD is not often diagnosed until it has reached advanced stages. According to the Global Initiative for Chronic Obstructive Lung Disease (GOLD), COPD is defined as air flow limitation that is not fully reversible, is gradually progressive, and is associated with an abnormal inflammatory lung response to noxious particles or gases (Rabe et al., 2007), making it hard to diagnose. Four out of five COPD patients have ever smoked or are current smokers, which puts smokers in a high risk population. Still there have been some attempts to screen the general population and not just high risk patients.

The main method of diagnosing a patient with COPD is office spirometry. When the subject blows into an office spirometer, it measures variables such as the forced expiratory volume in 1 second (FEV1) and forced volume capacity (FVC) and assigns a quality grade (A-F) that indicates the validity of the reading. A passing grade (A-C) is sometimes difficult to achieve for patients who have to blow into the device for as long as 10 seconds. Enright and Kaminsky (2003) suggest that instead of asking general practitioners to do spirometry, technicians should perform the test and then have general practitioners interpret the results.

One of the problems with screening for COPD is that there is no universally accepted definition of a COPD case, which makes estimating population prevalence difficult. GOLD defined a COPD case as someone whose FEV1/FVC < 0.7. The American Thoracic Society (ATS) recommends that cutoffs be age and gender specific, as lung capacity is a function of these two variables (Enright and Kaminsky, 2003). Hankinson et al. (1999) showed that FEV1 decreases with age and tends to be higher in males than in females. The third National Health and Nutrition Examination Survey (NHANES III) also has its own recommendation for COPD diagnosis. NHANES III surveyed a random sample of the U.S. population from 1988 to 1994 and collected spirometry data on 20,627 participants ages eight and older who came from a selection of races (Hankinson et al., 1999). The NHANES III study developed reference equations to determine the

lower limit of normal for FEV1/FVC for different populations. In the end, the significant covariates in the equations were gender, height, race, and age. With so many ways of defining a COPD case, estimates for population prevalence vary. Some reports calculate a COPD prevalence of 4.5% while others calculate a 21.1% prevalence (Wilt et al., 2005).

With all the difficulties of using spirometry to determine COPD, the U.S. Preventive Services Task Force (USPSTF) was formed in order to make a recommendation on whether using spirometry to screen for COPD is effective. The USPSTF reviewed COPD studies from 1966 to 2007. Lin et al. (2008) did a review of the studies and found no papers provided direct evidence on health outcomes associated with screening for COPD. They concluded that screening using spirometry would require testing hundreds of patients to identify a single exacerbation and would likely identify people with mild or moderate airflow obstruction who would not experience any adverse health benefits attributable to COPD, and was therefore not an effective screener for COPD in the general population.

Since the Agency for Healthcare Research and Quality (AHRQ) did not recommend spirometry for COPD screening (Qaseem et al., 2007; Lin et al., 2008) for healthy adults who do not report symptoms to a clinician, there has been a need to develop and evaluate alternatives to spirometry for detecting COPD. Nelson et al. (2012) staged a large COPD screening study in the general adult population to evaluate pocket spirometers. Ideally, screening tools should be affordable, simple, and accurate enough to avoid producing large numbers of false positives or false negatives (Marshall, 1996a,b). Pocket spirometers cost approximately $30 and require a short exhalation, making them ideal for large screening studies of the sort described in Nelson et al. (2012). These devices, however, have not been validated for accuracy as a population-wide screener. Both FEV1 and peak expiratory force (PEF), which is the maximum speed of exhalation, can be measured with a pocket spirometer. Nelson et al. (2012), selected PEF as the measurement to be used for screening in their general population study. The main question of interest is whether there is a significant difference between the performance of the pocket spirometer screener on PEF and FEV1 when used on the entire population of interest, and not just the study population.

We will first describe the study in Section 3.2. In Section 3.3 we report some descriptive summaries of the data, which provide the numbers we need in Section 3.4, where we describe our method to address our three aims. Section 3.5 describes the results. In Section 3.6 we use simulations to explore how our method works in different scenarios. Section 3.7 discusses the strengths and limitations of our method.

3.2 EXAMPLE

The study by Nelson et al. (2012) recruited subjects from events such as health fairs, health expositions, and national conventions of older adults in cities where a large number of adults were expected to attend. These subjects represented the population at risk for COPD. A "Mobile Spirometry Unit" trailer or 10×10 foot booths advertised free lung tests. Subjects provided informed consent, but otherwise their information was linked only with a non-descriptive ID number. Patients provided their gender, age, height, and race in order to better calculate the lower limit of normal of PEF for a person with those characteristics, also referred to as the predicted PEF. Then all patients answered a questionnaire used to identify whether they were at high risk for COPD and performed at least 3 maneuvers, i.e. breaths, into the pocket spirometer. Pocket spirometers from Vitalograph asma-1 Kansas City, MO were chosen because they met ATS accuracy standards for PEF and FEV1 and were efficient for a large number of screener tests due to their minimization of cross-contamination between patients by using disposable one-way mouth pieces.

Severe COPD is associated with low values of PEF and the study was designed to sample more heavily from those with low PEF values. In Stage 1 of sampling, all patients whose PEF was less than 70% predicted for their age, height and race were asked to complete a spirometry exam. In Stage 2 of sampling (which takes place after Stage 1), every 10^{th} person who did not meet the Stage 1 requirement of having a PEF less than 70% was asked to complete an office spirometry exam, which produced a 10% random sample of the at risk population. The office spirometer used was ndd EAsyOne Frontline, Zurich Switzerland, which is a diagnostic-quality office spirometer that meets ATS accuracy standards and minimizes cross-contamination by using disposable plastic mouthpieces (Nelson et al., 2012). The FEV1 and FVC values to determine COPD status (severe versus non-severe) were measured using the office spirometer. The technician recorded the maximum measure of three passing attempts (grades A-C).

Although the GOLD standard uses FEV1/FVC$< 70\%$ of predicted as the cutoff for determining a COPD case, an alternative measure was used. The GOLD standard assumes a post-bronchodilator spirometry, while pre-bronchodilator spirometry was used in this screener study. Therefore a less conservative definition of severe COPD cases was used. For this purpose, FEV6 was used to estimate FVC and additionally FEV1$< 60\%$ of predicted was required. The final definition for a severe COPD case by office spirometry were subjects with FEV1/FEV6$<$ lower limit of normal (given age, height, gender, and ethnicity) and FEV1$< 60\%$.

The population that the study wanted to screen with the pocket spirometers corresponds to potential COPD patients. Therefore never smokers with no asthma-like symptoms, women who were pregnant, and people who had tuberculosis, a current respiratory infection, or eye surgery, major surgery or heart attack during the previous 30 days were excluded. With exclusion of these groups, there is a better target population at risk for COPD without confounding from other factors accounting for restricted air flow. The initial plan to target the COPD risk population was to limit pocket spirometry to those subjects who had two or more risk factors according to the questionnaire, which asked about COPD risk factors, such as age, daily wheezing, productive cough, asthma, and recent smoking activity. Since it was determined that performing PEF on all participants would add little time or expense, all patients were screened on the pocket spriometer regardless of the number of risk factors identified on the questionnaire.

Although there was a clearly defined population to screen, not everyone wanted to be screened. There were patients who refused to do either pocket spirometry or office spirometry, while some were unable to perform office spirometry and achieve a passing grade. There were also people who had normal PEF values but who volunteered for spirometry. When calculating the sensitivity and specificity, it is important to weigh these volunteers properly. Since these people were self-selecting into the study, it is important that they were given a different weight from the rest of the patients in Stage 2 as they were not part of the random sample.

Since not everyone was asked or was able to perform office spirometry to produce an interpretable disease status, it is necessary to have weights according to the sampling scheme. The purpose of this analysis was to use an innovative application of sampling weights to evaluate the performance of the screener in identifying COPD cases, confirmed with the use of diagnostic quality spirometry. There are two main issues that the sampling scheme needs to address. First, there is verification bias. As noted in Begg and Greenes (1983), not everyone is sent on for disease verification, which will artificially inflate our sensitivity and deflate our specificity since a higher percentage of diseased patients will be evaluated in our study compared to the general population. Secondly, we need to address the fact that we have some patients whose disease status is uninterpretable even after being evaluated for the disease. Disregarding these patients whose disease status is unknown will lead to bias in our estimates (Matchar et al., 1990).

We had three main aims: 1) to generalize our results for estimates of the screener to the entire screened population, 2) determine if PEF$< 70\%$ was a good cutoff for identifying potential COPD cases, and 3) to determine if PEF was the best screener variable to differentiate between disease statuses. We were interested in providing weighted analysis of sensitivity and specificity so that the results could be more generally

applicable, and we were also interested in using the data for alternative analyses for which the study was not necessarily designed, such as the evaluation of a different screener variable.

3.3 DATA

Data were collected between June 2008 and December 2009. A total of 5,761 people provided demographic data and completed the risk assessment questionnaire. Of these, 5,638 people performed the peak flow screener, and this subset comprised the study population. 315 of the 5,638 subjects had peak flow $< 70\%$ predicted so they also were asked to performed office spirometry. 251 of the 315 completed office spirometry and of the 251, only 179 had spirometry maneuvers with adequate quality grades (A-C). In summary, of the 315 subjects whose pocket spiromety PEF $< 70\%$ predicted, 179 had passing maneuvers, 72 had failing maneuvers, and 64 did not complete the office spirometry. There were 5,323 participants with PEF $\geq 70\%$ predicted, of which 651 underwent spirometry. Of the 651 spirometry participants, 107 were volunteers who requested spirometry while the remainder was a random 10% sample. Only 550 of the 651 had passing spirometry maneuvers. Table 3.1 provides a more detailed break down of the sample that underwent both the screener test and office spirometry.

The study population of 5,638 participants provided a good representation of the population at risk that was the target. The majority of participants was female (n=3,262 57.9%). The predominant race was White/Hispanic (n=4,932 87.5%) while Blacks made up the remaining population (n=706, 12.5%). The average age was 54.41 years with a minimum of 12 and a maximum of 93. Furthermore, 84.2% (n=4,745) were over 40 years of age, which is considered a risk factor for COPD.

The questionnaire was based on other studies which used short questionnaires to detect adult smokers who have a high probability of having COPD. The risk factors considered in the questionnaire were the presence of daily wheezing and productive cough, asthma diagnosis, physical activity limitations due to breathing, particle (i.e. smoke, chemical, dust) exposure, smoking status, and if so, how recently smoking occurred. A total of 2,243 (39.8%) had been smokers with 641 (11.4%) having smoked in the past 6 months. Furthermore, 3,450 (61.2%) had particle exposure. The rest of the risk factors were found in roughly 20% of the population.

3.4 METHOD

3.4.1 EXTRAPOLATING TO THE POPULATION AT RISK

In order to generalize the office spirometry results so that they were more representative of the people who were screened, there needs to be a weighting scheme. In the group with abnormal PEF, which was defined to be less than 70% of predicted PEF, the 179 adequate spirometry results needed to represent the 315 that were supposed to complete spirometry. Therefore, each person in this group, where everyone was also asked to perform spirometry, was given a weight of 1.76. This assumes that those who did not complete spirometry are similar to those who failed to get a passing grade. In the group of 123 volunteers, whose whose PEF was greater than 70% but who self-selected into the office spirometry portion of the study, there were 107 acceptable office spirometry maneuvers, so they were given a weight of 1.15. The volunteers needed to be removed from the 550 acceptable spirometry results, as well as the 5,323 population count as they were not randomly selected. Therefore, the remaining 443 would represent the non-volunteering population of 5,200 (5,323-123) and were given a weight of 11.74. Table 3.2 shows how the weights were calculated. The use of weights to account for those whose disease status was not observed seems reasonable given that the mean of the PEF values for those who were able to produce an acceptable disease status is very similar to the mean of the PEF values for those who were unable to produce an acceptable disease score for volunteers, those below the cutoff, and those above the cutoff (Table 3.3).

Since we had sampling weights in our data set, it was important to make sure that our analysis correctly accounted for them in the variance. For calculation of confidence intervals using the sampling weights, we used SAS-callable SUDAAN, which is a sampling program that calculates variance taking into account the sampling design. We treated the study as a single population, and we assumed simple random sampling with replacement and unequal probabilities of selection in our design specification.

Using the weights, we also graphed the sensitivities and specificities of different screener cutoffs on a weighted Receiver Operator Curve (ROC). An ROC is used to show the trade-off between sensitivity and specificity as the cutoff of the continuous screener variable changes. In our study, sensitivity is defined as the probability of the pocket spirometer PEF measure being below the cutoff given that the person has COPD based on office spirometry analysis. Specificity is the probability of the screener being negative at the same cutoff given that the person does not have COPD. As the cutoff for PEF increases, our sensitivity increases, thus identifying more patients for office-grade spirometry, leaving fewer cases behind; however, at the same

time, our specificity decreases. The ROC plots the sensitivity by 1-specificity. A good measure has high sensitivity and specificity.

We used SUDAAN and SAS to make a graphical depiction of the ROC and the 95% confidence intervals of sensitivity and specificity about different cut points. We measured the confidence intervals of a range of cutoffs that went from 55% to 85% by increments of 5% to get an overall picture of how PEF would perform in the general population at differen cutoffs.

3.4.2 Evaluating cutoff

One way to determine an ideal cutoff is to use the cutoff that corresponds to the point that is farthest from the one-to-one line. The one-to-one line represents a test that is random, i.e. sensitivity and specificity add to 1. The point farthest away from this line would indicate the cutoff that has the best sensitivity given the tradeoff with specificity, assuming sensitivity and specificity are equally valued.

The study was designed to prioritize the negative predictive value of the screener. Researchers wanted to be confident that a negative diagnostic of the screener truly represented a non-severe COPD case. Therefore, the false negatives needed to be minimized while maintaining an acceptable true positive rate. With respect to sensitivity and specificity, maximizing the negative predictive value means minimizing the ratio of $\frac{\text{1-sensitivity}}{\text{specificity}}$ while maintaining an acceptable overall specificity, as shown below, where ρ is the disease prevalence in the general population

$$\begin{aligned}
\text{NPV} &= \text{P(True Negative}|\text{Test Negative)} \\
&= \frac{(1-\rho)\text{specificity}}{(1-\rho)\text{specificity} + \rho(1-\text{sensitivity})} \\
&= \left(1 + \frac{\rho}{1-\rho}\frac{1-\text{sensitivity}}{\text{specificity}}\right)^{-1}.
\end{aligned}$$

From Figure 3.2, we see that the ratio is smaller for larger values of sensitivity, so long as the sensitivity is not 1. Therefore cutoffs with larger sensitivities will also have larger negative predictive values.

3.4.3 PEF vs FEV1

In addition to assessing different cutoffs values for PEF as a screening tool, we wanted to analyze whether PEF was the best variable to use for the screener. Since the pocket spirometers measured both PEF and FEV1,

we analyzed the data as if we had used FEV1 as the screener variable. The FEV1 measure from the pocket spirometer has less accuracy than the office spirometer measured value, although the advantage is that the pocket spirometer is easier to use.

Another measure that ROC's provide is the area under the curve (AUC), which can be used as the measure of discrimination for the screener variables. DeLong et al. (1988) developed a non-parametric correlated AUC test that is applicable to comparisons of correlated measures in the study population (i.e. two screeners that were measured on the same person). However, for generalization to the general population, incorporation of sampling weights is needed. The Kawaguchi et al. (2011) variation of the DeLong et al. (1988) test statistic will be used to incorporate sampling weights. In Kawaguchi et al. (2011), they formulate the test statistic as a ratio of two estimators. The numerator is the estimated probability that a random pair of patients come from different disease groups and the diseased patient has a lower value for the k^{th} response variable (assuming low values are associated with disease). Here, the k^{th} response corresponds to PEF for $k = 1$ and FEV1 for $k = 2$. The denominator pertains to the estimated probability that a random pair of patients are from different disease groups (severe versus non-severe COPD). The weighted estimators of the numerator and denominator statistic can be written as:

$$\begin{aligned}
\hat{\theta}_{1k} &= \sum_{j=1}^{N} \sum_{j' \neq j}^{N} U_{1jj'k} \\
U_{1jj'k} &= \frac{w_j w_{j'} I[(t_j - t_{j'})(Y_{jk} - Y_{j'k}) > 0]}{w_j w_{j'}} \\
\hat{\theta}_{2k} &= \sum_{j=1}^{N} \sum_{j' \neq j}^{N} U_{2jj'k} \\
U_{2jj'k} &= \frac{w_j w_{j'} I[(t_j - t_{j'}) \neq 0]}{w_j w_{j'}}
\end{aligned}$$

where $t_j = 1$ if the subject has severe COPD and -1 if not, Y_{jk} is the value of the j^{th} patient' k^{th} response, and w_j is the j^{th} patient's sampling weight. The test statistic to determine if there is a significant difference between the two measures can be written as $\hat{\theta} = \frac{\hat{\theta}_{11}}{\hat{\theta}_{21}} - \frac{\hat{\theta}_{12}}{\hat{\theta}_{22}}$. Let $\hat{\theta}$ be a statistic that estimates the difference between the AUC's for PEF and FEV1, and $\hat{\theta}_k = \frac{\hat{\theta}_{1k}}{\hat{\theta}_{2k}}$ is a statistic that estimates the AUC for response k where $k = 1$ indicates the AUC for PEF and $k = 2$ indicates the AUC for FEV1.

We use the weighted estimates U_{1j*k}, U_{2j*k}, which are the respective weighted averages of $U_{1jj'k}$ and $U_{2jj'k}$ over all $j' \neq j$ to formulate $\hat{\theta}_k$ as the weighted mean of two estimates using the following calculations:

$$\hat{\theta}_k = \frac{\sum_{j=1}^{N}\sum_{j'\neq j}^{N} w_j w_{j'} U_{1jj'k} \Big/ \left(\sum_{j=1}^{N}\sum_{j'\neq j}^{N} w_j w_{j'}\right)}{\sum_{j=1}^{N}\sum_{j'\neq j}^{N} w_j w_{j'} U_{2jj'k} \Big/ \left(\sum_{j=1}^{N}\sum_{j'\neq j}^{N} w_j w_{j'}\right)}$$

$$\hat{\theta}_k = \frac{\left[\sum_{j=1}^{N} w_j \left\{\sum_{j'\neq j}^{N} w_{j'} U_{1jj'k} \Big/ \sum_{j'\neq j}^{N} w_{j'}\right\} \Big/ \sum_{j=1}^{N} w_j\right]}{\left[\sum_{j=1}^{N} w_j \left\{\sum_{j'\neq j}^{N} w_{j'} U_{2jj'k} \Big/ \sum_{j'\neq j}^{N} w_{j'}\right\} \Big/ \sum_{j=1}^{N} w_j\right]}$$

$$U_{1**k} = \sum_{j=1}^{N} w_j U_{1j*k} \Big/ \sum_{j=1}^{N} w_j$$

$$U_{2**k} = \sum_{j=1}^{N} w_j U_{2j*k} \Big/ \sum_{j=1}^{N} w_j$$

$$\hat{\theta}_k = \left(\sum_{j=1}^{N} w_j U_{1j*k} \Big/ \sum_{j=1}^{N} w_j\right) \Big/ \left(\sum_{j=1}^{N} w_j U_{2j*k} \Big/ \sum_{j=1}^{N} w_j\right)$$

$$= (U_{1**k}/U_{2**k}).$$

While calculating the weighted mean of $\hat{\theta}$ is easily implemented, calculating the variance is harder. The unweighted variance is calculated by creating a vector of values for the j^{th} patient ($j = 1, \ldots, N$), where the vector is expressed as $F_j = (U_{1j*1}, U_{1j*2}, U_{2j*1}, U_{2j*2})$. The first subscript of U_{ij*k} corresponds to the group that's being estimated: $i = 1$ if estimating the probability that for a random pair of patients in which one has severe COPD and one does not, the patient with severe COPD will have the larger test result, and $i = 2$ if estimating the probability that a randomly chosen pair will have different COPD statuses. The last subscript refers to the screener variable being measured. Let U_{1j*k} pertain to the probability that the j^{th} subject has a higher k^{th} response value than any other subject in the opposite COPD class, and can be thought of as the weighted average of $U_{1jj'k}$ over all values of $j' \neq j$. Let U_{2j*k} pertain to the probability that any

other subject is from the opposite COPD class. In terms of formulas we have

$$U_{ij*k} = \frac{\sum_{j' \neq j}^{N} w_{j'} U_{ijj'k}}{\sum_{j' \neq j}^{N} w_{j'}}$$

for $i = 1, 2$. Our statistic needs to incorporate both measures as a function of F so we have

$$\begin{aligned}
\hat{\theta} &= \hat{\theta}_1 - \hat{\theta}_2 \\
&= \frac{U_{1**1}}{U_{2**1}} - \frac{U_{1**2}}{U_{2**2}} \\
&= C_1 \exp\{A_1 \log(\bar{F})\}
\end{aligned}$$

where

$$C_1 = \begin{bmatrix} 1 & -1 \end{bmatrix},$$

$$A_1 = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \text{ and}$$

$$\bar{F} = \frac{\sum_{j=1}^{N} w_j F_j}{\sum_{j=1}^{N} w_j}$$

We can use Taylor series expansion of $\hat{\theta}$ to get

$$\text{Var}(\hat{\theta}) = \begin{bmatrix} 1 & -1 \end{bmatrix} A_1 * D_{a_2} * W_F * D_{a_2} * A_1' * \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

where D_{a_2} is a diagonal matrix whose diagonal is the vector $a_2 = \bar{F}^{-1}$ and W_F is the covariance matrix of $\bar{F}$ taking into account sampling weights.

Details for the unweighted covariance calculation of the average of F (V_F) shown below come from Davis and Quade (1968)

$$V_F = \frac{4}{N(N-1)} \sum_{j=1}^{N} (F_j - \bar{F})'(F_j - \bar{F}).$$

Since the weighted counterpart of $\frac{1}{N(N-1)} \sum_{j=1}^{N} (F_j - \bar{F})'(F_j - \bar{F})$ can be obtained from SAS-callable SUDAAN, it's multiplication by 4 produces W_F, with weights taken into account. We can test the null hypothesis for 2 measures by testing $H_0 : \theta = 0$ using the test statistic $X = \frac{\hat{\theta}^2}{\text{Var}(\hat{\theta})}$. By the central limit theorem, this statistic has a χ^2_1 distribution.

We can expand the methodology to accommodate the situation in which we have 3 screeners and want to determine whether at least one of the tests has an AUC that is significantly different from the others. In this situation $\bar{F} = (U_{1**1}, U_{1**2}, U_{1**3}, U_{2**1}, U_{2**2}, U_{2**3})$, the weighted covariance matrix of $\bar{F}$ (W_F) is $4\times$(Covariance matrix from SUDAAN). We still multiply by a factor of 4 since we are still comparing 2 states, diseased vs non-diseased, at a time.

Now our null hypothesis is $H_0 : \theta_1 - \theta_2 = \theta_1 - \theta_3 = 0$. Now $\theta = (\theta_1 - \theta_2, \theta_1 - \theta_3)'$

$$\text{Var}(\hat{\theta}) = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} A_1 * D_{a_2} * W_F * D_{a_2} * A_1' * \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}$$

where

$$A_1 = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}.$$

We can test the null hypothesis for 3 measures using the test statistic $X = \hat{\theta}' \text{Var}(\hat{\theta})^{-1} \hat{\theta}$. By the central limit theorem, this statistic has a χ^2_3 distribution.

3.5 RESULTS

3.5.1 EXTRAPOLATING TO THE POPULATION AT RISK

Of the 179 adequate spirometry results for those subjects whose PEF was $< 70\%$ predicted, 113 would be considered to have COPD according to spirometry. Of the 550 who had normal peak flow and adequate spirometry results, only 30 would be considered to have COPD according to spirometry. With the weights, 199 of the 315 with low PEF values have severe COPD by spirometry, and 267 of the 5,323 with normal PEF values would be expected to have severe COPD by spirometry. The un-weighted and weighted results are compared in Tables 3.1 and 3.4. The screener has a weighted sensitivity of 0.4270, and a weighted specificity of 0.9776 versus an unweighted sensitivity and specificity of 0.7902 and 0.8873, respectively. Using SUDAAN to calculate the weighted confidence intervals of our screener, the 95% confidence interval for weighted sensitivity is (0.3228, 0.5377) and for weighted specificity is (0.9711, 0.9826). Table 3.4 uses the weights to calculate the estimates of the population had everyone undergone both pocket and office spirometry.

In order to illustrate how different PEF cutoffs of $55\% - 85\%$ would have performed, we use Figure 3.1, which provides the ROC combined with confidence intervals. Table 3.5 provides the actual estimates for the range of cutoffs. The inclusion of the cutoff of 88% is explained in more detail in the following section.

3.5.2 EVALUATING CUTOFF

If we wanted to consider sensitivity and specificity as equally important, we could use the weighted ROC curve to find the point farthest away from the one-to-one line and to use the data to determine which cutoff that point corresponded to. We determined that a PEF cutoff of 88% would have increased sensitivity without too much of a loss in specificity (Figure 3.3).

If instead we wanted to maximize the negative predictive value of our screener while generalizing to the general population, we could find the cutoff point that was farthest from the $y = \sqrt{x}$. The point that was farthest from the curve corresponded to a PEF cutoff of 91% for the pocket spirometer. With the higher cut off, our sensitivity would have been 0.9510 and our specificity would have been 0.6809. The weighted sensitivity would have been 0.8455 and the weighted specificity would have been 0.7854. Tables 3.6 and 3.7 provide a comparison of the full 2×2 tables.

In order to maximize the negative predictive value of a screening test based on a pocket spirometer, we recommend using a cut off value of 91% instead of 70% for PEF. The reason why we were we wanted to maximize the negative predictive value even at the cost of more false positives is because the negative effects of a false positive were minimal. Sending more people to spirometry was considered less undesirable than failing to identify someone who might have COPD. 70% was already considered a conservative cutoff in the sense that we wanted mostly severe COPD cases in stage 1 of sampling, but the analysis from the $y = \sqrt{x}$ curve suggests we could have been less conservative.

Our analysis calculated the cutoff corresponding to the farthest point from the $y = \sqrt{x}$ curve in order to emphasize sensitivity, however if specificity were a higher priority, we could have calculated the cutoff corresponding to the $y = x^2$ curve instead. This would represent the case when we wanted to maximize the positive predictive value. One scenario in which this might occur would be when the diagnostic exam is prohibitively expensive or may be potentially dangerous. For COPD, however, the consequences of a false positive for the screener are minimal, as the consequences of a false positive, office spirometry, is neither costly nor harmful.

3.5.3 PEF vs FEV1

Another question of interest was whether PEF was the best variable to use for the screener using AUC as the measure of discrimination. The AUC is an estimate of the probability that the measure of someone in the non-diseased group will have a higher response variable value than someone in the diseased group, where the response variable is either PEF or FEV1 in this situation. The FEV1 measure from the pocket spirometer has less accuracy than the office spirometer measured value, although the advantage is that the pocket spirometer is easier to use. Using the same weights a weighted ROC curve using FEV1 as the screener was calculated, and the results are shown in Figure 3.4. There were 23 observations missing screener FEV1 data. The overall mean of the group (n=5,615, $\bar{FEV1} = 0.8470$) was used for these missing observations so that the same weights in the PEF situation could be used. The weighted AUC for PEF is 0.8740, while the weighted AUC for FEV1 is 0.9002.

The weighted test statistic to test the difference between correlated AUC curves is $\frac{0.02619^2}{0.0483} = 0.0142$, with a p-value of 0.9051 from a χ^2 distribution with 1 df. The data suggests that there is not enough evidence to show a significant difference between AUC using PEF and FEV1. As the AUC for FEV1 is higher, the data suggests that FEV1 on the pocket spirometer would have been a better predictor of severe COPD than PEF.

Table 3.1: PEF Screener versus COPD status by office spirometry, un-weighted.

	Non-severe COPD by spirometry	Severe COPD by spirometry	Total
PEF $\geq 70\%$	520	30	550
PEF $< 70\%$	66	113	179
Total	586	143	729

Table 3.2: Calculation of weights

Category	# observed disease status	# screened	Weight
PEF< 0.7	179	315	1.76
PEF≥ 0.7 and randomly selected	443	5200	11.74
PEF≥ 0.7 and volunteered	107	123	1.15

This is graphically confirmed by Figure 3.4. As we are interested in high sensitivity in order to have a high negative predictive value, the FEV1 test performs better than the PEF test for high sensitivities. However we will use a simulation to further explore the performance aspects of our weighted statistics.

3.6 SIMULATIONS LOOKING AT SAMPLING ASSUMPTIONS, NO VOLUNTEERS

We examined the assumptions of our estimator using four different simulations. In all the simulations, D was the true disease status for the 5,000 patients. $D \sim$ Bernoulli(π) where 1 indicated the patient had the disease, and π is the prevalence of the disease. Y_1 was the value of the first screener and Y_2 was the value of the second screener. $Y_1|D \sim N(\mu_0 + \mu_1 * D, \sigma_1^2)$ and $Y_2|D \sim N(\beta_0 + \beta_1 * D, \sigma_2^2)$. R_1 and R_2 simulated the sampling rules, where 1 indicated that the subject was sent on for disease verification. $R_1|Y_1, D = (Y_1 < c)$ where c was the cutoff for the screener that sent all subjects on to disease verification. R_2 was 1 for every 10^{th} individual who has $Y_1 \geq c$. We let W_1 and W_2 represent the whether or not the patient was able to get an interpretable disease status. If the disease was interpretable, W_1 or W_2 was 1. For those who had $Y_1 < c$, $W_1|R_1 = 1, Y_1, D \sim$ Bernoulli$(1/w_1)$ where the magnitude of $w_1 \in [1, \infty)$ represents how hard it is for the patients in the first phase of the sampling scheme to produce an interpretable result. If these patients are easily able to produce interpretable disease results, then w_1 is close to 1. $W_2|R_1 = 0, Y_1, D \sim$ Bernoulli$(1/w_2)$ is the variable indicating interpretable disease results in the second sampling stage. All simulations were repeated 510 times. These following values were assigned to the simulation parameters: $\hat{\pi} = 0.2$, $\hat{\mu}_0 = 1.016$, $\hat{\mu}_1 = -0.4$, $\hat{\sigma}_1 = 0.25^2$,$\hat{\beta}_0 = 0.832$, $\hat{\beta}_1 = -0.37$, $\hat{\sigma}_2 = 0.2^2$ and w_1 and w_2 were varied depending on the situation.

In the first set of simulations, we explored whether the impact of non-ignorable verification bias on our estimator. We made the probability of getting an interpretable disease status, W_1 and W_2, dependent on

Table 3.3: Comparison of PEF between 3 groups of subjects

		Volunteers	Below cutoff	Above cutoff, without volunteers
Observed disease status	#	107	179	443
	Mean	0.98	0.55	1.07
	Std Dev	0.21	0.12	0.21
Designated or volunteered for office spirometry, but unobserved disease status	#	16	136	85
	Mean	0.97	0.58	1.03
	Std Dev	0.2	0.13	0.19
Total group	#	123	315	5200
	Mean	0.98	0.56	1.09
	Std Dev	0.21	0.12	0.2
Not selected for office spirometry	#	.	.	4672
	Mean	.	.	1.1
	Std Dev	.	.	0.2

Table 3.4: PEF Screener versus COPD status by office spirometry, weighted

	Non-Severe COPD by spirometry	Severe COPD by spirometry	Total
PEF $\geq 70\%$	5056	267	5323
PEF $< 70\%$	116	199	315
Total	5172	466	5638

Table 3.5: Sensitivity and specificity with 95% confidence interval (CI) for different cutoffs of PEF

Cutoff % for PEF screener	Sensitivity	95% CI for Sensitivity	Specificity	95% CI for Specificity
55	0.2531	(0.1853, 0.3355)	0.9952	(0.9919, 0.9972)
60	0.3211	(0.2393, 0.4155)	0.9929	(0.9890, 0.9954)
65	0.3739	(0.2812, 0.4769)	0.9888	(0.9841, 0.9921)
70	0.4268	(0.3228, 0.5377)	0.9775	(0.9711, 0.9826)
75	0.5365	(0.4096, 0.6588)	0.9339	(0.9116, 0.9508)
80	0.6672	(0.5254, 0.7840)	0.894	(0.8652, 0.9172)
85	0.7451	(0.6005, 0.8504)	0.8428	(0.8083, 0.8720)

Table 3.6: Unweighted values if cutoff of 91% had been used

	Non-severe COPD by spirometry	Severe COPD by spirometry	**Total**
PEF $\geq 91\%$	399	7	406
PEF $< 91\%$	187	136	323
Total	586	143	729

Table 3.7: Weighted values if cutoff of 91% had been used

	Non-severe COPD by spirometry	Severe COPD by spirometry	**Total**
PEF $\geq 91\%$	4062	72	4134
PEF $< 91\%$	1110	394	1504
Total	5172	466	5638

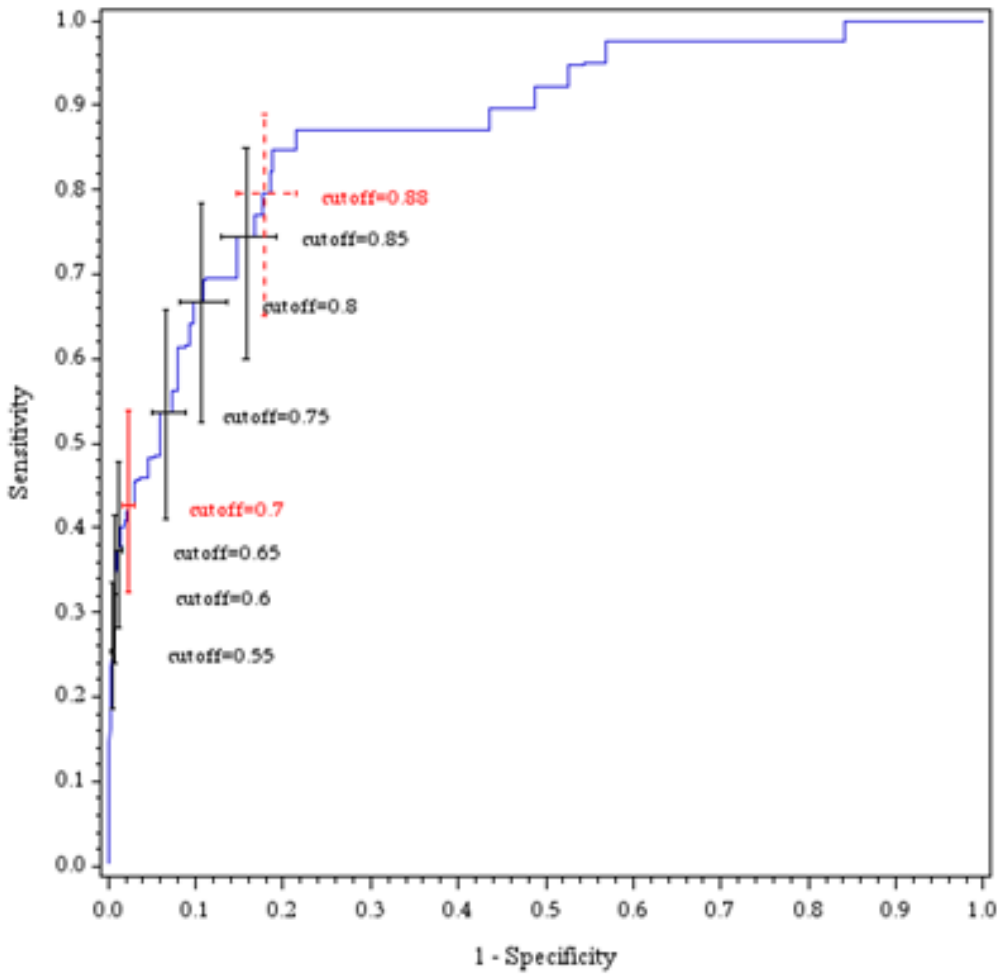

Figure 3.1: ROC with confidece intervals around a variety of cutoffs

the disease status itself. We had $W_1|R_1 = 1, Y_1, D \sim$ Bernoulli $\left(\frac{\exp\{\alpha_0+\alpha_1*D\}}{1+\exp\{\alpha_0+\alpha_1*D\}}\right)$ and W_2 had the same distribution

$(W_2|R_1 = 0, Y_1, D \sim$ Bernoulli $\left(\frac{\exp\{\alpha_0+\alpha_1*D\}}{1+\exp\{\alpha_0+\alpha_1*D\}}\right))$. As the magnitude of α_1 increases, the larger the effect of disease status on disease interpret-ability. We let $\alpha_0 = 0.75$ and α_1 vary from -3 to 3 by 1. No matter the magnitude of α_1, the coverage of the estimator remains high, as shown in Figure 3.5. The weight of the interpretability of diseased patients varied from 1.02 to 10.5. The weight of the non-diseased patients was 1.5. The same held true whether the verification of disease depended on the disease or the screener (Figure 3.5). This was done by replacing the disease status with the screener value and looking at a range of slope and intercept values. $W_1|R_1 = 1, Y_1, D \sim$ Bernoulli $\left(\frac{\exp\{\phi_0+\phi_1*D\}}{1+\exp\{\phi_0+\phi_1*D\}}\right)$ and $W_2|R_1 = 0, Y_1, D \sim$ Bernoulli $\left(\frac{\exp\{\phi_0+\phi_1*D\}}{1+\exp\{\phi_0+\phi_1*D\}}\right)$

In the next simulation, we looked at adjusting the sampling scheme. While the first part of the sampling scheme still sent those under a cutoff on for additional screener, the second sampling stage for those who had a screener variable above the cutoff dependend on the disease status. Instead of R_2 being 1 for every tenth observation for those whose screener was above the cutoff, $R_2|R_1 = 0, Y_1, D \sim$ Bernoulli $\left(\frac{\exp\{\eta_0+\eta_1*D\}}{1+\exp\{\eta_0+\eta_1*D\}}\right)$. Without the random sampling scheme in the second stage of sampling, the coverage of the estimator decreases dramatically.

In the final set of simulations, we explored having a situation in which $W_1|R_1 = 1, Y_1, D$ or $W_2|R_1 = 0, Y_1, D$ did not have the same distribution over their respective domains of Y_1. This would represent a situation in which the probability of disease interpretability depended on a certain cutoff of Y_1 (c_2), but the cutoff is not the same as the cutoff used in the study ($c \neq c_2$). If $c_2 < c$ then $W_1|Y_1 < c_2 < c, D \sim$ Bernoulli$(1/w_1)$, $W_1|c_2 \leq Y_1 < c, D \sim$ Bernoulli$(1/w_2)$, and $W_2|c \geq Y_1 \sim$ Bernoulli$(1/w_2)$. Alternatively, if $c_2 > c$ then $W_1|Y_1 < c < c_2, D \sim$ Bernoulli$(1/w_1)$, $W_2|c \geq Y_1 < c_2, D \sim$ Bernoulli$(1/w_1)$, and $W_2|Y_1 \leq c_2, D \sim$ Bernoulli$(1/w_2)$. Figure 3.6 represents the simulation set-up. Biologically this might happen if more severe cases, represented by either really high or really low values of Y_1 had a harder time producing interpretable results. In our simulation, we let estimates of $w_1, w_2 \in (1.2, 1.4, 1.6, 1.78, 1.8, 2, 3)$ and $c_2 \in (0.6, 0.7, 0.9, 1)$.

This alternation to the simulation created some situations in which our estimator did not have good 95% coverage. In situations for which $w_1 \neq w_2$ and $c \neq c_2$ our estimator of the difference in correlated AUC's had 92.3% 95% coverage (Figure 3.7), but of greater concerns is that the coverage could be as low as 78.2%. When $c = c_2$ the estimator had good 95% coverage. This would imply that the cutoff for disease interpretability was the same as the cutoff for sampling. For example, a disease verification for which obese patients might have a harder time producing interpretable results would want the BMI screener variable cutoff to be the same as the BMI screener cutoff for obeseity. The other scenario for which disease interpretability would be sufficient is when there is no difference in the distribution of disease interpretability given Y_1 ($w_1 = w_2$).

3.7 DISCUSSION

3.7.1 EXTRAPOLATING TO THE GENERAL POPULATION

Our post-hoc analysis suggests that the pocket spirometer's sensitivity is not large when extrapolated to the general population (weighted sensitivity of 0.4270 versus unweighted sensitivity of 0.7902), although specificity improves with extrapolation (0.8873 vs 0.9776). Differences between unweighted and weighted results are expected since we have verification bias. While other methods involve simulation or an estimate of the disease prevalence to extrapolate to the general population (Matchar et al., 1990; Begg and Greenes, 1983), our method does not require either and uses only data that we have collected. This is ideal for situations in which disease prevalence estimates are unknown, or for which the current literature vary greatly.

3.7.2 EVALUATING CUTOFF

According to our analysis, the cutoff of 70% may have been a conservative choice, in the sense that it is overly strict and we are not screening many of the severe COPD patients who are above the 70% cutoff. Cuoffs as high as 88% could have been used to increase specificity without too much of a decrease in sensitivity.

3.7.3 PEF VS FEV1

Our sampling scheme not only allows us to easily measure the AUC and correlated AUC between PEF and FEV1, but it also provides an easy way to test the difference between correlated AUC's in the presence of verification bias and unknown disease states. We will compare our method with the method of Rotnitzky et al. (2006).

In Rotnitzky et al. (2006) paper, the authors deal with verification bias by first modeling the probability of being sent on for disease verification given the test result and other measured covariates, and then modeling the probability of disease given that the disease was verified. Using the estimates from these models, they create an estimator for the disease status based on the model parameters, test result, and other measured covariates. The details are provided in the Appendix 2.

Using Rotnitzky et al. (2006) method, the difference in the AUC's is -0.0369, with a variance of 0.17.The test statistic is 0.0081, with a p-value of 0.9282. While both methods yield similar test statistics (ours was 0.0142), one of the advantages of using Sudaan to calculate the weighted variance of the test statistic is that it eliminates the modeling step required by Rotnitzky et al. (2006). We do not need to be concerned about

model misspecification, or missing covariate data. Furthermore, while our method may produce a smaller difference between the correlated AUC's, the variance is also smaller than the method of Rotnitzky et al. (2006).

A disadvantage of the Rotnitzky et al. (2006) method is that it cannot accomodate variables for which a parameter cannot be estimated. There are two situations in which this occurs in this particular example. The first situation is that Rotnitzky et al. (2006) method cannot discriminate between people whose disease verification was unknown because they were not selected for disease verification and the people whose disease verification was unknown because they were unable to produce a passable office spirometry reading. While this could be viewed as a potential non-ignorable disease situation, we still need the disease status to form parameter estimates. By having the people whose disease status was missing due to failure to produce an acceptable office spirometry reading being estimated by people who were sent on for disease verification, we get a more accurate estimate of their disease status since the results are being estimated from a more equivalent population. The second situation that yields unestimable parameter estimates is the volunteer population.

By stratifying the unknown disease status with the test results, we avoid having to simulate marginal sensitivity and specificities for our data as in Matchar et al. (1990). At the same time, this allows us to take into account their data when calculating variances.

These simulation results show us some strengths, weaknesses, and underlying assumptions of our method. Including a group for volunteers produces results that less biased than estimates not including volunteers. The simulations show us that the assumption that our sample is random for those whose screener variable is greater than the cutoff is extremely important. Coverage decreases dramatically when the random sample actually depends on disease status. The simulations also illustrates some assumptions our method makes about the disease interpretability mechanism. While the assumption of uniform disease interpretability is reasonable, it shows us that some knowledge of the interaction between disease and disease verification may rule out this method in certain situations. While we assume that the interpretablilty of the disease behaves consistently within each group, one must keep in mind that the group is chosen by an untried cutoff.

Our weighting scheme allows us to generalize our results, and our alternative methods for the ROC analysis allow us to emphasize sensitivity more than specificity. Furthermore, these methods are adaptable to situations that have different characteristics than COPD.

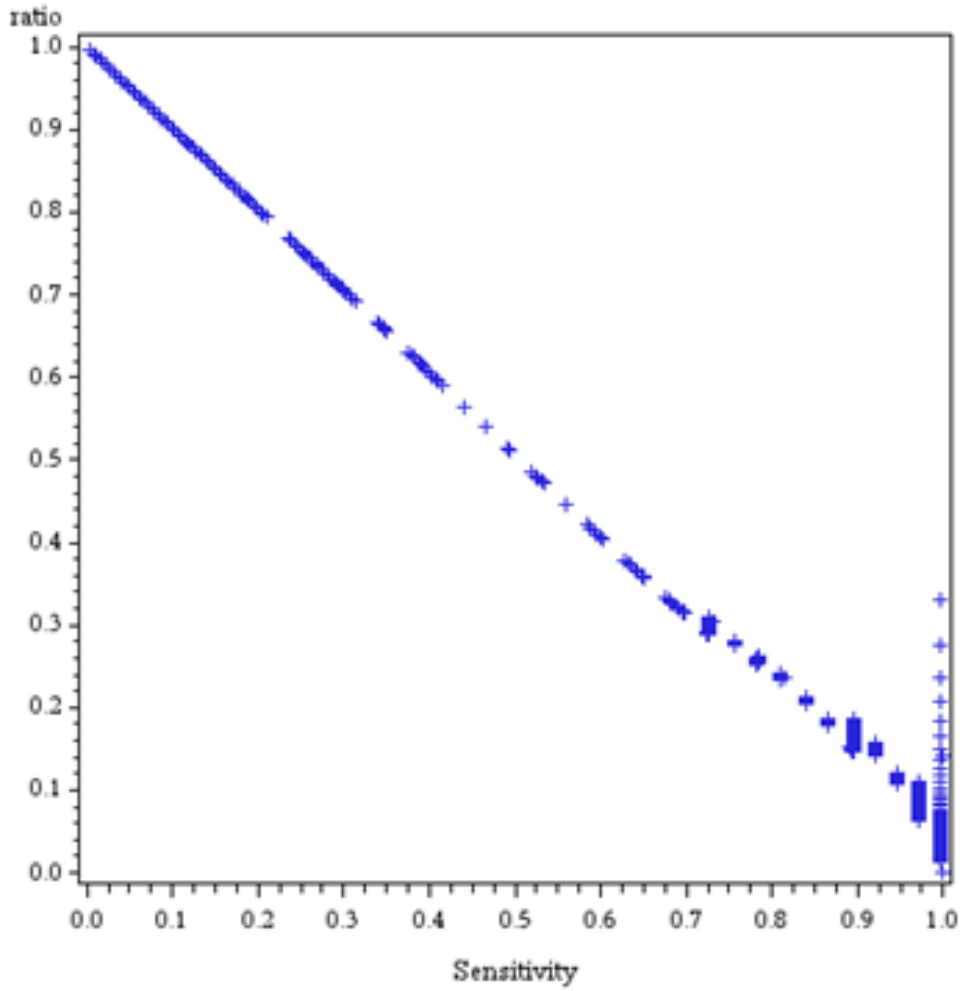

Figure 3.2: Ratio ($\frac{1-\text{sensitivity}}{\text{specificity}}$) vs sensitivity

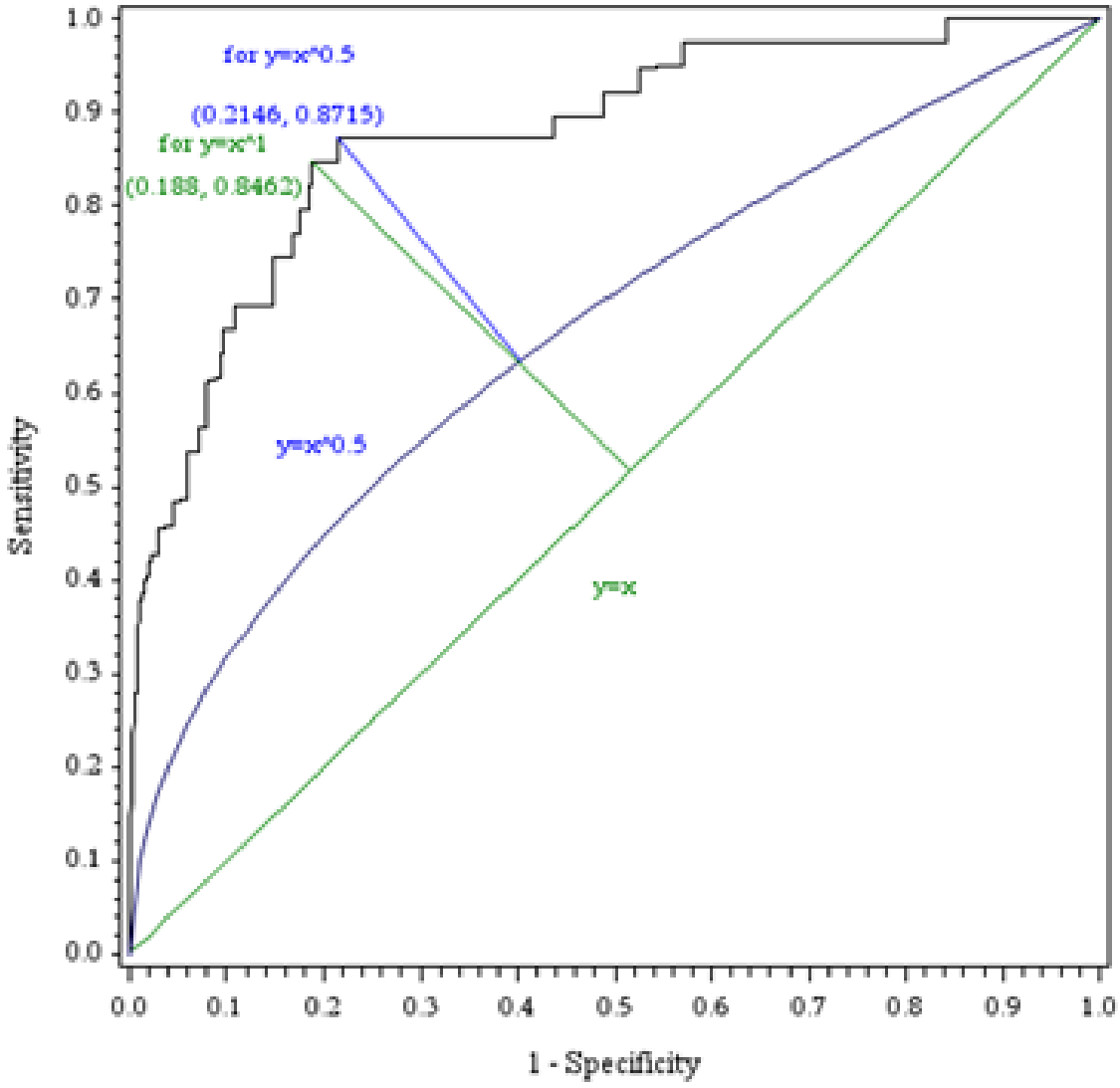

Figure 3.3: ROC of PEF to the $y = x$ line and the $y = \sqrt{x}$ curve

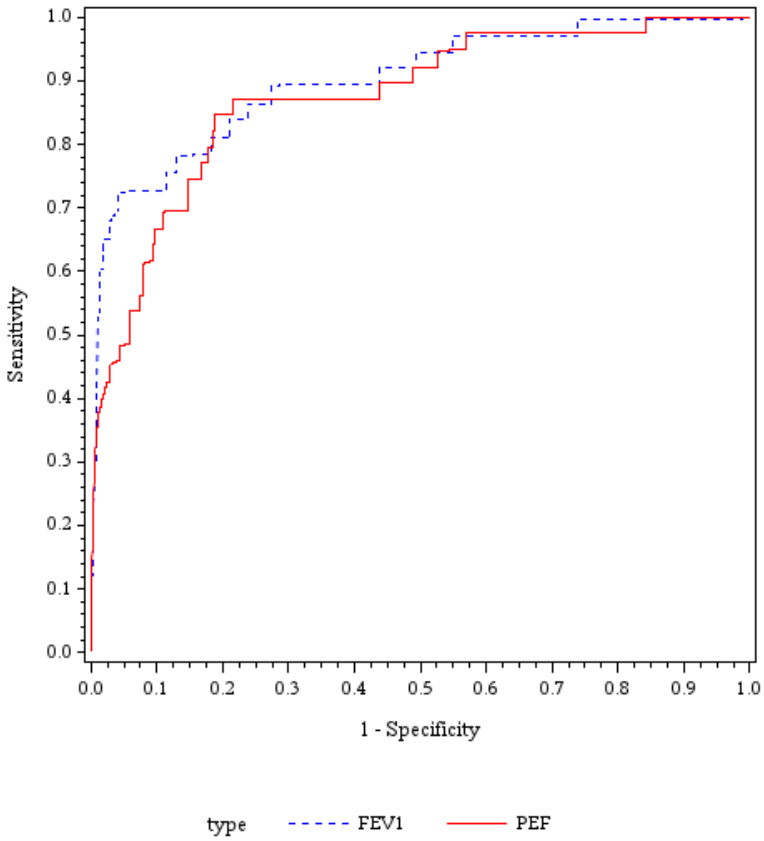

Figure 3.4: Weighted ROC of PEF and FEV1

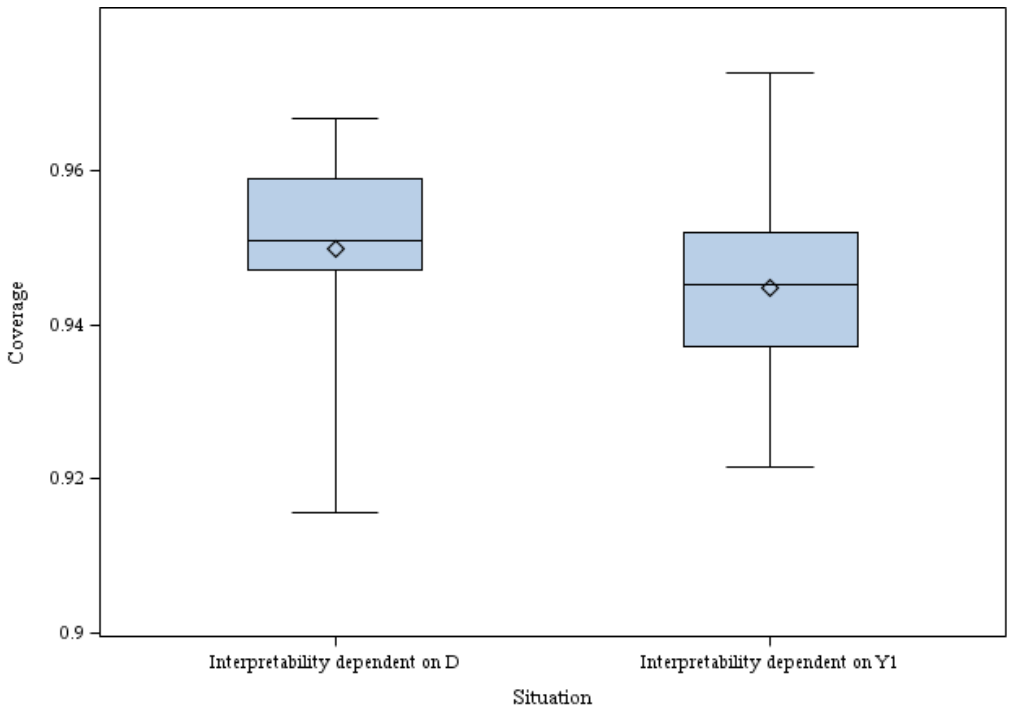

Figure 3.5: Summary of coverage of simulations exploring non-ignorable verification bias

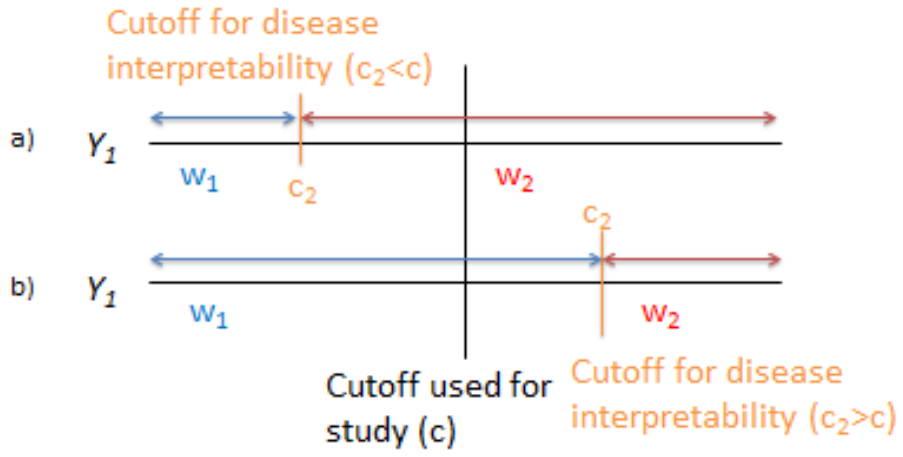

Figure 3.6: Graphic representation of components changed in the simulation

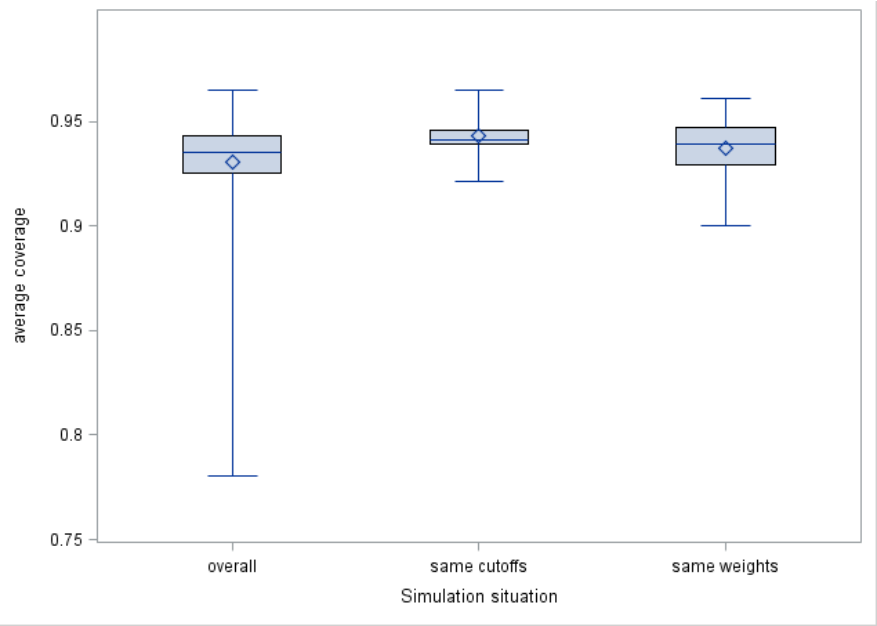

Figure 3.7: Summary of coverage of simulations exploring mismatch of study cutoff and disease interpretability cutoff

CHAPTER 4: RANDOMIZATION-BASED NON-PARAMETRIC COVARIANCE ADJUSTMENT FOR CATEGORICAL OUTCOMES WITH REPEATED MEASURES

For randomized studies, the main reason for adjusting by covariates is to increase precision of the treatment parameter. For randomized clinical trials, Robinson and Jewell (1991) showed that for logistic regression, adjusting for predictive covariates results in greater power for efficacy when testing for treatment effect. Furthermore, Rosenblum and van der Laan (2009) showed that a certain class of hypotheses are robust even when the model is mis-specified, meaning that even if we do not have the right covariates in the model, our hypothesis test will still have the correct type I error.

After adjusting for covariates, we still have multiple ways of modeling the treatment effect, depending on the outcome variable. Oftentimes, the outcome variable in clinical settings is a dichotomous variable, but there has been some work to extend methodology to an ordinal outcome. A few different methods deal with ordinal outcome variables. One such model is the proportional odds model developed by McCullagh (1980). This model requires that the effect of the covariates be the same regardless of which cutoff differentiates between a success and failure for the odds. This assumption is called the proportional odds assumption and the only way to determine if covariates have a proportional odds for the outcome is to look at the data. The partial proportional odds model uses both the ordinal nature of the outcome, and relaxes the proportional odds assumption for the covariates in the model. The advantage of this model is that for covariates for which proportional odds holds, we only need one parameter, but we still have flexibility for covariates which do not satisfy the proportional odds assumption.

In order to state an analysis plan that will be appropriate regardless of the data distribution, statisticians have developed semi-parametric and non-parametric methods that do not rely on distributional assumptions of the data. In Hastie and Tibshirani (1987), they developed non-parametric methods for logistic regression. They use a local scoring algorithm for the logistic model where estimates of the predicted probability for each subject is estimated with a back-fitting algorithm. In the back-fitting algorithm each covariate's effect is estimated using a scatterplot smoother on the residuals. The scatterplot smooths allow for a wide variety of covariate adjustments to take place. Alternatively, Zhang et al. (2008) developed a semi-parametric model for

logistic regression. They implement a three step algorithm. In the first step, the unadjusted treatment effect is calculated using estimating equations which produces a consistent, asymptotically normal estimator. Then using that estimate of treatment effect, each subject's contribution to the estimating equation is calculated for all treatment levels. Using this basis for the third and final step, they calculate the expected contribution to the estimating equation for each treatment given the covariates using parametric models. Therefore, each treatment effect has its own model and covariate effects. Using these parameters in the estimating equation, they estimate the final adjusted treatment effect. The adjustments of the estimating equation means that hypotheses testing also requires covariate adjustments to score equations and the like.

Our method is similar to Zhang et al. (2008) in that we first start with the unadjusted treatment estimate and then adjust it through some function of the covariates. However, instead of modeling covariate effects, we adjust the estimate non-parametrically through the correlation of the covariates with the treatment effects. Our method also differs in the hypothesis testing method in that a direct variance for the adjusted treatment effect is calculated making hypothesis testing similar to classical methods. We propose extending the non-parametric covariate adjustment methods that Tangen and Koch (1999) used in repeated measures randomized clinical trial with missing data for outcomes.

This chapter is organized as follows: in Section 4.1 we explain and demonstrate our methodology for binary outcomes on a respiratory data set. We expand our methodology to ordinal outcomes in Section 4.2 and address stratification variables in Section 4.3. In Section 4.4 we apply the method to a clincial trial evaluating two doses in the treatment of Cushing's disease, and finally, we discuss the strengths and limitations of our method in Section 4.5

4.1 METHOD

Our method consists of four steps. First, we model the treatment effect for outcomes nested within each visit. Let V be the number of visits, with some visits having randomly missing outcome data for some patients and M be the number of covariates we wish to adjust by, with all M covariates having complete data. The V treatment effects allow the treatment effect to differ by visit, which may be the case when there is a delay or tapering of treatment effect. We define the vector $\boldsymbol{d}$ as the vector of the V treatment by visit effects for outcomes stacked on top of the M differences of the means of the covariates of the treatment from

the placebo. Therefore $\boldsymbol{d}$ is a $(V + M) \times 1$ vector. We will refer to the M differences of the means of the covariates of the treatment from the placebo as the differences of the means of the covariates for brevity.

The second step is to calculate the covariance matrix of $\boldsymbol{d}$, which involves calculating three covariance matrices: the covariance matrix of the treatment effects for the V outcomes, the covariance matrix of the M differences of the means of the covariates, and the covariance matrix of the treatment effect for outcomes with the covariates. We can calculate the covariance matrix of the treatment parameters for outcomes by multiplying the dfbetas associated with the treatment effect (Hammill and Preisser, 2006), which is explained later in further detail. Next we calculate the usual covariance matrix of the differences of the means of the covariates based on sums of cross-products of deviations. The covariance matrix between the treatment effect for outcomes and the difference of the means of the covariates is derived by multiplying the dfbetas from the model for outcomes by their counterparts for the differences of the means of the covariates on the basis of comparability to corresponding estimators based on U-statistics (Saville and Koch, 2013). The covariance matrix of $\boldsymbol{d}$, $\boldsymbol{V_d}$, is a $(V + M) \times (V + M)$ matrix.

In the third step we force the differences in means of covariates to 0 through weighted least squares (WLS) regression. We regress $\boldsymbol{d}$, which contains both treatment effects for outcomes and differences in the means of the covariates, onto the space of only treatment parameters. We use WLS with the weights based on $\boldsymbol{V_d^{-1}}$. Forcing the differences in means of the covariates to 0 through WLS is appropriate due to randomization; any differences observed in the distribution of the covariates is due to chance. From WLS theory we obtain both the estimator for the adjusted treatment effects ($\boldsymbol{b}$) and its covariance matrix ($\boldsymbol{V_b}$).

Finally, since $\boldsymbol{b}$ is approximately multivariate-normally distributed when sample sizes are sufficiently large, we can test hypotheses using traditional contrast matrices and associated asympotitc theory for test statistics. We can conduct tests of homogeneity of treatment effects across visits, and assess the overall significance for treatment effects.

In the case of ordinal outcomes, we use partial proportional odds in order to avoid the proportional odds assumption. Let P be the number of categories. Our treatment is tested within visit and outcome cutoff designating favorable versus unfavorable outcome. The method proceeds as described previously, only the dimensions of the vector $\boldsymbol{d}$ are $[(P-1)V + M] \times 1$, and $\boldsymbol{V_d}$ is a $[(P-1)V + M] \times [(P-1)V + M]$ matrix. Additionally, we can test the proportional odds assumption using the appropriate contrast matrix.

When we have stratification and sufficiently large sample size in all strata, we can apply our method on each stratum. Each stratum will have its own adjusted treatment effect, and we then take a weighted average

(using weights such as Mantel-Haenszel's) to get an overall adjusted treatment effect and covariance matrix. The method requires at least 30 subjects in each treatment assignment of each stratum. This is because there needs to be enough subjects for the generalized estimating equations (GEE) used in longitudinal analysis to enable approximate normality for adjusted treatment effects and create consistent variance estimates. Since we use Chi-square tests in our hypothesis testing, we assume that $\boldsymbol{V_d}$ is essentially known through consistent estimation.

4.1.1 LOGISTIC LONGITUDINAL MODEL

We will look at the binary outcome variable for a trial with one treatment and one control group. Let $Y_{htj} = 1$ if the h^{th} subject for the t^{th} treatment group reported favorable outcomes at the j^{th} visit, and 0 otherwise. Let $t = 1$ if the subject is on the treatment and 2 if the subject is in the control group. Let $h = 1, \ldots, n$ where n is the total number of subjects in the study, and $j = 1, \ldots, V$ where V is the number of visits in the study. We will assume that the design of the trial forces subjects to have the same visit schedule.

First we will model reference intercepts for each visit and treatment effects nested in visits. Let $\pi_{tj} = Pr(Y_{tj} = 1)$ which is the probability that a subject in the t^{th} treatment group will have a favorable outcome at the j^{th} visit. Let α_j be the parameter that represents the intercept for the j^{th} visit, and β_j be the parameter that represents the treatment effect at the j^{th} visit. The longitudinal logistic model is

$$\begin{aligned} log\left(\frac{\pi_{tj}}{1-\pi_{tj}}\right) =& \alpha_1 * (j=1) + \ldots + \alpha_V * (j=V) \\ &+ \beta_1 * (j=1) * (t=1) + \ldots + \beta_V * (j=V) * (t=1) \end{aligned} \tag{4.1}$$

where ( ) is the indicator that the expression within the parenthesis is true. This model does not have an overall intercept, and each visit is allowed to have its own effect for treatment and control. For example, the model for the treatment effect at the first visit on the log odds of a favorable event is

$$log\left(\frac{\pi_{t1}}{1-\pi_{t1}}\right) = \alpha_1 + \beta_1 * (t=1)$$

where α_1 is the parameter for the log odds of a favorable outcome for the control group at visit 1 and $\alpha_1 + \beta_1$ is the parameter for the log odds of a favorable outcome for the treatment group at visit 1. Therefore, the vector of treatment effects is $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_V)$.

Let n_t be the number of subjects in the t^{th} treatment group and $n_1 + n_2 = n$. We create $\boldsymbol{R}_1 = (\boldsymbol{r}_{11}, \ldots, \boldsymbol{r}_{1n_1})'$ which is the $(n_1 \times V)$ matrix of unstandardized cluster-level dfbeta's for the treatment effect vector ($\boldsymbol{\beta}$) from the n_1 subjects on treatment. Let $\boldsymbol{R}_2 = (\boldsymbol{r}_{2n_1+1}, \ldots, \boldsymbol{r}_{2n})'$ be the $(n_2 \times V)$ matrix of unstandardized cluster-level dfbeta's for the treatment effect vector ($\boldsymbol{\beta}$) from the n_2 subjects on placebo. These are described in Hammill and Preisser (2006). The number of columns is V since we have a treatment parameter for each of the V visits. Therefore, as shown in Hammill and Preisser (2006) if $\boldsymbol{R}' = (\boldsymbol{R}_1', \boldsymbol{R}_2')$ then $\boldsymbol{R}'\boldsymbol{R}$ is the bias corrected covariance matrix estimator of $\boldsymbol{\beta}$. This bias corrected covariance matrix approximates the empirical covariance matrix and has more robust performance with moderate sample sizes. Therefore, $\boldsymbol{R}'\boldsymbol{R} = (\boldsymbol{R}_1'\boldsymbol{R}_1 + \boldsymbol{R}_2'\boldsymbol{R}_2)$.

The macro by Preisser and Qaqish (1996) requires that the data have a monotone missing pattern in addition to missing completely at random (MCAR). We use this macro's calculation of the dfbetas, which can be used to provide bias-corrected covariance matrices. However, if unstructured correlation between responses is used, the data can be similar to missing at random (MAR-like). In the event that the data does not have a monotone missing pattern, other methods of producing dfbetas can be used, such as PROC GENMOD in SAS.

Since we want to non-parametrically adjust $\boldsymbol{\beta}$ for covariates in a randomization based non-parametric way, we will need to adjust the covariance matrix of $\hat{\boldsymbol{\beta}}$ with the covariance matrix of the treatment effects with their covariates as well as the covariance matrix of the covariates themselves. Let $\boldsymbol{X}_t = (\boldsymbol{x}_{t1}, \ldots, \boldsymbol{x}_{tM})$ be the $n_t \times M$ matrix where $\boldsymbol{x}_{tm}$ is the $n_t \times 1$ column vector for the m^{th} covariate with the n_t patients in the t^{th} treatment stacked vertically. Let $\bar{\boldsymbol{x}}_t$ be the $M \times 1$ vector of means of the M covariates for the t^{th} treatment group. The covariance matrix for the means of the covariates of the t^{th} treatment group is

$$\begin{aligned} \boldsymbol{V}_{\bar{\boldsymbol{x}}_t} &= (\boldsymbol{X}_t - \mathbf{1}\bar{\boldsymbol{x}}_t')'(\boldsymbol{X}_t - \mathbf{1}\bar{\boldsymbol{x}}_t')/\{n_t(n_t-1)\} \\ &= \boldsymbol{C}_t'\boldsymbol{C}_t \end{aligned} \tag{4.2}$$

where $\boldsymbol{C}_t = (\boldsymbol{X}_t - \mathbf{1}\bar{\boldsymbol{x}}_t')/\sqrt{n_t(n_t-1)}$ and $\mathbf{1}$ is a $(n_t \times 1)$ vector of ones. Let $\boldsymbol{d} = (\hat{\boldsymbol{\beta}}', (\bar{\boldsymbol{x}}_{\mathbf{1}} - \bar{\boldsymbol{x}}_{\mathbf{2}})')'$ which is the $(V + M \times 1)$ column vector stacking the V log odds treatment effects for each visit with the M covariate differences of the averages of the 2 treatment groups. The covariance matrix for $\boldsymbol{d}$ is $\boldsymbol{V}_{\boldsymbol{d}}$ produced

as shown below on the basis of Saville and Koch (2013)

$$\boldsymbol{V}_d = \begin{bmatrix} (\boldsymbol{V}_{\hat{\boldsymbol{\beta}},1} + \boldsymbol{V}_{\hat{\boldsymbol{\beta}},2}) & (\boldsymbol{V}_{\hat{\boldsymbol{\beta}},\bar{\boldsymbol{x}}_1} - \boldsymbol{V}_{\hat{\boldsymbol{\beta}},\bar{\boldsymbol{x}}_2}) \\ (\boldsymbol{V}_{\hat{\boldsymbol{\beta}},\bar{\boldsymbol{x}}_1} - \boldsymbol{V}_{\hat{\boldsymbol{\beta}},\bar{\boldsymbol{x}}_2}) & (\boldsymbol{V}_{\bar{\boldsymbol{x}}_1} + \boldsymbol{V}_{\bar{\boldsymbol{x}}_2}) \end{bmatrix}$$
$$= \begin{bmatrix} \boldsymbol{R}_1'\boldsymbol{R}_1 + \boldsymbol{R}_2'\boldsymbol{R}_2 & \boldsymbol{R}_1'\boldsymbol{C}_1 - \boldsymbol{R}_2'\boldsymbol{C}_2 \\ \boldsymbol{C}_1'\boldsymbol{R}_1 - \boldsymbol{C}_2'\boldsymbol{R}_2 & \boldsymbol{C}_1'\boldsymbol{C}_1 + \boldsymbol{C}_2'\boldsymbol{C}_2 \end{bmatrix}.$$

We then use weighted least squares methodology to regress the parameter estimates for the V outcomes and M covariate differences onto the parameter space that has just the treatment parameter estimates for outcomes where the covariate differences are forced to be 0. Let $\boldsymbol{b}$ be the covariance adjusted estimate of treatment effect on the log odds ratios for the treatment effect at the V visits ($V \times 1$). Then $\boldsymbol{b}$ can be calculated from $\boldsymbol{d}$ with weighted least squares where $\boldsymbol{X} = [I_V, 0_{VM}]'$ where $\boldsymbol{I}_V$ is the $(V \times V)$ identity matrix and $\mathbf{0}_{VM}$ is the $(V \times M)$ matrix of 0's

$$\begin{aligned} \boldsymbol{b} &= (\boldsymbol{X}'\boldsymbol{V}_d^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}_d^{-1}\boldsymbol{d} \\ &= \hat{\boldsymbol{\beta}} - (\boldsymbol{V}_{\hat{\beta}_t,\bar{\boldsymbol{x}}_1} - \boldsymbol{V}_{\hat{\beta}_t,\bar{\boldsymbol{x}}_2})'(\boldsymbol{V}_{\bar{\boldsymbol{x}}_1} + \boldsymbol{V}_{\bar{\boldsymbol{x}}_2})^{-1}(\bar{x}_1 - \bar{x}_2). \end{aligned} \tag{4.3}$$

An estimator of the covariance matrix of $\boldsymbol{b}$ is $\boldsymbol{V}_{\boldsymbol{b}}$

$$\begin{aligned} \boldsymbol{V}_{\boldsymbol{b}} &= (\boldsymbol{X}'\boldsymbol{V}_d^{-1}\boldsymbol{X})^{-1} \\ &= (\boldsymbol{V}_{\hat{\beta},1} + \boldsymbol{V}_{\hat{\beta},2}) - (\boldsymbol{V}_{\hat{\beta},\bar{x}_1} - \boldsymbol{V}_{\hat{\beta},\bar{x}_2})'(\boldsymbol{V}_{\bar{x}_1} + \boldsymbol{V}_{\bar{x}_2})^{-1}(\boldsymbol{V}_{\hat{\beta},\bar{x}_1} - \boldsymbol{V}_{\hat{\beta},\bar{x}_2}) \end{aligned} \tag{4.4}$$

As the estimator $\boldsymbol{b}$ has an approximately multivariate normal distribution, we can test the homogeneity of the visit effects with a contrast statement. Let $\boldsymbol{C} = [\boldsymbol{I}_{(V-1)}, -\mathbf{1}_{(V-1)}]$. The test statistic for homogeneity of visit effects is

$$Q_{\text{visit}} = \boldsymbol{b}'C(CV_{\boldsymbol{b}}C')^{-1}\boldsymbol{C}\boldsymbol{b} \tag{4.5}$$

which has a χ^2 distribution with $V - 1$ df.

Table 4.1: Distribution of responses for Stokes et al. (2012)

Visit	Drug	Proportion of responders
1	Placebo	0.49
	Treatment	0.69
2	Placebo	0.39
	Treatment	0.7
3	Placebo	0.46
	Treatment	0.72
4	Placebo	0.44
	Treatment	0.61

4.1.2 Results for respiratory data set

To illustrate our methods, we use a data set from Stokes et al. (2012). We have 111 patients enrolled in a clinical trial comparing two treatments for a respiratory disease. Patients are stratified by two centers, and their outcome variable is a five-point scale for how they feel. Patients who felt terrible had a 0 recorded, and patients who felt excellent had a 4 recorded. This scale was recorded at baseline (as a covariate) and then at each of 4 visits (as outcomes). Other covariate data collected are age, gender, and center. There is no missing data for this study. The distributions of responses by visit and treatment are shown in Table 4.1.

Using the method previously described in the data set for Stokes et al. (2012), we can get a sense for how much of an impact the covariance adjustment has by comparing $\boldsymbol{d}$ to $\boldsymbol{b}$ along with their estimated covariance matrices (Equations 4.7 and 4.8 versus Equation 4.9). The test statistic of homogeneity of treatment effect across visits was 3.14 (with 3 df), which has a p-value of 0.3699. Since this result is compatible with treatment effects being homogeneous across visits, we can consider a model model with one treatment parameter for all visits. The covariate-adjusted effect of treatment on the log odds ratio of good or excellent outcomes is 0.95 (0.46, 1.43). The test statistic for the treatment effect is 14.32, with a p-value of 0.00015. This indicates that the treatment has a significant effect on the log-odds ratio for treatment vs placebo for the study population. A comparison of the unadjusted and adjusted odds ratios (OR) for treatment effect by visit is shown in Table 4.2, and a comparison for the overall treatment effect is shown in Table 4.3. Our unadjusted treatment effect estimate $\hat{\boldsymbol{\beta}}$ comes from the initial model (Equation 4.6). Both tables demonstrate that there is minimal adjustment to the estimate itself, but randomization-based non-parametric adjustment of covariates

produces reduction in the length of the confidence interval.

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha_1(j=1) + \alpha_2(j=2) + \ldots + \alpha_V(j=V) + \beta_{\text{all}} * (i=1) \tag{4.6}$$

Table 4.2: Comparison of OR for unadjusted and adjusted treatment effects by visit

	Unadjusted			Adjusted		
Visit	OR	Lower 95% CL	Upper 95% CL	OR	Lower 95% CL	Upper 95% CL
1	2.25	1.04	4.89	2.27	1.2	4.29
2	3.78	1.71	8.33	3.96	1.87	8.39
3	3.1	1.4	6.84	3.24	1.59	6.59
4	2.01	0.94	4.29	2.14	1.09	4.18

Table 4.3: Comparison of overall treatment OR

Method	OR	95% LCL	95% UCL
Unadjusted	2.69	1.46	4.96
Adjusted	2.58	1.71	3.89

$$\boldsymbol{d} = (0.8128, 1.3293, 1.1314, 0.6988, -0.7602, -0.1871, -0.0156, 0.0088)' \tag{4.7}$$

$$\boldsymbol{V}_d = \begin{bmatrix} 0.1618 & 0.0780 & 0.0703 & 0.0774 & 0.0245 & -0.0006 & 0.0470 & 0.0102 \\ 0.0780 & 0.1689 & 0.0936 & 0.0834 & -0.0533 & 0.0006 & 0.0281 & 0.0067 \\ 0.0703 & 0.0936 & 0.1691 & 0.0887 & -0.1552 & -0.0006 & 0.0360 & 0.0060 \\ 0.0774 & 0.0834 & 0.0887 & 0.1547 & -0.0930 & 0.0014 & 0.0333 & 0.0115 \\ 0.0245 & -0.0533 & -0.1552 & -0.0930 & 6.7936 & 0.0592 & 0.0216 & 0.0548 \\ -0.0006 & 0.0006 & -0.0006 & 0.0014 & 0.0592 & 0.0056 & -0.0009 & 0.0015 \\ 0.0470 & 0.0281 & 0.0360 & 0.0333 & 0.0216 & -0.0009 & 0.0402 & 0.0058 \\ 0.0102 & 0.0067 & 0.0060 & 0.0115 & 0.0548 & 0.0015 & 0.0058 & 0.0092 \end{bmatrix} \tag{4.8}$$

$$\boldsymbol{b} = \begin{bmatrix} 0.8202 \\ 1.3758 \\ 1.1756 \\ 0.7588 \end{bmatrix}, \boldsymbol{V_b} = \begin{bmatrix} 0.1054 & 0.0438 & 0.0273 & 0.0352 \\ 0.0438 & 0.1469 & 0.0653 & 0.0554 \\ 0.0273 & 0.0653 & 0.1313 & 0.0533 \\ 0.0352 & 0.0554 & 0.0533 & 0.1173 \end{bmatrix} \tag{4.9}$$

4.1.3 COMPARISON WITH TANGEN AND KOCH (1999) METHOD

Tangen and Koch (1999) discussed a similar method using the same basic concepts as our method, but with a different estimation method for the covariance matrix and with the requirement of no missing data. Tangen and Koch (1999) create a vector of the log odds ratios for treatment effects and differences in the means of the covariates and then uses weighted least squares to force the differences in the covariates to 0. However, instead of using dfbetas to calculate covariance terms related to the treatment effect, their method uses the difference in the observed value and the mean for each treatment group and Taylor series methods. The Taylor series method is an approximation so our method of using the dfbetas could be more accurate.

As we can see from Figure 4.1, the two methods produce very similar results. Table 4.4 suggests that our method produces slightly larger variance calculations and adjusted parameter estimates. The advantage of using the newer method is that the covariance matrix calculation takes into account possible missingness of the response through dfbeta calculations as well as the correlation between responses. However a limitation is that our use of GEE to calculate the dfbetas involves the assumption that missing responses are missing completely at random, or very similar to missing at random. Although Tangen and Koch (1999) and our method can handle stratification, we will manage center as a covariate.

4.2 PARTIAL PROPORTIONAL ODDS LONGITUDINAL MODEL

Now we will extend our method to ordinal outcomes. Let $Y_{htj} = p$ represent an outcome of p for the h^{th} subject for the t^{th} treatment group at the j^{th} visit, where $p = 1, 2, \ldots, P$ and the larger the value of p, the more favorable the outcome. We will assume that the design of the trial forces subjects to have the same visit schedule.

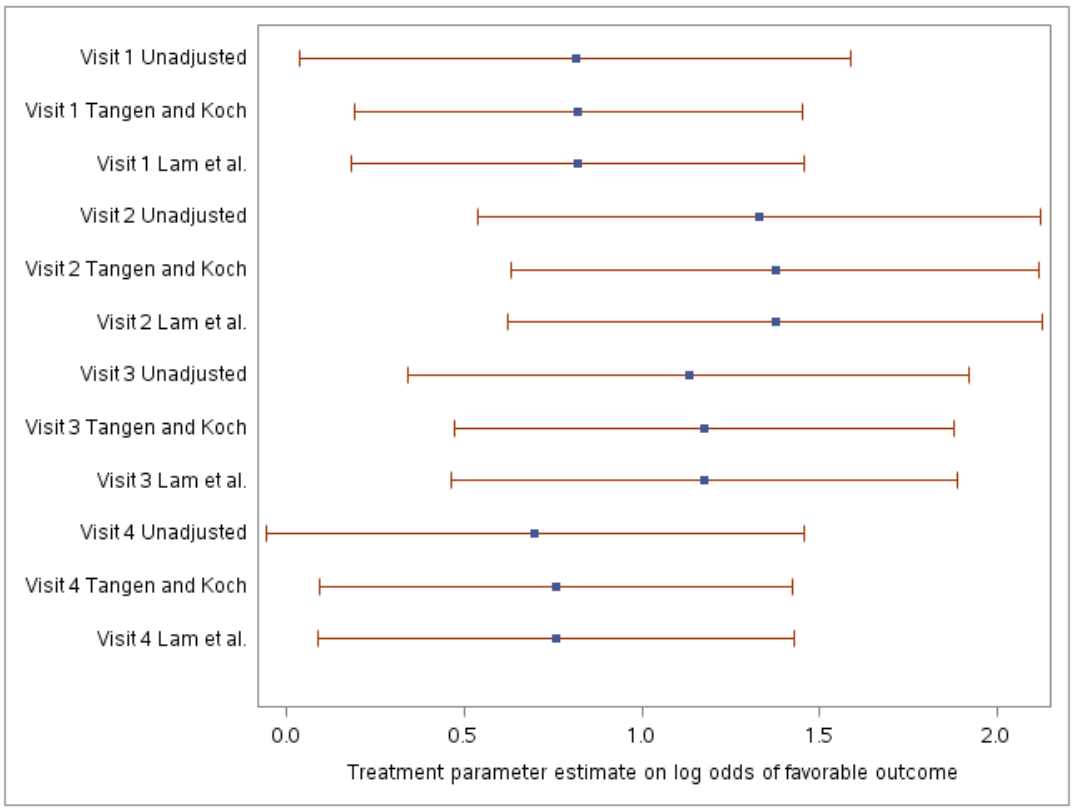

Figure 4.1: Forest plot comparing Tangen and Koch (1999) and current method

First we will model reference intercepts for visit by cumulative logit with treatment nested within visit by cumulative logit. Let $\gamma_{ptj} = Pr(Y_{tj} \geq p)$ which is the probability that a subject in the t^{th} treatment group will have an outcome at least as good as the p^{th} level$(p > 1)$ at the j^{th} visit. Let α_{pj} represent the j^{th} visit intercept for the log odds model for an outcome at least as good as p. Let β_{pj} be the treatment effect for the j^{th} visit for the log odds of an outcome at least as good as p. The longitudinal partial proportional odds model

Table 4.4: Table comparing treatment effects nested within visit for 2 methods

Visit	Method	Estimate	SE	LCL	UCL	p-value
1	Tangen and Koch	0.8201	0.3218	0.1894	1.4508	0.0108
	Lam and Koch	0.8202	0.3247	0.1837	1.4566	0.0115
2	Tangen and Koch	1.3753	0.3798	0.631	2.1197	0.0003
	Lam and Koch	1.3758	0.3833	0.6246	2.127	0.0003
3	Tangen and Koch	1.1751	0.359	0.4714	1.8788	0.0011
	Lam and Koch	1.1756	0.3623	0.4654	1.8857	0.0012
4	Tangen and Koch	0.7582	0.3395	0.0929	1.4235	0.0255
	Lam and Koch	0.7588	0.3426	0.0874	1.4302	0.0268

is

$$
\begin{aligned}
log\left(\frac{\gamma_{ptj}}{1-\gamma_{ptj}}\right) &= \alpha_{21}(p=2)(j=1)+\ldots+\alpha_{2V}(p=2)(j=V)+\ldots \\
&+\alpha_{P1}(p=P)(j=1)+\ldots+\alpha_{PV}(p=P)(j=V) \\
&+\beta_{21}(p=2)(j=1)(t=1)+\ldots+\beta_{2V}(p=2)(j=V)(t=1)+\ldots \\
&+\beta_{P1}(p=P)(j=1)(t=1)+\ldots+\beta_{PV}(p=P)(j=V)(t=1)
\end{aligned} \tag{4.10}
$$

Each visit and each cumulative logit is allowed to have its own specifications for treatment and control. $\boldsymbol{\alpha} = (\alpha_{21},\ldots,\alpha_{2V},\ldots,\alpha_{P1},\ldots,\alpha_{PV})'$ contains the estimators of the V visit intercepts for a cutoff of 2 followed by the estimators of the V visit intercepts for a cutoff of 3, etc. Likewise, $\boldsymbol{\beta} = (\beta_{21},\ldots,\beta_{2V},\ldots,\beta_{P1},\ldots,\beta_{PV})'$ is the $(P-1)V \times 1$ vector that contains the V treatment by visit effects for a cutoff of 2 followed by the V treatment by visit effects for a cutoff of 3, etc. For example, the model for the treatment at the first visit for $Pr(Y_{tj} \geq 2)$ is

$$
log\left(\frac{\gamma_{2t1}}{1-\gamma_{2t1}}\right) = \alpha_{21} + \beta_{21} * (t=1)
$$

where α_{21} is the parameter for the log odds of $Pr(Y_{tj} \geq 2)$ for the control group at visit 1 and $\alpha_{21}+\beta_{21}$ is the parameter for the log odds of $Pr(Y_{tj} \geq 2)$ for the treatment group at visit 1.

Let $\boldsymbol{R}_1 = (\boldsymbol{r}_{11},\ldots,\boldsymbol{r}_{1n_1})'$ be the $n_1 \times K$ matrix of un-standardized cluster-level dfbetas for patients in group 1, where $K=(P-1)V$, as described in Hammill and Preisser (2006) for the treatment effect vector

($\boldsymbol{\beta}$). Each $\boldsymbol{r}_{1h}$ where $h \in (1, \ldots, n_1)$ is a $K \times 1$ column vector that corresponds to the V treatment effects for all visits for the first cumulative logit followed by the V treatment effects for all visits for the second cumulative logit, etc. Let $\boldsymbol{R}_2 = (\boldsymbol{r}_{2n_1+1}, \ldots, \boldsymbol{r}_{2n})'$ be the $n_2 \times K$ matrix of un-standardized cluster-level dfbetas for patients in group 2. Therefore $\boldsymbol{R}' = (\boldsymbol{R}_1', \boldsymbol{R}_2')$, and $\boldsymbol{R'R}$ is the bias corrected covariance matrix estimator of $\boldsymbol{\beta}$. This bias corrected covariance matrix approximates the empirical covariance matrix and has better performance with moderate sample size. Therefore, $\boldsymbol{R'R} = (\boldsymbol{R}_1'\boldsymbol{R}_1 + \boldsymbol{R}_2'\boldsymbol{R}_2)$.

The non-parametric covariate adjustment is performed the same as in the logistic regression situation. The only difference is that the dimensions that referred to treatment effects are now larger by a factor of $P - 1$. Therefore, $\boldsymbol{d} = (\hat{\boldsymbol{\beta}}', (\bar{\boldsymbol{x}}_{\boldsymbol{1}} - \bar{\boldsymbol{x}}_{\boldsymbol{2}})')'$, which is a $[(P-1)V + M] \times 1$ column vector.

We can test the proportional odds assumption across cumulative logits using the following contrast specification

$$\boldsymbol{Q}_{\text{proportional odds}} = \boldsymbol{b'C'}(\boldsymbol{CV_bC'})^{-1}\boldsymbol{Cb} \tag{4.11}$$

where $\boldsymbol{C} = [\boldsymbol{I}_{P-2} \otimes \boldsymbol{I}_V, \boldsymbol{1}_{P-2} \otimes \boldsymbol{-I}_V]$. $\boldsymbol{I}_{P-2} \otimes \boldsymbol{I}_V$ is a block diagonal matrix where each of the $P - 2$ blocks is the identity matrix of dimension V. $\boldsymbol{1}_{P-2} \otimes \boldsymbol{-I}_V$ is a column matrix of $P - 2$ blocks where each block is $-\boldsymbol{I}_V$. If the proportional odds test is non-significant, the treatment effect can be regressed onto the parameter space that regresses the treatment by visit effects for all cutoffs into the same space. In otherwords, carry out a weighted regression of $\boldsymbol{d}$ on $\boldsymbol{X} = [(\boldsymbol{1}_{P-1} \otimes \boldsymbol{I}_V)' \; \boldsymbol{0}_{VM}]'$.

4.2.1 Application to respiratory data set

Before applying the partial proportional odds regression model, we must first check the number of observations in each type of outcome for the visit*treatment combinations. The poorest two outcomes (0 and 1) are sparsely populated, so we will not place a cutoff for the cumulative logits between them. Furthermore the two best outcomes (3 and 4) are clinically similar as they represent good and excellent outcomes so we will also collapse them. Our cutoffs are then 0-1 vs 2-4 (cutoff=2) and 0-2 vs 3-4 (cutoff=3) Using the previously specified methods, we can obtain the treatment adjusted estimates (Equation 4.14) which can be compared to the unadjusted estimates and differences in means (Equation 4.13). We can test whether or not the treatment

by visit effect satisfies the proportional odds assumption across the two cutoffs by testing

$$\begin{aligned}\eta_{21} &= \eta_{31}\\ \vdots &= \vdots\\ \eta_{24} &= \eta_{34}\end{aligned}$$

where η_{pj} is the covariance adjusted treatment effect parameter for cutoff p at visit j. In other words, b_{pj} is an estimator of η_{pj}.

The test statistic of the proportional odds assumption is 1.15 with a p-value of 0.8863, which suggests that the treatment by visit effects do not significantly differ across the two cutoffs. The adjusted treatment estimates by visit over both cutoffs is given in Equation 4.16. The test statistic for whether or not the treatment effects are different across the visits is 5.41, which has a p-value of 0.1440. This suggests we can use a model with one treatment parameter for all visits, as shown in Equation 4.12. The test statistic is 14.61, with a p-value of 0.0001. The data suggest that treatment is a significant predictor of the odds of more favorable versus less favorable outcomes. The covariate adjusted treatment effect on the log odds of more favorable outcomes versus less favorable outcomes is 0.94 (0.46, 1.43). A comparison of the unadjusted and adjusted OR for treatment effects by visit for each cutoff is given in Table 4.5.

$$\begin{aligned}log\left(\frac{\gamma_{ptj}}{1-\gamma_{ptj}}\right) &= \alpha_{21}(p=2)(j=1)+\ldots+\alpha_{2V}(p=2)(j=V)+\ldots\\ &+\alpha_{P1}(p=P)(j=1)+\ldots+\alpha_{PV}(p=P)(j=V)\\ &+\beta_1(j=1)(t=1)+\ldots+\beta_V(j=V)(t=1)\end{aligned} \tag{4.12}$$

Table 4.5: Comparison of unadjusted and adjusted method for partial proportional odds model

		Unadjusted			Adjusted		
Cutoff	Visit	OR	95% LCL	95% UCL	OR	95% LCL	95% UCL
2-4 vs 0-1	1	2.66	0.78	9.06	3.09	0.96	9.89
	2	4.88	1.51	15.73	5.58	1.79	17.36
	3	4.17	1.41	12.28	4.44	1.53	12.90
	4	1.79	0.71	4.51	1.89	0.76	4.67
3-4 vs 0-2	1	2.25	1.04	4.89	2.27	1.20	4.29
	2	3.78	1.71	8.33	3.96	1.87	8.39
	3	3.1	1.4	6.84	3.24	1.59	6.59
	4	2.01	0.94	4.29	2.14	1.09	4.18

$$\boldsymbol{d} = \begin{bmatrix} 0.9782 \\ 1.5847 \\ 1.4267 \\ 0.5798 \\ 0.8128 \\ 1.3293 \\ 1.1314 \\ 0.6988 \\ -0.7602 \\ -0.1871 \\ -0.0156 \\ 0.0088 \end{bmatrix} \tag{4.13}$$

$$\boldsymbol{b} = \begin{bmatrix} 1.1276 \\ 1.7187 \\ 1.4910 \\ 0.6346 \\ 0.8202 \\ 1.3758 \\ 1.1756 \\ 0.7588 \end{bmatrix} \tag{4.14}$$

$$\boldsymbol{V_b} = \begin{bmatrix} 0.3527 & 0.1470 & 0.1028 & 0.0986 & 0.0580 & 0.0506 & 0.0359 & 0.0332 \\ 0.1470 & 0.3358 & 0.0685 & 0.0803 & 0.0317 & 0.0861 & 0.0563 & 0.0629 \\ 0.1028 & 0.0685 & 0.2959 & 0.1149 & 0.0294 & 0.0598 & 0.0974 & 0.0688 \\ 0.0986 & 0.0803 & 0.1149 & 0.2139 & 0.0230 & 0.0364 & 0.0655 & 0.0818 \\ 0.0580 & 0.0317 & 0.0294 & 0.0230 & 0.1054 & 0.0438 & 0.0273 & 0.0352 \\ 0.0506 & 0.0861 & 0.0598 & 0.0364 & 0.0438 & 0.1469 & 0.0653 & 0.0554 \\ 0.0359 & 0.0563 & 0.0974 & 0.0655 & 0.0273 & 0.0653 & 0.1313 & 0.0533 \\ 0.0332 & 0.0629 & 0.0688 & 0.0818 & 0.0352 & 0.0554 & 0.0533 & 0.1173 \end{bmatrix} \tag{4.15}$$

$$\boldsymbol{b} = \begin{bmatrix} 0.8431 \\ 1.4080 \\ 1.2416 \\ 0.6780 \end{bmatrix}, \boldsymbol{V_b} = \begin{bmatrix} 0.0982 & 0.0455 & 0.0306 & 0.0334 \\ 0.0455 & 0.1301 & 0.0629 & 0.0536 \\ 0.0306 & 0.0629 & 0.1235 & 0.0598 \\ 0.0334 & 0.0536 & 0.0598 & 0.1068 \end{bmatrix} \tag{4.16}$$

4.3 STRATIFICATION

Clinical trials often have some factor for stratification of patients, such as center or a demographic or prognostic characteristic. While both covariates and stratified variables are managed in the same way in parametric models, they are not the same in randomization-based nonparametric adjustments (Tangen and Koch, 1999). In the nonparametric model, when we account for a stratified factor, we consider the

randomization being invoked separately within each stratum. Expanding the stratification methods of Tangen and Koch (1999) and Zink and Koch (2012), we have a method for stratification.

Suppose we stratify by the variable $s = 1, \ldots, S$. For simplification, we will assume our outcome variable is binary. Much like in the partial proportional odds model, we index the stratified variable as we did the cumulative logit. For the s^{th} stratum, the model is:

$$\begin{aligned} log\left(\frac{\pi_{stj}}{1-\pi_{stj}}\right) &= \alpha_{s1}(j=1) + \alpha_{s2}(j=2) + \ldots + \alpha_{sV}(j=V) + \ldots \\ &+ \beta_{s1}(j=1)(t=1) + \beta_{s2}(j=2)(t=1) \\ &+ \ldots + \beta_{sV}(j=V)(t=1) \end{aligned}$$

Each visit and each stratum level logit is allowed to have its own effect for treatment and control. For example, the model at the second visit for the third value of the stratification variable is

$$log\left(\frac{\pi_{3i2}}{1-\pi_{3i2}}\right) = \alpha_{32} + \beta_{32} * (t=1)$$

where α_{32} is the parameter for the log odds of $Pr(Y_{3i2} = 1)$ for the control group at visit 2 in the third stratum and $\alpha_{32} + \beta_{32}$ is the parameter for the log odds of $Pr(Y_{3i2} = 1)$ for the treatment group at visit 2. Let $\boldsymbol{\beta} = (\beta_{11}, \ldots, \beta_{1V}, \ldots, \beta_{S1}, \ldots, \beta_{SV})'$ which is a $SV \times 1$ column vector. The vector of treatment effects for the s^{th} stratum is $\boldsymbol{\beta}_s = (\beta_{s1}, \ldots, \beta_{sV})$; which is a $(V \times 1)$ vector.

4.3.1 Moderate to large strata

For large sample size in each stratum ($n_{st} \geq 30$, where n_{st} is the number of patients in the s^{th} stratum for the t^{th} treatment), we calculate the covariance-adjusted treatment effects for each stratum, and then take a weighted average of those treatment effects.

The covariate adjustment is different from the partial proportional odds case. Instead of $\boldsymbol{X}_t$ for each treatment, we have $\boldsymbol{X}_{st}$ which is the $(n_{st} \times M)$ matrix of the M covariates for the n_{st} patients in the s^{th} stratum for the t^{th} treatment. $\bar{\boldsymbol{x}}_{st}$ is the $(M \times 1)$ vector of averages for the M covariates in the s^{th} stratum and t^{th} treatment.

Therefore, we let

$$\boldsymbol{d}_s = \begin{bmatrix} \hat{\boldsymbol{\beta}}_{\boldsymbol{st}} \\ \bar{\boldsymbol{x}}_{s1} - \bar{\boldsymbol{x}}_{s2} \end{bmatrix} \hat{=} \begin{bmatrix} \boldsymbol{I}_V & \boldsymbol{0}_{VM} \end{bmatrix}' \begin{bmatrix} \boldsymbol{b_s} \end{bmatrix},$$

where $\hat{=}$ means estimates. $\boldsymbol{b}_s = [b_{s1t}, \ldots, b_{sVt}]'$ represents the treatment effects for the s^{th} stratum. The overall treatment effects $\bar{\boldsymbol{b}}$ are the weighted averages of the S adjusted treatment effects for the respective strata:

$$\bar{\boldsymbol{b}} = \sum_{s=1}^{S} \frac{w_s \boldsymbol{b_s}}{\sum_{s=1}^{S} w_s}$$

where $w_s = \frac{n_{s1} n_{s2}}{n_{s1} + n_{s2}}$ is the Mantel-Haenszel weight of the s^{th} stratum. The estimated covariance matrix, $\boldsymbol{V_{ds}}$ is produced from $\boldsymbol{H}_{st} = [\boldsymbol{R}_{st}, (-1)^{t-1}\boldsymbol{C}_{st}]$ which is the $n_{st} \times (V + M)$ matrix where the first V columns correspond to the n_{st} dfbetas for the V treatment effects for the s^{th} stratum and t^{th} treatment, and the last M columns refer to the n_{st} corresponding differences between the observed covariate levels and $\bar{\boldsymbol{x}}_{\boldsymbol{st}}$, i.e. with $\boldsymbol{H}'_s = [\boldsymbol{H}'_{s1}, \boldsymbol{H}'_{s2}]$

$$\boldsymbol{V}_{ds} = \boldsymbol{H}'_s \boldsymbol{H}_s = \begin{bmatrix} \boldsymbol{R}'_{s1} & \boldsymbol{R}'_{s2} \\ \boldsymbol{C}'_{s1} & -\boldsymbol{C}'_{s2} \end{bmatrix} \begin{bmatrix} \boldsymbol{R}'_{s1} & \boldsymbol{C}'_{s1} \\ \boldsymbol{R}'_{s2} & -\boldsymbol{C}'_{s2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}'_{s1}\boldsymbol{R}_{s1} + \boldsymbol{R}'_{s2}\boldsymbol{R}_{s2} & \boldsymbol{R}'_{s1}\boldsymbol{C}_{s1} - \boldsymbol{R}'_{s2}\boldsymbol{C}_{s2} \\ \boldsymbol{C}'_{s1}\boldsymbol{R}_{s1} - \boldsymbol{C}'_{s2}\boldsymbol{R}_{s2} & \boldsymbol{C}'_{s1}\boldsymbol{C}_{s1} + \boldsymbol{C}'_{s2}\boldsymbol{C}_{s2} \end{bmatrix}.$$

The estimated covariance matrix of $\bar{\boldsymbol{b}}$ is $\boldsymbol{V}_{\bar{\boldsymbol{b}}}$

$$\boldsymbol{V}_{\bar{\boldsymbol{b}}} = \sum_{s=1}^{S} \left(\frac{w_s}{\sum_{s=1}^{S} w_s} \right)^2 \boldsymbol{V_{bs}}$$

where $\boldsymbol{V_{bs}}$ calculated below, is the covariance matrix of $\boldsymbol{b}_s$, calculated from the weighted least squares regression for the s^{th} stratum

$$\boldsymbol{V_{bs}} = (\boldsymbol{X}'\boldsymbol{V}_{ds}^{-1}\boldsymbol{X})^{-1}.$$

4.3.2 SMALL TO MODERATE STRATA

For small to moderate sample size in each stratum, we construct weighted averages across the strata and then apply the weighted least squares regression on the weighted outcomes.

Therefore, we let

$$\sum_{s=1}^{S} \frac{w_s}{\sum_{s=1}^{S} w_s} \begin{bmatrix} \hat{\boldsymbol{\beta}}_{\boldsymbol{st}} \\ \bar{\boldsymbol{x}}_{s1} - \bar{\boldsymbol{x}}_{s2} \end{bmatrix} \hat{=} \begin{bmatrix} \boldsymbol{I}_V & \boldsymbol{0}_{VM} \end{bmatrix}' \begin{bmatrix} \boldsymbol{b_*} \end{bmatrix}$$

where $\boldsymbol{b}_* = [b_{1t}, \ldots, b_{Vt}]'$ represents the stratum-weighted, covariate-adjusted treatment effect for each visit.

Let $\boldsymbol{d_*} = \sum_{s=1}^{S} \frac{w_s}{\sum_{s=1}^{S} w_s} \begin{bmatrix} \hat{\boldsymbol{\beta}}_{\boldsymbol{st}} \\ \bar{\boldsymbol{x}}_{s1} - \bar{\boldsymbol{x}}_{s2} \end{bmatrix}$

The estimator of $\boldsymbol{b}_*$ is

$$\boldsymbol{b}_* = (\boldsymbol{X}'\boldsymbol{V}_{d*}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{d*}}^{-1}\boldsymbol{d_*}$$

where $\boldsymbol{V}_{d*}$ is defined below

$$\boldsymbol{V}_{d*} = \sum_{s=1}^{S} \left(\frac{w_s}{\sum_{s=1}^{S} w_s} \right)^2 \boldsymbol{V}_{ds}.$$

The estimator of the covariance matrix of $\boldsymbol{b}_*$ is $\boldsymbol{V}_{b_*}$ as given below,

$$\boldsymbol{V}_{b_*} = (\boldsymbol{X}'\boldsymbol{V}_{d*}^{-1}\boldsymbol{X})^{-1}.$$

Furthermore, our method has some flexibility if sample sizes are only moderately large. For example, we can have stratum effects, visit effects, and visit by treatment effects, which allows us to reduce our parameters by $2SV - (S - 1 + 2V)$ while still allowing treatment effects to vary by visit. The model would look like Equation 4.17

$$log\left(\frac{\pi_{stj}}{1-\pi_{stj}}\right) = \alpha_1(j=1) + \alpha_2(j=2) + \ldots + \alpha_V(j=V) + \ldots \quad (4.17)$$
$$+ \phi_1(s=1) + \phi_2(s=2) + \ldots + \phi_S(s=S-1)$$
$$+ \beta_1(j=1)(t=1) + \beta_2(j=2)(t=1)$$
$$+ \ldots + \beta_V(j=V)(t=1)$$

where $\boldsymbol{\alpha}$ would be the vector of visit intercepts estimated across all strata, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{S-1})$ would be the vector containing the stratum effects for being in the s^{th} stratum, $s = 1, \ldots, S-1$, relative to the S^{th} stratum, and $\boldsymbol{\beta}$ contains the treatment by visit effects estimated across all strata. This would be a good model to use when the treatment effect by visit is the same within all strata, but some strata have consistently higher or lower estimates of the log odds for placebo over all visits. If we had the conditions listed previously and also have homogeneous treatment effects for all visits, we can consider estimating the average of the treatment effects over all visits as opposed to treatment effects nested within visit, such as in Equation 4.18

$$log\left(\frac{\pi_{stj}}{1-\pi_{stj}}\right) = \alpha_1(j=1) + \alpha_2(j=2) + \ldots + \alpha_V(j=V) \quad (4.18)$$
$$+ \phi_1(s=1) + \phi_2(s=2) + \ldots + \phi_S(s=S-1)$$
$$+ \beta(t=1)$$

.

4.3.3 Application to the respiratory data set

With 27-29 subjects in each treatment by center combination, it is unclear whether to use the method for moderate to large or small to moderate strata. As there are no missing data in our data set, we will use the method for large sample size. The Mantel Haenszel weight for each center is 0.50. The adjusted treatment effects for center 1 are (0.66, 1.22, 1.11, 0.73), and the adjusted treatment effects for center 2 are (1.45, 1.76, 1.45, 1.07), as shown in Equation 4.21. Comparing these estimates to Equation 4.19, we see that the estimates have been adjusted more in center 1 than in center 2, possibly due to the better precision of the estimated

covariance matrix of the differences in the means (Equation 4.20). The treatment effect for each visit is higher for center 2 than it is for center 1, but the estimated covariance matrix for center 2 is also less precisely estimated than that of center 1 (Equation 4.22). The consistently higher estimates may support our decision to stratify by center as these populations may not be the same. The average age, percentage of females, and baseline level for center 2 is higher for both the treatment and placebo group than center 1's. In center 2, the placebo has a slightly lower baseline, suggesting that the placebo group in center 2 is somewhat worse than the treatment group at baseline.

The value of the test statistic that tests the homogeneity of the treatment effect across visits after adjusting for center stratification is 2.00, with a p-value of 0.57. Therefore, if we were to stratify by center our covariate-adjusted treatment effect on the log odds ratio is 1.14 (0.60, 1.69). The test statistic testing the significance of the treatment effect is 17.09, which has a p-value < 0.0001. This result is very similar to the treatment effect which managed center as a covariate (1.14 versus 0.95). The stratified value is larger than the non-stratified counterpart. The stratified treatment effect is for the log odds ratio for patients at the same center versus the non-stratified treatment effect is for the population average log odds ratio. The comparison of the unadjusted and adjusted method when managing center as a stratification variable is shown in Table 4.6.

Table 4.6: Comparison of unadjusted and adjusted OR with stratification by center

	Unadjusted			Adjusted		
Visit	OR	95% LCL	95% UCL	OR	95% LCL	95% UCL
1	2.55	1.11	5.89	2.86	1.38	5.92
2	4.01	1.77	9.09	4.43	2.00	9.81
3	3.25	1.44	7.32	3.59	1.69	7.63
4	2.16	0.97	4.81	2.45	1.15	5.23

$$\boldsymbol{d}_1 = \begin{bmatrix} 0.4224 \\ 1.0165 \\ 0.8789 \\ 0.5754 \\ -0.7637 \\ -0.0983 \\ -0.2095 \end{bmatrix}, \boldsymbol{d}_2 = \begin{bmatrix} 1.4615 \\ 1.7693 \\ 1.4816 \\ 0.9651 \\ -0.8624 \\ -0.2804 \\ 0.1706 \end{bmatrix} \tag{4.19}$$

$$\boldsymbol{V}_{\boldsymbol{d}_1} = \begin{bmatrix} 0.3125 & 0.1562 & 0.1684 & 0.1477 & -0.3132 & -0.0064 & 0.1131 \\ 0.1562 & 0.3292 & 0.1668 & 0.1465 & -0.3918 & 0.0004 & 0.0740 \\ 0.1684 & 0.1668 & 0.3238 & 0.1679 & -0.3559 & -0.0009 & 0.0885 \\ 0.1477 & 0.1465 & 0.1679 & 0.3346 & -0.3470 & -0.0026 & 0.0697 \\ -0.3132 & -0.3918 & -0.3559 & -0.3470 & 9.7230 & 0.0353 & -0.0534 \\ -0.0064 & 0.0004 & -0.0009 & -0.0026 & 0.0353 & 0.0077 & -0.0011 \\ 0.1131 & 0.0740 & 0.0885 & 0.0697 & -0.0534 & -0.0011 & 0.0767 \end{bmatrix} \tag{4.20}$$

$$\boldsymbol{V}_{\boldsymbol{d}_2} = \begin{bmatrix} 0.4733 & 0.1697 & 0.1202 & 0.2085 & 0.1408 & 0.0021 & 0.0599 \\ 0.1697 & 0.4215 & 0.2669 & 0.2019 & 0.0776 & 0.0008 & 0.0221 \\ 0.1202 & 0.2669 & 0.4183 & 0.2192 & -0.3947 & -0.0037 & 0.0413 \\ 0.2085 & 0.2019 & 0.2192 & 0.3879 & -0.3931 & 0.0028 & 0.0489 \\ 0.1408 & 0.0776 & -0.3947 & -0.3931 & 16.7011 & 0.1724 & 0.0018 \\ 0.0021 & 0.0008 & -0.0037 & 0.0028 & 0.1724 & 0.0139 & -0.0058 \\ 0.0599 & 0.0221 & 0.0413 & 0.0489 & 0.0018 & -0.0058 & 0.0708 \end{bmatrix}$$

$$\boldsymbol{b}_1 = \begin{bmatrix} 0.6595 \\ 1.2221 \\ 1.1119 \\ 0.7318 \end{bmatrix}, \boldsymbol{b}_2 = \begin{bmatrix} 1.4479 \\ 1.7612 \\ 1.4462 \\ 1.0663 \end{bmatrix} \tag{4.21}$$

$$\boldsymbol{V}_{\boldsymbol{b_1}} = \begin{bmatrix} 0.1380 & 0.0403 & 0.0315 & 0.0374 \\ 0.0403 & 0.2448 & 0.0705 & 0.0689 \\ 0.0315 & 0.0705 & 0.2125 & 0.0784 \\ 0.0374 & 0.0689 & 0.0784 & 0.2619 \end{bmatrix}, \boldsymbol{V}_{\boldsymbol{b_2}} = \begin{bmatrix} 0.4189 & 0.1495 & 0.0868 & 0.1653 \\ 0.1495 & 0.4140 & 0.2553 & 0.1867 \\ 0.0868 & 0.2553 & 0.3837 & 0.1778 \\ 0.1653 & 0.1867 & 0.1778 & 0.3348 \end{bmatrix} \tag{4.22}$$

$$\bar{\boldsymbol{b}} = \begin{bmatrix} 1.0503 \\ 1.4893 \\ 1.2776 \\ 0.8976 \end{bmatrix}, \boldsymbol{V}_{\bar{\boldsymbol{b}}} = \begin{bmatrix} 0.1380 & 0.0470 & 0.0293 & 0.0501 \\ 0.0470 & 0.1640 & 0.0807 & 0.0634 \\ 0.0293 & 0.0807 & 0.1483 & 0.0636 \\ 0.0501 & 0.0634 & 0.0636 & 0.1489 \end{bmatrix} \tag{4.23}$$

4.4 CUSHING'S DISEASE STUDY

Cushing's disease is a rare, life-threatening disease caused by an adrenocorticotropic hormone-secreting pituitary adenoma. Patients with Cushing's disease suffer from symptoms similar to chronic hypercortisolism and have impacted health and quality of life. The primary treatment for Cushing's disease is surgery, but for those patients who do not respond well to surgery or who are not viable candidates for surgery, there are few therapeutic options.

Pasireotide was explored as possible treatment for those suffering from Cushing's disease who do not respond well to surgery or for whom surgery is not an option, due to its high affinity for a particular somatostatin receptor (SSTR5), which is expressed at high levels in adrenocorticotropic-secreting tumors in Cushing's disease. Two doses were explored for efficacy (Colao et al., 2012): 600 μg twice daily (bid) and 900 μg bid. Due to ethical concerns, there was no placebo arm in this clinical trial. The data we were given are not the exact data from the trial, but a random bootstrap sample.

162 patients were enrolled in the clinical trial in a roughly one-to-one ratio, 80 at the pasireotide 600 μg bid dose and 82 at the 900 μg bid dose for 12 months. The urinary free cortisol (UFC) levels were measured every 30 days. After 3 months a patient's dose was increased by 300 μg if month 3 UFC$>$ min$(2 \times$ ULN$,$ baseline UFC$)$, where ULN is Upper Limit of Normal (145 nmol/24hr). The primary efficacy measure was the percent of responders for each dose at month 6. Responders were defined as patients whose UFC$<$ULN at month 6 without a dose increase at month 3. Those patients whose dose was increased were considered non-responders, regardless of their month 6 UFC levels. Furthermore, there is a secondary endpoint with the ordered outcome of controlled, partially controlled, and uncontrolled, where controlled was the most favorable response and uncontrolled was the least favorable. These outcomes are respectively defined as month 6 UFC $\leq$ ULN (regardless of dose increase), month 6 as UFC $>$ ULN but UFC decreased by at least 50% from baseline (regardless of dose increase), and neither controlled nor partially controlled at month 6.

Our post-hoc analysis will address the effect of treatment on the log odds of response as both binary and ordinal outcomes taking into account more than just the month 3 and month 6 UFC levels. In addition to sampling the UFC at baseline, there were 2 sampling visits before the month 3 visit, the month 3 visit, 2 sampling visits between the month 3 and month 6 visits, and the month 6 visit for a total of 6 sampling visits. As responder status is only defined for month 6, we defined a responder before or at month 3 as a patient whose UFC$<$ min$(2 \times$ ULN, baseline UFC$)$. If a patient was not a responder at month 3, they could not be a responder at later visits due to the increase in dose. If they were a responder at month 3, then for visits after month 3, patients were classified as a responder if their post month 3 UFC was lower than ULN. Furthermore, we will examine the data looking at the visit effect on treatment as well as the role of missing data.

When dealing with the secondary (ordinal) outcome, we defined the categories as controlled (UFC $\leq$ ULN (regardless of dose increase)), partially controlled (UFC $>$ ULN but UFC decreased by at least 50% from baseline (regardless of dose increase)), and uncontrolled (neither controlled nor partially controlled). The definition for the ordinal outcome is the same at all 6 visits.

Due to the limitations of the availability of the data, the only baseline covariate we have is the baseline UFC level. Since we have arbitrary missingness of the response variable, we used PROC GENMOD in SAS with a within subject specification on the repeated specification so as to account for the way responses are aligned when the correlations between responses are calculated. The dfbetas produced using PROC GENMOD and unstructured correlation are similar within a reasonable order of magnitude to the ones produced by Hammill and Preisser (2006), but they can allow for different missing patterns.

4.4.1 DICHOTOMOUS OUTCOME

In the data set, we have 162 subjects ($K = 162$) and six visits ($J = 6$). The covariate we will adjust is log(baseline UFC). The log transformation normalizes the UFC distribution. The high dose group has a log(baseline UFC) of 6.33 and the low dose has 6.52. Baseline has no missing UFC values, and a visit has at most 21.0% of the 162 responses missing. The distribution of responders by visit is shown in Table 4.7. The distribution of responders by visit and treatment is shown in Table 4.8.

Using the methods described previously, we compare the randomization-based non-parametric covariate adjustment method to the unadjusted method in Table 4.9. We can see there that the method tends to adjust the estimate of the treatment by visit effect, but the confidence intervals are narrower. Since the definition of responder status for visits 4-6 is different from the definition of responder at visits 1-3, a test of homogeneity

of treatment effect across visits should take this into account. Furthermore, it appears from Table 4.8 that there could be a potential delay in the treatment effect, so we will look at combining visits 2 and 3, and visits 4-6. Our test of homogeneity will test

$$\begin{aligned} H_0 : &\eta_2 = \eta_3 \\ &\eta_4 = \eta_5 = \eta_6 \end{aligned} \tag{4.24}$$

where η_j is the covariance adjusted treatment parameter at the j^{th} visit. The test statistic of homogeneity of treatment effect across visits managing visits 2-3 as distinct from visits 4-6 was 4.62 with 3 df, which has a p-value of 0.2020. The data suggest that the treatment effect is homogeneous within these 2 sets of visits. We also evaluate whether combining visits 1-3 and 4-6 versus visits 2-3 and 4-6, but the test statistic was 9.61 with 4 df, which had a p-value of 0.048. Since the data suggest that this combination is significantly different from 0, we will use the initial suggestion of combining visits 2 and 3, and combining visits 4-6.

The new treatment estimates and its estimated covariance matrix are shown in Equation 4.30, where $\boldsymbol{c} = (c_1, c_2, c_3)$ solves Equation 4.26. Let c_1 be the estimator of the adjusted treatment effect on the log odds of responder status at visit 1, c_2 be the estimator of the adjusted treatment effect on the log odds at visits 2 and 3, and c_3 be the estimator of the adjusted treatment effect on the log odds at visits 4, 5, and 6. The test statistic for all three treatment parameters being equal to 0 is 6.39 with 3 df and a p-value of 0.0942. The estimates of the adjusted treatment effects in the condensed version (Equation 4.30) suggest that there separate assessments are of interest. The test statistic for $c_3 = 0$ is 2.08 with 1 df and a p-value of 0.1492. This higher p-value could be due to the large estimated variance for c_3. Thus, there is not enough evidence to suggest that the two doses have significantly different odds of a patient being a responder.

$$\begin{aligned} H_0 : &\eta_1 = 0 \\ &\eta_2 = \eta_3 = 0 \\ &\eta_4 = \eta_5 = \eta_6 = 0 \end{aligned} \tag{4.25}$$

$$\boldsymbol{b} \hat{=} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} \tag{4.26}$$

Table 4.7: Distribution of binary response pooled over treatment

Visit	Missing	Non-Responder	Responder
1	18	65	79
2	30	61	71
3	31	60	71
4	30	97	35
5	34	97	31
6	32	101	29

Table 4.8: Distribution of binary response by visit and treatment

Visit	Dose	Missing	Non-Responder	Responder
1	Low	8	34	40
	High	10	31	39
2	Low	16	37	29
	High	14	24	42
3	Low	17	33	32
	High	14	27	39
4	Low	17	53	12
	High	13	44	23
5	Low	18	51	13
	High	16	46	18
6	Low	17	55	10
	High	15	46	19

Table 4.9: Comparison of randomization based non-parametric covariate adjustment with unadjusted treatment estimates

	Unadjusted			Adjusted		
Visit	OR Estimate	95% LCL	95% UCL	OR Estimate	95% LCL	95% UCL
1	0.99	0.51	1.90	0.85	0.49	1.60
2	2.25	1.13	4.48	1.97	1.00	3.88
3	1.43	0.72	2.83	1.21	0.62	2.36
4	1.96	0.90	4.27	1.75	0.80	3.81
5	1.42	0.65	3.10	1.27	0.58	2.78
6	2.22	0.97	5.10	2.00	0.87	4.61

$$\boldsymbol{d} = \begin{pmatrix} -0.0103 & 0.8118 & 0.3573 & 0.6721 & 0.3475 & 0.7995 & -0.1894 \end{pmatrix}', \tag{4.27}$$

$$\boldsymbol{V_d} = \begin{pmatrix} 0.1145 & 0.0682 & 0.0526 & 0.0484 & 0.0501 & 0.0502 & -0.0108 \\ 0.0682 & 0.1264 & 0.0635 & 0.0524 & 0.0543 & 0.0564 & -0.0093 \\ 0.0526 & 0.0635 & 0.1250 & 0.0764 & 0.0747 & 0.0728 & -0.0114 \\ 0.0484 & 0.0524 & 0.0764 & 0.1631 & 0.1167 & 0.1267 & -0.0079 \\ 0.0501 & 0.0543 & 0.0747 & 0.1167 & 0.1646 & 0.1363 & -0.0076 \\ 0.0502 & 0.0564 & 0.0728 & 0.1267 & 0.1363 & 0.1850 & -0.0072 \\ -0.0108 & -0.0093 & -0.0114 & -0.0079 & -0.0076 & -0.0072 & 0.0130 \end{pmatrix} \tag{4.28}$$

$$\boldsymbol{b} = \begin{pmatrix} -0.1674 \\ 0.6771 \\ 0.1922 \\ 0.5576 \\ 0.2365 \\ 0.6950 \end{pmatrix}, \boldsymbol{V_b} = \begin{pmatrix} 0.1055 & 0.0606 & 0.0432 & 0.0419 & 0.0438 & 0.0443 \\ 0.0606 & 0.1198 & 0.0554 & 0.0468 & 0.0489 & 0.0513 \\ 0.0432 & 0.0554 & 0.1151 & 0.0695 & 0.0680 & 0.0665 \\ 0.0419 & 0.0468 & 0.0695 & 0.1584 & 0.1121 & 0.1224 \\ 0.0438 & 0.0489 & 0.0680 & 0.1121 & 0.1602 & 0.1321 \\ 0.0443 & 0.0513 & 0.0665 & 0.1224 & 0.1321 & 0.1810 \end{pmatrix} \tag{4.29}$$

$$\boldsymbol{c} = \begin{pmatrix} -0.2328 \\ 0.4238 \\ 0.5224 \end{pmatrix}, \boldsymbol{V_c} = \begin{pmatrix} 0.1030 & 0.0515 & 0.0458 \\ 0.0515 & 0.0864 & 0.0588 \\ 0.0458 & 0.0588 & 0.1312 \end{pmatrix} \tag{4.30}$$

Since a placebo dose was considered as not ethical, the investigators considered a dose for which the lower 95% confidence limit of the proportion of responders was above 15% as having useful benefits in terms of UFC reduction. The 15% criterion was chosen as the cutoff by both an external Cushing's Advisory Board and the study's steering committee. As we see in Table 4.11, both doses showed significant signs of UFC reduction, with the high dose showing better UFC reduction.

Estimation of the proportion of responders was done by additionally including the six variables associated with visit intercepts in the vector for the treatment effects, and incorporating the corresponding dfbetas into

the covariance matrix calculations. If we let ζ_j be the parameter of the adjusted j^{th} visit intercept and a_j be its estimator, then if we are interested in simplifying the model, we can test

$$\begin{aligned} H_0 :&\zeta_2 = \zeta_3 \\ &\zeta_4 = \zeta_5 = \zeta_6 \\ &\eta_2 = \eta_3 \\ &\eta_4 = \eta_5 = \eta_6. \end{aligned}$$

$$\boldsymbol{d} = \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}} \\ (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \end{pmatrix} = \begin{pmatrix} 0.6721 \\ 0.3475 \\ 0.7995 \\ -0.0103 \\ 0.8118 \\ 0.3573 \\ 0.1876 \\ -0.2371 \\ -0.0275 \\ -1.3771 \\ -1.2758 \\ -1.6607 \\ -0.1894 \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} 0.5576 \\ 0.2365 \\ 0.6950 \\ -0.1674 \\ 0.6771 \\ 0.1922 \\ 0.2608 \\ -0.1961 \\ 0.0475 \\ -1.3175 \\ -1.2218 \\ -1.6307 \end{pmatrix} \tag{4.31}$$

$$\boldsymbol{c} = \begin{pmatrix} 0.3522 \\ -0.0796 \\ -1.3613 \\ -0.2943 \\ 0.4397 \\ 0.5571 \end{pmatrix}, \boldsymbol{V_c} = \begin{pmatrix} 0.0496 & 0.0271 & 0.0227 & -0.0468 & -0.0243 & -0.0210 \\ 0.0271 & 0.0472 & 0.0320 & -0.0252 & -0.0453 & -0.0309 \\ 0.0227 & 0.0320 & 0.0747 & -0.0212 & -0.0305 & -0.0737 \\ -0.0468 & -0.0252 & -0.0212 & 0.0993 & 0.0519 & 0.0463 \\ -0.0243 & -0.0453 & -0.0305 & 0.0519 & 0.0862 & 0.0583 \\ -0.0210 & -0.0309 & -0.0737 & 0.0463 & 0.0583 & 0.1300 \end{pmatrix} \tag{4.32}$$

The a_j are simplified similarly to the treatment effect since responder status of those on low dose also depends on the responder status at visit 3. The test statistic for homogeneity of visit and treatment effects within visits 2-3 and 4-6 is 6.36 with 6 df and a p-value of 0.3844. The data suggest we can simplify our model to have six parameters ($\boldsymbol{c} = (c_1, c_2, c_3, c_4, c_5, c_6)$), as shown in Equation 4.32. Let c_1 be the estimator for the visit 1 intercept on the log odds of responder status, c_2 be the estimator for visits 2 and 3 effect, c_3 be estimator for visits 4,5, and 6 effect; and let c_4 be the estimator for the high dose effect for visit 1, c_5 be the estimator for the high dose effect on visits 2 and 3, and c_6 be estimator of the high dose effect for visits 4, 5, and 6.

Let $\boldsymbol{V_c}[i, j]$ be the element in the i^{th} row and j^{th} column of matrix $\boldsymbol{V_c}$. The lower 95% CL of the estimates of the log odds of being a responder given dose and time can be estimated in terms of parameters as given in Equation 4.10. We can use the inverse logit function to find the estimated proportion.

Table 4.10: Lower 95% CL of odds of responders on low and high dose at 3 time points

Visit	Low Dose	High Dose
Visit 1	$c_1 - 1.96\sqrt{\boldsymbol{V_c}[1,1]}$	$c_1 + c_4 - 1.96\sqrt{\boldsymbol{V_c}[1,1] + \boldsymbol{V_c}[4,4] + 2\boldsymbol{V_c}[1,4]}$
Visit 2 and 3	$c_2 - 1.96\sqrt{\boldsymbol{V_c}[2,2]}$	$c_2 + c_5 - 1.96\sqrt{\boldsymbol{V_c}[2,2] + \boldsymbol{V_c}[5,5] + 2\boldsymbol{V_c}[2,5]}$
Visit 4, 5, and 6	$c_3 - 1.93\sqrt{\boldsymbol{V_c}[3,3]}$	$c_3 + c_6 - 1.96\sqrt{\boldsymbol{V_c}[3,3] + \boldsymbol{V_c}[6,6] + 2\boldsymbol{V_c}[3,6]}$

From Table 4.11, we see that the lower 95% CL of the proportion of responders for the high dose is greater than 15% at visit 6, but the low dose slightly misses this criterion. This indicates that the high dose is sufficiently effective and the low dose is not quite sufficient by a small margin.

Table 4.11: Lower 95% CL of proportion of responders on low and high dose at 3 time points

Visit	Low Dose	High Dose
Visit 1	0.4790	0.4007
Visit 2 and 3	0.3763	0.4887
Visit 4, 5, and 6	0.1305	0.2187

4.4.2 ORDINAL OUTCOME

The distribution of the ordinal response is shown in Table 4.12. When a subject is controlled, $Y_{htj} = 3$, and when a subject is uncontrolled, $Y_{htj} = 1$. Therefore, a cutoff of 3 means that we are modeling the log odds of having controlled UFC levels, while a cutoff of 2 means that we are modeling the log odds of having at least partially controlled UFC levels. When using PROC GENMOD to calculate the dfbetas, an independence working correlation had to be used to obtain model convergence. Therefore, we only have twelve treatment effects. The test statistic for the test of proportional odds is 23.56 with 6 df and a p-value of 0.0006, which suggests that the treatment by visit effect is significantly different across the two cutoffs. We can see this from Equation 4.34, as the estimates for the cutoff of 3 are positive, and the estimates for the cutoff of 2 are mostly negative. The test statistic for whether the treatment effects are homogeneous within two classes of visits (2-3 versus 4-6) for each cutoff is 11.57 with 6 df and a p-value of 0.1713. The data suggest that we can use a simplified model (Equation 4.33)

$$\boldsymbol{d} = \begin{pmatrix} \hat{\beta}_{21} \\ \vdots \\ \hat{\beta}_{26} \\ \hat{\beta}_{31} \\ \vdots \\ \hat{\beta}_{36} \\ (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \end{pmatrix} \hat{=} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} c_{21} \\ c_{22} \\ c_{23} \\ c_{31} \\ c_{32} \\ c_{33} \end{pmatrix}. \tag{4.33}$$

Let c_{p1} be the estimator of the adjusted high dose effect on log odds of responder status at visit 1 for the p^{th} cutoff, c_{p2} is the estimator of the adjusted high dose effect on log odds of responder status at visit 2 and 3 for the p^{th} cutoff, and c_{p3} is the estimator of the adjusted high dose effect on log odds of responder status at visit

Table 4.12: Distribution of ordinal responses by treatment

Visit	Dose	Controlled	Partially controlled	Uncontrolled
1	Low dose	44	50	54
	High dose	54	8	78
2	Low dose	28	40	64
	High dose	40	26	66
3	Low dose	26	40	64
	High dose	44	26	62
4	Low dose	28	32	64
	High dose	52	16	54
5	Low dose	28	40	52
	High dose	42	20	50
6	Low dose	22	32	58
	High dose	42	16	52

4, 5, and 6 for the p^{th} cutoff. The estimates of $\boldsymbol{c}$ are shown in Equation 4.35. Since we are only interested in the high dose effect at visit 6, we can test whether the parameters corresponding to visit 6 for each cutoff are equal to 0. Our test statistic is 8.22 with 2 df and a p-value of 0.0164. This suggests that the high dose has a significant effect on the log odds of more favorable outcome (i.e. more controlled) for one or both cutoffs at visit 6. The estimates for the high dose for an outcome of controlled versus partially controlled or uncontrolled are all positive, while the estimates for high dose for an outcome of at least partially controlled versus uncontrolled gradually increase. This trend in an outcome of controlled versus other suggests that the high dose has a positive significant effect on the log odds of controlled versus partially or uncontrolled, and a possible insignificant effect on the log odds of at least partially controlled versus uncontrolled.

$$\boldsymbol{d} = \begin{pmatrix} -0.7839 \\ -0.0606 \\ 0.0906 \\ 0.2951 \\ -0.0532 \\ 0.1807 \\ 0.3948 \\ 0.4793 \\ 0.6931 \\ 0.9349 \\ 0.6788 \\ 0.9269 \\ -0.1894 \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} -0.8269 \\ -0.1173 \\ 0.0496 \\ 0.2651 \\ -0.0920 \\ 0.0587 \\ 0.4484 \\ 0.4817 \\ 0.6476 \\ 0.9229 \\ 0.6704 \\ 0.8604 \end{pmatrix} \tag{4.34}$$

$$\boldsymbol{c} = \begin{pmatrix} -0.8519 \\ -0.0289 \\ 0.0800 \\ 0.4320 \\ 0.5164 \\ 0.7918 \end{pmatrix}, \boldsymbol{V_c} = \begin{pmatrix} 0.1179 & 0.0584 & 0.0511 & 0.0847 & 0.0407 & 0.0392 \\ 0.0584 & 0.0918 & 0.0607 & 0.0396 & 0.0625 & 0.0512 \\ 0.0511 & 0.0607 & 0.0914 & 0.0384 & 0.0584 & 0.0726 \\ 0.0847 & 0.0396 & 0.0384 & 0.1232 & 0.0579 & 0.0633 \\ 0.0407 & 0.0625 & 0.0584 & 0.0579 & 0.1266 & 0.0871 \\ 0.0392 & 0.0512 & 0.0726 & 0.0633 & 0.0871 & 0.1228 \end{pmatrix} \tag{4.35}$$

4.5 DISCUSSION

To summarize, our method consists of calculating the unadjusted (for covariates) treatment effects for outcomes nested within visits (and cutoffs if outcome is ordinal), and creating a vector of the treatment effects and the differences in the means of the covariates ($\boldsymbol{d}$). U-statistic-based quantities and dfbetas are used to calculate the covariance matrix $\boldsymbol{V_d}$ of $\boldsymbol{d}$. Then WLS regression is used to regress $\boldsymbol{d}$ only onto the space of treatment parameters for outcomes. Finally, hypothesis testing occurs with contrast matrices. While adjustment is non-parametric due to the distribution-free method of variance calculation of covariates, it still

yields a consistent, asymptotically normally distributed estimator of the treatment effects when sample sizes are reasonably large as is the case for many confirmatory clinical trials.

Our method provides an easy-to-implement non-parametric randomization based covariate adjustment for treatment effects for responses in randomized clinical trials with only a few requirements and limitations. Its only requirements are the population has been randomized to one of two groups and that we have large enough sample size for the GEE method to provide approximately normal estimates and consistent estimates for their covariance matrix. We need the GEE estimation to obtain the initial treatment effect estimates for responses and dfbetas for variance calculations. We rely on the randomization basis to force the differences of covariates to 0. One limitation of the method is that it only adjusts for covariates through their linear correlation with the treatment effect, and so it is possible that this method would not be good with respect to improving power for covariates which have a known non-linear relationship with the outcome (unless that relationship was taken into account).

When we have large sample size (100 subjects per treatment) and missing are less than 20%, then we want to use GEE with unstructured correlation so that our estimates are MAR-like. If we have monotone missingness, we can use the macro provided by Hammill and Preisser (2006) which provides estimates of bias-corrected covariance matrices. If we have intermittent missingness among the responses, we can use methods such as PROC GENMOD in SAS. If we have medium sample size, then we can use an independence specification for correlation if we suspect our data are MCAR. If we do not have a MCAR situation, we can use multiple imputation methods to fill the missing responses and an unstructured specification to generate our dfbetas for the covariance matrix estimation.

The requirements are easily met in a regulatory setting, and further development of the method could make it more appealing. One further development is to include more than two treatments. This could potentially be done similarly to the partial proportional odds model where the treatment effect vector is expanded to include other treatments as opposed to other cutoffs. Another possible extension would be to account for covariates with missing data.

CHAPTER 5: RANDOMIZATION-BASED NON-PARAMETRIC COVARIANCE ADJUSTMENT FOR TIME-TO-EVENT OUTCOMES WITH MISSING DATA

5.1 INTRODUCTION

Clinical trials often have time-to-event outcomes rather than cross-sectional or change from baseline endpoints. A popular method for analyzing such data is the Cox proportional hazards model (Cox, 1972), which estimates the effects of the covariates on the hazard ratio relative to an unspecified baseline hazard. It assumes that hazards are proportional between groups with respect to time. Covariate adjustment can provide increased power for tests of treatment versus control (Jiang et al., 2008), but it requires each covariate to meet the assumption of proportional hazards. One can do non-parametric tests of treatment versus placebo using logrank and Wilcoxon tests (Gehan, 1965; Peto and Peto, 1972), and randomization-based non-parametric covariate adjustment for tests of treatment versus placebo can be carried out using methods developed by Saville and Koch (2013) and Tangen and Koch (2001).

The most appealing feature of a method like the one proposed in Saville and Koch (2013) is that it requires few assumptions, produces interpretable estimates of hazard ratios with estimated covariance matrices or variances, and is computationally straight-forward to implement. The main requirement in the Saville and Koch (2013) method is that subjects are randomized to treatment and placebo. This paucity of assumptions is attractive in a regulatory clinical trial setting in which analysis plans must be stated *a priori* (LaVange et al., 2005). In contrast, the proportional hazards assumption complicates covariate adjustment in analysis using the Cox proportional hazards model because it is not known if the covariates will satisfy this assumption without looking at the data. Furthermore, the Cox proportional hazards model nor those of Saville and Koch (2013) and Tangen and Koch (2001) do not specify how to handle missing data among the covariates.

Our method for the comparison of randomized treatment to placebo requires the missing data to be MAR (and in some implementations MCAR), allows missing covariates to be both continuous and categorical, contains easily implemented computations, and does not require specification of either the missing data

distribution or mechanism for covariates. It also provides meaningful parameter estimates for treatment comparison to placebo as well as corresponding estimates of variance.

We follow the randomization based non-parametric covariate adjustment of treatment effect method of Saville and Koch (2013), except for the covariance matrix estimate for the covariates is produced via generalized linear models of the transformed covariate values. In Section 5.2 we review the method of Saville and Koch (2013) and extend it to the case of missing covariates. We apply our method to an oncology trial in Section 5.3 and make comparisons with other methods. Finally, in Section 5.4 we discuss strengths and limitations of our method.

5.2 METHOD

The general method of randomization-based non-parametric covariance adjustment consists of four steps. First, we model the treatment effect for outcomes. We have survival data for K outcomes for a randomized clinical trial evaluating one group, which we will call treatment, versus another, which we will designate as placebo. We will follow the notation used in Saville and Koch (2013), however in this case, covariates are allowed to be missing. Let M be the number of covariates for adjustment. We define a $(K + M) \times 1$ vector, $\boldsymbol{d}$, as the vector of the K estimates of treatment effects by outcomes stacked on top of the M estimates of the differences of the means of the covariates of the treatment from the placebo. We will refer to the M estimates for the differences of the means of the covariates of the treatment from the placebo as the differences of the means of the covariates for brevity.

In the second step, we estimate the covariance matrix of $\boldsymbol{d}$, which involves estimating three covariance matrices: the covariance matrix of the estimates of treatment effects for the K outcomes, the covariance matrix of the estimates of the differences of the means of the covariates, and the covariance matrix of the estimates of treatment effect for outcomes with those for the comparisons of the covariates. We can calculate the covariance matrix of the treatment effects for outcomes by using output associated with the time-to-event analysis (Wei et al., 1989), which is explained later in further detail. Next, we calculate the covariance matrix of the estimated differences of the means of the covariates and the covariance matrix of the treatment effects with the differences in the means of the covariates using one of three methods which are the main contributions of this paper. The covariance matrix of $\boldsymbol{d}$, $\boldsymbol{V_d}$, is a $(K + M) \times (K + M)$ matrix.

Third, we force the estimated differences in means of covariates to 0 through weighted least squares (WLS) regression. We regress $\boldsymbol{d}$, which contains both treatment effects for outcomes and covariate differences estimates, onto the space of only treatment parameters for outcomes. We use WLS with the weights based on$\boldsymbol{V_d}^{-1}$. Forcing the estimated differences in means of the covariates to 0 through WLS is appropriate due to the randomization since we assume these differences in the means of the covariates are random and the population receiving the treatment is the same as the population receiving the placebo. From WLS methods, we obtain both the estimator for the adjusted treatment effects ($\boldsymbol{b}$) and a consistent estimator for their covariance matrix ($\boldsymbol{V_b}$).

The fourth step is to do hypothesis testing using the approximately multivariate normal distribution of $\boldsymbol{b}$ (via sufficiently supportive sample size), with the corresponding expected value and covariance matrix estimated in the WLS in step 3. We can test hypothesis using traditional contrast matrices and calculate estimates of parameters from linear combinations of $\boldsymbol{b}$.

To carry out these four steps it will be helpful to explain some notation. Let T_{jk} be the survival time for the k^{th} outcome for the j^{th} patient. Let z_j be the j^{th} subject's indicator of treatment; $z_j = 1$ if subject j receives treatment and 0 otherwise. For each outcome event k, we will use a marginal Cox proportional hazards model to estimate the treatment effects. Our model for the estimate of treatment effects is

$$\lambda_{jk}(t|z_j) = \lambda_{0k}(t)\exp(\beta_k z_j). \tag{5.1}$$

In Equation 5.1, $\lambda_{0k}(t)$ is the baseline hazard for outcome k given the subject has survived until time t. Let β_k be the treatment effect on the log hazard ratio for the k^{th} outcome. We assume that we have independent random censoring, which is a MAR-like assumption. Let i index group, and $i = 1$ if we want to indicate the treatment group, and $i = 2$ if we want to indicate the placebo group. We assume the first n_1 patients are in group 1 (treatment), the last n_2 patients are in group 2 (placebo), and $n_1 + n_2 = n$, where n is the total number of patients in the study. Let $\boldsymbol{R} = (\boldsymbol{R}_1', \boldsymbol{R}_2')'$ where $\boldsymbol{R}_1 = (\boldsymbol{r}_{11}, \ldots, \boldsymbol{r}_{1n_1})'$ be the $n_1 \times K$ matrix of dfbeta residuals for group 1 obtained from the fitted unadjusted Cox proportional hazards model for each of the K events, and $\boldsymbol{R}_2 = (\boldsymbol{r}_{2n_1+1}, \ldots, \boldsymbol{r}_{2n})'$ be the $n_2 \times K$ matrix of dfbetas residuals for group 2. The components of $\boldsymbol{R}_i$ are $\boldsymbol{r}_{ij}$, a $K \times 1$ column vector of dfbeta residuals for the treatment effects for the K outcomes for patient j in group i. Let $\hat{\boldsymbol{\beta}}$ be the $K \times 1$ vector of estimated log hazard ratios for treatment versus placebo for the K outcomes.

The estimated differences in the means of the covariates is expressed as $\hat{\boldsymbol{\theta}}$. When all covariates have no missing data, $\hat{\boldsymbol{\theta}} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$, where $\bar{\boldsymbol{x}}_i = (\bar{x}_{i1}, \ldots, \bar{x}_{iM})$ is the vector of M baseline covariate means for group i, and $\bar{x}_{im}$ denotes the mean of covariate m for group i.

To calculate the covariance matrix of $\boldsymbol{d} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\theta}}')'$ we need to calculate $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$, $\boldsymbol{V}_{\hat{\boldsymbol{\beta}},\hat{\theta}}$, and $\boldsymbol{V}_{\hat{\theta}}$. Let $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = (\boldsymbol{R}_1'\boldsymbol{R}_1 + \boldsymbol{R}_2'\boldsymbol{R}_2)$ be the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$, $\boldsymbol{V}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}}$ be the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\theta}}$, and $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}}$ be the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$. When all covariates have no missing data, we use $\boldsymbol{C}_i = \frac{\boldsymbol{X}_i - \mathbf{1}_{n_i}\bar{\boldsymbol{x}}_i'}{\sqrt{n_i(n_i-1)}}$ where $\boldsymbol{X}_i$ is the $n_i \times M$ matrix of observed values for group i as discussed in Saville and Koch (2013); i.e., $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}} = (\boldsymbol{C}_1'\boldsymbol{C}_1 + \boldsymbol{C}_2'\boldsymbol{C}_2)$ and $\boldsymbol{V}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}} = (\boldsymbol{R}_1'\boldsymbol{C}_1 - \boldsymbol{R}_2'\boldsymbol{C}_2)$. Let $\mathbf{1}_{n_i}$ denote the $n_i \times 1$ column vector of 1's. Each subject contributes one row to $\boldsymbol{X}_i$, and the columns are the M covariates. To calculate $\boldsymbol{V}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}}$ and $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}}$ in the presence of missing data where we have U of the M covariates containing missing data, we suggest a method using dfbetas from a model of covariates to estimate covariance models.

5.2.1 DFBETAS FROM A SCALED MODEL OF COVARIATES

In order to obtain an estimate of the differences of the means of the covariates for the treatment group versus placebo comparison in the presence of missing data, we fit generalized linear models with repeated measures within subjects. We transform all the covariates onto a 0-1 scale by subtracting a plausible minimum possible value from the covariate value and dividing the difference by a plausible range of the covariate (Equation 5.2). For categorical variables with v levels, we create $v - 1$ indicators of level and the covariate can take the value 0 or 1 for the $v - 1$ indicators. We then use PROC GENMOD and generalized estimating equations (GEE) in SAS 9.4 to model the effect of treatment nested within covariate type on the transformed covariate values with the identity link, normal distribution, and unstructured working correlation matrix as in Equation 5.3. The use of unstructured covariance allows us to be reasonably sure that we have the correct correlation structure specified. When the correct correlation structure is specified, simulation studies (Preisser et al., 2002) show that the expected bias in treatment parameters is neglible even under MAR conditions. Then we use dfbeta residuals to obtain an estimated covariance matrix of the differences in the means of the covariates. The covariate dfbeta can be used to replace $\boldsymbol{C}_i$ in the complete data method with some transformation (Equation 5.2) to address the scaling to the 0-1 scale.

Let x_{imj} be the value of covariate m for patient j in group i. Let a_m be the minimum possible population value for covariate m and p_m be the maximum value. We define our covariate trasnformation as

$$f(x_{imj}) = \frac{x_{imj} - a_m}{p_m - a_m}, \tag{5.2}$$

which puts the covariates on the 0-1 scale and allows GEE to be more computationally straightforward. If x_{imj} is missing then $f(x_{imj})$ is missing. We want to model the value of the $M \times 1$ vector of covariates for patients in group i ($\boldsymbol{x}_{1m} = (x_{1m1}, \ldots, x_{1mn_1})'$, and $\boldsymbol{x}_{2m} = (x_{2mn_1+1}, \ldots, x_{2mn})'$) using the following model:

$$\begin{aligned} f(\boldsymbol{x}_{im}) =& \eta_1(m=1) + \ldots + \eta_M(m=M) + \\ & \phi_1(m=1)(i=1) + \ldots + \phi_m(m=M)(i=1). \end{aligned} \tag{5.3}$$

In Equation 5.3, η_m is the intercept for estimating a transformation of covariate m for a subject in group 2 and ϕ_m is the effect of being in group 1 on the estimate of the transformation of covariate m. We use SAS v9.4's GEE implementation to obtain the dfbetas associated with $\boldsymbol{\phi}$ (Preisser and Qaqish, 1996). Let $\boldsymbol{g}_{ij}$ be the $M \times 1$ vector of dfbetas corresponding to $\hat{\boldsymbol{\phi}}$ for patient j in group i. The $\boldsymbol{g}_{ij}$ are components of $\boldsymbol{G}_1 = (\boldsymbol{g}_{11}, \ldots, \boldsymbol{g}_{1n_1})'$, the $n_1 \times M$ matrix of dfbetas for patients in group 1, and $\boldsymbol{G}_2$, the corresponding $n_2 \times M$ matrix of dfbetas for patients in group 2.

To obtain the estimated differences in means of the covariates we compute $\hat{\boldsymbol{\theta}} = \boldsymbol{D}\hat{\boldsymbol{\phi}}$ where $\boldsymbol{D} = \boldsymbol{diag}(\boldsymbol{p} - \boldsymbol{a})$ is a diagonal $M \times M$ matrix with the length of the ranges of the M covariates along the diagonal, where $\boldsymbol{p} = (p_1, \ldots, p_M)$ and $\boldsymbol{a} = (a_1, \ldots, a_M)$. The estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ is $\text{Cov}(\boldsymbol{D}\hat{\boldsymbol{\phi}}) = \boldsymbol{D}\text{Cov}(\hat{\boldsymbol{\phi}})\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{G}_1'\boldsymbol{G}_1 + \boldsymbol{G}_2'\boldsymbol{G}_2)\boldsymbol{D}$ and $\boldsymbol{V}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}} = \boldsymbol{R}_1'\boldsymbol{G}_1\boldsymbol{D} + \boldsymbol{R}_2'\boldsymbol{G}_2\boldsymbol{D}$. In addition to the transformation of the dfbetas for the covariates by the range of the covariates, this formulation of the covariates differs from previous methods in that the terms are added instead of subtracted when calculating the estimated covariance matrix of the treatment effects for outcomes with the differences in the means of the covariates. This is because the dfbeta residuals of the covariate model estimate the scaled difference itself, there is no need to subtract the terms. The method then proceeds with WLS being performed on $\boldsymbol{d}$. The main feature of this method is the difference in the calculation of elements of $\boldsymbol{V}_{\boldsymbol{d}}$. We will compare our method of

handling missing covariate data with the traditional method of multiple imputation or imputation of the mean, so those methods will be discussed briefly.

5.2.2 MULTIPLE IMPUTATION

Another way to deal with missingness among baseline covariates is to create many data sets with imputed values. Let P be the number of imputed data sets to be created, and $\boldsymbol{R}'\boldsymbol{R}$ be the same estimate for all data sets as there are no missing treatment values. The p^{th} imputed data set has its own $\boldsymbol{C}_{ip} = \frac{\boldsymbol{X}_{ip} - \mathbf{1}_{n_i}\bar{\boldsymbol{x}}'_{ip}}{\sqrt{n_i(n_i-1)}}$ where $\boldsymbol{X}_{ip}$ is the $n_i \times M$ matrix of observed and imputed values for group i, imputation p. Each subject contributes one row to $\boldsymbol{X}_{ip}$, and the columns are the M covariates. Furthermore $\bar{\boldsymbol{x}}_{ip}$ is the column vector of covariate means for group i, imputation p, and $\mathbf{1}_{n_1}$ is a $n_i \times 1$ vector of 1's. Therefore,

$$\boldsymbol{d}_p = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \bar{\boldsymbol{x}}_{1p} - \bar{\boldsymbol{x}}_{2p} \end{pmatrix}, \boldsymbol{V}_{\boldsymbol{d}_p} = \begin{bmatrix} \boldsymbol{R}'\boldsymbol{R} & \boldsymbol{R}'_1\boldsymbol{C}_{1p} - \boldsymbol{R}'_2\boldsymbol{C}_{2p} \\ \boldsymbol{C}'_{1p}\boldsymbol{R}_1 - \boldsymbol{C}'_{2p}\boldsymbol{R}_2 & \boldsymbol{C}'_{1p}\boldsymbol{C}_{1p} + \boldsymbol{C}'_{2p}\boldsymbol{C}_{2p} \end{bmatrix}. \tag{5.4}$$

For each of the $p = 1, \ldots, P$ imputed data sets, let $\boldsymbol{b}_p$ be the adjusted treatment effect for the log hazard ratio of treatment to placebo for the p^{th} imputation. Furthermore $\boldsymbol{V}_{\boldsymbol{b}_p}$ is the estimated covariance matrix of $\boldsymbol{b}_p$. Using PROC MIANALYZE in SAS 9.4, we can combine our estimates of the P $\boldsymbol{b}_p$'s and $\boldsymbol{V}_{\boldsymbol{b}_p}$'s to obtain an overall $\boldsymbol{b}$ ($\boldsymbol{b}_*$) and $\boldsymbol{V}_{\boldsymbol{b}}$ ($\boldsymbol{V}_{\boldsymbol{b}_*}$), which we can then use for standard estimation and hypotheses tests of parameters. MIANALYZE combines parameter estimates and associated standard errors or covariance matrices of the imputed data sets and then can enable univariate inference for these parameters.

An alternative method would be to combine our estimates of the P $\boldsymbol{d}_p$'s and $\boldsymbol{V}_{\boldsymbol{d}_p}$'s to get an overall estimate of $\boldsymbol{d}$ ($\boldsymbol{d}_*$) and $\boldsymbol{V}_{\boldsymbol{d}}$ ($\boldsymbol{V}_{\boldsymbol{d}_*}$), which we can then use for WLS. Then $\boldsymbol{d}_*$ and $\boldsymbol{V}_{\boldsymbol{d}_*}$ can be used to estimate the adjusted treatment effects, $\boldsymbol{b}_{**}$ and its covariance matrix $\boldsymbol{V}_{\boldsymbol{b}_{**}}$.

5.2.3 MISSINGNESS INDICATORS

For the method using missingness indicators, we created a missingness vector for each of the U missing variables. Without loss of generality, we assume the first U covariates contain missing values. Therefore, $\boldsymbol{m}_{i1}, \ldots, \boldsymbol{m}_{iU}$ are the missingness indicators for $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iU}$, respectively, and $\boldsymbol{x}_{iu}$ are the $n_i \times 1$ values for covariate u for patients in group i. The missing values in $\boldsymbol{x}_{iu}$ are replaced with the group i sample mean for covariate u. This imputation allows us to use the non-missing covariate information from patient j without

affecting ϕ too much. Let $\boldsymbol{m}_{iu}$ be the $n_i \times 1$ column vector whose elements are 1 when the corresponding $\boldsymbol{x}_{iu}$ element is missing and 0 otherwise. The average of $\boldsymbol{m}_{iu}$ is the proportion of missing for $\boldsymbol{x}_{iu}$. In the case of categorical variables with missing values, a corresponding set of dummy variables can have missing values replaced by corresponding means. Ordinal variables can be managed as numerical and have the missing values replaced by the mean.

Let $\boldsymbol{d}$ be a $(K + M + U) \times 1$ vector with the K treatment effects on outcome followed by the M differences in the means of the covariates, followed by the U differences in the means of the proportion of missingness for the U covariates containing missing values. The methodology proceeds as usual with $\boldsymbol{C}_i$ now a $n_i \times (M + U)$ matrix where the missingness indicators are managed like a usual covariate. WLS produces the adjusted treatment effects on the log of the hazard ratios for the K outcomes.

5.3 EXAMPLE

We applied the randomization based non-parametric covariance adjustment of treatment effect (referred to as adjusted treatment effect for brevity) to data from the Eastern Cooperative Oncology Group, who carried out a phase III clinical trial for high dose interferon on multiple melanoma patients. Survival time (in years) was defined as the time between enrollment in the study and progression of the tumor or death, whichever came first. In this data set, we have $n = 285$ subjects and either their time to relapse or time to censoring. There were 196 relapses and 89 censored observations. Therefore, we have only one outcome of interest ($K = 1$). Summary statistics of the survival time are shown in Table 5.1. Kaplan-Meier curves for the treatment groups are shown in Figure 5.1. The $M = 3$covariates of interest are age (17-78 years), Breslow score (0.2-35), and size (0-476). Only the Breslow score, which measures the thickness of the tumor, and size of the the tumor have missing observations ($U = 2$). The distribution of missingness by covariate is shown in Table 5.2 and the missingness of each variable by treatment is shown in Table 5.3. Observations with a survival time of 0 were recorded as 0.01 instead since survival times are always positive, which seemed reasonable given that the second smallest survival time was 0.03.

5.3.1 DFBETAS FROM A SCALED MODEL OF COVARIATES

In order to reduce the influence of outliers for PROC GENMOD with the identity link in SAS for calculation of the dfbetas, Breslow and size were transformed by adding 1 and then taking the natural log . The addition

Table 5.1: Summary of survival time (yrs) by drug group and status

Drug Group	Status	N Obs	Min	Quartile 1	Median	Quartile 3	Max
Placebo	Censored	35	1.9	5.12	6.41	7.8	9.64
	Relapsed	105	0.03	0.23	0.44	1.36	8.26
Treatment	Censored	54	0.01	4.95	7.02	8.04	9.63
	Relapsed	91	0.05	0.34	0.71	1.52	5.17

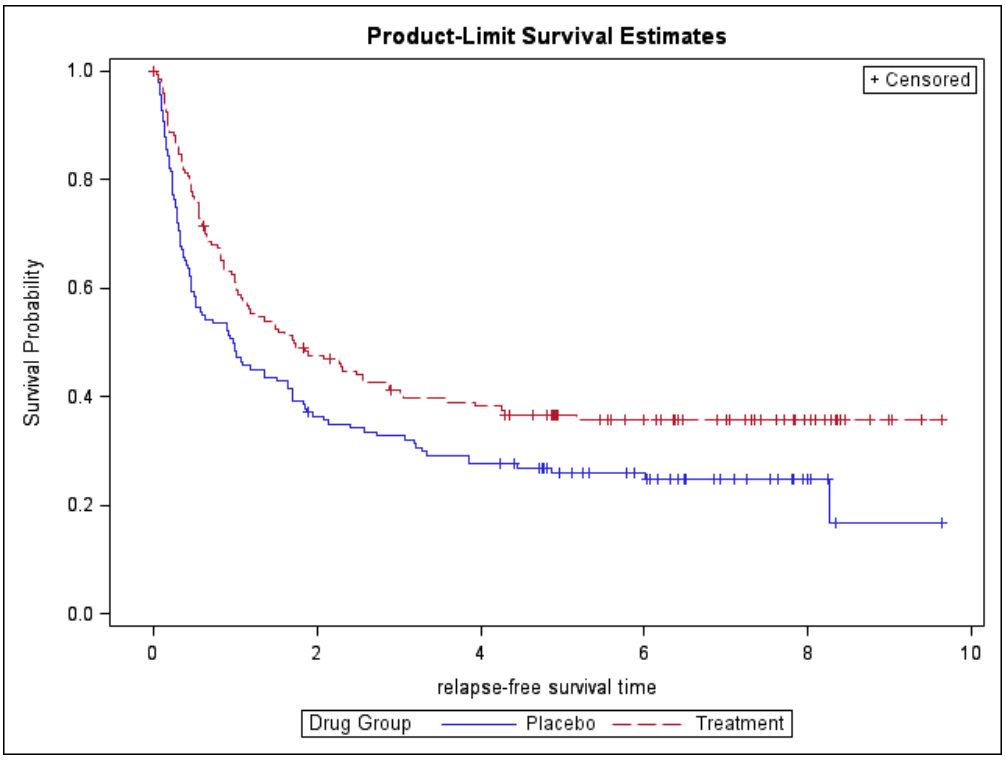

Figure 5.1: Kaplan-Meier plot of treatment versus placebo

of one is so that all transformed values are greater than or equal to 0. The square root and cubed root transformation were also considered but there were still long tails with the transformation so the $\log(x_{imj}+1)$ for $m's$ corresponding to Breslow and size were chosen. These transformed values were standardized using Equation 5.2. The population minimum, maximum, and range used to transform the variables onto the 0-1 scale are shown in Table 5.8. The randomization seems to have worked reasonably well as there does not appear to be large differences in the means of the covariates of the two treatment groups (Table 5.4). We also see from Equation 5.5 that the small covariance between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ means that not much will be gained by

Table 5.2: Distribution of missing observations by covariate

		Size		
		Present	Missing	Total
Breslow	Present	213	42	255
	Missing	19	11	30
	Total	232	53	285

Table 5.3: Number (percent) missing for covariates by treatment

Drug	N	Age	Breslow	Size
Placebo	140	0 (0%)	13 (9.3%)	26 (18.6%)
Treatment	145	0 (0%)	17 (11.7%)	27 (18.6%)
Total	285	0 (0%)	30 (10.5%)	53 (18.6%)

covariance adjustment

$$\boldsymbol{d} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\theta}} \end{pmatrix} = \begin{pmatrix} -0.357 \\ 1.038 \\ -0.098 \\ 0.089 \end{pmatrix}, \boldsymbol{V_d} = \begin{pmatrix} 0.0202 & 0.0137 & 0.0001 & 0.0004 \\ 0.0137 & 2.4221 & 0.0087 & 0.0240 \\ 0.0001 & 0.0087 & 0.0059 & 0.0014 \\ 0.0004 & 0.0240 & 0.0014 & 0.0162 \end{pmatrix}. \tag{5.5}$$

The adjusted treatment effect log hazard ratio is -0.363 (-0.641, -0.085). The unadjusted treatment effect on the log hazard scale compared to placebo is -0.357 (-0.638, -0.076). The parameter estimates themselves for the dfbeta and unadjusted method are reported in Table 5.9, and $\boldsymbol{V_b} = 0.020$ compared to the estimated variance of the unadjusted method of 0.0202. We have modest gains in the efficiency, although our estimate is farther from the null. The adjusted hazard ratio of the treatment is 0.695 (0.526 ,0.917), while the unadjusted hazard ratio is 0.700 (0.528, 0.927).

Table 5.4: Estimates of the differences in means (treatment-placebo) of covariates from PROC GENMOD on original scale

Variable	Difference Estimate	Standard Error
Age	1.038	1.545
Ln(Breslow+1)	-0.098	0.076
ln(Size+1)	0.089	0.126

5.3.2 MULTIPLE IMPUTATION METHOD

Using PROC MI in SAS 9.4 with MCMC, we created 30 imputed data sets that have imputed values for ln(Breslow+1) and ln(size+1). Because we had 11 subjects missing both variables, we used age which had no missing values to provide a basis for which to impute the other missing values for the other two covariates. We avoided using treatment to impute Breslow and size because it is preferable not to use treatment since any imbalances by treatment may be propagated in the imputed data set, and this would be in conflict with having randomization between groups.

Using the method of combining the $\boldsymbol{b}_p$, we get an overall adjusted treatment effect on the log hazard scale of -0.363 (-0.641, -0.085), compared to the unadjusted treatment effect on the log hazard scale of -0.357 (-0.638, -0.076). This translates to an adjusted hazard ratio of treatment to placebo of 0.695 (0.527, 0.918) versus an unadjusted hazard ratio of treatment to placebo of 0.6997 (0.5282, 0.9268). A summary over the 30 imputed data sets of the adjusted treatment effects on log-hazard ratios and its estimated variance is provided in Table 5.5. The average within imputation variance for b is 0.0210, and the average between imputation variance for b is 1.64×10^{-05}, which suggests that we have stable parameter estimates for b over all the imputations.

Table 5.5: Summary of imputed data sets

Variable	Mean	Std Dev	Min	Max
b_*	-0.3635	0.004	-0.3707	-0.3554
V_{b_*}	0.0201	2.40E-05	0.0201	0.0201

If we use the method of combining the $\boldsymbol{d}_p$ and then carrying out the WLS on $\boldsymbol{d}_*$, we obtain a very similar parameter estimate with a larger confidence interval. This is due to the larger values in $\boldsymbol{V}_{\boldsymbol{d}_*}$ than in $\boldsymbol{V}_{\boldsymbol{d}}$, as shown in Equations 5.6 and 5.5, which are due to imputation. The estimated treatment effects and differences in the means of the covariates and its estimated covariance matrix when doing multiple imputation is

$$\boldsymbol{d}_* = \begin{bmatrix} -0.3571 \\ 1.0379 \\ -0.0896 \\ 0.0731 \end{bmatrix}, \boldsymbol{V}_{\boldsymbol{d}_*} = \begin{bmatrix} 0.021924 & 0.014865 & 0.000144 & 0.000462 \\ 0.014865 & 2.610159 & 0.009348 & 0.026854 \\ 0.000144 & 0.009348 & 0.005697 & 0.001568 \\ 0.000462 & 0.026854 & 0.001568 & 0.014194 \end{bmatrix}. \tag{5.6}$$

The adjusted hazard ratio of the treatment compared to placebo is 0.695 (0.527, 0.918), while the unadjusted hazard ratio is 0.700 (0.528, 0.927). The adjusted treatment effect combining the $\boldsymbol{d}_*$ on the log hazard scale is -0.363 (-0.653, -0.074), compared to the adjusted treatment effect combining the $\boldsymbol{b}_p$ on the log hazard scale compared to placebo which is -0.363 (-0.641, -0.085). Not only is the standard error on the adjusted treatment effect larger when combining the $\boldsymbol{d}_p$ than when combining the $\boldsymbol{b}_p$, it is also larger than the standard error of the unadjusted treatment effect (Table 5.9). The estimated treatment effects and differences in the means of the covariates and its estimated covariance matrix when doing multiple imputation is

$$\boldsymbol{d}_* = \begin{bmatrix} -0.3571 \\ 1.0379 \\ -0.0896 \\ 0.0731 \end{bmatrix}, \boldsymbol{V}_{\boldsymbol{d}_*} = \begin{bmatrix} 0.021924 & 0.014865 & 0.000144 & 0.000462 \\ 0.014865 & 2.610159 & 0.009348 & 0.026854 \\ 0.000144 & 0.009348 & 0.005697 & 0.001568 \\ 0.000462 & 0.026854 & 0.001568 & 0.014194 \end{bmatrix}. \tag{5.7}$$

Each of the imputations had the Cox proportional hazards model with all three covariates (age, ln(Breslow+1), and ln(size+1)) fit to it:

$$\lambda_{jk}(t|z_j) = \lambda_{0k}(t)\exp(\boldsymbol{\beta}_k'\boldsymbol{x}_j) \tag{5.8}$$

where $\boldsymbol{x}_j$ is a vector of covariates containing the indicator of age, ln(Breslow+1), and ln(size+1) of the j^{th} patient and $\boldsymbol{\beta}_k$ is an $(M+1)\times 1$ vector of covariate and treatment effects. The results were combined using PROC MIANALYZE. We can then determine how predictive the covariates themselves are from Table 5.6. Ln(Breslow+1) and ln(size+1) are not significant, and age, which is minimally significant, does not have a large effect size. This suggests that adjusting by the covariates may only yield modest gains in efficiency.

Table 5.6: Full model results from Equation 5.8 on imputations of original data on original covariate scale

Variable	Estimate	Standard Error	p-value
Age	0.0116	0.0054	0.0322
ln(Breslow+1)	-0.0715	0.14	0.6065
ln(Size+1)	-0.0461	0.0721	0.5224
Treatment	-0.3934	0.1398	0.0049

5.3.3 MISSING INDICATORS METHOD

Since we have two variables with missingness, we need to create two indicator variables of missingness, one for ln(Breslow+1) ($m = 2$) and one for ln(size+1) ($m = 3$). We can see the difference in the proportion of missingess in Breslow and size between the treatment and placebo are very small (0.024 and 0.000493, respectively as in Table 5.3). The overall means used to impute missing values for the covariates are shown in Table 5.8. The adjusted treatment effect on the log hazard scale compared to placebo is -0.369 (-0.647, -0.092) versus -0.357 (-0.638, -0.076) for the unadjusted method. The estimated standard error of the adjusted treatment parameter is 0.141, compared to the estimated variance of the unadjusted treatment effect of 0.143. The adjusted hazard ratio of the treatment compared to placebo is 0.691 (0.538, 0.912), while the unadjusted hazard ratio is 0.700 (0.528, 0.927). The estimated treatment effects and differences in the means of the covariates and its estimated covariance matrix when using missingness indicators is

$$\boldsymbol{d} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2 \\ \bar{\boldsymbol{m}}_{12} - \bar{\boldsymbol{m}}_{22} \\ \bar{\boldsymbol{m}}_{13} - \bar{\boldsymbol{m}}_{23} \end{pmatrix} = \begin{pmatrix} -0.3571 \\ 1.0379 \\ -0.0899 \\ 0.0575 \\ 0.0244 \\ 0.0005 \end{pmatrix}$$

$$V_d = \begin{pmatrix} 0.0202 & 0.0137 & 0.0001 & 0.0003 & 0.0003 & -0.0003 \\ 0.0137 & 2.4050 & 0.0084 & 0.0206 & 0.0019 & -0.0024 \\ 0.0001 & 0.0084 & 0.0047 & 0.0011 & 0.0000 & -0.0001 \\ 0.0003 & 0.0206 & 0.0011 & 0.0108 & -0.0002 & 0.0000 \\ 0.0003 & 0.0019 & 0.0000 & -0.0002 & 0.0013 & 0.0003 \\ -0.0003 & -0.0024 & -0.0001 & 0.0000 & 0.0003 & 0.0021 \end{pmatrix}. \tag{5.9}$$

Table 5.7: Distribution of the missingness of size by treatment group

	Size		
Drug	Present	Missing	Total
Placebo	114	26	140
Treatment	118	27	145
Total	232	53	285

Table 5.8: Population based minimum and maximum for covariates (# in parenthesis is the un-logged value)

Variable	Min	Max	Mean	Range
Age	0	100	47.12	100
ln(Breslow+1)	0 (1)	4.62 (100)	1.33	4.62
ln(Size+1)	0 (1)	6.91 (1000)	1.23	6.91

5.3.4 COMPARING ALL THREE METHODS

As we can see from Figure 5.2, there is good agreement between the adjusted estimated treatment effect and the corresponding variance estimates for the three methods. Table 5.9 shows that the method using dfbetas has good agreement with the method of imputation. Comparing the methods with the unadjusted estimate, we see that the estimate of the treatment effect is close to the unadjusted estimate, and we have variance reduction.

We also wanted to compare the three methods across bootstrap samples of the original data. We created 100 bootstrap samples with replacement of the 285 observations from our original data set. We applied all three methods to see how well the methods agreed across bootstrap samples. Figures 5.3 and 5.4 are box plots of the results across the bootstrap samples, where the imputation method has 30 imputations. Overall, we see that the methods agree with each other across the 100 bootstrap samples for the estimation of both b and V_b. While there is not much change between the adjusted and unadjusted estimates of b, we see that our method of adjustment using dfbetas, multiple imputation, and indicators, all have smaller estimated variance than the unadjusted estimate of the variance.

We also fit the full Cox proportional hazards model with covariates on the multiple imputation of the bootstraps. We used the same model as before (Equation 5.8). The estimates and standard errors of the covariates (including treatment) were then combined in PROC MIANALYZE. The p-values for the covariates combined over 100 bootstrap samples are shown in Figure 5.5. The covariates age, ln(Breslow+1), and ln(size+1) do not appear to be consistently significant given the wide range of p-values. Treatment, however, appears to be significant in most bootstrap samples. Scatter plots of the estimates of b and its standard error

using our method by dfbetas and multiple imputation show that the methods have good agreement between each other over a range of values (Figures 5.6 and 5.7).

1

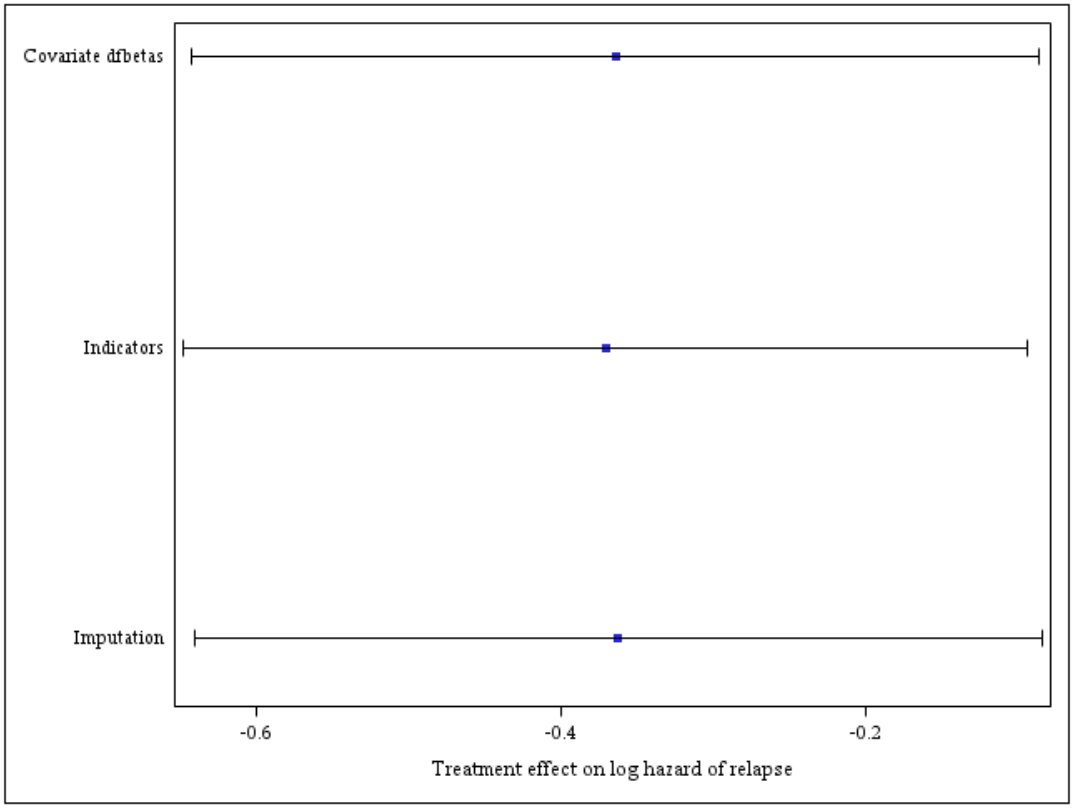

Figure 5.2: Forest plot comparing three methods to handle missingness

5.4 DISCUSSION

This article handled the issue of missing data for baseline covariates in randomization-based non-parametric covariance adjustment of evaluation of treatment effects by using subject specific dfbetas of treatment effects of a model of the treatment nested within type of covariates on a transformation of covariate values on the 0-1 scale. Comparing the method of covariate dfbetas to multiple imputation and single imputation using the variable's sample mean combined with missingness indicator variables suggests good agreement between the three methods.

Table 5.9: Comparing estimates of three methods

Method	Trt Est	Std Err	LCL	UCL	p-value
Covariate dfbetas	-0.363	0.142	-0.641	-0.085	0.011
Imputation combining $\boldsymbol{b}_p$	-0.363	0.142	-0.641	-0.085	0.010
Imputation combining $\boldsymbol{d}_p$	-0.363	0.148	-0.653	-0.074	0.014
Indicators	-0.369	0.141	-0.647	-0.092	0.009
Unadjusted	-0.357	0.143	-0.638	-0.076	0.013

Table 5.10: Summary of b across 100 bootstraps

Type	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
Imputation	-0.820	-0.443	-0.341	-0.248	-0.026
Indicators	-0.825	-0.436	-0.312	-0.242	-0.026
Covariate dfbetas	-0.820	-0.445	-0.339	-0.245	-0.023

The covariate dfbeta method requires the MAR assumption of covariates through the use of GEE for creation of the dfbetas as well as independent random censoring of the survival time. To check this assumption, the methods used in Zhao et al. (2014) may be applied. When unstructured correlation is used for the working correlation of the covariate model, a MAR-like assumption is sufficient. The method is efficient compared to the unadjusted method, which we see in the estimated variance reduction. The method also works well with moderate sample size. The limitation of our method is that it may not work well when there is extensive missing data. In that case, $\boldsymbol{D}\hat{\boldsymbol{\phi}}$ may not be a good estimate of $\boldsymbol{\theta}$.

The multiple imputation method also requires the MAR assumption as the process of combining estimates and standard errors or covariance structures requires it. Large sample sizes may be needed, but it can handle extensive missingness among the baseline covariates.

The method of using missingness indicators does not necessarily require the data to be MAR. It can only handle moderate missing data, and it could potentially be less efficient by reducing the correlation of the covariates with the outcome. Furthermore, when many covariates contain missing data, large sample size may be needed, since more missingness indicators are included in the formulation of the differences of the means of the covariates.

In confirmatory clinical trials, baseline covariates usually are completely observed as some data collection is required to assess whether the patient meets inclusion/exclusion criteria. If baseline covariates are missing, they are typically only minimally missing. Since the covariates are measured at baseline, they typically satisfy the MAR assumption. Missingness among biomarkers, if present, is usually due to malfunctions of laboratory machines and therefore do not depend on the unobserved data. Furthermore, it is sufficient that the

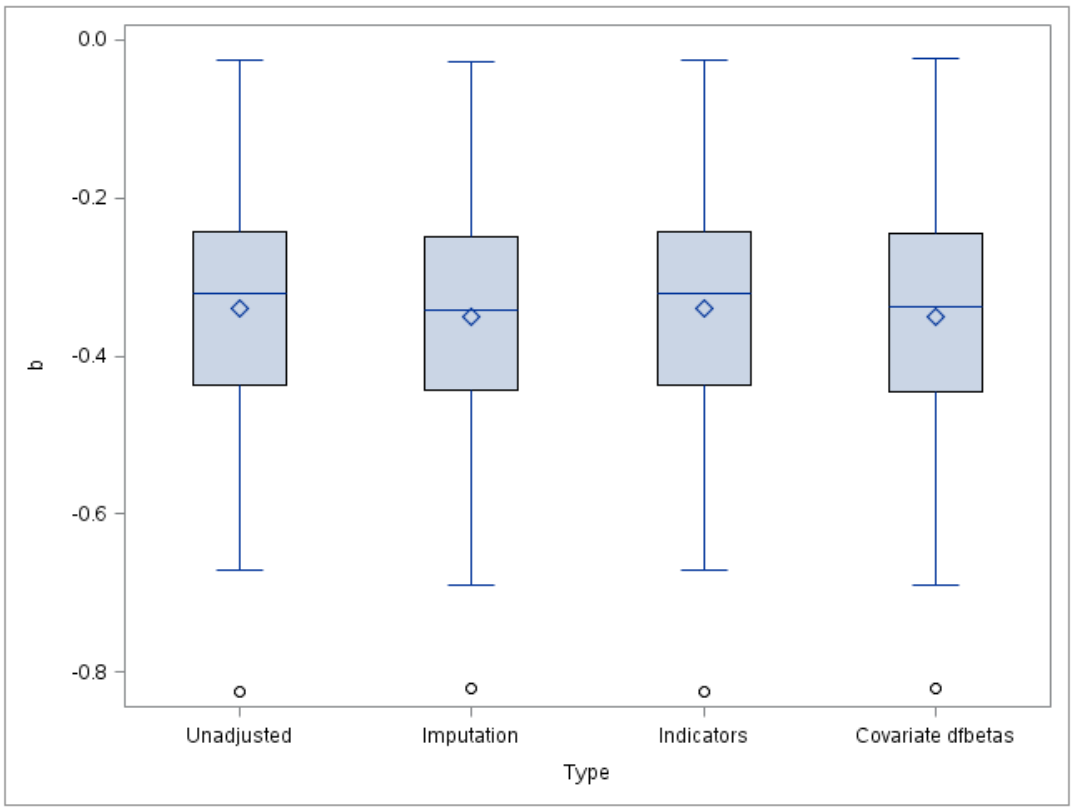

Figure 5.3: Boxplot of b across 100 bootstraps by method

dfbeta method work for minimally missing data (usually considered as less than 20% missing) since a useful covariate would need to be present to contribute to the gains in efficiency in a MAR situation. If there were a useful covariate with many missing values, that could be managed using the method of indicators. The covariate would have a dummy variable created for each category and treat the missing value as the reference. Even if the covariate were MNAR, such as being below the limit of detection for a laboratory machine, the indicator method would suffice.

These three methods each have their strengths and weaknesses. For regulatory clinical trial settings, using the covariate dfbetas may be the most commonly useful method due to its efficiency even in a moderate sample size. Multiple imputation should be used when sample size is large and there is extensive missing

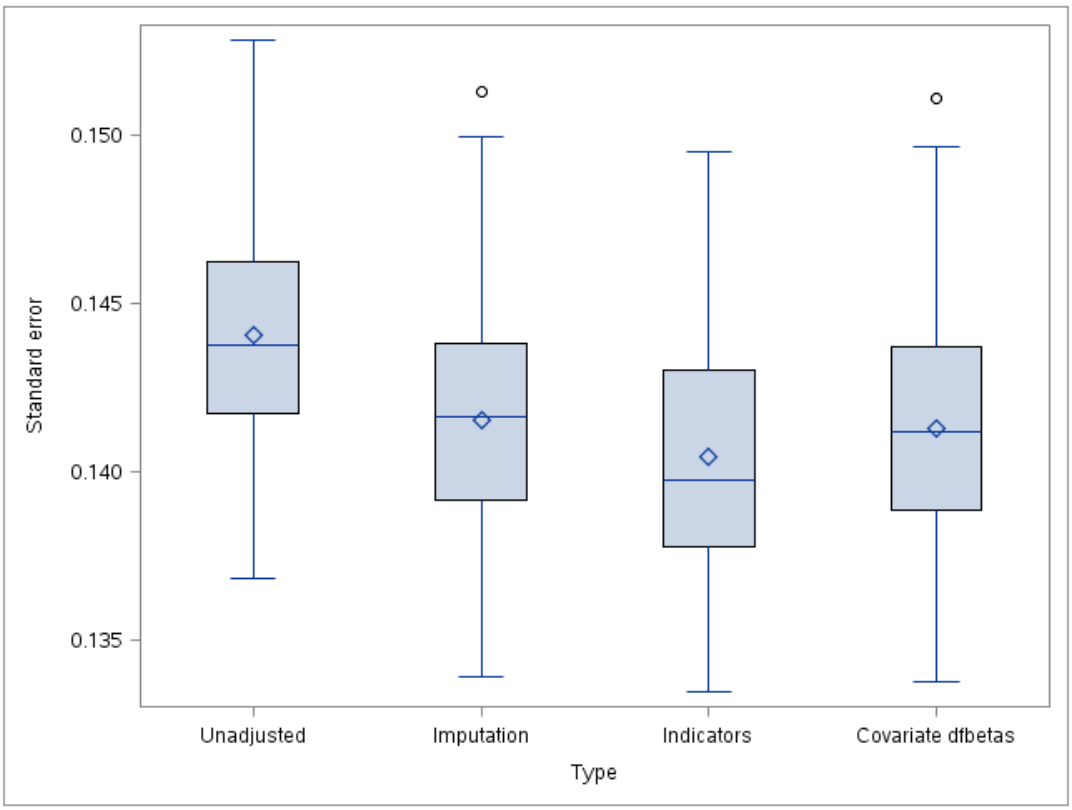

Figure 5.4: Boxplot of estimated standard error of b across 100 bootstraps by method

data, and the method of indicators could be appropriate when the MAR assumption may not be satisfied and the data has a large sample size.

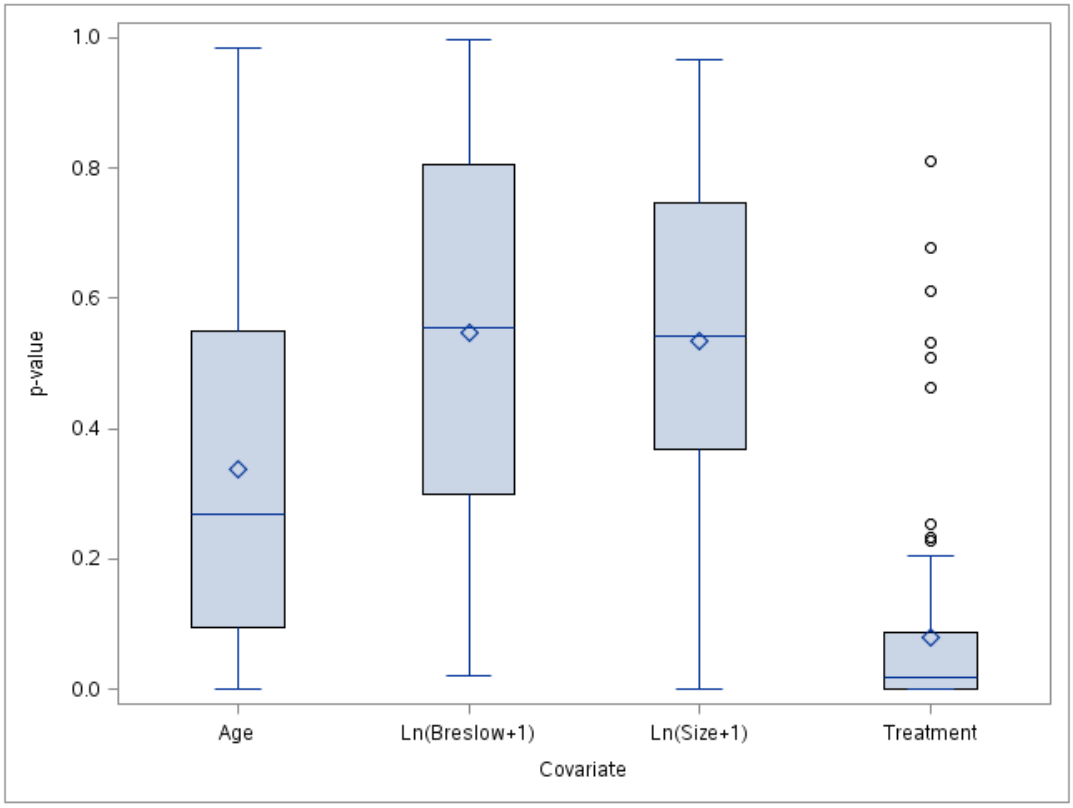

Figure 5.5: Boxplot of p-values of covariates across 100 bootstraps

Table 5.11: Summary of V_b across 100 bootstraps

Type	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
Imputation	0.01794	0.01938	0.02007	0.02069	0.02289
Indicators	0.01782	0.01899	0.01954	0.02046	0.02236
Covariate dfbetas	0.0179	0.01928	0.01994	0.02066	0.02283

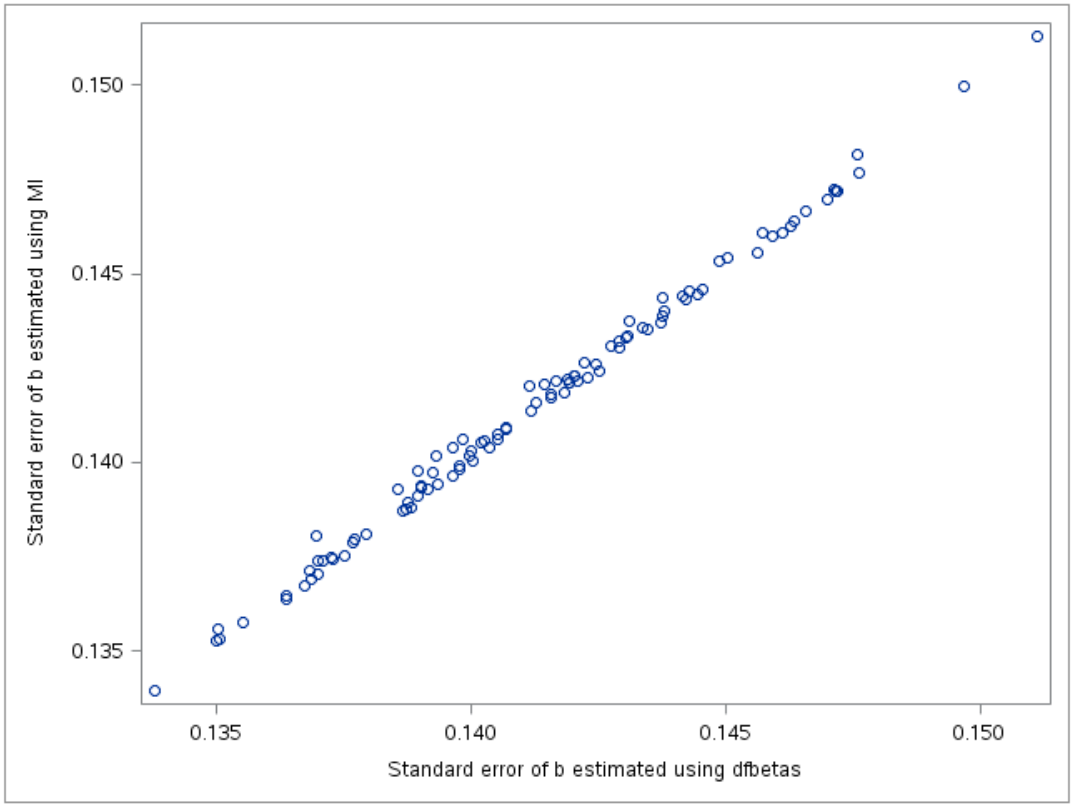

Figure 5.6: Scatter plot comparing standard errors of multiple imputation method with dfbeta method over 100 bootstraps

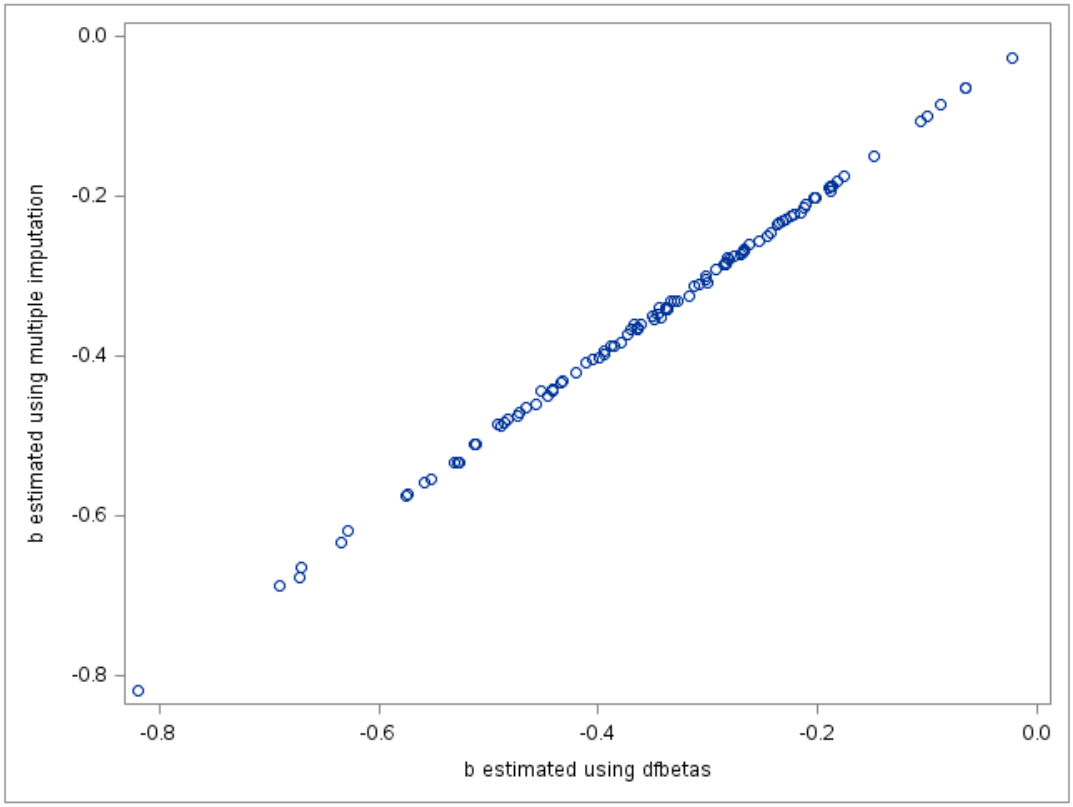

Figure 5.7: Scatter plot comparing treatment estimates of multiple imputation method with dfbeta method over 100 bootstraps

CHAPTER 6: RANDOMIZATION-BASED NON-PARAMETRIC COVARIANCE ADJUSTMENT FOR TIME-TO-EVENT OUTCOMES WITH MULTIPLE TREATMENT GROUPS

6.1 INTRODUCTION

Previous non-parametric covariance adjustment methodology has focused on only one treatment versus reference group, but we extend the method to make multiple treatment comparisons. Some clinical trials have multiple doses for which they wish to assess efficacy. There could be some indication that one dose is moderately effective, but there could be questions as to whether a larger dose could be more effective or if a smaller dose might be as adequaely effective with fewer safety issues.

We follow the randomization based non-parametric covariate adjustment of treatment effect method of Saville and Koch (2013), except for the covariance matrix estimate for the covariates is done via generalized linear models of the transformed covariate values. The dfbetas then take the place of the deviations from the mean. In Section 6.2 we review the method of Saville and Koch (2013), and extend it to the case of missing covariates and multiple treatment groups. We apply our method to a neurological disease dose-finding trial which we describe in Section 6.3 and provide results in Section 6.5. In Section 6.4 we compare our method with that of Hussey (2012). Finally, in Section 6.6 we discuss strengths and limitations of our method.

6.2 METHOD

The general method of randomization-based non-parametric covariance adjustment consists of four steps. We will follow the notation used in Saville and Koch (2013); however, in this case, baseline covariates are allowed to have missingness and more than one treatment will be evaluated. First we model the unadjusted treatment effect for outcomes. In our randomized clincial trial, we have survival data for K outcomes evaluating T different treatments. We will compare treatments $1, 2, \ldots, T-1$ to treatment T. The main difference between our method and that of Saville and Koch (2013) is that we have $T-1$ times more treatment parameters to allow for pairwise comparisons with group T (as well as for contrasts among groups), and we have $T-1$

times more differences in the means of the covariates to allow for comparisons of different treatment groups' means to the means of covariates of group T (or for contrasts among groups). Let M be the number of covariates for adjustment. We define the vector $\boldsymbol{d}$ as the vector of the $K(T-1)$ treatment effects by outcomes stacked on top of the $M(T-1)$ estimates for the differences of the means of the covariates of the treatments from treatmeng group T. The differences of the means of the covariates are estimated as some of the baseline covariates may be missing. Therefore $\boldsymbol{d}$ is a $(K+M)(T-1) \times 1$ column vector.

In the second step, we want to estimate the covariance matrix of $\boldsymbol{d}$ because we will use it in a weighted least squares regression which occurs in the third step. The covariance matrix of $\boldsymbol{d}$ is comprised of three different covariance matrices: the covariance matrix of the treatment effects for the K outcomes, the covariance matrix of the estimates for the differences of the means of the covariates for the $T-1$ comparisons, and the covariance matrix of the treatment effect for outcomes with those for the comparisons of the covariates. We can estimate the covariance matrix of the treatment parameters for outcomes by calculating the sum of cross products of the dfbetas associated with the treatment effects (Wei et al., 1989). Next we calculate the covariance matrix of the estimated differences of the means of the covariates and the covariance matrix of the treatment effects for outcomes with the estimated differences in the means of the covariates using dfbetas of a generalized linear model of the covariates transformed onto a 0-1 scale which we describe in some detail later. The covariance matrix of $\boldsymbol{d}$, $\boldsymbol{V_d}$, is a $(K+M)(T-1) \times (K+M)(T-1)$ matrix.

Third we force the estimated differences in means of covariates to 0 through weighted least squares (WLS) regression. We regress $\boldsymbol{d}$, which contains both estimates of treatment effect for outcomes and estimates of the differences in the means of the covariates onto the space of only treatment parameters. We use WLS with the weights based on $\boldsymbol{V_d^{-1}}$. Forcing the estimated differences in means of the covariates to 0 through WLS is appropriate due to randomization since these differences for covariates are random and the populations receiving the T treatments are the same for them. From WLS methods, we obtain both the estimator for the adjusted treatment effects ($\boldsymbol{b}$) and and a consistent estimator for their covariance matrix ($\boldsymbol{V_b}$).

The fourth step is to do hypothesis testing using the approximately multivariate normal distribution of $\boldsymbol{b}$ (via sufficiently supportive sample size), with the corresponding expected value and covariance matrix estimated in the WLS in Step 3. We can test hypotheses using traditional contrast matrices and calculate estimates of parameters from linear combinations of $\boldsymbol{b}$.

To carry out these four steps we must first explain some notation. Let T_{ijk} be the survival time for the k^{th} outcome for the j^{th} patient on the i^{th} treatment. Let $\boldsymbol{z}_j = (z_{j1}, \ldots, z_{jT-1})'$ be a vector of the j^{th}'s subject's

indicator of treatment; z_{ji} is 1 if the patient receives treatment i, and 0 otherwise. Thus the position of the 1 indicates which treatment patient j receives and if patient j receives treatment T then $\boldsymbol{z}_j$ is a vector of 0's. For each outcome k, we will use a marginal Cox proportional hazards model to estimate the unadjusted treatment effects. Our model for the estimate of treatment effects is

$$\lambda_{jk}(t|\boldsymbol{z}_j) = \lambda_{0k}(t)\exp(\boldsymbol{z}_j'\boldsymbol{\beta}_k)$$

where $\lambda_{0k}(t)$ is the baseline hazard for outcome k given the subject has survived until time t. Let $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{(T-1)k})'$ be the $(T-1)\times 1$ vector of the $T-1$ treatment effects on the log hazards ratios for the k^{th} outcome, and let β_{ik} be the treatment effect for the k^{th} outcome for the comparison between the i^{th} and T^{th} treatment. Thus, we have pairwise comparisons of different treatment groups with the reference group T for the k^{th} outcome.

To express our model for the K outcomes and the $T-1$ treatment effects, let $\boldsymbol{\lambda}_j(t|\boldsymbol{z}_j) = (\lambda_{j1}(t|\boldsymbol{z}_j), \ldots, \lambda_{jK}(t|\boldsymbol{z}_j))'$ be the vector of the K hazards for patient j, and let $\boldsymbol{\lambda}_0(t|\boldsymbol{z}_j) = (\lambda_{01}(t), \ldots, \lambda_{0K}(t))'$ be the vector of baseline hazards for the K outcomes. Therefore, our model for the unadjusted treatment effects for the K outcomes and T treatments is

$$\boldsymbol{\lambda}_j(t|\boldsymbol{z}_j) = \lambda_0(t)\exp\{[\boldsymbol{z_j}' \otimes \boldsymbol{I}_K]\boldsymbol{\beta}\} \tag{6.1}$$

where $\boldsymbol{\beta} = (\beta_{11}, \ldots, \beta_{1K}, \ldots, \beta_{(T-1)1}, \ldots, \beta_{(T-1)K})'$ is the $K(T-1)\times 1$ column vector of all the treatment parameters for all outcomes, $\boldsymbol{I}_K$ is the $K \times K$ identity matrix, and $\otimes$ is the kronecker product. Thus $\boldsymbol{z_j}' \otimes \boldsymbol{I}_K$ is a matrix of $1 \times (T-1)$ diagonal blocks where the i^{th} block is $z_{ji} \times \boldsymbol{I}_K$. $\boldsymbol{\beta}$ is arranged so that outcomes are nested within treatment. We assume that we have independent random censoring, which satisfies a MAR-like assumption.

We next need notation to refer to the treatment effects dfbetas. Let i be the index of treatment group, n_i be the number of patients receiving treatment i, and $\sum_{i=1}^{T} n_i = n$, where n is the total number of patients in the study. Without loss of generality, we will index the patients from 1 to n where patients will be grouped by treatment. Patients 1 to n_1 are in treatment 1, and patients indexed from $([\sum_{t=1}^{i-1} n_t] + 1)$ to $(\sum_{t=1}^{i} n_t)$, are in

treatment group i, where i ranges from 1 to T. $\boldsymbol{R} = (\boldsymbol{R}_1', \boldsymbol{R}_2', \ldots, \boldsymbol{R}_T')'$is the $n \times K(T-1)$ matrix where

$$\boldsymbol{R}_1 = \begin{bmatrix} r_{11(1)} & r_{12(1)} & \cdots & r_{1K(1)} & \cdots & r_{\{T-1\}1(1)} & \cdots & r_{\{T-1\}K(1)} \\ r_{11(2)} & r_{12(2)} & \cdots & r_{1K(2)} & \cdots & r_{\{T-1\}1(2)} & \cdots & r_{\{T-1\}K(2)} \\ \vdots & & & & & & & \vdots \\ r_{11(n_1)} & r_{12(n_1)} & \cdots & r_{1K(n_1)} & \cdots & r_{\{T-1\}1(n_1)} & \cdots & r_{\{T-1\}K(n_1)} \end{bmatrix}. \tag{6.2}$$

In Equation 6.2, $\boldsymbol{R}_1$ is the $n_1 \times K(T-1)$ matrix of dfbeta residuals for treatment 1 obtained from the unadjusted Cox proportional hazards model with $T-1$ parameters for each of the K events. Each row represents a subject's effect on $\boldsymbol{\beta}$ from Equation 6.1. Because of the way we have grouped our patients by treatment, we can express the $n_i \times K(T-1)$ matrix of dfbetas for patients in treatment $i \in (1, \ldots, T)$ in general terms as shown in Equation 6.3:

$$\boldsymbol{R}_i = \begin{bmatrix} r_{11([\sum_{t=1}^{i-1} n_t]+1)} & \cdots & r_{1K([\sum_{t=1}^{i-1} n_t]+1)} & \cdots & r_{\{T-1\}1([\sum_{t=1}^{i-1} n_t]+1)} & \cdots & r_{\{T-1\}K([\sum_{t=1}^{i-1} n_t]+1)} \\ r_{11([\sum_{t=1}^{i-1} n_t]+2)} & \cdots & r_{1K([\sum_{t=1}^{i-1} n_t]+2)} & \cdots & r_{\{T-1\}1([\sum_{t=1}^{i-1} n_t]+2)} & \cdots & r_{\{T-1\}K([\sum_{t=1}^{i-1} n_t]+2)} \\ \vdots & & & & & & \vdots \\ r_{11(\sum_{t=1}^{i} n_t)} & \cdots & r_{1K(\sum_{t=1}^{i} n_t)} & \cdots & r_{\{T-1\}1(\sum_{t=1}^{i} n_t)} & \cdots & r_{\{T-1\}K(\sum_{t=1}^{i} n_t)} \end{bmatrix}. \tag{6.3}$$

In Equation 6.3 $r_{ik(j)} \hat{=} \hat{\beta}_{ik(j)} - \hat{\beta_{ik}}$, where $\hat{\beta}_{ik(j)}$ is the estimate of β_{ik} when subject j is deleted. In other words $r_{ik(j)}$ is the dfbeta of the j^{th} patient on parameter β_{ik}. The estimated covariance matrix $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$ for $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{R}'\boldsymbol{R}$.

The estimated differences in the means of the covariates is expressed as $\hat{\boldsymbol{\theta}}$, and when all covariates have no missing data, $\hat{\boldsymbol{\theta}} = ((\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_T)', \ldots, (\bar{\boldsymbol{x}}_{T-1} - \bar{\boldsymbol{x}}_T)')'$, where $\bar{\boldsymbol{x}}_i = (\bar{x}_{i1}, \ldots, \bar{x}_{iM})'$ is the vector of M baseline covariate means for treatment i with $\bar{x}_{im}$ being the mean of covariate m for patients in treatment group i.

To estimate the covariance matrix of $\boldsymbol{d} = (\hat{\boldsymbol{\beta}}', \hat{\theta}')'$ we need to calculate $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$, $\boldsymbol{V}_{\hat{\boldsymbol{\beta}},\hat{\theta}}$, and $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}}$ where $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$, $\boldsymbol{V}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\theta}}$, and $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{R}'\boldsymbol{R}$ is the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$. When all covariates have no missing data and only one treatment group is evaluated relative to placebo, we used $\boldsymbol{C}_i = \frac{\boldsymbol{X}_i - \mathbf{1}_{n_i}\bar{\boldsymbol{x}}_i'}{\sqrt{n_i(n_i-1)}}$ where $\boldsymbol{X}_i$ is the $n_i \times M$ matrix of observed values for patients in treatment group i. Each subject contributes one row to $\boldsymbol{X}_i$, and the columns are the M

covariates. $\mathbf{1}_{n_i}$ is the $n_i \times 1$ column vector of 1's. To calculate $\boldsymbol{V}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}}$ and $\boldsymbol{V}_{\hat{\theta}}$ in the presence of missing data we need to model the effect of treatment group on the mean of each of the M covariates.

In order to obatin an estimate of the differences of the means of the covariates for the $T-1$ treatment group comparisons in the presence of missing data we use generalized linear models with repeated measures within subjects. We transform all the covariates onto a 0-1 scale by subtracting a plausible minimum possible value from the covariate value and dividing the difference by a plausible range of the covariate (Equation 6.4). We then use generalized estimating equations to estimate the effect of treatment nested within covariate type on the transformed covariate values in a model with the identity link, normal distribution, and unstructured working correlation matrix, and then using dfbetas (Preisser and Qaqish, 1996) to obtain an estimated covariance matrix of the differences in the means of the covariates. The covariate dfbetas then replace the $\boldsymbol{C}_i$ in the complete data method, with some transformation to reverse the scaling to the 0-1 scale.

Let x_{imj} be the value of covariate m for patient j in treatment group i. Let a_m be a plausible minimum possible population value for covariate m and p_m be a plausible maximum value. Our covariate transformation is defined as

$$f(x_{imj}) = \frac{x_{imj} - a_m}{p_m - a_m}. \tag{6.4}$$

This transformation of covariates to the 0-1 scale allows generalized estimating equations (GEE) in PROC GENMOD to be more computationally straightforward, and if x_{imj} is missing then $f(x_{imj})$ is missing. We want to model the value of the $M \times 1$ vector of covariates for patients in treatment group i ($\boldsymbol{x}_{im}$) using the following marginal mean model:

$$\begin{aligned} E\left[f(\boldsymbol{x}_{im})\right] =& \eta_1(m=1) + \ldots + \eta_M(m=M) + \\ & \phi_{11}(i=1)(m=1) + \ldots + \phi_{1M}(i=1)(m=M) + \ldots \\ & \phi_{(T-1)1}(i=T-1)(m=1) + \ldots + \phi_{(T-1)M}(i=T-1)(m=M). \end{aligned} \tag{6.5}$$

In Equation 6.5, η_m is the intercept for estimating a transformation of covariate m for a subject in treatment group T and ϕ_{im} is the effect of being in treatment group i instead of T on the estimate of the transformation of covariate m. We use PROC GENMOD with an unstructured covariance matrix to obtain the dfbetas

associated with $\boldsymbol{\phi} = (\phi_{11}, \ldots, \phi_{(T-1)M})$, and we use the dfbetas in the matrix $\boldsymbol{G} = (\boldsymbol{G}_1', \boldsymbol{G}_2', \ldots, \boldsymbol{G}_T')'$ which is an $n \times M(T-1)$ matrix.

Let $\boldsymbol{G}_1$ be the $n_1 \times M(T-1)$ matrix in Equation 6.6 for dfbetas for $\boldsymbol{\phi}$ for the subjects in treatment group 1 so that

$$\boldsymbol{G}_1 = \begin{bmatrix} g_{11(1)} & g_{12(1)} & \cdots & g_{1M(1)} & \cdots & g_{\{T-1\}1(1)} & \cdots & g_{\{T-1\}M(1)} \\ g_{11(2)} & g_{12(2)} & \cdots & g_{1M(2)} & \cdots & g_{\{T-1\}1(2)} & \cdots & g_{\{T-1\}M(2)} \\ \vdots & & & & & & & \vdots \\ g_{11(n_1)} & g_{12(n_1)} & \cdots & g_{1M(n_1)} & \cdots & g_{\{T-1\}1(n_1)} & \cdots & g_{\{T-1\}M(n_1)} \end{bmatrix}. \tag{6.6}$$

More generally, $\boldsymbol{G}_i$ is the $n_i \times (T-1)M$ matrix in Equation 6.7 for dfbetas for patients in treatment group $i \in (1, \ldots, T)$ and is written as

$$\boldsymbol{G}_i = \begin{bmatrix} g_{11([\sum_{t=1}^{i-1} n_t]+1)} & \cdots & g_{1M([\sum_{t=1}^{i-1} n_t]+1)} & \cdots & g_{\{T-1\}1([\sum_{t=1}^{i-1} n_t]+1)} & \cdots & g_{\{T-1\}M([\sum_{t=1}^{i-1} n_t]+1)} \\ g_{11([\sum_{t=1}^{i-1} n_t]+2)} & \cdots & g_{1M([\sum_{t=1}^{i-1} n_t]+2)} & \cdots & g_{\{T-1\}1([\sum_{t=1}^{i-1} n_t]+2)} & \cdots & g_{\{T-1\}M([\sum_{t=1}^{i-1} n_t]+2)} \\ \vdots & & & & & & \vdots \\ g_{11(\sum_{t=1}^{i} n_t)} & \cdots & g_{1M(\sum_{t=1}^{i} n_t)} & \cdots & g_{\{T-1\}1(\sum_{t=1}^{i} n_t)} & \cdots & g_{\{T-1\}M(\sum_{t=1}^{i} n_t)} \end{bmatrix} \tag{6.7}$$

and $g_{im} \approx \hat{\phi}_{im(j)} - \hat{\phi}_{im}$, which is the dfbeta for patient j on $\hat{\phi}_{im}$.

To obtain the estimated differences in means of the covariates, we compute $\hat{\boldsymbol{\theta}} = \boldsymbol{D}\hat{\boldsymbol{\phi}}$ where $\boldsymbol{D} = \boldsymbol{I}_{T-1} \otimes \boldsymbol{diag}(\boldsymbol{p} - \boldsymbol{a})$ is a block diagonal matrix of dimension $M(T-1) \times M(T-1)$. $\boldsymbol{I}_{T-1}$ is the $(T-1) \times (T-1)$ identity matrix, and $\boldsymbol{diag}(\boldsymbol{p} - \boldsymbol{a})$ is a diagonal matrix for which the diagonal is the vector $(\boldsymbol{p} - \boldsymbol{a})$; $\boldsymbol{p} = (p_1, \ldots, p_M)$ and $\boldsymbol{a} = (a_1, \ldots, a_M)$ so $(\boldsymbol{p} - \boldsymbol{a})$ is the $M \times 1$ vector of ranges for the M covariates. Thus $\boldsymbol{D}$ has $(T-1)$ diagonal blocks which are the $M \times M$ diagonal matrices with the lengths of the ranges of the M covariates along the diagonal. The estimated covariance matrix $\boldsymbol{V}_{\hat{\boldsymbol{\theta}}}$ of $\hat{\boldsymbol{\theta}}$ is shown in

Equation 6.8

$$\begin{aligned} \boldsymbol{V}_{\hat{\boldsymbol{\theta}}} &= \text{Cov}(\boldsymbol{D}\hat{\boldsymbol{\phi}}) \\ &= \boldsymbol{D}\text{Cov}(\hat{\boldsymbol{\phi}})\boldsymbol{D} \\ &= \boldsymbol{D}(\boldsymbol{G}'\boldsymbol{G})\boldsymbol{D} \end{aligned} \tag{6.8}$$

Also,

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}} = \boldsymbol{R}'\boldsymbol{G}\boldsymbol{D}.$$

The main feature of this method is the difference in the calculation of elements of $\boldsymbol{V}_d$ as in Equation 6.9.

$$\boldsymbol{V}_{\boldsymbol{d}} = \begin{bmatrix} \boldsymbol{R}'\boldsymbol{R} & \boldsymbol{R}'\boldsymbol{G}\boldsymbol{D} \\ \boldsymbol{D}\boldsymbol{G}'\boldsymbol{R} & \boldsymbol{D}\boldsymbol{G}'\boldsymbol{G}\boldsymbol{D} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{T} \boldsymbol{R}_i'\boldsymbol{R}_i & \sum_{i=1}^{T} \boldsymbol{R}_i'\boldsymbol{G}_i\boldsymbol{D} \\ \sum_{i=1}^{T} \boldsymbol{D}\boldsymbol{G}_i'\boldsymbol{R}_i & \sum_{i=1}^{T} \boldsymbol{D}\boldsymbol{G}_i'\boldsymbol{G}_i\boldsymbol{D} \end{bmatrix}. \tag{6.9}$$

The calculation for the covariance matrix has a very similar form to the work developed by Hussey (2012). The method then proceeds with WLS being performed on $\boldsymbol{d}$ as shown in Equation 6.10

$$\boldsymbol{d} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\theta}} \end{pmatrix} \hat{=} \begin{bmatrix} \boldsymbol{I}_{K(T-1)} \\ \boldsymbol{0}_{M(T-1)\times K(T-1)} \end{bmatrix} \boldsymbol{b} = \boldsymbol{X}\boldsymbol{b} \tag{6.10}$$

where $\boldsymbol{I}_{(T-1)K}$ is the $(T-1)K \times (T-1)K$ identity matrix and $\boldsymbol{0}_{(T-1)M\times(T-1)K}$ is a $(T-1)M \times (T-1)K$ matrix of 0's. Thus, with WLS, $\boldsymbol{b}$ is obtained as shown in Equation 6.11. The WLS estimate of $\boldsymbol{b}$ is

$$\begin{aligned} \boldsymbol{b} =& (\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{d}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{d}}^{-1}\boldsymbol{d} \\ =& \hat{\boldsymbol{\beta}} - (\boldsymbol{R}'\boldsymbol{G}\boldsymbol{D})(\boldsymbol{D}\boldsymbol{G}'\boldsymbol{G}\boldsymbol{D})^{-1}\boldsymbol{D}\hat{\boldsymbol{\phi}} \\ =& \hat{\boldsymbol{\beta}} - (\boldsymbol{R}'\boldsymbol{G})(\boldsymbol{G}'\boldsymbol{G})^{-1}\hat{\boldsymbol{\phi}}. \end{aligned} \tag{6.11}$$

A consistent estimator for $\boldsymbol{V_b}$ for the covariance matrix of $\boldsymbol{b}$ is

$$\boldsymbol{V_b} = (\boldsymbol{X}'\boldsymbol{V_d}^{-1}\boldsymbol{X})^{-1} = \boldsymbol{R}'\boldsymbol{R} - \boldsymbol{R}'\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{R} \tag{6.12}$$

6.3 NEUROLOGICAL DISORDER DATA

Our data come from a double-blind randomized clinical trial evaluating the effect of multiple doses of riluzole on time to disease progression of Amyotrophic Lateral Sclerosis (ALS). In the trial, disease progression (an event) was defined as death, requirement of tracheostomy, or intubation with artificial ventilation leading to tracheostomy. In total, there were 959 patients randomized and 431 events. The 3 doses of 50 mg, 100 mg, and 200 mg of riluzole were compared to placeo (0 mg). Using the notation in the previous method, $K = 1$ and $T = 4$. Patients in this analysis were followed for 18 months and have a maximum survival time of 549 days. The trial is described in more detail in Lacomblez et al. (1996) and has been analyzed using different methods of non-parametric covariance adjustment by Hussey (2012); Tangen and Koch (2001). Previous studies have focused on 100 mg as the main dose of interest.

The specification for covariates was the same as in Hussey (2012) for methods that he developed for time-to-event data with T treatment groups. In this regard, he used the same covariates he had identified using stratified conditional step-wise logistic regression for 2 treatment groups (100 and 200 mg dose versus 0 mg dose with the 50 mg dose subjects omitted). In the study, patients were stratified by three regions and two sites of onset for six total strata. The three regions were France and Belgium, other European countries, and North America and the two onset sites were bulbar versus limb. It also evaluated covariates that are associated with patients surviving past 12 months, with a conditional stepwise logistic regression for the 12 strata by pooled treatment groups with an entry criterion of $p < 0.005$. Conditioning on strata by pooled treatment enables a selection of covariates that are predictive of having an event by 12 months independent of treatment. There were 6 patients who were censored before 12 months but who did not suffer the event. These patients were managed as not experiencing the outcome. The covariates chosen were age (years), disease duration (years), weight (kgs), muscle testing score (22 items on a 5 point scale), visual analogue scale fatigue score, and vital capacity ratio at inclusion. For our method, the ranges we chose for the covariates are shown in Table 6.1.

Table 6.1: Ranges used for transformation of covariates

Variable	Min	Max
Age (yrs)	0	100
Disease Duration (yrs)	0	15
Weight (kgs)	0	500
Muscle testing (22 items on 5 point scale)	0	110
Visual Analogue Scale Fatigue	0	100
Vital Capacity Ratio	0	200

Table 6.2: Imputed means for varaiables by disease onset site

Site onset	Variable	Mean
Limb	Age at inclusion (years)	55.73
	Disease duration (years)	1.97
	Weight (kgs)	68.89
	Muscle Testing Total Score	84.35
	Visual Analogue Scale Fatigue (inclusion)	48.43
	Vital Capacity Ratio (inclusion)	90.53
Bulbar	Age at inclusion (years)	58.89
	Disease duration (years)	1.38
	Weight (kgs)	65.15
	Muscle Testing Total Score	97.91
	Visual Analogue Scale Fatigue (inclusion)	37.28
	Vital Capacity Ratio (inclusion)	83

The overall missingness percentage for the selected covariates was 1.53%, and the mean values of the covariates given their site of disease onset were used for imputation (Table 6.2) in analyses reported by Hussey (2012). The distribution of missingness by treatment group is shown in Table 6.3 and suggests that the number of missing observations is roughly the same for all 4 treatment groups. We will analyze the data with both the 1.53% missingness and with the imputed values so that we can compare this method with Hussey (2012). For all 6 covariates, the largest absolute difference in the mean of any of the 4 treatment groups for the imputed data set versus the non-imputed data set was 0.17 for visual analogue scale fatigue score. Given the means of the variables, we do not expect to see large differences between the imputed and non-imputed data sets.

Table 6.3: Distribution of missingness by treatment group

Variable	50 mg	100 mg	200 mg	Placebo	Overall
Age at inclusion (years)	0	0	0	0	0
Disease duration (years)	0	0	0	0	0
Weight (kgs)	9	10	7	8	34
Muscle Testing Total Score	4	1	3	0	8
Visual Analogue Scale Fatigue	11	10	11	10	42
Vital Capacity Ratio	2	2	0	0	4

6.4 COMPARISON OF THE PROPOSED METHOD WITH HUSSEY (2012)

Hussey (2012) investigated non-parametric randomization based methods for multiple treatment effects in multivariate time-to-event outcomes with no missing data for the baseline covariates. When referring to components calculated using the method of Hussey (2012), we will use a subscript of H for differentiation.

The vector of treatment effects and baseline covariates is

$$\boldsymbol{d}_H = \begin{pmatrix} \hat{\beta}_1^{(1)} \\ \vdots \\ \hat{\beta}_K^{(1)} \\ \vdots \\ \hat{\beta}_1^{(T-1)} \\ \vdots \\ \hat{\beta}_K^{(T-1)} \\ \bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_T \\ \bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_T \\ \vdots \\ \bar{\boldsymbol{x}}_{T-1} - \bar{\boldsymbol{x}}_T \end{pmatrix} \tag{6.13}$$

where $\beta_k^{(i)}$ is the treatment effect of being in group i on outcome k and is equivalent to our definition of β_{ik}; $\bar{\boldsymbol{x}}_i$ is defined the same way we have defined it previously. We let $\boldsymbol{\beta}_H = (\beta_1,^{(1)}, \ldots, \beta_K^{(1)}, \ldots, \beta_1^{(T-1)}, \ldots, \beta_K^{(T-1)})$.

The covariance matrix of $\hat{\boldsymbol{\beta}}_H$ is

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}_H} = \begin{bmatrix} \boldsymbol{R}^{(1)\prime}\boldsymbol{R}^{(1)} & \boldsymbol{R}^{(1)\prime}\boldsymbol{R}^{(2)} & \dots & \boldsymbol{R}^{(1)\prime}\boldsymbol{R}^{(T-1)} \\ & \boldsymbol{R}^{(2)\prime}\boldsymbol{R}^{(2)} & \dots & \boldsymbol{R}^{(2)\prime}\boldsymbol{R}^{(T-1)} \\ & & \ddots & \vdots \\ & & & \boldsymbol{R}^{(T-1)\prime}\boldsymbol{R}^{(T-1)} \end{bmatrix} \tag{6.14}$$

where $\boldsymbol{R}^{(u)} = (\boldsymbol{R}_i^{(u)}, \dots, \boldsymbol{R}_{T-1}^{(u)})$ with $\boldsymbol{R}_i^{(u)}$ is the $n_i \times K$ matrix of dfbeta residuals for K events for the t^{th} treatment comparison for the n_i subjects from group i. Furthermore, $\boldsymbol{R}^{(u)\prime}\boldsymbol{R}^{(s)} = \sum_{i=1}^{T} \boldsymbol{R}_i^{(u)\prime}\boldsymbol{R}_i^{(s)}$ is the robust estimate of the $K \times K$ covariance matrix of $\hat{\boldsymbol{\beta}}^{(u)}$ and $\hat{\boldsymbol{\beta}}^{(s)}$ where $\hat{\boldsymbol{\beta}}^{(u)} = (\hat{\beta}_1^{(u)}, \dots, \hat{\beta}_K^{(u)})$. Since $\boldsymbol{R}^{(u)\prime}\boldsymbol{R}^{(s)} = \sum_{i=1}^{T} \boldsymbol{R}_i^{(u)\prime}\boldsymbol{R}_i^{(s)}$ is equivalent to the subset of $\boldsymbol{R'R}$ composed of rows $(u(K-1)+1, \dots, u(K-1)+(T-1))$ and columns $(s(K-1)+1, \dots, s(K-1)+(T-1))$, then $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}_H}$ is equivalent to $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$

Let $\bar{\boldsymbol{x}}_H = ((\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_T)', \dots, (\bar{\boldsymbol{x}}_{T-1} - \bar{\boldsymbol{x}}_T)')'$ be the $M * (T-1) \times 1$ vector of M differences of means of the covariates for the $T-1$ treatment comparisons. The covariance matrix for $\bar{\boldsymbol{x}}_H$ is

$$\boldsymbol{V}_{\bar{\boldsymbol{x}}_H} = \begin{bmatrix} \boldsymbol{V}_{\bar{\boldsymbol{x}}_1} + \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} & \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} & \dots & \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} \\ & \boldsymbol{V}_{\bar{\boldsymbol{x}}_2} + \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} & \dots & \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} \\ & & & \boldsymbol{V}_{\bar{\boldsymbol{x}}_{T-1}} + \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} \end{bmatrix} \tag{6.15}$$

which is estimated by the covariance matrix $\boldsymbol{V}_{\hat{\theta}} = \boldsymbol{DG'GD}$. Note that $\boldsymbol{DG'GD} = \sum_{i=1}^{T} \boldsymbol{DG_i'G_iD}$. Since $\boldsymbol{G}_i$ contains only the patients in treatment group i, then for $i = 1, \dots, T-1$ the non-zero columns are $\{1 + M(i-1), \dots, M + M(i-1)\}$ and the rest of the columns are effectively 0. Therefore, $\boldsymbol{DG_i'G_iD} \hat{=} \boldsymbol{diag}(\boldsymbol{e}_{T-1}(i)) \otimes \boldsymbol{V}_{\bar{\boldsymbol{x}}_i}$, where $\boldsymbol{diag}(\boldsymbol{e}_{T-1}(i))$ is a diagonal matrix for which the diagonal is $\boldsymbol{e}_{T-1}(i)$, a $(T-1) \times 1$ vector of 0's with a 1 in element i, and $\hat{=}$ means estimates. For $i = T$, the patients are in the placebo group, so deleting them adjusts the estimates of $\boldsymbol{\eta}$ which affects all values of $\boldsymbol{\phi}$. Therefore, $\boldsymbol{G}_T$ is non-zero in all columns. The effect of deleting a patient in the reference group affects ϕ_{im} the same for $i = 1, \dots, T-1$ when m is held constant, so $\boldsymbol{DG_T'G_TD}$ has block matrix structure made of one matrix

repeated for all blocks. Therefore $\boldsymbol{DG_T'G_TD} \hat{=} \begin{bmatrix} \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} & \dots & \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} \\ \vdots & \vdots & \vdots \\ \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} & \dots & \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} \end{bmatrix}$, and the covariance matrix of

$$\begin{aligned} \boldsymbol{DG'GD} &= \sum_{i=1}^{T} \boldsymbol{DG_i'G_iD} \\ &\approx \begin{bmatrix} \boldsymbol{V}_{\bar{\boldsymbol{x}}_1} & 0 & \dots & 0 \\ 0 & \boldsymbol{V}_{\bar{\boldsymbol{x}}_2} & \dots & 0 \\ & \ddots & & \\ & & & \boldsymbol{V}_{\bar{\boldsymbol{x}}_{T-1}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} & \dots & \dots & \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} & \dots & \dots & \boldsymbol{V}_{\bar{\boldsymbol{x}}_T} \end{bmatrix} = \boldsymbol{V}_{\bar{\boldsymbol{x}}_H} \end{aligned} \tag{6.16}$$

agrees with the covariance matrix of Hussey (2012).

Let $\boldsymbol{R}_{iM} = (\boldsymbol{R}_i^{(1)}, \dots, \boldsymbol{R}_i^{(T-1)})$ be the $n_i \times K * (T-1)$ matrix of dfbeta residuals for the K events and $T-1$ treatment comparisons for the n_i patients on treatment i. The covariance matrix of $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}_H, \bar{\boldsymbol{x}}_i} - \boldsymbol{V}_{\hat{\boldsymbol{\beta}}_H, \bar{\boldsymbol{x}}_T} = \boldsymbol{R}_{iM}'\boldsymbol{C}_i - \boldsymbol{R}_{TM}'\boldsymbol{C}_T$, where $\boldsymbol{C}_i$ is defined earlier in Section 6.2. Our covariance matrix estimate is $\boldsymbol{R'GD}$. Due to the columns of 0's in $\boldsymbol{R}_i$ and $\boldsymbol{G}_i$ for $i < T$, and the repetitive nature of $\boldsymbol{R}_T$ and $\boldsymbol{G}_T$

$$\boldsymbol{R'GD} = \sum_{i=1}^{T} \boldsymbol{R}_i'\boldsymbol{G}_i\boldsymbol{D} \tag{6.17}$$

where $\boldsymbol{R}_i'\boldsymbol{G}_i\boldsymbol{D} \hat{=} \boldsymbol{e}_{T-1}(i) \otimes \boldsymbol{V}_{\hat{\boldsymbol{\beta}}, \bar{\boldsymbol{x}}_i}$ for $i < T$, and $\boldsymbol{R}_T'\boldsymbol{G}_T\boldsymbol{D} \hat{=} \boldsymbol{1}_{(T-1)} \otimes \boldsymbol{V}_{\hat{\boldsymbol{\beta}}, \bar{\boldsymbol{x}}_T}$

$$\begin{aligned} \boldsymbol{V}_{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}} &= \boldsymbol{R'GD} \\ &\approx \begin{bmatrix} \boldsymbol{V}_{\hat{\boldsymbol{\beta}}, \bar{\boldsymbol{x}}_1} & \dots & \boldsymbol{V}_{\hat{\boldsymbol{\beta}}, \bar{\boldsymbol{x}}_{T-1}} \end{bmatrix} + \begin{bmatrix} -\boldsymbol{V}_{\hat{\boldsymbol{\beta}}, \bar{\boldsymbol{x}}_T} & \dots & -\boldsymbol{V}_{\hat{\boldsymbol{\beta}}, \bar{\boldsymbol{x}}_T} \end{bmatrix} = \boldsymbol{V}_{\hat{\boldsymbol{\beta}}, \bar{\boldsymbol{x}}_H}. \end{aligned} \tag{6.18}$$

Our method of estimating the covariance matrix of $\boldsymbol{d}$, which uses estimates of a generalized linear model, agrees with the method of Hussey (2012) which uses U-statistics kernel functions.

While Hussey (2012) uses the imputed data set, our results are reported for the data set with missing covariate values. Still our estimates of $\boldsymbol{d}$ and $\boldsymbol{d}_H$ are very similar (6.19). The estimate of $\boldsymbol{V}_{\boldsymbol{d}}$ and $\boldsymbol{V}_{\boldsymbol{d}_H}$ is provided in the appendix. There is only a slight difference in the adjusted estimates of treatment effect of $\boldsymbol{b}$ vs $\boldsymbol{b}_H$ (6.20) and a slight different in the adjusted estimated covariance matrices (6.21).

6.5 EXAMPLES

We can see from Table 6.4 that while almost half of the patients on the placebo experienced an event (defined as death, a tracheostomy, or intubation with artificial ventilation leading to tracheostomy), the percentage is smaller in all the drug doses. Furthermore, the median survival time for those who had an event is highest for the 100 mg dose (329 days versus 289.5, 282, and 256.5 days), suggesting that the 100 mg dose lengthens time to disease progression.

To explore the proportional hazards assumption for the Cox proportional hazards model, we will examine the Kaplan-Meier curve (Figure 6.1). There appears to be overlap for the first 100 days or so, and then the treatments begin to separate from the placebo. The 200 mg dose appears to cross with both the 50 mg and 100 mg dose at different points in time near the end of the study. While not perfectly satisfying the proportional hazards assumption, the separation achieved in the middle of the study seems adequately compatible.

The distributions of the covariates by treatment group is shown in Tables 6.5 and 6.6. We can see that the distributions of the covariates are very similar for all 3 doses and placebo, suggesting that any impact of covariance adjustment will mostly be through the extent of variance reduction as that corresponds to the correlations of the covariates with the outcomes.

We can see from Table 6.7 that the non-parametric covariance adjustment adjusts the effect of all three doses away from the null. Since we are interested in pairwise comparisons of the dose with placebo, we can test the 3 parameters associated with treatment effect using the Hochberg method. The Hochberg method is a step-up method for multiple comparisons (Hochberg, 1988). The data suggests that the 100 mg and 200 mg dose are significantly different from 0, while there is not enough evidence to make the same conclusion about the 50 mg dose.

We see from Figure 6.2 that we have some variance reduction over the unadjusted method without much change in the estimates themselves, and our method agrees quite well with that of Hussey (2012). This is re-enforced by Table 6.7.

The non-parametric covariance adjustment is greatest for the 200 mg dose. This is possibly due to the larger mean visual analogue scale fatigue score the 200 mg dose has compared to placebo (48.07 vs 43.31). This was the largest difference in the means of the covariates that was observed for all covariates and all treatment groups.

The comparison with the most variance reduction from randomization-based covariance adjustment is that for 100 mg. The unadjusted standard error is 0.1347 while the adjusted standard error is 0.1101. This is an 18.2% reduction in variability. The 50 mg dose has an estimated 16.6% reduction, and the 200 mg dose has a 16.8% reduction. The stronger treatment parameter estimate of the 100 mg dose combined with its reduced estimated variance well confirm the efficacy of the 100 mg dose.

1

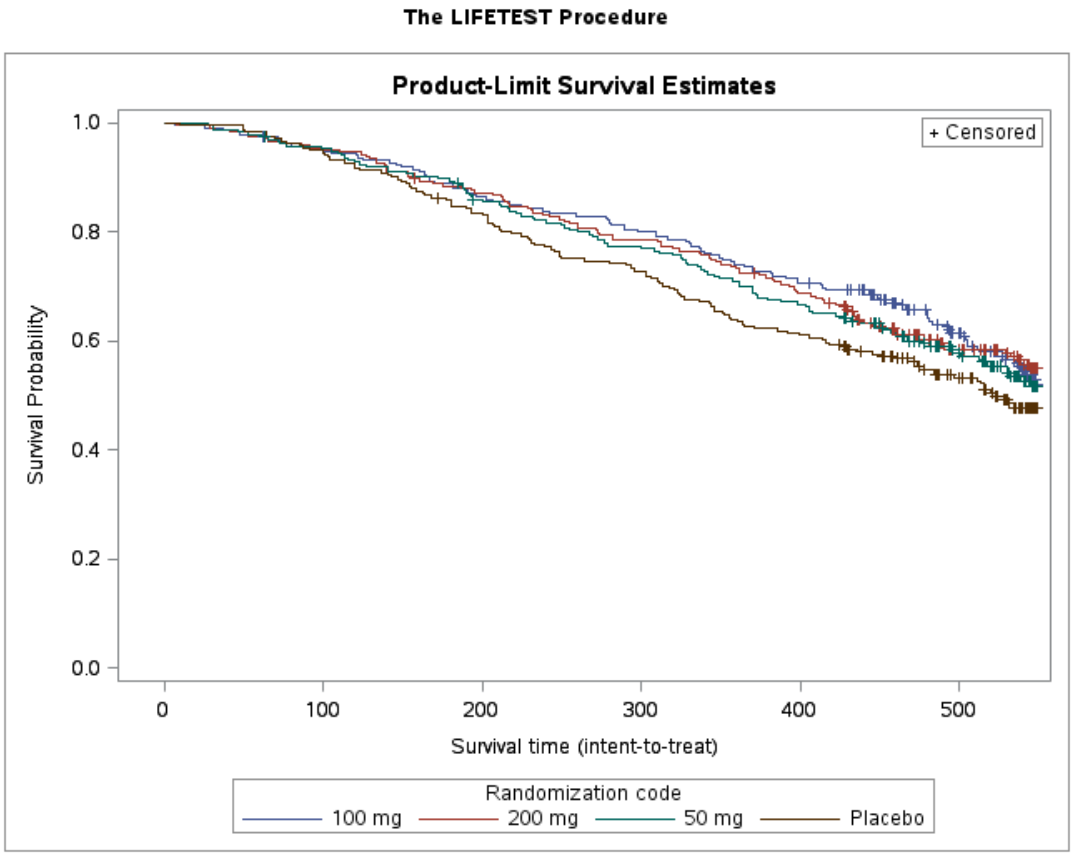

Figure 6.1: Kaplan-Meier plot of 3 doses versus placebo

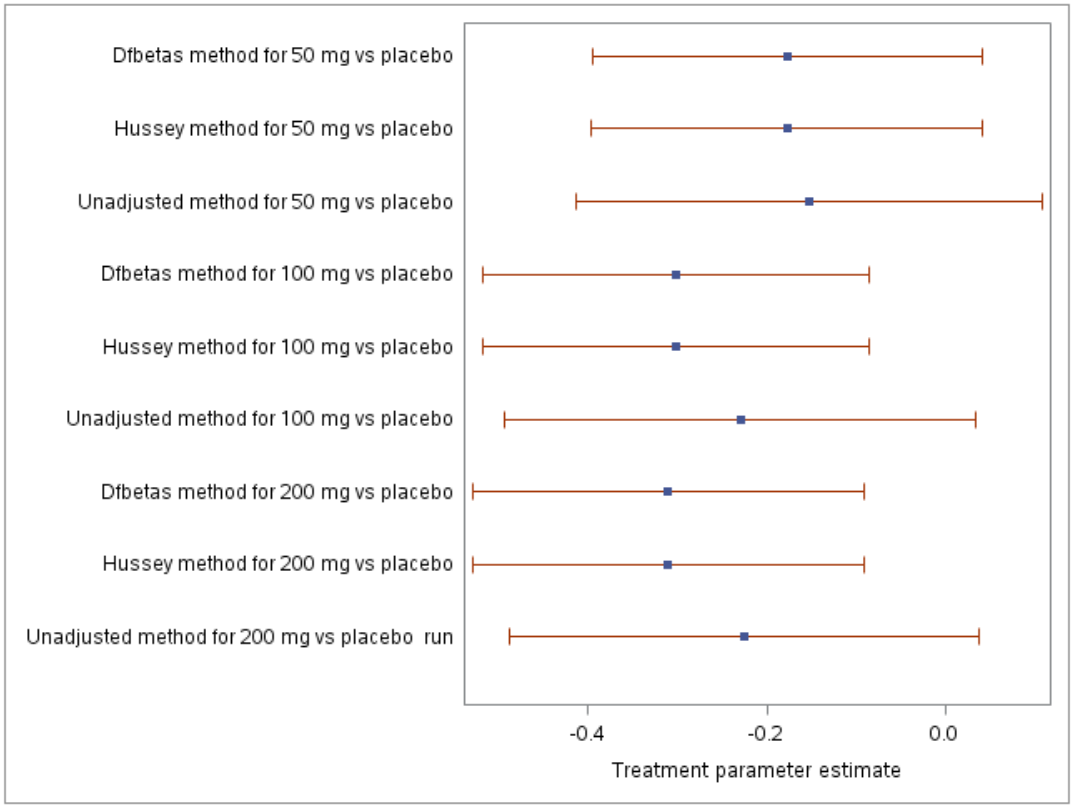

Figure 6.2: Forest plot of treatment doses by method

Table 6.4: Distribution of survival time (days) by treatment and censoring

Dose	Censoring indicator	#	Min	Q1	Median	Q3	Max
50 mg	Required tracheostomy	106	27	186	289.5	405	546
	Censored	131	63	487	541	549	549
100 mg	Required tracheostomy	102	6	170	329	458	549
	Censored	134	62	493	541.5	549	549
200 mg	Required tracheostomy	103	7	160	282	410	545
	Censored	141	157	478	544	549	549
Placebo	Required tracheostomy	120	10	160.5	256.5	368	534
	Censored	122	172	492	542	549	549

$$\boldsymbol{d} = \begin{pmatrix} -0.15284 \\ -0.22981 \\ -0.22595 \\ 1.105224 \\ 0.021829 \\ -0.52947 \\ 0.76529 \\ 0.903731 \\ 0.91071 \\ 0.919176 \\ -0.15251 \\ -0.04213 \\ 1.169933 \\ 1.181435 \\ 0.77418 \\ 0.836167 \\ -0.05602 \\ -1.07567 \\ 0.834009 \\ 4.844518 \\ 0.565 \end{pmatrix}, \boldsymbol{d}_H = \begin{pmatrix} -0.15284 \\ -0.22981 \\ -0.22595 \\ 1.105224 \\ 0.021829 \\ -0.48319 \\ 0.661182 \\ 0.844358 \\ 0.924232 \\ 0.919176 \\ -0.15251 \\ -0.01381 \\ 1.180201 \\ 1.170219 \\ 0.807384 \\ 0.836167 \\ -0.05602 \\ -1.04977 \\ 0.794585 \\ 4.56865 \\ 0.565 \end{pmatrix} \tag{6.19}$$

$$\boldsymbol{b} = \begin{pmatrix} -0.176859 \\ -0.301994 \\ -0.311058 \end{pmatrix}, \boldsymbol{b}_H = \begin{pmatrix} -0.177648 \\ -0.301312 \\ -0.310445 \end{pmatrix} \tag{6.20}$$

$$\boldsymbol{V_b} = \begin{pmatrix} 0.0123679 & 0.0052555 & 0.0052517 \\ 0.0052555 & 0.0121377 & 0.0052493 \\ 0.0052517 & 0.0052493 & 0.012499 \end{pmatrix}, \boldsymbol{V_{b_H}} = \begin{pmatrix} 0.0124148 & 0.0052611 & 0.0052586 \\ 0.0052611 & 0.0121542 & 0.0052524 \\ 0.0052586 & 0.0052524 & 0.012531 \end{pmatrix} \tag{6.21}$$

Table 6.5: Distribution of covariates by treatment

Dose	N Obs	Label	Min	1st Q	Median	3rd Q	Max
50 mg	237	Age (years)	26	50	59	66	75
		Disease duration (years)	0	1	2	3	5
		Weight (inclusion)	37	58	65	76	103
		Muscle Testing Total Score	34	82	92	100	110
		Visual Analogue Scale Fatigue	0	19	45	71	100
		Vital Capacity Ratio	55	73	86	102	157
100 mg	236	Age (years)	24	51	59	65	75
		Disease duration (years)	0	1	1	2	5
		Weight (inclusion)	40	59	67	77	117
		Muscle Testing Total Score	28	80	92	101	110
		Visual Analogue Scale Fatigue	0	19	46	70	100
		Vital Capacity Ratio	57	73	86	103	134
200 mg	244	Age (years)	26	48	59	65	75
		Disease duration (years)	0	1	1	2	5
		Weight (inclusion)	38	59	67	74	107
		Muscle Testing Total Score	26	79	91	100	110
		Visual Analogue Scale Fatigue	0	22	49	76	100
		Vital Capacity Ratio	44	73	85	102	148
Placebo	242	Age (years)	27	48	57	65	81
		Disease duration (years)	0	1	2	3	14
		Weight (inclusion)	32	59	68	78	97
		Muscle Testing Total Score	24	79	91	101	110
		Visual Analogue Scale Fatigue	0	18	48	64	98
		Vital Capacity Ratio	52	73	84	101	135

6.6 DISCUSSION

In summary, we use the outcome dfbetas from the unadjusted treatment effects with the transformed covariate dfbetas from treatment nested within covariates to find the covariance matrix of $\boldsymbol{d}$. This vector is regressed

Table 6.6: Distribution of mean of covariates by treatment group

Variable	50 mg	100 mg	200 mg	Placebo
Age at inclusion (years)	57.09	56.91	56.82	55.99
Disease duration (years)	1.86	1.68	1.78	1.83
Weight (inclusion)	67.64	68.11	67.06	68.13
Muscle Testing Total Score	88.6	89.06	88.66	87.86
Visual Analogue Scale Fatigue (incl.)	44.17	44.48	48.07	43.31
Vital Capacity Ratio at Inc	88.55	88.43	88.21	87.64

Table 6.7: Comparison of unadjusted, Hussey (2012), and dfbeta methods

Dose vs Placebo	Method	Estimate	Std Error	LCL	UCL	p-value
50 mg	Unadjusted	-0.1528	0.1333	-0.4141	0.1084	0.2516
50 mg	Hussey	-0.1777	0.1114	-0.3960	0.0407	0.1109
50 mg	Dfbetas	-0.1769	0.1112	-0.3948	0.0411	0.1118
100 mg	Unadjusted	-0.2298	0.1347	-0.4938	0.0342	0.0880
100 mg	Hussey	-0.3013	0.1103	-0.5174	-0.0852	0.0063
100 mg	Dfbetas	-0.3020	0.1102	-0.5179	-0.0861	0.0061
200 mg	Unadjusted	-0.2260	0.1344	-0.4893	0.0374	0.0926
200 mg	Hussey	-0.3104	0.1119	-0.5299	-0.0910	0.0056
200 mg	Dfbetas	-0.3111	0.1118	-0.5302	-0.0919	0.0054

onto the space of adjusted treatment parameters by a weighted least squares regression to force the differences in the means of the covariates to 0. Hypotheses tests are conducted using the approximately multivariate normal distribution of $\boldsymbol{b}$. With multiple comparisons, we simply extend both the treatment parameter estimates by a factor of $T-1$ and the differences in the means of the covariates by a factor of $T-1$. This allows for the $T-1$ treatment groups to be compared to treatment T.

The covariate dfbeta method requires the MAR-like assumption of covariate missingness through the use of GEE for creation of the dfbetas as well as independent random censoring of the survival time. GEE requires MCAR, but gives essentially equivalent results to corresponding maximum likelihood methods that require MAR, in the case of multivariate normal covariates. To check the latter assumption of independent random censoring of the survival time, the methods used in Zhao et al. (2014) could be applied. Furthermore, we assume the sample size is sufficiently large that $\boldsymbol{b}$ is multivariate normally distributed.

Our method provides computationally straight-forward methods for handling missing baseline covariate data and agrees well with the method by Hussey (2012) that requires complete data and uses U-statistic kernels for justification of the covariance matrix. Furthermore, there are less components involved with calculating

the estimate of $\boldsymbol{V}_d$. Using cluster-level dfbetas from a model of the covariate values with treatment nested within the covariate type is an elegant way of estimating the differences of the means of the covariates when the covariates have missing data. It is readily available in existing statistical packages.

While traditional modeling methods consider the effect of a treatment when all other covariates in the model are the same, this method does the opposite of that. Instead of comparing treatment effects at the same covariate level, we adjust the treatment effect to simulate the situation in which the means of covariates are the same for the treatment populations. This is justified by the randomization of subjects to treatment groups. If there are large covariate imbalances, we may see large shifts in the estimates for treatment comparison, as in the case of the 200 mg dose in the ALS study.

While both the 100 and 200 mg dose have significant effects on the estimated time-to-event compared to placebo, the 100 mg dose had a somewhat smaller estimated variance and an estimated effect similar to the 200 mg dose. These properties make it a useful dose. Even when we performed the analysis on the imputed data we saw the same results. The imputed data set results are so similar to the non-imputed data set results that they have been omitted for brevity.

Future directions could examine the effect of stratification on the analysis. This study had six strata that were not taken into account. Furthermore, as this non-parametric randomization-based methodology is well suited to confirmatory clinical trials due to its few requirements, sample size calculations using this methodology may be attractive as well.

CHAPTER 7: FUTURE DIRECTIONS

7.1 ROC ANALYSIS IN THE PRESENCE OF VERIFICATION BIAS

In our simulation of estimators for correlated AUC's in ROC analysis in the presence of verification bias we explored how the weights performed when there were no volunteers. Future studies could conduct simulations in which there were a group of those screened who volunteer as opposed to being randomly chosen.

The set up would be similar to the without volunteer situation. In our simulation with volunteers, we model the probability of volunteering on a risk factor, such as whether or not a patient has smoked within the last six months. For all 5000 patients in each iteration, let $S \sim$ Bernoulli(π_s), where $s = 1$ if the patient smoked within the last six months and $s = 0$ otherwise. As before, the disease (D) is modeled as $D \sim$ Bernoulli(π). The first screener variable, which can now depend on both disease and risk factor can be modeled as $Y_1|D, S \sim N(\mu_0 + \mu_1 D + \mu_2 S + \mu_3 DS, \sigma_1^2)$ and the second screener can be modeled as $Y_2|D, S \sim N(\beta_0 + \beta_1 D + \beta_2 S + \beta_3 DS, \sigma_2^2)$. For the sampling variables, let $R_1|Y_1, D, S = (Y_1 < c)$ and $R_2|R_1 = 0, Y_1, D, S \sim$ Bernoulli$(1/10)$. This is different than the implementation in the study, where every tenth subject whose screener value was greater than the cutoff was selected, but this is easier to simulate and effectively the same. The volunteers would be represented by $R_3|R_2 = 0, R_1 = 0, Y_1, D, S \sim$ Bernoulli$(logit^{-1}(\phi_0 + \phi_1 S))$. When $R_3 = 1$ then the patient volunteered to go on for disease verification. Different values of ϕ_0 and ϕ_1 would affect the percentage of volunteers. Letting μ_3 and β_3 be 0 would indicate there was no interaction between the disease and risk factor status of the patient. Interpretability (W_1, W_2) would be modeled as $W_1|R_1 = 1, Y_1, D, S \sim$ Bernoulli$(1/w_1)$ and $W_2|R_1 = 0, (R_2 = 1 \text{or} R_3 = 1), Y_1, D, S \sim$ Bernoulli$(1/w_2)$ which would assume that the probability of interpretability of the disease is the same for those whose screener value was above the cutoff, regardless of whether they were randomly chosen or volunteered. We would have 510 iterations for varying values of ϕ_0 and ϕ_1.

Once the data generation is defined for a variety of values of ϕ_0 and ϕ_1, we would calculate the estimated differences in AUC and the true AUC since we have the true disease status for all 5000 patients for each

iteration. We could calculate the bias of all the iterations as well as the 95% confidence interval coverage when volunteers are given a separate category, exploring the effect of volunteers.

7.2 RANDOMIZATION-BASED NON-PARAMETRIC COVARIANCE ADJUSTMENT FOR TIME-TO-EVENT OUTCOMES WITH MISSING DATA EXPANSIONS

Since we used GEE to calculate the estimated differences of the means of the covariates we need to assume that our covariates are MCAR. For baseline covariates, this is a pallet-able assumption. If we assume our data is MAR-like, the unstructured correlation structure suggests that our bias is negligible (Preisser et al., 2002). If instead of using GEE, we had multivariate normally distributed covariates after transformation to the 0-1 scale, we could use maximum likelihood methods which only require that the data be MAR. Extensions to our method of randomization-based non-parametric covariance adjustment with missing covariate data could use SAS's PROC MIXED when calculating $\hat{\phi}$ and $\boldsymbol{G}$.

In clinical trials, there is still more to develop with respect to sensitivity analysis in survival models, weighted ROC analysis in the presence of verification bias, and randomization-based non-parametric covariance adjustment of treatment effects. This dissertation gives a few examples in these areas.

APPENDIX 1: BAYESIAN SENSITIVITY ANALYSIS IN SURVIVAL MODELS SUPPLEMENT

To calculate $FI_{RI}[v]$ in the piecewise exponential setting, we need to calculate $G^{-1/2}(\omega_\phi(0))$ using spectral decomposition. $G(\omega_\phi(0)) = \begin{bmatrix} 1e^{-12} & 4.5e^{-19} \\ 4.5e^{-19} & 1e^{-12} \end{bmatrix}$ which we will express as $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$. The eigenvalues are $\frac{a+c \pm \sqrt{(a+c)^2-4(ac-b^2)}}{2}$, which we will express as λ_1 and λ_2, respectively. They have corresponding eigenvectors $-\frac{a-c \mp \sqrt{(a+c)^2-4(ac-b^2)}}{2b}$, which will be expressed as v_1 and v_2, respectively. Let $Q = \begin{bmatrix} 1 & 1 \\ v_1 & v_2 \end{bmatrix}$.

$$G^{-1/2}(\omega_\phi(0)) = Q \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}^{-1/2} Q^{-1} \tag{A2.1}$$

We use this matrix in the calculation of FI using equation (2.5).

APPENDIX 2: ROC ANALYSIS IN THE PRESENCE OF VERIFICATION BIAS

Let X_i be the disease status of the i^{th} patient. Let Y_i be the test result of the i^{th} patient, and let V_i be a vector of covariates for the i^{th} patient. Let the entire population be $i = 1 \ldots n$. Everyone in this population has a test result, but only a sub-sample have known X_i. Model

$$log\left\{\frac{Pr(R=0|Y,V)}{Pr(R=1|Y,V)}\right\} = h(Y,V) + q(Y,V)X, \tag{A4.1}$$

then model

$$log\left\{\frac{Pr(X=1|R=1,Y,V)}{Pr(X=0|R=1,Y,V)}\right\} = m(Y,V). \tag{A4.2}$$

The double robust estimator for the i^{th} patient's disease status is

$$X_{DR}(\gamma,\mu) = P(R_i,Y_i,V_i|\mu) + U_i(\gamma|\mu) \tag{A4.3}$$

where

$$U_i(\gamma,\mu) = R_i[\{X_i - P(1,Y_i,V_i|\mu)\} + \exp\{h(Y_i,V_i|\gamma) + q(Y_i,V_i)X_i\} * \{X_i - P(0,Y_i,V_i|\mu)\}] \tag{A4.4}$$

$$\begin{aligned} P(R_i,Y_i,V_i) = Pr(X_i=1|R_1=1,Y_i,V_i)* \\ \left\{R_i + \left(\frac{(1-R_i)\exp\{q(Y_i,V_i)\}}{1-Pr(X_i=1|R_i=1,Y_i,V_i)*(1-\exp\{q(Y_i,V_i)\})}\right)\right\}. \end{aligned} \tag{A4.5}$$

We then substitute $[1+\exp\{-m(Y_i,V_i|\mu)\}]^{-1}$ for $Pr(X_i=1|R_1=1,Y_i,V_i)$.

In our situation, since a random sample of the population is sent on for disease verification if their PEF 70%, then we can treat the probability of disease verification as independent of the disease status itself, given that we know the test result. The random sample allows us to treat the verification bias a ignorable, simplifying the estimation of γ and μ. The model we used for disease verification was:

$$log\left\{\frac{Pr(R=0|Y,V)}{Pr(R=1|Y,V)}\right\} = \gamma_0 + \gamma_1 I(PEF<.7) + \gamma_2 I(PEF<.7)*age + \gamma_3 * age \tag{A4.6}$$

The model we used for disease given verification was:

$$log\left\{\frac{Pr(X=1|R=1,Y,V)}{Pr(X=0|R=1,Y,V)}\right\} = \mu_0 + \mu_1 PEF + \mu_2 I(Q3=Y) + \mu_3 I(Q6=Y) \qquad \text{(A4.7)}$$

Using the parameter estimates of γ and μ, the test result, and covariate information, we can form an estimate of X, the disease status for people without a known disease status.

APPENDIX 3: RANDOMIZATION-BASED NON-PARAMETRIC COVARIANCE ADJUSTMENT FOR TIME-TO-EVENT OUTCOMES WITH MULTIPLE TREATMENT GROUPS

$\boldsymbol{V_d}$ USING DFBETA METHOD

Estimates for $\boldsymbol{V_d}$ and $\boldsymbol{V_{d_M}}$ are provided below.

$$\boldsymbol{V_d} = \begin{pmatrix} \boldsymbol{V_{\hat{\beta}}} & \boldsymbol{V_{\hat{\beta},\hat{\theta}}} \\ & \boldsymbol{V_{\hat{\theta}}} \end{pmatrix} \tag{A6.1}$$

$$\boldsymbol{V_{\hat{\beta}}} = \begin{pmatrix} 0.018002 & 0.008604 & 0.008603 \\ 0.008604 & 0.01798 & 0.008604 \\ 0.008603 & 0.008604 & 0.018337 \end{pmatrix} \tag{A6.2}$$

$$\boldsymbol{V_{\hat{\theta}}} = \begin{pmatrix} \boldsymbol{A_{\hat{\theta}}} & \boldsymbol{D_{\hat{\theta}}} & \boldsymbol{D_{\hat{\theta}}} \\ \boldsymbol{D_{\hat{\theta}}} & \boldsymbol{B_{\hat{\theta}}} & \boldsymbol{D_{\hat{\theta}}} \\ \boldsymbol{D_{\hat{\theta}}} & \boldsymbol{D_{\hat{\theta}}} & \boldsymbol{C_{\hat{\theta}}} \end{pmatrix} \tag{A6.3}$$

$$\boldsymbol{V}'_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}} = \begin{pmatrix} 0.046479 & 0.029494 & 0.029418 \\ -0.00339 & -0.00147 & -0.00147 \\ -0.03041 & -0.0166 & -0.01657 \\ -0.04393 & -0.02299 & -0.02296 \\ 0.052747 & 0.029801 & 0.029968 \\ -0.05872 & -0.02811 & -0.02827 \\ 0.029421 & 0.045011 & 0.029396 \\ -0.00147 & -0.00298 & -0.00147 \\ -0.01655 & -0.03173 & -0.01654 \\ -0.02294 & -0.03797 & -0.02287 \\ 0.030181 & 0.067293 & 0.030342 \\ -0.02827 & -0.0602 & -0.02813 \\ 0.029398 & 0.02938 & 0.046725 \\ -0.00147 & -0.00146 & -0.00191 \\ -0.01654 & -0.01645 & -0.02901 \\ -0.02284 & -0.02288 & -0.048 \\ 0.030164 & 0.030171 & 0.079207 \\ -0.02839 & -0.02831 & -0.0676 \end{pmatrix} \tag{A6.4}$$

$$\boldsymbol{A}_{\hat{\boldsymbol{\theta}}} = \begin{pmatrix} 1.0357 & -0.0005 & -0.2546 & 0.0192 & 0.1302 & 0.0411 \\ -0.0005 & 0.0150 & 0.0108 & -0.0470 & 0.0437 & -0.0095 \\ -0.2546 & 0.0108 & 1.4710 & 0.1832 & 0.0870 & -0.0976 \\ 0.0192 & -0.0470 & 0.1832 & 2.2605 & -0.9764 & 0.5635 \\ 0.1302 & 0.0437 & 0.0870 & -0.9764 & 7.2307 & 0.0857 \\ 0.0411 & -0.0095 & -0.0976 & 0.5635 & 0.0857 & 2.9051 \end{pmatrix} \tag{A6.5}$$

$$\boldsymbol{B}_{\hat{\boldsymbol{\theta}}} = \begin{pmatrix} 1.0486 & 0.0086 & -0.2487 & -0.0085 & 0.3395 & 0.0608 \\ 0.0086 & 0.0142 & -0.0059 & -0.0503 & 0.0532 & -0.0180 \\ -0.2487 & -0.0059 & 1.5328 & 0.1155 & -0.2240 & -0.1068 \\ -0.0085 & -0.0503 & 0.1155 & 2.1006 & -1.0599 & 0.4948 \\ 0.3395 & 0.0532 & -0.2240 & -1.0599 & 7.1173 & -0.1329 \\ 0.0608 & -0.0180 & -0.1068 & 0.4948 & -0.1329 & 2.9451 \end{pmatrix} \tag{A6.6}$$

$$\boldsymbol{C}_{\hat{\boldsymbol{\theta}}} = \begin{pmatrix} 1.0301 & 0.0066 & -0.2216 & -0.0485 & 0.3138 & -0.0305 \\ 0.0066 & 0.0142 & -0.0001 & -0.0425 & 0.0541 & -0.0156 \\ -0.2216 & -0.0001 & 1.2950 & 0.1349 & -0.1383 & -0.0519 \\ -0.0485 & -0.0425 & 0.1349 & 2.0548 & -1.0428 & 0.4020 \\ 0.3138 & 0.0541 & -0.1383 & -1.0428 & 7.1749 & -0.2920 \\ -0.0305 & -0.0156 & -0.0519 & 0.4020 & -0.2920 & 2.9272 \end{pmatrix} \tag{A6.7}$$

$$\boldsymbol{D}_{\hat{\boldsymbol{\theta}}} = \begin{pmatrix} 0.5471 & 0.0029 & -0.1336 & 0.0269 & 0.0624 & 0.0290 \\ 0.0029 & 0.0084 & -0.0022 & -0.0267 & 0.0319 & -0.0114 \\ -0.1336 & -0.0022 & 0.7331 & 0.0909 & -0.0591 & -0.0638 \\ 0.0269 & -0.0267 & 0.0909 & 1.1705 & -0.5584 & 0.2909 \\ 0.0624 & 0.0319 & -0.0591 & -0.5584 & 3.3133 & -0.2017 \\ 0.0290 & -0.0114 & -0.0638 & 0.2909 & -0.2017 & 1.3775 \end{pmatrix} \tag{A6.8}$$

$$\boldsymbol{V}_{\boldsymbol{d}_H} = \begin{pmatrix} \boldsymbol{V}_{\hat{\boldsymbol{\beta}}_H} & \boldsymbol{V}_{\hat{\boldsymbol{\beta}},\bar{\boldsymbol{x}}_H} \\ & \boldsymbol{V}_{\bar{\boldsymbol{x}}_H} \end{pmatrix} \tag{A6.9}$$

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}_H} = \begin{pmatrix} 0.0180017 & 0.0086037 & 0.008603 \\ 0.0086037 & 0.0179803 & 0.008604 \\ 0.0086025 & 0.0086039 & 0.018337 \end{pmatrix} \tag{A6.10}$$

$$\boldsymbol{V}_{\bar{\boldsymbol{x}}_H} = \begin{pmatrix} \boldsymbol{A}_{\bar{\boldsymbol{x}}_H} & \boldsymbol{D}_{\bar{\boldsymbol{x}}_H} & \boldsymbol{D}_{\bar{\boldsymbol{x}}_H} \\ \boldsymbol{D}_{\bar{\boldsymbol{x}}_H} & \boldsymbol{B}_{\bar{\boldsymbol{x}}_H} & \boldsymbol{D}_{\bar{\boldsymbol{x}}_H} \\ \boldsymbol{D}_{\bar{\boldsymbol{x}}_H} & \boldsymbol{D}_{\bar{\boldsymbol{x}}_H} & \boldsymbol{C}_{\bar{\boldsymbol{x}}_H} \end{pmatrix} \tag{A6.11}$$

$$\boldsymbol{V}'_{\hat{\boldsymbol{\beta}}_{\boldsymbol{H}},\bar{\boldsymbol{x}}_H} = \begin{pmatrix} 0.0463816 & 0.0294333 & 0.029357 \\ -0.003386 & -0.001466 & -0.00147 \\ -0.029235 & -0.016041 & -0.01601 \\ -0.041727 & -0.02293 & -0.0229 \\ 0.0490974 & 0.0281865 & 0.028352 \\ -0.058322 & -0.028052 & -0.0282 \\ 0.0293601 & 0.0449168 & 0.029335 \\ -0.001462 & -0.002969 & -0.00147 \\ -0.015998 & -0.030219 & -0.01598 \\ -0.022894 & -0.037975 & -0.02282 \\ 0.0285645 & 0.0638801 & 0.028714 \\ -0.028209 & -0.059724 & -0.02806 \\ 0.0293374 & 0.029319 & 0.046628 \\ -0.001463 & -0.001461 & -0.00191 \\ -0.01598 & -0.015887 & -0.02782 \\ -0.022792 & -0.022832 & -0.04772 \\ 0.0285538 & 0.0285714 & 0.075022 \\ -0.028326 & -0.028247 & -0.06747 \end{pmatrix} \tag{A6.12}$$

$$\boldsymbol{A}_{\bar{\boldsymbol{x}}_H} = \begin{pmatrix} 1.0314 & -0.0005 & -0.2453 & 0.0098 & 0.1164 & 0.0391 \\ -0.0005 & 0.0149 & 0.0105 & -0.0474 & 0.0417 & -0.0091 \\ -0.2453 & 0.0105 & 1.3653 & 0.1523 & 0.0729 & -0.0906 \\ 0.0098 & -0.0474 & 0.1523 & 2.1928 & -0.8964 & 0.5363 \\ 0.1164 & 0.0417 & 0.0729 & -0.8964 & 6.6084 & 0.1173 \\ 0.0391 & -0.0091 & -0.0906 & 0.5363 & 0.1173 & 2.8685 \end{pmatrix} \tag{A6.13}$$

$$\boldsymbol{B}_{\bar{\boldsymbol{x}}_H} = \begin{pmatrix} 1.0442 & 0.0086 & -0.2416 & -0.0094 & 0.3246 & 0.0611 \\ 0.0086 & 0.0141 & -0.0057 & -0.0500 & 0.0522 & -0.0180 \\ -0.2416 & -0.0057 & 1.4153 & 0.1062 & -0.2139 & -0.1017 \\ -0.0094 & -0.0500 & 0.1062 & 2.0859 & -1.0397 & 0.4915 \\ 0.3246 & 0.0522 & -0.2139 & -1.0397 & 6.5457 & -0.1215 \\ 0.0611 & -0.0180 & -0.1017 & 0.4915 & -0.1215 & 2.9071 \end{pmatrix} \tag{A6.14}$$

$$\boldsymbol{C}_{\bar{\boldsymbol{x}}_H} = \begin{pmatrix} 1.0259 & 0.0065 & -0.2155 & -0.0495 & 0.2956 & -0.0304 \\ 0.0065 & 0.0141 & 0.0003 & -0.0423 & 0.0518 & -0.0156 \\ -0.2155 & 0.0003 & 1.2127 & 0.1303 & -0.1219 & -0.0492 \\ -0.0495 & -0.0423 & 0.1303 & 2.0282 & -1.0243 & 0.4035 \\ 0.2956 & 0.0518 & -0.1219 & -1.0243 & 6.5796 & -0.2731 \\ -0.0304 & -0.0156 & -0.0492 & 0.4035 & -0.2731 & 2.9151 \end{pmatrix} \tag{A6.15}$$

$$\boldsymbol{D}_{\bar{\boldsymbol{x}}_H} = \begin{pmatrix} 0.5448 & 0.0029 & -0.1305 & 0.0268 & 0.0566 & 0.0288 \\ 0.0029 & 0.0084 & -0.0020 & -0.0266 & 0.0313 & -0.0114 \\ -0.1305 & -0.0020 & 0.6840 & 0.0858 & -0.0589 & -0.0596 \\ 0.0268 & -0.0266 & 0.0858 & 1.1657 & -0.5483 & 0.2897 \\ 0.0566 & 0.0313 & -0.0589 & -0.5483 & 3.0537 & -0.1810 \\ 0.0288 & -0.0114 & -0.0596 & 0.2897 & -0.1810 & 1.3719 \end{pmatrix} \tag{A6.16}$$

BIBLIOGRAPHY

Alonzo, T., Pepe, M., and Lumley, T. (2003). Estimating disease prevalence in two-phase studies. *Biostatistics*, 4:313–326.

Alonzo, T. A. and Pepe, M. S. (2005). Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54:173–190.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, 12(4):387–415.

Begg, C. (1987). Biases in the assessment of diagnostic tests. *Statistics in Medicine*, 6:411–423.

Begg, C. and Greenes, R. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, 39(1):207–215.

Begg, C. B., Greenes, R. A., and Iglewicz, B. (1986). The influence of uninterpretablility on the assessment of diagnostic tests. *Journal of Chronic Diseases*, 39:575–584.

Berger, J. (1990). Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25:303–328.

Berger, J. (1994). An overview of robust Bayesian analysis. *Test*, 3:5–58.

Berger, J. and Berliner, L. M. (1986). Robust Bayes and empirical Bayes analysis with ϵ-contaminated priors. *Annals of Statistics*, 14:461–486.

Chen, H. Y. and Little, R. J. A. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 94:896–908.

Cho, H., Ibrahim, J., Sinha, D., and Zhu, H. (2009). Bayesian case influence diagnostics for survival models. *Biometrics*, 65:116–124.

Colao, A., Petersenn, S., Newell-Price, J., Findling, J. W., Gu, F., Maldonado, M., Schoenherr, U., Mills, D., Salgado, L. R., , and Biller, B. M. (2012). A 12 month phase 3 study of pasireotide in Cushing's disease. *New England Journal of Medicine*, 366:914–924.

Cook, R. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, Series B*, 48:133–169.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34:187–220.

Davis, C. and Quade, D. (1968). On comparing the correlations within two pairs of variables. *Biometrics*, 24:987–995.

Davison, A. and Tsai, C.-L. (1992). Regression model diagnostics. *International Statistical Review/ Revue Internationale de Statistique*, 60:337–353.

DeLong, E., DeLong, D., and Clarke-Pearson, D. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44:837–845.

Dunson, D. and Herring, A. (2003). Bayesian inferences in the Cox model for order restricted alternatives. *Biometrics*, 59:918–925.

Dunson, D. and Park, H.-H. (2008). Kernel stick breaking processes. *Biometrika*, 95:307–323.

Enright, P. L. and Kaminsky, D. A. (2003). Strategies for screening for chronic obstructive pulmonary disease. *Respiratory Care*, 48:1194–1203.

Gehan, E. (1965). A generalized Wilcoxon test for comparing arbitrary single-censored samples. *Biometrika*, 52:203–223.

Gelfand (1999). fda. *Biometrics*, 55(4):1295–1299.

Gelfand, A., Dey, D., and Chang, H. (1992). Model determination using predictive distributions, with implementation via sampling-based methods. (Disc: P 160-167). pages 147–159, oxford. Oxford University Press.

Green, D. and Swets, J. (1966). *Signal detection theory and psychophysics*. Wiley.

Gustafson, P. and Wasserman, L. (1995). Local sensitivity diagnostics for Bayesian inference. *Annals of Statistics*, 23:2153–2167.

Hammill, B. G. and Preisser, J. S. (2006). A SAS/IML software program for gee and regression diagnostics. *Computational Statistics and Data Analysis*, 51:1197–1212.

Hankinson, J. L., Odencrantz, J. R., and Fedan, K. B. (1999). Spirometric reference values from a sample of the general U.S. population. *American Journal of Respiratory Critical Care Medicine*, 159:179–187.

Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843.

Hastie, T. and Tibshirani, R. (1987). Non-parametric logistic and proportional odds regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36:260–276.

Herring, A., Ibrahim, J., and Lipsitz, S. (2002). Frailty models with missing covariates. *Biometrics*, pages 98–109.

Herring, A., Ibrahim, J., and Lipsitz, S. (2004). Non-ignorable missing covariate data in survival analysis: A case-study of an international breast cancer study group trial. *Journal of the Royal Statistical Society-Series C Applied Statistics*, 53:293–310.

Herring, A. H. and Ibrahim, J. G. (2001). Likelihood-based methods for missing covariates in the Cox proportional hazards model. *Journal of the American Statistical Association*, 96:292–302.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–802.

Hosmer, D. W. J. and Lemeshow, S. (1999). *Applied survival analysis*. Wiley, New York.

Hussey, M. A. (2012). *Extensions of nonparametric randomization-based analysis of covariance*. PhD thesis, University of North Carolina, Chapel Hill.

Ibrahim, J., Chen, M., and Kim, S. (2008). Bayesian variable selection for the Cox regression model with missing covariates. *Life time Data Analysis*, 14:496–520.

Ibrahim, J., M.-H., C., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer, New York.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346.

Ibrahim, J. G. and Lipsitz, S. R. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society, Series B*, 61:173–190.

Jiang, H., Symanowski, J., Paul, S., Qu, Y., Zagar, A., and Hong, S. (2008). The type I error and power of non-parametric logrank and Wilcoxon tests with adjustment for covariates–a simulation study. *Statistics in Medicine*, 27:5850–5860.

Kawaguchi, A., Koch, G., , and Wang, X. (2011). Stratified multivariate Mann-Whitney estimators for the comparison of two treatments with randomization based covariance adjustment. *Statistics in Biopharmaceutical Research*, 3:217–231.

Koch, G. G., Tangen, C. M., Jung, J.-W., and Amara, I. A. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine*, 17:1863–1892.

Lacomblez, L., Bensimon, G., Leigh, P. N., Guillet, P., and Meininger, V. (1996). Dose-ranging study of riluzole in amyotrophic lateral sclerosis. *Lancet*, 347:1425–1431.

LaVange, L., Durham, T. A., and Koch, G. G. (2005). Randomization-based nonparametric methods for the analysis of multicentre trials. *Statistical Methods in Medical Research*, 14:281–301.

Leong, T., Lipsitz, S., and Ibrahim, J. (2001). Incomplete covariates in the Cox model with applications to biological marker data. *Journal of Royal Statistical Society-C Applied Statistics*, 50:467–484.

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

Lin, D. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of American Statistical Association*, 88:1341–1349.

Lin, K., Watkins, B., Johnson, T., Rodriguez, J. A., and Barton, M. B. (2008). Screening for chronic obstructive pulmonary disease using spirometry: Summary of the evidence for the u.s. preventive services task force. *Annals of Internal Medicine*, 148:535–543.

Lipsitz, S. R. and Ibrahim, J. G. (1998). Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics*, 54:1002–1013.

Little, R. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Mahabadi, S. E. and Ganhali, M. (2013). An index of local sensitivity to non-ignorability for parametric survival models with potential non-random missing covariate: an application to the SEER cancer registry data. *Journal of Applied Statistics*, 39:2327–2348.

Marshall, K. G. (1996a). Prevention: How much harm? how much benefit? 1. Influence of reporting methods on perception of benefits. *CMAJ*, 154(10)):1493–1499.

Marshall, K. G. (1996b). Prevention: How much harm? how much benefit? 3. Physical, psychological and social harm. *CMAJ*, 155(2):169–176.

Matchar, D. B., Simel, D. L., Geweke, J. F., and Feussner, J. R. (1990). A Bayesian method for evaluating medical test operating characteristics when some patients' conditions fail to be diagnosed by the reference standard. *Medical Decision Making*, 10:102–111.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 42:109–142.

McCulloch, R. (1989). Local model influence. *Journal of the American Statistical Association*, 84:473–478.

Nelson, S., LaVange, L., Nie, Y., Walsh, J., Enright, P., Marinez, F., Mannino, D., , and Thomashow, B. (2012). Questionnaires and pocket spirometers provide an alternative approach for COPD screening in the general population. *CHEST*, 142(2):358–366.

Paik, M. C. and Tsai, W. Y. (1997). On using the Cox proportional hazards model with missing covariates. *Biometrika*, 84:579–593.

Peterson, B. and Harrell, F. E. J. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39:205–217.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:205–207.

Preisser, J. S., Lohman, K. K., and Rathouz, P. J. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, 21:3035–3054.

Preisser, J. S. and Qaqish, B. F. (1996). Deletion diagnostics for generalized estmating equations. *Biometrika*, 83:551–562.

Qaseem, A., Snow, V., Shekelle, P., and et al. (2007). Screening for chronic obstructive pulmonary disease using spirometry: a clinical practice guideline from the American college of physicians. *Annals of Internal Medicine*, 147(9):633–638.

Rabe, K., Hurd, S., Anzueto, A., Barnes, P., Buist, S., Calverley, P., Fukuchi, Y., Jenkins, C., Rodriguez-Roisin, R., van Weel, C., , and Zielinski, J. (2007). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine*, 176(6):532–555.

Robinson, L. D. and Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, 58:227–240.

Rosenblum, M. and van der Laan, M. J. (2009). Consultant's forum: Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics*, 65:937–945.

Rotnitzky, A., Faraggi, D., and Schisterman, E. (2006). Double robust estimate of the area under the receiver-operating characteristic curve in the presence of verification bias. *Journal of the American Statistical Association*, 101(475):1276–1288.

Saville, B. R. and Koch, G. G. (2013). Estimating covariance-adjusted log hazard ratios in randomized clinical trials using cox proportional hazards models and nonparametric randomization based analysis of covariance. *Journal of Biopharmaceutical Statistics*, 23:477–490.

Schacht, A., Bogaerts, K., Bluhmki, E., and Lesaffre, E. (2008). A new nonparametric approach for baseline covariate adjustment for two-group comparative studies. *Biometrics*, 64:1110–11116.

Schluchter, M. and Jackson, K. (1989). Log-linear analysis of censored survival data with partially observed covariates. *Journal of the American Statistical Association*, 84:42–52.

Sinha, D., Ibrahim, J., and Chen, M.-H. (2003). A Bayesian justification of Cox's partial likelihood. *Biometrika*, 90:629–641.

Stokes, M. E., Davis, C. S., and Koch, G. G. (2012). *Categorical Data Analysis Using the SAS System*. Wiley, Cary, NC, 3 edition.

Stram, D. O., Wei, L., and Ware, J. H. (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. *Journal of the American Statistical Association*, 83:631–637.

Tangen, C. M. and Koch, G. G. (1999). Complementary nonparametric analysis of covariance for logistic regression in a randomized clinical trial setting. *Journal of Biopharmaceutical Statistics*, 9:45–66.

Tangen, C. M. and Koch, G. G. (2001). Non-parametric analysis of covariance for confirmatory randomized clinical trials to evaluate dose-response relationships. *Statistics in Medicine*, 20:2585–2607.

Wei, L., Lin, D., and Weissfeld, L. (1989). Analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Associations*, 84:1065–1073.

Wei, W. H. and Su, J. S. (1999). Model choice and influential cases for survival studies. *Biometrics*, 55(4):1295–1299.

Wilt, T. J., Niewoehner, D., Kim, C. B., Kane, R. L., Linabery, A., Tacklind, J., and et al. (2005). Use of spirometry for case finding, diagnosis, and management of chronic obstructive pulmonary disease (COPD). *Rockville, MD: Agency for Healthcare Research and Quality 2005. AHR publication no. 05-E017-2*. Prepared by the Minnesota Evidence-based Practice Center under contract no. 290-02-0009.

Zeger, S. L., Liang, K.-Y., and Self, S. G. (1985). The analysis of binary longitudinal data with time-independent covariates. *Biometrika*, 72:31–38.

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxilliary covariates. *Biometrics*, 64:707–715.

Zhao, Y., Herring, A. H., Zhou, H., Ali, M. W., and Koch, G. G. (2014). A multiple imputation method for sensitivity analyses of time-to-event data with possibly informative censoring. *Journal of Biopharmaceutical Statistics*, 24:229–253.

Zhou, H. and Pepe, M. S. (1995). Auxiliary covariate data in failure time regression. *Biometrika*, 82:299–314.

Zhu, H., Ibrahim, J., Lee, S., and Zhang, H. (2007). Pertrubation selection and influence measures in local influence analysis. *Annals of Statistics*, 35:2565–2588.

Zhu, H., Ibrahim, J., and Tang, N. (2010). Bayesian local influence analysis. *Annals of Statistics*.

Zink, R. C. and Koch, G. G. (2012). NParCov3: a SAS/IML macro for non-parametric randomization-based analysis of covariance. *Journal of Statistical Software*, 50.