A WILCOXON-TYPE STATISTIC FOR REPEATED BINARY MEASURES WITH DROPOUTS AND POSSIBLE MULTIPLE OUTCOMES

by

Okan Umit Elci

BS in Statistics, Dokuz Eylul University, Turkey, 2000

MS in Statistics, Iowa State University, 2004

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Okan Umit Elci

It was defended on

May 27, 2010

and approved by

Sati Mazumdar, PhD, Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Susan S. Sereika, PhD, Associate Professor, Nursing-Health and Community Systems
School of Nursing, University of Pittsburgh

Stewart J. Anderson, PhD, Professor, Department of Biostatistics,
Graduate School of Public Health, University of Pittsburgh

Dissertation Advisor:
Howard E. Rockette, PhD, Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

A WILCOXON-TYPE STATISTIC FOR REPEATED BINARY MEASURES WITH DROPOUTS AND POSSIBLE MULTIPLE OUTCOMES

Okan Umit Elci, PhD

University of Pittsburgh, 2010

In clinical trials, we often compare two treatment groups using repeated binary measures over time. In such trials, we may encounter missing observations, adverse side effects, or non-responsiveness to therapy which for ethical reasons, may result in increased medical intervention beyond the protocol therapy. We developed a family of statistical tests based on the Wilcoxon statistic which orders the vectors of repeated binary observations and events where the ordering is determined by 'clinical relevance'. For some scenarios, clinically meaningful ordering of the vectors may be defined by a natural algorithm, while for other scenarios the ordering is obtained from a group of clinicians. We present the statistical development of the proposed method, effects of the variability of rankings among clinicians, examples of the application of the proposed method using data from a clinical trial on otitis media, and simulation studies comparing the statistical power of the proposed method to more traditional methods of analysis. Our simulation studies indicate that the proposed method is competitive with and, for some scenarios, is preferable to the traditional methods. Although the proposed method is not applicable to every situation, we believe that for some diseases and scenarios, this simple method is noteworthy in the sense that it can be adjusted to extremely complex situations if vectors can be hierarchically ordered in a reasonable fashion, it can be focused on alternatives that have high clinical relevance, and it can be readily adapted to accommodate non-protocol 'outcomes' and

missing data. The public health relevance of this study is that clinically meaningful results can be targeted in clinical trials.

Keywords: Longitudinal data, binary outcomes, Wilcoxon test statistic, ordering, and natural algorithm.

TABLE OF CONTENTS

LIST OF TABLES

PREFACE

I would like to express my special thanks to the members of my dissertation committee who helped and supported me during the process of my study. Dr. Howard Rockette, my academic advisor, deserves my exceptional thanks for his support and invaluable advice in assisting me to determine my career goals. Lastly, I would like to thank to my wonderful family and friends for their support and encouragement.

1.0 INTRODUCTION

Longitudinal studies where measurements are taken repeatedly on the same subjects throughout a period of time are commonly used in the health sciences, biological, biomedical, and epidemiological areas. Since repeated observations are taken from the same subject, they are expected to be correlated with each other over time. These longitudinal serial correlations among the measurements made on the same subject should be taken into account in the analysis of the data. In such studies, the response at each point in time may be continuous, categorical, or binary (e.g., yes/no, success/failure, positive/negative, and presence/absence). In particular, binary outcomes are sometimes collected in clinical trials when the purpose is to examine the effect of a treatment over time. In some studies, continuous outcomes are dichotomized by a threshold value and methods for analyzing binary outcomes are applied. In this work, our main interest will be dichotomous outcomes collected in a study with repeated measures over time. As motivating examples, we use a group of clinical trials that have been conducted to evaluate the efficacy of treatment for otitis media.

Numerous statistical techniques have been developed for analysis of longitudinal binary data. One of the earlier methods used to analyze longitudinal binary data is to sum the repeated binary observations for each individual over time and to compare the proportion of responses between treatment groups using *t*-test. Random effects models (mixed effects logistic regression models) (*Molenberghs and Verbeke* 2005) and marginal models (logistic regression models using generalized estimating equations (GEE)) (*Liang and Zeger* 1986) are two commonly used alternatives in modeling longitudinal binary data. The former permits heterogeneity across the subjects, so the regression coefficients are different between subjects. Thus, random effects

models are called subject-specific models. The random intercept logistic regression model is the simplest form of this modeling for analysis of longitudinal binary measurements. Marginal models are relevant methods for analyzing repeated binary measures, when the main objective of a study is to examine the effect of covariates on the population mean. These models assume that the regression coefficients are the same for all of the subjects and regression coefficients are interpreted for the population level not at the individual level. Hence, they are called population-averaged models. A logistic regression model using GEE, which is regarded as a marginal model, is the longitudinal extension of a logistic regression model for binary responses. Random effects models and marginal models consider correlation among the observations obtained from the same subjects.

In some clinical trials, measurements from subjects are collected at specified time points for each treatment group (e.g. drug and placebo) during the treatment period and the main interest is to examine the efficacy of a new treatment over placebo. Data measured at the last time point may be used for primary efficacy assessment. For example, proportions of positive responses to treatment groups based on binary outcomes are compared to evaluate the efficacy of a new treatment. In this situation, the comparison is based on the measurements at the last time point. Standard statistical methods such as, Fisher's exact test, chi-squared test of independence, t-test, and binary logistic regression can be used to compare two treatment groups (*Ali and Talukder* 2005).

Interest may be focused primarily on understanding the time trend of a treatment, such as the effect of a drug during the early or late stage of the treatment. Another interest is to establish the overall treatment efficacy rather than a time trend of a regimen. A binary logistic regression model using GEE or a logistic regression model with random effects, mentioned above, are two

approaches to evaluate the time trend and the overall efficacy of a treatment. These approaches use the entire longitudinal data collected at each time point.

Methods using comparisons between pairs of subjects and ranking scores have been used to evaluate the differences between two treatment groups. As proposed by Gehan (1965), data from pairs of subjects are compared for analysis of time to the occurrence of a particular event in clinical trials. Moye *et al*. (1992) recommended comparing each of the subjects in one group (placebo) with each of the subjects in another group (treatment) by integrating additional measurements or variables taken over time in the analysis of time-to-event data. Follmann *et al*. (1992) subjectively ranked the patients in clinical trials with multiple outcomes using clinically meaningful information based on the subjective strategy of ordering the patients. A group of people who are experts in the area of study was asked to rank the patients with several outcomes measured over time based on the clinically relevant importance in the analysis of time-to-event data from a cardiovascular clinical trial and then techniques using the ranking scores were applied to assess the treatment effect.

Our main purpose in this study is to introduce a novel statistical method in situations where we wish to compare two treatments using all of the time points, the outcome is binary, and the time points are pre-specified. Our proposed method is based on ordering the entire vector of repeated observations rather than considering the individual components of the vector. Ideally, vectors can be ordered based on clinical relevance. The two treatment groups are then compared by applying statistical techniques handling ranked measurements such as Wilcoxon rank-sum test (*Wilcoxon* 1945) or Mann-Whitney U test (*Mann and Whitney* 1947) and regression methods using the ranked scores. This can be extended to missing data and multiple outcomes.

1.1 MISSING DATA ISSUES

It is uncommon that observations from study subjects at all specified time points will be obtained in longitudinal studies. Thus, missing data frequently occur in these studies. The presence of missing data can cause misleading inferences and incorrect decisions such as biased and inefficient estimates and poor confidence intervals. Missing data can also reduce statistical power.

Ignoring the missing data violates the strict principle of "intention-to-treat" analysis comparing the treatment groups to which they were randomly assigned regardless of the treatment they actually received (*Houck et al.* 2004; *Bubbar et al.* 2006; *Li et al.* 2006; *and Liu et al.* 2006). Intention-to-treat analyses includes all subjects irrespective of whether they received the treatment or they violate the study protocol (e.g., subjects drop out of the study or have inadequate adherence). When the missing data are not handled in an appropriate way, one may obtain erroneous conclusions.

Therefore, it is important to examine and account for the reasons for the missing data prior to choosing an appropriate statistical method to analyze data. It is crucial to understand the patterns of missing data and the mechanisms that result in the data being missing in order to handle missing data in the analysis.

1.1.1 Missing Data Mechanisms

According to the terminology of Little and Rubin (*Little* and *Rubin* 2002), there are three types of missing data mechanisms. A missing data process is called *missing completely at random* (MCAR) when the missingness is not associated with either observed or unobserved outcomes and is not related to any variables. When the data are MCAR and all available data are used in

the analysis, valid inferences can be obtained. It should be emphasized that MCAR is a strong assumption and is usually difficult to satisfy in practice.

Data are said to be *missing at random* (MAR) if the missingness is related to the observed data but does not depend on the potentially unobserved data. MAR is a less restrictive type of missing data and more often, but not always, a reasonable assumption in many applications. When different withdrawal rates for each treatment groups are assumed, data are more likely to be MAR, rather than MCAR (*Liu* and *Gould* 2002). When unequal drop-out rates in the treatment groups are assumed, missingness depends on the treatment groups. It is considered as covariate-dependent dropout. MCAR and MAR are considered as ignorable missingness.

When the missing data are not random and depend on the unobserved data, missing data are called *missing not at random* (MNAR). One almost always obtains biased estimates and/or invalid results using methods which do not take into account the missing values in the analysis when the missing data process is MNAR. Results produced by standard methods of analyzing longitudinal data are not valid when missing data are MNAR. Special care is needed to analyze the data with MNAR. It is required that observed and missing data are jointly modeled to avoid misleading inferences. MNAR is considered as non-ignorable missing data. Second paragraph.

1.1.2 Approaches For Handling Missing Data

Performing analyses using the methods accounting for missing data are of importance in the analysis of data with missing observations. There are different ways to deal with missing data. One of the popular choices is a “complete case” (CC) analysis including only subjects with complete data. Subjects with missing values are excluded from the analysis. Advantage of CC analysis is that any standard statistical methods can be performed on the complete cases.

However, CC analysis produces biased estimates unless the strong MCAR assumption is satisfied (*Little* and *Rubin* 2002). Even though the stringent MCAR assumption holds, one disadvantage of CC analysis is that excluding the subjects with missing observations reduces the sample size leading to the loss of efficiency. Moreover, excluding the subjects with incomplete data violates the principle of intention-to-treat when assessing treatment efficacy or effectiveness.

Another commonly used method to handle missing data in longitudinal studies is the last observation carried forward (LOCF) approach, where the last observed value is replaced for all of that subject's subsequent values that are not obtained. The LOCF approach is a simple data imputation method and satisfies the intention-to-treat principle by including all subjects regardless of missing data. However, the LOCF method makes the very strong and unrealistic assumption about missing data (i.e., subject's response profile remains unchanged at the time of the last observed value prior to dropout). It can produce very biased results and tends to decrease variability which influences the plausibility of parameter estimates due to the assumption of no further change in the profile (*Tang et al.* 2005). Furthermore, the method is typically applied to dropouts and does not address subjects with only intermittent missing data.

For longitudinal data with binary responses, another approach might be the worst and best case analysis which is simple and easy to apply for the imputation of missing data. For example, subjects in placebo groups are likely to withdraw from the study because of lack of treatment improvement and thus, missing observations for these subjects may be imputed as worst response. If subjects who have sufficient benefit from the treatment discontinue the study, missing observations for these subjects may be imputed as best response. When subjects in the treatment group drop out from the study as a result of an adverse effect or lack of any beneficial

treatment effect, best response to all missing responses in the placebo group and worst response to all missing responses in the treatment group can be assigned but such an extreme analysis would eliminate the beneficial effect of the treatment. It is usually recommended for sensitivity analysis to evaluate the effect of treating missing data in different ways regarding the robustness of the results. (*Minini et al.* 2004)

GEE analysis using all available observations from subjects requires the MCAR assumption. GEE yields valid and consistent estimates under a strong MCAR. Complete case and GEE analyses produce invalid estimates when missing data are MAR but not MCAR. Likelihood-based methods (random-effects models) also use all available data obtained from subjects and yield unbiased results when the less restrictive assumption of MAR is satisfied.

Gould (1980) proposed a statistical method for analysis of longitudinal data comparing two treatment groups based on the observations taken at the last time point from clinical trials in the presence of missing data. The main idea in Gould's method is to consider the information about the reasons for withdrawals in the analysis. Response outcomes obtained at the last planned occasion and the reasons for withdrawals from two treatment groups are plausibly ordered and then statistical methods handling the ranked observations (i.e., Mann Whitney test) are performed in comparison of two treatment groups.

One method to handle missing data is the generalized Wilcoxon test for time-to-event data proposed by Gehan (1965). In clinical studies where the outcome of interest is the time to an event, patients may not complete the study and are censored before the time to occurrence of an event. The relative rank of a censored and non-censored observation is only known if the time to event for a subject who is censored or lost to follow-up is greater than the time of event for the non-censored subject. Gehan's generalization of the Wilcoxon statistic is a widely used approach

and is useful for comparing two survival curves for randomly right-censored data. All pairs of subjects in the two groups are compared in terms of the pattern of events and censored measurements.

Multiple imputation (*Rubin* 1987) provides a modern and simulation-based approach for handling missing data. In multiple imputation, several full data sets are generated in which randomly selected observed values are selected from an appropriately defined subset and substituted for missing data. These completed data sets are separately analyzed and results from each generated data set are combined to yield an overall result. A multiple imputation procedure assumes that missing data are MAR. The assumption of MAR is required to produce valid inferences. Statistical methods requiring complete data can be used by a multiple imputation approach. Popularity of using multiple imputation method increases as developments in computer and technology advance.

1.2 ISSUES OF CLINICAL INTERVENTIONS WHICH ALTER OUTCOMES

In some clinical trials, two treatment groups are compared to assess the effect of therapy and comparison is usually made based on one single primary outcome of interest. In such studies, unplanned occurrences may arise and it is common to give non-protocol clinical interventions to the patients due to development of illness or the results of the therapy such as drug allergy, serious adverse experience, and inadequate effectiveness of treatment. Serious adverse effect of a drug in some patients may prevent safe use of treatment which might necessitate change of treatment. Unsatisfactory effect of treatment may also require another intervention. Moreover, giving another therapy might cause ethical issues and violates the study protocol. This necessity

of giving another therapy may result in a dramatic change in the primary outcome or inappropriate outcome and interfere with results of the analysis. To assess the effects of therapy, it may be important to include such events in the analysis. Applying analyses without considering these occurrences may fail to reflect the overall effects of treatment.

For example, in a clinical trial to compare an antibiotic with placebo in children with chronic effusion (fluid in the ear), tube insertion might be needed because of lack of therapeutic effect of the treatment. This non-protocol intervention impacts on the primary outcome. Tube insertion reduces the occurrence of effusion. If tube insertion is performed on a substantial number of children in the placebo group, chronic effusion will be eliminated and the overall response will be satisfactory for these children. The placebo group could actually have a lower average number of days of chronic effusion than the treatment group because of a large number of inserted tubes. Yet clinically this is not considered a good outcome.

In such cases, it is difficult to accommodate these occurrences in the analysis. Incorporating such occurrences into the analysis of data from clinical trials in a way that preserves the clinical relevance of the outcomes would be an important contribution.

1.3 SPECIFIC AIMS

The primary purpose of this methodological research is to develop a new method of analyzing repeated binary data by extending the Wilcoxon test statistic to the vectors of the repeated observations. The performance of the proposed method was compared with existing methods in the analysis of longitudinal binary outcomes. Since most clinical trials compare two interventions or medications and there are natural extensions to more than two groups, we

considered comparison of two treatment groups in this study. Analysis of longitudinal categorical outcomes and comparing more than two groups can be regarded for future areas of research. The following specific aims were considered in this study.

1.3.1 Comparison of Methods for Longitudinal Binary Responses with Complete Data

We first formalized the method for selected orderings of the set of vectors. We contrasted our proposed method against three existing methods in the absence of missing data. We validated the Type I error and compared the statistical power of the four statistical methods using computer simulations. Even though most clinical trials have missing data and the case of a complete data set might be considered unrealistic, we felt this is an important first step in our evaluation. It is also regarded as beneficial to consider complete data because full data sets are provided when multiple imputation approach is used to handle missing data. Moreover, complete data sets are analyzed when complete case analysis or LOCF approaches are chosen to handle missing data even though they have disadvantages and can produce invalid results under particular circumstances.

1.3.2 Comparison of Methods for Longitudinal Binary Responses with Missing Data

Missing data occur frequently in longitudinal studies. Statistical methods can produce unrealistic and invalid results provided that the missingness has not been properly examined or handled in the analysis of the data. Biased and inefficient estimates may be obtained when the statistical procedures accounting for missing data are not performed. The proposed procedure has several natural extensions which incorporate missing data including a generalized Gehan's Wilcoxon

test statistic (*Gehan, E. A.* 1965) which has been used for analysis of censored data in clinical trials.

We modified Gould's idea to account for missing data by ranking withdrawals due to outcome related reasons such as lack of efficacy, adverse experiences, recovery or dramatic improvement of the treatment.

Thus, we evaluated the performance of the proposed method and compared it to the other traditional methods for longitudinal data with binary responses in the presence of missing data with a simulation study. We considered both the MCAR and the MAR missing data process in the comparison of the four statistical procedures.

1.3.3 Development of General Framework for Summarizing Methods of Ranking Vectors of Observations

We developed a general framework for summarizing methods of ranking vectors of observations. These were developed based on selection of mathematical functions to mimic standard approaches and ranking selected by physicians. If many clinicians are available to order the vectors of binary responses with multiple outcomes and adverse effects and homogeneity among their rankings is not obtained, it might be difficult to interpret the results and to apply the proposed method to the data to be analyzed. We examined the effects of the variability of rankings among clinicians and presented a statistical approach to efficiently use this heterogeneity in the analysis of the data. If a sufficient number of clinicians were not available to reach reasonable agreement of rankings, we developed strategies to use the information available from a small group of clinicians to employ the proposed method.

1.3.4 Consideration of Clinical Interventions Which Alter Outcomes

Often in clinical trials, we encounter unpleasant side effects or non-responsiveness to therapy which for ethical reasons, may lead to increased medical intervention beyond the protocol therapy. Therefore, it is necessary to give non-protocol therapy because of a poor clinical outcome with protocol therapy. This may result in patients for whom the primary outcome is drastically altered or for which the primary outcome is no longer appropriate or meaningful. Standard techniques have difficulty incorporating these occurrences. Our proposed method addresses such a problem as long as the 'need for intervention' can be clinically ranked relative to the original vector of outcomes. We developed necessary statistics in order that the proposed method can be applied to such 'interventions' and evaluated the effect on the overall analysis.

2.0 LONGITUDINAL BINARY RESPONSES WITH COMPLETE DATA

Our proposed method is to compare the overall evaluation of treatment between two groups by transforming the information on the entire vector of repeated outcomes into a ranking of vectors in terms of clinical relevance and then applying an appropriate statistical procedure dealing with ranked measurements such as Wilcoxon rank-sum test or Mann-Whitney U test or the regression methods based on ranks. Our approach requires a rank ordering of all the subjects in both groups based on the input from the clinicians. The agreement among the clinicians is an important consideration in the analysis of the method.

2.1 PROPOSED METHODOLOGY

Suppose that we have two treatment groups of subjects, A and B, to be compared. Assume that n_A and n_B subjects are assigned to treatment A (*Trt*-A) and treatment B (*Trt*-B), respectively and measurements are taken from all $N = n_A + n_B$ subjects over time in longitudinal study with k time points. Let x_{it} denote the response from the i^{th} subject within the *Trt*-A and y_{jt} denote the response from the j^{th} subject within the *Trt*-B at time t $(t = 1, 2, \dots, k;\ i = 1,2, \dots, n_A;\ j = 1,2, \dots, n_B)$. First, we assume that all measurements are observed from all subjects that is, there are no missing data. Without loss of generality, let 1 indicate presence of disease and 0 indicate absence of disease. Thus,

$\boldsymbol{X}_i = (x_{i1}, x_{i2}, \dots, x_{ik})_{1xk}$ is the sequence of observed outcomes regarding the i^{th} subject within the *Trt*-A with

$$x_{it} = \begin{cases} 1 & if\ disease \\ 0 & if\ no\ disease \end{cases} \quad i = 1,2,\dots,n_A,\ t = 1,2,\dots,k$$

$\boldsymbol{Y}_j = \left(y_{j1}, y_{j2}, \dots, y_{jk}\right)_{1xk}$ is the sequence of observed outcomes regarding the j^{th} subject within the *Trt*-B with

$$y_{jt} = \begin{cases} 1 & if\ disease \\ 0 & if\ no\ disease \end{cases} \quad j = 1,2,\dots,n_B,\ t = 1,2,\dots,k$$

Since k repeated binary responses are measured, there are 2^k possible response profiles over time. The proposed method is based on the assumption that the vectors can be ordered depending on their clinical relevance. We define a matrix with 2^k rows and k columns. Each row of this matrix represents one of the 2^k possible response profiles. It should be noted that some of the rows might share the same ranked score depending on their clinical importance. We designate this matrix as $\boldsymbol{S}$. Let $\boldsymbol{S}^{(m)}$ indicate the m^{th} row of $\boldsymbol{S}$ and *Rank* ($\boldsymbol{S}^{(\mathrm{m})}$) denotes the assigned score to $\boldsymbol{S}^{(m)}$. Define

$\delta_i = Rank(S^{(\mathrm{r})})$ if $\boldsymbol{X}_{\mathrm{i}} \equiv S^{(\mathrm{r})}$, $\quad i = 1,\dots,n_A;\ r = 1,\dots,2^k$

$\xi_j = Rank(S^{(\mathrm{s})})$ if $\boldsymbol{Y}_{\mathrm{j}} \equiv S^{(\mathrm{s})}$, $\quad j = 1,\dots,n_B;\ s = 1,\dots,2^k$

Hence, $\delta_1, \dots, \delta_{n_A}$ are values of the subjects in *Trt*-A and $\xi_1, \dots, \xi_{n_B}$ are values of the subjects in *Trt*-B. The Mann Whitney form of the test can be defined as

$$U_{ij} = \begin{cases} 0 & if\ \boldsymbol{X}_i < \boldsymbol{Y}_j \text{ or } \delta_i < \xi_j \\ 0.5 & if\ \boldsymbol{X}_i = \boldsymbol{Y}_j \text{ or } \delta_i = \xi_j \\ 1 & if\ \boldsymbol{X}_i > \boldsymbol{Y}_j \text{ or } \delta_i > \xi_j \end{cases}$$

It can be more convenient to work with the following form of the Mann Whitney test.

$$U_{ij}^{*} = \begin{cases} -1 & if\ \boldsymbol{X}_i < \boldsymbol{Y}_j \text{ or } \delta_i < \xi_j \\ 0 & if\ \boldsymbol{X}_i = \boldsymbol{Y}_j \text{ or } \delta_i = \xi_j \\ 1 & if\ \boldsymbol{X}_i > \boldsymbol{Y}_j \text{ or } \delta_i > \xi_j \end{cases}$$

It is easy to see that $U_{ij}^{*} = 2U_{ij} - 1$ and we can calculate the statistic U or U^{*}.

We can calculate the Wilcoxon (1945) test statistic W. $\delta_1, \ldots, \delta_{n_A}, \xi_1, \ldots, \xi_{n_B}$ can be ordered from lowest to highest score and the Wilcoxon statistic W is the sum of the ranks of the δ's in the combined ordered arrangement of δ's and ξ's. When ties are present, ties can be replaced by the average of ranks that the set of tied values would have been assigned if the values were distinct.

$$\mathrm{W} = \sum_{i=1}^{n_A} Rank(\delta_i) = \sum_{i=1}^{n_A} \left\{ \sum_{j=1}^{n_B} \left[I_{\{\delta_i > \xi_j\}} + 0.5\, I_{\{\delta_i = \xi_j\}} \right] + \sum_{j=1}^{n_A} \left[I_{\{\delta_i > \delta_j\}} + 0.5\, I_{\{\delta_i = \delta_j\}} \right] \right\}$$

Each subject (or vectors of binary responses) in *Trt*-A can be compared with each subject in *Trt*-B. The Mann-Whitney (1947) statistic, U, is the number of times a ξ_j ($1 \leq j \leq n_B$) in the *Trt*-B precedes a δ_i ($1 \leq i \leq n_A$) in the *Trt*-A in the combined ranking of the two treatment groups.

$$U = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} U_{ij} = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \left\{ I_{\{\delta_i > \xi_j\}} + 0.5 I_{\{\delta_i = \xi_j\}} \right\}$$

or

$$U^{*} = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} U_{ij}^{*} = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \left\{ I_{\{\delta_i > \xi_j\}} - I_{\{\delta_i < \xi_j\}} \right\}$$

where $I_{\{v<0\}}$ is a set indicator with $I_{\{v<0\}} = 1$ if $v < 0$ and 0 otherwise.

When there are no ties, the U^{*} statistic is related to the Wilcoxon statistic W and Mann-Whitney statistic U. It can be shown that

$$U^* = 2W - n_A(n_A + n_B + 1)$$

$$U^* = 2U - n_A n_B$$

Large values of U or U^* indicate that there is a difference between treatment groups.

2.2 SIMULATION STUDY

We conducted a simulation study to assess the empirical performance of the proposed method versus three commonly used methods for analysis of repeated binary measures: 1) Two-sample *t*-test comparing the average of the proportion of negative (positive) responses over time between two treatment groups; 2) logistic regression modeling using GEE (GEE); and 3) mixed effects logistic regression model with random intercept (MERI). Since often investigators are interested in a consistent treatment effect across time between two treatment groups, we first did not include treatment by time interaction term in the fitting models with GEE and MERI in our simulation study. For completeness, we also compared the proposed method with GEE and MERI when the interaction term of treatment by time is considered.

2.2.1 Statistical Models

We now describe the three approaches applied to each of the simulated longitudinal binary responses in this section. Suppose that a total of *N* subjects are randomized to one of the two treatment groups (A and B). Assume that the number of subjects in *Trt*-A (n_A) and *Trt*-B (n_B) are equal and that observations from each subject are taken at four time points.

Let $Y_{it} = 1$ if the measurement taken from i^{th} subject at time point *t* is positive (presence of disease) response and $Y_{it} = 0$ otherwise; $i = 1, \dots, N = n_A + n_B$ and $t = 1, 2, 3, 4$. Define a treatment indicator as $Trt_i = 0$ if subject *i* is randomized to the *Trt*-A and $Trt_i = 1$ if subject *i* randomized to the *Trt*-B. To simplify notation, we assume that subjects $i = 1, \dots, n_A$ are randomized to *Trt*-A and subjects $i = n_A + 1, \dots, n_A + n_B$ are randomized to *Trt*-B.

Two-sample t-test comparing the mean of the proportions of negative responses

The repeated binary outcomes for each subject over time are summarized as a proportion of positive responses and two treatment groups are compared based on the proportion of positive responses using the two-sample *t*-test. Let p_{Ai} and p_{Bi} be the proportion of positive responses for the i^{th} subject in *Trt*-A and *Trt*-B over the four time points, respectively. The mean of the proportions of positive responses for subjects in *Trt*-A is

$$p_A = \frac{\sum_{i=1}^{n_A} p_{Ai}}{n_A}$$

and the mean of the proportions of positive responses for subjects in *Trt*-B is

$$p_B = \frac{\sum_{i=n_A+1}^{N} p_{Bi}}{n_B}$$

The null hypothesis of no treatment effect can be tested to compare two treatment groups in terms of average proportion of positive responses over time using the two-sample *t*-test: $H_0: p_A = p_B$

Logistic Regression Model using GEE

The marginal probability of positive response for the i^{th} subject at a given time point *t* (p_{it}) was modeled as a logistic function of time and treatment group assuming no treatment by time interaction, that is,

$$ln\left(\frac{p_{it}}{1-p_{it}}\right) = \beta_0 + \beta_1 Time_{it} + \beta_2 Trt_i$$

where

$Time_{it} = 1, 2, 3, 4$ for the four time points,

Trt_i is the treatment indicator which equals to 0 for *Trt*-A and 1 for *Trt*-B, and

$$p_{it} = P\,r(Y_{it} = 1|Trt_i).$$

β_0 and β_1 represent the logit of p_{it} for *Trt*-A and the linear changes in the logit of positive response over time, respectively, and β_2 indicates the difference between treatment groups over time. Since the objective is to evaluate the overall treatment effect, β_2 is of primary interest in the simulation study. When a logistic regression model using GEE was applied to each simulated dataset, different working correlation structures, such as independent, exchangeable, auto-regressive, and unstructured correlation structures were used. We tested the null hypothesis of no treatment difference, that is, H_0: $\beta_2 = 0$

Mixed Effects Logistic Regression Model with Random Intercept

Mixed-effects logistic regression models with a random intercept were considered for modeling each simulated binary data, that is,

$$ln\left(\frac{p_{it}}{1-p_{it}}\right) = \beta_0 + \beta_1 Time_{it} + \beta_2 Trt_i + \zeta_i$$

where the random subject effects ζ_i are distributed as $N(0, \sigma_\zeta^2)$ and ζ_i independent across subjects *i*, given a random intercept and $p_{it} = Pr(Y_{it} = 1|Trt_i, \zeta_i)$, and β_0, β_1, and β_2 represent the same parameters as in logistic regression model using GEE above. We assumed that Y_{it} are independently distributed given p_{it} as $Y_{it}|p_{it} \sim Binomial(1, p_{it})$.

Similar to logistic regression model using GEE, our interest is to test the null hypothesis of no treatment effect over time. We tested the null hypothesis of no treatment difference, that is, $H_0: \beta_2 = 0$

2.2.2 Simulation Design

The power and type I error rates of the four statistical procedures were compared through computer simulations under a range of scenarios including different marginal probabilities of positive response and different correlation structures among the binary observations.

The type I empirical rates were computed as the proportion of time the null hypothesis of no treatment effect was rejected at the two-sided nominal α=0.05. If the empirical type I error was close to the nominal error (α=0.05), the test was considered valid. The statistical power for selected alternative hypotheses (i.e., treatment effect) was calculated as the proportion of rejections of the false null hypothesis of no treatment effect assuming a two-sided type I error of 0.05.

2.2.3 Data Generation

Correlated binary outcomes were generated given the marginal probabilities and correlation structure using the method of Park *et al*. (1996). This method creates correlated binary data from correlated Poisson variables with no requirement of a complex numerical procedure to be solved. The only limitation of this technique is that negatively correlated binary random variables cannot be generated. However, it is not unusual to assume nonnegative correlations among binary observations taken repeatedly from the same subject over time. Demirtas (2004) presented an R routine for generating correlated binary variables using an algorithm developed by Park *et al.*

(1996). 2000 simulated data sets from a hypothetical longitudinal binary data of different sets of parameters were generated using the R function written by Demirtas (2004). In these simulations, we restricted attention to four time points. The response at each time point was either positive (=1) or negative (=0). In each simulation, correlated binary outcomes with 4 repeated measurements were generated for the specified number of subjects for each of the two treatment groups separately given the marginal expectations (i.e., marginal probabilities of positive response at four time points are $p_A = (p_{A1}, p_{A2}, p_{A3}, p_{A4})$ and $p_B = (p_{B1}, p_{B2}, p_{B3}, p_{B4})$ for *Trt*-A and *Trt*-B, respectively) and the correlation structures among responses. The number of subjects was equal for both treatment groups with $n = 60$ subjects for each group.

Six sets of marginal expectations of positive responses at four time points were used for comparing statistical power: (1) no change over time for placebo group while there is an improvement for drug group ($p_A = (.6, .6, .6, .6)$ and $p_B = (.6, .5, .4, .3)$); (2) no change at the beginning of the study for both groups but the placebo group worsens while the drug group gets better later in the study ($p_A = (.7, .7, .8, .9)$ and $p_B = (.7, .7, .6, .5)$); (3) placebo group gradually worsens but the drug group gets better over time ($p_A = (.5, .6, .6, .7)$ and $p_B = (.5, .4, .4, .3)$); (4) both groups show good progress but the drug group shows faster improvement ($p_A = (.8, .7, .7, .6)$ and $p_B = (.8, .6, .5, .4)$); (5) placebo group gets worse while the drug group does not show any change over time ($p_A = (.4, .5, .6, .7)$ and $p_B = (.4, .4, .4, .4)$); and (6) no change over time for the placebo group while the effect of drug is immediate and is maintained over time ($p_A = (.7, .7, .7, .7)$ and $p_B = (.5, .5, .5, .5)$).

A within-subject independent correlation structure (no correlation among the observations, i.e., repeated measurements may be taken at long time intervals resulting in negligible correlations among the measurements) for weakly dependent binary responses,

exchangeable correlation structure (constant correlation among the observations) with correlation coefficient being 0.3 for moderately dependent binary responses, and a first-order auto-regressive, AR(1), correlation structure with correlation coefficient being 0.6 for strongly dependent binary responses

i.e.,

$Corr\left(Y_j, Y_k\right) = 0.6\ for\ |j-k| = 1,$

$Corr\left(Y_j, Y_k\right) = 0.36\ for\ |j-k| = 2,$ and $Corr\left(Y_j, Y_k\right) = 0.216\ for\ |j-k| = 3$were assumed for the association among the responses. No missing values were generated for this simulation study. Thus, the correlation structures were as follows:

Independent correlation structure:

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

Exchangeable correlation structure:

$$R = \begin{bmatrix} 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 \end{bmatrix},$$

AR(1) correlation structure:

$$R = \begin{bmatrix} 1 & 0.6 & 0.36 & 0.216 \\ 0.6 & 1 & 0.6 & 0.36 \\ 0.36 & 0.6 & 1 & 0.6 \\ 0.216 & 0.36 & 0.6 & 1 \end{bmatrix}.$$

2.2.4 Strategies for Ordering Vectors of Binary Responses

Since four repeated binary responses were generated, there are $2^4 = 16$ possible response profiles over time. For each simulated data set, we ordered the vectors based on both the number of

positive responses (1 = disease), and the time to first/last appearance of disease. Although we will eventually obtain input from clinicians on some real life examples, for now we considered two different strategies for ordering the observed vectors. For the first ordering strategy to rank vectors of observations of subjects (strategy I), we have used a ranking which gives priority to the number of disease episodes and then breaks ties by ranking as a poorer outcome those with the *earliest* episode. For the second ordering strategy (strategy II), we have used a ranking which gives priority to the number of disease episodes, as for strategy I, but then breaks ties by ranking as a poorer outcome those with the *latest* episode (Table 2.1). For both ordering strategies, a vector of four positive responses at all time points receives the lowest ranked score and a vector of four positive responses at all time points receives the highest ranked score. We compared two treatment groups with regard to the ordered vectors using the Wilcoxon rank-sum test.

2.2.5 Simulation Results

The results for the type I error rates are presented in Table 2.2 and the results for the power are presented in Table 2.3.

Type I Error Rate

Table 2.2 presents the percentages of the type I error rates for each procedure when the marginal probabilities of positive response (undesirable outcome i.e., having disease) are the same and different at all four time points for both treatment groups using different correlation structures. Table 1 was produced under three different correlation structures: (1) no correlation between the repeated binary responses, (2) exchangeable correlation structure with correlation coefficients being 0.3, and (3) AR(1) correlation structure with correlation coefficient being 0.6. Simulation results showed that none of the procedures yielded high type I error rates and all procedures

produced type I error rates around the nominal value (α=0.05) when independent correlation structure was adopted. All methods were comparable with respect to type I error rate in the absence of correlation among the repeated binary responses.

The proposed method using strategies I and II, t-test, and GEE with different correlation structures preserved the type I error well under the null hypothesis whereas MERI tended to produce slightly higher type I error rate when exchangeable correlation structure was occupied. GEE with unstructured correlation structure yielded a little higher type I error rate for the situation where the marginal probabilities of positive response were the same at the four time points (Patterns 1, 3, and 4) compared to the proposed method, t-test and GEE with other correlation structures.

We did not detect apparent difference with respect to type I error rates over the methods when the AR(1) correlation structure with correlation coefficient being 0.6 was used. However, MERI model produced slightly higher type I error rates than the proposed method, t-test, and GEE with different correlation structures except unstructured (incorrect) correlation structure which yielded slightly higher type I error rates for some scenarios such as patterns 1, 4, 6, and 8. The reason of obtaining higher type I error rate with MERI can be explained by the fact that an increase in the correlation among the observations inflates the type I error rate. (*Hedeker and Gibbons* 2006; *Liu and Wu* 2008). Diggle *et al*. (2002) also showed that positive correlation increases the type I error rate for repeated binary responses.

Overall, the proposed method yielded desirable Type I error rates around the nominal 5 percent and performed well under different scenarios (different marginal probabilities and correlation structures).

Statistical Power

Statistical power of the competing methods from simulations are reported in Table 2.3 when the correlation structures are independent, exchangeable, and AR(1) with different marginal probabilities of positive response at four time points for each treatment group.

Table 2.1: Ordering of 16 possible vectors based on the two different strategies

STRATEGY I									
Number of disease episodes									
0		1		2		3		4	
R	**Profile**	**R**	**Profile**	**R**	**Profile**	**R**	**Profile**	**R**	**Profile**
16	[0 0 0 0]	12	[1 0 0 0]	6	[1 1 0 0]	2	[1 1 1 0]	1	[1 1 1 1]
		13	[0 1 0 0]	7	[1 0 1 0]	3	[1 1 0 1]		
		14	[0 0 1 0]	8	[1 0 0 1]	4	[1 0 1 1]		
		15	[0 0 0 1]	9	[0 1 1 0]	5	[0 1 1 1]		
				10	[0 1 0 1]				
				11	[0 0 1 1]				
Algorithm:									
1. Rank first by number of disease episodes									
2. Break ties by considering *earlier* disease as poorer outcome									
STRATEGY II									
Number of disease episodes									
0		1		2		3		4	
	Profile		**Profile**		**Profile**		**Profile**		**Profile**
16	[0 0 0 0]	15	[1 0 0 0]	11	[1 1 0 0]	5	[1 1 1 0]	1	[1 1 1 1]
		14	[0 1 0 0]	10	[1 0 1 0]	4	[1 1 0 1]		
		13	[0 0 1 0]	9	[1 0 0 1]	3	[1 0 1 1]		
		12	[0 0 0 1]	8	[0 1 1 0]	2	[0 1 1 1]		
					[0 1 0 1]				
					[0 0 1 1]				
Algorithm:									
1. Rank first by number of disease episodes									
2. Break ties by considering *later* disease as poorer outcome									

(1=positive, 0=negative), R=Rank score. Lower rank score is poorer outcome.

When the marginal expectations of positive response are p_A= (.4, .4, .4, .4) and p_B= (.3, .3, .3, .3), the proposed method had slightly lower power than the other methods in the case of no correlation among observations. The proposed method yielded statistical power comparable to other procedures except MERI, which kept a little higher power than the other methods, in the presence of correlation among the responses (exchangeable and AR-1 correlation structures).

When the marginal expectations of positive response are p_A= (.8, .8, .8, .8) and p_B= (.7, .7, .7, .7), our proposed method with two different ordering schemes had slightly lower statistical power than the other methods under the scenario of assuming no correlation among the responses. However, our proposed method yielded similar power compared with MERI and higher power than the *t*-test and GEE when we adapted the exchangeable correlation structure. The proposed method showed slightly higher or similar power in detecting treatment differences under the alternative hypothesis compared with the *t*-test and GEE, with correct or incorrect correlation structures for the scenario of applying AR(1).

Under the scenario of no change over time for the placebo group but fast and constant improvement for treatment group (i.e., p_A= (.7, .7, .7, .7) and p_B= (.5, .5, .5, .5) or p_A= (.5, .5, .5, .5) and p_B= (.3, .3, .3, .3)), all procedures produced similar power regardless of assuming different correlation structures.

In some clinical trials, placebo-treated subjects are expected to worsen over time and no improvement is observed for subjects randomized to drug group. In order to demonstrate this situation in our simulation study, we chose scenarios where the marginal expectations of positive response for placebo and drug groups are p_A= (.4, .5, .6, .7) and p_B = (.4, .4, .4, .4) or p_A= (.6, .7, .8, .9) and p_B= (.6, .6, .6, .6), respectively. Under these scenarios, the proposed

method with strategy II had higher statistical power than the competitive methods in the absence of correlation among the responses but the proposed method with strategy I had the lowest power. When we assumed that repeated measurers are correlated, the proposed method with strategy II had still higher power than the other procedures and loss in power was higher for other procedures than the proposed method with strategy II.

When the marginal probabilities of positive response difference between two treatment groups is high at the beginning of the trial and this difference gradually decreases over time and equalize at the last time points (i.e., p_A= (.7, .6, .5, .4) and p_B = (.4, .4, .4, .4)), the proposed method with strategy I produced statistical power higher than the other procedures and the proposed method with strategy II produced the lowest power in the absence or presence of correlation among the responses. Note that our proposed method produces different power depending on the ordering strategy of the vectors.

For p_A= (.3, .5, .6, .8) and p_B = (.8, .6, .5, .3) using different correlation structures, our proposed method with strategy I and strategy II had noticeably higher statistical power compared to other methods. Other procedures yielded power around the nominal 5 percent. It should be noted that GEE and MERI methods were not designed to be sensitive to interactions.

Under the scenario p_A= (.6, .6, .6, .6) and p_B = (.6, .5, .4, .3), the proposed method with strategy II yielded higher power than the other procedures regardless of employing different assumptions of correlation among the repeated observations. Statistical power produced by the proposed method with strategy I was lower than that produced by the other procedures.

Although our proposed method with strategy I had lower power than the other procedures, the method with strategy II yielded higher power compared to other procedures under the scenario

where the marginal probabilities are p_A= (.8, .7, .7, .6) and p_B = (.8, .6, .5, .4), p_A= (.5, .6, .6, .7) and p_B = (.5, .4, .4, .3) or p_A= (.7, .7, .8, .9) and p_B = (.7, .7, .6, .5).

Table 2.2: Type I Error Rate (%): Simulation results for comparing the methods with full data

	Placebo	**Drug**	**PROPOSED**		***t*-test**	**GEE**				**MERI**
	p_1 p_2 p_3 p_4	**p_1 p_2 p_3 p_4**	**I**	**II**		**IND**	**EX**	**AR(1)**	**UN**	
Independent Correlation Structure										
1	.3 .3 .3 .3	.3 .3 .3 .3	5.2	4.7	4.6	4.6	4.6	4.6	5.2	4.2
2	.4 .4 .4 .4	.4 .4 .4 .4	5.2	5.4	5.2	5.4	5.4	5.2	5.7	5.0
3	.5 .5 .5 .5	.5 .5 .5 .5	4.8	4.5	4.8	4.9	4.9	5.0	5.1	4.5
4	.7 .7 .7 .7	.7 .7 .7 .7	4.2	5.0	4.2	4.2	4.4	4.2	4.9	4.1
5	.6 .4 .6 .4	.6 .4 .6 .4	5.3	5.0	5.4	5.8	5.8	5.7	5.7	4.9
6	.4 .5 .6 .7	.4. 5 .6 .7	5.8	5.7	5.2	5.4	5.4	5.3	5.8	5.0
7	.6 .7 .8 .9	.6 .7 .8 .9	5.3	5.5	5.4	5.4	5.5	5.6	5.8	5.0
8	.7 .6 .5 .4	.7 .6 .5 .4	5.0	5.2	4.8	5.1	5.1	5.0	5.4	4.8
9	.6 .6 .4 .4	.6 .6 .4 .4	5.0	5.1	4.8	5.0	5.0	5.0	5.2	4.2
Exchangeable Correlation Structure										
1	.3 .3 .3 .3	.3 .3 .3 .3	5.1	5.4	5.0	5.0	5.1	5.2	5.7	6.1
2	.4 .4 .4 .4	.4 .4 .4 .4	4.9	5.8	5.5	5.5	5.5	5.6	6.0	6.4
3	.5 .5 .5 .5	.5 .5 .5 .5	5.4	5.8	5.7	5.8	5.8	5.8	6.2	6.8
4	.7 .7 .7 .7	.7 .7 .7 .7	4.9	5.4	4.9	5.4	5.4	5.7	5.8	6.4
5	.6 .4 .6 .4	.6 .4 .6 .4	4.8	4.8	4.6	4.6	4.6	4.6	5.1	5.4
6	.4 .5 .6 .7	.4. 5 .6 .7	5.7	5.4	5.6	5.8	5.8	5.6	6.0	6.4
7	.6 .7 .8 .9	.6 .7 .8 .9	5.8	4.7	4.8	5.2	5.6	5.2	5.6	6.2
8	.7 .6 .5 .4	.7 .6 .5 .4	4.9	5.2	5.0	5.1	5.1	5.3	5.4	6.0
9	.6 .6 .4 .4	.6 .6 .4 .4	4.7	4.4	4.6	4.7	4.7	5.2	5.3	6.0
AR(1) Correlation Structure										
1	.3 .3 .3 .3	.3 .3 .3 .3	5.0	5.4	4.7	4.7	5.0	5.4	5.8	6.0
2	.4 .4 .4 .4	.4 .4 .4 .4	4.7	5.2	5.0	5.0	5.2	5.2	5.8	6.3
3	.5 .5 .5 .5	.5 .5 .5 .5	4.4	4.7	4.9	5.0	5.0	5.2	5.4	5.6
4	.7 .7 .7 .7	.7 .7 .7 .7	5.4	5.0	5.4	5.5	5.6	5.4	6.0	6.4
5	.6 .4 .6 .4	.6 .4 .6 .4	5.2	5.4	5.1	5.2	5.1	5.1	5.4	6.1
6	.4 .5 .6 .7	.4. 5 .6 .7	4.9	4.5	5.2	5.2	5.2	5.1	5.7	5.8
7	.6 .7 .8 .9	.6 .7 .8 .9	5.4	4.8	5.0	5.2	5.2	5.4	5.6	6.1
8	.7 .6 .5 .4	.7 .6 .5 .4	5.4	5.4	5.1	5.4	5.3	5.5	6.0	6.1
9	.6 .6 .4 .4	.6 .6 .4 .4	5.0	5.4	5.2	5.2	5.4	5.2	5.5	6.4

Abbreviations: PROPOSED, proposed method based on the ordering of strategy I and II in Table 2.1; *t*-test, *t*-test comparing the means of the proportions of positive responses between the two groups; GEE, logistic regression model using GEE; MERI, mixed effects logistic regression model with random intercept; IND, independent correlation structure; EX, exchangeable correlation structure; AR(1), first-order auto-regressive correlation structure; UN, unstructured correlation structure; p_t, marginal probability of positive response at time t, t = 1, 2, 3, 4.

We also compared our proposed method with GEE and MERI when the time by treatment interaction term was considered. We reported the simulation results for this comparison in Table 2.4. As expected, GEE and MERI methods gained power in some scenarios (i.e., p_A= (.3, .5, .6, .8) and p_B = (.8, .6, .5, .3), p_A= (.4, .5, .6, .7) and p_B = (.4, .4, .4, .4), p_A= (.6, .6, .6, .6) and p_B = (.6, .5, .4, .3)) compared to results in Table 2.3. Under these scenarios, powers of these methods are comparable with the proposed method when independent correlation structure is used and are higher than that of the proposed method in the presence of correlation among observations. For some scenarios (i.e., p_A= (.4, .4, .4, .4) and p_B = (.3, .3, .3, .3), p_A= (.8, .8, .8, .8) and p_B = (.7, .7, .7, .7), p_A= (.7, .7, .7, .7) and p_B = (.5, .5, .5, .5)), there is a decrease in power for GEE and MERI methods. Under these scenarios, the proposed method yielded higher power than these methods.

In summary, the proposed method is competitive with and, for some scenarios, is preferable to the traditional methods. As the simulation study showed, different ordering strategies of the vectors of binary response give different results.

Table 2.3: Power (%): Simulation results for comparing the methods without missing data

	Placebo	Drug	PROPOSED		*t*-test	GEE				MERI
	p_1 p_2 p_3 p_4	p_1 p_2 p_3 p_4	I	II		IND	EX	AR(1)	UN	
Independent Correlation Structure										
1	.3 .5 .6 .8	.8 .6 .5 .3	56.9	54.4	5.0	5.2	5.2	5.2	7.8	3.6
2	.4 .6 .7. 8	.4 .4 .4 .4	96.9	100	100	100	100	100	100	99.0
3	.4 .5 .6 .7	.4 .4 .4 .4	69.4	97.7	92.0	92.2	92.2	92.2	92.7	91.5
4	.7 .6 .5 .4	.4 .4 .4 .4	97.2	69.6	91.6	91.8	91.6	91.7	91.8	91.2
5	.6 .7 .8 .9	.6 .6 .6 .6	96.9	100	100	100	100	100	100	99.9
6	.8 .7 .7. 6	.8 .6 .5 .4	48.4	73.2	64.8	65.2	65.2	65.4	66.0	63.9
7	.5 .6 .6. 7	.5 .4 .4 .3	92.3	99.9	99.4	99.4	99.4	99.4	99.2	99.4
8	.7 .7 .8 .9	.7 .7 .6 .5	40.1	81.2	68.0	68.7	68.8	68.4	69.3	66.9
9	.6 .6 .6 .6	.6 .5 .4. 3	70.0	97.0	91.4	91.7	91.7	91.7	91.8	91.0
10	.4 .4 .4 .4	.3 .3 .3 .3	58.8	58.6	63.8	64.4	64.4	64.8	65.1	63.1
11	.8 .8 .8 .8	.7 .7 .7 .7	64.8	66.2	70.6	71.3	71.4	71.4	71.4	70.2
12	.7 .7 .7 .7	.5 .5 .5 .5	99.0	99.2	99.6	99.6	99.6	99.6	99.4	99.6
13	.5 .5 .5 .5	.3 .3 .3 .3	99.0	99.0	99.7	99.7	99.7	99.7	99.7	99.7
Exchangeable Correlation Structure										
1	.3 .5 .6 .8	.8 .6 .5 .3	30.6	31.4	4.8	5.0	5.0	5.0	8.5	5.8
2	.4 .6 .7. 8	.4 .4 .4 .4	83.4	98.9	96.3	96.2	96.3	95.6	96.2	97.2
3	.4 .5 .6 .7	.4 .4 .4 .4	47.1	81.2	69.4	69.6	70.2	69.3	70.8	73.2
4	.7 .6 .5 .4	.4 .4 .4 .4	81.2	48.4	69.4	69.7	70.2	69.2	71.4	72.8
5	.6 .7 .8 .9	.6 .6 .6 .6	49.4	88.0	73.5	74.3	69.6	73.3	71.8	78.3
6	.8 .7 .7. 6	.8 .6 .5 .4	50.8	69.4	56.1	56.2	61.4	54.2	61.8	59.6
7	.5 .6 .6. 7	.5 .4 .4 .3	71.4	95.3	88.2	88.4	88.4	87.1	89.2	89.2
8	.7 .7 .8 .9	.7 .7 .6 .5	52.6	92.2	74.8	74.8	74.2	79.4	79.8	78.8
9	.6 .6 .6 .6	.6 .5 .4. 3	53.2	83.4	68.0	68.2	68.7	68.1	69.3	69.6
10	.4 .4 .4 .4	.3 .3 .3 .3	35.2	35.0	36.3	36.8	36.6	36.6	37.8	39.3
11	.8 .8 .8 .8	.7 .7 .7 .7	51.8	51.8	46.0	46.3	46.2	45.2	46.4	52.6
12	.7 .7 .7 .7	.5 .5 .5 .5	91.6	91.4	89.6	89.6	89.6	89.1	89.6	92.0
13	.5 .5 .5 .5	.3 .3 .3 .3	89.4	90.4	91.4	91.6	91.6	90.2	91.4	92.2
AR(1) Correlation Structure										
1	.3 .5 .6 .8	.8 .6 .5 .3	32.6	32.0	5.4	5.6	5.5	5.4	9.8	6.2
2	.4 .6 .7. 8	.4 .4 .4 .4	71.2	96.0	89.2	89.4	90.2	89.7	89.6	90.7
3	.4 .5 .6 .7	.4 .4 .4 .4	33.5	74.4	54.0	54.5	56.8	58.8	58.9	58.8
4	.7 .6 .5 .4	.4 .4 .4 .4	68.8	38.2	53.3	53.4	55.4	57.4	60.0	56.5
5	.6 .7 .8 .9	.6 .6 .6 .6	34.5	71.9	57.6	58.4	52.2	60.6	61.1	60.4
6	.8 .7 .7. 6	.8 .6 .5 .4	62.6	76.4	69.4	69.3	72.2	65.1	69.0	71.4
7	.5 .6 .6. 7	.5 .4 .4 .3	74.4	96.4	90.8	90.8	90.8	89.8	91.2	92.0
8	.7 .7 .8 .9	.7 .7 .6 .5	41.2	87.1	72.0	72.4	72.0	79.5	81.4	74.7
9	.6 .6 .6 .6	.6 .5 .4. 3	64.1	87.4	79.8	80.0	79.8	76.4	78.0	81.7
10	.4 .4 .4 .4	.3 .3 .3 .3	36.0	36.0	36.8	37.0	36.8	36.5	38.0	40.4
11	.8 .8 .8 .8	.7 .7 .7 .7	44.1	43.4	43.0	43.4	43.2	42.4	43.8	48.2
12	.7 .7 .7 .7	.5 .5 .5 .5	91.4	90.0	90.6	90.8	90.8	90.5	91.1	92.4
13	.5 .5 .5 .5	.3 .3 .3 .3	87.3	87.6	88.6	88.7	88.5	88.4	89.3	89.9

Table 2.4: Power (%): Simulation results for comparing the methods without missing data when treatment by time interaction is considered

	Placebo	Drug	PROPOSED		*t*-test	GEE				MERI
	$p_1\ p_2\ p_3\ p_4$	$p_1\ p_2\ p_3\ p_4$	I	II		IND	EX	AR(1)	UN	
Independent Correlation Structure										
1	.3 .5 .6 .8	.8 .6 .5 .3	55.4	54.8	5.3	100	100	100	100	100
2	.4 .6 .7. 8	.4 .4 .4 .4	97.4	100	99.8	100	100	100	100	100
3	.4 .5 .6 .7	.4 .4 .4 .4	71.5	97.3	91.4	97.6	97.6	97.6	97.6	97.6
4	.7 .6 .5 .4	.4 .4 .4 .4	96.8	69.4	90.0	96.3	96.3	96.2	96.4	96.6
5	.6 .7 .8 .9	.6 .6 .6 .6	72.2	98.9	95.4	99.2	99.2	99.3	99.2	99.3
6	.8 .7 .7. 6	.8 .6 .5 .4	66.8	88.2	82.8	84.6	84.6	84.6	85.0	84.7
7	.5 .6 .6. 7	.5 .4 .4 .3	92.0	100	99.2	99.8	99.8	99.8	99.8	99.8
8	.7 .7 .8 .9	.7 .7 .6 .5	67.0	99.4	95.0	99.8	99.8	99.8	99.8	99.8
9	.6 .6 .6 .6	.6 .5 .4. 3	68.8	97.3	91.7	96.8	96.8	96.8	97.0	96.8
10	.4 .4 .4 .4	.3 .3 .3 .3	58.6	58.1	63.7	55.4	55.4	55.6	56.1	53.6
11	.8 .8 .8 .8	.7 .7 .7 .7	67.4	66.4	71.4	61.6	61.6	61.4	61.6	60.2
12	.7 .7 .7 .7	.5 .5 .5 .5	98.8	98.7	99.2	98.4	98.4	98.4	98.4	98.4
13	.5 .5 .5 .5	.3 .3 .3 .3	98.6	98.9	99.2	98.5	98.5	98.5	98.6	98.4
Exchangeable Correlation Structure										
1	.3 .5 .6 .8	.8 .6 .5 .3	32.2	32.8	5.8	100	100	100	100	100
2	.4 .6 .7. 8	.4 .4 .4 .4	82.2	98.9	96.8	100	100	100	100	100
3	.4 .5 .6 .7	.4 .4 .4 .4	49.0	81.8	69.2	93.4	93.4	93.4	93.4	94.7
4	.7 .6 .5 .4	.4 .4 .4 .4	80.6	48.4	68.8	92.5	92.5	91.8	92.1	93.6
5	.6 .7 .8 .9	.6 .6 .6 .6	49.5	88.2	74.8	98.0	97.9	98.0	98.2	98.6
6	.8 .7 .7. 6	.8 .6 .5 .4	51.7	69.4	57.0	71.3	72.0	65.8	70.7	71.6
7	.5 .6 .6. 7	.5 .4 .4 .3	71.0	95.5	88.6	98.9	98.9	99.2	99.0	99.0
8	.7 .7 .8 .9	.7 .7 .6 .5	49.2	91.3	72.6	99.8	99.8	99.7	99.8	99.8
9	.6 .6 .6 .6	.6 .5 .4. 3	53.6	83.6	66.7	92.6	92.6	90.9	92.6	92.8
10	.4 .4 .4 .4	.3 .3 .3 .3	36.1	36.0	37.6	30.1	30.1	30.0	31.4	33.2
11	.8 .8 .8 .8	.7 .7 .7 .7	50.6	50.6	45.0	33.1	33.2	32.0	33.8	40.0
12	.7 .7 .7 .7	.5 .5 .5 .5	93.1	92.2	91.2	83.0	83.0	82.1	83.6	86.1
13	.5 .5 .5 .5	.3 .3 .3 .3	89.4	89.8	91.0	84.0	84.0	82.6	84.0	85.4
Auto Regressive Correlation Structure										
1	.3 .5 .6 .8	.8 .6 .5 .3	33.9	31.0	5.0	100	100	100	100	100
2	.4 .6 .7. 8	.4 .4 .4 .4	71.9	95.8	90.0	99.9	99.8	99.7	99.8	99.9
3	.4 .5 .6 .7	.4 .4 .4 .4	33.4	70.6	52.4	93.1	92.8	90.5	91.0	95.2
4	.7 .6 .5 .4	.4 .4 .4 .4	67.0	37.4	51.8	85.5	85.4	87.2	85.4	89.0
5	.6 .7 .8 .9	.6 .6 .6 .6	34.0	71.8	56.6	97.4	97.4	96.1	96.6	98.5
6	.8 .7 .7. 6	.8 .6 .5 .4	61.2	76.2	68.8	75.4	77.6	71.0	74.2	81.2
7	.5 .6 .6. 7	.5 .4 .4 .3	72.4	96.4	89.4	99.6	99.6	99.6	99.6	99.8
8	.7 .7 .8 .9	.7 .7 .6 .5	42.8	87.7	72.0	99.8	99.8	99.7	99.8	99.8
9	.6 .6 .6 .6	.6 .5 .4. 3	63.6	87.6	79.5	94.0	94.1	93.6	94.2	96.2
10	.4 .4 .4 .4	.3 .3 .3 .3	35.2	35.1	35.8	27.6	27.6	28.1	29.0	33.8
11	.8 .8 .8 .8	.7 .7 .7 .7	48.8	46.7	47.6	36.3	36.3	36.2	37.6	44.9
12	.7 .7 .7 .7	.5 .5 .5 .5	90.6	90.2	90.9	83.9	83.8	82.8	84.0	87.2
13	.5 .5 .5 .5	.3 .3 .3 .3	89.2	89.2	90.2	83.8	83.8	82.8	84.2	87.0

2.3 POWER CALCULATION OF THE PROPOSED METHOD

The proposed method uses a Wilcoxon-type test statistic applied to vectors of observations and as the number of replicates increases, there is no clear concept of an increase in effective sample size as exists for many of the tests for repeated measures. Therefore it is not clear that the power of the Wilcoxon test statistic will be competitive with standard techniques. We believe that it is important to examine the relationship of the power of the proposed method to the number of repeated measurements.

We consider power calculations for a clinical trial that is designed to compare two treatment groups (treatment A and treatment B) on a binary response of disease that will be assessed longitudinally over time. We give an illustration under a scenario where the marginal probability of disease is the same at all time points but different for each group and repeated observations taken from each subject are not correlated (independent correlation structure).

The Wilcoxon test statistic can be viewed as a test based on stochastic dominance or the probability that the rank of a randomly selected subject assigned to treatment A will be lower than that of a randomly selected subject assigned to treatment B. For simplicity, lower rank indicates worse outcome. We calculate the exact probability that a subject assigned to treatment A has a worse rank score than a subject assigned to treatment B for specified marginal probabilities and assuming independence as a function of the number of replicates. Let r = the number of repeated measurements taken from each subject. We assume that the probability of observing disease at each time point is $p_A = 0.8$ and $p_B = 0.7$ for treatment A and B, respectively. For $r = 1$, the probability that the rank of randomly chosen subject from treatment A is lower than that of randomly chosen from treatment B is the product of the probability that a

subject assigned to treatment A has disease and the probability that a subject assigned to treatment B does not have disease

$$Pr(A < B) = (0.8)(0.3) = 0.24$$

We consider the ordering of vectors based on the strategy I given in *Section* 2.2.4 to calculate the probability $Pr(A < B)$ for $r > 1$. Based on this strategy, more time points with disease is worse outcome and rank of a subject who has disease earlier is lower than that of a subject who has disease later if the subjects have the same number of time points with disease. For $r = 2$, for example, a subject who had disease at both time points ($[1, 1]$) has the lowest rank and a subject who had no diseases at all time points ($[0, 0]$) is the highest rank. A subject whose profile is $[1, 0]$ has lower rank score than a subject whose profile is $[0, 1]$. As p_A= (0.8, 0.8) andp_B= (0.7, 0.7) for treatment A and B, respectively, for $r = 2$,

$$Pr(A < B) = Pr([1,1])Pr([1,0]) + Pr([1,1])Pr([0,1]) + Pr([1,1])Pr([0,0]) + Pr([1,0])Pr([0,0]) + Pr([0,1])Pr([0,0]) + Pr([1,0])Pr([0,1])$$

$$Pr(A < B) = (0.64)(0.21) + (0.64)(0.21) + (0.64)(0.09) + (0.16)(0.09) + (0.16)(0.09) + (0.16)(0.21) = 0.3888$$

We calculated $Pr(A < B)$ for $r = 3, 4, 5$ by writing a simple program in R package and present the results in Table 2.5. The probability that a subject in treatment A has lower rank than a subject in treatment B increases as the number of repeated measurements increases as shown in Table 2.5. It indicates that the power for testing treatment differences with the proposed method which uses a Wilcoxon test statistic increases with the increasing of the number of repeated measures.

The power of the Wilcoxon test statistic was computed using the PROC POWER in SAS 9.2 for different values of r. PROC POWER uses the O'Brien-Castelloe approximation to

calculate the power for an equal number of subjects of 60 per group ($n_A = n_B = 60$) and one-tailed alpha level of 0.025. We also estimated the power for the given sample sizes and one-sided Type I error of 0.025 through simulation study using 2000 simulation data sets. For each simulation, independent binary outcomes were generated given the marginal probabilities (e.g., $p_A = (0.8, 0.8)$ and $p_B = (0.7, 0.7)$ for $r = 2$ and $p_A = (0.8, 0.8, 0.8)$ and $p_B = (0.7, 0.7, 0.7,)$ for $r = 3$) for r repeated measurements as demonstrated in *Section* 2.2.3. The power estimates computed using SAS 9.2 and using the simulation study are compared in Table 2.5. Table 2.5 examines the relationship between the power of the Wilcoxon test statistic and the number of repeated observations and shows that power estimates increases as the number of repeated measurements taken from each subject increases.

Table 2.5: Statistical power calculation for a study with r repeated binary measures for an equal sample size of 60 per group and one-tailed alpha=0.025

	Number of repeated binary measurements			
	2	**3**	**4**	**5**
Pr (A < B)	0.39	0.48	0.55	0.59
Power (%)	40.9	54.9	66.8	6.3
Simulation (%)	39.4	57.1	66.2	75.8

2.4 CONCLUSION

We conducted simulation studies to compare the performance of our proposed method to that of commonly used methods for analysis of longitudinal binary data when there are no missing data. Simulation studies indicated that none of the approaches are uniformly better than the others. The type I error for all methods is reasonably close to the nominal value except MERI method

which produced higher type I error rates under the assumption of exchangeable and AR(1) correlation structures. There are some situations where our approach performs better/worse in terms of statistical power than the other approaches depending on the strategy for ordering vectors and the difference in the alternative hypotheses. In most situations, power for testing treatment differences with the proposed method versus the other standard methods is comparable and our proposed method is competitive with other methods. Note that our proposed method in the strategies of ranking we employed detects some types of interaction. Therefore, when interaction terms are not included in the GEE or MERI models, the proposed method tends to have slightly higher statistical power for alternatives that are different only for main effects but are very much inferior if the differences are due to an interaction. When an interaction term is included in these models, they lose statistical power for main effects and the proposed test often has superior power for main effect. We believe that our approach is feasible and it will readily be adaptable to missing data and multiple outcomes (intervention with other treatments). It is also more adjustable in distinguishing ‘clinically relevant difference’.

3.0 LONGITUDINAL BINARY RESPONSES WITH MISSING DATA

Although many longitudinal studies are conducted to obtain measurements from all subjects at each of several time points, occurrence of missing data is common in such studies. In practice, all measurements from all subjects at pre-specified time points are not entirely observed. Withdrawal and loss to follow-up of subjects are one of the main concerns in longitudinal research. It is important to examine the missing data in order to draw valid and realistic results in the analysis of the data. As mentioned in *Section* 1.1, there are different methods of handling missing data in the analysis of repeated binary responses. Our proposed method allows several routes to be followed to incorporate missing data in the analysis of longitudinal binary outcomes. We restrict attention here to settings in which the pattern of missing data is monotone even though the proposed methodology can be applied to dataset in the presence of intermittent missing values. The motivation for the monotone missing data pattern is that it is more straightforward to address the dropouts compared to intermittent missing data.

3.1 SET-UP AND NOTATION

We introduce notations to be used throughout this chapter. We assume that n_A and n_B subjects are assigned to *Trt*-A and *Trt*-B, respectively and binary measurements are taken from all $N = n_A + n_B$ subjects over time in longitudinal study with k time points. To simplify notation, we assume that the design plan was to observe all subjects at the same number of time points. Let $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ and $\boldsymbol{Y}_j = (y_{j1}, y_{j2}, \ldots, y_{jk})$ denote the $1xk$ complete vectors of repeated

binary responses for subject i in the *Trt*-A and subject j in the *Trt*-B, respectively (i.e., outcomes that would have been observed if there were no missing data), with x_{it} representing the binary response observed for the i^{th} subject in the *Trt*-A and y_{jt} representing the binary responses observed for the j^{th} subject in the *Trt*-B at time t ($t = 1,2,\ldots,$ k, $i = 1,2,\ldots, n_A$, $j = 1,2,\ldots, n_B$). When some of the observations from subjects are not observed, we partition the vectors $\boldsymbol{X}_i$ and $\boldsymbol{Y}_j$ into $\boldsymbol{X}_i = (\boldsymbol{X}_i^{O(t-1)}, \boldsymbol{X}_i^{M(t)})$ and $\boldsymbol{Y}_j = (\boldsymbol{Y}_j^{O(t-1)}, \boldsymbol{Y}_j^{M(t)})$, respectively, where $\boldsymbol{X}_i^{O(t-1)} = (x_{i1}, x_{i2}, \ldots, x_{i,t-1})$ denotes the observed part of $\boldsymbol{X}_i$ before time t at which dropout occurs and $\boldsymbol{X}_i^{M(t)} = (x_{it}, x_{i,t+1}, \ldots, x_{ik})$ denotes the unobserved missing part of $\boldsymbol{X}_i$ after occurrence of dropout at time t; similarly, $\boldsymbol{Y}_j^{O(t-1)} = (y_{j1}, y_{j2}, \ldots, y_{j,t-1})$ denotes the observed part of $\boldsymbol{Y}_j$ before time t at which dropout occurs and $\boldsymbol{Y}_j^{M(t)} = (y_{jt}, y_{j,t+1}, \ldots, y_{ik})$ denotes the unobserved missing part of $\boldsymbol{Y}_j$ after occurrence of dropout at time t. Without loss of generality, 1 represents the presence of disease, 0 represents the absence of disease, and $\cdot$ indicates a missing observation. If the measurement is observed it takes a value 0 or 1 and if the measurement is not observed, it is considered as missing. In this situation,

$$x_{is} = \begin{cases} 1, & if\ disease \\ 0, & if\ no\ disease \\ \cdot, & missing \end{cases} \qquad i = 1,2,\ldots,n_A;\ s = 1,2,\ldots,k$$

$$y_{js} = \begin{cases} 1, & if\ disease \\ 0, & if\ no\ disease \\ \cdot, & missing \end{cases} \qquad j = 1,2,\ldots,n_B;\ s = 1,2,\ldots,k$$

When we restrict the missingness to dropouts, there are k possible points at which a dropout can occur. Hence, there are $2^k - 1$ possible response profiles in case of the occurrence of dropout. If we assume that no dropout occurs at the 1^{st} time point, there are $2^k - 2$ possible missingness profiles over time.

Since there are 2^k possible response profiles over time for a complete sequence and $2^k - 2$ possible profiles over time for an incomplete sequence, there are $2^{k+1} - 2$ possible vectors for complete and incomplete sequences together. In other words, the maximum number of binary vectors to be obtained cannot exceed $2^{k+1} - 2$. We define a scalar dropout indicator $D_s = t$ representing the time t at which dropout occurs between time t and $(t+1)$ for $t = 2, \dots, k$ for s^{th} vector. For an incomplete vector of binary responses, D_s denotes the time at which dropout occurs. For a complete vector of binary responses, $D_s = k + 1$ indicates no dropout (i.e. all measurements are observed).

We define a matrix with $(2^{k+1} - 2)$ rows and k columns. Each row of this matrix represents one of the possible response profiles with or without missing data and a ranked score will be assigned to each row. We designate this matrix as $\boldsymbol{Z}$. We will give details about how to create the $\boldsymbol{Z}$ matrix when missing data are encountered.

For k time points, $\boldsymbol{Z}_s = (z_{s1}, z_{s2,} \dots, z_{sk})$ with $s = 1,2, \dots, 2^{k+1} - 2$ denoting one of the possible $2^{k+1} - 2$ binary representations for longitudinal binary outcomes in the presence or absence of missing data.

If some responses are not observed, we partition $\boldsymbol{Z}_s$ into two sub-vectors such that $\mathbf{Z}_s^{\mathrm{O}(t-1)}$ represents the sub-vector of observed measurements before time t at which loss to follow-up occurs and $\boldsymbol{Z}_s^{\mathrm{M}(t)}$ represents the sub-vector of unobserved measurements after occurrence of dropout at time t.

$$\boldsymbol{Z}_s = \left(\boldsymbol{Z}_s^{\mathrm{O}(t-1)}, \boldsymbol{Z}_s^{\mathrm{M}(t)}\right) \tag{3.1}$$

where

$$\boldsymbol{Z}_s^{\mathrm{O}(t-1)} = (z_{s1}, \dots, z_{s,t-1}) \qquad t = 2, \dots, k$$

$$\boldsymbol{Z}_s^{\mathrm{M}(t)} = (z_{st}, \dots, z_{sk})$$

Let $\mathbf{C}$ be a matrix with 2^k rows and k columns consisting of all $\mathbf{C}_s$, $s = 1, \dots, 2^k$. Each row of $\mathbf{C}$ represents one of the possible response profiles without missing observations as shown below ($\mathbf{C}_s$ denotes s^{th} row of $\mathbf{C}$ with k binary outcomes and $\mathbf{C}_{s,p}$ denotes the element of the s^{th} row and p^{th} column of $\mathbf{C}$):

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_{2^k} \end{pmatrix} \tag{3.2}$$

with

$$\boldsymbol{C}_{s,p} = \begin{cases} 1, & (h-1)2^{k-p} < s < h2^{k-p} \quad h = 1,3, \dots, 2^p - 1; \; s = 1,2, \dots, 2^k; p = 1,2, \dots, k \\ 0, & otherwise \end{cases}$$

Let $\boldsymbol{M}$ be a matrix with $2^k - 2$ rows and k columns consisting of response sequences including at least one missing observation. Each row of $\boldsymbol{M}$ denotes one of the possible response profiles with the occurrence of loss to follow-up as shown below:

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{M}_1 \\ \vdots \\ \boldsymbol{M}_{2^k-2} \end{pmatrix} \tag{3.3}$$

with

$$\boldsymbol{M}_{s,p} = \begin{cases} 1, & G + (h-1)2^{k-d-p} < s \le G + h2^{k-d-p} \\ 0, & G + h2^{k-d-p} < s \le G + (h+1)2^{k-d-p} \\ \cdot, & otherwise \end{cases} \qquad \begin{matrix} d = 1,2, \dots, k-1 \\ p = 1,2, \dots, k-d \\ h = 1,2, \dots, 2^p - 1 \\ s = 1,2, \dots, 2^k - 2 \end{matrix}$$

where

$G = 2^k\left(2^{-1} + 2^{-2} + \cdots 2^{-(d-1)}\right) I_{(d \ge 2)}$ with indicator function I

The rows of the matrix $\boldsymbol{M}$ with $2^k - 2$ rows and k columns are added after the rows of $\boldsymbol{C}$ matrix with 2^k rows and k columns to obtain $\boldsymbol{Z}$ matrix with $2^{k+1} - 2$ rows and k columns such as

$$\boldsymbol{Z}_{(2^{k+1}-2)xk} = \begin{pmatrix} \boldsymbol{C}_{2^k xk} \\ \boldsymbol{M}_{2^{k-2}xk} \end{pmatrix} \tag{3.4}$$

As an illustration for $k = 4$, $\boldsymbol{C}$ matrix with $2^4 = 16$ rows and 4 columns, $\boldsymbol{M}$ matrix with $2^4 - 2 = 14$ rows and 4 columns, and $\boldsymbol{Z}$ matrix with $2^{4+1} - 2 = 30$ rows and 4 columns would be

$$\boldsymbol{C} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \boldsymbol{M} = \begin{pmatrix} 1 & 1 & 1 & \cdot \\ 1 & 1 & 0 & \cdot \\ 1 & 0 & 1 & \cdot \\ 1 & 0 & 0 & \cdot \\ 0 & 1 & 1 & \cdot \\ 0 & 1 & 0 & \cdot \\ 0 & 0 & 1 & \cdot \\ 0 & 0 & 0 & \cdot \\ 1 & 1 & \cdot & \cdot \\ 1 & 0 & \cdot & \cdot \\ 0 & 1 & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot \end{pmatrix} \quad \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{C} \\ \boldsymbol{M} \end{pmatrix} \tag{3.5}$$

3.2 SCORE ASSIGNMENT TO THE VECTORS WITH MISSING DATA

We assume that ranked scores are determined by a natural algorithm or obtained from a group of clinicians for all possible response profiles in the situation where no lost to follow-up has occurred. In other words, a group of clinicians ordered the complete sequences of binary

responses. We did not ask them to consider vectors with missing data in the ordering of the binary outcomes. Hence, we only have ranked scores of fully observed repeated responses from clinicians. We also do not incorporate the reasons of withdrawals from the study when assigning a score to the vectors including missing data. We basically use the matrix $\boldsymbol{C}$ which is obtained based on subjective agreement of a group of clinicians as a reference to assign a score to each row of the $\boldsymbol{M}$ matrix. We indicate a ranked score of a vector of binary responses, $\boldsymbol{C}_i$ as $Rank\{\boldsymbol{C}_i\}$ or $R\{\boldsymbol{C}_i\}$.

The idea is to assign scores to the first 2^{k-1} rows of $\boldsymbol{M}$ using the assigned scores of each row of $\boldsymbol{C}$ and sequentially assign scores to the other rows of $\boldsymbol{M}$ using the scores assigned to the first 2^{k-1} rows of $\boldsymbol{M}$ calculated by using the scores of the rows of $\boldsymbol{C}$. In order to apply this procedure, we need to calculate probabilities or weights in the following equation.

Dropout at $k-1$

$$R\{\boldsymbol{M}_s\} = Pr[\boldsymbol{M}_s = \boldsymbol{C}_{2s-1}]R\{\boldsymbol{C}_{2s-1}\} + Pr[\boldsymbol{M}_s = \boldsymbol{C}_{2s}]R\{\boldsymbol{C}_{2s}\}, \qquad s = 1,2,\dots,2^{k-1} \tag{3.6}$$

Dropout at $k-d$

$$R\{\boldsymbol{M}_F\} = Pr[\boldsymbol{M}_{\mathrm{F}} = \boldsymbol{M}_L]R\{\boldsymbol{M}_L\} + Pr[\boldsymbol{M}_F = \boldsymbol{M}_T]R\{\boldsymbol{M}_T\} \tag{3.7}$$

where

$$F = \left[2^k\left(2^{-1} + 2^{-2} + \dots + 2^{-(d-1)}\right)\mathrm{I}_{(\mathrm{d}\geq 2)} + \mathrm{s}\right]$$
$$L = \left[2^k(2^{-1} + 2^{-2} + \dots + 2^{-(d-2)})I_{(d-1\geq 2)} + 2s - 1\right] \qquad with \qquad d = 2,3,\dots,k-1,\quad s = 1,2,\dots,2^{k-d}$$
$$T = \left[2^k(2^{-1} + 2^{-2} + \dots + 2^{-(d-2)})I_{(d-1\geq 2)} + 2s\right]$$

In the above equation (3.7), k represents the number of replicates, d represents the number of missing observations in the vector, and $k-d$ indicates the time lost to follow-up.

In *Section* 3.2.1, we describe two techniques to assign a score to a vector of binary responses with missing data (row of $\boldsymbol{M}$) using the scores assigned to the vectors of binary

responses without missing data (rows of $\boldsymbol{C}$). Score of a vector with missing value can be assigned by estimating the probability that a response was positive (negative) if it had not been missing. In other words, if the probabilities in equations (3.6) and (3.7) are estimated, score to be assigned to a vector with missing data can be calculated.

In *Section* 3.2.2, we briefly discuss Gehan's Wilcoxon statistic which had been proposed to analyze time to event data and Gould's method which incorporates information about reason of withdrawals in the analysis. We modified these two approaches to apply the proposed method to assign score to the vectors with dropouts.

3.2.1 Stochastic Approach

We provide strategies for calculating the probabilities in (3.6) and (3.7) to estimate the scores of the vectors with dropouts.

3.2.1.1 Crude Substitution

One approach to assign a score to the vector of outcomes with missing values is to form a weighted average of the ranks that could theoretically have occurred if there were no missing data. Probabilities or weights in (3.6) and (3.7) can be determined based on a priori set probabilities.

The outcome at any time point may not be obtained from the subjects for different reasons. If the missing response was observed, it would have the value 0 or 1. We illustrate this approach in the case of four observations collected from each subject over time. For example, we assume that missingness occurs at the last time point. The score to be assigned to a vector with a missing

observation at the 4^{th} time point would be one of the two scores from the two possibly observed vectors of responses if the observation at the last time point would have been observed as outcome is binary taking one of the two outcomes. Therefore, we can assign a score to this vector between the two scores. Let [1, 0, 0, •] denotes the $1x4$ vector of incomplete binary data for a given subject with missing value occurring at the last time point (• indicates missing value). If the last observation from this subject was not missing, the complete outcome vector for this subject would be [1, 0, 0, **0**] or [1, 0, 0, **1**]. If 1 indicates a positive (undesirable or disease) outcome and a higher score is better, the score to be assigned to the vector with missing observation could be higher than the score assigned to [1, 0, 0, 1] and lower than the one assigned to [1, 0, 0, 0]. This approach simply uses the fact that outcome is dichotomous and does not take into account the reasons of withdrawals. A score to be assigned to [1, 0, 0, •] would be a simple weighted average of the scores assigned to [1, 0, 0, 0] and [1, 0, 0, 1]. We could, for example, assign equal weights to the occurrence of a 0 or 1. A modification of the technique would be to give priority to the outcomes that previously occurred in this individual. We refer to this approach as the "crude substitution approach".

When ties are present, previous responses could be used to break the tied values. We illustrate a situation where ties can occur by extending the above example. Suppose we ordered the vectors of complete binary responses observed at all time points based on the information from clinicians or a nested set of criteria and assigned a score to [0, 1, 1, 0] which is higher than the score assigned to [1, 0, 0, 1] and lower than the one assigned to [1, 0, 0, 0]. Assume that applying the above method results in tied rank for the two vectors, [1, 0, 0, •] and [0, 1, 1, 0]. We can break the tied values by comparing these two vectors based on the observations measured in the first three time points. As 1 indicates positive response, [1, 0, 0] which is the first three

observations of [1, 0, 0, •] would be better result than the [0, 1, 1] which is the first three observations of [0, 1, 1, 0]. Therefore, we can assign a higher score to [1, 0, 0, •] than [0, 1, 1, 0].

Moreover, we actually applied equal weight to the vectors ([1, 0, 0, 0] and [1, 0, 0, 1]) to assign a score to the vector with missing data in the first example above. In this hypothetical example, we assumed that the occurrence of 1 and that of 0 at the last time point are equal regardless of the previously observed responses.

We illustrate how to assign scores to the vectors with missing data, rows of the $\boldsymbol{M}$ matrix in (3.5), by using the assigned scores to the rows of $\boldsymbol{C}$ based on strategy I in Table 2.1. We first assign scores to the first eight rows of $\boldsymbol{M}$ by using equation (3.6) and scores assigned to the rows of $\boldsymbol{C}$ and then assign scores to the other rows of $\boldsymbol{M}$ by using equation (3.7). In the same way, scores can be assigned based on strategy II.

The first row of $\boldsymbol{M}$, $\boldsymbol{M}_1$ = (1 1 1 •), would have been $\boldsymbol{C}_1$ = (1 1 1 1) or $\boldsymbol{C}_2$ = (1 1 1 0) if a missing observation had been observed. The rank of this vector would be

$R\{\boldsymbol{M}_1\} = Pr[\ \boldsymbol{M}_1=(1\ 1\ 1\ 1)\ ]R\{\boldsymbol{C}_1\} + Pr[\boldsymbol{M}_1=(1\ 1\ 1\ 0)]R\{\boldsymbol{C}_2\}$

The last observation is more likely to be a 1 rather than 0 based on the previous responses in this vector and therefore, $Pr[\ \boldsymbol{M}_1= (1\ 1\ 1\ 1)\ ]$ would be higher than $Pr[\ \boldsymbol{M}_1= (1\ 1\ 1\ 0)\ ]$. We assigned a score of 1 to $\boldsymbol{C}_1$ and 2 to $\boldsymbol{C}_2$ based on the strategy I in Table 2.1 If we assume that $Pr[\ \boldsymbol{C}_1= (1\ 1\ 1\ 1)\ ]$ = 0.6 which is higher than $Pr[\ \boldsymbol{M}_1= (1\ 1\ 1\ 0)\ ]$ = 0.4, the rank of $\boldsymbol{M}_1$ is

$R\{\boldsymbol{M}_1\} = 0.6*1 + 0.4*2 = 1.4$

The second row of $\boldsymbol{M}$, $\boldsymbol{M}_2$ = (1 1 0 •), would have been $\boldsymbol{C}_3$ = (1 1 0 1) or $\boldsymbol{C}_4$ = (1 1 0 0) if a missing observation had been observed. Based on the previous responses, we can assign equal weights to the vectors. As $Pr[\ \boldsymbol{M}_2= (1\ 1\ 0\ 1)\ ] = Pr[\ \boldsymbol{M}_2= (1\ 1\ 0\ 0)\ ]$ = 0.5 and assigned scores to $\boldsymbol{C}_3$ and $\boldsymbol{C}_4$ are 3 and 6, respectively, rank of $\boldsymbol{M}_2$ is

$R\{\boldsymbol{M}_2\} = Pr[\ \boldsymbol{M}_2 = (1\ 1\ 0\ 1)\ ]R\{\boldsymbol{C}_3\} + Pr[\boldsymbol{M}_2=(1\ 1\ 0\ 0)]R\{\boldsymbol{C}_4\} = 0.5*3 + 0.5*6 = 4.5$

Similarly, we assigned scores to the other vectors with one missing value (rows 3-8 of $\boldsymbol{M}$) below.

$R\{\boldsymbol{M}_3=(1\ 0\ 1\ \bullet)\} = Pr[\boldsymbol{M}_3=(1\ 0\ 1\ 1)]R\{\boldsymbol{C}_5\} + Pr[\boldsymbol{M}_3=(1\ 0\ 1\ 0)]R\{\boldsymbol{C}_6\} =0.5*4 + 0.5*7 = 5.5$

$R\{\boldsymbol{M}_4=(1\ 0\ 0\ \bullet)\} = Pr[\boldsymbol{M}_4=(1\ 0\ 0\ 1)]R\{\boldsymbol{C}_7\} + Pr[\boldsymbol{M}_4=(1\ 0\ 0\ 0)]R\{\boldsymbol{C}_8\} =0.4*8 + 0.6*12=10.4$

$R\{\boldsymbol{M}_5=(0\ 1\ 1\ \bullet)\} = Pr[\boldsymbol{M}_5=(0\ 1\ 1\ 1)]R\{\boldsymbol{C}_9\} + Pr[\boldsymbol{M}_5=(0\ 1\ 1\ 0)]R\{\boldsymbol{C}_{10}\}=0.6*5 + 0.4*9=6.6$

$R\{\boldsymbol{M}_6=(0\ 1\ 0\ \bullet)\} = Pr[\boldsymbol{M}_6=(0\ 1\ 0\ 1)]R\{\boldsymbol{C}_{11}\} + Pr[\boldsymbol{M}_6=(0\ 1\ 0\ 0)]R\{\boldsymbol{C}_{12}\}=0.5*10 + 0.5*13=11.5$

$R\{\boldsymbol{M}_7=(0\ 0\ 1\ \bullet)\} = Pr[\boldsymbol{M}_7=(0\ 0\ 1\ 1)]R\{\boldsymbol{C}_{13}\} + Pr[\boldsymbol{M}_7=(0\ 0\ 1\ 0)]R\{\boldsymbol{C}_{14}\}=0.5*11 + 0.5*14 = 12.5$

$R\{\boldsymbol{M}_8=(0\ 0\ 0\ \bullet)\} = Pr[\boldsymbol{M}_8=(0\ 0\ 0\ 1)]R\{\boldsymbol{C}_{15}\} + Pr[\boldsymbol{M}_8=(0\ 0\ 0\ 0)]R\{\boldsymbol{C}_{16}\}=0.4*15 + 0.6*16=15.6$

By using (3.7) and the ranks of first eight rows of $\boldsymbol{M}$, we assigned scores to the last six rows of $\boldsymbol{M}$ as follows.

$R\{\boldsymbol{M}_9=(1\ 1\ \bullet\ \bullet)\} = Pr[\boldsymbol{M}_9=(1\ 1\ 1\ \bullet)]R\{\boldsymbol{M}_1\} + Pr[\boldsymbol{M}_9= (1\ 1\ 0\ \bullet)]R\{\boldsymbol{M}_2\}=0.6*1.4 + 0.4*4.5=2.6$

$R\{\boldsymbol{M}_{10}=(1\ 0\ \bullet\ \bullet)\} = Pr[\boldsymbol{M}_{10}=(1\ 0\ 1\ \bullet)]R\{\boldsymbol{M}_3\} + Pr[\boldsymbol{M}_{10}=(1\ 0\ 0\ \bullet)]R\{\boldsymbol{M}_4\}=0.5*(5.5+10.4)=7.95$

$R\{\boldsymbol{M}_{11}=(0\ 1\ \bullet\ \bullet)\} = Pr[\boldsymbol{M}_{11}=(0\ 1\ 1\ \bullet)]R\{\boldsymbol{M}_5\} + Pr[\boldsymbol{M}_{11}=(0\ 1\ 0\ \bullet)]R\{\boldsymbol{M}_6\}=0.5*(6.6+11.5)=9.1$

$R\{\boldsymbol{M}_{12}=(0\ 0\ \bullet\ \bullet)\} = Pr[\boldsymbol{M}_{12}=(0\ 0\ 1\ \bullet)]R\{\boldsymbol{M}_7\} + Pr[\boldsymbol{M}_{12}=(0\ 0\ 0\ \bullet)]R\{\boldsymbol{M}_8\}=0.4*12.5+0.6*15.6=14.4$

$R\{\boldsymbol{M}_{13}=(1\ \bullet\ \bullet\ \bullet)\} = Pr[\boldsymbol{M}_{13}=(1\ 1\ \bullet\ \bullet)\ ]R\{\boldsymbol{M}_9\} + Pr[\boldsymbol{M}_{13}=(1\ 0\ \bullet\ \bullet)]R\{\boldsymbol{M}_{10}\}=0.5*(2.6+7.95)=5.3$

$R\{\boldsymbol{M}_{14}=(0\ \bullet\ \bullet\ \bullet)\} =$

$Pr[\boldsymbol{M}_{14}=(0\ 1\ \bullet\ \bullet)\ ]R\{\boldsymbol{M}_{11}\}+Pr[\boldsymbol{M}_{14}=(0\ 0\ \bullet\ \bullet)]R\{\boldsymbol{M}_{12}\}=0.5*(9.1+14.4)=11.75$

3.2.1.2 Logistic Regression Substitution

In this section, we describe another strategy to assign scores to the vectors with missing data using the observed data and scores assigned to the possible vectors without missing data. We estimate the probability that a response is positive (negative) at a given time point by applying a

conditional logistic regression model given the previous responses and treatment variable. We fit conditional logistic regression models based on the non-missing data to predict expected values for response being positive (disease) at each time point separately.

Consider a situation in which a binary response is obtained at each visit of $t = 1, \dots, 4$ and $\boldsymbol{Y}_i = (Y_{i1},\ Y_{i2}\ ,\ Y_{i3},\ Y_{i4})$. Y_{it} indicates the response variable for the i^{th} subject at time t $(i = 1, \dots, N)$. A measurement Y_{it} in a longitudinal sequence is described as a function of previous outcomes, $Y_i^{\ O(t-1)} = \left(Y_{i1}, \dots, Y_{i,t-1}\right)$. Missing values occurring after the first time point can be predicted by conditional estimates of the probability that the unobserved value would be equal to 1 or 0 if it was observed. To compute these probabilities, we assume that the probability of a positive response is determined by previously observed responses and a treatment indicator. For example, a standard logistic regression model based on the available data at time $t = 2$ given the non-missing data at time $t = 1$ and treatment indicator can be fitted to estimate the probability of positive response at time $t = 2$. Therefore, we fit the following logistic regression models to estimate the probabilities of positive response at points where missing data occurred given the treatment indicator and the previous response,

$$logit(Pr(Y_{i2} = 1|Trt_i, y_{i1})) = \beta_0 + \beta_1 Trt_i + \beta_2 y_{i1}$$

$$logit(Pr(Y_{i3} = 1|Trt_i, y_{i2})) = \alpha_0 + \alpha_1 Trt_i + \alpha_2 y_{i2}$$

$$logit(Pr(Y_{i4} = 1|Trt_i, y_{i3})) = \gamma_0 + \gamma_1 Trt_i + \gamma_2 y_{i3}.$$

Then

$$Pr(Y_{i2} = 1|Trt_i, y_{i1}) = \frac{e^{\beta_0 + \beta_1 Trt_i + \beta_2 y_{i1}}}{1 + e^{\beta_0 + \beta_1 Trt_i + \beta_2 y_{i1}}}$$

$$Pr(Y_{i3} = 1|Trt_i, y_{i2}) = \frac{e^{\alpha_0 + \alpha_1 Trt_i + \alpha_2 y_{i2}}}{1 + e^{\alpha_0 + \alpha_1 Trt_i + \alpha_2 y_{i2}}}$$

$$Pr(Y_{i4} = 1|Trt_i, y_{i3}) = \frac{e^{\gamma_0 + \gamma_1 Trt_i + \gamma_2 y_{i3}}}{1 + e^{\gamma_0 + \gamma_1 Trt_i + \gamma_2 y_{i3}}}$$

We assume that the probability of a positive response at a given time point is entirely determined by the available responses observed at the previous time point. The estimated parameters are used to predict the probability of positive response for missing observations.

For illustration, we demonstrate how to apply this approach to assign scores to the vectors with missing data. Here we empirically calculate the probabilities in (3.6) by using the available observations from the dataset. As we considered the first row of $\boldsymbol{M}$, $\boldsymbol{M}_1$ = (1 1 1 •), we estimate the probability that $\boldsymbol{M}_1$ is equal to $\boldsymbol{C}_1$ = (1 1 1 1) or $\boldsymbol{C}_2$ = (1 1 1 0) which is the probability that the last observation is 1 or 0. We estimate this probability by fitting the logistic regression model mentioned above and substitute into the following equation

$$R\{\boldsymbol{M}_1\} = Pr[\ \boldsymbol{M}_1=(1\ 1\ 1\ 1)\ ]R\{\boldsymbol{C}_1\}+Pr[\boldsymbol{M}_1=(1\ 1\ 1\ 0)]R\{\boldsymbol{C}_2\}$$

Similarly, we estimate the probabilities of positive response at each time point where the observation was not obtained by fitting logistic regression models at each time point to assign scores to the other vectors with missing data.

This approach differs from the 'crude substitution approach' in S*ection* 3.2.1.1 since the assigned scores to the vectors with missing data depend on the empirical frequency distributions of known outcomes. We refer to this approach as "logistic regression substitution approach".

It should be noted that the reasons of missingness are not known and we assume that we do not have any information about reasons of occurrence of dropouts.

3.2.2 Nonstochastic Approach

The strategies considered here are not required to assume or estimate probabilities as in *Section* 3.2.1. Some vectors with missing data can still be partially ranked because its relative score

compared to another vector does not change regardless of the 'true' outcomes for the missing data. Also in some clinical trials, subjects may drop out of the study without completing all schedules visits. If reasons for withdrawals are known, the information about withdrawals can be incorporated in the analysis.

3.2.2.1 Modification of Gehan's Wilcoxon Test

In clinical trials where the primary outcome is time to occurrence of an event (e.g., death), some subjects are censored before completing the study. To incorporate subjects who were censored into the analysis of data, Gehan (1965) proposed a method for integrating information about dropping out of the study. The Gehan-Wilcoxon test is formulated on the basis of comparisons of all pairs of subjects where each pair contains one individual from each group. Even with incomplete data the relative rank of some pairs can be determined. For example, in a clinical trial comparing two treatment groups in terms of time to the occurrence of a death, if censored time of a subject X is greater than the time of death for a subject Y, X is ranked higher than Y because it is known that the time of death for subject X is greater than that for subject Y. We can modify Gehan's Wilcoxon test to apply the proposed method for analyzing longitudinal binary data. Consider a situation where binary response is measured over four time points. A subject X dropped out of the study after the second time point and [1, 1, •, •] denotes the profile for subject X. Another subject Y completed the study and the observed profile of this subject is [1, 0, 0, 0]. A subject X is ranked lower than a subject Y regardless of missing values because it is known that X would have worse outcome than Y even those missing values were assumed to be disease free.

3.2.2.2 Modification of Gould's Method

When we have information about reasons for discontinuing the trial, we can incorporate this information in the ranking. Follow-up measurements which are scheduled to be observed after the initiation of treatment may be missing for some subjects due to treatment-related reasons or the progression of the disease. Some subjects withdraw from the study as a result of an adverse effect, lack of efficacy of the treatment, recovery, and external reasons irrelevant to the progress of the disease or unwillingness to continue the study. For example, in placebo-controlled trials, subjects in the placebo group who experience little or no improvement may drop out from the study to look for better treatment or subjects in the drug group who recover considerably may withdraw from the study because they think that it is not necessary to continue the study. Such subjects with missing observations can be included in the analysis of the data by assigning a rank that represents a better or worse score relative to those actually observed based on the reason of withdrawal or the time of leaving the study.

One method for analysis of data with missing observations proposed by *Gould* (1980) is to incorporate the information about reasons of withdrawals into a rank ordering of subjects. Gould proposed an approach for analyzing longitudinal data with continuous responses when the outcomes observed at the last time point are used to make comparisons between two treatment groups without using the previous observations. If subjects prematurely withdraw from the study due to outcome related reasons (i.e., lack of improvement, adverse experiences, and beneficial effect of treatment), informatively missing observations occur and so, a pre-scheduled outcome at the end of the study cannot be measured for these subjects. Gould (1980) suggested ordering these withdrawals by the drop-out reasons and using rank tests to compare the treatment groups based on the rank scores produced by incorporating subjects with missing data.

If it is known that subjects who did not complete the study have better or worse responses than those who completed the study (e.g. cured subjects or lack of treatment effect), relevant scores that may be used in an analysis are assigned to such subjects.

We can modify Gould's idea to account for missing data in the analysis of longitudinal binary responses by ranking withdrawals based on the reasons such as treatment-related adverse events and a satisfactory effect of treatment Subjects who have withdrawn from the study due to treatment-related reasons can be ranked as better/worse values based on their missing data information and reasons for withdrawals.

Non-completers can be considered as 'poor' outcomes if they withdraw from the study due to an adverse side effect or treatment failure or they can be regarded as 'good' outcomes if reasons of leaving the study are satisfactory effect of treatment or cure. Subjects withdrawn from the study at the same time point might receive the same score if their withdrawal reasons are same. Thus, these subjects share the same tied rank value. These tied ranks can be broken by ordering these vectors on the basis of the responses obtained prior to withdrawal. Moreover, another variable measured at baseline or during the study can be used to break the ties.

3.3 PROPOSED METHOD OF ANALYSIS

Define

$$\delta_i = Rank\left(\boldsymbol{Z}^{(r)}\right) \quad \text{if } \boldsymbol{X}_i = \left(\boldsymbol{X}_i^{O(t-1)}, \boldsymbol{X}_i^{M(t)}\right) \equiv \boldsymbol{Z}^{(r)}, \quad i = 1,2,\ldots, n_A; r = 1,2,\ldots, 2^{k+1} - 2$$

$$\xi_j = Rank\left(\boldsymbol{Z}^{(s)}\right) \quad \text{if } \boldsymbol{Y}_{\mathrm{j}} = (\boldsymbol{Y}_{\mathrm{j}}^{\mathrm{O(t-1)}}, \boldsymbol{Y}_{\mathrm{j}}^{\mathrm{M(t)}}) \equiv \boldsymbol{Z}^{(\mathrm{s})}, \quad j = 1,2,\ldots, n_B; s = 1,2,\ldots, 2^{k+1} - 2$$

Hence, $\delta_1, \ldots, \delta_{n_A}$ are values of the subjects in *Trt*-A and $\xi_1, \ldots, \xi_{n_B}$ are values of the subjects in *Trt*-B.

We can compare each of the n_A subjects in *Trt*-A with each of the n_B subjects in *Trt*-B using the relative order of the two vectors of outcomes in the presence of missing data. The assigned score comparing subject i in *Trt*-A with subject j in *Trt*-B is denoted as U_{ij}. Thus, the comparison of the two treatment groups can be defined by a series of scores U_{ij}, $i = 1,2, \ldots, n_A$, $j = 1,2, \ldots, n_B$. We assume that U_{ij} takes the values 0, 0.5 or 1 if the i^{th} subject is worse, the same or better than the j^{th} subject, respectively. It can be defined as

$$U_{ij} = \begin{cases} 0 & if\ (\boldsymbol{X}_i^{O(t-1)}, \boldsymbol{X}_i^{M(t)}) < (\boldsymbol{Y}_j^{O(t-1)}, \boldsymbol{Y}_j^{M(t)})\ or\ \delta_i < \xi_j \\ 0.5 & if\ (\boldsymbol{X}_i^{O(t-1)}, \boldsymbol{X}_i^{M(t)}) = (\boldsymbol{Y}_j^{O(t-1)}, \boldsymbol{Y}_j^{M(t)})\ or\ \delta_i = \xi_j \\ 1 & if\ (\boldsymbol{X}_i^{O(t-1)}, \boldsymbol{X}_i^{M(t)}) > (\boldsymbol{Y}_j^{O(t-1)}, \boldsymbol{Y}_j^{M(t)})\ or\ \delta_i > \xi_j \end{cases}$$

We compare two treatment groups by applying a Wilcoxon rank-sum test to the ranked scores, $\delta_1, \ldots, \delta_{n_A}, \xi_1, \ldots, \xi_{n_B}$. Rank scores can be ordered from lowest to highest score and the Wilcoxon statistic W is the sum of the ranks of the δ's in the combined ordered arrangement of δ's and ξ's. When ties are present, ties can be replaced by the average of ranks that the set of tied values would have been assigned if the values were distinct.

$$\mathrm{W} = \sum_{i=1}^{n_A} Rank(\delta_i) = \sum_{i=1}^{n_A} \left\{ \sum_{j=1}^{n_B} \left[I_{\{\delta_i > \xi_j\}} + 0.5\, I_{\{\delta_i = \xi_j\}} \right] + \sum_{j=1}^{n_A} \left[I_{\{\delta_i > \delta_j\}} + 0.5\, I_{\{\delta_i = \delta_j\}} \right] \right\}$$

The Mann-Whitney (1947) statistic, U, is the number of times a ξ_j ($1 \leq j \leq n_B$) in the *Trt*-B precedes a δ_i ($1 \leq i \leq n_A$) in the *Trt*-A in the combined ranking of the two treatment groups.

$$U = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} U_{ij} = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \left\{ I_{\{\delta_i > \xi_j\}} + 0.5 I_{\{\delta_i = \xi_j\}} \right\}$$

where $I_{\{v<0\}}$ is a set indicator with $I_{\{v<0\}} = 1$ if $v < 0$ and 0 otherwise.

Large values of U indicate that there is a difference between treatment groups

3.4 SIMULATION STUDY

We conducted a simulation study comparing the performance of the proposed method and the frequently applied three standard methods using simulated data sets with various primary assumptions about the complete data sets and missing data mechanisms. Statistical models corresponding to each method of analysis were described in *Section* 2.2.1. For the *t*-test comparing the average proportions of positive responses between two treatment groups, proportions for each subject in both treatment groups were calculated based on the observed responses in which the number of positive responses was divided by the number of observed measurements, not the number of time points. The methods were compared with respect to type I error rates and power.

3.4.1 Simulation Design

The simulation study consisted of two different phases: (1) the data-generating phase and (2) analysis phase. In the first step, repeated binary responses were generated for each treatment group (placebo and drug) separately and missing values were imposed on these binary measurements. In the second step, data with missing values were analyzed by our proposed method and other commonly used methods under a wide range of scenarios and different missing data assumptions.

3.4.2 Data Generation

We assumed that measurements were collected at four time points (t_1, t_2, t_3, and t_4). We simulated correlated binary responses given the marginal probabilities of positive (disease) responses and independent correlation structures, exchangeable correlation structures with correlation coefficient being 0.3, and AR(1), correlation structures with correlation coefficient being 0.6 (i.e., $Corr\ (Y_j, Y_k) = 0.6$, $Corr(Y_j, Y_k) = 0.36$, and $Corr(Y_j, Y_k) = 0.216$ for $|j - k| = 1,2,3$, respectively) using the method of *Park et al*. (1996) as mentioned in *Section* 2.2.3. The response at each time point was either positive (= 1) or negative (= 0) if it was observed. The same sets of marginal expectations and correlation structures for the association among the measurements were considered as we used in *Section* 2.2.3.

We assumed that the treatment indicator (A indicates placebo group and B indicates drug group) is completely measured and that dropout could occur at the 2^{nd}, 3^{rd}, or 4^{th} time points. When a measurement is missing from any subject at any given time point, all subsequent measurements for that subject are assumed to be missing (monotone missing data pattern). 2000 simulated data sets of 60 subjects per treatment group for four time points were generated for each setting and a missingness mechanism imposed to the generated data. We applied all methods to the same simulated dataset.

We subjected each simulated data set to various missing data patterns and proportion of withdrawals. After simulating full longitudinal binary responses for each treatment group, we randomly deleted responses (MCAR) at specified rates and deleted stochastically on the basis of the value of the covariate (treatment group) or on the basis of previously observed response (MAR) after the first time point to create an incomplete data set. We assumed that measurements from each subject were obtained at the first assessment. We first imposed missing values on

measurements at the second time point (Y_2), then imposed missing values on those at the third time point (Y_3) for the remaining subjects and then imposed missing values on those at the last time point (Y_4) for the subjects who stayed in the study separately for each treatment group. Two separate scenarios were used for the MAR mechanism: The likelihood of missing values on the response is related to the treatment group for the first scenario and is correlated with the previous response prior to dropping out for the second scenario.

For MCAR, the probability of drop-out at time $t = 2,3,4$ did not depend on any variables or responses. After generating a complete data set, data were randomly deleted at rates of 10%, 15%, and 25% for the 2^{nd}, 3^{rd}, and 4^{th} time points, respectively. The subjects were equally likely to drop out of the study for both treatment groups.

Let Y_{it} be the observed binary response for subject $i = 1, \dots, N$ and occasion $t = 1, 2, 3, 4$. $M_{it} = 0$ denotes a missing response and $M_{it} = 1$ denotes an observed response.

$$Pr(M_{i2} = 0 | M_{i1} = 1) = 0.10$$

$$Pr(M_{i3} = 0 | M_{i2} = 1) = 0.15$$

$$Pr(M_{i4} = 0 | M_{i3} = 1) = 0.25$$

For the first scenario of MAR (MAR-1), rates of missingness are different for each treatment group at the 2^{nd}, 3^{rd}, and 4^{th} time points. Unequal drop-out rates in treatment and placebo groups were assumed. This MAR situation represents a reasonable scenario where missing data might predominantly occur in one treatment group relative to other group. It illustrates that more subjects in the placebo group than the treatment group drop out the study due to a lack of efficacy of the treatment. It results in different rates of missingness between the treatment groups. Thus, we assumed that proportions of withdrawals are higher in *Trt*-A than *Trt*-B at each time point. Rates of missigness of 10%, 15%, and 25% for the placebo group and

6%, 10%, and 16% for the treatment group at the 2^{nd}, 3^{rd}, and 4^{th} time points, respectively were considered.

$$Pr(M_{i2} = 0|M_{i1} = 1) = \begin{cases} 0.10, & placebo \\ 0.06, & treatment \end{cases}$$

$$Pr(M_{i3} = 0|M_{i2} = 1) = \begin{cases} 0.15, & placebo \\ 0.10, & treatment \end{cases}$$

$$Pr(M_{i4} = 0|M_{i3} = 1) = \begin{cases} 0.25, & placebo \\ 0.16, & treatment \end{cases}$$

For the second scenario of MAR (MAR-2), the likelihood of missingness occurring at a given time point was dependent on the previous response. The same dropout probabilities are used for both groups. The probability of drop-out at each time points after the first time point were gradually increased by 10%, 15%, and 25%, respectively. The marginal probabilities at four time points are $p_A = (p_{A1}, p_{A2}, p_{A3}, p_{A4})$ and $p_B = (p_{B1}, p_{B2}, p_{B3}, p_{B4})$ for *Trt*-A and *Trt*-B, respectively. The dropout model was specified as the probability of withdrawal at time point *j* given the response at the previous time point $j-1$ is modeled using a logistic regression in the form of

$$logit\left[Pr\left(M_{ij} = 0 \middle| M_{i,j-1} = 1\right)\right] = \tau_{j-1} + \tau_y Y_{i,j-1} \; for \; j = 2, 3, 4$$

From the above model, we calculated the intercept given the dependency of the missing value (τ_y) and the probability of positive response at the previous time point for $j = 2, 3, 4$.

$$\tau_{j-1} = logit\left[Pr\left(M_{ij} = 0 \middle| M_{i,j-1} = 1\right)\right] - \tau_y Y_{i,j-1}$$

We set the equal rates of dropout at the 2^{nd}, 3^{rd}, and 4^{th} time points for both treatment groups at 10%, 15%, and 25%, respectively. The missing pattern follows from the equations.

$$logit[Pr(M_{i2} = 0|M_{i1} = 1)] = \begin{cases} -2.197 + \tau_y p_{A1}, & placebo \\ -2.197 + \tau_y p_{B1}, & treatment \end{cases}$$

$$logit[Pr(M_{i3} = 0|M_{i2} = 1)] = \begin{cases} -1.735 + \tau_y p_{A2}, & placebo \\ -1.735 + \tau_y p_{B2}, & treatment \end{cases}$$

$$logit[Pr(M_{i4} = 0|M_{i3} = 1)] = \begin{cases} -1.099 + \tau_y p_{A3}, & placebo \\ -1.099 + \tau_y p_{B3}, & treatment \end{cases}$$

τ_y = 0.8 was chosen for strong dependency of drop-out on the previous response in our simulation runs. Non-response rates for each treatment group were calculated from the missing data pattern above given the dependence of responses to missingness (τ_y) and marginal probabilities p_A and p_B to impose missing data mechanism on the complete data.

3.4.3 Assigning Scores to the Vectors with Missing Data

We considered the same strategies for ordering the observed vectors without missing data as shown in Table 2.1 in *Section* 2.2.4. We used these two strategies to assign scores to the vectors with missing data. Details regarding the score assignment to the vectors with missing data were given in *Section* 3.2.1.1. Assigned scores to the possible vectors with missing data, which are rows of matrix $\boldsymbol{M}$ in (3.5), based on strategy I and strategy II are shown in Table 3.1.

3.4.4 Simulation Results

The simulation results for the Type I error rate are reported in Table 3.2, Table 3.3, and Table 3.4 and the simulation results for the power are reported in Table 3.5, Table 3.6, and Table 3.7 when the missing data mechanism is MCAR, MAR-1, and MAR-2, respectively. Results comparing the proposed method with GEE and MERI methods in terms of power are presented in Table 3.8 when the time by treatment interaction term was included under the assumption of MAR-2 missing data mechanism and the assumption of AR(1) correlation structure.

Table 3.1: Assigned scores to the vectors with missing data

	Rank Score	
	Strategy	
Profile	**I**	**II**
[1 1 1 •]	1.4	2.6
[1 1 0 •]	4.5	7.5
[1 0 1 •]	5.5	6.5
[1 0 0 •]	10.4	12.2
[0 1 1 •]	6.6	4.8
[0 1 0 •]	11.5	10.5
[0 0 1 •]	12.5	9.5
[0 0 0 •]	15.6	14.4
[1 1 • •]	2.64	4.56
[1 0 • •]	7.95	9.35
[0 1 • •]	9.05	7.65
[0 0 • •]	14.36	12.44
[1 • • •]	5.3	6.96
[0 • • •]	11.71	10.1

(1 = positive, 0 = negative, and • = missing)

Type I Error Rate

The type I error rates of the proposed method and the competing methods are given in Table 3.2 under different selection scenarios of marginal expectations at four time points and using three different correlation structures when the missing data mechanism was assumed to be MCAR. Under a variety of scenarios using different correlation structures, none of the methods differed substantially from the type I error rate except the proposed logistic regression substitution approach which noticeably inflated the type I error rate in all cases regardless of the missing data mechanism. When we use this approach to estimate the probability that an outcome is positive at a given time point, conditional logistic regression model excludes subjects missing the observation at that point and is likely to produce biased estimates of treatment effects. Therefore, the probabilities in equations (3.6) and (3.7) may not properly estimated and the type I error rates

produced by this approach are greatly inflated. Since the logistic regression substitution approach did not control the type I error rate at the nominal level of 0.05, we excluded this approach in the comparison of the methods, therefore, when we use the term “proposed method”, it will refer to “crude substitution approach”.

The type I error rates for the proposed approach with the two different strategies (I and II) are close to or around the nominal 5 percent level regardless of the different correlation structures. The type I error rates for MERI are inflated when exchangeable or AR(1) correlation structure was assumed. In some cases, GEE with unstructured correlation structure yielded slightly higher type I error rate than the nominal level of 0.05. When compared to other methods, the proposed method (using both strategies) produced lower type I error rate in almost all cases.

Under MAR-1, all methods produced desirable type I error rate around the nominal level when independent correlation structure was assumed, as reported in Table 3.3. Similar to the MCAR mechanism, the type I error rate for MERI was inflated when an exchangeable or AR(1) correlation structure was assumed. As explained in *Section* 2.2.3, MERI produces higher type I error rate as correlation among the observations within subject increases. GEE produced a slightly higher type I error rate in the scenario of $p_A = p_B = (.6, .4, .6, .4)$ when the correlation structures were assumed to be exchangeable and AR(1). The proposed method exhibit inflation of the type I error rates for $p_A = p_B = (.6, .7, .8, .9)$ under the all three correlation structure.

Under MAR-2, as shown in Table 3.4, simulation results demonstrated that all methods controlled the type I error rate at the nominal level, even the MERI method which inflated the type I error rate under MCAR and MAR-1. The type I error rates for all methods were lower when MAR-2 was assumed compared to MAR-1 and MCAR. The type I error rate for the proposed method with strategy II was slightly lower than that for the other methods in all

scenarios. The type I error rates for the proposed method with strategy I and other methods were comparable.

For the scenario of $p_A = p_B = (.6, .7, .8, .9)$, a high type I error rate was produced by the proposed method (strategy I and especially strategy II) in the situation where the missing data mechanism was assumed to be MAR-1 while much lower type I error rate was produced under MCAR and MAR-2.

Based on these results, we can conclude that the proposed method with crude substitution approach controlled the type I error rate around nominal level for almost all scenarios in the presence of missing data.

Statistical Power

The logistic substitution approach artificially yielded high power under the alternative hypothesis at the price of a largely inflated type I error rate under the null hypothesis in all scenarios (different marginal probabilities and different correlation structures) regardless of the missing data mechanism. Therefore, this method was excluded when the methods were further discussed in terms of power.

The statistical powers of the methods under MCAR are presented in Table 3.5. In the situation where $p_A = (.3, .5, .6, .8)$ and $p_B = (.8, .6, .5, .3)$ and different correlation structures were assumed, the proposed method with strategy II dramatically reduced power while the other methods especially the *t*-test and the proposed method with strategy I, which yielded the highest power compared to other methods, Under MCAR, the *t*-test tended to reduce power more than the other methods for almost all scenarios except those where $p_A = (.7, .7, .7, .7)$ and $p_B = (.5, .5, .5, .5)$, $p_A = (.5, .5, .5, .5)$ and $p_B = (.3, .3, .3, .3)$, and $p_A = (.7, .6, .5, .4)$ and

Table 3.2: Type I Error (%): Simualtion results for comparing the methods with missing data. MCAR, missingness with equal dropout rates for both groups

Dropout rates are equal for both groups: (0, 10%, 15%, 25%) n = (60, 54, 46, 34)												
	Placebo	**Drug**	**Crude**		**Logistic**		***t*-test**	**GEE**				**MERI**
	$p_1\ p_2\ p_3\ p_4$	$p_1\ p_2\ p_3\ p_4$	**I**	**II**	**I**	**II**		**IND**	**EX**	**AR(1)**	**UN**	
Independent Correlation Structure												
1	.3 .3 .3 .3	.3 .3 .3 .3	4.8	4.2	9.2	20.2	5.0	5.4	5.4	5.2	5.9	5.0
2	.4 .4 .4 .4	.4 .4 .4 .4	5.6	5.1	10.0	19.1	5.0	5.9	6.1	6.2	6.6	5.4
3	.5 .5 .5 .5	.5 .5 .5 .5	4.6	4.1	8.6	18.4	4.6	4.6	4.6	4.6	5.0	4.0
4	.7 .7 .7 .7	.7 .7 .7 .7	4.6	4.2	10.5	20.9	4.6	6.2	6.0	6.2	6.5	5.5
5	.6 .4 .6 .4	.6 .4 .6 .4	5.0	4.6	9.7	18.6	5.8	5.3	5.4	5.3	6.0	4.4
6	.4 .5 .6 .7	.4. 5 .6 .7	5.6	4.3	9.8	18.8	5.5	5.6	5.7	5.7	6.4	5.2
7	.6 .7 .8 .9	.6 .7 .8 .9	4.2	2.4	6.8	15.1	4.8	4.6	4.8	4.6	5.2	4.5
8	.7 .6 .5 .4	.7 .6 .5 .4	4.6	5.2	8.8	17.6	4.2	4.8	4.8	4.8	5.2	4.4
9	.6 .6 .4 .4	.6 .6 .4 .4	4.6	4.4	9.4	18.0	5.6	5.6	5.4	5.6	5.8	4.9
Exchangeable Correlation Structure												
1	.3 .3 .3 .3	.3 .3 .3 .3	4.6	4.6	9.4	14.8	5.1	5.8	5.8	5.8	6.2	6.8
2	.4 .4 .4 .4	.4 .4 .4 .4	5.9	5.8	9.9	14.8	5.8	6.4	6.0	6.0	6.3	7.0
3	.5 .5 .5 .5	.5 .5 .5 .5	4.7	4.8	9.4	13.6	4.6	5.6	5.0	5.2	5.4	6.2
4	.7 .7 .7 .7	.7 .7 .7 .7	3.4	2.4	8.4	15.6	4.7	5.2	4.6	4.8	5.4	5.6
5	.6 .4 .6 .4	.6 .4 .6 .4	5.3	4.6	9.1	12.0	5.3	5.0	5.2	5.2	5.8	6.1
6	.4 .5 .6 .7	.4. 5 .6 .7	4.9	4.6	9.4	13.6	4.9	5.8	5.2	5.3	6.4	6.6
7	.6 .7 .8 .9	.6 .7 .8 .9	3.6	2.5	6.2	11.4	4.7	5.0	5.2	5.4	5.6	6.4
8	.7 .6 .5 .4	.7 .6 .5 .4	5.1	5.4	9.6	14.7	5.2	5.6	5.3	5.4	5.6	6.4
9	.6 .6 .4 .4	.6 .6 .4 .4	4.8	4.4	9.2	14.4	4.8	5.1	5.0	5.0	5.2	6.1
Auto-Regressive Correlation Structure												
1	.3 .3 .3 .3	.3 .3 .3 .3	3.8	3.6	8.2	13.8	4.8	5.3	5.2	5.4	5.6	6.2
2	.4 .4 .4 .4	.4 .4 .4 .4	4.4	4.8	8.8	14.0	4.9	5.4	5.0	4.6	5.0	6.3
3	.5 .5 .5 .5	.5 .5 .5 .5	5.5	4.8	8.5	13.7	5.5	5.6	5.0	5.3	5.7	6.3
4	.7 .7 .7 .7	.7 .7 .7 .7	5.2	4.4	8.2	12.8	5.2	5.4	5.6	5.4	6.0	6.8
5	.6 .4 .6 .4	.6 .4 .6 .4	4.7	5.0	8.6	12.6	4.5	4.8	4.9	4.9	5.2	5.9
6	.4 .5 .6 .7	.4. 5 .6 .7	5.2	5.1	8.6	12.6	5.3	5.4	5.6	5.8	6.2	7.0
7	.6 .7 .8 .9	.6 .7 .8 .9	3.6	2.2	6.0	8.8	4.4	4.0	4.2	4.5	5.1	5.0
8	.7 .6 .5 .4	.7 .6 .5 .4	5.6	5.1	10.6	13.6	5.4	5.6	5.6	5.8	6.2	7.2
9	.6 .6 .4 .4	.6 .6 .4 .4	4.0	4.4	8.8	13.9	3.9	4.1	4.6	4.4	4.7	5.1

Abbreviations: Crude, proposed method by applying "crude substitution approach" for score assignment to the vectors with missing data based on the ordering of strategies I and II in Table 2.1; Logistic, proposed method by applying "logistic regression approach" for score assignment to the vectors with missing data based on the ordering of strategies I and II in Table 2.1; *t*-test, *t*-test comparing the means of the proportions of positive responses between the two groups; GEE, logistic regression model using GEE; MERI, mixed effects logistic regression model with random intercept; IND, independent correlation structure; EX, exchangeable correlation structure; AR(1), first-order auto-regressive correlation structure; UN, unstructured correlation structure; p_t, marginal probability of positive response at time t, t = 1, 2, 3, 4.

Table 3.3: Type I Error Rate (%): Simulation results for comparing the methods with missing data. MAR-1, missingness with different dropout rates for each group

Dropout rates: Placebo: (0, 10%, 15%, 25%) n = (60, 54, 46, 34)												
Drug : (0, 6%, 10%, 1 n = (60, 56, 50, 42)												
	Placebo	**Drug**	**Crude**		**Logistic**		***t*-test**	**GEE**				**MERI**
	p_1 p_2 p_3 p_4	**p_1 p_2 p_3 p_4**	**I**	**II**	**I**	**II**		**IND**	**EX**	**AR(1)**	**UN**	
Independent Correlation Structure												
1	.3 .3 .3 .3	.3 .3 .3 .3	5.4	5.4	8.5	15.2	5.6	5.5	5.4	5.5	5.9	5.0
2	.4 .4 .4 .4	.4 .4 .4 .4	5.0	5.9	8.6	16.1	4.8	5.2	5.2	5.2	5.9	4.6
3	.5 .5 .5 .5	.5 .5 .5 .5	5.0	4.5	8.2	14.5	4.9	4.8	4.8	5.0	5.1	4.4
4	.7 .7 .7 .7	.7 .7 .7 .7	5.1	5.6	7.8	17.0	4.8	4.7	4.8	4.8	5.2	4.2
5	.6 .4 .6 .4	.6 .4 .6 .4	4.5	5.0	7.6	14.2	4.8	4.8	4.8	5.0	5.1	4.4
6	.4 .5 .6 .7	.4. 5 .6 .7	4.6	4.8	7.8	14.3	4.0	4.4	4.2	4.4	5.0	4.2
7	.6 .7 .8 .9	.6 .7 .8 .9	6.4	7.2	7.0	13.2	6.0	5.6	5.6	5.7	5.8	5.3
8	.7 .6 .5 .4	.7 .6 .5 .4	5.4	5.4	9.0	14.4	5.0	5.3	5.4	5.4	6.0	4.9
9	.6 .6 .4 .4	.6 .6 .4 .4	5.4	5.2	8.2	15.4	5.6	5.5	5.4	5.6	6.0	5.4
Exchangeable Correlation Structure												
1	.3 .3 .3 .3	.3 .3 .3 .3	4.6	5.2	7.8	10.6	4.7	4.2	4.8	4.0	5.4	5.5
2	.4 .4 .4 .4	.4 .4 .4 .4	5.0	4.7	7.8	11.6	5.2	5.2	5.1	4.9	5.7	6.2
3	.5 .5 .5 .5	.5 .5 .5 .5	4.6	4.4	7.9	11.8	4.6	4.6	4.3	4.5	5.0	5.6
4	.7 .7 .7 .7	.7 .7 .7 .7	5.9	6.2	9.0	14.2	5.6	6.4	6.0	6.4	6.5	7.0
5	.6 .4 .6 .4	.6 .4 .6 .4	5.8	5.6	9.2	13.2	6.0	6.2	6.1	6.2	6.6	7.2
6	.4 .5 .6 .7	.4. 5 .6 .7	5.0	5.2	8.0	10.6	4.5	5.0	4.6	4.8	5.2	5.6
7	.6 .7 .8 .9	.6 .7 .8 .9	6.4	6.9	7.6	10.9	4.8	5.5	5.8	5.8	6.0	6.6
8	.7 .6 .5 .4	.7 .6 .5 .4	4.8	5.0	8.2	11.4	5.5	4.8	5.0	5.2	5.5	5.8
9	.6 .6 .4 .4	.6 .6 .4 .4	5.2	4.9	8.4	11.0	5.3	5.0	5.4	5.2	6.2	6.4
Auto-Regressive Correlation Structure												
1	.3 .3 .3 .3	.3 .3 .3 .3	6.0	5.8	7.8	12.2	4.5	5.6	5.1	5.0	5.4	6.2
2	.4 .4 .4 .4	.4 .4 .4 .4	4.8	4.4	6.8	10.0	4.7	4.6	4.7	4.4	5.2	5.8
3	.5 .5 .5 .5	.5 .5 .5 .5	5.8	5.2	8.0	11.2	5.4	5.6	5.6	5.2	5.8	6.8
4	.7 .7 .7 .7	.7 .7 .7 .7	5.5	5.2	8.2	10.9	4.8	5.1	5.0	5.2	5.5	6.5
5	.6 .4 .6 .4	.6 .4 .6 .4	5.6	5.5	8.6	12.2	5.6	6.4	6.2	6.0	6.5	7.5
6	.4 .5 .6 .7	.4. 5 .6 .7	5.2	5.6	7.8	10.4	5.4	5.4	5.2	5.5	6.2	6.6
7	.6 .7 .8 .9	.6 .7 .8 .9	6.6	7.8	8.1	9.8	6.2	5.3	5.4	5.8	6.6	7.2
8	.7 .6 .5 .4	.7 .6 .5 .4	5.4	5.6	8.8	11.2	5.0	5.3	5.3	5.2	6.3	6.5
9	.6 .6 .4 .4	.6 .6 .4 .4	4.4	5.2	7.2	11.6	4.7	4.8	4.5	4.6	5.0	5.4

Abbreviations are same as in Table 3.2

Table 3.4: Type I Error Rate (%): Simulation results for comparing the methods with missing data. MAR-2, missingness with same dropout rates for each group

Dropout rates are equal for both groups: (0, 10%, 15%, 25%) n = (60, 54, 46, 34)												
	Placebo	**Drug**	**Crude**		**Logistic**		***t*-test**	**GEE**				**MERI**
	p_1 p_2 p_3 p_4	**p_1 p_2 p_3 p_4**	**I**	**II**	**I**	**II**		**IND**	**EX**	**AR(1)**	**UN**	
Independent Correlation Structure												
1	.3 .3 .3 .3	.3 .3 .3 .3	3.0	2.0	7.8	15.0	4.1	3.2	3.2	3.2	3.6	2.8
2	.4 .4 .4 .4	.4 .4 .4 .4	4.4	3.4	8.4	14.2	4.5	4.0	4.1	4.1	4.0	3.6
3	.5 .5 .5 .5	.5 .5 .5 .5	4.2	3.0	7.4	14.6	4.2	3.5	3.6	3.6	3.7	3.2
4	.7 .7 .7 .7	.7 .7 .7 .7	3.0	2.9	7.2	16.0	3.8	3.6	3.6	3.6	3.7	3.2
5	.6 .4 .6 .4	.6 .4 .6 .4	4.4	3.2	8.4	14.0	4.2	4.4	4.4	4.4	4.4	3.4
6	.4 .5 .6 .7	.4. 5 .6 .7	4.8	3.2	10.0	16.9	4.6	4.5	4.4	4.3	4.6	3.5
7	.6 .7 .8 .9	.6 .7 .8 .9	2.9	1.0	6.8	14.2	4.1	4.5	4.8	4.8	4.9	4.4
8	.7 .6 .5 .4	.7 .6 .5 .4	4.0	3.4	8.4	13.6	3.6	4.0	4.0	3.9	3.9	3.6
9	.6 .6 .4 .4	.6 .6 .4 .4	4.0	3.0	9.0	14.4	4.2	3.6	4.0	3.6	4.2	3.3
Exchangeable Correlation Structure												
1	.3 .3 .3 .3	.3 .3 .3 .3	3.5	3.4	7.0	11.0	5.0	3.8	4.2	4.2	4.4	5.0
2	.4 .4 .4 .4	.4 .4 .4 .4	4.2	4.0	8.0	10.2	4.6	4.4	5.0	4.9	5.4	6.0
3	.5 .5 .5 .5	.5 .5 .5 .5	4.6	4.3	9.0	10.8	5.4	4.7	5.2	4.8	5.4	6.0
4	.7 .7 .7 .7	.7 .7 .7 .7	3.0	2.4	8.6	14.4	5.2	5.2	5.2	5.2	5.4	5.6
5	.6 .4 .6 .4	.6 .4 .6 .4	4.4	3.5	8.1	11.6	4.7	4.0	4.4	4.0	4.5	5.4
6	.4 .5 .6 .7	.4. 5 .6 .7	4.8	3.5	9.2	11.9	4.5	5.0	4.8	4.8	5.2	6.0
7	.6 .7 .8 .9	.6 .7 .8 .9	3.5	1.6	7.2	12.0	4.8	5.6	5.7	5.3	6.0	6.8
8	.7 .6 .5 .4	.7 .6 .5 .4	3.5	3.2	8.2	10.0	4.8	3.6	4.4	4.2	4.8	5.2
9	.6 .6 .4 .4	.6 .6 .4 .4	4.2	3.0	8.2	10.0	4.0	3.3	4.4	3.8	4.4	4.9
Auto-Regressive Correlation Structure												
1	.3 .3 .3 .3	.3 .3 .3 .3	3.2	3.0	6.3	10.0	4.3	3.4	3.9	4.2	4.2	4.6
2	.4 .4 .4 .4	.4 .4 .4 .4	3.4	2.8	6.8	10.6	4.4	3.4	4.2	4.1	4.6	5.2
3	.5 .5 .5 .5	.5 .5 .5 .5	4.0	3.6	7.0	10.2	3.8	3.9	4.1	4.2	4.4	4.8
4	.7 .7 .7 .7	.7 .7 .7 .7	4.0	3.6	7.4	10.4	5.6	4.6	5.2	4.8	5.8	6.2
5	.6 .4 .6 .4	.6 .4 .6 .4	4.4	4.2	6.8	9.8	4.8	4.2	4.2	4.0	4.4	5.2
6	.4 .5 .6 .7	.4. 5 .6 .7	4.3	3.6	7.0	11.4	4.4	4.2	4.3	4.4	4.6	5.3
7	.6 .7 .8 .9	.6 .7 .8 .9	3.4	1.9	6.8	8.7	4.9	4.2	4.5	4.4	5.0	5.9
8	.7 .6 .5 .4	.7 .6 .5 .4	4.0	3.7	8.3	9.4	4.8	4.1	4.8	4.8	5.1	5.4
9	.6 .6 .4 .4	.6 .6 .4 .4	4.2	3.3	8.0	10.6	3.8	4.4	4.1	4.1	4.6	5.0

Abbreviations are same as in Table 3.2.

$p_B = (.4, .4, .4, .4)$ regardless of assumption of correlation structures when compared to the results in Table 2.2.

When the marginal probabilities were $p_A = (.4, .4, .4, .4)$ and $p_B = (.3, .3, .3, .3)$ with AR(1) correlation structure and the marginal probabilities were $p_A = (.8, .8, .8, .8)$ and

$p_B = (.7, .7, .7, .7)$ assuming both exchangeable and AR(1) correlation structures, the proposed approach with strategy II had slightly lower power than the other methods whereas it yielded statistical power comparable to other procedures in the absence of missing data (Table 2.2). In view of the results on Table 2.2, the results produced under MCAR (Table 3.5) were parallel in terms of comparisons of the power of the methods and the decrease in power for the proposed method with strategy II was less than the decrease in power for other methods.

The results are reported in Table 3.6 when the dropout rates were different in each treatment group (MAR-1). We compared the results under MAR-1 (Table 3.6) to those under MCAR (Table 3.5). There were some scenarios where the results under MCAR were different than those under MAR-1.

For the scenario of $p_A = (.6, .7, .8, .9)$ and $p_B = (.6, .6, .6, .6)$ assuming independent and AR(1) correlation structures, the power of the proposed method with strategy II reduced while GEE and MERI methods increased power under MAR-1 compared with MCAR and hence, the proposed method with strategy II, GEE, and MERI are comparable in terms of power. For the same scenario with the assumption of exchangeable correlation structure, the proposed method with strategy II demonstrated lower power than GEE and MERI under MAR-1 whereas it had higher power than other methods under MCAR. For the scenario where $p_A = (.7, .7, .8, .9)$ and $p_B = (.7, .7, .6, .5)$, the proposed method with strategy II yielded lower power when independent correlation structure was assumed and it yielded similar or slightly higher power compared with

GEE and MERI methods when exchangeable and AR(1) correlation structure were employed under MAR-1. The proposed method with strategies I and II produced higher power when the marginal probabilities were $p_A = (.4, .4, .4, .4)$ and $p_B = (.3, .3, .3, .3)$ and produced lower power when the marginal probabilities were $p_A = (.8, .8, .8, .8)$ and $p_B = (.7, .7, .7, .7)$

compared with other methods regardless of correlation structures. In other scenarios, results under MAR-1 were similar to those under MCAR. *t*-test tended to produce similar or higher power under MAR-1 compared to MCAR in all scenarios except p_A = (.3, .5, .6, .8) and p_B = (.8, .6, .5, .3) in which it produced lower power.

The results are demonstrated in Table 3.7 when the missing data mechanism was assumed to be MAR-2. The results under MAR-2 were very similar to those under MCAR. All methods produced very close statistical power in detecting the treatment difference under the alternative hypothesis when both missing data mechanism of MCAR and MAR-2 were considered.

Table 3.8 reported the simulation results to compare the proposed method with GEE and MERI approaches when the treatment by time interaction term is included for applying these methods. As expected and observed in *Section* 2.2.5, GEE and MERI methods achieved more power in some scenarios (i.e., p_A= (.3, .5, .6, .8) and p_B = (.8, .6, .5, .3), p_A= (.4, .5, .6, .7) and p_B = (.4, .4, .4, .4), p_A= (.6, .6, .6, .6) and p_B = (.6, .5, .4, .3)) compared to results in Table 3.7. Under these scenarios, powers of these methods are comparable with the proposed method when independent correlation structure is used and are higher than that of the proposed method in the presence of correlation among observations. For some scenarios (i.e., p_A= (.4, .4, .4, .4) and p_B = (.3, .3, .3, .3), p_A= (.8, .8, .8, .8) and p_B = (.7, .7, .7, .7), p_A= (.7, .7, .7, .7) and p_B = (.5, .5, .5, .5)), there is a decrease in power for GEE and MERI methods compared to results in Table 3.7.

Table 3.5: Power (%): Simulation results for comparing the methods with missing data. MCAR, missingness with same dropout rates for each group

	Placebo	Drug	Crude		Logistic		*t*-test	GEE				MERI
	p_1 p_2 p_3 p_4	p_1 p_2 p_3 p_4	I	II	I	II		IND	EX	AR(1)	UN	
Independent Correlation Structure												
1	.3 .5 .6 .8	.8 .6 .5 .3	85.4	6.0	60.0	72.2	64.2	31.4	29.4	32.0	36.2	28.0
2	.4 .6 .7. 8	.4 .4 .4 .4	86.1	99.6	96.5	100	86.5	97.6	97.6	97.6	97.1	97.1
3	.4 .5 .6 .7	.4 .4 .4 .4	47.0	82.5	70.4	96.3	50.9	71.4	71.6	71.4	71.0	69.2
4	.7 .6 .5 .4	.4 .4 .4 .4	96.6	78.1	96.3	77.2	92.4	92.8	93.0	92.6	92.4	92.7
5	.6 .7 .8 .9	.6 .6 .6 .6	47.2	87.4	71.8	98.3	53.8	76.6	76.8	76.4	77.2	74.7
6	.8 .7 .7. 6	.8 .6 .5 .4	33.4	51.6	53.4	79.4	32.9	45.8	45.7	45.9	46.6	44.3
7	.5 .6 .6. 7	.5 .4 .4 .3	75.7	97.4	92.1	100	77.0	93.1	93.0	93.1	93.0	92.7
8	.7 .7 .8 .9	.7 .7 .6 .5	23.6	53.2	44.4	86.6	28.4	42.0	42.2	41.9	43.4	40.5
9	.6 .6 .6 .6	.6 .5 .4. 3	46.6	83.1	70.8	97.2	50.2	70.2	70.6	70.2	70.6	68.4
10	.4 .4 .4 .4	.3 .3 .3 .3	49.3	47.0	62.6	68.2	46.9	55.1	55.3	55.4	55.6	53.8
11	.8 .8 .8 .8	.7 .7 .7 .7	52.6	47.6	69.2	75.9	53.3	61.4	61.7	61.6	62.4	60.4
12	.7 .7 .7 .7	.5 .5 .5 .5	96.9	96.4	98.6	98.6	95.4	98.0	97.9	98.0	97.8	97.8
13	.5 .5 .5 .5	.3 .3 .3 .3	97.0	96.4	99.0	8.6	94.9	98.4	98.4	98.5	98.1	98.4
Exchangeable Correlation Structure												
1	.3 .5 .6 .8	.8 .6 .5 .3	65.2	5.8	39.9	48.8	46.8	20.4	27.6	27.3	37.2	29.4
2	.4 .6 .7. 8	.4 .4 .4 .4	69.2	93.6	86.2	98.9	75.5	87.4	84.8	84.0	84.4	87.7
3	.4 .5 .6 .7	.4 .4 .4 .4	34.0	62.2	53.0	83.8	39.8	51.0	48.1	50.2	50.5	52.6
4	.7 .6 .5 .4	.4 .4 .4 .4	84.0	54.0	83.0	55.6	81.8	75.1	79.9	78.4	80.4	82.2
5	.6 .7 .8 .9	.6 .6 .6 .6	34.2	66.8	57.4	92.8	42.6	57.6	50.1	53.9	52.4	59.0
6	.8 .7 .7. 6	.8 .6 .5 .4	38.6	55.5	56.2	76.7	34.8	42.0	44.2	41.1	44.2	43.9
7	.5 .6 .6. 7	.5 .4 .4 .3	54.7	84.8	74.1	96.2	59.7	71.8	69.4	68.2	70.0	72.1
8	.7 .7 .8 .9	.7 .7 .6 .5	27.2	67.8	54.0	94.8	32.8	47.4	42.6	49.2	49.4	48.2
9	.6 .6 .6 .6	.6 .5 .4. 3	38.5	67.3	57.2	85.8	38.3	49.2	46.0	47.1	47.9	48.3
10	.4 .4 .4 .4	.3 .3 .3 .3	31.1	30.4	40.6	46.1	31.7	33.6	33.8	32.8	34.2	36.3
11	.8 .8 .8 .8	.7 .7 .7 .7	39.4	33.8	54.8	60.1	39.4	40.1	40.6	40.2	42.0	46.0
12	.7 .7 .7 .7	.5 .5 .5 .5	87.8	85.5	92.0	92.0	84.4	86.0	87.2	85.8	87.2	89.2
13	.5 .5 .5 .5	.3 .3 .3 .3	85.0	83.8	89.8	89.6	85.6	86.1	86.9	86.7	87.7	88.3
Auto Regressive Correlation Structure												
1	.3 .5 .6 .8	.8 .6 .5 .3	70.8	7.0	46.1	52.4	50.9	20.6	31.4	28.6	35.6	33.0
2	.4 .6 .7. 8	.4 .4 .4 .4	47.6	86.7	71.2	97.2	53.2	69.2	65.6	68.6	68.9	68.2
3	.4 .5 .6 .7	.4 .4 .4 .4	19.2	49.8	35.3	78.0	22.0	31.4	29.5	32.8	33.1	31.6
4	.7 .6 .5 .4	.4 .4 .4 .4	73.3	39.4	70.0	44.0	68.3	60.3	67.1	67.2	70.6	68.9
5	.6 .7 .8 .9	.6 .6 .6 .6	19.2	50.0	36.1	78.6	24.7	34.8	28.8	36.2	37.3	35.7
6	.8 .7 .7. 6	.8 .6 .5 .4	50.9	65.4	66.5	80.6	46.4	55.4	57.6	52.2	55.9	57.8
7	.5 .6 .6. 7	.5 .4 .4 .3	53.5	87.8	74.4	97.2	60.2	74.6	70.0	71.4	72.5	73.6
8	.7 .7 .8 .9	.7 .7 .6 .5	24.4	64.4	49.7	91.4	31.4	44.4	40.8	49.7	52.2	44.4
9	.6 .6 .6 .6	.6 .5 .4. 3	48.2	77.1	68.0	92.0	50.6	63.2	59.4	57.4	59.0	63.4
10	.4 .4 .4 .4	.3 .3 .3 .3	31.8	28.6	41.8	47.2	33.0	34.0	34.6	34.6	35.7	38.3
11	.8 .8 .8 .8	.7 .7 .7 .7	37.6	34.2	50.0	52.6	39.0	40.4	40.8	40.0	41.8	45.4
12	.7 .7 .7 .7	.5 .5 .5 .5	85.5	83.0	90.2	90.2	85.6	85.6	86.6	85.4	86.8	88.2
13	.5 .5 .5 .5	.3 .3 .3 .3	83.6	81.9	89.9	89.8	83.9	85.2	85.6	85.6	86.8	87.9

Abbreviations are same as in Table 3.2.

Table 3.6: Power (%): Simulation results for comparing the methods with missing data. MAR-1, missingness with different dropout rates for each group

Dropout rates: Placebo: (0, 10%, 15%, 25%) Drug : (0, 6%, 10%, 16%)												
	Placebo	**Drug**	**Crude**		**Logistic**		***t*-test**	**GEE**				**MERI**
	p_1 p_2 p_3 p_4	p_1 p_2 p_3 p_4	**I**	**II**	**I**	**II**		**IND**	**EX**	**AR(1)**	**UN**	
Independent Correlation Structure												
1	.3 .5 .6 .8	.8 .6 .5 .3	83.8	10.2	60.4	70.2	50.8	23.5	21.3	24.2	28.5	20.0
2	.4 .6 .7. 8	.4 .4 .4 .4	88.8	96.4	97.2	100	89.8	98.7	98.8	98.8	98.3	98.4
3	.4 .5 .6 .7	.4 .4 .4 .4	51.0	86.7	71.8	96.8	54.4	77.0	76.8	76.8	77.0	75.0
4	.7 .6 .5 .4	.4 .4 .4 .4	97.4	78.0	97.2	74.4	93.8	93.3	93.2	93.5	93.0	93.0
5	.6 .7 .8 .9	.6 .6 .6 .6	41.2	80.0	71.1	99.0	56.0	82.0	82.0	81.8	81.5	80.0
6	.8 .7 .7. 6	.8 .6 .5 .4	32.8	52.8	49.2	76.2	46.0	48.0	47.9	47.8	49.2	46.0
7	.5 .6 .6. 7	.5 .4 .4 .3	78.9	98.5	91.8	99.8	83.6	94.6	94.4	94.5	94.2	94.2
8	.7 .7 .8 .9	.7 .7 .6 .5	17.6	42.2	41.4	83.6	33.1	47.4	48.0	47.6	48.1	45.4
9	.6 .6 .6 .6	.6 .5 .4. 3	52.8	87.6	71.4	96.0	61.6	76.6	76.8	76.5	76.0	75.2
10	.4 .4 .4 .4	.3 .3 .3 .3	58.2	59.2	63.4	68.2	48.8	55.3	55.6	55.6	95.8	54.4
11	.8 .8 .8 .8	.7 .7 .7 .7	42.0	29.4	65.7	69.6	57.2	65.0	64.9	65.2	65.2	63.5
12	.7 .7 .7 .7	.5 .5 .5 .5	96.2	94.9	98.7	98.5	95.8	98.4	98.3	98.4	98.4	98.2
13	.5 .5 .5 .5	.3 .3 .3 .3	97.9	98.0	98.9	98.7	95.8	98.4	98.4	98.4	98.4	98.3
Exchangeable Correlation Structure												
1	.3 .5 .6 .8	.8 .6 .5 .3	61.7	6.4	40.0	47.2	35.6	16.0	22.0	20.5	29.4	23.4
2	.4 .6 .7. 8	.4 .4 .4 .4	68.2	91.8	83.4	98.4	75.0	87.6	85.8	84.7	86.0	88.0
3	.4 .5 .6 .7	.4 .4 .4 .4	38.0	62.6	53.2	83.2	43.4	54.8	52.7	53.2	54.1	56.3
4	.7 .6 .5 .4	.4 .4 .4 .4	84.3	53.8	82.8	54.0	81.2	74.9	78.4	76.7	79.6	81.2
5	.6 .7 .8 .9	.6 .6 .6 .6	26.5	51.4	49.3	88.2	43.6	59.7	52.7	56.9	56.0	61.4
6	.8 .7 .7. 6	.8 .6 .5 .4	34.2	52.0	50.3	72.6	43.2	43.1	45.6	40.9	45.7	45.0
7	.5 .6 .6. 7	.5 .4 .4 .3	56.9	86.2	71.9	95.2	63.8	74.0	71.2	70.8	73.1	74.6
8	.7 .7 .8 .9	.7 .7 .6 .5	21.4	56.4	47.5	91.8	40.2	53.4	49.2	56.0	55.4	54.0
9	.6 .6 .6 .6	.6 .5 .4. 3	38.0	71.4	54.4	85.2	43.0	50.7	47.8	48.0	48.6	50.2
10	.4 .4 .4 .4	.3 .3 .3 .3	41.8	41.7	47.7	50.0	35.8	35.4	36.8	36.2	38.3	40.0
11	.8 .8 .8 .8	.7 .7 .7 .7	25.0	16.2	43.6	48.8	39.6	41.8	41.6	41.2	43.2	46.8
12	.7 .7 .7 .7	.5 .5 .5 .5	84.0	80.0	91.0	90.0	86.2	86.2	87.6	86.4	88.0	89.7
13	.5 .5 .5 .5	.3 .3 .3 .3	87.1	87.0	90.6	89.8	84.8	85.6	86.7	85.8	87.3	87.8
Auto Regressive Correlation Structure												
1	.3 .5 .6 .8	.8 .6 .5 .3	64.2	11.0	46.0	52.4	39.4	17.3	25.0	23.6	30.7	26.6
2	.4 .6 .7. 8	.4 .4 .4 .4	50.0	87.6	69.8	97.6	53.5	71.0	68.3	69.6	69.8	70.3
3	.4 .5 .6 .7	.4 .4 .4 .4	25.4	56.4	39.2	80.4	24.8	36.0	34.0	37.6	36.8	36.6
4	.7 .6 .5 .4	.4 .4 .4 .4	78.6	49.2	74.8	46.8	71.6	61.0	66.8	67.1	70.2	69.1
5	.6 .7 .8 .9	.6 .6 .6 .6	18.4	43.4	34.2	73.9	30.1	42.6	37.4	43.0	43.1	42.4
6	.8 .7 .7. 6	.8 .6 .5 .4	51.0	67.2	62.4	78.7	56.2	56.0	56.6	51.1	56.2	57.4
7	.5 .6 .6. 7	.5 .4 .4 .3	59.6	91.8	74.0	97.8	64.8	75.6	71.5	71.6	73.0	75.0
8	.7 .7 .8 .9	.7 .7 .6 .5	19.8	58.0	41.4	87.7	36.2	48.2	44.2	53.6	55.2	47.0
9	.6 .6 .6 .6	.6 .5 .4. 3	53.4	82.0	67.4	91.3	59.2	65.0	61.9	59.3	62.2	65.6
10	.4 .4 .4 .4	.3 .3 .3 .3	41.4	42.2	46.0	49.6	33.0	33.4	33.3	31.6	35.1	36.8
11	.8 .8 .8 .8	.7 .7 .7 .7	26.4	19.6	39.3	41.7	41.1	41.6	41.5	40.9	42.6	47.9
12	.7 .7 .7 .7	.5 .5 .5 .5	85.0	81.6	89.4	89.4	86.6	86.3	87.4	86.0	87.8	89.1
13	.5 .5 .5 .5	.3 .3 .3 .3	89.2	88.2	91.8	91.2	85.4	85.6	85.7	85.0	86.5	88.4

Abbreviations are same as in Table 3.2.

Table 3.7: Power (%): Simulation results for comparing the methods with missing data. MAR-2, missingness with same dropout rates for each group

Dropout rates are equal for both groups : (0, 10%, 15%, 25%)								n = (60, 54, 46, 34)				
	Placebo	**Drug**	**Crude**		**Logistic**		***t*-test**	**GEE**				**MERI**
	p_1 p_2 p_3 p_4	p_1 p_2 p_3 p_4	**I**	**II**	**I**	**II**		**IND**	**EX**	**AR(1)**	**UN**	
Independent Correlation Structure												
1	.3 .5 .6 .8	.8 .6 .5 .3	86.9	5.0	61.6	73.6	68.3	33.3	29.6	33.8	38.0	28.8
2	.4 .6 .7. 8	.4 .4 .4 .4	86.6	99.6	97.4	100	85.2	98.0	98.0	98.0	97.4	97.5
3	.4 .5 .6 .7	.4 .4 .4 .4	46.6	85.1	72.8	99.0	47.6	71.0	71.6	71.2	71.4	69.5
4	.7 .6 .5 .4	.4 .4 .4 .4	97.4	78.8	97.7	80.6	94.0	95.3	95.1	95.2	94.8	95.0
5	.6 .7 .8 .9	.6 .6 .6 .6	47.6	88.2	75.8	99.6	53.2	79.3	79.4	79.2	78.5	77.4
6	.8 .7 .7. 6	.8 .6 .5 .4	50.4	74.8	54.6	79.4	54.0	67.5	67.4	67.4	67.6	65.8
7	.5 .6 .6. 7	.5 .4 .4 .3	77.5	98.4	93.4	100	79.0	94.4	94.5	94.4	94.0	94.0
8	.7 .7 .8 .9	.7 .7 .6 .5	40.8	90.7	47.7	89.4	51.6	76.1	76.2	76.0	77.3	74.1
9	.6 .6 .6 .6	.6 .5 .4. 3	51.9	85.9	73.0	98.3	55.2	72.8	73.2	73.0	73.0	71.6
10	.4 .4 .4 .4	.3 .3 .3 .3	51.0	49.6	65.4	72.0	48.4	55.4	55.8	55.2	56.2	54.6
11	.8 .8 .8 .8	.7 .7 .7 .7	52.8	49.4	72.4	77.4	56.3	64.0	64.0	64.1	63.6	62.5
12	.7 .7 .7 .7	.5 .5 .5 .5	97.4	97.8	99.3	99.3	96.2	98.6	98.6	98.6	98.8	98.6
13	.5 .5 .5 .5	.3 .3 .3 .3	97.6	97.2	99.0	99.3	95.4	98.4	98.3	98.4	98.4	98.2
Exchangeable Correlation Structure												
1	.3 .5 .6 .8	.8 .6 .5 .3	66.2	4.7	38.6	53.0	47.6	18.4	27.7	26.0	38.9	29.5
2	.4 .6 .7. 8	.4 .4 .4 .4	67.2	93.6	87.5	99.2	71.8	87.8	84.4	84.0	84.0	86.8
3	.4 .5 .6 .7	.4 .4 .4 .4	31.9	58.8	53.4	86.0	36.0	48.6	45.4	46.8	47.0	50.0
4	.7 .6 .5 .4	.4 .4 .4 .4	84.6	52.9	84.5	57.4	80.4	76.6	80.8	78.8	81.3	83.0
5	.6 .7 .8 .9	.6 .6 .6 .6	30.8	61.7	57.0	93.8	42.1	56.5	49.6	53.4	50.0	57.2
6	.8 .7 .7. 6	.8 .6 .5 .4	38.8	57.8	56.4	79.0	38.6	42.9	46.7	41.6	46.6	45.1
7	.5 .6 .6. 7	.5 .4 .4 .3	54.0	85.8	73.7	96.8	56.9	71.5	68.4	67.9	70.0	70.8
8	.7 .7 .8 .9	.7 .7 .6 .5	28.2	69.2	57.7	97.2	37.5	51.4	47.8	54.0	52.6	51.9
9	.6 .6 .6 .6	.6 .5 .4. 3	36.0	69.6	56.7	87.8	38.1	46.5	44.8	45.0	46.0	47.2
10	.4 .4 .4 .4	.3 .3 .3 .3	31.8	30.2	40.2	45.9	32.1	32.1	33.4	32.5	34.3	36.2
11	.8 .8 .8 .8	.7 .7 .7 .7	38.6	33.2	58.1	64.1	41.2	42.9	43.8	43.2	44.4	48.6
12	.7 .7 .7 .7	.5 .5 .5 .5	86.8	85.2	93.2	93.4	85.1	85.8	87.2	85.8	86.8	89.2
13	.5 .5 .5 .5	.3 .3 .3 .3	87.3	87.1	91.8	93.0	86.7	88.0	88.9	87.8	89.0	90.3
Auto Regressive Correlation Structure												
1	.3 .5 .6 .8	.8 .6 .5 .3	68.9	6.4	41.9	54.0	46.8	19.6	30.2	27.4	33.0	31.0
2	.4 .6 .7. 8	.4 .4 .4 .4	48.6	87.4	72.5	98.2	51.8	69.4	64.4	67.6	67.4	67.5
3	.4 .5 .6 .7	.4 .4 .4 .4	19.2	52.1	37.0	84.2	21.0	29.6	28.0	32.0	31.4	31.2
4	.7 .6 .5 .4	.4 .4 .4 .4	72.2	39.3	70.7	42.2	68.0	61.2	68.9	68.0	70.0	69.9
5	.6 .7 .8 .9	.6 .6 .6 .6	20.0	50.8	40.0	84.0	27.2	38.4	32.1	39.2	39.1	37.8
6	.8 .7 .7. 6	.8 .6 .5 .4	50.9	67.0	67.4	81.8	48.4	55.3	58.2	52.8	56.2	57.8
7	.5 .6 .6. 7	.5 .4 .4 .3	54.0	89.8	74.0	97.3	59.0	73.6	68.0	69.0	70.4	72.3
8	.7 .7 .8 .9	.7 .7 .6 .5	22.0	66.4	49.8	93.2	31.2	46.2	42.4	52.4	54.3	45.0
9	.6 .6 .6 .6	.6 .5 .4. 3	49.3	77.8	66.5	92.0	52.6	62.2	58.4	56.9	58.6	62.2
10	.4 .4 .4 .4	.3 .3 .3 .3	31.2	28.1	40.1	46.1	31.6	32.6	33.4	33.2	33.8	37.1
11	.8 .8 .8 .8	.7 .7 .7 .7	34.8	32.1	48.8	52.8	39.4	38.5	40.2	39.8	41.7	45.2
12	.7 .7 .7 .7	.5 .5 .5 .5	87.6	85.9	92.2	91.4	87.4	87.9	88.4	87.6	88.6	89.7
13	.5 .5 .5 .5	.3 .3 .3 .3	83.3	82.2	88.8	90.0	82.7	84.6	85.4	84.8	85.6	87.4

Abbreviations are same as in Table 3.2.

Table 3.8: Power (%): Simulation results for comparing the methods with missing data and treatment by time interaction. MAR-2, missingness with same dropout rates for each group

Dropout rates are equal for both groups : (0, 10%, 15%, 25%) n = (60, 54, 46, 34)										
	Placebo	**Drug**	**PROPOSED**		***t*-test**	**GEE**				**MERI**
	p_1 p_2 p_3 p_4	p_1 p_2 p_3 p_4	**I**	**II**		**IND**	**EX**	**AR(1)**	**UN**	
Independent Correlation Structure										
1	.3 .5 .6 .8	.8 .6 .5 .3	87.4	4.0	65.0	100	100	100	100	100
2	.4 .6 .7. 8	.4 .4 .4 .4	87.5	99.4	87.2	99.8	99.8	99.8	99.6	99.8
3	.4 .5 .6 .7	.4 .4 .4 .4	47.6	85.5	45.4	87.6	87.8	87.8	88.5	87.4
4	.7 .6 .5 .4	.4 .4 .4 .4	97.0	78.6	93.2	95.6	95.6	95.6	95.6	95.7
5	.6 .7 .8 .9	.6 .6 .6 .6	49.8	88.1	55.4	95.7	95.8	95.8	96.0	95.8
6	.8 .7 .7. 6	.8 .6 .5 .4	50.7	73.8	53.7	70.2	70.5	70.2	71.2	70.4
7	.5 .6 .6. 7	.5 .4 .4 .3	77.2	97.8	78.4	98.7	98.6	98.6	98.7	98.6
8	.7 .7 .8 .9	.7 .7 .6 .5	37.5	89.7	51.2	98.4	98.4	98.4	98.4	98.2
9	.6 .6 .6 .6	.6 .5 .4. 3	48.9	85.4	52.7	88.2	88.0	88.0	88.0	87.7
10	.4 .4 .4 .4	.3 .3 .3 .3	50.8	48.2	45.8	42.8	42.8	43.0	44.0	41.2
11	.8 .8 .8 .8	.7 .7 .7 .7	55.1	48.6	56.4	53.2	53.1	53.0	53.4	51.8
12	.7 .7 .7 .7	.5 .5 .5 .5	97.6	96.9	95.8	96.6	96.6	96.6	96.4	96.2
13	.5 .5 .5 .5	.3 .3 .3 .3	97.1	97.6	95.4	96.2	96.1	96.0	96.0	95.8
Exchangeable Correlation Structure										
1	.3 .5 .6 .8	.8 .6 .5 .3	65.0	4.6	47.8	100	100	100	100	100
2	.4 .6 .7. 8	.4 .4 .4 .4	66.8	93.6	70.6	98.2	98.8	98.5	98.9	89.3
3	.4 .5 .6 .7	.4 .4 .4 .4	33.6	60.1	38.6	77.4	80.1	78.8	81.6	82.1
4	.7 .6 .5 .4	.4 .4 .4 .4	82.6	54.2	79.4	90.4	90.4	89.6	90.4	91.6
5	.6 .7 .8 .9	.6 .6 .6 .6	30.4	63.8	40.8	89.4	90.0	89.2	90.4	91.9
6	.8 .7 .7. 6	.8 .6 .5 .4	39.2	57.4	39.3	55.3	60.3	54.9	60.5	60.3
7	.5 .6 .6. 7	.5 .4 .4 .3	54.8	86.3	59.0	94.2	95.3	95.8	95.3	95.6
8	.7 .7 .8 .9	.7 .7 .6 .5	25.2	65.9	34.5	95.4	96.0	95.4	96.0	96.4
9	.6 .6 .6 .6	.6 .5 .4. 3	38.0	69.4	38.8	77.3	80.4	78.6	80.2	81.0
10	.4 .4 .4 .4	.3 .3 .3 .3	32.0	30.0	32.6	22.6	24.6	23.8	26.0	26.8
11	.8 .8 .8 .8	.7 .7 .7 .7	37.9	31.6	39.6	29.8	30.9	29.4	31.8	36.2
12	.7 .7 .7 .7	.5 .5 .5 .5	87.7	85.6	85.4	77.0	78.1	77.2	79.5	82.2
13	.5 .5 .5 .5	.3 .3 .3 .3	86.1	85.2	85.4	76.9	79.0	77.1	79.2	81.0
Auto Regressive Correlation Structure										
1	.3 .5 .6 .8	.8 .6 .5 .3	70.2	6.7	50.2	100	100	100	100	100
2	.4 .6 .7. 8	.4 .4 .4 .4	47.8	88.4	51.4	97.5	97.6	97.4	97.2	99.2
3	.4 .5 .6 .7	.4 .4 .4 .4	17.8	51.6	20.0	74.7	76.4	71.7	72.8	84.0
4	.7 .6 .5 .4	.4 .4 .4 .4	71.4	38.2	66.8	84.0	83.8	86.1	84.8	88.4
5	.6 .7 .8 .9	.6 .6 .6 .6	19.2	51.1	25.8	89.3	88.6	86.0	86.2	92.3
6	.8 .7 .7. 6	.8 .6 .5 .4	49.8	66.4	48.2	60.4	67.0	62.4	66.2	72.8
7	.5 .6 .6. 7	.5 .4 .4 .3	51.5	88.2	57.4	96.1	96.8	96.6	96.8	98.0
8	.7 .7 .8 .9	.7 .7 .6 .5	22.4	66.8	30.8	98.6	98.6	97.6	97.4	99.0
9	.6 .6 .6 .6	.6 .5 .4. 3	48.0	78.4	50.8	83.6	85.8	85.8	86.2	90.2
10	.4 .4 .4 .4	.3 .3 .3 .3	30.4	28.3	31.5	21.8	23.4	23.9	24.9	29.3
11	.8 .8 .8 .8	.7 .7 .7 .7	39.8	36.4	43.8	32.7	34.1	32.6	35.4	41.3
12	.7 .7 .7 .7	.5 .5 .5 .5	87.0	85.4	85.8	76.9	78.2	77.8	79.2	82.6
13	.5 .5 .5 .5	.3 .3 .3 .3	85.6	83.6	83.5	77.2	76.8	76.8	78.4	81.7

Abbreviations are same as in Table 3.2 except that Proposed, proposed method by applying "crude substitution approach" for score assignment to the vectors with missing data based on the ordering of strategies I and II.

3.5 CONCLUSION

We compared the performance of our proposed method to other methods for analysis of longitudinal binary data in the presence of missing data via simulation study. The type I error rate of the proposed method with logistic regression substitution was considerably inflated so this method is not preferable. The type I error rate for MERI method was slightly inflated when the correlation structure is assumed to be exchangeable or AR(1). Other methods produced type I error rate reasonably close to the nominal value of 0.05.

Under MCAR, MAR-1, and MAR-2 missing data mechanisms, simulation studies comparing the proposed method with other methods demonstrated that none of the approaches are uniformly better than the others in terms of power. All methods yielded very similar results for all scenarios under MCAR and MAR-2. Results obtained under MAR-1 were different for some scenarios. All method lost power in the presence of missing data but the lost in power is greater for *t*-test than other methods. Our proposed approach performs better/worse than the other approaches for some situations depending on strategies of ordering vectors and assigning scores to the vectors with missing data. The proposed method with crude substitution may be preferred method for some scenarios in which the proposed method gives higher empirical power and maintains control over the type I error rate.

In most situations, our proposed method is competitive with other methods when data are not complete. We believe that our approach will be adaptable to multiple outcomes (intervention with other treatments, adverse side effects, and therapy with multiple effects).

4.0 METHODS OF RANKING VECTORS OF OBSERVATIONS

Our proposed method is based on applying the Wilcoxon test statistic to vectors of repeated binary observations and events. The main requirement of applying the proposed method is to order vectors of observations in terms of meaningful importance. The ordering is determined by 'clinical relevance'. For some scenarios, clinically meaningful ordering of the vectors may be defined by a 'natural algorithm'; while for other scenarios, the ordering could be obtained from a group of clinicians.

If an agreement among clinicians with respect to ordering of the vectors of repeated observations is reached, it is reasonable to apply the proposed approach to the data to be analyzed. If there is no consensus among clinicians about orderings of vectors due to differences in opinion among clinicians, it might not be easy to interpret the results and to employ the proposed method in the analysis of the data (*Follmann et al*. 1992). On the other hand, it might be beneficial to obtain different rankings produced by a group of experts who have diverse backgrounds. In this case, data can be analyzed in different ways and a proper perspective can be obtained on the impact of variability of clinician assessment on study conclusions.

In this chapter, we discuss natural algorithm and clinicians' orderings with respect to obtaining the rankings of vectors of repeated observations. We apply the ordering algorithms to a clinical trial of the efficacy of treatment for children with acute otitis media (AOM). We measure the variability of orderings among clinicians and examine the effects of this heterogeneity in the analysis of the otitis media clinical trial data.

4.1 NATURAL ALGORITHM

To demonstrate achievement of ordering vectors of observations, we present an illustration with a clinical trial comparing two treatment groups (e.g. drug versus placebo) to evaluate the effect of antibiotic treatment for children with acute otitis media to prevent AOM and to clear Middle Ear Effusion (MEE). Binary outcomes (1 = AOM or MEE, 0 = Disease-free) are obtained from each child for four time points which result in 16 possible profiles over time.

We first rank the vectors according to the number of disease-free time points. A vector with 4 disease-free time points ranks higher than the one with 3 disease-free time points which ranks higher than a vector with 2 disease-free time points and so on. We assume, without loss of generality, that the lower rank score indicates poorer outcome. After applying the first criteria, 4 vectors with one disease-free time points, 6 vectors with two disease-free time points, and 4 vectors with three disease-free time points share separately the same tied rank scores. These tied ranks can be broken on the basis of time to occurrence of disease. We follow two different strategies to break ties.

Some trials expect to observe a rapid effect of antibiotic treatment and successful initial effect of therapy is important for these trials. Therefore, we break ties by considering earlier disease as a poorer outcome. Among the subjects who have the same number of time points with diseases, if the disease occurred earlier for the subject, then this subject has lower rank score than the one who had disease later (Table 4.1). For some trials, it can be important to examine subjects without disease at or near the end of the study so we can break ties by considering *later* diseases as poorer outcome. In this case, a subject who has disease earlier is assigned higher rank score than the one who has disease later (Table 4.2).

Table 4.1: Nested criteria for ordering of 16 possible outcomes, R=Rank

Number of disease-free time points									
0		1		2		3		4	
R	**Profile**	**R**	**Profile**	**R**	**Profile**	**R**	**Profile**	**R**	**Profile**
1	[1 1 1 1]	2	[1 1 1 0]	6	[1 1 0 0]	12	[1 0 0 0]	16	[0 0 0 0]
		3	[1 1 0 1]	7	[1 0 1 0]	13	[0 1 0 0]		
		4	[1 0 1 1]	8	[1 0 0 1]	14	[0 0 1 0]		
		5	[0 1 1 1]	9	[0 1 1 0]	15	[0 0 0 1]		
				10	[0 1 0 1]				
				11	[0 0 1 1]				
Algorithm:									
1. Rank first by number of disease-free time points									
2. Break ties by considering *earlier* disease as poorer outcome									

Lower rank is *poorer* outcome.

Table 4.2: Nested criteria for ordering of 16 possible outcomes, R: Rank

Number of disease-free time points									
0		1		2		3		4	
R	**Profile**	**R**	**Profile**	**R**	**Profile**	**R**	**Profile**	**R**	**Profile**
1	[1 1 1 1]	5	[1 1 1 0]	11	[1 1 0 0]	15	[1 0 0 0]	16	[0 0 0 0]
		4	[1 1 0 1]	10	[1 0 1 0]	14	[0 1 0 0]		
		3	[1 0 1 1]	9	[1 0 0 1]	13	[0 0 1 0]		
		2	[0 1 1 1]	8	[0 1 1 0]	12	[0 0 0 1]		
				7	[0 1 0 1]				
				6	[0 0 1 1]				
Algorithm:									
1. Rank first by number of disease-free time points									
2. Break ties by considering *later* disease as poorer outcome									

Note: Lower rank is *poorer* outcome.

Such a nested set of criteria should be very robust in assigning vector ranks and permits a priority to be assigned as to what type of differences are most important. Applying different

criteria generates different orderings. Note that a vector with 4 disease-free time points ranks as the best outcome and a vector with 4 disease episodes ranks as the worst outcome based on the both ordering criteria. Table 4.1 and Table 4.2 present ordering of possible outcomes based on the two different principles.

4.2 CLINICIANS' RANKINGS

While ordering the vectors of repeated binary observations can be obtained by natural algorithm as the one just described, another approach to rank the outcomes is to ask a group of people who are expert in the area of study to order the vectors. Experts' opinion and knowledge play an important role in the application of our proposed method. Experts can rank the vectors in order of clinically meaningful and relevant importance.

4.2.1 Guidelines for Clinicians

We collaborated with three clinicians who are knowledgeable about otitis media at Children's Hospital of Pittsburgh of University of Pittsburgh Medical Center (UPMC). We asked them 3 scenarios related to otitis media. The following instructions were provided to them and each clinician independently ranked the outcomes:

We are developing a statistical test to use in randomized clinical trials that will give more weight to clinical considerations. It is based on ranking a set of outcomes from lowest to highest in regard to which is "clinically the least or most desirable". We are giving you a scenario and we would like you to answer the given questions. Note that there is no correct answer and "ties" are allowed in ranking your set of outcomes i.e. two different outcomes might be considered as

clinically equivalent. This will be asked to a number of clinicians because the extent of clinical variability is one of the important issues for us.

Scenario 1: Children entering the study with MEE are given a treatment and followed for a period of time. Four measurements, one measurement/month, are taken for each patient over the four months. A determination is made at each visit whether child has effusion ("1") in either ear or no effusion ("0"). For example, [1 0 0 1] would indicate effusion at the 1st and 4th visit and no effusion at visit 2 and 3 for a patient. We have the following 16 possible outcomes for the 4 months of measurements which can be obtained from each patient at the end of the study. We would like you to rank these outcomes from 1 (the least desirable or worst) to 16 (the most desirable or best). Some of the outcomes might be clinically equivalent so that they might have the same ranking score. Is there any "algorithm" you used to rank outcomes?

Table 4.3: Possible outcomes with 4 time points

1	[1 1 1 1]	5	[0 1 1 1]	9	[0 1 1 0]	13	[0 1 0 0]
2	[1 1 1 0]	6	[1 1 0 0]	10	[0 1 0 1]	14	[0 0 1 0]
3	[1 1 0 1]	7	[1 0 1 0]	11	[0 0 1 1]	15	[0 0 0 1]
4	[1 0 1 1]	8	[1 0 0 1]	12	[1 0 0 0]	16	[0 0 0 0]

Thank you for your help.

4.2.2 Clinicians' Responses

All three clinicians responded to questions and returned their answers to us. We present the responses of clinicians. Table 4.4 demonstrates the considerations clinicians applied to order the 16 vectors for 4 measurements. They ordered the vectors based on these criteria. Scores assigned to the vectors by clinicians are shown in Table 4.5.

Table 4.4: Considerations applied by three clinicians for ordering the vectors

Clinician A
1. The longer time without relapse the better.
2. An ideal treatment should prevent relapse.
3. Consecutive months of MEE worse than MEE for the same number of months but separated by no MEE.
Clinician B
1. More visits with MEE worse.
2. Those with MEE at the last visit would be worse than those without MEE at the last visit within each number of visits with MEE.
Clinician C
1. The more "0" the better.
2. The less continuous the period of "1" the better.
3. Prefer to see resolution at last visit.

As presented in Table 4.4, they reached an agreement that more time points with diseases are worse outcome. They first ordered the vectors based on the number of time points with diseases. For example, all clinicians assigned the highest score to the vector with no disease and the lowest score to the vector with diseases over four time points if higher score is better. They followed different strategies to break ties for vectors which have the same number of time points with diseases. While clinician B and C ordered the vectors with three diseases (profiles 2-5) similarly, clinician A ordered them in a different way as shown in Table 4.5. For ordering vectors with one disease (profiles 12-15), clinician B and C considered later diseases as poorer outcome and ordered the vectors almost identical while clinician A regarded vectors with earlier disease as poorer outcome and ordered the vectors opposite way. Clinicians assigned the same scores to the vectors when they considered that they are indeterminate. For example, Clinician C assigned the same score to some of the vectors with 2 diseases (profiles 6, 8, 9, and 10).

Table 4.5: Rank scores assigned by clinicians

		Clinician (original)			Clinician (adjusted)		
	Profile	**A**	**B**	**C**	**A**	**B**	**C**
1	[1 1 1 1]	1	1	1	1	1	1
2	[1 1 1 0]	2	4	5	2	5	5
3	[1 1 0 1]	3	3	4	4	3.5	4
4	[1 0 1 1]	3	3	3	4	3.5	3
5	[0 1 1 1]	3	2	2	4	2	2
6	[1 1 0 0]	4	6	7	6	7	8.5
7	[1 0 1 0]	5	5	8	8	6	11
8	[1 0 0 1]	5	8	7	8	10	8.5
9	[0 1 1 0]	5	9	7	8	11	8.5
10	[0 1 0 1]	6	7	7	10	8.5	8.5
11	[0 0 1 1]	7	7	6	11	8.5	6
12	[1 0 0 0]	8	13	11	12	15	15
13	[0 1 0 0]	9	12	10	13	14	13.5
14	[0 0 1 0]	10	11	10	14	13	13.5
15	[0 0 0 1]	11	10	9	15	12	12
16	[0 0 0 0]	12	14	12	16	16	16

4.2.3 Clinicians Agreement

The degree of agreement between each pair of clinicians can be measured using Spearman's rank correlation. Table 4.6 shows correlations between the pairs of clinicians for rankings of 16 vectors.

High degree of agreement between clinicians was observed. Correlations between pairs of clinicians range from 0.887 to 0.936. The reason of observing high correlations among the clinicians is partly because they considered the more time points with diseases as worse outcome. For example, all three clinicians agreed that a child with 4 visits of disease ranks lower than the one with 3 visits of disease. All clinicians did not follow the same criteria for breaking ties so the reduction in correlation coefficients may have been due to this disagreement. For example, one of them preferred to see resolution at the earlier visits while two of them to see

resolution at the later visits within the same number of diseases, therefore, they broke ties by using different rules. In some instances, they could not clearly rank some of the profiles which led them to assign the same rank to the some vectors.

Table 4.6: Correlations (Spearman) among clinicians

Clinician	A	B	C
A	1		
B	.911	1	
C	.887	.936	1

4.3 APPLICATION TO A CLINICAL STUDY

We illustrate an application of our proposed method in the analysis of otitis media clinical trial data in which repeated observations were obtained at 5 unequally spaced time points over a period of time. We evaluate the ordering strategies consisting of a natural algorithm based on two different criteria and experts' rankings produced by three clinicians. We also compare our proposed approach with the other three standard methods.

4.3.1 Data Description

The motivating data of this study come from a randomized trial evaluating the effect of a treatment of acute otitis media for children between 7 months and 12 years of age. The trial, reported by Mandel, *et al*. (1995), compared 20 days of antimicrobial treatment versus the standard 10 days of amoxicillin treatment for children with AOM to prevent recurrences of AOM and to resolve the occurrence of middle ear effusion and also to determine if it is beneficial to change the antimicrobial agent after initial 10-day treatment with amoxicillin.

The study was designed as a placebo-controlled, double-blinded, randomized clinical trial. Patients were stratified according to their age (7-23 months, 2-6 years, and 7-12 years), laterality of effusion (unilateral or bilateral), and history of AOM during the previous year (0, 1-2, 3+ episodes). In the first 10 days of the trial, all patients were given an initial dose of amoxicillin. In the second 10 days of the study (days 11 through 20), available two hundred and sixty seven children with AOM at entry were randomly assigned to receive one of the three treatment medications: 1) 88 children continued taking amoxicillin, 2) 86 children received amoxicillin-clavulanete (Augmentin), and 3) 93 children received a placebo. Follow-up assessments of examination of the ears, nose, and throat of each patient were scheduled at 10, 20, 30, 60, and 90 days after entry. Since all subjects were given the same treatment in the first 10 days, we assumed that the 20-day assessment time point was considered as the first time point so that there are 4 follow-up time points included in the analysis.

Of the patients lost to follow up prior to the 10-day assessment (2 in the Amoxicillin group, 4 in the Augmentin group, and 4 in the placebo group), 3 patients discontinued the study with unknown reasons and 7 patients were released from the study due to recurrence of symptoms of acute infection before the end of the initial 10-day course of amoxicillin. 3 patients were withdrawn before the 20-day assessment, one in each treatment group. A total of 13 patients were excluded from the analysis. Table 4.7 displays the overview of the missing data patterns by treatment group. Although the total number of patients in this trial was 267, the number of subjects who had complete data at all time points was 201. In addition to 13 patients excluded from the study, 30 patients were lost to follow-up after the 20-day assessment: 6 of them before 30-day, 16 of them before 60-day, and 8 of them before 90-day assessments. 23

patients missed only one of four visits during the period of the trial: 8 missed at the 20-day, 9 missed at the 30-day, and 6 missed at the 60-day assessments.

Table 4.7: Overview of missing data: Number of subjects in each missing data pattern

Assessment (days)				Number of Subjects by Treatment Group			
20	30	60	90	Amoxicillin	Augmentin	Placebo	**Total**
Completers							
O	O	O	O	68	67	66	201
Dropouts							
O	O	O	M	1	3	4	8
O	O	M	M	6	3	7	16
O	M	M	M	2	2	2	6
M	M	M	M	1	1	1	3
Intermittent							
M	O	O	O	5	2	1	8
O	M	O	O	1	2	6	9
O	O	M	O	2	2	2	6

(O: Observed, M: Missing)

Table 4.8 summarizes frequency of the observed profiles for four time points in the dataset. Of the 74 patients who completed all scheduled visits without any diseases, 37%, 35%, and 28% of them were in the Amoxicillin, Augmentin, and Placebo groups, respectively. Among the 61 patients whose measurements at all time points were observed and who had only one visit with disease, the proportion of patients in Amoxicillin, Augmentin, and Placebo groups were 31%, 39%, and 30%, respectively. Of the 31 patients who had 2 visits with disease and 2 visits without disease, the proportion of patients in Placebo (48%) was higher than the proportions of patients in Amoxicillin (29%) and Augmentin (23%) groups. The more patients were observed in Placebo (39%) and Amoxicillin (35%) groups compared to those in Augmentin group (26%) among the patients who had only one visit without disease over four time points.

Table 4.8: Frequencies: Observed profiles by treatment groups

	Assessment (days)				Frequency			
	20	30	60	90	Treatment Groups			
	Profile				Amoxicillin	Augmentin	Placebo	Total
1	0	0	0	0	27	26	21	74
2	0	0	0	1	5	6	7	18
3	0	0	1	0	7	8	4	19
4	0	1	0	0	3	8	0	11
5	1	0	0	0	4	2	7	3
6	0	0	1	1	3	2	2	7
7	0	1	0	1	2	1	0	3
8	0	1	1	0	1	0	1	2
9	1	0	0	1	1	3	1	5
10	1	0	1	0	0	0	3	3
11	1	1	0	0	2	1	8	11
12	0	1	1	1	1	3	2	6
13	1	0	1	1	1	0	0	1
14	1	1	0	1	2	0	3	5
15	1	1	1	0	4	3	4	11
16	1	1	1	1	5	4	3	12
17	0	0	1	•	0	0	1	1
18	0	1	1	•	0	3	0	3
19	1	0	1	•	0	0	3	3
20	1	1	1	•	1	0	0	1
21	0	0	•	•	0	2	0	2
22	0	1	•	•	4	0	3	7
23	1	0	•	•	1	0	0	1
24	1	1	•	•	1	1	3	5
25	0	•	•	•	0	2	1	3
26	1	•	•	•	2	0	1	3
27	•	0	0	0	2	1	1	4
28	•	0	1	1	1	1	0	2
29	•	1	1	1	2	0	0	2
30	0	•	0	0	1	2	1	4
31	0	•	1	1	0	0	1	1
32	1	•	0	0	0	0	2	2
33	1	•	0	1	0	0	1	1
34	1	•	1	1	0	0	1	1
35	0	0	•	0	2	0	1	3
36	0	1	•	0	0	1	0	1
37	1	0	•	0	0	1	0	1
38	1	1	•	0	0	0	1	1
39	•	1	•	•	0	0	1	1

Note: 1 = Disease; 0 = Disease-free; • = Missing

All possible profiles with missing values were not observed in this dataset (e.g., [1 • 1 •] was not observed) and there was no particular profile with missing values that occurred mainly over groups.

4.3.2 Data Analysis Results

Since our purpose is to illustrate the application of the proposed method in analyzing data, we restrict attention to the comparison of two treatment groups (Augmentin versus Placebo). We considered the rankings produced by natural algorithm and clinicians' opinion to order to the vectors and employed the proposed method to the data using each of these orderings.

We applied the approach given in *Section* 3.2.2.1 to assign scores to the vectors with missing observations using the scores assigned to the complete vectors presented in Table 4.1, Table 4.2, and Table 4.3. We used the fact that missing response value would have been 1 (disease) or 0 (no disease) if it was measured. We assumed that the probability of observing disease and that of observing no disease are equal if that missing value had been obtained. For example, to assign a score to the vector [0 0 1 •], the probability of obtaining [0 0 1 **1**] is equal to that of obtaining [0 0 1 **0**] and thus, simple average of the scores of [0 0 1 1] and [0 0 1 0] is assigned to [0 0 1 •]. In other words, the probabilities in equations (3.6) and (3.7) are chosen as 0.5. Table 4.9 demonstrates the scores assigned to the vectors with missing values using the scores in Table 4.1, Table 4.2, and Table 4.3.

The analyses present the results of comparing the treatment effect using the proposed method with different ordering strategies achieved by natural algorithm and clinicians' opinion and the *t*-test comparing the proportions of diseases between the two treatment groups are presented in Table 4.10. Table 4.11 shows the analyses results from the GEE and Mixed Effects

Table 4.9: Scores assigned to the profiles with missing values

Observed		Natural Algorithm		Clinicians		
Profile	**Possibilities**	**1**	**2**	**A**	**B**	**C**
[1 1 1 •]	{ R(1111) + R(1110) } / 2	1.5	1.5	1.5	2.5	3
[1 1 0 •]	{ R(1101) + R(1100) } / 2	4.5	7.5	3.5	4.5	5.5
[1 0 1 •]	{ R(1011) + R(1010) } / 2	5.5	6.5	4	4	5.5
[1 0 0 •]	{ R(1001) + R(1000) } / 2	10	11.5	6.5	10.5	9
[0 1 1 •]	{ R(0111) + R(0110) } / 2	7	5.5	4	5.5	4.5
[0 1 0 •]	{ R(0101) + R(0100) } / 2	11.5	10.5	7.5	9.5	8.5
[0 0 1 •]	{ R(0011) + R(0010) } / 2	12.5	9.5	8.5	9	8
[0 0 0 •]	{ R(0001) + R(0000) } / 2	15.5	14	11.5	12	10.5
[1 1 • •]	{ R(111 •) + R(110 •) } / 2	3	4.5	2.5	3.5	4.25
[1 0 • •]	{ R(101 •) + R(100 •) } / 2	7.75	9.25	5.25	7.25	7.25
[0 1 • •]	{ R(011 •) + R(010 •) } / 2	9.25	8	5.75	7.5	6.5
[0 0 • •]	{ R(001 •) + R(000 •) } / 2	14	11.75	10	10.5	9.25
[• 1 • •]	{ R(11 • •) + R(01 • •) }/ 2	6.125	6.25	4.125	5.5	5.375
[1 • • •]	{ R(11 • •) + R(10 • •) }/ 2	5.375	6.875	3.875	5.375	5.75
[0 • • •]	{ R(01 • •) + R(00 • •) }/ 2	11.625	9.875	7.875	9	7.875
[• 1 1 1]	{ R(1111) + R(0111) } / 2	3	1.5	2	1.5	1.5
[• 0 1 1]	{ R(1011) + R(0011) } / 2	7.5	4.5	5	5	4.5
[• 0 0 0]	{ R(1000) + R(0000) } / 2	14	15.5	10	13.5	11.5
[1 • 1 1]	{ R(1111) + R(1011) } / 2	2.5	2	2	2	2
[1 • 0 0]	{ R(1100) + R(1000) } / 2	9	13	6	9.5	9
[0 • 1 1]	{ R(0111) + R(0011) } / 2	8	4	5	4.5	4
[0 • 0 0]	{ R(0100) + R(0000) } / 2	14.5	15	10.5	13	11
[1 • 0 1]	{ R(1101) + R(1001) } / 2	5.5	6	4	5.5	5.5
[1 1 • 0]	{ R(1110) + R(1100) } / 2	4	8	3	5	6
[1 0 • 0]	{ R(1010) + R(1000) } / 2	9.5	12.5	6.5	9	9.5
[0 1 • 0]	{ R(0110) + R(0100) } / 2	11	11.5	7	10.5	8.5
[0 0 • 0]	{ R(0010) + R(0000) } / 2	15	14.5	11	12.5	11

R ([]): rank score of the vector []

Model using the available data. While Wilcoxon tests produced significant treatment differences based on the ranking of natural algorithm 1, clinician A, and clinician B, the treatment comparison was not significant from the analysis when the ordering was achieved by natural algorithm 2. A borderline statistically significant difference (χ_1^2 =3.22, p=0.0725) was found between the two treatment groups when the analysis was performed based on the ordering of

clinician C. A two-sample *t*-test comparing the average of the proportion of disease over time yielded a significant treatment difference with p=0.0369.

Table 4.10: Data analysis results from the proposed method and two sample t-test

	Method of Analysis					
	Proposed Method					***t*-test**
	Natural Algorithm		Clinician			
	1	2	A	B	C	
Test Statistic	5.29	2.73	4.76	4.72	3.22	2.10
p value	0.0215	0.0987	0.0292	0.0299	0.0725	0.0369

Based on the main effects models testing the overall treatment effect over time, GEE and mixed effects logistic regression with random intercept methods yielded statistically significant treatment effect with p=0.0273 and p=0.0477, respectively. GEE and mixed effects logistic regression approaches produced significant treatment by time interaction effect (Table 4.11).

The results based on the rankings of clinician A and B were similar because the rankings of these two clinicians are highly correlated. While the proposed method produced significant treatment difference based on the ordering of clinician A, the result was not significant based on the ordering of clinician C. The reason may be because clinician A and clinician B applied different algorithm to break ties. Even though high correlation between the rankings of clinician B and C is observed, the treatment difference was significant based on the rankings of clinician B but not clinician C. The reason of obtaining different results from clinician B and clinician C can be explained by the fact that they broke the ties for vectors with 2 diseases in a different way. Different ordering strategies produced different results as the proposed method yielded different

results based on the natural algorithm 1 and 2. As a summary, the levels of statistical significance of the proposed method were competitive at least with those of the other standard methods.

Table 4.11: Data analysis results from GEE and mixed effects logistic regression model

Main Effects Model						
	Marginal Model GEE			Mixed Effects Model		
	Estimate	SE	p value	Estimate	SE	p value
Treatment (Augmentin)	-0.48	0.2191	0.0273	-0.56	0.2801	0.0477
Time (days)	-0.003	0.0032	0.2924	-0.004	0.0036	0.2432
Main and Interaction Effects Model						
	Marginal Model GEE			Mixed Effects Model		
Treatment (Augmentin)	-1.17	0.3795	0.0021	-1.41	0.4584	0.0024
Time (days)	-0.01	0.0046	0.0338	-0.013	0.0051	0.0148
Treatment*Time	0.014	0.0062	0.0242	0.018	0.0074	0.0174

4.4 CONCLUSION

We presented several methods of ordering the vectors of observations. Ordering strategies obtained by natural algorithm and from a group of clinicians were applied to the otitis media clinical trial. We compared the results obtained from the proposed method using natural algorithm and clinicians' ordering schemes with those produced by other methods. Results of the analysis show that the proposed method provides results similar to those obtained from other methods. We can conclude that our proposed method is competitive with other methods. Also, our proposed approach is adaptable to missing data.

As shown in Table 4.1 and Table 4.5, clinician A ranked the vectors very similar to the natural algorithm 1 which orders the vectors by the number of time points with disease and the earlier time to occurrence of disease. As presented in Table 4.10, the proposed method using

these two orderings produced very similar results. If clinician B and C had followed the same strategies as clinician A to order the vectors, one general algorithm would be created (i.e., natural algorithm 1) and this algorithm would be used for the analysis. In other words, when a good agreement among the clinicians with respect to ordering of the vectors is attained, general algorithm can be developed based on the clinicians' ordering (*Bjorling et al*. 1997). This established algorithm can be used as a reference for similar studies.

The feasibility of our proposed method depends on the degree of agreement among clinicians with respect to the ordering of the vectors of repeated binary observations. When clinicians' rankings are inconsistent with each other, it might be difficult to apply the proposed approach. One could use a specific ordering algorithm to evaluate the treatment effect, draw conclusion about treatment use, and interpret the results based on this ordering scheme. In this situation, consensus among the clinicians is not important (*Brittain et al*. 1997).

In clinical trials, some subjects withdraw the study and the reasons for dropping out can be available for these subjects. For example, subjects leave the study due to a lack of treatment effect or adverse effects. Even though we did not incorporate the reasons of withdrawal in the analysis of otitis media trial, these informative missing data can be easily incorporated into our approach by ranking the vectors with missing data according to the reasons of withdrawal (Modification of Gould's Method, *Section* 3.2.2.2). Our approach can handle this type of missing data while incorporating informative missing data into the analysis may be sophisticated for other methods (GEE, MERI).

Even though all methods had similar results for this one dataset, it may not be the case in general. The simulation results showed that depending on the pattern of differences our proposed method can be a lot worse or a lot better than conventional methods. Present dataset had

interactions over time. Therefore we believe that for some diseases and some situations, there will be differences where our method will have an advantage because clinically important differences are being targeted.

5.0 REPEATED BINARY MEASURES WITH MULTIPLE OUTCOMES

Clinical trials are often planned to compare two treatments (e.g., drug compared with a placebo) using repeated binary measurements over time and in general, a treatment effect is evaluated based on one response variable of interest. Additionally, several outcomes can be observed in such studies and these outcomes may cause different clinical influences on subjects. It may not be appropriate to evaluate the overall treatment effect without accounting for these occurrences. Using analyses based on single outcome may not capture all aspects of outcome to assess the overall effects of therapies under study and it may increase the risk of drawing improper conclusions. This chapter focuses on the proposed method in evaluating a treatment effect from longitudinal binary data with multiple outcomes.

5.1 MULTIPLE OUTCOMES AND EVENTS

In clinical studies, multiple outcomes may arise (e.g., AOM, MEE, or none in otitis media trial) and many symptoms may occur due to the progression of the disease in addition to single primary outcome (1=disease, 0=disease-free). Moreover, some undesirable effects of the therapy which could have an impact on patients' quality of life may occur besides the benefits of the treatment. In some studies in which the effect of treatment declines over time or patients are unresponsive to the assigned therapy, it might be necessary to change the therapy. Also, some subjects prematurely withdraw the study and comparison of treatments may be affected if the patterns or circumstance of dropouts are different between the treatment groups. To effectively

evaluate a treatment effect, it is important to incorporate occurrences developing during the study such as different responses, serious side-effects, and 'need for clinical intervention' which can influence outcomes.

In the presence of such occurrences, an overall evaluation of the treatment effect is not adequately determined by examining individual response because the related information about the treatment effect from various events is not considered (*Gray and Brookmeyer* 1998). While treatment groups are statistically compared in terms of main outcome, examination of qualitative findings and other possible responses that might reveal clinically meaningful difference must be included in the analyses. The proposed method allows the assessment of overall treatment effect in presence of these occurrences. It has the flexibility to integrate information across multiple events. The objective of this chapter is to discuss the proposed method for analyzing this type of longitudinal data and to demonstrate how to adapt the proposed method in distinguishing "clinically relevant difference". We illustrate adjustment of the proposed method with an example using data from otitis media clinical trial.

5.1.1 Adverse Effects and Need for Clinical Intervention

In clinical trials, patients may experience several occurrences during the course of the trial such as serious adverse effects, insufficient effectiveness of treatment, and allergic reaction to the therapy. Although a treatment has a positive effect on a primary outcome, it can have a severe negative effect on another body system. It may be required to give a non-protocol clinical intervention or another treatment to the patients who experience serious side effects or are unresponsive to the medical treatment. The requirement of giving another therapy or intervention may result in a drastic change in the primary outcome and interfere with the study results. On the

other hand, a therapy, for example, may cause side effects yet has generally positive effect. If subjects who drop out due to adverse effects of therapy are not included in the analysis, favorable information about the effectiveness of the treatment may be ignored.

When treatment and placebo are being compared, it is not uncommon that the placebo group has a lower rate of improvement of progression of disease than the treatment group. Therefore, non-protocol intervention may be necessary for subjects receiving placebo. Different rates of interventions in the treatment groups may result in the improper comparison of treatment groups and an incorrect conclusion that treatment is not preferable to placebo may be reached. For example, in a clinical trial of otitis media comparing the effect of antibiotic in children with acute otitis media, children in placebo group are more likely to need tube insertion than those in the treatment group to prevent fluid in their ears due to insufficient therapeutic effect. If a considerable number of children in the placebo group receives tube, effusion will not be observed and a desirable response will be obtained for these children. Ignoring a large number of tube insertions in the placebo group can lead to a misleading conclusion that treatment is not effective. Nevertheless, tube placement clinically is not regarded as a good outcome.

In placebo-controlled studies, it is not unexpected that subjects in the treatment group may have lack of therapeutic effects or intolerable side effects. It might be required to give another therapy to the subjects who have experienced adverse effects of study treatment. Responses from subjects who had been randomly assigned to treatment group but received different therapy because of unpleasant side effects would not be similar to the responses from the subjects who did not change the assigned therapy. This results in difficulties in comparing treatment groups.

In such trials, it is important to accommodate these occurrences and qualitative measures, which are observed during the trial, to draw proper inferences and results. The relative benefits and side-effects of treatment must be weighed to evaluate the overall effect of the therapy. It would be useful to incorporate such events into the analysis in a way that preserves the clinical relevance of the outcomes. The proposed approach can be adaptable to such occurrences.

5.1.2 Clustered Data

In some studies, clustering of observational units arises and this induces dependence among the responses of the same cluster. For example, the left and right eyes of individual patients are evaluated for the examination of eye illnesses in the ophthalmology studies. In otitis media clinical trials where patients are followed over time after the initiation of antimicrobial treatment, measurements are obtained separately from each ear of a subject to assess the symptoms of acute infection. Since ears are clustered within subjects, the data are doubly nested. In this situation, two types of correlation are inherent: the correlation between ears or eyes of the same subject and the correlation between the measurements taken on each subject at different time points. Type I error rate is inflated or biased results may be obtained if an analysis ignores the correlation between the ears clustered within the same subjects (*Hedeker and Gibbons* 2006). Methods must account for the correlations that exist between measurements taken from each subject both at the same time and across time.

Analyzing this type of data would be a challenge for some of the standard methods. For example, applying the GEE procedure to this kind of dataset could be appropriate however, this method would require a complicated correlation structure (*Lefkopoulou et al*. 1989). Even though the GEE method is robust against misspecification of correlation structure, it is important to

choose an appropriate correlation structure before performing GEE as incorrect choice of correlation structure causes a reduction in efficiency (*Hedeker and Gibbons* 2006). Our proposed method is a simpler approach and adaptable to handle this sort of data nested twice.

For example, in a clinical trial for assessing the efficacy of antibiotic in children with AOM, a child who has experienced unsatisfactory treatment effects and has had *bilateral* effusion is more likely to have hearing loss which may damage language and cognitive development during early childhood than the one who has *unilateral* effusion because unaffected ear can prevent complete hearing loss (*Bluestone and Klein* 2001).

Furthermore, otitis media with effusion (OME) is defined as an inflammation of the middle ear with fluid without signs or symptoms of acute infection. Bilateral OME for 3 months or more may result in tube placement regardless of hearing loss. If a child with unilateral OME for 6 months or more has significant hearing loss, language and learning problems, or middle ear abnormalities, placement of a tube would be appropriate to prevent OME and to improve hearing to reduce the risk for language and learning problems (*Alper et al*. 2004). Since tube placement, which is clinically considered a poor outcome, is required for children with long standing bilateral OME more than for those with unilateral OME, bilateral effusion is a different and worse outcome than the unilateral effusion, thus, they should be considered as separate outcomes. The proposed method has the ability to differentiate the difference between these two outcomes so that it accounts for correlation that occurs between observations within subject.

5.1.3 Categorical Data

Even though clinical trials are designed to measure the effect of the treatment on a single outcome, treatments have different impacts on patients. It is of importance to distinguish the

possible outcomes a patient may have in such trials. The effect of treatment is sometimes evaluated on each separate outcome using standard approaches. For example, in a trial of treatments for otitis media, therapy is given to subjects to improve several different outcomes such as prevention of AOM and resolution of MEE in the middle ear. Treatment effect is evaluated based on each individual response variable. One could count separately the total number of AOM and MEE episodes over time for each subject and compare the proportions of time with AOM and MEE between the two treatment groups to test the null hypothesis of no treatment effect on each outcome. Even though evaluating individual responses gives useful information about the treatment effect, it does not provide a single overall evaluation of the treatment effect. This crude analysis also results in the inflation of the experimentwise error rate.

Since AOM is a worse outcome than MEE, one could consider treating outcome as ordinal data and apply GEE or mixed effects models to analyze longitudinal ordinal data. However, these analyses may not have enough statistical power to detect the effects of interest between two treatment groups if adequate frequencies of AOM or MEE are not observed. Infrequent sparse cells can be collapsed into one category due to the rare occurrence of AOM and MEE but collapsing categories can result in a loss of information. The proposed method does not suffer from such limitations. We can adapt our approach to analyze data with this type of outcome to detect clinically meaningful treatment effect without sacrificing information.

5.2 ADAPTATION OF THE PROPOSED METHOD

In the previous section, we discussed the possible outcomes which can occur in clinical studies besides a main outcome. Our proposed method can be adapted to accommodate these

occurrences as long as vectors of binary observations with multiple outcomes and adverse effects can be clinically ranked. If clinicians can order the vectors in a clinically relevant manner and consensus among clinicians about rankings is obtained, the proposed method can be easily applied to the data. In the absence of agreement among clinicians, interpretation of the results may not be easy and the proposed method may not be beneficially applied. However, a method could be employed incorporating the variability among clinicians (e.g. bootsraping clinical response) or enough weight might be obtained to develop an empirical algorithm for ranking outcomes (e.g. the disagreements might be considered as ties) We present clinicians' responses for complex situations and use the information from them to employ the proposed method to the otitis media data.

5.2.1 Clinicians' Opinion

As discussed in the previous chapter, we obtained input from clinicians about ordering binary outcomes in Scenario 1, *Section* 4.2.1. We asked them to order vectors of observations for more complex scenarios. We present the rest of the instructions provided to three clinicians to order the vectors with multiple outcomes.

5.2.1.1 Guidelines for Clinicians

Scenario 2: Consider Scenario 1 with two modifications:

1) The trial is six months

2) If the treatment is not effective, the child receives tubes.

Thus we could have outcomes like [1 1 1 1 T T] which would indicate effusion at the first four visits and tube is inserted by visit five. How would you incorporate tubes into the ranking process? What are the clinical criteria you use to place tubes?

Scenario 3: We now address a more complicated algorithm. Specifically, we incorporate AOM into Scenario 1. We address this more complicated scenario in two steps:

Step 1 – For a fixed pattern in Scenario 1, e.g. [1101], is there a simple way to order the outcomes with AOM superimposed within this MEE pattern? Thus, order

1	[1 1 0 1]
2	[A 1 0 1]
3	[1 A 0 1]
4	[1 1 0 A]
5	[A A 0 1]
6	[A 1 0 A]
7	[1 A 0 A]
8	[A A 0 A]

where A indicates AOM.

Is there a general algorithm you are applying?

Step 2 – Are there any suggestions you have for ordering AOM superimposed on different patterns?

e.g. Pattern 1111 and 1110

1111 $\leftrightarrow$ A1A0 less MEE but more AOM

Any insight you have as to a general algorithm for these types of comparisons would be useful.

Thank you for your help. We believe this could eventually result in a better procedure to analyze studies of patients with MEE.

5.2.1.2 Clinicians' Responses

Table 5.1 presents the responses from three clinicians for Scenario 2 about tube intervention. All clinicians agreed fairly well with each other regarding the tube insertion. It can be interpreted as tube placement is considered a poor outcome because patients who are unresponsive to medical treatment are likely to be given a tube. Also, it is worth noting that bilateral effusion is a different outcome than unilateral effusion since one of the criteria for placing tube for patients is to have 3-6 months of effusion depending on the bilateral or unilateral. Therefore, it is important to account for this distinction between bilateral and unilateral effusion in the analysis.

Table 5.1: Responses from 3 clinicians for Scenario 2

Clinician A
Tube would be worse than 6 months of effusion.
Clinician B
Functionally, [1111-11] would be (or may be) worse than [1111-TT], as hearing would most likely be within the normal range with tubes but may not be with effusion. The minimum criteria for undergoing tympanostomy tube insertion for OME is $\geq$3 months of bilateral effusion or $\geq$6 months unilateral effusion, unresponsive to medical treatment and not improving. Such factors as age of the child, hearing status, season, presence of developmental/school problems, also enter into the decision of whether or not to recommend tube placement.
Clinician C
1. $\geq$ 3-6 months continuous effusion especially if bilateral not improving and especially if significant hearing deficit 2. Might also lean more about tubes if there was apparent and relevant patient discomfort. In reality, clinically this is more likely to occur in the presence of superimposed recurrent AOM.

Table 5.2 illustrates the ordering of outcomes with AOM superimposed to MEE by three clinicians. Considerations suggested by them for ordering 8 vectors are given in Table 5.3. They reached a consensus about comparing AOM and MEE. They preferred MEE to AOM and agreed that AOM is worse outcome than MEE. All clinicians first order the vectors according to the

number of AOM which shows a high degree of consensus about importance of observing AOM or MEE.

Based on clinicians' considerations, a vector with 3 AOM is the worst outcome and a vector without AOM is the best outcome among these 8 trajectories. While Clinicians A and B did not break the ties in terms of time to occurrence of AOM, Clinician C considered the time to occurrence of AOM in the ordering and broke the ties based on this criterion. It is stated that other factors such as age and period of language development influence ordering the vectors with AOM and MEE.

Table 5.2: Rank scores for Scenario 3 by clinicians

		Clinician		
	Profile	**A**	**B**	**C**
1	[A A 0 A]	1	1	1
2	[1 A 0 A]	3	3	2
3	[A 1 0 A]	3	3	4
4	[A A 0 1]	3	3	3
5	[1 1 0 A]	6	6	7
6	[1 A 0 1]	6	6	6
7	[A 1 0 1]	6	6	5
8	[1 1 0 1]	8	8	8

A consensus may not be achieved when comparing a vector including one type of outcome with a vector including another type of outcome when there are different numbers of points of diseases. For example, it might be difficult to order vectors [1100] and [A000] as clinician C addressed in Table 5.3. One question to ask clinicians would be how many time points with MEE would be better, worse or equal to how many time points with AOM. For example, three time points with MEE over four time points ([1110]) is worse outcome than the 1 time point with AOM ([A000]) based on the evaluation of clinician C. As a result, consensus

among clinicians may not be reached if rankings are applied to different type of outcomes. In this situation, the same scores are assigned to some vectors in the absence of agreement.

Table 5.3: Responses from clinicians for Scenario 3 and Step 2

Clinician A
AOM is worse than MEE. It's painful. I would rather have a child with OME for many months before opting for AOM. The age of the child also matters. If it is during language development and if the MEE is bilateral, it is worse.
Clinician B
AOM is more disruptive to the family and child than OME (in most cases), so the more the episodes of AOM, the worse the course.
Clinician C
A1A0 is worse than 1110. What about A000 vs. 1100? i.e. is any acute worse by diffusion? How much effusion? A000 vs. 1100 is difficult but feel A000 is worse because it requires antibiotic with attendant costs, potential adverse effects, etc patient discomfort to AOM. But I would consider 1110 worse than A000 because of continuous effusion.

5.3 AN ILLUSTRATION: OTITIS MEDIA TRIAL

Algorithms can be developed to take into account simultaneously whether disease status was none, fluid (MEE), infection (AOM); unilateral or bilateral disease; number of time points of disease and early disease versus late disease. Some hierarchy including all possible outcomes is defined and the outcomes are considered from top to bottom to assign scores to the subjects. We create a set of criteria for ranking vectors of observations with multiple outcomes for otitis media trial based on the clinicians' responses presented above. We illustrate how our proposed method can be adjusted to the complex scenarios using this information. Table 5.4 depicts a set of

criteria. Based on the criteria given in Table 5.4, Bilateral AOM would be the worst outcome which can be observed at any time point.

Table 5.4: Criteria for ranking outcomes

1. AOM is worse outcome than MEE
2. Bilateral AOM (MEE) is worse than unilateral AOM (MEE)
3. Tube placement is considered as poor outcome

Children with AOM entering the otitis media trial where subjects are assessed over 4 time points may have one of the three possible outcomes at any time point: AOM, MEE, or none in their one or both ears. Seven possible outcomes which can be observed in the otitis media trial are given in Table 5.5. If multiple outcomes were not considered and only one response is evaluated as disease (1 = AOM, MEE, unilateral or bilateral) or not disease (0 = none), patients 1-4 and patients 5-7 would seperately share the same tied rank value. However, these tied ranks are broken on the basis of distinguishing these possible outcomes. For example, patient 2 who had *bilateral* effusion for the first two time points but remained effusion free for the last two time points would be considered as having worse outcome than patient 1 who had *unilateral* effusion for the first two points and stayed effusion free for the last two time points. Therefore, patient 2 is assigned a lower rank score than patient 1 when higher rank score is better. In the same manner, patient 4 is assigned a lower score than patient 3 because bilateral AOM are observed for patient 4 while unilateral AOM are observed for patient 3 in the first two time points. Therefore, a child with *bilateral* effusion ranks lower than a child with *unilateral*

effusion. These rank scores reflect relative ordering of the outcomes. Patients can be assigned different scores with respect to unilateral or bilateral outcome.

Even though patient 1 and 3 did not have disease-free observations for the first two time points, patient 3 experienced 2 AOM while patient 1 experienced 2 MEE. Since AOM is regarded as worse outcome than MEE, patient 3 has a lower rank score than patient 1. Our proposed method can incorporate categorical outcomes in the analysis. Among the first four patients, patient 4 is assigned the lowest rank score due to the both AOM and bilateral. Comparing patient 3 and 6 may be indeterminate. The reason is that clinicians might not agree that 2 AOM is worse than 3 MEE although 2 AOM is worse outcome than 2 MEE. Comparing patient 4 and 7 may be uncertain due to the same reason. Therefore, the same score can be assigned to the vectors if they are not precisely ordered. Patient 5 would have the lowest score among these 7 trajectories because the tube was placed after the 3rd time points. The reason of observing no disease at the 4th time point for patient 5 is tube insertion which interferes with outcome.

Table 5.5: Representation of trajectories from 7 patients form otitis media trial

Patient ID	Trajectories	Score
1	[MEE MEE 0 0]	4
2	[MEE-b MEE-b 0 0]	3
3	[AOM AOM 0 0]	2
4	[AOM-b AOM-b 0 0]	1
5	[MEE MEE-b MEE-b **T** 0]	0
6	[MEE MEE MEE 0]	2
7	[MEE-b MEE-b MEE 0]	1

MEE-b denotes bilateral MEE, AOM-b denotes bilateral AOM, and T denotes tube insertion.

5.4 CONCLUSION

We discussed how to adapt our proposed method in studies where several outcomes and occurrences are observed in addition to main outcome. The proposed method has the flexibility to incorporate these occurrences in the analysis but it might be difficult to employ the proposed method for some situations where there are different numbers of points of diseases and different type of outcomes because ordering vectors can be burdensome for clinicians. Even though clinicians rank the vectors in clinical relevant manner, there may not be a consensus among clinicians. However, the proposed method can still be applied by using specific individual ordering and in this case; interpretation is made based on this particular ordering. In some situations, enough weight might be calculated for each vector from clinicians' ranking to develop an algorithm or tied values are assigned in the absence of agreement.

On the other hand, the proposed method can be adjusted to extremely complex situations if a 'hierarchy' of criteria of ordering vectors can be applied. If vectors can be ordered in reasonable way, the proposed method can be easily applied. Also it can be readily adapted to accommodate non-protocol 'outcomes' (tube insertion in the otitis media trial).

6.0 DISCUSSION

The purpose of this study is to develop a family of statistical tests based on the Wilcoxon test statistic which orders the vectors of repeated binary observations and events where the ordering is determined by “clinical relevance”. Our simulation studies indicate that the proposed method has statistical power competitive with and, for some scenarios, is preferable to conventional methods in the absence and presence of missing data. The real data analysis (otitis media trial) also shows that the proposed method and other methods give similar results indicating that the proposed method is appropriate and adaptable to missing data.

Although the proposed method is not applicable to every situation, we believe that for some diseases and scenarios, this easy-to-apply, simple method is noteworthy in the sense that it can be adjusted to extremely complex situations if vectors can be hierarchically ordered in a reasonable fashion, it can be focused on alternatives that have high clinical relevance, and it can be readily adapted to accommodate non-protocol “outcomes” and missing data. Another advantage of the proposed approach is that no distributional assumptions are made and no assumptions are required regarding correlation among the observations.

While the proposed method has advantages, there are a few limitations. Some outcomes may be difficult to order or clinicians order the vectors in different ways which result in disagreement among clinicians. However, same scores might be assigned to the vectors which are not easily ordered or disagreement among clinicians can be considered as ties. Another disadvantage of the proposed method might be that for long term follow-up, ranking of all possible ‘theoretical’ outcomes may not be feasible or may be burdensome for clinicians. For

example, there are 2^8 = 256 possible outcomes if the number of time points is 8. However, in real life, all possible outcomes may not be observed (e.g. *Section* 4.3.1 otitis media clinical trial).

The proposed method with *crude substitution* approach was found to have statistical power comparable to other methods and to control the type I error rate reasonably close the nominal value of 0.05 depending on the ordering strategies in the presence of missing data in our simulation studies. However, simulation studies suggest that the proposed method with *logistic regression substitution* approach is not preferable due to the inflation of type I error rate. We believe that the logistic regression substitution approach can be improved by further investigations to control the type I error rate.

BIBLIOGRAPHY

Ali, M. W., Talukder, E. (2005). Analysis of Longitudinal Binary Data with Missing Data due to Dropouts. *Journal of Biopharmaceutical Statistics*, 15, 993-1007.

Alper, C., Bluestone, C. D., Casselbrant, M., Dohar, J., Mandel, E. (2004). Advanced Therapy of Otitis Media, *BC Decker Inc*.

Bluestone, C. D. and Klein, J. O. (2001). Otitis Media in Infants and Children, *3rd Edition, BC Decker Inc*.

Brittain, E., Palensky, J., Blood, J., Wittes, J. (1997). Blinded Subjective Rankings as a Method of Assessing Treatment Effect: A Large Sample Example from the Systolic Hypertension in the Elderly Program (SHEP). *Statistics in Medicine* 16(6): 681-93.

Bubbar, V. K., Kreder, H. J. (2006). Topics in training: The Intention-to-Treat Principle: A Primer for the Orthopaedic Surgeon. *Journal of Bone and Joint Surgery*, 88, 2097-2099.

Demirtas, H. (2004). Pseudo-Random Number Generation in R for commonly used Multivariate Distributions *Journal of Modern Applied Statistical Methods* 3(2): 485-497.

Diggle, P.J., Heagerty, P., Liang, K., Zeger, S.L. (2002). Analysis of Longitudinal Data.(2nd ed.):Oxford University Press.

Follmann, D., J. Wittes, et al. (1992). The use of subjective rankings in clinical trials with an application to cardiovascular disease. *Statistics in Medicine* 11(4): 427-37.

Gehan, E. A. (1965). A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-censored Samples. *Biometrika*, 52, 203-223.

Gould, A. L. (1980). A New Approach to the Analysis of Clinical Drug Trials with Withdrawals. *Biometrics*, 36, 721-727.

Gray, S. M. and Brookmeyer, R. (1998). Estimating a Treatment Effect from Multidimensional Longitudinal Data. *Biometrics* 54: 976-988.

Hedeker, D., Gibbons, R.D. (2006). Longitudinal Data Analysis. *Wiley: New Jersey.*

Houck, P. R., Mazumdar, S., Sengul, T. K., Tang, G., Mulsant, B. H., Pollock, B. G., Reynolds, C.F. (2004). Estimating Treatment Effects from Longitudinal Clinical Trial Data fith Missing Values: Comparative Analyses using Different Methods. *Psychiatry Research*, 129, 209-215.

Lefkopoulou, M., Moore, D., and Ryan, L. (1989). The Analysis of Multiple Correlated Binary Outcomes: Application to Rodent Teratology Experiments. *Biopharmaceutical Statistics* 84(407): 810-815

Li, X., Mehrotra, D. V., Barnard, J. (2006) Analysis of Incomplete Longitudinal Binary Data using Multiple Imputation. *Statistics in Medicine*, 25, 2107-2124

Liang, K. Y., Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73, 12–22.

Little, T., Rubin, D. B. (2002). Statistical Analysis with Missing Data (2nd ed). *Wiley: New York.*

Liu, G. and A. L. Gould (2002). Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *J Biopharm Stat* 12(2): 207-26.

Liu, H., Wu, T. (2008). Sample Size Calculation and Power Analysis of Changes in Mean Response over Time. *Communications in Statistics-Simulation and Computation*, 37:1785-1798

Liu, M., Wei, L., Zhang, J. (2006). Review of Guidelines and Literature for Handling Missing Data in Longitudinal Clinical Trials with a Case Study. *Pharmaceut. Statist.*, 5, 7-18.

Mandel, E. M., M. L. Casselbrant, et al. (1995). Efficacy of 20- versus 10-day antimicrobial treatment for acute otitis media. *Pediatrics* 96(1): 5-13.

Mann, H. B., Whitney, D.R. (1947). On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*, 18, 50-60.

Molenberghs, G., Verbeke, G. (2005). Models for Discrete Longitudinal Data. *Springer Series in Statistics*, Springer.

Minini, P., Chavance, M. (2004). Sensitivity Analysis of Longitudinal Binary Data with Non-monotone Missing Values. *Biostatistics*, 5, 4, 531-544.

Moye, L. A., B. R. Davis, et al. (1992). Analysis of a clinical trial involving a combined mortality and adherence dependent interval censored endpoint. *Statistics in Medicine* 11(13): 1705-17.

O'Brien, P. (1992). Comments. *Statistics in Medicine*, 11, 447-449.

Park, C. G., Park, T. and Shin, D. W. (1996). A Simple Method for Generating Correlated Binary Variates. *The American Statistician* 50(4): 306-310.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. *Wiley: New York*.

Tang, L., Song, J., Belin, T. R., Unutzer, J. (2005). A Comparison of Imputation Methods in A Longitudinal Randomized Clinical Trial. *Statistics in Medicine*, 24, 2111-2128.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics*, 1, 80-83.

.