SENSITIVITY ANALYSES OF TIME-TO-EVENT DATA WITH POSSIBLY INFORMATIVE CENSORING FOR CONFIRMATORY CLINICAL TRIALS

Yue Zhao

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

Chapel Hill
2012

Approved by:

Dr. Gary G. Koch
Dr. Amy H. Herring
Dr. Haibo Zhou
Dr. John S. Preisser
Dr. Wayne D. Rosamond

Abstract

YUE ZHAO: Sensitivity Analyses of Time-to-event Data With Possibly Informative Censoring for Confirmatory Clinical Trials
(Under the direction of Dr. Gary G. Koch and Dr. Amy H. Herring)

We presents a multiple imputation method for sensitivity analysis of continuous time-to-event data with possibly informative censoring. The imputed time for censored values is drawn from the failure time distribution conditional on the time of follow-up discontinuation. A variety of specifications regarding the post-withdrawal tendency of having events can be incorporated in the imputation through a hazard ratio parameter for discontinuation versus continuation of follow-up. Multiply imputed data sets are analyzed with the primary analysis method, and the results are then combined using the methods of Rubin.

We then introduce covariate-adjusted sensitivity analysis within the established framework. For the illustrative example in the previous paper (chapter 2), we compare three methods of analysis for time-to-event data, and then we illustrate how to incorporate these methods into the proposed sensitivity analysis for covariate adjustment. The three methods are the multivariable Cox proportional hazards model, non-parametric ANCOVA, and inverse probability weighting with propensity scores. The assumptions, statistical issues, and features for these methods are discussed.

Lastly we extend the underlying principle of the proposed sensitivity analysis to grouped time-to-event data. Various post-withdrawal assumptions are specified through a conditional odds ratio of failure for the discontinued vs. retained patients, so that the counts of withdrawals are redistributed to the failure counts in the following time intervals or to the counts censored at the end of study, as if all the withdrawers completed

follow-up. The hypothetical survival profile estimates and the inferences on treatment effects (i.e., the incidence density ratio, the odds ratio, the Mann-Whitney probability, and the Mantel-Haenszel criterion) are produced by matrix operations with the covariance estimators obtained using the linear Taylor's series approximations. Therefore there is no need to perform the multiple imputation procedures for the missing outcomes (i.e., probabilistically assign the patients to a failure status in the time intervals following their withdrawals).

The methods are straightforward to implement with SAS macros. The interpretation of the sensitivity parameters is transparent and easily conveyed to clinical reviewers.

Acknowledgments

My most heartfelt thanks go to my mentor and co-advisor, Dr. Gary G. Koch, who coached me through my research, exposed me to scientific thinking, and allowed me the opportunity to pursue my career goals. I will be forever grateful to him for his extreme patience, generosity and encouragement.

I would like to thank my co-advisor, Dr. Amy H. Herring, for her guidance and advice, and for providing me financial assistance when I needed it the most. I also appreciate the time and effort that Drs. John S. Preisser, Benjamin R. Saville, and Haibo Zhou have devoted to my academic progress. In addition, I would like to give my special thanks to Dr. Wayne D. Rosamond for participating on my committee.

Finally, I thank all the people who have helped me during my graduate study and supported me through the completion of my dissertation, particularly Dr. Mirza W. Ali who provided some motivating background for the topic addressed by the second chapter and Dr. Suzanne Edwards who generously provided data for the illustration example in the second and third chapters.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
GEE	Generalized estimating equation
ITT	Intent-to-treat
IPW	Inverse probability weights
KM	Kaplan-Meier
LTF	Loss-to-follow-up
LOCF	Last observation carried forward
MAR	Missing at random
MCAR	Missing completely at random
MI	Multiple imputation
MNAR	Missing not at random
MRM	Mixed-Effect Regression Model
NPANCOVA	Non-parametric ANCOVA
PM	Pattern mixture
PP	Per-protocol
PS	Propensity score

Chapter 1

Literature review

1.1 Introduction

The efficacy of a new treatment for providing better outcome, for improving patient's condition, or for preventing disease recurrence is typically evaluated in randomized clinical trials in which patients are followed over time. Two types of data can be collected to assess the efficacy of the new treatment: (1) Longitudinal data from the repeated measurements of a criterion on the same subject at multiple visits over time. This criterion may be a measure of functionality, physiological performance, symptoms, or general well-being. (2) Time-to-event data, where the endpoint event may be death, disease progression or recurrence.

In a perfect clinical trial, all randomized patients would be completers, who complete the study as planned without any violation of protocol. Completers for longitudinal data are the patients who attend every visit. And completers for time-to-event data are the patients who either have an event during the follow-up period, or complete the follow-up period without the event. If everybody in the trial is essentially a completer and has data for all measured primary outcome variables, the data analysis and interpretation of results would be fairly straightforward. In the longitudinal case, the

average effect of treatment across multiple visits or the effect at the last visit could be analyzed using standard methods, such as the repeated measures mixed-effect models. For time-to-event data, the Kaplan-Meier (KM) curve, Log-rank test, and possibly the Cox proportional hazards model are most often used. However, a ubiquitous problem in all clinical trials, regardless of the outcomes, is missing data due to patients discontinuing study treatment before study completion. For longitudinal data, missing due to withdrawal means a patient has missing status for all visits after a certain time point. In the time-to-event analysis, it means a patient's follow-up time is censored before the end of the study. If we are able to make the assumption of missing at random (MAR), which will be explained in the subsequent section, the usual methods mentioned previously will provide valid analysis and interpretable results. However, one could never know for certain whether the MAR assumption is appropriate, and hence there may need to be other methods to address the robustness of conclusions about the treatment effect when departure from the MAR assumption is possible. Such analysis is sometimes called *sensitivity analysis*, and it will be explained in the subsequent section.

Extensive efforts have been made to establish appropriate methods for analyzing incomplete data and performing sensitivity analysis for longitudinal clinical trials. Here, we want to focus on the issue of withdrawal in the time-to-event scenario and discuss methods for sensitivity analysis for regulatory settings. Section 1.2 begins with reviewing methods available in the longitudinal data settings, and provides the concept of withdrawal and discusses the impact of missing data in longitudinal data analyses. Section 1.3 introduces missing data mechanisms in the context of withdrawal. Section 1.4 summarizes the common reasons for withdrawal. Mixed-effect models and multiple imputation are the major analytic approaches to deal with missing data in longitudinal clinical trials; and they are described in Section 1.5 and 1.6, respectively. Section 1.7

discusses the concept and the necessity of sensitivity analysis, via selection and pattern mixture models. Available sensitivity analysis strategies under the intent-to-treat (ITT) principle for longitudinal clinical trials are presented in Section 1.8. Finally, the recent development of sensitivity analysis for handling dependent and informative censoring is reviewed in section 1.9.

1.2 Withdrawal in Longitudinal Clinical Trials

A defining feature of longitudinal clinical trials is that the repeated measurements on the same individual during the course of treatment allow a direct study of change in the treatment effect over time. Typically, a baseline (or pre-treatment) measurement is taken on all patients who are then randomized. Measurements of the outcome variable are then taken repeatedly at the same set of occasions for all participants in the study. At the last planned time point, the treatment difference from placebo (i.e., the efficacy of the experimental drug) can be assessed in terms of change in outcomes from baseline or average rate of change in outcomes. Although efforts are made to collect data on every individual in the trial at each time point of follow-up, some patients could stop adhering to the protocol or discontinue their study treatment for reasons beyond the control of investigators. Consequently, the subsequent follow-up data will be missing (i.e., *monotonic* missing). If a patient is no longer seen after a certain follow-up visit, we say the data is missing due to *withdrawal*. In the literature, *dropout* and *loss to follow-up* are also used synonymously.

In many areas of clinical research, withdrawal is the major reason for missing data and most directly affects the interpretation of trial results. For instance, how to analyze and interpret longitudinal data often depends on unverifiable assumptions about the underlying missing data mechanism. Therefore, different assumptions and associated

analysis methods may lead to conflicting statistical inferences about treatment benefit (Permutt and Pinheiro, 2009). Furthermore, incorrectly addressing missing data may bias parameter estimates, inflate Type I and Type II error rates, and degrade the performance of confidence intervals, thereby leading to incorrect conclusions about treatment effect (Collins et al., 2001).

1.3 Missing Data Mechanisms

To understand how statistical inferences may be affected by missing data, it is important to understand the missingness mechanisms. The following terminology is based on the standard missing data framework of Little and Rubin (2002).

In the context of a longitudinal clinical trial, we assume that k measurements are to be obtained at times $t_1, ..., t_k$ for n independent subjects. For subject i, $i = 1, ..., n$, a set of measurements y_{ij} $(j = 1, ..., k)$ is collected, and we define the following:

- $\boldsymbol{y_i} = (y_{i1}, ..., y_{ik})$ is a $(1 \times k)$ complete data vector of outcomes for subject i, possibly with monotonic missing data.
- $\boldsymbol{r_i}$ is the missing data indicator. Specifically, let $r_{ij} = 1$ if y_{ij} is observed; and let $r_{ij} = 0$ if y_{ij} is missing.
- Given $\boldsymbol{r_i}$, $\boldsymbol{y_i}$ can be partitioned into $(\boldsymbol{y_i^o}, \boldsymbol{y_i^m})$, corresponding to the observed and the missing part of $\boldsymbol{y_i}$.

Then, the full data density can be factored into two parts as follows:

$$f(\boldsymbol{y_i^o}, \boldsymbol{y_i^m}, \boldsymbol{r_i}|X_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y_i^o}, \boldsymbol{y_i^m}|X_i, \boldsymbol{\theta}) f(\boldsymbol{r_i}|\boldsymbol{y_i^o}, \boldsymbol{y_i^m}, X_i, \boldsymbol{\psi}). \tag{1.1}$$

Here, X_i is the design matrix for observed covariates, and $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ denote parameter vectors for the measurement and missingness mechanism, respectively. The first factor

is the marginal density of the measurement process, and the second one is the density of the missingness process conditional on the outcomes.

Under a *missing completely at random* (**MCAR**) mechanism, the missingness is assumed to be unrelated to either the observed data or the missing outcomes, i.e.,

$$f(\boldsymbol{r_i}|\boldsymbol{y_i^o}, \boldsymbol{y_i^m}, X_i, \boldsymbol{\psi}) = f(\boldsymbol{r_i}|\boldsymbol{\psi}). \tag{1.2}$$

Therefore, 1.1 simplifies to

$$f(\boldsymbol{y_i^o}, \boldsymbol{y_i^m}, \boldsymbol{r_i}|X_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y_i^o}, \boldsymbol{y_i^m}|X_i, \boldsymbol{\theta}) f(\boldsymbol{r_i}|\boldsymbol{\psi}), \tag{1.3}$$

indicating the independence of the measurement and missingness mechanisms. The joint distribution of $\boldsymbol{y_i^o}$ and $\boldsymbol{r_i}$ becomes

$$f(\boldsymbol{y_i^o}, \boldsymbol{r_i}|X_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y_i^o}|X_i, \boldsymbol{\theta}) f(\boldsymbol{r_i}|\boldsymbol{\psi}). \tag{1.4}$$

Hence, in whatever way the observed data are analyzed (whether using a frequentist or likelihood method), the missingness mechanism is *ignorable*. For instance, under MCAR, analysis restricted to cases for which all measurements were recorded (i.e., a complete-case analysis) will yield unbiased estimates. Furthermore, MCAR also implies that the distribution of unobserved outcomes after withdrawal is the same for those who do and do not withdraw, and the outcomes for those who withdraw has the same distribution as the target population (Fitzmaurice, 2003). Unfortunately, those assumptions are often not realistic.

Under a *missing at random* (**MAR**) mechanism, $\boldsymbol{r_i}$ depends on $\boldsymbol{y_i}$ only through its observed part $\boldsymbol{y_i^o}$. Thus, conditional on the observed data (i.e., the fixed covariates X_i

and $\boldsymbol{y_i^o}$), the missingness is independent of the missing outcomes:

$$f(\boldsymbol{r_i}|\boldsymbol{y_i^o}, \boldsymbol{y_i^m}, X_i, \boldsymbol{\psi}) = f(\boldsymbol{r_i}|\boldsymbol{y_i^o}, X_i, \boldsymbol{\psi}). \tag{1.5}$$

The full data density is partitioned as

$$f(\boldsymbol{y_i^o}, \boldsymbol{y_i^m}, \boldsymbol{r_i}|X_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y_i^o}, \boldsymbol{y_i^m}|X_i, \boldsymbol{\theta}) f(\boldsymbol{r_i}|\boldsymbol{y_i^o}, X_i, \boldsymbol{\psi}). \tag{1.6}$$

At the level of observed data, we obtain

$$\begin{aligned} f(\boldsymbol{y_i^o}, \boldsymbol{r_i}|X_i, \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\boldsymbol{y_i^o}, \boldsymbol{y_i^m}|X_i, \boldsymbol{\theta}) f(\boldsymbol{r_i}|\boldsymbol{y_i^o}, X_i, \boldsymbol{\psi}) \mathrm{d}\boldsymbol{y_i^m} \\ &= f(\boldsymbol{y_i^o}|X_i, \boldsymbol{\theta}) f(\boldsymbol{r_i}|\boldsymbol{y_i^o}, X_i, \boldsymbol{\psi}). \end{aligned} \tag{1.7}$$

If $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are disjoint, then the likelihood can be factorized into two distinct components. The missingness is *ignorable* within the likelihood framework, and inference can be based solely on the marginal density of the observed data, $f(\boldsymbol{y_i^o}|X_i, \boldsymbol{\theta})$. Apparently, under the MAR process, the distribution of the outcomes for those who withdraw is not the same as the distribution in population. But conditional on the observed outcomes prior to the withdrawal occasion, the distribution of the unobserved outcomes following withdrawal is the same for those who do and do not withdraw at that occasion. However, this aspect of the assumption cannot be tested from the data at hand (Fitzmaurice, 2003).

Covariate-dependent missing is a special situation, where the missingness only depends on the fixed covariates X_i, but not on the outcomes $\boldsymbol{y_i}$, that is,

$$f(\boldsymbol{r_i}|\boldsymbol{y_i^o}, \boldsymbol{y_i^m}, X_i, \boldsymbol{\psi}) = f(\boldsymbol{r_i}|X_i, \boldsymbol{\psi}). \tag{1.8}$$

Molenberghs and Kenward (2007) viewed it as a special case of MCAR. But Little (1995) suggested the term 'MCAR' should be reserved for the case where the missingness dose not depend on either $\boldsymbol{y_i}$ or X_i, as shown in (1.2). When the missingness mechanism is only associated with the observed data (MAR), conditional on the fully observed data regardless of either the fixed covairates X_i or the observed outcomes $\boldsymbol{y_i}$, the missingness is completely at random (MCAR). Hence, it is reasonable to refer assumption (1.8) as 'covariate-dependent MAR' and refer assumption (1.5) as 'outcome-dependent MAR' (DeSouza et al., 2009).

Finally, when a *missing not at random* (**MNAR**) mechanism operates, the missingness depends on the unobserved outcomes $\boldsymbol{y_i^m}$, perhaps in addition to $\boldsymbol{y_i^o}$. The joint distribution of $\boldsymbol{y_i}$ and $\boldsymbol{r_i}$ (1.1) cannot be further simplified. Conditional on the past outcomes prior to the withdrawal occasion, the distribution of the future outcomes following withdrawal is different for those who do and do not withdraw. Clearly, the missingness mechanism is *non-ignorable*, and the distribution of the unobserved outcomes for those who withdraw is not estimable from the data on those observed after withdrawal. Therefore, the inference can only be made by the joint models of measurement and missingness processes, with model assumptions about missing data mechanism. Such models can be formulated in either the *selection* model or the *pattern-mixture* model framework. For selection models, the full data density is factored as the marginal density of the measurement process and the density of the missingness process, conditional on the outcomes (as shown in 1.1). While the pattern-mixture models use the reverse factorization,

$$f(\boldsymbol{y_i^o}, \boldsymbol{y_i^m}, \boldsymbol{r_i}|X_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y_i^o}, \boldsymbol{y_i^m}|\boldsymbol{r_i}, X_i, \boldsymbol{\theta}) f(\boldsymbol{r_i}|X_i, \boldsymbol{\psi}), \tag{1.9}$$

allowing a different response model for each pattern of the missing value.

For likelihood-based inference, MCAR and MAR with separate parameterization

of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are often called *ignorable*, because unbiased estimates of parameters can be obtained from observed data. The same requirement hold for Bayesian inference. But for non-likelihood frequentist inference, such as analysis of variance (ANOVA) and generalized estimating equation (GEE) method, MCAR is the only sufficient condition for ignorability. Either MNAR or MAR with parameters in common is called *non-ignorable*. When missingness is non-ignorable, the inference based on the likelihood ignoring the missingness process is biased (Molenberghs and Kenward, 2007).

1.4 Withdrawal Reasons in Clinical Trials

In general, no statistical test is available to assess from which mechanism the missing data arise. Hence, in longitudinal clinical trials, an intuitive way to explore missingness is to look at the reasons for withdrawals. Siddiqui et al. (2009) summarized the common reasons as follows: (i) recovery, (ii) lack of improvement, (iii) treatment-related side-effects, (iv) unpleasant study procedure, (v) intercurrent health problems, and (vi) external factors unrelated to the trials. For example, MCAR might result from a patient dropping out because of relocation (vi). If a patient was observed doing poorly and then decided to discontinue participation, then the withdrawal was related to the outcome of interest and was explained by the observed data. In this case, it is reasonable to assume MAR. An example of MNAR could be an instance in which a patient had been doing well and was then lost to follow-up due to a worsened condition after the last observed visit (Mallinckrodt et al., 2003).

However, so far, there is no formal classification of withdrawal reasons into the three missingness mechanisms defined by Little and Rubin (2002). This might partially be due to the fact that the definition of missingness mechanisms is based on the relationship between outcome and independent variables, and this relationship could vary from case to case (Siddiqui et al., 2009). For instance, withdrawals due to adverse events

are probably not MNAR in many cases because all the relevant data probably were observed. But whether classifying it as MCAR or MAR depends on the specific situation. In many clinical trial settings, extensive efforts are made to observe all the possible outcomes and the factors that influence withdrawal. Thus, the MAR assumption is much more plausible than the MCAR assumption, because the observed data could explain much of the missingness in many scenarios (Mallinckrodt et al., 2003). In principle, clinical trials by their very design seek to minimize the amount of MNAR. But, the possibility of MNAR can never be ruled out, and a mixture of different missingness mechanisms in a clinical dataset often takes place.

Within the framework of pattern mixture models, we develop sensitivity analysis methods for time-to-event data with possibly informative censoring. The chapter 2 presents a multiple imputation (MI) method for sensitivity analysis of continuous time-to-event data, invoking multiple imputation of the missing failure times due to withdrawals. In the chapter 3, we discuss the covariate-adjusted sensitivity analysis within the established framework. In the chapter 4, the underlying principle for this type of sensitivity analysis is extended to grouped time-to-event data.

1.5 Mixed-Effect Regression Model (MRM)

Under the assumption of an ignorable missingness mechanism, the likelihood-based mixed-effect models provide valid analysis for incomplete data from longitudinal clinical trials. When outcomes are continuous and Gaussian, linear mixed models of Laird and Ware (1982)are the sensible choice. And generalized linear mixed models could be used to analyze endpoints that are of non-Gaussian type. Using the maximum likelihood estimator, the missing values are treated as unknown random variables to be averaged over by integration and removed from the likelihood.

The linear mixed models assume that the vector $\boldsymbol{Y_i}$ of k repeated continuous measurements for the ith subject satisfies

$$\begin{aligned} \boldsymbol{Y_i}|\boldsymbol{b_i} &\sim N(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i}, \Sigma_i) \\ \boldsymbol{b_i} &\sim N(\mathbf{0}, G) \end{aligned} \tag{1.10}$$

where $\boldsymbol{\beta}$ is the p-dimensional vector of fixed effects; $\boldsymbol{b_i}$ is the q-dimensional vector of subject-specific random effects; and $X_i(k \times p)$ and $Z_i(k \times q)$ are the corresponding design matrices. It then can be easily derived that, marginally,

$$\boldsymbol{Y_i} \sim N(X_i\boldsymbol{\beta}, V_i) \text{ and } V_i = Z_iGZ_i' + \Sigma_i. \tag{1.11}$$

Conditional on $\boldsymbol{b_i}$, the outcomes Y_{ij}'s, $j = 1, 2, ..., k$, are usually assumed to be mutually independent. In this case, model residuals are uncorrelated, that is, $\Sigma_i = \sigma^2\boldsymbol{I}_{k\times k}$.

The generalized linear mixed models combine generalized linear model concepts with ideas from linear mixed models. It is assumed that, the conditional distributions of outcomes Y_{ij}', given $\boldsymbol{b_i}$, are independent, and in the form of the exponential-family

$$\begin{aligned} f_i(y_{ij}|\boldsymbol{b_i}, \boldsymbol{\beta}, \phi) &= exp\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y_{ij}, \phi)\} \\ \boldsymbol{b_i} &\sim N(\mathbf{0}, \boldsymbol{G}) \end{aligned}$$

with

$$\eta(\mu_{ij}) = \eta(E(Y_{ij}|\boldsymbol{b_i})) = \boldsymbol{x_{ij}'}\boldsymbol{\beta} + \boldsymbol{z_{ij}'}\boldsymbol{b_i}$$

for a known link function $\eta(\cdot)$, where ϕ is a scale parameter, and θ_{ij} is the canonoical parameter and satisfies $\mathrm{d}\psi(\theta_{ij})/\mathrm{d}\theta_{ij}$=$E(Y_{ij}|\boldsymbol{b_i})$=$\mu_{ij}$. Unlike linear mixed models, the marginal likelihood of the generalized linear model often involves the approximation of integration over the q-dimensional random effects. This is because, in most situations,

the integrals can not be solved analytically.

Typically, clinical trials are conducted to assess the fixed effects (i.e. the difference in treatment effects), rather than the subject-specific random effects. To accommodate this focus, the marginal linear mixed model (1.11) is often implemented in analyzing the continuous outcome measures with pre-specified time points. In this formulation, the random effects are modeled as a part of the marginal covariance matrix V, and the fixed effects could include treatment effect, time trend, treatment-by-time interaction and other covariates, such as baseline measurement or risk factors. Mallinckrodt et al. (2008) referred to this model as mixed-effects model for repeated-measures analysis (MMRM). A common implementation of MMRM is a cell mean model with unstructured within-subject error covariance structure V. More specifically, the outcome measure on the ith patient with treatment t at the jth occasion is modeled as $y_{ijt} = \mu_{jt} + \varepsilon_{ij}$, where μ_{jt} is the group mean of treatment t at the jth occasion and ε_{ij} is the jth element of the residual vector $\boldsymbol{\varepsilon_i}$ with $\boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, V)$. This version of MMRM is well suited to the general characteristics of clinical trials, i.e., common scheduled measurements for all patients and relatively a small number of measurement occasions. In addition, the unstructured covariance modeling relaxes the assumption about within patient correlation and often provides the best fit to the data.

As we previously noted, in many longitudinal clinical trial settings, the MAR assumption is always more plausible than the MCAR assumption. A MAR method is valid if data are MCAR or MAR, but MCAR methods are valid only if data are MCAR. The MMRM analysis uses all available data (observed outcomes) to provide information about the missing outcomes, via the within patient correlation structure. Under MAR, MMRM projects what would have happened if withdrawals continued to adhere to the protocol, and seeks to estimate the intervention effect that would be seen if all patients undertook the intervention as per the protocol (Carpenter and Kenward, 2008). Thus,

broadly speaking, MMRM analysis, as well as other MAR model-based methods, could be considered as a per-protocol(PP) analysis for the intent-to-treat(ITT) population.

Also, the MMRM analysis can be easily implemented using standard software, such as the SAS Procedure Mixed (SAS Institute, Cary, NC). A series of simulation studies have been conducted to compare the performance of MMRM with last observation carried forward (LOCF), in evaluating treatment efficacy under different missing data mechanisms (Mallinckrodt et al., 2008; Siddiqui et al., 2009). LOCF is an *ad hoc* method commonly used to impute missing values due to withdrawal for the primary efficacy analysis of clinical trials. Under MCAR or MAR, the MMRM analysis was able to estimate the true treatment difference with a negligible bias and control the type I error rate at a nominal level, whereas LOCF underestimated the standard error, and yielded increased bias and inflated type I error. Furthermore, in the presence of MNAR mechanism or a mixture of the three missing mechanisms, the MMRM analysis was also superior to LOCF in minimizing estimation bias and controlling type I and II error rates. In general, it has been recognized that the likelihood based MMRM is a robust and appropriate approach to handle missing data for primary efficacy analysis.

However, a shortcoming for MMRM and almost every model-based approach is that ignorable missingness depends on correct model spiflication. Missingness that might be non-ignorable given one model could be ignorable given another (Mallinckrodt et al., 2008). For instance, if withdrawal depends only on an observed variable, say treatment, and treatment is included in the analytic model, then the mechanism giving rise to the withdrawal would be ignorable (MCAR), and statistical analysis results would be unbiased. Whereas if treatment is not included in the analytic model, the missingness mechanism would be non-ignorable (MNAR), consequently the inference based on such a model would be biased.

1.6 Multiple Imputation (MI)

Multiple imputation was first introduced by Rubin (1987) to handle missing responses in a sample survey. Under ignorable missingness, it is another feasible approach for analyzing incomplete data. Three steps are involved in MI analysis:

1. The missing values ($\boldsymbol{y^m}$) are filled M times by an imputation model to generate M complete data sets. Each value is a random draw from the conditional distribution of the missing value given the observed data, in such a way that the imputations properly represent the information about the missing value in the imputation model.

2. Each of the M complete data sets is then analyzed using the method that would have been appropriate if the data had been complete (i.e., the analysis model).

3. The estimates of the desired quantities and associated standard errors from separate M analyses are combined into a single inference .

The MAR assumption needs to apply only for the first step, which can be carried out by the SAS procedure MI. The statistical inference performed in the second step will be valid, if the missingness given the imputation model is ignorable. The third step could be conducted by the SAS procedure MIANALYZE using the multiple imputation technology given by Rubin (1987).

More specifically, suppose that θ is the parameter of interest to be estimated using a analysis model. If complete data were available, the inference about θ given large samples would typically be based on the point estimate $\hat{\theta}(\boldsymbol{y^o, y^m})$, the variance estimate $\hat{V}_\theta(\boldsymbol{y^o, y^m})$, and the appropriate normal approximation

$$\frac{(\hat{\theta} - \theta)}{\sqrt{\hat{V}_\theta}} \sim N(0, 1). \tag{1.12}$$

Let $\hat{\theta}^m$ and V^m denote the point and variance estimates from the mth imputed data set ($m = 1, 2, ..., M$). The MI estimate for θ is simply the average of estimates from the M imputed complete data sets,

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}^m. \tag{1.13}$$

Also, define $W = \frac{1}{M} \sum_{m=1}^{M} V^m$ to be the average of the M within-imputation variances, and $B = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}^m - \bar{\theta})^2$ to be the between-imputation variance. Then, the variance estimator of $\bar{\theta}$ is given by the sum of W and B multiplied by a finite sample correction,

$$V = W + \frac{M+1}{M} B. \tag{1.14}$$

Confidence interval estimates and hypothesis tests are based on the approximation of a t distribution

$$V^{-\frac{1}{2}} (\bar{\theta} - \theta) \sim t_\nu \tag{1.15}$$

with degrees of freedom $\nu = (M-1)(1 + r^{-1})^2$, where $r = (1 + M^{-1})B/W$. Thus, a $100(1-\alpha)\%$ confidence interval estimate for θ is $\bar{\theta} \pm t_{\nu,(1-\frac{\alpha}{2})} V^{\frac{1}{2}}$; and a two-sided p-value for the null hypothesis $H_0 : \theta = 0$ is obtained by comparing $\bar{\theta}/V^{\frac{1}{2}}$ with the distribution of t_ν.

Notice that if $\boldsymbol{y^m}$ carried no information about θ, the imputed data estimates $\hat{\theta}^m$ would be identical and V would reduced to W. Therefore, r, the ratio of the between-imputation to the within-imputation variation, measures the relative increase in variance due to missing information. When there is no missing information about θ the values of r and B are both zero. The rate of missing information in the system can be

obtained by comparing the spread of the distribution in (1.12) to the distribution in (1.15) as

$$\gamma = \frac{r + 2/(\nu + 3)}{1 + r}. \tag{1.16}$$

Unlike maximum likelihood (ML) based mixed-effect models, the MI approach handles missing data in a imputation step entirely separate from the analysis. The model used to impute missing values can differ from the model used for inference (Liu and Gould, 2002). An advantage about the MI method, compared to the ML mixed-effect model approach, is that the imputation model can include considerably more variables predictive of missingness, therefore increasing the plausibility of the MAR assumption (Mallinckrodt et al., 2003). Besides the variables that are potential causes or correlates of the missingness itself, the imputation model could also include the variables that are simply correlated with the variables that have missing values regardless of their relation to the missingness mechanism. Collins and colleagues referred to this class of variables as auxiliary variables (Collins et al., 2001). Their simulation study indicated that incorporating the auxiliary variables not only increased efficiency and statistical power under MAR, but also reduced the bias and moved the situation closer to MAR when missingnes was non-ignorable. In clinical trial settings, the auxiliary variables, such as postbaseline or time varying covariates correlated with treatment could be included in the imputation model. But those variables cannot be included in the analysis model or the MMRM model because of their confounding with the treatment effect (Mallinckrodt et al., 2008).

From a theoretical standpoint, it is possible to add auxiliary variables to the ML mixed-effect model. However, without a careful plan, it could produce undesirable effects, such as altering the meaning of model or computational complexity. Furthermore, the statistical software that is currently available for ML missing data approaches also

do not facilitate the use of auxiliary variables. Using MI, it is much easier to incorporate auxiliary variables into the analysis via the imputation step. For this reason, the MI method is practically more robust against misspecification of the missingness mechanism than the direct likelihood method.

1.7 Sensitivity Analysis

1.7.1 Why Sensitivity Analysis

When analyzing incomplete data, additional assumptions about missing data have to be made in order to choose sensible statistical methods. Those assumptions break down into two broad categories as follows (Carpenter and Kenward, 2008). The first type of assumptions focuses on the missingness mechanism, specifically, how the probability of missingness depends on the observed and unobserved data. This leads to the selection model approach. The second type of assumptions focuses on the distribution of missing data given the observed data, that is, whether or how the distribution of unobserved outcomes is different from the distribution of outcomes for those who have no missing data, and how this difference depends on the patterns of the missingness. This is the extension of pattern mixture models. In section 1.3, the connections between the missingness mechanisms and the assumptions of missing data distribution were discussed in the context of MCAR, MAR, and MNAR.

Depending on the statistical approach we adopt to handle the missing data, we need to make assumptions about either the missingness mechanism or the missing data distribution. Unfortunately, neither of those assumptions can be validated definitely from the observed data. For instance, MCAR indicates the distribution of outcomes prior to withdrawal is the same for those who do and do not withdraw at that occasion, which can be suggested by the observed data. But we cannot be completely sure that data

are MCAR, since we do not observe the missing outcomes. Nevertheless, the observed data can rule out MCAR, if there is any relationship between the observed data and the occurrence of missing data (Carpenter and Kenward, 2008). In clinical trials, MCAR is most unlikely, and MAR is often reasonable; however, the possibility of MNAR is impossible to be ruled out. Results from likelihood-based MAR methods without consideration of their limitations could be misleading (Mallinckrodt et al., 2003). Thus, In many circumstances, analysis valid under MNAR assumption is required. However, one obvious and fundamental problem is that the conclusion of the MNAR analysis is conditional on the appropriateness of the assumed model (Mallinckrodt and Kenward, 2009), and MNAR model assumptions are not verifiable from the observed data. More importantly, the consequences of model miss-specification are more severe with MNAR methods than with MAR methods. Hence, no single MNAR approach can be considered definitive or well suited for the primary analysis in confirmatory clinical trials.

Given the above discussion, no universally best applicable approach for handling missing data can be recommended. Investigators should make effort in exploring the impact of missing data assumptions on the results of analysis. For longitudinal clinical trial data, it has been a broad consensus among statisticians in the pharmaceutical industry and academia that likelihood-based methods implemented under the MAR framework are the sensible choice for the primary analysis, and a series of MNAR analyses should be implemented as *sensitivity analysis* to assess the robustness of the primary analysis result to the possible departure from the MAR assumption (Mallinckrodt and Kenward, 2009).

In presence of incomplete data, *sensitivity analysis* is generically defined as a way to explore the impact of missing data assumptions on statistical inferences and scientific

conclusions (Molenberghs, 2009). In clinical trial settings where interest is in the treatment efficacy hypothesis, a pre-defined sensitivity analysis approach was recommended by Carpenter and Kenward (2008) as follows. Either before the trial is conducted, or during a blind review of the data, investigators identify a set of missing data models (i.e., assumptions about missingness mechanism or missing data distribution) to address the impact of clinically plausible departure from MAR. A sensible analysis is then planned under each missing data model. After the blinding is broken, a series of analyses are performed. The results from such analyses would reflect a range of conclusions under assumed models, and therefore demonstrate the robustness of the inference about treatment comparisons to different missing data assumptions.

1.7.2 Selection and Pattern-mixture Models

The sensitivity of inference to missing data assumptions is typically investigated via selection or pattern-mixture models under MNAR framework. Carpenter and Kenward (2008) illustrated the application of both approaches using clinical trial examples.

Selection models (Diggle and Kenward, 1994; Little and Rubin, 2002) describe the full data likelihood as the product of the marginal density of the measurement process and the density of the missingness process conditional on outcomes (shown in 1.1). Besides the model for the measurement process as used in MAR analysis, a model explaining missingness has to be fitted at the same time in the analysis. Carpenter and Kenward (2008) gave an example of the missingness process model for withdrawal as

$$\begin{aligned} \text{logit}\Pr(R_{ij} = 0 | R_{i(j-1)} = 1) &= \alpha_j + \delta Y_{ij} \\ \Pr(R_{ij} = 0 | R_{i(j-1)} = 0) &= 1, \end{aligned} \tag{1.17}$$

where Y_{ij} denotes the response of interest for the ith subject at the jth visit, $i = 1, ..., n$

and $j = 1, 2, ..., k$. Here, the log odds of subject i withdrawal at the jth visit depends not only on that visit (α_j), but also on the response at that visit. $\delta = 0$ corresponds to a MAR assumption, and a positive value of δ implies odds of withdrawal at a visit increases with the value of response at that visit. Using MCMC methods in winBUGS (Spiegelhalter *et al.*,1999), this model can be fitted in conjunction with a mixed model for the measurement process. However, the estimated value of δ and its standard error depend critically on unverifiable assumptions for the missing data distribution. Thus, estimating δ is usually not recommended (Carpenter and Kenward, 2008). Rather, investigators should identify a set of plausible values for δ, then fit the model with each of these values, and explore how the estimated treatment effect varies as a sensitivity analysis. In many applications, model (1.17) could be extended to include other variables, such as, treatment and response from previous visit. Theoretically, variables that are included in a mixed model to maximize the plausibility of MAR should be included to make the selection model more plausible and make the treatment estimates less sensitive to MNAR. However, those complicated models are sometimes difficult to fit and interpret.

An alternative approach to sensitivity analysis is the pattern-mixture (PM) models (Little, 1993). PM models are based on factorization of the full data likelihood as the product of the measurement process conditional on the withdrawal pattern and the missingness process (displayed in 1.9). An important feature of PM models is that the parameters are overspecified, hence additional restrictions are needed to make parameters identifiable (Molenberghs and Kenward, 2007). One general rule is to set the non-estimable parameters of incomplete patterns equal to the parameters or functions of the parameters that describe the distribution of completers (Little, 1994; Mallinckrodt et al., 2003). In practice, the random-effect PM model formulated by Little (1995) is often used to analyze incomplete clinical trial data. To implement this approach,

analytical models are fitted separately for different groups, defined by withdrawal patterns (e.g. early or late withdrawals, and completers), and then the overall estimate is obtained as a weighted average of the pattern-specific estimates (Siddiqui et al., 2009; Ali and Siddiqui, 2000).

A natural way to perform sensitivity analysis under the PM framework is to assume that the model for missing data is a modification of the model for observed data. By varying this modification, a range of sensitivity analyses could be easily performed. Carpenter and Kenward (2008) illustrated this type of sensitivity analysis using an example with a single response from normal distribution. For the control arm, assume that the observed responses come from $normal(\mu_C, \sigma^2)$ and the unobserved responses have shifted mean $(\mu_C + \delta_C)$. Let π_C denote the probability of withdrawal, then the average response in the control arm is $(1 - \pi_C)\mu_C + \pi_C(\mu_C + \delta_C)$. Likewise, for the intervention arm, μ_I, δ_I, and π_I are defined analogously, then the average response in the intervention arm is $(1 - \pi_I)\mu_I + \pi_I(\mu_I + \delta_I)$. The averaged treatment effect, Δ, is then

$$\begin{aligned} \Delta &= ((1 - \pi_I)\mu_I + \pi_I(\mu_I + \delta_I)) - ((1 - \pi_C)\mu_C + \pi_C(\mu_C + \delta_C)) \\ &= (\mu_I - \mu_C) + (\delta_I\pi_I - \delta_C\pi_C). \end{aligned} \tag{1.18}$$

Note that $(\mu_I - \mu_C)$ is the treatment effect in completers, δ_I and δ_C are the parameters describing the degree of informative missingness, and $\delta_I = \delta_C = 0$ implies MCAR. Using a Bayesian approach, the sensitivity analysis could be performed via a series of plausible prior elicitation about the joint distribution of δ_I and δ_C.

1.8 Intent-to-treat Analysis and Sensitivity Analysis

1.8.1 Intent-to-treat and Per-protocol Analysis

Intent-to-treat(ITT) has been accepted as a fundamental principle for establishing efficacy/effectvieness in randomized confirmatory clinical trials. ITT analysis attempts to evaluate the effect of the original treatment strategy rather than directly assess the effect of treatment itself (Flyer and Hirman, 2009). It includes all randomized patients, regardless of adherence to the protocol or premature withdrawal. In a typical ITT analysis, the treatment effect is measured with patients assigned to the treatment as randomized, rather that to the treatment actually received (Little and Yau, 1996).

In contrast to ITT analysis, per-protocol (PP) analysis incorporates actual treatment usage and compliance into a direct measurement of treatment effect. In the presence of missing data due to withdrawals, PP analysis seeks to estimate the treatment effect that would have been seen if all patients completed their assigned treatment. In this regard, PP analysis is consistent with the MAR assumption, under which the missing data problem in longitudinal data can be properly addressed by ML based mixed-effect models (Carpenter and Kenward, 2008; Flyer and Hirman, 2009).

The only ITT analysis approach that could address missing data is to continue following patients after they stop adhering to the intervention protocol, regardless whatever treatment they then receive. However, practical issues in data collection and trial design make its implementation difficult (Flyer and Hirman, 2009), for instance, patients may withdraw consent. According to the ITT principle, data from patients who stop treatment but do not begin other treatments should reflect the long-term benefit of the initial treatment assignment. However, data collected from patients that received non-study treatments will be very difficult to interpret. Because of those reasons, many clinical trials are designed to not collect data on patients who discontinue their study

treatment. Very often, when patients stop complying with the protocol, they withdraw from the study, leading to missing responses.

1.8.2 PM Model with Longitudinal Data via MI

Because of the distinction in the underlying hypothesis between ITT and PP analyses, methods to address the missing data issue for ITT analysis should be different correspondingly from those for PP analysis. We should consider that the distribution of responses is different for those who do and do not continue with the intervention protocol, and even after taking into account the information in observed data, the withdrawal mechanism may still depend on the unseen response (Carpenter and Kenward, 2008). Furthermore, if the withdrawal is likely to be associated with a change in treatment regime, then a different model is needed for those who withdraw. However, we may not actually have enough information to model patients' treatment adherences and behaviors following withdrawal. Therefore, the missing data issue following withdrawal in the ITT setting should be viewed as a MNAR analysis under specific assumptions (Carpenter and Kenward, 2008).

Under MNAR, a range of missingness process models could be consistent with the observed data. Thus, there is no longer a definitive ITT analysis (Carpenter and Kenward, 2008). Different post-withdrawal behaviors should be incorporated into sensitivity analyses as part of the ITT analysis. A natural approach for such a sensitivity analysis is multiple imputation using PM models conditional on all relevant observed data and different missing patterns to reflect treatment change (or discontinuation) after dropout. Little and Yau (1996) proposed a PM imputation model including the treatment dosage actually received after withdrawal, upon which multiple imputations were created by sequential Bayesian draws from the parameter posterior distribution and the missing value predictive distribution conditional on the drawn parameters. The

sensitivity analysis was then carried out for a range of plausible alternative assumptions about dosage after withdrawal. A key feature of this method is that the imputation model differs from the model used for ITT analysis. Under the ITT principle, the analysis model must not include compliance information following randomization, but the imputation model has no problem to incorporate such compliance information (Carpenter and Kenward, 2008).

Furthermore, under simple but reasonable assumptions about the consequence of treatment regime change on future behavior, the PM imputation model can be constructed from components estimated from a MAR mixed-effect model, therefore greatly simplifying the modeling process (Carpenter and Kenward, 2008). This approach was implemented in a SAS macro by Roger (2008). Suppose we are interested in ITT treatment effects at the final visit of an active drug group and a placebo group, and assume that patients discontinue their treatment after withdrawal. Then, it is reasonable to apply a MAR model to the placebo group, and assume that, for the active treatment group, the post-withdrawal behavior of response conditional on the past is probably the same as those in the placebo group. Thus, MAR model estimations from the placebo group can be used to construct a MNAR PM model to impute the future responses of withdrawals in the active drug group. Based on this idea, a variety of PM models can be constructed under various plausible assumptions about post-withdrawal behaviors in the active drug group, to assess the robustness of the primary analysis result. For example, one PM model hypothesis could be that the change in mean for post-withdrawal responses in the active drug group is the same as the change in mean in the placebo group, while another model could assume that the mean of post-withdrawal responses in the active drug group become the mean in the placebo group. This method and the standard MAR imputation only differ with respect to their implications for unobserved responses from the active drug group. And this difference, in turn, leads to different

treatment comparison estimates (Carpenter and Kenward, 2008).

1.9 Informative Censoring and Sensitivity Analysis

Survival analysis is used in medical research for analyzing time-to-event (survival) data, which is the time duration from a well-defined time origin to the occurrence of a particular event. A distinguishing feature of time-to-event data is that they are usually censored or incomplete in some way. The survival time is said to be (right) censored if the observation period was cut off before the event occurred. In such case, we do not know when (or, indeed, whether) the patient will experience an event, and we only know that the patient has not had the event at the last observed time.

In a typical setting for analyzing censored survival data, there is a potential censor time C and a potential survival time T for each individual. We only observe time $Y = \min(T, C)$ and the censor indicator $\delta = I(T \leq C)$, which indicates whether a failure is observed. Under this condition, the distribution of T is not identifiable unless we make further assumptions. An important assumption in most survival analysis is that the censoring mechanism is *non-informative* or *ignorable*. This means that the instantaneous failure rate given subject survival to time t is not changed by additional information that the subject was uncensored up to that time t (Leung et al., 1997; Lagakos, 1979). In another words, the censoring of an observation does not provide any information regarding the prospect of survival time of that particular subject beyond the censor time. The contribution of a censored observation to the likelihood is just the probability that survival time T exceeds the censor time c. Therefore, the censoring mechanism is irrelevant for inference about the distribution of T.

In general, censoring due to study termination is called end-of-study or administrative censoring, as apposed to loss-to-follow-up (LTF) censoring when patients withdraw during the study period (Leung et al., 1997). It is usually reasonable to assume that the

administrative censoring is independent of the survival time. Lagakos (1979) proved that independent censoring is a special case of the noninformative censoring; hence the administrative censoring imposes no problem to the standard survival analysis procedures. However, it is not always proper to make non-informative assumptions about the loss-to-follow-up censoring. When the probability of censoring depends on the survival time, the censoring mechanism is said to be *informative* (Kalbfleisch and Prentice, 2002), and the inference based on the standard methodologies is no longer valid (Collett, 2003). If patients who withdraw are at higher risk of subsequently having an event (i.e., the survival time and the censor time are positively correlated), the Kaplan-Meier estimator would overestimate the survival function of T. Similarly, if the withdrawals are at lower risk of failure (i.e., the survival time and the censor time are negatively correlated), then the survival function would be underestimated (Leung et al., 1997). Consequently, the main issue, related to the informative censoring in clinical trials, is the potential bias in comparison of survival functions between treatment groups.

When censoring is informative or dependent, the focus of analysis is usually the joint distribution of T and C. Unfortunately, this joint distribution is not identifiable from the observed data Y and δ. Tsiatis (1972) showed that if we have a model with dependent risks, then a proxy model with independent T and C always exists, and those two models give the same joint distribution of observed Y and δ. Therefore, unless we make additional assumptions, it is impossible to test the noninformative assumption, or to estimate the degree of dependence between failure time and censoring mechanism.

Given that no definite model can be fit under dependent censoring in practice, many references have proposed sensitivity analysis about informative censoring in a competing risk framework. Additional assumptions and models were introduced on the dependence structure between the failure time and the dependent censoring process in order to

make the problem identifiable. Scharfstein and Robins (2002) proposed classes of semi-parametric models for the cause-specific hazard of censoring indexed by censoring bias functions. Their method allows adjustment for informative censoring due to measured prognostic factors for failure time, and simultaneously quantifies the sensitivity of the inference to residual dependence between failure and censoring due to unmeasured factors. Siannis et al. (2005) studied the sensitivity analysis for informative censoring in parametric survival models. The association between T and C was modeled by the conditional distribution of C given the (possibly unobserved) T through a bias function and a dependence parameter. The sensitivity analysis was developed on a range of the dependence parameters for small values. Recently, Ruan and Gray (2008) investigated the effect of the dependence between dependent withdrawal and disease progression time using a conditional probability model incorporating a set of sensitivity parameters. They linked the sensitivity analysis to the log-rank test via constructing log-rank-type score statistics from the estimated distributions, and they explored the impact of the sensitivity parameters on the inference by examining how much dependence between the disease progression and withdrawal times would need to be present to ultimately change the conclusion from the statistic for inference about the comparison of two treatment groups.

1.10 Summary

Within the framework of pattern mixture models, we develop sensitivity analysis methods for time-to-event data with possibly informative censoring. The chapter 2 presents a multiple imputation (MI) method for sensitivity analysis of continuous time-to-event data, invoking multiple imputation of the missing failure times due to withdrawals. In the chapter 3, we discuss the covariate-adjusted sensitivity analysis within the established framework. In the chapter 4, the underlying principle for this

type of sensitivity analysis is extended to grouped time-to-event data.

Chapter 2

Sensitivity Analyses of withdrawals in Time-to-Event Data

2.1 Introduction

An essential property of confirmatory clinical trials is the randomization of patients so that the control and the test treatment have statistically equivalent distributions for known and unknown baseline characteristics that may have potential associations with the outcome of interest (NRC, 2010; CHMP, 2010). However, a ubiquitous and inevitable problem that can undermine the comparability of randomized treatment groups is potential bias from the nature and extent of missing data for patients who prematurely discontinue their planned follow-up period for the assigned treatment (or the study) without further assessment. In view of this problem, the design of many clinical trials specifies continued follow-up of patients after premature termination of the assigned treatment for such reasons as adverse events, lack of compliance, lack of efficacy, or protocol deviations. A rationale for this practice is that it provides potentially useful information about the experiences of these patients for their remaining follow-up time until their planned (or premature) discontinuation from the study (Flyer and Hirman, 2009; Walton, 2009). However, the role of this information can be unclear when patients receive effective rescue treatment after discontinuing their assigned

treatment (Flyer and Hirman, 2009). For example, the comparison of regimens that begin with test treatment or placebo followed by effective rescue therapy after their discontinuation could erroneously suggest that an ineffective test treatment is effective solely because it forces more patients to switch to rescue therapy than placebo (Permutt and Pinheiro, 2009). Thus, in such situations, analyses for the comparison of the assigned treatments may need to ignore any unclear information subsequent to their discontinuation and thereby proceed with the corresponding experiences of patients as if missing.

Analytical strategies for drawing inferences from incomplete data rely on untestable assumptions about the missing data distributions and the missingness mechanism (NRC, 2010; CHMP, 2010). Little and Rubin (2002) classified the missing data mechanism into three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). When the data are MCAR, the missing data are unrelated to the observed and unobserved study variables, and so the observed data are statistically representative for the experiences of all randomized patients. In practice, however, MCAR is usually an unrealistic assumption. When the data are MAR, the missingness depends only upon the observed study variables. That is, conditional on the observed study variables, the probability of missing does not depend on the values of the missing data. When the missingness probability also depends on the values of the missing data, the data are said to be MNAR. In many situations, the MAR paradigm is realistic for the primary analysis in confirmatory clinical trials (Zhang, 2009; Mallinckrodt et al., 2008). However, the observed data can never rule out the possibility of MNAR. Therefore, sensitivity analyses exploring the implications of departures from the primary MAR assumption are always of interest to assess the robustness of the treatment effect inferences.

We consider randomized clinical trials where a time-to-event is the primary outcome.

Conventional methods such as the Kaplan-Meier estimation of survival curves (Kaplan and Meier, 1958), the logrank or Wilcoxon tests (Mantel, 1966; Gehan, 1965), and the Cox proportional hazards model (Cox, 1972) are frequently employed to describe time-to-event distributions and to assess treatment effects. Missing data for a time-to-event occurs for patients who prematurely discontinue follow-up for the assigned treatment (or the study) prior to the occurrence of the event or the end of their planned follow-up period (or the administrative closing date of the study). One way to address this type of missing data is to censor the follow-up times of such patients at their times of premature discontinuation. Such censoring is non-informative in a sense like the MAR assumption (Heitjan, 1994) when the assumption of its independence from the possibly unobserved time-to-event applies; i.e., the possibly unknown true time to the event for a patient is the same regardless of whether or not it is actually observed (or whether censoring occurs or not prior to it). Unfortunately, the conventional MAR-like methods ignore the fact that the patients who discontinue the assigned treatment no longer receive it after discontinuation. Instead, they attempt to estimate what would be expected for the study if all patients remained on their assigned treatments until the occurrence of the event or the end of their planned follow-up period (Flyer, 2009).

Alternatively, discontinuation from treatment can be specified as clinical failure when the event of interest is unfavorable (Flyer and Hirman, 2009). In this case, one has a composite endpoint (i.e., time to the event of interest or discontinuation), and it expresses the time period for which a patient has had favorable experience with treatment. The application of this method to both the control and the test treatment groups produces what can be called the worst-case analysis, because patients who discontinue treatment are managed as having much higher risk of a future event than other patients (Rothmann et al., 2009). In contrast, the method in which a control patient who discontinues treatment has their follow-up time censored at the time of

discontinuation and such a test treatment patient is managed as having the event is known as a worst-comparison analysis (Rothmann et al., 2009). The result from the worst-comparison analysis provides a stringent boundary on the impact of patients who discontinued treatment. Both the worst-case analysis and the worst-comparison analysis have potentially unclear relevance for a study because they both make unrealistic assumptions (Wittes, 2009). Usually, they are not designated as the primary analysis, but they can be used as sensitivity analyses with the worst-comparison analysis invoking maximal stress to the robustness of the study results (Walton, 2009). If the study conclusions are not altered by such methods, then one is reassured regarding the validity of the primary MAR-like analysis. Nevertheless, many studies will not maintain robustness to such sensitivity analyses. Hence, these methods are often criticized as unrealistically stringent and potentially problematic for a promising therapy to show effectiveness (Yan et al., 2009).

For longitudinal data with discontinuing patients, Little and Yau (1996) proposed multiple imputation of the missing responses on the basis of models incorporating actual treatment doses that might apply, or imputed doses under a variety of plausible assumptions. Recently, using a similar basic approach, Roger (2008) developed a sensitivity analysis, where the estimates from a mixed-effects model in the placebo group were used to provide information about possible future behaviors of discontinued patients from the test treatment. In this paper, we propose a related sensitivity analysis for time-to-event data. On the basis of Kaplan-Meier (KM) estimators (or Cox proportional hazards model counterparts), patients who discontinue their assigned treatment (or follow-up) have multiple imputations for their experiences during their unobserved remaining times until the planned end of their follow-up period (as if they continued to be followed). The imputed data sets, having only administrative censoring of follow-up for patients who did not have the event by the end of their planned follow-up period,

can then be analyzed by the standard methods for right censored time-to-event data. A key feature of this multiple imputation method for sensitivity analyses is a corresponding hazard ratio parameter θ for how the conditional survival distribution for the missing extent of follow-up can allow for different post-discontinuation behaviors of patients from the placebo and the test treatment groups. One can then investigate the impact of departures from the primary missingness assumption (i.e., non-informative independent censoring) by summarizing the treatment effect as a function of θ over a plausible range. This multiple imputation method is an extension and modification of the work by Taylor et al. (2002), where the conditional KM estimators were used to impute failure times for survival analyses under a specification for non-informative censoring. The implementation of this method is illustrated with data from a clinical trial in psychiatry.

2.2 Clinical trial examples

For illustrative purposes, we consider time-to-event data based on a clinical trial pertaining to maintenance treatment for bipolar disorder (Calabrese et al., 2003). For reasons related to the confidentiality of the data from this clinical trial, the example in this paper is based on a random sample (with replacement) of 150 patients with the test treatment and 150 patients with placebo. The study design for this clinical trial had an 8 to 16 weeks run-in period within which all patients received test treatment. Eligible patients who tolerated and adhered to this therapy were randomized to the test treatment or to the placebo, and then followed for up to 76 weeks as the planned follow-up period. Accordingly, this study had a randomized withdrawal design, and the primary efficacy endpoint was the time-to-intervention for any mood episode.

A total of 97 (32.33%) patients discontinued the study prematurely (35% on placebo and 29% on test treatment). Cumulative proportions of discontinued patients are shown

in Figure 2.1 (which has the convention of managing the patients who completed the study with the primary event as having imputed follow-up of 76 weeks without premature discontinuation). Discontinuations predominantly occurred before 35 weeks with higher cumulative proportions for the placebo group. The documented reasons for discontinuation are summarized in Table 2.1, although except perhaps for "adverse events", they are not informative about possible missing data mechanisms. The cumulative proportions of discontinuation by those reasons are displayed for each treatment arm in Figure A.1 of the Appendix. For an informal evaluation of the association of discontinuation with treatments, patients' demographics, and baseline psychiatric assessments, we used logistic regression models for the odds of discontinuation versus completion of the study (either with the primary outcome or completion of 76 weeks of follow-up without it). As shown in Table 2.2, neither the unadjusted (from univariate regression on each individual variable) nor the adjusted (from multivariate regression on all the variables) odds ratios have p-values below 0.05 for any of the baseline variables or treatments. However, in view of the substantial extent of discontinuations, sensitivity analyses to address the robustness of conclusions to the management of missing information are of interest.

Table 2.1: Discontinuations and the corresponding reasons by treatment groups

Disposition	Overall		Treatment group			
			Placebo		Test treatment	
	N	%	N	%	N	%
Completed study without episode	46	15.33	15	10.00	31	20.67
Intervention for a mood episode	157	52.33	82	54.67	75	50.00
Discontinued study prematurely	97	32.33	53	35.33	44	29.33
Adverse event	24	8.00	16	10.67	8	5.33
Consent withdraw	28	9.33	13	8.67	15	10.00
Lost to follow-up	20	6.67	8	5.33	12	8.00
Protocol violation	10	3.33	3	2.00	7	4.67
Other (including missing data)	15	5.00	13	8.67	2	1.33

The primary time-to-event analysis for this example has censoring of follow-up time

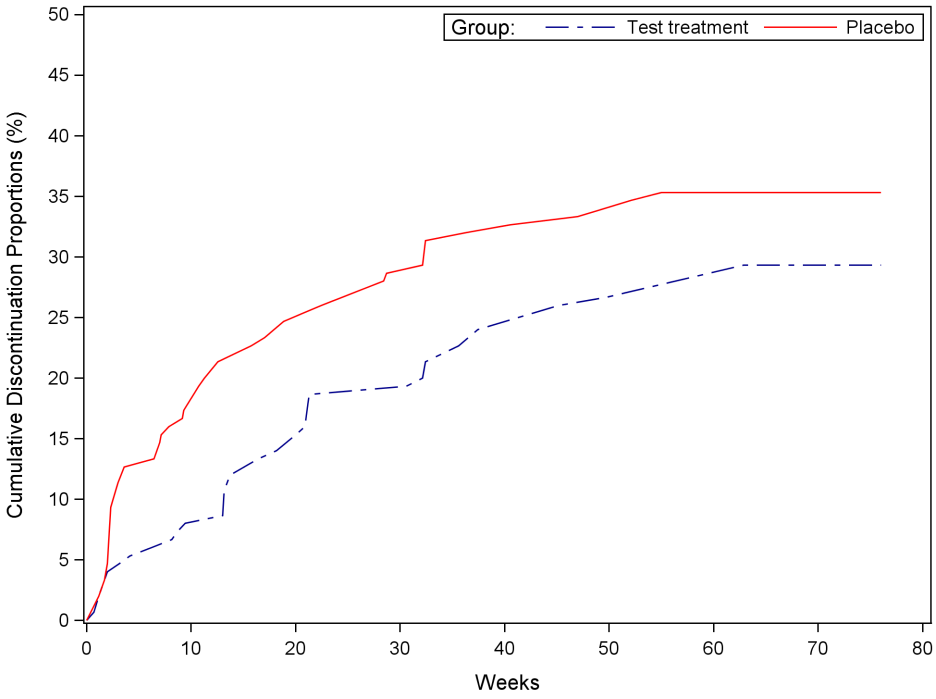


Figure 2.1: Cumulative discontinuation proportions by treatment groups

for patients with premature discontinuation of treatment, and so it has the MAR-like assumption of non-informative independent censoring. The previously noted worst-case analysis and the worst-comparison analysis serve as sensitivity analyses. The Cox proportional hazards model with one explanatory variable for treatment is used to obtain an unadjusted hazard ratio. The non-parametric logrank and Wilcoxon tests are also used to compare the test treatment and placebo. The results from these analyses are summarized in rows (1A), (1B), and (1C) of Table 2.3, and they are interpretable as indicating superiority of the test treatment. The worst-case analysis provides stronger results in favor of test treatment, whereas the worst-comparison analysis shows no treatment difference. The worst-case analysis tends to overstate the difference in favor of the test treatment because the placebo group has more prematurely discontinued patients. Conversely, the worst-comparison analysis excessively understates the difference in favor of test treatment by unrealistically managing all of its patients with

premature discontinuation as having events at the time of discontinuation. Therefore, more realistic approaches are worthy of consideration for sensitivity analyses to address robustness of conclusions for a clinical trial like this example to possibly informative censoring of time-to-event data.

Table 2.2: Unadjusted and adjusted odds ratios for discontinuation

Baseline characteristics	Univariate logistic regression			Multivariate logistic regression		
	Coef.[a] (StdErr.[b])	OR (95% CI)	P value	Coef. (StdErr.)	OR (95% CI)	P value
Treatment (test vs. placebo)	-0.275 (0.248)	0.76 (0.47, 1.23)	0.2671	-0.242 (0.253)	0.79 (0.48, 1.29)	0.3398
Age (1 year increment)	-0.012 (0.010)	0.99 (0.97, 1.01)	0.2423	-0.011 (0.011)	0.99 (0.97, 1.01)	0.2941
Gender (female vs. male)	0.111 (0.247)	1.12 (0.69, 1.81)	0.6527	-0.010 (0.258)	0.99 (0.60, 1.64)	0.9695
Pre-rand[c] CGI-I score[d]	0.301 (0.208)	1.35 (0.90, 2.03)	0.1474	0.584 (0.337)	1.79 (0.93, 3.47)	0.0828
Pre-rand CGI-S score	0.021 (0.168)	1.02 (0.74, 1.42)	0.9018	-0.348 (0.277)	0.71 (0.41, 1.21)	0.2085
Pre-rand GAS score	-0.002 (0.012)	1.00 (0.98, 1.02)	0.8719	0.010 (0.017)	1.01 (0.98, 1.04)	0.5580
Pre-rand MRS 11 item total score	-0.012 (0.047)	0.99 (0.90, 1.08)	0.7962	-0.036 (0.050)	0.96 (0.87, 1.06)	0.4690
Pre-rand MRS 17 item total score	0.028 (0.029)	1.03 (0.97, 1.09)	0.3433	0.036 (0.038)	1.04 (0.96, 1.12)	0.3529

[a]Coefficient

[b]Standard Error

[c]Pre-randomization

[d]1 unit increment for all score variables

Table 2.3: Analyses of treatment comparisons for delaying time-to-intervention for any mood episode

Analysis method	Semi-parametric analysis (Cox PH model)						
	time period	Coefficient	Std Err	HR (95% CI)	P value	Log-rank[a]	Wilcoxon[a]
1. Original primary and sensitivity analyses							
(1A) MAR-like[b]	Whole period	-0.393	0.161	0.675 (0.493, 0.925)	0.0144	0.0149	0.0048
	0-3 weeks	-0.728	0.284	0.483 (0.277, 0.842)	0.0103		
	4-5 weeks	-0.700	0.356	0.497 (0.247, 0.998)	0.0495		
	6-20 weeks	-0.096	0.342	0.908 (0.465, 1.774)	0.7782		
	21-76 weeks	0.138	0.367	1.148 (0.559, 2.357)	0.7065		
(1B) Worst case	Whole period	-0.484	0.127	0.616 (0.481, 0.790)	0.0001	0.0001	<0.0001
(1C) Worst comparison	Whole period	0.033	0.144	1.033 (0.779, 1.371)	0.8212	0.7983	0.3827
2. Kaplan-Meier multiple imputation (KMMI) method at $L = 50$							
(2A) $\theta = 1$	Whole period	-0.322	0.160	0.724 (0.530, 0.991)	0.0436	0.0451	0.0109
3. Proportional hazard multiple imputation (PHMI) method at $L = 50$							
(3A) $\theta = 1$	Whole period	-0.388	0.158	0.678 (0.497, 0.925)	0.0143	0.0145	0.0053
4. KMMI method with bootstrap at $L = 500$							
(4A) $\theta = 1$	Whole period	-0.336	0.170	0.714 (0.512, 0.997)	0.0480	0.0507	0.0114
5. PHMI method with bootstrap at $L = 500$							
(5A) $\theta = 1$	Whole period	-0.396	0.160	0.673 (0.492, 0.921)	0.0134	0.0140	0.0053

[a]P values of hypothesis test

[b]MAR-like: censoring patients at the time of discontinuation

2.3 Method

In most clinical trials with time-to-event data, a primary analysis that has censoring of follow-up time for patients with premature discontinuation of treatment is generally reasonable. The primary MAR-like assumption for such analysis is non-informative independent censoring. The proposed sensitivity analysis in this paper addresses the implications of departures from this assumption by imputing different outcomes for the patients with premature discontinuation. It thereby enables assessment of the robustness of the results from the primary analysis with censoring of follow-up times for patients with premature discontinuation.

Consideration is first given to a Kaplan-Meier multiple imputation (KMMI) procedure and its separate invocation for the placebo group and the test treatment group. For this purpose, we describe the KMMI strategy for a single treatment group with n patients who have the same planned follow-up time t^*. For the ith patient, we observe time $Y_i = \min(T_i, C_i)$, where T_i and C_i are the potential time-to-event and time to premature discontinuation (or censoring) for the patient. We define the censoring indicator $\delta_i = I(T_i \leq C_i)$, so that the data can be summarized by (Y_i, δ_i) for $i = 1, 2, \ldots, n$. We assume that a study has events observed at M distinct times $(t_1 < t_2 < \cdots < t_M)$, and it has premature discontinuation of patients observed at K distinct times $(c_1 < c_2 < \cdots < c_K)$. Also, there may be more than one patient with the same times at risk y_i (i.e., occasionally tied t's or c's), and we assume that $y = t^*, \delta = 0$ for at least one patient who completes the entire planned follow-up time without the event (and has administrative censoring of their follow-up time at t^*).

2.3.1 Kaplan-Meier Multiple Imputation Strategy

To establish the notation further, let k index the censoring times before t^*. Let $t_{k,0}$ denote the latest failure time prior to c_k (or equal to it) when $t_1 \leq c_k$, and let $t_{k,0} = 0$

if $t_1 > c_k$. Let $t_{k,j}$ denote the jth failure time after $c_k, j = 1, 2, \ldots, J_k$, when $c_k < t_M$. Note that the possible values of J_k range from 1 to M, depending on the position of c_k with respect to the order of the t_m's ($m = 1, 2, \ldots, M$): J_k equals M if $c_k < t_1$, and J_k equals 1 if $t_{M-1} \leq c_k < t_M$. From the data (Y_i, δ_i), we obtain the Kaplan-Meier (KM) estimates $\hat{S}(t)$ for the survival distribution for the event times, and it has support on the observed failure times $(t_1, t_2, \ldots, t_M)$.

First, we estimate the survival rates for all $K + 1$ censoring times (t^* and c_k's, $k = 1, 2, \ldots, K$). For a censoring time c_k followed by at least one failure time (i.e., $c_k < t_M$), the estimate of the survival function $\hat{S}(c_k)$ is defined by the straightforward convention of linear interpolation as follows:

$$\begin{aligned} \hat{S}(c_k) &= \hat{S}(t_{k,0}) - \frac{c_k - t_{k,0}}{t_{k,1} - t_{k,0}} \times (\hat{S}(t_{k,0}) - \hat{S}(t_{k,1})) \\ &= \frac{(t_{k,1} - c_k)\hat{S}(t_{k,0}) + (c_k - t_{k,0})\hat{S}(t_{k,1})}{(t_{k,1} - t_{k,0})}. \end{aligned} \tag{2.1}$$

In (2.1), linear interpolation is used for computational convenience and for transparent interpretation. For the planned administrative censoring time $t^* > t_M$, (2.1) is not applicable because there is not a KM estimate for $\hat{S}(t^*)$. Nevertheless, with motivation from a suggestion in Brown et al. (1974) to use an exponential model to extrapolate $\hat{S}(t_M)$ to $\hat{S}(t^*)$, we use an exponential model for the conditional survival function for the last f events (e.g., $f = 5$) as in (2.2)

$$\hat{S}((t_M - t_{M-f})|t > t_{M-f}) = \frac{\hat{S}(t_M)}{\hat{S}(t_{M-f})} = \exp\{-h \times (t_M - t_{M-f})\}, \tag{2.2}$$

to determine the corresponding hazard h from which $\hat{S}(t^*)$ is computed as shown in

(2.3).

$$\begin{aligned}\hat{S}(t^*) &= \hat{S}(t_M) \times \hat{S}((t^* - t_M)|t > t_M) \\ &= \hat{S}(t_M) \times \exp\{h \times (t^* - t_M)\}.\end{aligned} \tag{2.3}$$

For a censoring time c_k after the last failure time (i.e., $t_M < c_k < t^*$), (2.3) similarly provides $\hat{S}(c_k) = \hat{S}(t_M) \times \exp\{h \times (c_k - t_M)\}$.

Secondly, we construct the estimated conditional failure time distribution for each patient with premature discontinuation. A fixed hazard ratio θ for a patient with premature discontinuation having an event after their censoring time c_k relative to the patients still remaining on their assigned treatment is introduced as the sensitivity parameter. Thus, under the proportional hazards assumption, the estimated survival function at time t (after c_k) equals $\hat{S}(t)^\theta$. For a patient with premature discontinuation at $c_k < t_M$, the estimated conditional probability of having the event in the time interval $[t_{k,j}, t_{k,j+1}]$, for $j = 1, 2, \ldots, (J_k - 1)$, is given by

$$\hat{f}_{k,j}(\theta) = \frac{\hat{S}(t_{k,j})^\theta - \hat{S}(t_{k,j+1})^\theta}{\hat{S}(c_k)^\theta}. \tag{2.4}$$

For the interval $[c_k, t_{k,1}]$ and $[t_{k,J_k}, t^*]$, the estimated conditional probabilities are

$$\hat{f}_{k,0}(\theta) = \frac{\hat{S}(c_k)^\theta - \hat{S}(t_{k,1})^\theta}{\hat{S}(c_k)^\theta} \text{ and } \hat{f}_{k,J_k}(\theta) = \frac{\hat{S}(t_{k,J_k})^\theta - \hat{S}(t^*)^\theta}{\hat{S}(c_k)^\theta}, \tag{2.5}$$

respectively. Correspondingly, for a patient with premature discontinuation at c_k with $t_M \leq c_k < t^*$, the estimated conditional probability of having the event in the time interval $[c_k, t^*]$ is given by

$$\hat{f}_{k,0}(\theta) = \frac{\hat{S}(c_k)^\theta - \hat{S}(t^*)^\theta}{\hat{S}(c_k)^\theta} \tag{2.6}$$

Thus, the estimate for the conditional cumulative incidence function for a patient with premature discontinuation at c_k to have the event by the time t in $[t_{k,j} < t < t_{k,j+1}]$, for $j = 1, 2, 3, \ldots, J_k$ with $t_{k,J_k+1} = t^*$ by convention, can be obtained by cumulative summation of the $\hat{f}_{k,j}(\theta)$ for the respective time intervals as shown in (2.7).

$$\hat{F}_{k,j}(\theta) = \sum_{j'=0}^{j} \hat{f}_{k,j'} = 1 - \frac{\hat{S}(t_{k,j+1})^{\theta}}{\hat{S}(c_k)^{\theta}} \tag{2.7}$$

Under this formulation, $\theta > 1$ (or < 1) implies a higher (or lower) hazard after c_k for patients with premature discontinuation at c_k than for patients with continued follow-up after c_k. Also, $\theta = 1$ specifies that patients with premature discontinuation and those with continued follow-up on the assigned treatment have the same tendency to experience an event in the future, and so it is MAR-like (and in harmony with non-informative independent censoring). Through the Cox proportional hazards model, the primary analysis can produce an estimate $\hat{\phi}$ of the hazard ratio for the effect of test treatment versus placebo under the MAR-like assumption of non-informative independent censoring for patients with premature discontinuation. However, even if this assumption is realistic, $\hat{\phi}$ pertains to what would be expected if the patients with premature discontinuation had hypothetically continued with their assigned treatments after discontinuation. Although such a perspective may be realistic for the placebo patients, it would usually be optimistic for the test treatment patients since those patients are no longer receiving test treatment after premature discontinuation. Thus, sensitivity analyses to address the implications of this issue are of interest.

One way to proceed with sensitivity analyses is to use multiple imputation with respect to the estimated conditional cumulative incidence functions in (2.7) to impute times to event for the patients with premature discontinuation in each treatment group. For the placebo group, one would typically use $\theta_P = 1$ under the realistic assumption

that its patients with premature discontinuation would have comparable experience after discontinuation to their counterparts without premature discontinuation, although other specifications of θ_P are feasibly optional. The test treatment group would usually have $\theta_T > \theta_P$ specified, and with $\theta_P = 1, \theta = (\theta_T/\theta_P) = \theta_T$ becomes a single parameter for calibrating sensitivity analyses. The choice of θ can either be arbitrary such as $1.05, 1.10, 1.15$, etc. or it can be values in a range (L, U), where $(1/U, 1/L)$ is a range of hazard ratios from previous related studies or clinical judgment for the comparison of effective medicines with placebo. For example, if previous related studies supported $(1/U, 1/L) = (0.60, 0.75)$, then one could consider θ in the range $(1.333, 1.667)$ for the extent to which a test treatment patient with premature discontinuation at c_k has a higher hazard after c_k than their counterparts with continuation of test treatment after c_k (in view of their treatment after discontinuation being more like placebo than an effective treatment).

With the conditional failure time distributions defined in (2.7), the multiple imputation scheme is as follows:

1. Generate a random number p from the uniform distribution between 0 and 1, and for computational convenience, use linear interpolation to impute failure times (although exponential model interpolations are alternatively feasible).

2. Suppose a patient has premature discontinuation at $c_k < t_M$:

   - If $0 \leq p \leq \hat{f}_{k,0}(\theta) = \hat{F}_{k,0}(\theta)$, then impute failure time $t_k^{(l)}$ between c_k and $t_{k,1}$ as $c_k + (t_{k,1} - c_k)\frac{p}{\hat{f}_{k,0}(\theta)}$, where l indicates the l-th imputation set.
   - If $\hat{F}_{k,j}(\theta) \leq p \leq \hat{F}_{k,j+1}(\theta)$ for $j = 0, 1, 2, 3, \ldots, (J_k - 1)$, then impute failure time $t_k^{(l)}$ between $t_{k,j+1}$ and $t_{k,j+2}$ as $\left(t_{k,j+1} + (t_{k,j+2} - t_{k,j+1}) \times \frac{p - \hat{F}_{k,j}(\theta)}{\hat{F}_{k,j+1}(\theta) - \hat{F}_{k,j}(\theta)}\right)$ where $t_{k,J_k+1} = t^*$ by convention.
   - If $p > \hat{F}_{k,J_k}(\theta)$, then manage the patient as having no event by the end of

follow-up time t^*.

3. Suppose a patient has premature discontinuation between t_M and t^*, so that $(t_M < c_k < t^*)$: If $p \leq \hat{f}_{k,0}(\theta)$, then impute failure time $t_k^{(l)}$ between c_k and t^* as $c_k + (t^* - c_k)\frac{p}{\hat{f}_{k,0}(\theta)}$; otherwise, manage the patient as having no event by the end of follow-up time t^*.

4. The imputation procedure is repeated to form L imputed data sets.

The tied c_k's can be processed separately. Thus, each complete data set has no patients with premature discontinuation, and so one can apply the conventional survival analysis methods for the primary analysis with only administrative censoring of follow-up at time t^*.

In reality, most clinical trials recruit patients over a period of time and have a common closing date. Therefore, patients always have different planned follow-up times and correspondingly different administrative censoring times for when they could complete the study without the event. Such staggered patient entry can be addressed by letting t_k^* denote the planned follow-up time for the kth patient with premature discontinuation at $c_k (c_k < t_k^*)$. For t_k^* between two consecutive failure times (t_m, t_{m+1}), the applicable survival function can be estimated at c_k and t^* in an analogous fashion to (2.1). For t_k^* after the last failure time $(t_k^* > t_M)$, the applicable survival function at c_k and t^* can be estimated by the method described for (2.3). In this way, the conditional failure time distribution can be constructed from $(2.4) - (2.7)$ according to a prematurely discontinued patient's planned follow-up time t_k^*. The multiple imputation can then be performed in the same fashion as discussed previously.

The proposed method does not seek inferences for the hypothetically true parameters for treatment effects, but rather addresses the sensitivity issues associated with the

unobserved outcomes of discontinued patients. For this purpose, the multiple imputation process regards the observed information being fixed, that is K and M, as well as the corresponding times to event and times to premature discontinuation. In the context of Bayesian multiple imputation, Rubin (1987) refers to this type of imputation as 'improper', because it does not account for the uncertainty associated with the sample estimates (i.e. KM estimates or Cox proportional hazards model counterparts). A way to address such uncertainty is to generate the L data sets by separate conditional failure time distributions estimated from independent nonparametric bootstrap resamples (with replacement) of the original data.

2.3.2 Parameter Estimations

The method for combining results from L imputed data sets follows well established rules (Rubin, 1987; Rubin and Schenker, 1991), and it can be applied easily by the SAS procedure MIANALYZE. Let β be a scalar parameter such as a survival rate or a cumulative hazard for a specific time point or a coefficient in the Cox proportional hazards model (i.e., the log hazard ratio) that can be estimated from the complete data. Let $\hat{\beta}^{(l)}$ denote the point estimate for β and let $\hat{V}_{\beta}^{(l)}$ denote its variance estimate from the lth data set. The overall multiple imputation (MI) estimate of β is obtained by averaging the estimates from the L complete-data analyses, $\bar{\beta} = (1/L)\sum_{l=1}^{L}\hat{\beta}^{(l)}$, and its estimated variance is the sum of the within-imputation variance $\bar{V}_{\beta} = (1/L)\sum_{l=1}^{L}\hat{V}_{\beta}^{(l)}$ and the product of the between-imputation variance $B_{\beta} = (L-1)^{-1}\sum_{l=1}^{L}(\hat{\beta}^{(l)} - \hat{\beta})^2$ and the finite sample correction shown in (2.8).

$$\hat{V}_{\bar{\beta}} = \bar{V}_{\beta} + (1 + L^{-1})B_{\beta} \tag{2.8}$$

Given sufficiently large sample size for the complete data to support an approximately standard normal $N(0,1)$ distribution for its hypothetical version of $(\hat{\beta}-\beta)\hat{V}_{\hat{\beta}}^{-1/2}$, for which missing data prevents availability, confidence intervals for β (and p-values for corresponding statistical tests) can be based on $(\bar{\beta}-\beta)\hat{V}_{\bar{\beta}}^{-1/2}$ having the t-distribution with approximate degrees of freedom (d.f.) as shown in (2.9).

$$\begin{aligned} \text{d.f.} &= (L-1)\left(1+\left(\frac{(1+L^{-1})\,B_\beta}{\bar{V}_\beta}\right)^{-1}\right)^2 \\ &= (L-1)(1+R^{-1})^2 \end{aligned} \tag{2.9}$$

Here, R expresses the relative increase in variance due to missing information. The fraction of missing information about β is estimated as

$$\gamma = \frac{R+2/(\text{df}+3)}{(1+R)}. \tag{2.10}$$

For non-parametric hypothesis testing with the logrank (or Wilcoxon) statistic, $\hat{\beta}^{(l)}$ is the difference between test treatment and placebo for means of logrank or Wilcoxon scores, and $\hat{V}_\beta^{(l)}$ is its estimated variance under the null hypothesis of no difference between test treatment and placebo. It then follows that $\bar{\beta}(\hat{V}_{\bar{\beta}}^{-1/2})$ approximately has the t-distribution with d.f. as in (2.9). Alternatively, $\hat{Z}^{(l)} = \hat{\beta}^{(l)}/\hat{V}_\beta^{(l)}$ can serve as $\beta^{(l)}$ with corresponding $\hat{V}_Z^{(l)} = 1$, in which case the statistical test would be based on $\bar{Z}(\hat{V}_{\bar{Z}}^{-1/2})$ with $\hat{V}_{\bar{Z}} = (1+(1+L^{-1})B_Z)$ with $B_Z = \sum_{l=1}^{L}(\hat{Z}^{(l)}-\bar{Z})^2/(L-1)$ (Taylor et al., 2002).

The term $L^{-1}B_\beta$ in (2.8) and the use of the t-distribution rather than a normal distribution widen the resulting interval estimates to account for replication variability incurred by using $L < \infty$ (Schafer, 1999). Schafer (1999) suggests that unless the fraction of missing information γ is unduly large, the widening is not substantial, and

MI inferences can be quite efficient even when L is small (usually less than 10). Nevertheless, in practice, the appropriate number of imputations should be investigated more closely, especially when the fraction of missing information is large (Horton and Lipsitz, 2001).

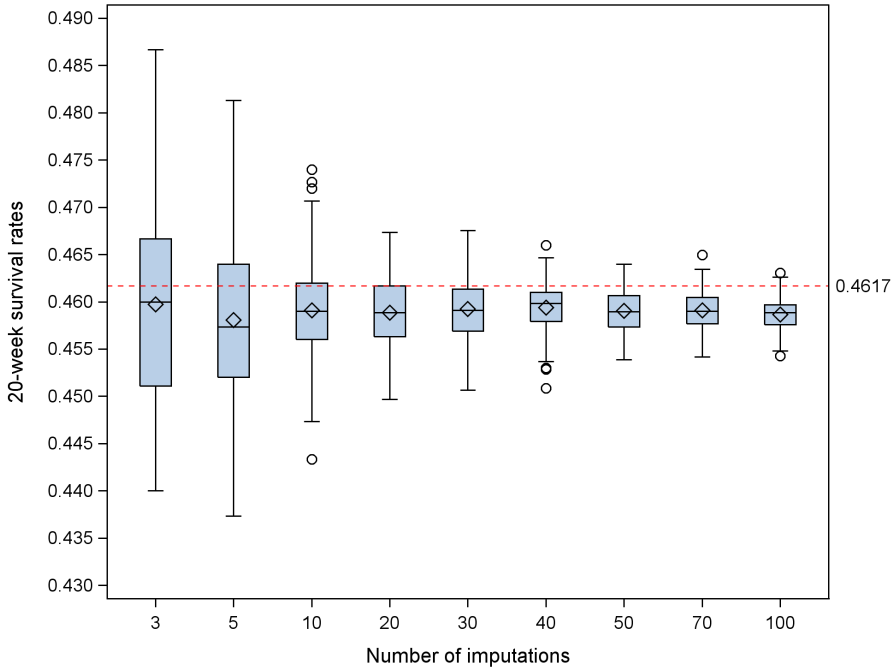
2.4 Results

2.4.1 Performance of KMMI method under $\theta = 1$

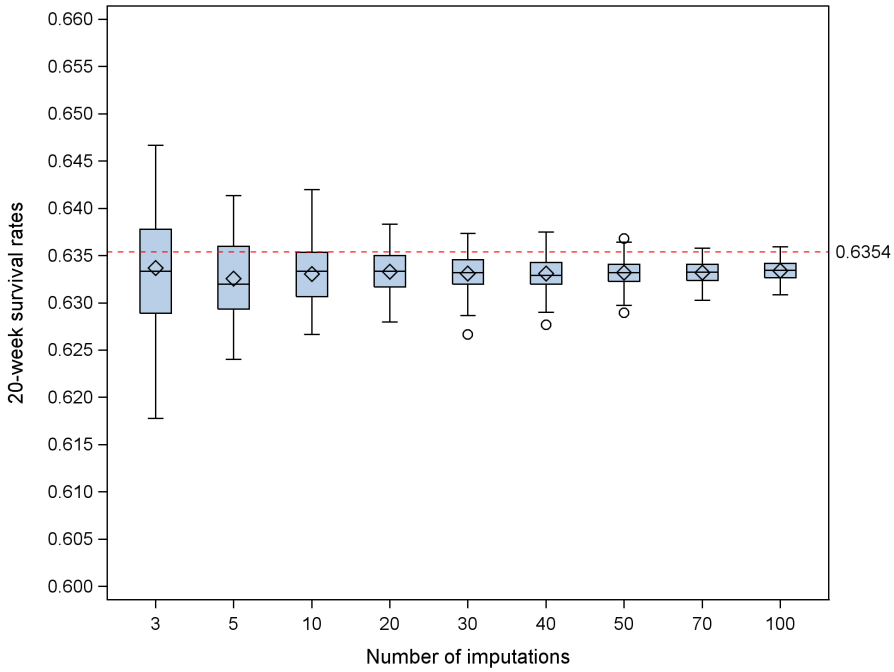
In this section, we consider the performance of the KMMI method with $\theta = 1$ for the clinical trial in Section 2.2. With this specification, the imputed data are produced from the same conditional failure time distributions as estimated by the KM method with censoring of the follow-up times of patients with premature discontinuation, and it thereby has the same MAR-like assumption of non-informative independent censoring. To apply this method, we proceed in accordance with Horton and Lipsitz (2001) to determine the appropriate number of imputations (L) by evaluating the stability of an estimator and its standard errors (SE) with respect to the different L's. Multiple imputations are performed separately for each of the two treatment groups with $\theta = 1$, and 100 replicates of imputations are produced for each of the following numbers of imputations ($L = 3$, 5, 10, 20, 30, 40, 50, 70, and 100); and so there are seven different sets of imputations. The variability of the estimates for the survival function at the 20th week are summarized in boxplots in Figure 2.2 relative to the conventional KM estimates (with censoring of follow-up times for patients with premature discontinuation). The relative variance increase due to missing data ($R = (1 + L^{-1})B_\beta/\bar{V}_\beta$) of the corresponding estimates are summarized in Figure 2.3. Compared to the conventional KM estimates, the mean values of estimates from the KMMI method are somewhat

smaller for all seven sets of imputations. Also, the KMMI estimates and the corresponding R in Figure 2.2 and Figure 2.3 (for the 20-week survival rate) are not stable for small numbers of imputations (i.e., $L \leq 10$). The variability of the MI estimates becomes smaller as the number of imputations increases, and stabilizes near $L = 50$ or higher for both treatment groups. Thus, 50 imputations is a reasonable choice for the amount of missing information which this example has. Although a comprehensive simulation study could shed more light on the choice of L for different extents of missing data, such research is beyond the scope of this paper. Nevertheless, for any real study, the specification of at least a moderately high value of $L \geq 50$ should be considered, especially given the simplicity of the computations even for large L.

We apply multiple imputation (MI) with $L = 50$ henceforth. The conventional KM curves for both treatment groups are shown in Figure 2.4a with their counterparts from averaging the KM estimates for 50 data sets imputed by the KMMI method. The corresponding cumulative hazard curves (via the Aalen-Nelson estimator) are shown in Figure 2.4b. The relationships shown for the KMMI method are almost identical to their conventional counterparts. In row (2A) of Table 2.3, results from the KMMI method are shown for the hazard ratio for the effect size of the test treatment versus placebo from the unadjusted Cox proportional hazards model (which only includes treatments) as well as for the p-values for the logrank test and the Wilcoxon test. Interestingly, the estimated hazard ratio from the KMMI method is closer to unity (and so is a smaller effect size) and has a somewhat larger p-value than its conventional counterpart with the use of censoring (HR=0.724 with p=0.0436 for KMMI versus HR=0.675 with p=0.0140 for conventional). This disagreement between the inference for the effect of the test treatment from the KMMI method with $\theta = 1$ and conventional counterparts with censoring could be a consequence of non-proportional hazards during the follow-up period. As can been seen from the survival curves in Figure 2.4a and the cumulative

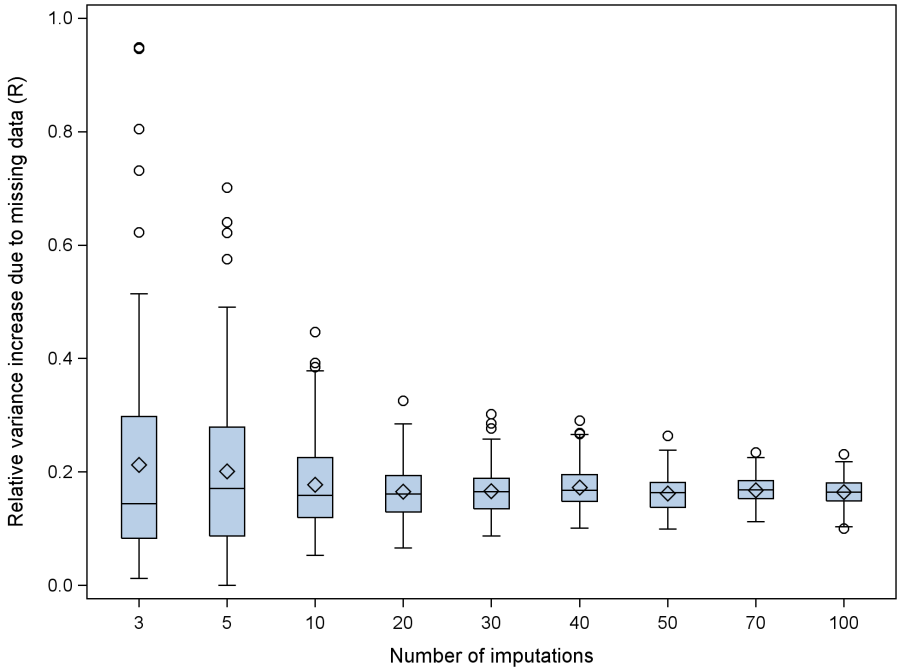


(a) Placebo group

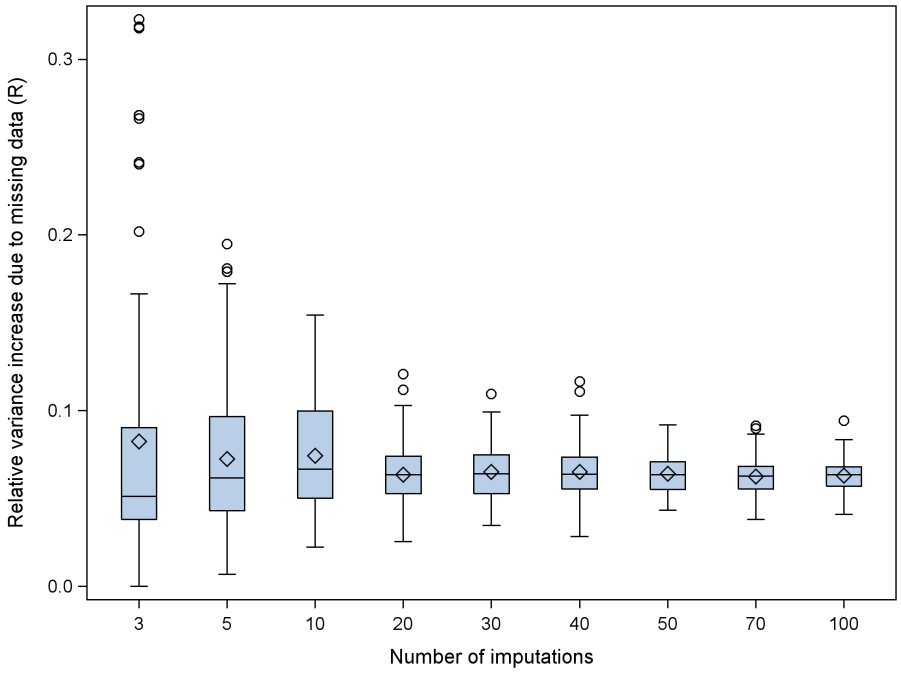


(b) Test treatment group

Figure 2.2: Distributions of 20 weeks survival rates for 100 replications of different numbers of imputations. The conventional KM estimates are indicated with the horizontal line.



(a) Placebo group



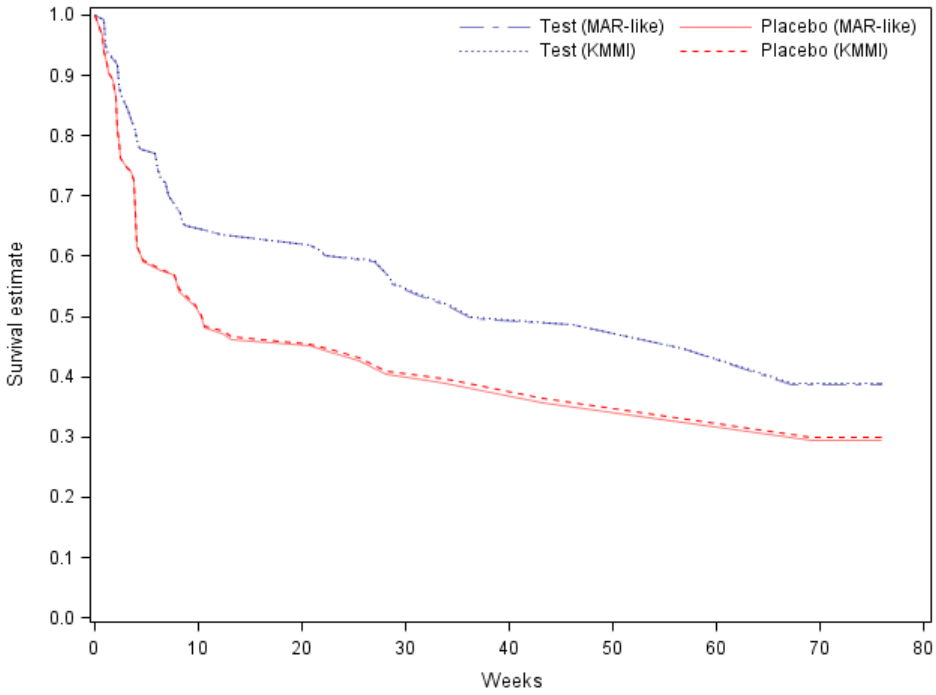
(b) Test treatment group

Figure 2.3: Distributions of relative variance increase due to missing data (R) of 20 weeks survival rates for 100 replications of different numbers of imputations.

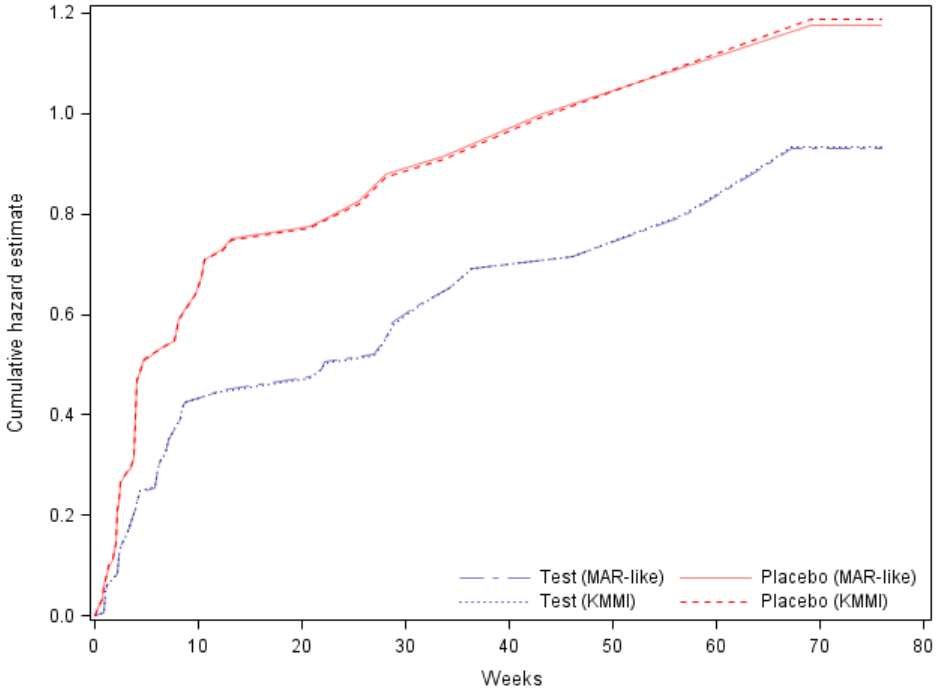
hazard curves in Figure 2.4b, the difference between the two treatment groups is most clearly evident during the early stage of the follow-up and less apparent later. This issue is explored further by partitioning the follow-up period into four distinct intervals with approximately equal numbers of events, and then producing conventional interval-specific hazard ratio estimates for each of them from an unadjusted Cox proportional hazards model. The results of such analysis in Table 2.3(1) suggest much stronger effect sizes for test treatment during $0-6$ weeks than $6-76$ weeks, and so they are contrary to the hazard ratio being constant during the entire follow-up period.

When treatment is the only explanatory variable in the Cox proportional hazards regression model, its estimated effect size is approximately an average of log HR over the entire follow-up period. When there are many patients with premature discontinuation, the estimation of the average log HR through conventional methods with censoring may tend to be mainly influenced by events during the earlier part of the follow-up period (where the effect sizes for test treatment are stronger for this example). The KMMI method eliminates censoring during the follow-up period by imputing potential times to event for every patient with premature discontinuation, and it thereby puts more weight on what happens during the latter part of the follow-up period (where the effect sizes for test treatment are smaller for this example), and so it produces a smaller effect size for test treatment (in the sense of an estimated hazard ratio that is closer to unity). Thus, this example suggests that the sensitivity analysis with $\theta = 1$ for the KMMI method can be useful for evaluating the implications of non-proportional hazards during the follow-up period.

An alternative structure for multiple imputation is provided by the Breslow estimators of the survival distributions for the placebo and test treatment groups from the Cox proportional hazards model with treatment as the only explanatory variable, and it can have implementation through its counterparts for (2.1) - (2.7). As shown in Table 2.3



(a) Survivor Curves



(b) Cumulative hazard Curves

Figure 2.4: Comparison of the results from the conventional (MAR-like) and the KMMI method

rows (1A) and (3A), the proportional hazards multiple imputation (PHMI) method under $\theta = 1$ provides very similar results as the conventional methods with censoring, mainly because both operate under the MAR-like assumption of non-informative independent censoring and both have the proportional hazards assumption.

We further consider an imputation with non-parametric bootstrap resampling so as to add extra between-imputation variability and thereby to be in better harmony with a 'proper' imputation. Consequently L may need to be much larger than 50, in order to provide appropriate precision for estimation. Both the KMMI and the PHMI methods proceed with an additional bootstrap step for $L = 50$, $L = 100$, and $L = 500$. The results of MI with the bootstrap for $L = 500$ are relatively consistent with the methods without the bootstrap for $L = 50$ (see Table A.1 for details). The bootstrap KMMI method uses separate samples with replacement for each treatment group, and its results for $L = 500$ (Table 2.3.4A) are slightly weaker compared with its counterparts without the bootstrap for $L = 50$. The PHMI method with the bootstrap uses samples with replacement from the combined treatment groups. As shown in Table 2.3.5A, when performed for $L = 500$, the PHMI with the bootstrap produces comparable results to the PHMI without the bootstrap for $L = 50$. The imputation methods with and without the bootstrap arise from different paradigms. The imputation methods with the bootstrap are based on Bayes theory and relate the posterior distribution given the observed data to the complete posterior distribution given no missing data in a random sample of a target population, and therefore they add more complexity to the imputation process. Alternatively, the methods without the bootstrap address the uncertainty of missing data in the context of the observed information being known and fixed. Depending on the purpose of the sensitivity analysis, either process can be applied. For this paper, we emphasize the sensitivity analysis using the MI methods without the bootstrap for $L = 50$.

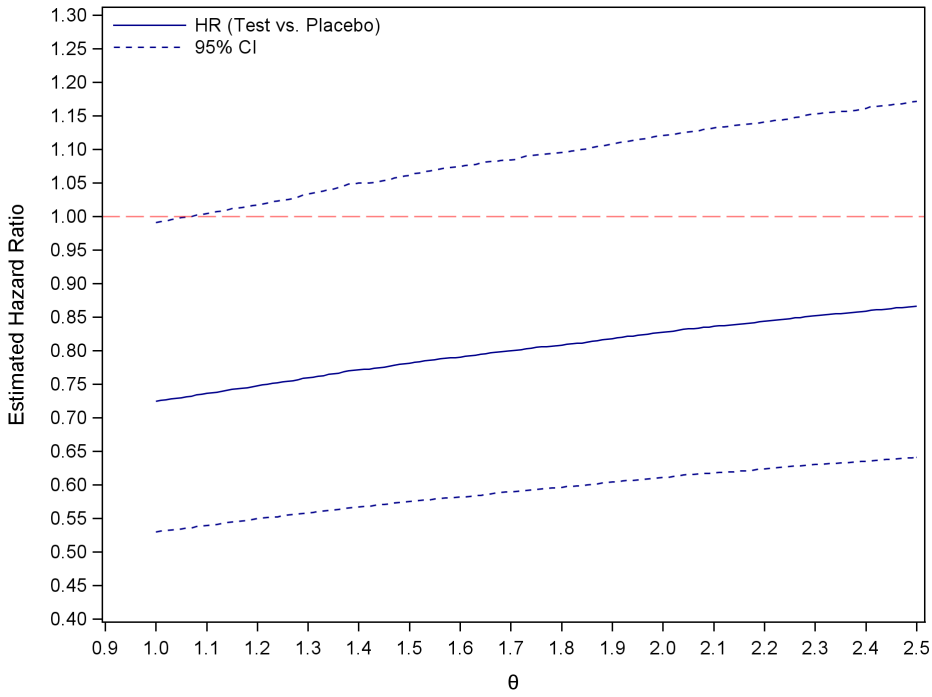
2.4.2 Sensitivity analysis

The sensitivity analyses proceed with varying θ (for the test treatment) in a plausible range from 1 to 2.5 (with $\theta = 1$ for placebo) to determine how the assessment of the treatment effect changes for the different extents of imputed events for patients with premature discontinuation at specific times c_k versus patients with continued follow-up beyond those times. In this regard, $\theta = 2.5 = (1/0.4)$, and 0.4 might represent a reasonably large effect size for a clearly effective treatment versus placebo in the published clinical literature for maintenance treatments of bipolar disorder. On this basis, it is a reasonable choice for the upper bound of the sensitivity parameter θ in terms of how much more rapidly the patients that had premature discontinuation would have the event compared to those that did not; in this regard, it is useful to note that $\theta = \infty$ corresponds to the worst comparison analysis. The value of θ is varied by 0.01 increments from 1 to 2.5, leading to 150 treatment effect assessments. Contour plots of the hazard ratio estimates and the p-values for treatment comparisons are then constructed as a function of the sensitivity parameter θ. We implement both the KMMI method and the PHMI method in these sensitivity analyses. The multiple imputation results from the Cox proportional hazards models as well as the logrank and Wilcoxon tests are combined using the method described in Section 2.3.2.

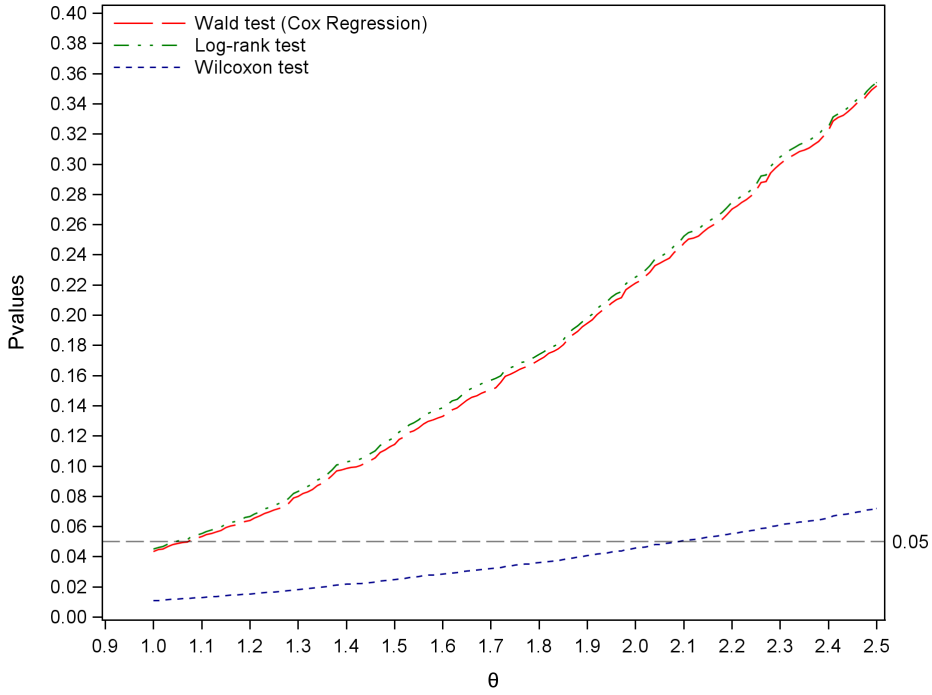
The sensitivity analysis results using the KMMI method are summarized in Figure 2.5. The values of θ plotted against the estimated hazard ratios with 95% confidence intervals are shown in Figure 2.5a, and the p-values obtained from the Wald test from the Cox proportional hazards model, the logrank test, and the Wilcoxon test are shown in Figure 2.5b. The magnitude of the estimated treatment effect moves closer to the null (i.e., $\theta = 1$) as the value of θ increases. The HR estimates for test treatment versus placebo have a range from 0.724 (for $\theta = 1$) to 0.867 (for $\theta = 2.5$). The corresponding p-values from the Wald test vary substantially over the range of θ, indicating

that the assumptions for patients with premature discontinuation can substantially influence study conclusions. As expected, the p-values from the Wald test agree with those from the logrank test, and they are larger than those from the Wilcoxon test. In order to have $p \leq 0.05$ with the Wald test (or the logrank test), $\theta \leq 1.08$ (or 1.05) is needed, with this specification being only slightly more stringent than the MAR-like assumption of non-informative independent censoring (or $\theta = 1$). For the Wilcoxon test, $p \leq 0.05$ applies with $\theta \leq 2.08$, and so it has better robustness to assumptions about patients with premature discontinuation of treatment for this example than the Wald test or the logrank test. Since the Wilcoxon test receives relatively more weight than the logrank test for early failures and relatively less weight for later failures, it is more able to detect the early hazard differences for this example than the logrank test. As shown in Figure 2.4b and Table 2.3(1), the estimated treatment effect is much stronger (i.e., hazard ratios are further away from 1) in the earlier part of the follow-up.

The results of sensitivity analyses with the PHMI method are shown in Figure 2.6. Because the PHMI method invokes the possibly unrealistic proportional hazards assumption, it suggests better robustness for the conclusions from the Cox proportional hazards model and the logrank test than the KMMI method. For $p \leq 0.05$ with the Wald test (or the logrank test), $\theta \leq 1.59$ (or 1.58) is needed; also, for the Wilcoxon test, $p \leq 0.05$ applies for all $\theta \leq 2.5$. In general, the PHMI method may not always suggest stronger conclusions than the KMMI method. When the differences between the test treatment and the placebo are more substantial during the latter part of the follow-up period than the early part, the KMMI method with $\theta = 1$ could lead to stronger conclusions (i.e., estimated hazard ratios further from 1 and smaller p-values), while the PHMI method under $\theta = 1$ would tend to produce similar results as conventional analyses with censoring of follow-up time for patients with premature discontinuation.

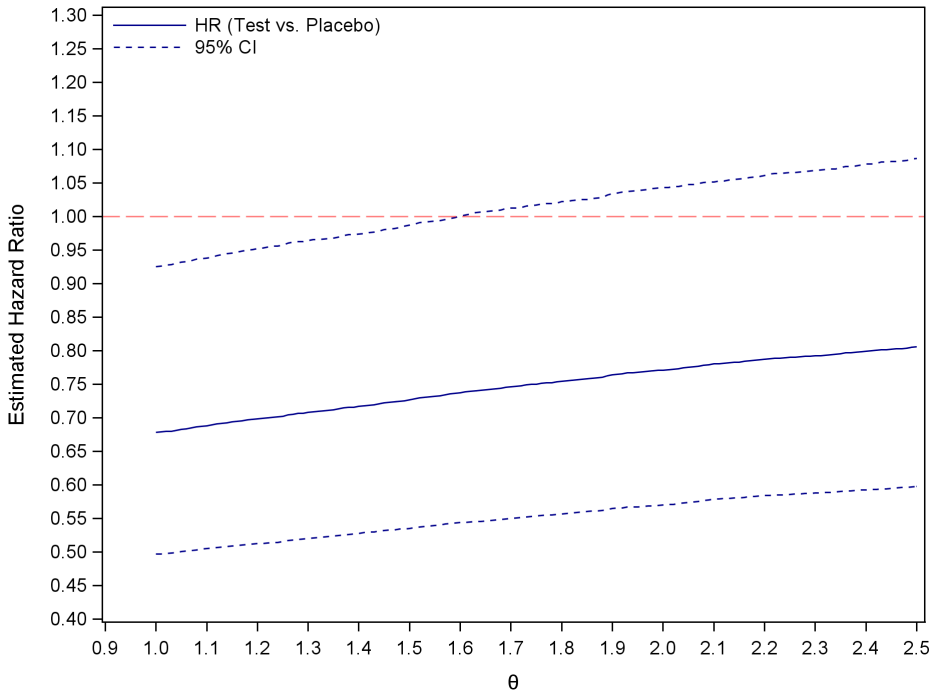


(a) Hazard ratio (HR) with pointwise 95% CI from Cox regression model using KMMI method

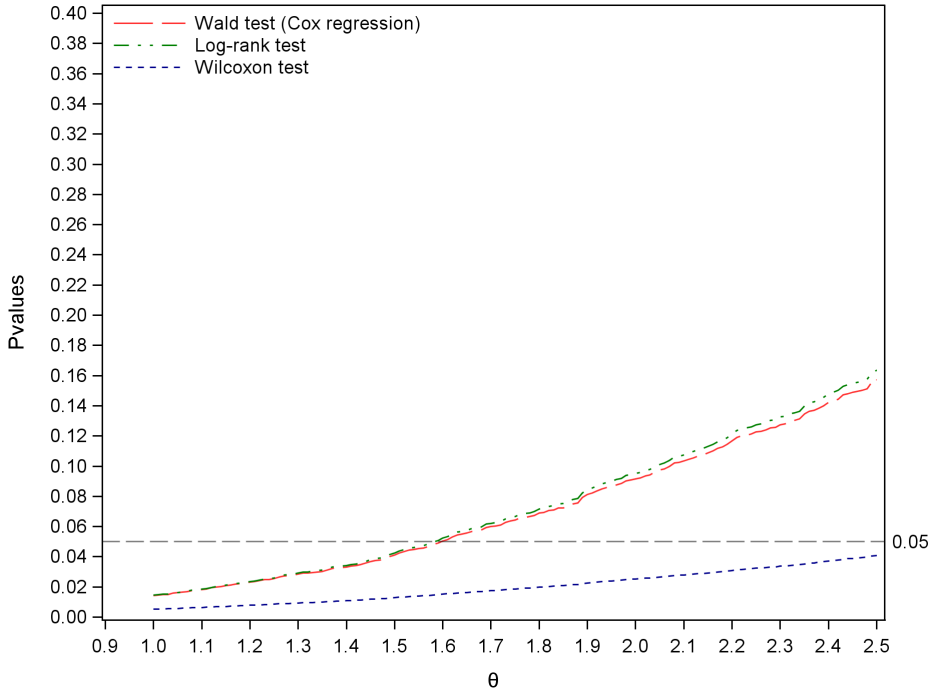


(b) P values from 3 different hypothesis tests using KMMI method

Figure 2.5: Sensitivity analysis results using KMMI method



(a) Hazard ratio (HR) with pointwise 95% CI from Cox regression model using PHMI method



(b) P values from 3 different hypothesis tests using PHMI method

Figure 2.6: Sensitivity analysis results using PHMI method

Therefore, the sensitivity analysis based on the KMMI method may provide more accurate assessment than the PHMI method.

Sensitivity analyses with both the KMMI and the PHMI methods can be useful for reviewers to understand the robustness of conclusions for treatment effects to the assumptions of non-informative independent censoring and proportional hazards. The degree to which conclusions are stable across a reasonable range of θ provides an indication of the confidence that can be placed on them. Opinions on possible values of θ can be based on knowledge from other studies for similar interventions. An investigation of the differences between baseline characteristics of completers and patients with premature discontinuation can be useful as well as the reasons for discontinuation. If such information suggests that only unrealistic values of θ would alter study conclusions, then the results of a primary analysis with conventional methods could be considered robust from a clinical perspective. When the inference about treatment effects could be overturned for plausible values of θ, then it should be viewed with caution.

2.5 Discussion

Analysis of incomplete data is a challenge for most clinical trials. Often, MAR-like assumptions about the missing data mechanism can be reasonable for primary analyses. However, the possibility of MNAR is difficult to rule out, particularly when patients with test treatment lose its benefit after discontinuation, and so sensitivity analyses for alternative ways to address missing data become of interest.

In time-to-event analyses, patients with premature discontinuation have their follow-up time censored at the time of discontinuation, and the usual assumption is non-informative independent censoring. As right-censoring is a special case of coarsened data, the assumption of non-informative independent censoring can be generalized to 'coarsened at random', which extends the concept of MAR to coarsened data (Heitjan,

1994). The MNAR issue for time-to-event data is to account properly for censoring that may be informative. Most sensitivity analyses in the literature assess the effect of various assumptions concerning the dependence between failure and censoring times (Scharfstein and Robins, 2002; Siannis et al., 2005; Ruan and Gray, 2008). However, clinical reviewers can have difficulty in understanding the interpretation of the sensitivity parameters in those analyses, and this can make the specification of reasonable ranges for the sensitivity parameter challenging.

This paper discusses sensitivity analyses for time-to-event data, and its suggested methods can have several appealing features in regulatory clinical trial settings. First, they enable direct exploration of the effect of departures from the non-informative independent censoring assumption for conventional methods (such as Cox proportional hazards models, logrank tests and Wilcoxon tests) through a sensitivity parameter that connects the unobserved outcomes and the observed outcomes, i.e., a hazard ratio for a discontinued patient having an event after discontinuation relative to the patients remaining on their assigned treatment. The multiple imputation strategy is straightforward because the predictive distributions are specified directly, and they do not depend on the models for assumed missingness mechanisms. The interpretation of the sensitivity parameter is transparent in the sense that the parameter is based on a standard criterion for analyzing time-to-event data, and consequently may be more understandable to reviewers. Secondly, the sensitivity analysis accounts for all randomized patients. The specifications for post-discontinuation experience are intended to address the question for what the long-term benefit of initial assignment would be if patients with premature discontinuation were followed to the end of the study without other treatment. In addition, the influence of departures from the non-informative independent censoring assumption with respect to patients with premature discontinuation can be assessed either simultaneously with the proportional hazards assumption

by the KMMI method or separately in its own right by the PHMI method. Thirdly, the sensitivity analysis is based on multiple imputation of missing outcomes, and therefore it provides a simple way of generating statistical inference without the need of special software and programming. All of the analyses presented in this paper can be produced by standard SAS PROC procedures and SAS macros. Finally, the proposed sensitivity analysis anchors on a primary MAR-like assumption, and then can have calibration toward the worst comparison analysis through how it penalizes premature discontinuation for the teat treatment. The method can be specified a priori and does not require any post hoc (i.e., data driven) revisions. Therefore, this type of sensitivity analysis to address the missing information from censored follow-up times could be attractive in the regulatory environment.

The sensitivity analysis illustrated here was performed for a continuous time-to-event endpoint. However, the methodology and underlying principles can be extended to categorical (or interval censored) time-to-event data. Furthermore, the proposed PHMI strategy can be modified to incorporate the information of patients' baseline risk factors. One can estimate the failure time distributions separately for subpopulations defined by baseline covariates and treatments through the multivariable Cox proportional hazards model that includes treatments and the set of covariates as explanatory variables. The conditional failure time distributions can then be used for risk adjusted multiple imputations. Currently, the discussed MI strategies invoke separate imputations for each of the two groups with its corresponding survival distribution estimates. An alternative approach is to impute times to event for both treatment groups using the information in the placebo group. The details of this method and its corresponding results are discussed in the Appendix 3. However, it may not address robustness as stringently as the methods that are the main focus of this paper, when the placebo group has a higher proportion of discontinuations than the test treatment

group.

Typically, the design of a confirmatory trial should account for the loss of power from patients with premature discontinuation (NRC, 2010). An often used approach is simply to inflate the initially planned sample size by the reciprocal of one minus the anticipated premature discontinuation rate, but it may only be reasonable if missing information is MCAR. Power calculations should be based on more plausible MAR-like assumptions, and perhaps accommodating the situations of MNAR and the potentially reduced effect size estimation in sensitivity analyses. However, those concerns usually cannot be addressed analytically in sample size calculations. The multiple imputation strategy presented in the current sensitivity analysis method can be adapted for simulation-based power calculations to assess the effect of missing data on sample size.

Chapter 3

Covariate-Adjusted Sensitivity Analysis for Time-to-event Data

3.1 Introduction

Missing data exist in practically all clinical trials. A major source of missing data is from patients discontinuing their assigned treatment and then withdrawing from the study. The extent to which missing data impact statistical inferences depends on the process (i.e., the mechanism) leading to the missingness. Little and Rubin (2002) outlined the following missing data framework: (1) data are missing completely at random (MCAR) if the missingness does not depend on either the observed or unobserved data; (2) data are considered missing at random (MAR) when the missingness only depends on the observed data; (3) data are missing not at random (MNAR) if the missingness also depends on the unobserved data. If the parameters of the measurement process and the missing data process are distinct under the MAR mechanism, the missing data mechanism is said to be ignorable for likelihood-based inference since unbiased (or consistent) parameter estimates can be obtained from the observed data (Mallinckrodt et al., 2008).

In many clinical trials, MAR can be reasonable and hence it is often chosen as the main assumption for the primary analysis (Mallinckrodt et al., 2008; Zhang, 2009).

However, missing mechanism can be more more complex than the ideal MAR assumption in practice. the possibility of MNAR can never be ruled out. Therefore, a prudent analyst should always conduct sensitivity analyses to assess the robustness of the treatment effect inferences to various alternative missing data assumptions (NRC, 2010). Zhao et al. (2012) recently introduced a method for sensitivity analysis for missing outcomes in time-to-event data, for which the primary analytical strategy has the MAR-like assumption of non-informative independent censoring. Based on the Kaplan-Meier (KM) estimator or its Cox proportional hazards (PH) model (Cox, 1972) counterparts, Zhao et al. (2012) employed multiple imputation of potential times to event for withdrawal patients to produce the inference if they were followed off treatment until the end of study. The departure from the primary MAR-like assumption was addressed by a sensitivity parameter that captures the difference in the post-discontinuation tendency of developing an event. When treatment effects are evaluated with the standard methods without covariate adjustment, application of such a sensitivity analysis is straightforward (Zhao et al., 2012).

Although the unadjusted analysis provides valid treatment comparisons in randomized studies, covariate-adjusted analysis is often implemented to increase statistical power or to offset the influence of random imbalance between treatment groups for the covariates with possibly strong relationships with the primary outcome (Tangen and Koch, 2000). One concern regarding the appropriateness of covariate adjustment with the Cox regression model is whether the proportional hazards assumption is hold for each variable in the model. In addition, incorrect model specifications may produce biased estimates for the regression coefficients (Tangen and Koch, 2000). One way to avoid those issues is to account for the covariates with the randomization based analysis of covariance (ANCOVA). Through the weighted least squares methodology (Grizzle

et al., 1969), non-parametric approaches have been proposed to provide covariate adjustment for inferences on incidence density ratios (Tangen and Koch, 2000) or hazard ratios (Moodie et al., 2011) for multiple non-overlapping time intervals. Recently, Saville and Koch (2012) discussed a randomization based method to estimate the covariate-adjusted population average hazard ratio with Cox regression models. Using the covariance matrix estimates of the unadjusted log hazard ratio from the Cox regression model and the group differences in means of baseline covariates, and they implemented the weighted least squares methodology to produce a covariate-adjusted log hazard ratio by forcing the differences in means for covariables to zero. The nice feature of this approach it incorporates the usual Cox regression model estimates into the non-parametric ANCOVA (NPANCOVA) paradigm, hence it avoids the proportional hazards assumption for the adjusted covariates and avoids possibly data driven model refinements. Consequently, it could be a more appealing strategy for the primary analysis in regulatory environments.

As an alternative to the conventional Cox multivariate regression models, inverse probability weights (IPW) are commonly employed to balance covariates across treatment groups in estimating risk-adjusted effects in comparative effectiveness studies with observational data (Cole and Hernán, 2004; Curtis et al., 2007). To implement such an approach, analysts first estimate, using a logistic regression, the predicted probability that an individual receives their own treatment conditional on the set of their observed covariates. This predicted probability of exposure to one of the treatments is called a propensity score (PS) (Rosenbaum and Rubin, 1983). Then each subject receives a weight by the inverse of this probability to create comparable pseudo populations that have similar distributions for those covariates (Robins et al., 2000). The average covariate-adjusted treatment effect can be easily produced by comparing the re-weighted pseudo populations for treatment groups through standard methods, such

as the conventional Cox regression model with treatments as the only factor. Therefore, similar in spirit to the method of Saville and Koch (2012), the IPW approach with PS also relax the strong assumption of proportional hazards for the multivariable Cox regression model.

In this article, we discuss how to implement covariate adjustment in the sensitivity analysis proposed in Zhao et al. (2012) for time-to-event data. When data are from randomized clinical trials, one can regard the patients of each treatment group as a random sample from the study population. As a result, one straightforward way to perform MI is through KM estimates, i.e., the KM-MI method in Zhao et al. (2012), mainly because the KM curve is a valid estimator of the survivor profile for randomized treatments. For each imputed data set, the covariate-adjusted log hazard ratio can be obtained with the method of Saville and Koch (2012). The final treatment estimate can be obtained from these estimates using Rubin (1987)'s formulas. Alternatively, one can impute data via the Breslow estimator from the Cox proportional hazards model that includes treatment and the set of the covariates, and then proceed with analysis by the same Cox regression model. Although this procedure is logically consistent and applicable for either randomized or observational studies, the issues with the Cox model assumptions could influence the acceptability of this approach in the regulatory setting, especially when those assumptions are not supported by the observed data. Under less stringent assumptions, we propose a new strategy for sensitivity analysis that employs IPW to account for covariates in the imputation of failure times for patients with premature discontinuation of treatment. As opposed to requiring randomized data as in the approach of Saville and Koch (2012), the covariate-adjusted sensitivity analysis strategy invoking IPW is applicable for both randomized and observational data. In this paper, we will discuss these methods in the context of an illustrative clinical trial in psychiatry.

3.2 Covariate-Adjusted Hazard Ratio Estimation

3.2.1 Nonparametric ANCOVA

Saville and Koch (2012) proposed a non-parametric, randomization-based ANCOVA (NPANCOVA) method to obtain covariate-adjusted log hazard ratios. Let $h = 1, 2$ index the test and the control group with n_h patients in group h; and let $\boldsymbol{r}_h$ be the corresponding dfbeta residual ($n_h \times 1$) vector obtained from the unadjusted Cox proportional hazards model with treatments as the only factor. The ith element of $\boldsymbol{r}_h$ is the change in the log hazard ratio estimate ($\hat{\beta}$) for comparing test treatment versus control when the ith observation in group h is omitted, and it can be approximated by $-\boldsymbol{\mathcal{I}}(\hat{\beta})^{-1}\mathcal{S}_{h_i}$, where $\boldsymbol{\mathcal{I}}(\hat{\beta})$ is the observed information matrix, and $\mathcal{S}_{h_i}$ is the ith score vector residual. Therefore, for $\boldsymbol{r} = (\boldsymbol{r}_1', \boldsymbol{r}_2')'$, $\boldsymbol{r}'\boldsymbol{r} = (\boldsymbol{r}_1'\boldsymbol{r}_1 + \boldsymbol{r}_2'\boldsymbol{r}_2)$ approximates the robust sandwich variance for $\hat{\beta}$ (Wei et al., 1989; Lin and Wei, 1989). Let $\boldsymbol{X}_h = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iq})$ be the $(n_h \times q)$ matrix of q baseline covariates for group h; and let $\bar{\boldsymbol{x}}_h = (\bar{x}_{h1}, \ldots, \bar{x}_{hq})'$ be the vector of means for the q baseline covariates for group h with the corresponding covariance shown in (3.1),

$$\boldsymbol{V}_{\bar{\boldsymbol{x}}_h} = (\boldsymbol{X}_h - \mathbf{1}\bar{\boldsymbol{x}}_h')'(\boldsymbol{X}_h - \mathbf{1}\bar{\boldsymbol{x}}_h')/\left(n_h(n_h-1)\right) = \boldsymbol{C}_h'\boldsymbol{C}_h \tag{3.1}$$

where $\boldsymbol{C}_h = (\boldsymbol{X}_h - \mathbf{1}\bar{\boldsymbol{x}}_h')/\sqrt{n_h(n_h-1)}$ and $\mathbf{1}$ is a $(n_h \times 1)$ vector of ones. Let $\boldsymbol{d} = (\hat{\beta}, (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)')'$ be the vector of the unadjusted log hazard ratio estimate for treatments and the differences in means for the baseline covariates for the test treatment and the control groups. Then the covariance matrix of $\boldsymbol{d}$ is obtained via the sums of cross products of $\boldsymbol{r}_h$ and $\boldsymbol{C}_h$ as shown in (3.2). The mathematical derivations were discussed

by Saville and Koch (2012).

$$\boldsymbol{V_d} = \begin{bmatrix} (\boldsymbol{r}_1'\boldsymbol{r}_1 + \boldsymbol{r}_2'\boldsymbol{r}_2) & (\boldsymbol{r}_1'\boldsymbol{C}_1 - \boldsymbol{r}_2'\boldsymbol{C}_2) \\ (\boldsymbol{r}_1'\boldsymbol{C}_1 - \boldsymbol{r}_2'\boldsymbol{C}_2) & (\boldsymbol{C}_1'\boldsymbol{C}_1 + \boldsymbol{C}_2'\boldsymbol{C}_2) \end{bmatrix} \tag{3.2}$$

With the NPANCOVA approach discussed in Koch et al. (1998), the covariate-adjusted estimate for the log hazard ratio can be obtained via the weighted least squares regression (Grizzle et al., 1969) for the model $E_A(\boldsymbol{d}) = \boldsymbol{Z}\delta$, where $E_A(\boldsymbol{d})$ is the asymptotic expected value for $\boldsymbol{d}$, $\boldsymbol{Z} = [1, \boldsymbol{0}_q']'$ is the matrix to specify the adjusted analysis, and δ is the regression coefficient. With $\boldsymbol{Z}$ to force the difference in means for covariates to zero, the covariate-adjusted log hazard ratio estimate is $\hat{\delta} = (\boldsymbol{Z}'\boldsymbol{V_d}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{V_d}^{-1}\boldsymbol{d}$ and the corresponding variance estimator is $V_\delta = (\boldsymbol{Z}'\boldsymbol{V_d}^{-1}\boldsymbol{Z})^{-1}$. When the sample sizes for each group are sufficiently large for $\boldsymbol{d}$ to have an approximately multivariate normal distribution, confidence intervals and p-values of corresponding statistical tests for the covariate-adjusted log hazard ratio can be based on $(\hat{\delta} - \delta)V_\delta^{-1/2}$ having an approximately normal distribution. The rationale for randomization based covariance adjustment is the expected absence of differences between the test treatment and the control groups for means of the covariables. A related criterion for evaluating the extent of random imbalances between the test treatment and control groups is Q_0 in (3.3), which approximately has a chi-square distribution with q degrees of freedom.

$$Q_0 = (\boldsymbol{d} - \boldsymbol{Z}\hat{\delta})'\boldsymbol{V_d}^{-1}(\boldsymbol{d} - \boldsymbol{Z}\hat{\delta}) \tag{3.3}$$

3.2.2 Inverse probability weights using propensity score

In observational studies, direct adjustment or standardizations apply population weights to subclass means in an effort to estimate population quantities from a sample that is not representative of the population. Motivated by the Horvitz-Thompson

estimator, Rosenbaum (2000) proposed a model-based direct adjustment using inverse propensity score weights. Later, in the counterfactual framework, Robins et al. (2000) discussed an analogous approach (i.e.,inverse probability-of-treatment weighted estimators) to estimate causal treatment effects in observational studies, defined as an average response difference for the entire population if every individual had received one treatment versus the other.

Let $a = (0, 1)$ denote the control and the test treatment, and let $\boldsymbol{X}$ denote the set of q covariates to be adjusted, so that we have $(a_g, \boldsymbol{x}_g)$ observed for the gth patient in the study population with size n. For each patient g, the probability of receiving the test treatment conditional on the corresponding observed covariates, i.e., $p_g = \text{pr}(a = 1|\boldsymbol{X} = \boldsymbol{x}_g)$, is estimated by a logistic regression model

$$\text{logit } p_g = \gamma_0 + \boldsymbol{\gamma}\boldsymbol{x}_g. \tag{3.4}$$

Rosenbaum and Rubin (1983) refer to those p_g's as propensity scores and the logistic models for generating p_g's as propensity score models. Analysts can then assign each patient g a weight w_g equal to the inverse of the conditional probability of receiving the observed treatment, that is $w_g = 1/\hat{p}_g$ for the patient in the test treatment group and $w_g = 1/(1 - \hat{p}_g)$ for the patient in the control group. The effect of IPW is to create a pseudo population consisting of w_g copies of each subject g so that the pseudo sample size for either treatment group approximately equals the total sample size (Robins et al., 2000), i.e., $\sum a_g w_g \approx \sum(1 - a_g)w_g \approx n$. In practice, standardized weights (sw_g) shown in (3.5) are often implemented to construct a pseudo population for the combined treatment groups with the pseudo sample size equal to the total sample size,

i.e., $\sum a_g w_g + \sum(1 - a_g)w_g = n$.

$$sw_g = \frac{w_g}{(\sum_{g=1}^{n} w_g)/n} = \frac{n \times w_g}{\sum_{g=1}^{n} w_g} \tag{3.5}$$

If the model in (3.4) is correctly specified and provides a good fit of the data, the marginal distributions of the q covariates in either group are similar to those in the entire study population (i.e., the combined treatment groups). In addition, when all q covariates are categorical and included in a saturated propensity score model, the pseudo population for either treatment group is completely unrelated with the treatment assignment, and it has exactly the same sample size and marginal distributions of those adjusted covariates as the entire study population. Therefore, assessing the extent to which IPW balances the treatment groups is often useful for developing an appropriate propensity score model. Discussions of PS usages in IPW and other risk-adjustment approaches for observational studies can be found in Curtis et al. (2007) and Glynn et al. (2006).

Given an appropriate balance of covariates between the two treatment groups, the covariate-adjusted log hazard ratio ($\hat{\beta}_{\text{IPW}}$) can be estimated by fitting a (univariate) Cox regression model with treatments as the only factor and with weighing each patient by sw_g. The robust sandwich variance estimator is used to perform Wald tests and to obtain the confidence interval of $\hat{\beta}_{\text{IPW}}$. With the same weights, a covariate-adjusted cumulative survival function estimator can be produced separately for each treatment group via the corresponding cumulative hazards estimated by the weighted version of the Breslow estimator. By this way, the survivor curves with non-proportion hazards nature can be obtained from the SAS PHREG procedure, and these estimates provide good approximation to the weighted KM estimator that cannot be produced from the SAS LIFETEST procedure.

3.3 Sensitivity Analysis using Multiple Imputation

3.3.1 Unadjusted multiple imputation

The Kaplan-Meier Multiple Imputation (KMMI) strategy, implemented separately within individual treatment groups, was described in Zhao et al. (2012). Briefly, we assume that a randomized group has events observed at M distinct times ($t_1 < t_2 < \cdots < t_M$), and it has premature discontinuation of patients observed at K distinct times ($c_1 < c_2 < \cdots < c_K$). With k indexing the censoring times, $t_{k,0}$ denotes the latest failure time prior to c_k and $t_{k,j}$ denotes the jth failure time after c_k. The imputation scheme is as follow:

1. Obtain the Kaplan-Meier (KM) estimates $\hat{S}(t)$ for the survival distribution with support of $t \in (t_1, t_2, \ldots, t_M)$. For the end of follow-up time t^* or the censor times after the last failure time (i.e., $t_M < c_k < t^*$), an exponential model is used to extrapolate $\hat{S}(t_M)$ to $\hat{S}(t^*)$ or the corresponding $\hat{S}(c_k)$.

2. With the survivor rate $\hat{S}(c_k)$ (for the patient discontinuing treatment at the time $c_k \leq t_M$) defined by a linear interpolation of $\hat{S}(t_{k,0})$ and $\hat{S}(t_{k,1})$, the probability of having an event in the time interval $[t_{k,j}, t_{k,j+1}]$ conditional on not having the event by the time c_k is given by

$$\hat{f}_{k,j}(\theta) = \frac{\hat{S}(t_{k,j})^\theta - \hat{S}(t_{k,j+1})^\theta}{\hat{S}(c_k)^\theta}, \tag{3.6}$$

   where the *sensitivity parameter* θ is a fixed hazard ratio for a patient with premature discontinuation having an event after the censoring time c_k relative to the patients still remaining on their assigned treatment.

3. The discontinued patients have their censoring times replaced by the failure times

drawn at random from their corresponding conditional distributions with cumulative density function

$$\hat{F}_{k,j}(\theta) = 1 - \frac{\hat{S}(t_{k,j+1})^{\theta}}{\hat{S}(c_k)^{\theta}}. \tag{3.7}$$

More specific details are in in Zhao et al. (2012).

4. The imputation procedure is repeated to form L imputed data sets.

The imputed data sets do not have any patient with premature discontinuation, and so analysts can apply the conventional analysis methods for time-to-event data with the censoring only at the end of follow-up time t^*.

A usual way to perform sensitivity analyese is to perform KMMI in each group under various values of the θ that address different post-discontinuation tendencies of having events. The principle for specifying the sensitivity parameter was discussed previously (Zhao et al., 2012). Briefly, analysts could specify θ_T larger than that for the placebo group to penalize the premature discontinuation for the test treatment. With specifying $\theta_P = 1$ for the placebo group to approximate the non-informative independent censoring specification that patients with premature discontinuation would have comparable experience after discontinuation to their counterparts without premature discontinuation, $\theta = (\theta_T/\theta_P) = \theta_T$ for the test treatment becomes a single parameter for calibrating sensitivity analyses.

3.3.2 Covariate-adjusted multiple imputation

The covariate-adjusted multiple imputations can proceed in two different ways. The first method is covariate-adjusted proportional hazards multiple imputation (PHMI). For this method, the imputation scheme (2) - (4) in section 3.3.1 is implemented for every prematurely discontinued patient using a patient specific survival distribution

estimated by the Breslow estimator from the Cox proportional hazards model with treatments and the set of covariates. The covariate-adjusted hazard ratios can then be obtained from imputed data sets by fitting the same Cox regression model for the MI process.

An alternative approach to adjust for covariates is to estimate the covariate-adjusted survival distributions for the placebo and the test treatment groups from the pseudo populations with balanced covariate distributions constructed from the IPW method described in section 3.2.2, and then to follow the MI process (2) - (4) in section 3.3.1 with the covariate-adjusted counterparts. Using the same set of weights for estimating survival distributions, a univariate Cox model with treatments as the only factor can be applied to estimate the covariate-adjusted hazard ratio for each imputed data set.

3.3.3 Parameter estimation

Following well established rules (Rubin, 1987; Rubin and Schenker, 1991), the method for combining results from L imputed data sets can be applied easily by the SAS procedure MIANALYZE. Let β be the log hazard ratio that would be estimated from the complete data. Let $\hat{\beta}^{(l)}$ denote the point estimate for β and let $\hat{V}_{\beta}^{(l)}$ denote its variance estimate from the lth data set.

The overall multiple imputation (MI) estimate of β is obtained by averaging the estimates from the L complete-data analyses, $\bar{\beta} = (1/L)\sum_{l=1}^{L}\hat{\beta}^{(l)}$, and its estimated variance is the sum of the within-imputation variance $\bar{V}_{\beta} = (1/L)\sum_{l=1}^{L}\hat{V}_{\beta}^{(l)}$ and the product of the between-imputation variance $B_{\beta} = (L-1)^{-1}\sum_{l=1}^{L}(\hat{\beta}^{(l)} - \hat{\beta})^2$ and a finite sample correction shown in (3.8).

$$\hat{V}_{\bar{\beta}} = \bar{V}_{\beta} + (1 + L^{-1})B_{\beta} \tag{3.8}$$

Given sufficiently large sample size for the complete data to support an approximately standard normal $N(0,1)$ distribution for its hypothetical version of $(\hat{\beta}-\beta)\hat{V}_{\hat{\beta}}^{-1/2}$ when there were no missing data, confidence intervals for β (and p-values for corresponding statistical tests) can be based on $(\bar{\beta}-\beta)\hat{V}_{\bar{\beta}}^{-1/2}$ having a t-distribution with approximate degrees of freedom (d.f.) as shown in (3.9).

$$\begin{aligned} \text{d.f.} &= (L-1)\left(1+\left(\frac{(1+L^{-1})\,B_\beta}{\bar{V}_\beta}\right)^{-1}\right)^2 \\ &= (L-1)(1+R^{-1})^2 \end{aligned} \tag{3.9}$$

Here, R expresses the relative increase in variance due to missing information.

3.4 Application

3.4.1 Clinical trial example

We illustrate the proposed methods with a clinical trial for maintenance treatment for bipolar disorder (Calabrese et al., 2003). For reasons related to the confidentiality of the data from this clinical trial, the application uses a data set of 300 patients (150 patients with the test treatment and 150 patients with the placebo) from a random sample (with replacement) from the true study population. The same data set was also used previously in Zhao et al. (2012). After an 8 to 16 weeks run-in period within which all patients received test treatment, eligible patients who tolerated and adhered to the therapy were randomized to the test treatment or to the placebo, and then followed for 76 weeks. Accordingly, this study had a randomized withdrawal design, and the primary efficacy endpoint was the time-to-intervention for any mood episode. A total of 97 (32.33%) patients discontinued the study treatment prematurely (35% on the placebo and 29% on the test treatment). Of 300 patients, 75 patients (50.0%) on the

test treatment and 82 patients (54.7%) on the placebo had the event of intervention for any mood episode.

Seven covariables had a *priori* specification as being of interest in the analysis plan and in the protocol for this clinical trial. Two of them are patients' demographics, and the rest of them are baseline psychiatric assessments related to disease progression in previous studies. The distributions of these covariables are presented in Table 3.1. The extent of random imbalance between treatments is summarized for each covariate with the standardized difference (i.e., the difference between means divided by the square root of the average of the two sample variances) and the two-sided p-value from the Wilcoxon rank sum test for the association between the covariate and the treatment assignment. The standardized difference (Std. Diff.) represents the difference in means between two groups in units of the standard deviation (STD), and some authors suggest that Std. Diff. $< 10\%$ likely expresses a negligible imbalance (Austin et al., 2010). Table 3.2 describes the associations between the covariates and the primary endpoint, as assessed by the Cox regression models stratified on the treatment. Under the assumption of non-informative independent censoring, the univariate analyses for each individual covariate and the multivariate regression analysis were used to evaluate associations for the covariates. Of the five covariates with Std. Diff. $\geq 10\%$, the pre-randomized (pre-rand) MRS 11 item total score has strong association with the primary endpoint (p-value of 0.002 in the univariate analysis and p-value of 0.003 in the multivariate analysis), whereas the pre-rand CGI-I score, the pre-rand CGI-S score, and the pre-rand GAS score have weak associations with the outcome ($0.05 \leq$ p-values ≤ 0.15). The pre-rand CGI-I score has the largest Std. Diff. of 23.6% with p-value < 0.05 for the Wilcoxon test of imbalance. Although the random imbalance criterion in (3.3) does not contradict the expected balance of covariables from randomization (p-value=0.257), the possibility of random imbalance is suggested. The distribution for

the pre-rand CGI-I score favors the placebo group, but the random imbalance of the pre-rand CGI-S score, the pre-rand GAS score, and the pre-rand MRS 11 item total score could make the test treatment group to have better outcome.

Table 3.1: Distribution of patients' baseline characteristics

(N) Covariables	Overall (300) Mean	Test treatment (150) Mean (STD)	Placebo (150) Mean (STD)	Std Diff (%) (Test - Placebo)	p-values (Wilcoxon test)
Age	42.9	42.7 (11.4)	43.2 (13.3)	-5	0.5499
Female (%)	49.7	45.3 (50.0)	54.0 (50.0)	-17	0.1340
Pre-rand CGI-I score	1.67	1.60 (0.58)	1.74 (0.61)	-24	0.0468
Pre-rand CGI-S score	2.02	1.99 (0.71)	2.06 (0.77)	-10	0.3519
Pre-rand GAS score	75.1	76.0 (9.8)	74.2 (11.0)	17	0.2398
Pre-rand MRS 11 item total score	1.59	1.47 (2.51)	1.72 (2.80)	-10	0.7930
Pre-rand HAMD 17 item total score	5.84	5.69 (4.23)	5.99 (4.17)	-7	0.4217

Table 3.2: Association of patients' baseline characteristics and the primary outcome (assessed with Cox model)

Covariables	Univariate analysis		Multivariate analysis	
	Coefficient (SE)	p-value	Coefficient (SE)	p-values
Age	-0.001 (0.007)	0.9323	0.001 (0.007)	0.8904
Female (proportion)	0.046 (0.162)	0.7760	-0.004 (0.164)	0.9819
Pre-rand CGI-I score	0.055 (0.128)	0.6693	-0.380 (0.219)	0.0822
Pre-rand CGI-S score	0.149 (0.104)	0.1533	0.260 (0.165)	0.1152
Pre-rand GAS score	-0.013 (0.008)	0.0928	-0.015 (0.011)	0.1513
Pre-rand MRS 11 item total score	0.073 (0.025)	0.0038	0.077 (0.026)	0.0030
Pre-rand HAMD 17 item total score	0.023 (0.019)	0.2391	0.003 (0.025)	0.9005

3.4.2 Covariate-adjusted analyses with MAR-like assumption

Analyses first proceed with the censoring of follow-up times for patients with premature discontinuation of their assigned treatment, and so they have the MAR-like assumption of non-informative independent censoring. The robust sandwich variance estimator is used for hypothesis testing and to obtain confidence intervals throughout the application. The analysis results are shown in Table 3.3. With a Cox PH model with one explanatory variable for treatments (i.e., univariate Cox model), the unadjusted log hazard ratio (HR) for comparing test treatment versus placebo, is estimated by -0.393 with standard error (SE) of 0.1597 and p-value of 0.0138, indicating superiority of the test treatment. The multivariable Cox model, with the assumption of proportional hazards for treatment and all seven covariates, produces a larger estimate for the treatment effect (covariate-adjusted log HR of -0.410) , a larger SE (0.167) and a slightly larger p-value of 0.0142 than the unadjusted Cox regression counterparts. When adjusting for the covariates via the NPANCOVA method, the estimated covariate-adjusted log HR is somewhat closer to the null (-0.374) than the unadjusted Cox estimates. With a slightly reduced SE (0.1562), NPANCOVA produces a somewhat larger p-value (0.0167). The decreased treatment effect after covariate adjustment with NPANCOVA is probably due to the random covariate imbalance favoring the test treatment group in the unadjusted analysis. Conversely, the Cox model with covariate adjustment often produces a point estimate for the treatment effect that is further from the null, mainly because it pertains to patients with the same profile of covariates in contrast to the population average nature of the unadjusted estimate or the adjusted estimate produced by the NPANCOVA method (Tangen and Koch, 2000; Jiang et al., 2008; Saville and Koch, 2012).

To apply inverse probability weights (IPW) and propensity scores (PS), we first fit a multivariable propensity model with the structure shown in (3.4). The propensity

model has treatment as the outcome and the set of seven covariates as explanatory variables. After weighing each subject by the sw_g defined in (3.5), the extent to which the IPW balances the treatment groups is assessed, and the corresponding results are summarized in Table 3.4. The standardized weights generate a pseudo sample size of 300 for the combined treatment groups (about 150 for each group). Within the pseudo population, the covariate means for the test treatment and the placebo groups are very similar (Std Diff $\leq 1\%$ and p-value > 0.9) and they are all approximating the overall population covariate means (shown in Table 3.1). As shown in Table 3.3, the IPW method produces a covariate-adjusted log HR estimate which is similar in value to that using the NPANCOVA estimator, but a larger SE, and consequently a larger p-value.

Table 3.3: Covariate-adjusted analyses for treatment effects under the MAR-like assumption

Method	Parameter (SE)	HR (95% CI)	p-values
Univariate Cox model	-0.3931 (0.1597)	0.675 (0.494, 0.923)	0.0138
Multivariable Cox model	-0.4097 (0.1670)	0.664 (0.479, 0.921)	0.0142
NPANCOVA	-0.3738 (0.1562)	0.688 (0.507, 0.935)	0.0167
IPW	-0.3787 (0.1638)	0.685 (0.497, 0.944)	0.0208

Table 3.4: Characteristics of the pseudo population created by standardized weights using IPW method

(Pseudo N) Covariables	Overall (300) Mean	Test treatment (150.15) Mean (STD)	Placebo (149.85) Mean (STD)	Std Diff (%) (Test - Placebo)	p-values
Age	43.0	43.1 (11.4)	43.0 (13.0)	1	0.9505
Female (%)	49.7	49.8 (50.0)	49.5 (50.0)	0.5	0.9734
Pre-rand CGI-I score	1.68	1.68 (0.59)	1.68 (0.59)	-0.4	0.9788
Pre-rand CGI-S score	2.03	2.03 (0.70)	2.03 (0.59)	-0.4	0.9825
Pre-rand GAS score	75.2	75.2 (9.6)	75.1 (10.9)	1	0.9383
Pre-rand MRS 11 item total score	1.60	1.61 (2.86)	1.59 (2.61)	0.8	0.9602
Pre-rand HAMD 17 item total score	5.90	5.90 (4.29)	5.90 (4.24)	-0.03	0.9988

Table 3.5: Sensitivity analysis with specification of $\theta = 1$

MI method	Analysis method	Parameter (SE)	HR (95% CI)	p-values
1. Unadjusted PHMI	Univariate (unadjusted) Cox model	-0.3885 (0.1578)	0.678 (0.498, 0.924)	0.0139
2. Unadjusted PHMI	Multivariable Cox model	-0.3725 (0.1626)	0.689 (0.500, 0.948)	0.0222
3. Covariate-adjusted PHMI	Multivariable Cox model	-0.3886 (0.1609)	0.678 (0.495, 0.930)	0.0159
4. Unadjusted KMMI	Univariate (unadjusted) Cox model	-0.3225 (0.1591)	0.724 (0.530, 0.990)	0.0429
5. Unadjusted KMMI	Multivariable Cox model	-0.3055 (0.1643)	0.737 (0.534, 1.02)	0.0632
6. Unadjusted KMMI	NPANCOVA	-0.3092 (0.1579)	0.734 (0.528, 1.00)	0.0505
7. Unadjusted KMMI	Univariate Cox model with IPW	-0.3126 (0.1633)	0.732 (0.531, 1.01)	0.0558
8. Covariate-adjusted KMMI with IPW	Univariate Cox model with IPW	-0.3172 (0.1640)	0.728 (0.528, 1.00)	0.0534

Table 3.6: Key steps and assumptions in the performance of sensitivity analyses under $\theta = 1$

	Multiple imputation strategy		Analysis for imputed data	
	Methods for obtaining survivor estimates	PH assumption for survival distributions	Analysis method	PH assumption for treatment groups
1.	Breslow estimator from univariate Cox model with treatment as explanatory variable	Required for treatment groups	Univariate Cox model with treatment as explanatory variable	Not required
2.	Breslow estimator from univariate Cox model with treatment as explanatory variable	Required for treatment groups	Multivariable Cox model with treatment and the seven covariables	Required
3.	Breslow estimator from multivariable Cox model with treatment and covariables	Required for treatment groups and covariates	Multivariable Cox model with treatment and the seven covariables	Required
4.	KM estimator for individual groups	Not required	Univariate Cox model with treatment as the only explanatory variable	Not required
5.	KM estimator for individual groups	Not required	Multivariable Cox model with treatment and the seven covariables	Required
6.	KM estimator for individual groups	Not required	Non-parametric ANCOVA	Not required
7.	KM estimator for individual groups	Not required	Univariate Cox model with treatment only and weighing each patient by sw_i	Not required
8.	Weighted version of KM estimator using sw_i (approximated by the weighted Breslow estimator)	Not required	Univariate Cox model with treatment only and weighing each patient by sw_i	Not required

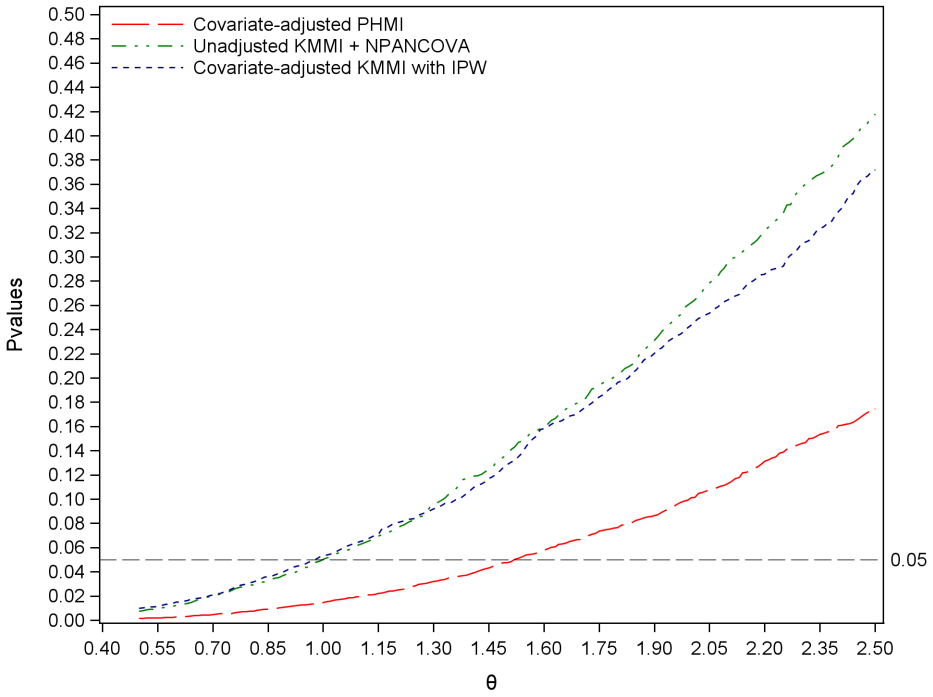


Figure 3.1: Sensitivity analyses with covariate adjustment

3.4.3 Sensitivity analyses with covariate adjustment

We first implement the sensitivity analysis without and with covariate adjustment under $\theta = 1$. With this specification, the imputed data are produced from the conditional failure time distributions estimated with the censoring of the follow-up times of patients with premature discontinuation, and they thereby have the MAR-like assumption of non-informative independent censoring. We perform the multiple imputations (MI) with $L = 50$ for the amount of missing information in this example. The justification for $L = 50$ was discussed in Zhao et al. (2012). Table 3.5 presents the covariate-adjusted (log) hazard ratios obtained from the combinations of the MI strategies and the covariate adjustment methods for the imputed data sets. An important component that differentiates various MI procedures is the survival distribution estimates, from which the conditional failure time distribution for imputation is constructed. Table 3.6

summarizes the methods for estimating the survival distributions, along with the other key steps and assumptions for the corresponding sensitivity analyses presented in Table 3.5. For the ease of comparisons, the unadjusted (log) hazard ratios were also estimated using the KMMI and the PHMI methods without covariate adjustment.

When imputed data sets are produced by the unadjusted PHMI method under $\theta = 1$, the unadjusted (log) hazard ratio (row 1) is very similar to that obtained via the conventional unadjusted Cox model with the censoring of follow-up times for discontinued patients (Table 3.4 row 1), mainly because both under the MAR-like assumption of non-informative censoring and both have the proportional hazards assumption. With the data sets imputed by the unadjusted PHMI method, a multivariate (i.e., covariate-adjusted) Cox regression model produces a smaller treatment effect estimate and a larger SE (row 2) than the unadjusted counterparts. Interestingly, the covariate-adjusted PHMI method (row 3) produces a covariate-adjusted log hazard ratio of -0.3886 that is comparable in value to the unadjusted log hazard ratio estimate of -0.3885 from the unadjusted PHMI method (row 1). And its corresponding SE estimate is in between the SE estimates from the unadjusted and adjusted Cox regression analyses for the data imputed by the unadjusted PHMI method (row 1 and 2).

When data are from randomized clinical trials, one could regard the patients of each treatment group as a random sample from the study population. Therefore, it is appropriate to apply either unadjusted or covariate-adjusted analysis to the data sets imputed by the KMMI method without covariate adjustment. The unadjusted log hazard ratio from the unadjusted KMMI method (row 4) under $\theta = 1$ is closer to the null and has a somewhat larger p-value than its conventional counterpart with the use of censoring (log HR=-0.3225 with p=0.0429 for unadjusted KMMI versus log HR=-0.3931 with p=0.0138 for the conventional method), due to the non-proportional hazards for the follow-up period (i.e., a much stronger effect size for the test treatment

during the early period in the example). With the same set of imputed data, the estimators for covariate-adjusted (log) hazard ratio from the multivariable Cox regression (row 5), the NPANCOVA (row 6), and the univariate Cox model with IPW (row 7) produce even smaller (i.e., closer to the null) treatment effects than the unadjusted Cox model estimator (row 4), which leads to larger p-values (> 0.05) for all three covariate-adjusted methods. Among those covariate adjustment analyses, only the NPANCOVA method generates a SE estimate smaller than that of the unadjusted Cox regression, whereas the multivariable Cox regression produces the smallest effect size and the largest SE estimates, and hence has the largest p-value. The covariate-adjusted KMMI strategy imputes failure times based on the survival distributions estimated from the balanced pseudo populations created by IPW. The data sets imputed under $\theta = 1$ are then analyzed using the univariate Cox model with the same weights. As shown in Table 3.5 row 8, it provides a slightly stronger adjusted result than the unadjusted KMMI method paired with covariate adjustment using IPW (row 7). Except for the unadjusted and adjusted PHMI methods supporting superiority of the test treatment, most of the sensitivity analyses with covariate adjustment only show marginal benefits for the test treatment under $\theta = 1$.

We then conduct the covariate-adjusted sensitivity analyses, i.e., the unadjusted KMMI with NPANCOVA (Table 3.5 and 3.6 row 6), the covairate-adjusted PHMI (Table 3.5 and 3.6 row 3), and the covariate-adjusted KMMI with IPW (Table 3.5 and 3.6 row 8), by varying the sensitivity parameter θ ($= \theta_T/\theta_P$) for the test treatment group. For sensitivity analyses in the regulatory setting, one would usually have $\theta_P = 1$ and $\theta_T > \theta_P > 1$ to penalize premature discontinuations for the test treatment. The choice of θ can be values in a range of (L, U), where $(1/U, 1/L)$ is a range of hazard ratios from previous related studies or clinical judgment for the comparison of effective medicines with placebo. Here, we set a lower bound < 1 and vary the value of θ by an

0.01 increment in a range from 0.5 to 2.5, mainly because two of the three sensitivity analysis methods fail to show the superiority of the test treatment under $\theta = 1$ for this example. The p-values for the covariate-adjusted (log) hazard ratios from the unadjusted KMMI with NPANCOVA, the covairate-adjusted PHMI, and the covariate-adjusted KMMI with IPW are plotted as functions of the sensitivity parameter θ in Figure 3.1. In order to have $p < 0.05$ via the unadjusted KMMI with NPANCOVA or the covariate-adjusted KMMI with IPW, $\theta < 1$ or < 0.98 is needed, which may not seem to be a reasonable assumption for the post-discontinuation behavior of patients with the test treatment. For $p < 0.05$ with the covariate-adjusted PHMI , $\theta < 1.52$ is needed, suggesting better robustness to assumptions about patients with premature discontinuation of treatment for this example. Compared with the unadjusted hazard ratio estimates obtained from the unadjusted PHMI method with the specification of $\theta > 1$ (presented in Zhao et al. (2012)), the covariate-adjusted PHMI method produces slightly weaker results, i.e. treatment effect estimates that are closer to the null and have larger SE estimates and larger p-values.

3.5 Summary

Covariate adjustment plays an important role in the analysis of observational studies and randomized clinical trials. In observational studies where the equivalence of comparison groups cannot be controlled by randomization, covariance analysis adjusts for inherent differences among comparison groups so that bias may be reduced (Koch et al., 1982, 1998). For randomized studies, covariate adjustment may provide more powerful statistical tests (relative to their unadjusted counterparts) for the comparison between treatment groups (Koch et al., 1982, 1998). In this paper, we discussed three covariance analysis methods for time-to-event data through an example from a clinical trial for a maintenance treatment of bipolar disorder, in which substantial premature

discontinuations of treatment occurred. The goal of this paper is to illustrate how to adapt those methods of covariate adjustment to the sensitivity analysis for assessing the robustness of conclusions to the management of missing information.

The multivariable Cox proportional hazards model is commonly employed for covariate adjustment in both observational and randomized studies. However, the appropriate application depends on several assumptions, such as correct model specification and proportional hazards for each variable in the model. When the proportional hazards assumption is not satisfied, the type I error is inflated for the Cox model with adjustment for covariables that are related to the outcome (Jiang et al., 2008). With adjustment for covariates, the treatment parameter estimates from the Cox model are often further from the null, and the corresponding SE estimates are always larger than the unadjusted counterparts. Therefore, the efficiency of the null hypothesis test of no treatment effect may not be clear for covariate adjustment (Hauck et al., 1998). To implement the covariate adjustment with multivariable Cox models in the sensitivity analysis for missing data, the imputed data sets are generated and analyzed by the Cox proportional hazards model with treatment and the set of covariates to be adjusted; and this could lead to the same issues as previously noted and cause concerns for interpreting the adjusted results, especially in the regulatory setting.

The NPANCOVA method proposed by Saville and Koch (2012) has random assignment of treatments as the principal assumption and avoids the major issues associated with the multivariable Cox proportional hazards model. Unlike the multivariable Cox model, the NPANCOVA method is more likely to preserve the type I error under non-proportional hazards and is more robust for different model assumptions (Jiang et al., 2008; Saville and Koch, 2012). The covariate-adjusted hazard ratio produced by NPANCOVA has the interpretation of a population average treatment effect, in contrast to the subpopulation (defined by adjusted covariates) specific estimates provided by the

multivariable Cox model. If adjusted covariates explain some of the variation in the response variable, the NPANCOVA method could generate more powerful statistical tests through variance reduction (Koch et al., 1982, 1998). In addition, the covariate adjustment with NPANCOVA induces equivalent comparison groups by offsetting random imbalances between treatment groups for covariables with noteworthy associations with the outcome of interest, and thereby it provides clarification of the degree to which the detected difference between randomized groups for the response variable is due to treatment rather than random imbalances for covariates. For the application data set with more random covariate imbalances favoring the test treatment group to have better response, the NPANCOVA method produces a covariate-adjusted log hazard ratio closer to the null, and a larger p-value than the counterparts produced by the unadjusted Cox regression even with the SE reduction. The covariate-adjusted sensitivity analysis with NPANCOVA invokes the unadjusted KMMI process, and therefore it generates somewhat weaker results than its conventional counterpart with the use of censoring for this particular example.

To reduce bias for comparing treatment effects in observational studies, the inverse probability weights (IPW) balances the distributions of the covariables to be adjusted across treatment groups by creating a pseudo population for each group that has covariable distributions comparable to those of the combined treatment groups. For a univariate Cox model with only treatment for the re-weighted pseudo population, the IPW method produces a covariate-adjusted hazard ratio with a population average nature similar to that from the NPANCOVA method, and it does not require the proportional hazards assumption needed for covariates by the conventional multivariable Cox regression. As shown in the example (Table 3.3 and 3.5), the IPW method reduces the bias due to the random covariate imbalance for treatment comparisons, but it produces larger SE estimates and larger p-values, and consequently, it provided more

conservative hypothesis tests than the NPANCOVA method. A comprehensive simulation study may be able to shed more light on its performance for covariate adjustment. Although the IPW method is usually used in observational studies, its appropriateness in the covariate-adjusted sensitivity analysis for missing data can be justified for both observational studies and confirmatory clinical trials.

Chapter 4

Sensitivity Analysis for Withdrawals in Grouped Time-to-event Data

4.1 Introduction

Grouped time-to-event data often arise in longitudinal clinical trials, where patients are evaluated repeatedly by diagnostic procedures at a specific set of follow-up times until the event of interest occurs or until completion of the entire follow-up period. Examples include assessing progression free survival in oncology by biopsy or imaging, examining patients with an endoscope for ulcer healing or recurrence, or occurrence of certain psychological characteristics which may not be immediately obvious to the patients, and thereby need clinical evaluation. In these situations, one can only determine the time interval during which the event occurs, but the exact time of failure is unknown. Further description of examples generating the grouped survival data is given in Johnson and Koch (1978), and Laird and Olivier (1981).

The general framework for analyzing grouped survival data was discussed in detail by Koch et al. (1972), Johnson and Koch (1978), and Deddens and Koch (1988). Within the context of longitudinal clinical trials, some patients may discontinue the study prior to its completion (i.e., withdraw or dropout) without having an observed event for

reasons such as protocol violations, adverse events, worsening of symptoms, or unrelated illness or some other reasons. Such patients always complicate the analysis and the interpretation of the data for efficacy comparisons of treatment groups. Moreover, the incomplete follow-up due to withdrawal can reduce the comparability of treatment groups provided by randomization, consequently it can undermine the validity of the trial and can lead to ambiguous study conclusions (NRC, 2010). Therefore, a central issue for grouped survival data analysis is how to take into account the incomplete follow-up information appropriately, and some methods are described in references such as Elashoff and Koch (1991) and Somerville et al. (2009).

One way to manage the withdrawal patients is to designate them as 'not having event' through the end of study. This approach gives rise to the crude event rate during a time period as the ratio of the number of subjects with the event versus the number of all randomized patients (Elashoff and Koch, 1991; Somerville et al., 2009). Although this convention has an intention-to-treat (ITT) spirit, it underestimates the event rate by unrealistically presuming all withdrawal patients would not experience any event through the entire study period. Another approach is to view the withdrawals as having the event, and this leads to overstated event rates. When one treatment group has a greater prevalence of withdrawals than the other, the treatment comparison using either method could be misleading and would need to be interpreted with caution (Koch et al., 1984).

Another way to manage patients who prematurely withdraw during an interval is the actuarial method that excludes them from the risk set for that interval and subsequent intervals, since their actual status after their last visit is unknown (Koch et al., 1984; Elashoff and Koch, 1991). This convention assumes withdrawals have no association with the tendency of having an event, in other words, the subsequent unobserved event rate for discontinued patients is the same as the observed event rate for patients who

remain in the study. Similar in spirit to the usual life-table estimates that assume non-informative independent censoring, the actuarial survival rate through the end of a specific interval is the product of the proportion of patients with no event for it and those for all preceding intervals. Although the actuarial method is adopted in most analysis strategies, it is useful to note that it is in fact a per-protocol analysis, because the method seeks to estimate the event rates would have been if the withdrawers had remained under study treatment.

Since the survival status of discontinued patients remains unknown subsequent to their last follow-up visit, there is no single correct way of computing the effective number of patients at risk during a particular time interval (Koch et al., 1972). For this reason, event rates are often summarized using multiple methods to demonstrate the robustness of primary conclusions, and one should always evaluate the sensitivity of the primary analysis results to alternative ways of managing incomplete follow-up data (Koch et al., 1984; Somerville et al., 2009).

Although the actuarial method for survival rates analysis is based on the product of conditional probability parameters, Koch et al. (1972) have shown that the grouped survival data can also be arranged in a contingency table format so that the underlying probability model can be written as a product multinomial distribution with unconditional probability parameters. Using the same computational framework, we develop a sensitivity analysis to assess the implication of withdrawals to the conclusion concerning the treatment comparison. Governed by the intention-to-treat principle, patients who discontinue study treatment prematurely should not be censored, but rather managed as if followed for the event of interest, at a possible higher event rate. The proposed method computes the hypothetical unconditional multinomial distributions of grouped survival time for all randomized patients, as if all the discontinued patients had been followed to the end of study in the absence of their assigned treatment. A sensitivity

parameter θ is introduced to the calculation as the conditional odds ratio of having an event in a time interval following withdrawal, to reflect the different event rates of patients who withdraw from study and patients who remain in the study. Different values of θ can be specified separately for the test and control treatment groups to cover a wide range of possible post-withdrawal experiences as alternatives to the primary non-informative independent censoring assumption. The extent to which the treatment inference changes over a range of θ allows a more complete assessment of the robustness of primary results in terms of different managements of withdrawal patients. The application of this proposed sensitivity analysis is illustrated using data from a maintenance trial for ulcer disease.

4.2 Methods

4.2.1 Data structure

Grouped survival data can usually be analyzed in a categorical data framework according to the occurrence of the follow-up status of patients with respect to a set of mutually exclusive time intervals (with possibly unequal lengths). Consider a study comparing a test and a control treatment to prevent the occurrence of a medical condition over t consecutive time intervals, where patients were evaluated at the end of the kth interval , for $k = 1, 2, \ldots, t$, until they experience the event of interest (i.e., fail) shown in (4.1). For each treatment group, the observed data can be arranged in the contingency table format. Here, f_k and w_k represent, respectively, the number of patients who fail or withdraw in the kth time interval, the quantity w_{t+1} represents the number of patients who are followed and survive (i.e. event-free) through all t time

intervals, and $n = \sum_{k=1}^{t} f_k + \sum_{k=1}^{t+1} w_k$ is the total number of patients in that group.

Fail during follow-up				Withdraw during follow-up				Censor	Total
1	2	$\cdots$	t	1	2	$\cdots$	t	(at end)	
f_1	f_2	$\cdots$	f_t	w_1	w_2	$\cdots$	w_t	w_{t+1}	n

(4.1)

One can consider each treatment group to have a multinomial distribution in which each patient has one of the $(2t+1)$ mutually exclusive outcomes. Then, the relevant model for observing a specific set of frequencies f_k and w_k is given in (4.2),

$$\phi = \frac{n!}{\prod_{k=1}^{t} f_k! \prod_{k=1}^{t+1} w_k!} \left(\prod_{k=1}^{t} \pi_{1k}^{f_k} \prod_{k=1}^{t+1} \pi_{2k}^{w_k} \right) \tag{4.2}$$

where π_{1k} is the probability that an individual will fail at some time during the kth time interval for $k = 1, 2, \ldots, t$, π_{2k} is the probability that an individual will withdraw in the kth time interval for $k = 1, 2, \ldots, t$, π_{2k}, and $\pi_{2(t+1)}$ is the probability that an individual will be followed through the entire t intervals and not fail. From the properties of the multinomial distribution, the unbiased estimator of $\boldsymbol{\pi}$, $\boldsymbol{\pi}$= $(\pi_{11}, \ldots, \pi_{1t}, \pi_{21}, \ldots, \pi_{2t}, \pi_{2(t+1)})'$, is $\boldsymbol{a}$ in (4.3).

$$\hat{\boldsymbol{\pi}} = \boldsymbol{a} = \frac{1}{n}(f_1, f_2, \ldots, f_t, w_1, w_2, \ldots, w_t, w_{t+1})' \tag{4.3}$$

Moreover, a consistent estimator for the variance-covariance matrix of $\boldsymbol{a}$ is

$$\hat{\text{Var}}(\boldsymbol{a}) = \boldsymbol{V_a} = \frac{1}{n}[\boldsymbol{D_a} - \boldsymbol{a}\boldsymbol{a'}] \tag{4.4}$$

where $\boldsymbol{D_a}$ is a $(2t+1) \times (2t+1)$ diagonal matrix with the elements of the vector $\boldsymbol{a}$ on the main diagonal.

4.2.2 Survival/failure probability estimation

In order to specify a survival/failure profile for each treatment group, we adopt the convention that patients are known to be event-free only to their last visits, and are censored at the beginning of the time interval in which they withdraw. One can conditionally view the f_k as following uncorrelated binomial distributions with parameters n_k and λ_k, where $n_k = \sum_{j=k}^{t} f_j + \sum_{j=k+1}^{t+1} w_j$ is the number of patients at risk during the kth interval, and λ_k is the conditional probability for failing in the kth interval given survival of all preceding intervals. From the properties of the binomial distribution, $\hat{\lambda}_k = h_k = f_k/n_k$, is an unbiased estimator for λ_k, and the variance of h_k can be estimated consistently by

$$\hat{\mathrm{Var}}(h_k) = \frac{1}{n_k}(h_k)(1 - h_k). \tag{4.5}$$

Since the conditional binomial distributions for the f_k within each time interval are uncorrelated, one can note that $\mathrm{Cov}(h_k, h_{k'}) = 0$ for $k \neq k'$.

It is also useful to express the h_k's, for $k = 1, 2, \ldots, t$, simultaneously in matrix notation as compound functions of the vector $\boldsymbol{a}$, via a sequential series of linear, logarithmic, and exponential transformations

$$\boldsymbol{h} = (h_1, h_2, \ldots, h_t)' = \exp[\boldsymbol{A_2} \log_e(\boldsymbol{A_1 a})], \tag{4.6}$$

where $\log_e(\ )$ denotes the element-wise vector operation that transforms a vector to the corresponding vector of natural logarithms, and $\exp[\ ]$ denotes the element-wise vector operation that transforms a vector to the corresponding vector of exponentiated values, and $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$ are matrices for linear transformations (for which the structures are described in **??**). The principal reason for expressing $\boldsymbol{h}$ in this matrix framework is that it facilitates the construction of the consistent estimator for the covariance matrix

of $\boldsymbol{h}$ using the first order linear Taylor series approximation, as discussed in Koch et al. (1972). After some matrix algebra, we show that the covariance matrix $\boldsymbol{V_h}$ estimated by this approach is actually a diagonal matrix with $\frac{1}{n_k}h_k(1-h_k)$ (the quantity shown in 4.5) being the kth diagonal element. The details for obtaining $\boldsymbol{V_h}$ are given in Appendix 4. It is not difficult to note that the matrix method employing the first order linear Taylor series approximation simply produces the appropriate variance and covariance quantities, with the advantage that a large number of parameters can be processed simultaneously.

By the definition of conditional probability, the actuarial (i.e., life table) estimator for the probability that an individual will fail in the kth time interval is computed as

$$p_1 = h_1 \text{ and } p_k = h_k \prod_{j=1}^{k-1}(1-h_j), \text{ for } k = 2,\, 3,\, \ldots,\, t, \tag{4.7}$$

when assuming non-informative independent censoring for withdrawal patients. Let $\boldsymbol{p} = (p_1, p_2, \ldots, p_t)'$, and so the consistent covariance matrix estimator $\boldsymbol{V_p}$ can be obtained through the linear Taylor's series approximations as shown in Appendix 5. With the specification in (4.7), the following relationship can be established,

$$\begin{aligned}
\frac{p_1}{1-p_1} &= \frac{h_1}{1-h_1} \\
\frac{p_2}{1-p_1-p_2} &= \frac{(1-h_1)h_2}{1-h_1-(1-h_1)h_2} = \frac{h_2}{1-h_2} \\
\frac{p_3}{1-p_1-p_2-p_3} &= \frac{(1-h_1)(1-h_2)h_3}{1-h_1-(1-h_1)h_2-(1-h_1)(1-h_2)h_3} \\
&= \frac{(1-h_2)h_3}{1-h_2-(1-h_2)h_3} = \frac{h_3}{1-h_3} \\
&\vdots
\end{aligned}$$

As a deduction, we obtain

$$\frac{p_k}{1-\sum_{j=1}^{k} p_j} = \frac{h_k}{1-h_k} \text{ , for } k = 1,\ 2,\ \ldots,\ t, \tag{4.8}$$

where the left-hand side of the equation is the estimated marginal odds for failing in the kth time interval (versus surviving beyond it), and the right-hand side of the equation is the estimated conditional odds for failing in the same interval given survival of all proceeding intervals.

The actuarial method assumes that the experience of the censored patients following their withdrawals is represented by the patients remaining in the risk set. This non-informative independent censoring assumption is in a sense like the missing at random (MAR) assumption in the language of missing data. Under this assumption, one can impute the failure times for the withdrawals w_1, w_2, ..., and w_t, using the failure probability distribution estimated from the observed data. The marginal failure probability distribution (p_1, p_2, ..., and p_t) is employed to impute failures for the withdrawals w_1. And for w_k, $k = 2, \ldots, t$, the imputations are based on the failure probabilities conditional on survival through the first $k-1$ intervals as estimated by

$$\frac{1}{1-\sum_{j=1}^{k-1} p_j}(p_k\ ,\ p_{k+1}\ ,\ \cdots\ ,\ p_t)\ .$$

Henceforth, one can redistribute the counts of w_k into the time intervals of k, $k+1$, ..., t. The counterpart estimates to the p_k's in (4.7), are easily obtained via (4.9) for $k = 1, 2, \ldots, t$, as if all the randomized patients were followed to the end of the study.

$$
\begin{aligned}
q_1 = p_1 &= \frac{1}{n}\left(f_1 + p_1 w_1\right) \qquad (4.9)\\
q_2 = p_2 &= \frac{1}{n}\left(f_2 + p_2 w_1 + \frac{p_2}{1-p_1} w_2\right)\\
q_k = p_k &= \frac{1}{n}\left(f_k + p_k w_1 + \frac{p_k}{1-p_1} w_2 + \frac{p_k}{1-p_1-p_2} w_3 + ... + \frac{p_k}{1-\sum_{j=1}^{k-1} p_j} w_k\right)\\
&= \frac{1}{n}\left(f_k + \sum_{g=1}^{k} \frac{p_k w_g}{1-\sum_{j=1}^{g-1} p_j}\right), \text{ with convention } \sum_{j=1}^{0} p_j = 0\\
&\vdots\\
q_t = p_t &= \frac{1}{n}\left(f_t + \sum_{g=1}^{t} \frac{p_t w_g}{1-\sum_{j=1}^{g-1} p_j}\right)
\end{aligned}
$$

By substitution of (4.7) into (4.9), the marginal failure probability distribution (i.e. q_k's for $k = 1, 2, \ldots, t$) can then be expressed as compound functions of the elements of the vectors $\boldsymbol{a}$ and $\boldsymbol{h}$ as shown in (4.10).

$$
\begin{aligned}
q_1 &= \frac{1}{n}\left[f_1 + h_1 w_1\right] \qquad (4.10)\\
q_2 &= \frac{1}{n}\left[f_2 + (1-h_1) h_2 w_1 + h_2 w_2\right)\\
q_k &= \frac{1}{n}\left[f_k + h_k \prod_{j=1}^{k-1}(1-h_j) w_1 + h_k \prod_{j=2}^{k-1}(1-h_j) w_2 + \cdots + h_k(1-h_{k-1}) w_{k-1} + h_k w_k\right]\\
&= \frac{1}{n}\left\{f_k + h_k\left[\sum_{g=1}^{k-1}\left(w_g \prod_{j=g}^{k-1}(1-h_j)\right) + w_k\right]\right\}\\
&\vdots\\
q_t &= \frac{1}{n}\left\{f_t + h_t\left[\sum_{g=1}^{t-1}\left(w_g \prod_{j=g}^{t-1}(1-h_j)\right) + w_t\right]\right\}\\
q_{(t+1)} &= 1 - \sum_{k=1}^{t} q_k \qquad (4.11)
\end{aligned}
$$

Also, $q_{(t+1)}$ in (4.11) is the estimated probability that an individual would have an event

after the tth time interval (i.e., the probability of survival through all t intervals).

4.2.3 General framework of sensitivity analysis

In discrete time-to-event analysis, the hazard function $\lambda(t_j)$ is a non-zero probability of experiencing an event at a time t_j, conditional upon the event of interest not occurring prior to that time point. Cox (1972) proposed an extension of the proportional hazards model to discrete time by working with the conditional odds of failing at each time t_j given survival up to that point

$$\frac{\lambda(t_j)}{1-\lambda(t_j)} = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\frac{\lambda_0(t_j)}{1-\lambda_0(t_j)},$$

where $\lambda(t_j)$ is the hazard at time t_j for an individual with covariates $\boldsymbol{x_i}$, $\lambda_0(t_j)$ is the baseline hazard rate at time t_j, and $\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ is the odds ratio associated with covariates $\boldsymbol{x_i}$. The logit of the hazard rate (or the logit of the conditional probability of having an event at time t_j given event-free up to that time) can then be the basis for fitting logistic regression models. The similar concept can be applied to the grouped survival data analysis. In this regard, we could view the marginal/conditional odds ratio for failing, i.e., the ratio of the quantities in (4.8) for one group to the other, as a useful criterion for evaluating the treatment difference.

For most clinical trials, withdrawals are likely to be associated with a change in treatment regime, i.e., discontinuation of the study treatment, or even switching to a rescue therapy. Consequently, the experience of developing an event for the remaining patients will no longer represent the response of the entire treatment group if every patient were followed without withdrawal. For this reason, let $\tilde{h}_k$ denote the probability of failing in the kth time interval conditional on survival of all preceding intervals for patients who drop out before or in the kth interval. Then, a different survival profile

for discontinued patients can be specified through a conditional odds ratio (θ) of failure for discontinued vs. retained patients as

$$\frac{\tilde{h}_k}{1-\tilde{h}_k} = \theta \frac{h_k}{1-h_k} \ , \ \theta \in (0, \infty). \tag{4.12}$$

Solving equation (4.12), $\tilde{h}_k$ has the functional form

$$\tilde{h}_k = \frac{\theta h_k}{1+(\theta-1)h_k} \ , \text{ for } k = 1,\ 2,\ \ldots,\ t, \tag{4.13}$$

where the odds ratio θ is treated as a specified parameter in the estimation of $\tilde{h}_k$. To simplify the notation, let $h_{k,\theta}$ denote the value of $\tilde{h}_k$ at a fixed value of θ, such that h_k defined in (4.5) can be viewed as a special case of $h_{k,\theta}$ under $\theta = 1$. Then, a corresponding covariance matrix for $\boldsymbol{h}_\theta = (h_{1\theta}, h_{2\theta}, \ldots, h_{t\theta})'$ can be estimated using the linear Taylor's series approximations (Koch et al., 1972) (see Appendix 6 for details). Replacing h_k's in (4.10) with $h_{k\theta}$ leads to the analogous estimates of q_k's in (4.14).

$$\begin{aligned} q_{k\theta} &= \frac{1}{n}\left[f_k + h_{k\theta} \sum_{g=1}^{k-1} \left(w_g \prod_{j=g}^{k-1} (1-h_{j\theta}) \right) + w_k h_{k\theta} \right] \\ q_{(t+1)\theta} &= 1 - \sum_{k=1}^{t} q_{k\theta} \end{aligned} \tag{4.14}$$

In order to construct the covariance matrix estimators for the failure probabilities estimated via imputing withdrawals' failure time intervals, we express q_k's and $q_{k\theta}$'s in matrix form as

$$\boldsymbol{q} = (q_1, q_2, \ldots, q_t, q_{t+1})' \text{ and } \boldsymbol{q}_\theta = (q_{1\theta}, q_{2\theta}, \ldots, q_{t\theta}, q_{(t+1)\theta})'.$$

The consistent estimators $\boldsymbol{V_q}$ and $\boldsymbol{V}_{\boldsymbol{q}_\theta}$ for their covariance matrices can then be obtained using the linear Taylor's series approximations. The details of the estimation

are described in Appendix 7.

In our formulation, θ is the sensitivity parameter to control the difference in the probability of experiencing an event between the withdrawal patients and the patients retained in the study. When $\theta = 1$, the estimators in (4.10) or (4.14) generate failure probabilities and the corresponding covariance estimates that are the same as the usual actuarial method. When $\theta >$ (or $<$) 1, the discontinued patients are assumed to have higher (or lower) probability of failure following their withdrawal compared to the remaining patients, thus the corresponding failure probability distribution ($q_{1\theta}$, $q_{2\theta}$, ..., $q_{t\theta}$, $q_{(t+1)\theta}$) can be estimated via (4.14). By this way, an imputed data set can be created to mimic the complete ITT analysis, in which the discontinued patients are followed off treatment until the end of the study. For withdrawals in the placebo group, the non-informative independent censoring assumption of $\theta_C = 1$ is reasonable in the sense that patients typically function under similar study conditions after dropout. However a $\theta_C > 1$ can be specified to address the possibility that the post-dropout experience of placebo patients is less favorable than with a MAR-like assumption, especially when an active drug is used for control. For the test treatment group, $\theta_T > \theta_C$ (> 1) can be specified so as to address larger departures from the independent censoring assumption. With $\theta_C = 1$, $\theta = (\theta_T/\theta_C) = \theta_T$ becomes a single parameter for calibrating sensitivity analyses, and therefore the sensitivity analyses can be carried out over a plausible range of θ that connects the specifications of patients' post-dropout behaviors and the response of interest from the observed data.

The proposed method should have reasonably good asymptotic properties in terms of the type I error of statistical tests and the coverage of confidence intervals when the sample size is sufficiently large (e.g., for each group, the number of withdrawals is at least 10 and the number of failures is at least 10 within each time interval, and there is at least 10 people who complete all the intervals without the event).

4.2.4 Criteria for treatment effect comparison

The treatment effect comparison between groups can be addressed through incidence density ratio (IDR), odds ratio (OR), Mann-Whitney probability, or Mantel-Haenszel test statistics calculated using the $q_{k\theta}$'s of the test and control groups. To establish notations, let $i = 1, 2$ index the test and control treatments, assuming n_i patients in group i. Let $k = 1, 2, \ldots, t$ index a set of time intervals, with t denoting the last time interval in the follow-up period. Let θ_i denote the sensitivity parameter in the ith group. Let $q_{ik\theta_i}$ denote the estimated probability that a patient with ith treatment will experience an event in the kth time interval via redistribution of the withdrawal counts. Let $\boldsymbol{q}_{i\theta_i} = (q_{i1\theta_i}, q_{i2\theta_i}, \ldots, q_{it\theta_i}, q_{i(t+1)\theta_i})'$; and let $\tilde{\boldsymbol{n}}_{i\theta_i} = n_i \boldsymbol{q}_{i\theta_i} = (\tilde{n}_{i1\theta_i}, \tilde{n}_{i2\theta_i}, \ldots, \tilde{n}_{it\theta_i}, \tilde{n}_{i(t+1)\theta_i})'$ be the redistributed numbers/counts of events within the time intervals.

Incidence density ratio and odds ratio

Incidence density (ID) is the number of events that occur in a time interval divided by the amount of person-time at risk for the same interval (Lavange et al., 1994; Tangen and Koch, 2000). For time-to-event data from the exponential distribution, ID is the maximum likelihood estimator (MLE) of the hazard rate. When grouped survival data are assumed to have different exponential distributions within individual time intervals (i.e., a piecewise exponential model), an ID can be calculated for each time interval separately to estimate the interval-specific hazard rate. In general, the exponential distribution assumption is not necessary to support ID as a useful descriptive statistic to compare treatment groups for categorical time-to-event data. The rationale of using odds ratio (OR) as a measure for the treatment effect comparison has been described in the section (4.2.3). The computation and interpretation of these two measurements share the similar principle, hence they are discussed here simultaneously.

For the purpose of simplicity, we assume that all the time intervals have equal length of 1 unit of time span and the occurrence of events is assessed at the end of each interval. The ID and the odds of having an event in the kth time interval for group i are computed as

$$\hat{\gamma}_{ik\theta_i} = \frac{n_i q_{ik\theta_i}}{\sum_{j=k}^{t+1} n_i q_{ij\theta_i}} = \frac{q_{ik\theta_i}}{\sum_{j=k}^{t+1} q_{ij\theta_i}} \tag{4.15}$$

and

$$\hat{\phi}_{ik\theta_i} = \frac{q_{ik\theta_i}}{1 - \sum_{j=1}^{k} q_{ij\theta_i}} = \frac{q_{ik\theta_i}}{\sum_{j=k+1}^{t+1} q_{ij\theta_i}} \tag{4.16}$$

respectively. One could then form the interval-specific $\log_e$ incidence density ratio (IDR) of the test ($i = 1$) to the control ($i = 2$) group

$$\hat{\eta}_k = \log_e \frac{\hat{\gamma}_{1k\theta_1}}{\hat{\gamma}_{2k\theta_2}} = \log_e q_{1k\theta_1} - \log_e(\sum_{j=k}^{t+1} q_{1j\theta_1}) - \log_e q_{2k\theta_2} + \log_e(\sum_{j=k}^{t+1} q_{2j\theta_2}) \tag{4.17}$$

as a relevant basis for the treatment comparison. Similarly, the interval-specific $\log_e$ OR of the test ($i = 1$) to the control ($i = 2$) group is

$$\hat{\psi}_k = \log_e \frac{\hat{\phi}_{1k\theta_1}}{\hat{\phi}_{2k\theta_2}} = \log_e q_{1k\theta_1} - \log_e(\sum_{j=k+1}^{t+1} q_{1j\theta_1}) - \log_e q_{2k\theta_2} + \log_e(\sum_{j=k+1}^{t+1} q_{2j\theta_2}) \tag{4.18}$$

Letting $\hat{\boldsymbol{\eta}}$=($\hat{\eta}_1$, $\hat{\eta}_2$, …, $\hat{\eta}_t$) and $\hat{\boldsymbol{\psi}}$=($\hat{\psi}_1$, $\hat{\psi}_2$, …, $\hat{\psi}_t$), the consistent variance-covariance estimators $\boldsymbol{V}_{\hat{\boldsymbol{\eta}}}$ and $\boldsymbol{V}_{\hat{\boldsymbol{\psi}}}$ can be obtained through linear Taylor's series approximation (see Appendix 8 for details). In this paradigm, if the test and the control groups have identical treatment effect, their corresponding $\hat{\eta}_k$ and $\hat{\psi}_k$ will approximately have expected value of zero, and therefore the departure of those quantities from the null represents the treatment difference between the two groups (Tangen and Koch, 2000).

Let vector $\boldsymbol{d}$ denote either $\hat{\boldsymbol{\eta}}$ or $\hat{\boldsymbol{\psi}}$. When the sample sizes for both groups are sufficiently large, the vector $\boldsymbol{d}$ has approximately a multivariate normal distribution with $\boldsymbol{V_d}$ as its essentially known covariance matrix. The variation in the elements of the vector $\boldsymbol{d}$ can be analyzed by fitting linear regression models of the form

$$E_A(\boldsymbol{d}) = \boldsymbol{X}\boldsymbol{\beta} \tag{4.19}$$

where $\boldsymbol{X}$ is a pre-specified design matrix with full rank for the model structure, $\boldsymbol{\beta}$ is the corresponding vector of unknown regression coefficients, and 'E_A' means asymptotic expectation. The weighted least squares asymptotic regression (Grizzle et al., 1969) is used to determine the estimator for $\boldsymbol{\beta}$ via

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V_d}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V_d}^{-1}\boldsymbol{d}, \tag{4.20}$$

and a consistent estimator for the covariance matrix of $\boldsymbol{b}$ is given by

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}'\boldsymbol{V_d}^{-1}\boldsymbol{X})^{-1} \tag{4.21}$$

If the data are adequately described by this model, a test of hypothesis H_0: $\boldsymbol{C\beta}$=0, can be performed with Wald statistics

$$\boldsymbol{Q_C} = \hat{\boldsymbol{\beta}}'\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}\boldsymbol{C}')^{-1}\boldsymbol{C}\hat{\boldsymbol{\beta}} \tag{4.22}$$

which has approximately a χ^2-distribution with d.f.=rank($\boldsymbol{C}$) in large samples under H_0. For the model with identity matrix $\boldsymbol{I}_{t\times t}$ as $\boldsymbol{X}$, an overall test of any treatment difference among the intervals can be obtained via $\boldsymbol{C} = \boldsymbol{I}_{t\times t}$. One can also use $\boldsymbol{C}$=$[-\boldsymbol{1}_{(t-1)\times 1}, \boldsymbol{I}_{(t-1)\times(t-1)}]$ to evaluate the homogeneity of IDRs (or ORs) across all of the intervals, i.e., the proportional ID (or odds) assumption. When the overall

treatment effect is of interest regardless of departures from homogeneity, $\boldsymbol{C}=[\mathbf{1}_{1\times t}]$ can be used to test whether the average $\log_e$ IDR$= \frac{1}{t}\sum_{i=1}^{t}\eta_i$ (or $\log_e$ OR$= \frac{1}{t}\sum_{i=1}^{t}\psi_i$) across the intervals equals zero. However, this test should be interpreted with caution, particularly when intervals have IDRs (ORs) with opposite direction. Alternatively, a common $\log_e$ IDR (or $\log_e$ OR) across intervals can be estimated as a single parameter β via the simplified model for which $\boldsymbol{X}=[\mathbf{1}_{t\times 1}]$.

Mann-Whitney probability

For two random variables X and Y on the same support, the Mann-Whitney probability, $Pr(Y \geq X)$ or $\xi = Pr(Y > X) + \frac{1}{2}Pr(Y = X)$, is a general measure of effect size to characterize the degree of separation of their distributions. In terms of the treatment difference in randomized clinical trials, ξ estimates the probability that a randomly selected patient receiving the new treatment has a better response than a randomly selected patient who receives the control treatment (Acion et al., 2006). This probability measurement is applicable to both continuous and ordinal data without requirement of assumptions on distributions, and at same time it maintains meaningful interpretation across a variety of outcome measures and sample distributions. Therefore, it has been advocated as an intuitive non-parametric approach to measure the size of effects (Acion et al., 2006; Newcombe, 2006).

For categorical time-to-event data, ξ is the probability that a patient randomly chosen from the test treatment ($i = 1$) group does not have the event of interest earlier than the patient randomly chosen from the control ($i = 2$) group, i.e. $Pr(T_1 \geq T_2)$, where T_i denotes the time to the event of interest in the ith group. In the present setting where the follow-up period consists of t distinctive time intervals, the Mann-Whitney

probability for the sensitivity analysis can be estimated as

$$\hat{\xi} = \sum_{k=1}^{t+1} q_{1k\theta_1}(\sum_{j=1}^{k-1} q_{2j\theta_2} + \frac{1}{2}q_{2k\theta_2}) = \sum_{k=1}^{t+1} q_{1k\theta_1}(\sum_{j=1}^{k} q_{2j\theta_2} - \frac{1}{2}q_{2k\theta_2}). \tag{4.23}$$

If the two treatments are equally effective, the chance of having a better response in the new treatment would be $\frac{1}{2}$, i.e. $\xi = \frac{1}{2}$. As the new treatment shows more benefit than the control, $\hat{\xi}$ moves towards 1, and as the new treatment shows less benefit, $\hat{\xi}$ moves towards 0. To perform the Wald hypothesis test of H_0: $\xi=\frac{1}{2}$ and calculate the confidence interval of $\hat{\xi}$, a consistent variance estimator of $\hat{\xi}$ can be constructed using linear Taylor's series approximation (see Appendix 9 for details).

Mantel-Haenszel criterion

Mantel (1966) extended the Mantel-Haenszel methodology (Mantel and Haenszel, 1959) to compare two sets of survival patterns. In order to implement the Mantel-Haenszel test, the imputed counts of all study intervals for two treatment groups (i.e., $\tilde{\boldsymbol{n}}_{1\theta_1}$ and $\tilde{\boldsymbol{n}}_{2\theta_2}$) are summarized in t (2×2) contingency tables, one for each time interval of the form

	Failed	Survived	Total
Group 1	$\tilde{n}_{1k}$	$\sum_{k'=(k+1)}^{(t+1)} \tilde{n}_{1k'}$	$\sum_{k'=k}^{(t+1)} \tilde{n}_{1k'}$
Group 2	$\tilde{n}_{2k}$	$\sum_{k'=(k+1)}^{(t+1)} \tilde{n}_{2k'}$	$\sum_{k'=k}^{(t+1)} \tilde{n}_{2k'}$

One can then compute the Mantel-Haenszel criterion for imputed data as

$$\text{MHE} = \left(\sum_{k=1}^{t} \tilde{n}_{1k} - \sum_{k=1}^{t} E(\tilde{n}_{1k})\right)^2 = \left(\sum_{k=1}^{t} \tilde{n}_{1k} - \sum_{k=1}^{t} \frac{(\tilde{n}_{1k} + \tilde{n}_{2k})\sum_{k'=k}^{(t+1)} \tilde{n}_{1k'}}{\sum_{k'=k}^{(t+1)}(\tilde{n}_{1k'} + \tilde{n}_{2k'})}\right)^2 \tag{4.24}$$

Under the null hypothesis of no treatment difference, the Mantel-Haenszel statistics Q_{MH} = MHE/Var(MHE) has approximately a χ^2 distribution with df of 1, which

allows simultaneous comparison of survival patterns as a whole. The calculation of the numerator and denominator of Q_{MH} are given in Appendix 10.

4.3 Application

The proposed sensitivity analysis method is illustrated for a maintenance trial to compare a test drug and an active control for prevention of duodenal ulcer recurrence (Elashoff and Koch, 1991). Endoscopic assessments for ulcer recurrence were made for patients with previously healed ulcer at protocol-scheduled visits of months 4, 8, and 12 without regard for symptoms. The final status of patients can be classified into one of the following three categories:(1) the patient had completed the study without recurrence; (2) the patient withdrew from further evaluation due to study discontinuation for some reasons (e.g. protocol violation, poor compliance, lost to follow-up); (3) the patient experienced a recurrence (i.e., treatment failure) during the follow-up period (Koch et al., 1984). The data from such a maintenance study are summarized in Table 4.1. An issue complicating the analysis is how to deal with the patients who discontinued the assigned treatment and withdrew from the study without having an observed ulcer recurrence. Those patients can cause ambiguity in the study conclusion because of their unknown recurrence status subsequent to their last endoscopic evaluation. As noted in Table 4.1, the overall withdrawal rate for the active control is 25.3%(61/241) and that for the test drug is 23.5%(57/243). In addition, the withdrawal rate in the first time interval is higher for the active control group (18.3% for the active control vs. 14.9% for the test drug), whereas the test drug group has a slightly higher withdrawal rates in the second and third intervals (5.0% and 2.1% for the active control vs. 5.8% and 2.9% for the test drug). With uneven withdrawal rates for the two treatment groups, multiple ways for managing these withdrawals need to be used to assess their impact on conclusions (Elashoff and Koch, 1991).

Table 4.1: Data for endoscopic assessment in 12-month maintenance trial for duodenal ulcer

	Number with recurrence first seen in interval			Number withdrawn from risk during interval			Number with no occurence	
Treatment	0-4 M	4-8 M	8-12 M	0-4 M	4-8 M	8-12 M	by 12 M	Total
Active control	40	24	6	44	12	5	110	241
Test drug	17	11	16	36	14	7	142	243

We first consider the performance of the proposed method under $\theta_C = \theta_T = 1$. With this specification, the withdrawal counts redistribution has the MAR-like assumption of non-informative independent censoring that is also specified in the usual actuarial method. For the purpose of comparisons, the data were also analyzed with the actuarial method and the crude rate approach. Table 4.2 summarizes the estimates of the recurrent probabilities for individual time intervals and the corresponding accumulative percentages with recurrence obtained by these three methods. The estimates for the interval-specific ($\log_e$) IDR and OR for the test drug over the active control are shown in Table 4.3 and 4.4, respectively, along with the corresponding inferences via the linear model with $\boldsymbol{X} = \boldsymbol{I}_{t\times t}$ as the design matrix. Table 4.5 presents the common ($\log_e$) IDR and OR for the test drug vs. the active control estimated by the linear model with $\boldsymbol{X} = \boldsymbol{1}_t$ as the design matrix. The treatment effect is also assessed by the Mann-Whitney probability and the Mantel-Haenszel criterion, and corresponding results are displayed in Table 4.6.

Except for the Mantel-Haenszel criterion, the proposed sensitivity analysis with the specification of $\theta_C = \theta_T = 1$ produces the exact same results as the actuarial method on the statistical estimates and inferences. The interval-specific IDRs and ORs for the test drug vs. the active control are not homogeneous across all three intervals (p-values for homogeneity < 0.05 for all three different managements of withdrawals). And more specifically, the test drug has smaller ulcer recurrence rates than the active

control for the first and second time intervals (i.e., the interval-specific $\log_e$ IDR and OR estimates < 0) with p-values < 0.05, but there is an inconclusive suggestion of more ulcer recurrence within the third interval (i.e., the interval-specific $\log_e$ IDR and OR estimates > 0) for which p-values ≈ 0.15. The common $\log_e$ IDR or OR estimator with the design matrix of $\boldsymbol{X} = \mathbf{1}_t$ is an inverse weighted estimator, and so is weighted roughly proportional to the number of events in the interval. As shown in Table 4.1, the three intervals have very different numbers of events, which averaging the interval-specific IDRs or ORs with equal weight would ignore. Therefore, using the common (average) estimators is considered more appropriate. With the corresponding results shown in Table 4.5, the common IDR of 0.52 (95% CI 0.36, 0.76) and the common OR of 0.49 (95% CI 0.32, 0.74) with corresponding p-values of 0.0008 suggest the superiority of the test treatment. For the actuarial method and the counts redistribution approach with $\theta_C = \theta_T = 1$, the Mann-Whitney probability (i.e., the probability that a patient randomly chosen from active control have the event of interest earlier than the patient randomly chosen from the test drug group) estimate is 0.58 (95% CI 0.54, 0.63) with p-value of 0.0003. Consistently, the Mantel-Haenszel tests from both methods also favor the test treatment, but the proposed sensitivity analysis under $\theta_C = \theta_T = 1$ produces somewhat less significant result than the actuarial method (Q_{MH}=10.9 with p-value of 0.001 for the proposed method vs. Q_{MH}=11.7 with p-value of 0.0006 for the actuarial method).

Results from the crude rate approach are also presented for comparisons. The crude interval-specific and cumulative rates for ulcer recurrence are smaller than their actuarial counterparts (Table 4.2). As noted in Table 4.3 and 4.4, the crude rate approach produces larger interval-specific IDR and OR estimates than the actuarial method for all three intervals, which give rise to the common $\log_e$ IDR and OR that

are closer to the null and have larger p-values (Table 4.5). Correspondingly, the p-values from the Wald test with Mann-Whitney probability and the Mantel-Haenszel test with the crude rate approach are also larger than those obtained from the actuarial method and the proposed method with the specification of $\theta_C = \theta_T = 1$ (Table 4.6).

We then consider the sensitivity analyses with separate sensitivity parameters for the active control and the test drug, i.e., θ_C and θ_T. One way to proceed is to specify $\theta_C > 1$ to address the probability that the post-withdrawal experience is less favorable than the primary assumption of non-informative independent censoring for control, and then specify $\theta = \theta_T/\theta_C > 1$ to address the larger departure from the primary assumption for the test drug. Therefore, there is a value of $\theta_T = \theta_C \times \theta$ for each specified θ_C. The choices of θ_C and θ can be arbitrary, such as 1, 1.5, 2, and 2.5, where the value of 2.5 for θ might represent a reasonably large difference in the post-withdrawal tendency of having ulcer recurrence, given the common OR estimate of having events for the active control vs. test drug was about 2.1 (=1/0.486) under $\theta_C = \theta_T = 1$ (and 0.486 is the estimated common OR in Table 4.5). The results from such a sensitivity analysis are summarized in Table 4.7. The treatment effect estimate is closer to the null as the value of θ_C and θ_T gets larger. Given a particular specification for θ_C and θ_T, the Wald test with Mann-Whitney probability produces the smallest p-values, whereas the Mantel-Haenszel test produces the largest p-value. For all the scenarios, the p-values are < 0.05, suggesting that the study conclusion is robust to the assumption of non-informative independent censoring.

Table 4.2: Interval-specific and cumulative rate for ulcer recurrence obtained from different managements of withdrawals (i.e., crude rate, life table, and sensitivity analysis with $\theta_C = \theta_T = 1$)

		Crude rate		Actuarial method or sensitivity analysis with $\theta_C = \theta_T = 1$	
Treatment	Interval	Interval-specific rate (SE)	Cumulative rate (SE)	Interval-specific rate (SE)	Cumulative rate (SE)
Active control	0-4M	0.166 (0.024)	0.166 (0.024)	0.203 (0.029)	0.203 (0.029)
	4-8M	0.100 (0.019)	0.266 (0.028)	0.132 (0.025)	0.335 (0.034)
	8-12M	0.025 (0.010)	0.291 (0.029)	0.034 (0.014)	0.369 (0.035)
Test drug	0-4M	0.070 (0.016)	0.070 (0.016)	0.082 (0.019)	0.082 (0.019)
	4-8M	0.045 (0.013)	0.115 (0.021)	0.057 (0.017)	0.140 (0.025)
	4-8M	0.066 (0.016)	0.181 (0.025)	0.087 (0.021)	0.227 (0.030)

Table 4.3: Interval-specific ($\log_e$) incidence density ratios and statistical inferences via the linear model with design matrix $\boldsymbol{X} = \boldsymbol{I}_{t\times t}$

Withdrawal Managements	Interval	$\log_e$ IDR (SE)	IDR (95% CI)	p-values	p-values for homogeneity
Crude rate	0-4M	-0.864 (0.275)	0.422 (0.246, 0.722)	0.0017	0.0055
	4-8M	-0.898 (0.351)	0.408 (0.205, 0.811)	0.0106	
	8-12M	0.786 (0.468)	2.19 (0.878, 5.49)	0.0928	
Actuarial method or	0-4M	-0.905 (0.272)	0.405 (0.237, 0.689)	0.0009	0.0070
sensitivity analysis	4-8M	-0.974 (0.346)	0.378 (0.192, 0.745)	0.0049	
with $\theta_C = \theta_T = 1$	4-8M	0.672 (0.463)	1.96 (0.790, 4.85)	0.1466	

Table 4.4: Interval-specific ($\log_e$) odds ratios and statistical inferences via the linear model with design matrix $\boldsymbol{X} = \boldsymbol{I}_{t\times t}$

Withdrawal Managements	Interval	$\log_e$ OR (SE)	OR (95% CI)	p-values	p-values for homogeneity
Crude rate	0-4M	-0.973 (0.305)	0.378 (0.208, 0.688)	0.0014	0.0042
	4-8M	-0.975 (0.378)	0.377 (0.180, 0.792)	0.0099	
	8-12M	0.829 (0.490)	2.29 (0.877, 5.99)	0.0906	
Actuarial method or	0-4M	-1.05 (0.309)	0.351 (0.192, 0.644)	0.0007	0.0051
Sensitivity analysis	4-8M	-1.09 (0.383)	0.336 (0.159, 0.712)	0.0044	
with $\theta_C = \theta_T = 1$	4-8M	0.726 (0.495)	2.07 (0.783, 5.45)	0.1430	

Table 4.5: Common ($\log_e$) incidence density ratios and odds ratios via the linear model with design matrix $\boldsymbol{X} = \mathbf{1}_t$

Withdrawal Managements	$\log_e$ IDR (SE)	IDR (95% CI)	p-values
Crude rate	-0.584 (0.196)	0.558 (0.380, 0.820)	0.0030
Actuarial method	-0.649 (0.194)	0.522 (0.357, 0.764)	0.0008
sensitivity analysis with $\theta_C = \theta_T = 1$	-0.649 (0.194)	0.522 (0.357, 0.764)	0.0008
	$\log_e$ OR (SE)	OR (95% CI)	
Crude rate	-0.631 (0.214)	0.532 (0.350, 0.809)	0.0032
Actuarial method	-0.722 (0.216)	0.486 (0.318, 0.742)	0.0008
Sensitivity analysis with $\theta_C = \theta_T = 1$	-0.722 (0.216)	0.486 (0.318, 0.742)	0.0008

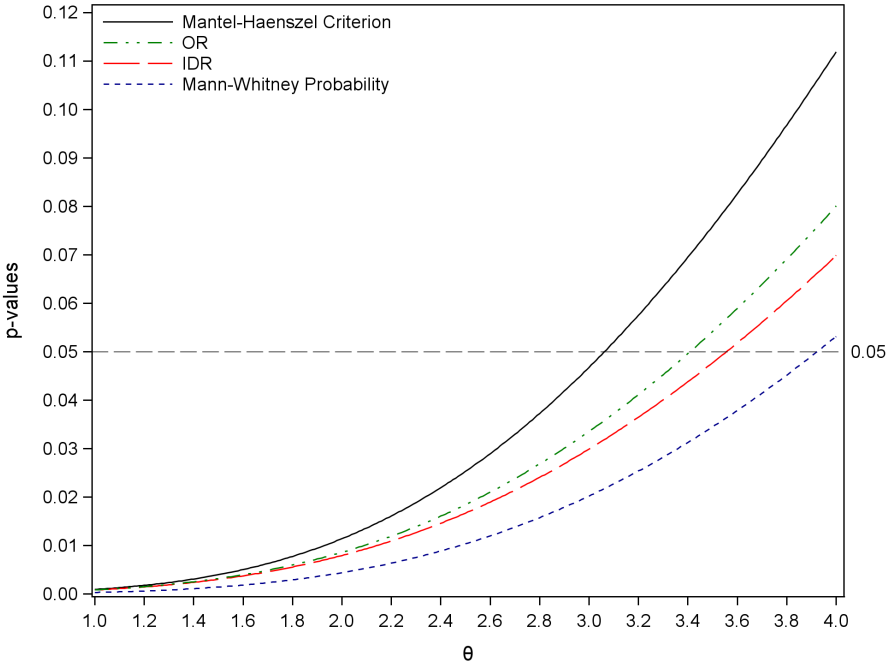
Table 4.6: Mann-Whitney probability and the Mantel-Haenszel criterion by different managements of withdrawals

	Mann-Whitney probability			Mantel-Haenszel test/criterion
Withdrawal Managements	Estimates (SE)	95% CI	p-values	Q_{MH}(p-values)
Crude rate	0.562 (0.0193)	(0.525, 0.600)	0.0012	8.97 (0.0027)
Actuarial method	0.584 (0.0233)	(0.538, 0.630)	0.0003	11.7 (0.0006)
Sensitivity analysis with $\theta_C = \theta_T = 1$	0.584 (0.0233)	(0.538, 0.630)	0.0003	10.9 (0.0010)

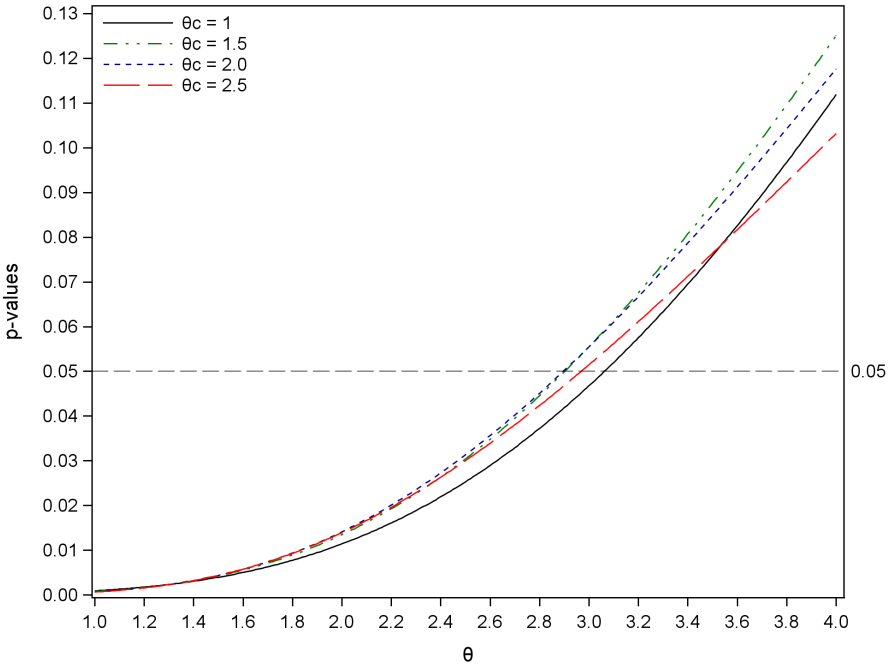
Table 4.7: Sensitivity analysis

θ_C	θ (θ_T/θ_C)	θ_T	$\log_e$ IDR		$\log_e$ OR		Mann-Whitney prob.		Mantel-Haenszel criterion
			$\log_e$ IDR (SE)	p-value	$\log_e$ OR (SE)	p-value	Est.(SE)	p-value	Q_{MH}(p-values)
1	1	1	-0.6493 (0.1941)	0.0008	-0.7222 (0.2164)	0.0008	0.5840 (0.233)	0.0003	10.9 (0.0010)
	1.5	1.5	-0.5727(0.1931)	0.0030	-0.6373 (0.2162)	0.0032	0.5762 (0.0239)	0.0014	8.29 (0.0040)
	2	2	-0.5093 (0.1919)	0.0080	-0.5663 (0.2156)	0.0086	0.5694 (0.0244)	0.0044	6.40 (0.0114)
	2.5	2.5	-0.4558 (0.1905)	0.0167	-0.5060 (0.2147)	0.0184	0.5635 (0.0248)	0.0104	5.00 (0.0253)
1.5	1	1.5	-0.6514 (0.1920)	0.0007	-0.7320 (0.2159)	0.0007	0.5898 (0.0244)	0.0002	11.3 (0.0008)
	1.5	2.25	-0.5601 (0.1900)	0.0032	-0.6297 (0.2149)	0.0034	0.5801 (0.0251)	0.0014	8.16 (0.0043)
	2	3	-0.4889 (0.1878)	0.0092	-0.5489 (0.2134)	0.0101	0.5719 (0.0256)	0.0049	6.11 (0.0135)
	2.5	3.75	-0.4320 (0.1855)	0.0199	-0.4838 (0.2118)	0.0223	0.5651 (0.0259)	0.0119	4.68 (0.0305)
2	1	2	-0.6459 (0.1895)	0.0007	-0.7321 (0.2146)	0.0006	0.5939 (0.0252)	0.0002	11.5 (0.0007)
	1.5	3	-0.5469 (0.1866)	0.0034	-0.6200 (0.2128)	0.0036	0.5827 (0.0258)	0.0013	8.11 (0.0044)
	2	4	-0.4735 (0.1834)	0.0098	-0.5359 (0.2105)	0.0109	0.5739 (0.0262)	0.0048	6.02 (0.0141)
	2.5	5	-4173 (0.1804)	0.0207	-0.4707 (0.2081)	0.0237	0.5667 (0.0265)	0.0117	4.63 (0.0314)
2.5	1	2.5	-0.6368 (0.18693)	0.0007	-0.7268 (0.2130)	0.0006	0.5966 (0.0256)	0.0002	11.6 (0.0007)
	1.5	3.75	-0.5343 (0.1830)	0.0035	-0.6099 (0.2103)	0.0037	0.5846 (0.0263)	0.0013	8.11 (0.0044)
	2	5	-0.4616 (0.1791)	0.0100	-0.5257 (0.2073)	0.0112	0.5755 (0.0266)	0.0046	6.04 (0.0141)
	2.5	6.25	-0.4079 (0.1756)	0.0202	-0.4629 (0.2043)	0.0235	0.5683 (0.0268)	0.0109	4.71 (0.0301)

Alternatively, the sensitivity analysis can proceed with varying θ in a range (L, U) for a given θ_C to assess how the treatment effect changes for different post-withdrawal tendencies of having the event for the test treatment group vs. the control. The values of $(1/U, 1/L)$ can be a range of odds ratios from previous related studies or the clinical judgement for comparison of effective medicine with control. For example, if the odds ratio for an active drug vs. placebo falls in the range of (0.5, 0.8), then one could consider $\theta_T = \theta_C \times \theta$ in the range of (1.25, 2) for the extent to which a test treatment patient with premature discontinuation has higher odds to have the event than such a patient without discontinuation for the time intervals following withdrawal. Here, we proceed with varying the value of θ from 1 to 4 by 0.01 increments for $\theta_C = 1$. In this regards, the odds ratio of $0.25 = 1/4$ (corresponding to $\theta = \theta_T = 4$) represents a large effect size for a clearly effective treatment vs. control that most test drugs might not exceed. And it is also useful to note that $\theta_C = \theta_T = 0$ corresponds to the crude rate approach and $\theta_C = \theta_T = \infty$ corresponds to the approach that view the withdrawals as having the event. The contour plot of the p-values for treatment comparisons is constructed as a function of the sensitivity parameter θ for each of the four criteria. As shown in Figure 4.1a, $p \leq 0.05$ applies with $\theta \leq 3.07$ for the Mantel-Haenszel test, $\theta \leq 3.41$ for the common OR inference, $\theta \leq 3.56$ for the common IDR inference, and $\theta \leq 3.93$ for the Mann-Whitney probability approach. Therefore, all four criteria suggest reasonably good robustness of the study conclusion to the assumption of non-informative independent censoring. Such a sensitivity analysis can be implemented for a set of θ_C separately. Figure 4.1b shows such a sensitivity analysis with the Mantel-Haenszel test as the criterion for treatment effect comparison. With the specification of θ_C=1.5 , 2, and 2.5, the sensitivity analyses suggest slightly weaker conclusions than that with $\theta_C = 1$, i.e., in order to have $p \leq 0.05$, $\theta \leq 2.91$ with $\theta_C = 1.5$, $\theta \leq 2.90$ with $\theta_C = 2$, and $\theta \leq 2.97$ with $\theta_C = 2.5$ are needed.



(a) Sensitivity analysis with the specification of $\theta_C = 1$



(b) Sensitivity analysis with the Mantel-Haenszel criterion

Figure 4.1: Contour plots of sensitivity analysis

4.4 Summary and discussion

In this article, we developed a sensitivity analysis method for grouped time-to-event data with possible informative censoring. The method explores the effect of departures from the non-informative independent assumption in terms of differences in the failure probability distributions of withdrawals and patients remained on their assigned treatment. A conditional odds ratio (θ) of failure for the discontinued vs. retained patients is incorporated as the sensitivity parameter to specify various post-withdrawal hypotheses for the tendency of having the event of interest. The hypothetical survivor profiles are estimated by redistributing the withdrawal counts to the failure counts in the time intervals following their last visits or to the counts censored at the end of study, as if the missing failure times were imputed for the withdrawals. The treatment effect and the corresponding covariance estimates have closed analytical forms, therefore there is no need to perform the multiple imputation procedures for the missing outcomes (i.e., probabilistically assign the patients to a failure status in the time intervals following their withdrawals). Under the specification of $\theta_C = \theta_T = 1$, the proposed method produces the same estimates as the actuarial method, while being naturally comparable to the crude rate estimates that have all randomized patients as the denominator. Hence, the interpretation of the sensitive analysis results is straightforward for the non-statisticians and clinical reviewers.

We also presented methods for comparing treatment effects that are applicable in the framework of the proposed sensitivity analysis. To extend the Cox model to the discrete time-to-event data, the interval-specific odds ratios (ORs) are analyzed through the weighted least squares (WLS) asymptotic regression. And the counterpart WLS regression on the incidence density ratios (IDRs) is a direct application of the piecewise exponential model. One advantage of such a piecewise model is that the piecewise exponential (i.e., constant) hazards approximate reasonably well almost any shape of

nonparametric baseline hazards. The common $\log_e$ OR or IDR, produced by the linear model with single coefficient, has the interpretation of a population average treatment effect. However, an equal weighted average of the interval-specific $\log_e$ ORs or IDRs may be difficult to interpret when the assumption of proportional odds or incidence densities is not appropriate. In contrast, the non-parametric methods provide valid inferences regardless of departures from the homogeneity of ORs or IDRs. The Mantel-Haenszel test for grouped data is equivalent to the logrank test for comparing survival curves for ungrouped data (Koch et al., 1985), whereas the Mann-Wittney probability (ξ) is related to the Wilcoxon (rank sum) test. The Wald hypothesis test of H_0: $\xi=\frac{1}{2}$ is more able to detect the early treatment differences than the logrank test, because it receives relatively more weight than the logrank test for early failures and relatively less weight for later failures.

A challenging issue in the design of clinical trials is how to account for the effects of missing outcomes (i.e., withdrawals) on inferences for the treatment comparison. The power calculation should accommodate the loss of statistical power due to the reduction of information, and more importantly, the biased estimation of the treatment effect when MAR assumption is in question. Often, those considerations cannot be addressed analytically, but through relatively involved simulation studies. With the proposed sensitivity analysis, the implication of departures from the MAR-like assumption to the statistical power and the sample size can be assessed analytically through various specifications for the sensitivity parameter and the interval-specific failure and withdrawal rates. Hence, this could be an additional feature that makes our method desirable to employ in the regulatory environment.

Chapter 5

Discussion

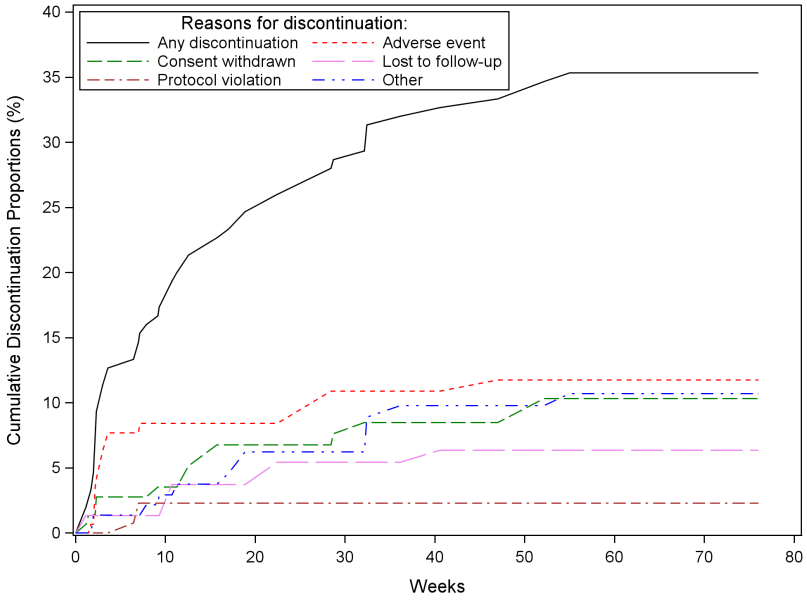
The intention-to-treat (ITT) principle was espoused by the Food and Drug Administration (FDA) and International Conference on Harmonization (ICH) for unbiased assessments of the efficacy/effectiveness for new therapies. The essence of the ITT principle is to evaluate all patients randomized into a study as scheduled and include all patients in the final analysis as randomized, even if they stop their assigned treatments prematurely (Lachin, 2000). However, missing outcomes due to premature withdrawals often create difficulties for implementing this principle. Even when the MAR assumption is appropriate, the usual MAR approaches provide effectively per-protocol (PP) analyses for the ITT population, because they attempt to estimate what would have happened for outcomes if all patients had been staying on their assigned treatment. In contrast, the ITT principle requires the missing data analysis strategies to focus on what would have happened if the patients had been followed in the absence of treatment. In harmony with this principle, we developed a series of sensitivity analyses for time-to-event data (continuous and grouped). The proposed methods assess the implications of departures from the non-informative independent censoring assumption that is often specified in the primary analysis. A major feature of our approaches is that they anchor on the primary missing data assumption and directly address the sensitivity of

results to the primary missing data assumption by specifying different post-withdrawal experiences of having the event of interest. The sensitivity parameters for calibrating these specifications are standard criteria for treatment comparisons. Consequently, the results from such sensitivity analyses are more informative for clinical judgement with regard to the robustness of the study conclusion from the primary analysis.

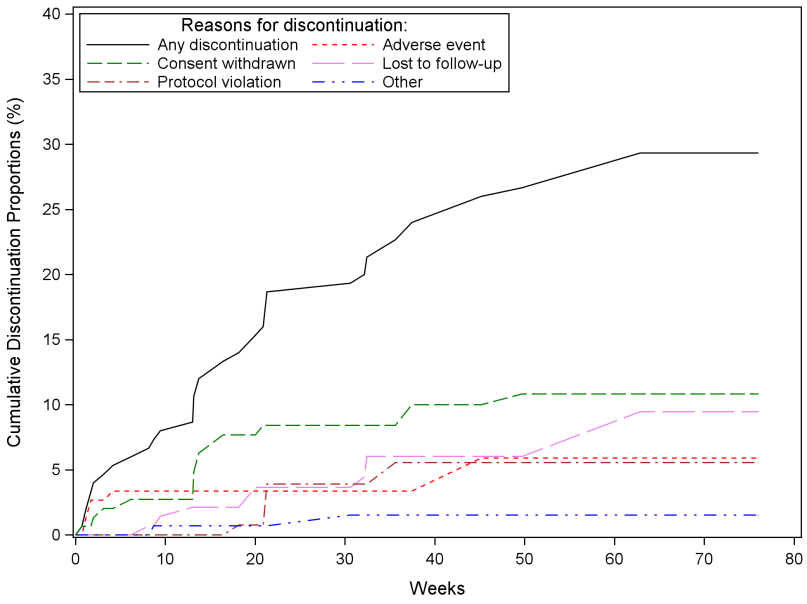
Furthermore, our approaches for sensitivity analyses can be extended to multiple failure time data. For a clinical trial example in cardiovascular diseases, the primary endpoint could be time to MI or all-cause death, whichever occurs first, and patients with a MI could have continued follow-up to the end of the study for mortality. In addition to the composite primary endpoint, investigators also wish to perform separate analysis on time-to-death and time-to-MI. In this situation, there are essentially 3 types of censoring: (1) patients withdrew without any event (censored for both outcomes); or (2) patients had a MI, then withdrew without being able to observe the event of death (censored for the time-to-death); or (3) patients died without MI (censored only for the time-to-MI). For the patients with type (1) censoring, both time-to-MI and time-to-death need to be imputed. For the patients with type (2) censoring, only the time-to-death needs to be imputed. And for patients with type (3) censoring, one could consider the death and the MI occurred at the same time or could impute the time-to-MI as if that patient might have lived to the end of the study. This principle is also applicable for the grouped multivariate survival data.

Appendix 1:

Cumulative discontinuation proportions by reasons



(a) Placebo group



(b) Test treatment group

Figure A.1: Cumulative discontinuation proportions by documented reasons

Appendix 2:

KMMI and PHMI methods with bootstrap resampling

Table A.1: KMMI and PHMI methods with or without bootstrap resampling at $\theta = 1$

Analysis method	Semi-parametric analysis (Cox PH model)			
	Coefficient (Std Err)	P value	Log-rank[1]	Wilcoxon[a]
1. $L = 50$				
KMMI without bootstrap	-0.322 (0.160)	0.0436	0.0451	0.0109
KMMI with bootstrap	-0.325 (0.180)	0.0727	0.0770	0.0167
PHMI without bootstrap	-0.388 (0.158)	0.0143	0.0145	0.0053
PHMI with bootstrap	-0.389 (0.165)	0.0191	0.0204	0.0068
2. $L = 100$				
KMMI without bootstrap	-0.328 (0.159)	0.0394	0.0410	0.0101
KMMI with bootstrap	-0.347 (0.178)	0.0519	0.0550	0.0127
PHMI without bootstrap	-0.394 (0.158)	0.0124	0.0126	0.0048
PHMI with bootstrap	-0.399 (0.164)	0.0150	0.0157	0.0056
3. $L = 500$				
KMMI without bootstrap	-0.332 (0.156)	0.0332	0.0345	0.0091
KMMI with bootstrap	-0.336 (0.170)	0.0480	0.0507	0.0114
PHMI without bootstrap	-0.398 (0.156)	0.0106	0.0108	0.0044
PHMI with bootstrap	-0.396 (0.160)	0.0134	0.0140	0.0053

Appendix 3:

Alternative KMMI strategy for sensitivity analysis

An alternative way to perform sensitivity analysis is to use the information in the placebo group to impute times to event for both treatment groups using the KMMI approach. To generate one set of the L imputed data, one could first impute failure time for discontinued patients in the placebo group under certain specification of θ_P; the KM estimates obtained from those complete data in the placebo group are then used to perform imputation through (1) - (7) for the discontinued patients in the test treatment group. In this MI procedure, the sensitivity parameter θ only needs to be specified for the placebo group. Besides choosing $\theta_p = 1$ to approximate a MAR-like assumption, $\theta_p > 1$ can be used to address the possibility that the post-discontinuation experience is less favorable than the patients remaining on their assigned treatment. The results of sensitivity analysis at $L = 50$ under various specifications of θ_P are shown in Table A.2. For this particular example, The estimated treatment effect when $\theta_p = 1$ is slightly weaker than those from applying the KMMI method within individual treatment groups with $\theta = 1$ for both (Table 3.2A). As the value of θ_p increases, results in favor of the test treatment become stronger, because the placebo group has more prematurely discontinued patients than the test treatment group, and thereby $\theta_p > 1$ penalizes the placebo group more.

Table A.2: Alternative KMMI strategy for sensitivity analysis

θ_P	Semi-parametric analysis (Cox PH model)					
	Coefficient	Std Err	HR (95% CI)	P value	Log-rank[2]	Wilcoxon[a]
1	-0.315	0.155	0.730 (0.538, 0.990)	0.0430	0.0440	0.0131
1.1	-0.332	0.154	0.717 (0.530, 0.971)	0.0313	0.0322	0.0097
1.2	-0.346	0.153	0.708 (0.524, 0.956)	0.0241	0.0248	0.0075
1.3	-0.361	0.150	0.697 (0.519, 0.935)	0.0160	0.0164	0.0053
1.4	-0.368	0.150	0.692 (0.516, 0.928)	0.0140	0.0143	0.0043
1.5	-0.384	0.151	0.681 (0.507, 0.916)	0.0110	0.0113	0.0033

Appendix 4:

Conditional probability of failing ($\boldsymbol{h}$)

We construct a $(2t+1)\times 1$ vector $\boldsymbol{a}$ corresponding to the categorical data structure in section 4.2.1,

$$\boldsymbol{a} = \frac{1}{n}(\boldsymbol{f}', \boldsymbol{w}')' = \frac{1}{n}(f_1, f_2, \ldots, f_t, w_1, w_2, \ldots, w_t, w_{t+1})',$$

with $n = \sum_{k=1}^{t} f_k + \sum_{k=1}^{t+1} w_k$. And we assume that the vector $n\boldsymbol{a}$ follows a multinomial distribution with

$$E(n\boldsymbol{a}) = n\boldsymbol{\pi} \text{ and } \mathrm{Var}(n\boldsymbol{a}) = n[\boldsymbol{D_\pi} - \boldsymbol{\pi\pi'}],$$

where $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{1t}, \pi_{21}, \ldots, \pi_{2t}, \pi_{2(t+1)})'$ and $\boldsymbol{D_\pi}$ is a $(2t+1)\times(2t+1)$ diagonal matrix with the elements of the vector $\boldsymbol{\pi}$ on the main diagonal. As a result, $\mathrm{Var}(\boldsymbol{a})$ can be consistently estimated via

$$\boldsymbol{V_a} = \frac{1}{n}[\boldsymbol{D_a} - \boldsymbol{aa'}] \tag{A.1}$$

where $\boldsymbol{D_a}$ is a $(2t+1)\times(2t+1)$ diagonal matrix with the elements of the vector $\boldsymbol{a}$ on the main diagonal.

We then formulate $\boldsymbol{h} = (h_1, h_2, \ldots, h_t)'$, where $h_k = f_k/n_k$ and $n_k = \sum_{j=k}^{t} f_j + \sum_{j=k+1}^{t+1} w_j$ for $k = 1, 2, \ldots, t$, in the form of compound functions of the vector $\boldsymbol{a}$ as

$$\boldsymbol{h} = \exp[\boldsymbol{A_2}\log_e(\boldsymbol{A_1 a})], \tag{A.2}$$

where $\log_e(\ )$ denotes the element-wise vector operation that transforms a vector to the corresponding vector of natural logarithms, and $\exp[\ ]$ denotes the element-wise vector operation that transforms a vector to the corresponding vector of exponentiated values, and matrices $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$, shown in (A.3) and (A.4) respectively, are matrices for linear

transformations.

$$\boldsymbol{A_1} = \begin{bmatrix} \boldsymbol{I}_t & \boldsymbol{0}_{t,(t+1)} \\ \boldsymbol{T}_t & \boldsymbol{0}_t \;\; \boldsymbol{T}_t \end{bmatrix} \tag{A.3}$$

$$\boldsymbol{A_2} = \begin{bmatrix} \boldsymbol{I}_t & -\boldsymbol{I}_t \end{bmatrix} \tag{A.4}$$

Here, $\boldsymbol{I}_t$ denotes an identity matrix, $\boldsymbol{0}$ represents a vector or a matrix with all elements equal to 0, and $\boldsymbol{T}_t$ is a $t \times t$ upper triangular matrix with all non-zero elements equal to 1.

$$\boldsymbol{T}_t = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 0 & 1 & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \tag{A.5}$$

The top half of $\boldsymbol{A_1}$ generates the numerators of the h_k (i.e. f_k) and the bottom half of $\boldsymbol{A_1}$ generates the denominators of the h_k (i.e. n_k), then $\boldsymbol{A_2} \log_e(\ )$ generates the $\log_e(h_k)$, and finally $\exp[\ ]$ generates the h_k.

We apply the first order linear Taylor series approximation to obtain the corresponding covariance matrix estimate. The first partial derivatives matrix $\boldsymbol{H_1}$ of $\boldsymbol{h}$ with respect to $\boldsymbol{a}$ can be written as the product of the first derivative matrices of sequential operations in accordance with the chain rule,

$$\boldsymbol{H_1} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{a}} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{a_3}} \frac{\partial \boldsymbol{a_3}}{\partial \boldsymbol{a_2}} \frac{\partial \boldsymbol{a_2}}{\partial \boldsymbol{a_1}} \frac{\partial \boldsymbol{a_1}}{\partial \boldsymbol{a}} = \boldsymbol{D_h} \boldsymbol{A_2} \boldsymbol{D_1}^{-1} \boldsymbol{A_1} \tag{A.6}$$

with

$$\boldsymbol{a_1} = \boldsymbol{A_1 a},\ \boldsymbol{a_2} = \log_e(\boldsymbol{a_1}),\ \boldsymbol{a_3} = \boldsymbol{A_2 a_2},\ \text{and}\ \boldsymbol{h} = \exp(\boldsymbol{a_3})$$
$$\frac{\partial \boldsymbol{a_1}}{\partial \boldsymbol{a}} = \boldsymbol{A_1},\ \frac{\partial \boldsymbol{a_2}}{\partial \boldsymbol{a_1}} = \boldsymbol{D_1}^{-1},\ \frac{\partial \boldsymbol{a_3}}{\partial \boldsymbol{a_2}} = \boldsymbol{A_2},\ \text{and}\ \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{a_3}} = \boldsymbol{D_h},$$

where $\boldsymbol{D_1}^{-1}$ is a diagonal matrix with the reciprocals of the elements of the $(2t+1)\times 1$ vector $\boldsymbol{a_1}$ on the main diagonal and $\boldsymbol{D_h}$ is a diagonal matrix with the elements of $\boldsymbol{h}$ on the main diagonal. It then follows from the method discussed in Koch et al. (1972) and Stokes et al. (2012) that

$$\boldsymbol{V_h} = \boldsymbol{H_1 V_a H_1'} = \boldsymbol{D_h A_2 D_1}^{-1} \boldsymbol{A_1 V_a A_1}' \boldsymbol{D_1}^{-1} \boldsymbol{A_2}' \boldsymbol{D_h} \tag{A.7}$$

is a consistent estimator for the covariance matrix of $\boldsymbol{h}$. With standard vector and matrix operations, one can show that (A.7) is a diagonal matrix with $h_k(1-h_k)/n_k$ on the kth element. The following paragraph offers the detailed derivation.

Define vector $\boldsymbol{n} = (n_1, n_2, \ldots, n_t)'$ containing the number of patients at risk for each of the t intervals. It is easily noticed that

$$\boldsymbol{a_1} = \boldsymbol{A_1 a} = \frac{1}{n}(\boldsymbol{f}', \boldsymbol{n}')' = (\boldsymbol{b}', \boldsymbol{m}')' \text{ with } \boldsymbol{V_{a_1}} = \boldsymbol{A_1 V_a A_1'} \tag{A.8}$$

By calculating the following variance and covariance estimates

$$\begin{aligned}
\mathrm{Var}(b_k) &\hat{=} b_k(1-b_k)/n \\
\mathrm{Var}(m_k) &\hat{=} m_k(1-m_k)/n \\
\mathrm{Cov}(b_k, m'_k) &\hat{=} b_k(1-m'_k)/n \quad \text{for } k' \leqslant k \\
&\hat{=} -b_k m'_k/n \qquad \text{for } k' > k \\
\mathrm{Cov}(m_k, m_{k'}) &\hat{=} m_{k'}(1-m_k)/n \text{ for } k' > k \\
\mathrm{Cov}(b_k, b_{k'}) &\hat{=} -b_k b_{k'}/n \text{ for } k' > k \ ,
\end{aligned} \tag{A.9}$$

we can identify the covariance structure of $\boldsymbol{V_{a_1}}$ being

$$\boldsymbol{V_{a_1}} = \frac{1}{n}\begin{bmatrix} [\boldsymbol{D_b} - \boldsymbol{b}\boldsymbol{b}'] & [\boldsymbol{D_b}\boldsymbol{T_t}' - \boldsymbol{b}\boldsymbol{m}'] \\ [\boldsymbol{T_t}\boldsymbol{D_b} - \boldsymbol{m}\boldsymbol{b}'] & [\boldsymbol{T_t}\boldsymbol{D_m} + \boldsymbol{D_m}\boldsymbol{T_t}' - \boldsymbol{D_m} - \boldsymbol{m}\boldsymbol{m}'] \end{bmatrix} \tag{A.10}$$

where $\boldsymbol{T_t}$ was shown in (A.5), $\boldsymbol{D_b}$ and $\boldsymbol{D_m}$ are diagonal matrices with the elements of $\boldsymbol{b}$ and $\boldsymbol{m}$ on the main diagonal, respectively. Next, we determine that the covariance structure of $\boldsymbol{V_{a_2}}$ is

$$\begin{aligned}
\boldsymbol{V_{a_2}} &= \boldsymbol{D_1}^{-1}\boldsymbol{V_{a_1}}\boldsymbol{D_1}^{-1} \\
&= \begin{bmatrix} \boldsymbol{D_b}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D_m}^{-1} \end{bmatrix} \boldsymbol{V_{a_1}} \begin{bmatrix} \boldsymbol{D_b}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D_m}^{-1} \end{bmatrix} \\
&= \frac{1}{n}\begin{bmatrix} \boldsymbol{D_b}^{-1} & \boldsymbol{T_t}'\boldsymbol{D_m}^{-1} \\ \boldsymbol{D_m}^{-1}\boldsymbol{T_t} & [\boldsymbol{D_m}^{-1}\boldsymbol{T_t} + \boldsymbol{T_t}'\boldsymbol{D_m}^{-1} - \boldsymbol{D_m}^{-1}] \end{bmatrix} - \mathbf{1}_{2t}\mathbf{1}'_{2t} \ ,
\end{aligned} \tag{A.11}$$

where $\mathbf{0}$ represents a $(t \times t)$ matrix with all elements equal to 0, and $\mathbf{1}_{2t}$ represents a $(2t \times 1)$ vector with all elements equal to 1. After some matrix algebra, it then follows

that

$$\boldsymbol{V_{a_3}} = \boldsymbol{A_2 V_{a_2} A_2}' = \frac{1}{n}[\boldsymbol{D_d}^{-1} - \boldsymbol{D_m}^{-1}] \ . \tag{A.12}$$

Since $\boldsymbol{D_h} = \boldsymbol{D_d D_m}^{-1} = \boldsymbol{D_m}^{-1}\boldsymbol{D_d}$, $\boldsymbol{V_h}$ can be expressed as

$$\begin{aligned} \boldsymbol{V_h} &= \boldsymbol{D_h V_{a_3} D_h} = \frac{1}{n}\boldsymbol{D_m}^{-1}\boldsymbol{D_d}[\boldsymbol{D_d}^{-1} - \boldsymbol{D_m}^{-1}]\boldsymbol{D_m}^{-1}\boldsymbol{D_d} \\ &= \frac{1}{n}\boldsymbol{D_m}^{-1}[\boldsymbol{D_h} - \boldsymbol{D_h D_h}] = \boldsymbol{D_n}^{-1}[\boldsymbol{D_h} - \boldsymbol{D_h D_h}]. \end{aligned} \tag{A.13}$$

Thus, we show that

$$\boldsymbol{V_h} = \text{Diag}\{\frac{h_k(1-h_k)}{n_k}\}. \tag{A.14}$$

Appendix 5:

Variance/covariance estimate for $\boldsymbol{p}$

The life table estimates for the failure probability distribution $\boldsymbol{p} = (p_1, p_2, \ldots, p_t)'$ can be expressed as the compound function of vector $\boldsymbol{h}$ shown in A.15.

$$\boldsymbol{p} = \exp[\boldsymbol{B_2} \log_e(\boldsymbol{B_1 h} + \boldsymbol{C_1})] \tag{A.15}$$

The $\boldsymbol{B_1}$, $\boldsymbol{B_2}$, and $\boldsymbol{C_1}$ are defined in A.16, A.17, and A.18, respectively, where $\boldsymbol{L_1}$ is a $t \times t$ lower triangular matrix with all non-zero elements equal to 1.

$$\boldsymbol{B_1} = \begin{bmatrix} \boldsymbol{I}_t \\ -\boldsymbol{I}_t \end{bmatrix} \tag{A.16}$$

$$\boldsymbol{B_2} = \begin{bmatrix} \boldsymbol{I}_t & (\boldsymbol{T}_t' - \boldsymbol{I}_t) \end{bmatrix} \tag{A.17}$$

$$\boldsymbol{C_1} = \begin{bmatrix} \boldsymbol{0}_t \\ \boldsymbol{1}_t \end{bmatrix} \tag{A.18}$$

With the chain rule for matrix differentiation of compound vector functions, a consistent covariance estimator $\boldsymbol{V_p}$ can be obtained as shown in A.19, where $\boldsymbol{D_2}^{-1}$ is a diagonal matrix with the reciprocals of the elements of the $2t \times 1$ vector $\boldsymbol{B_1 h} + \boldsymbol{C_1} = [\boldsymbol{h}', (\boldsymbol{1}_t - \boldsymbol{h}')]'$ on the main diagonal and $\boldsymbol{D_p}$ is a diagonal matrix with the elements of the vector $\boldsymbol{p}$ on the main diagonal.

$$\boldsymbol{V_p} = \boldsymbol{D_p B_2 D_2}^{-1} \boldsymbol{B_1 V_h B_1}' \boldsymbol{D_2}^{-1} \boldsymbol{B_2}' \boldsymbol{D_p} \tag{A.19}$$

Appendix 6:

Variance/covariance estimate for $\boldsymbol{h}_\theta$

Based on the relationship established in (4.13), $\boldsymbol{h}_\theta = (h_{1\theta}, h_{2\theta}, \ldots, h_{t\theta})'$ can be generated via the compound exponential and log-linear transformation of the vector $\boldsymbol{h}$

$$\boldsymbol{h}_\theta = \exp[\boldsymbol{A_2} \log_e(\boldsymbol{A_3 h} + \boldsymbol{C_2})], \tag{A.20}$$

with $\boldsymbol{A_2} = [\boldsymbol{I}_t - \boldsymbol{I}_t]$ as in (A.4), and $\boldsymbol{A_3}$ and $\boldsymbol{C_2}$ being defined by (A.21) and (A.22), respectively,

$$\boldsymbol{A_3} = \begin{bmatrix} \theta \\ (\theta - 1) \end{bmatrix} \otimes \boldsymbol{I}_t \tag{A.21}$$

$$\boldsymbol{C_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \otimes \boldsymbol{1}_t \tag{A.22}$$

where $\otimes$ is the right Kronecker product matrix operation such that the matrix on the right multiplies each element of the matrix on the left. From (A.20), one can easily note that the vector $\boldsymbol{h}_\theta$ is formulated in terms of a sequential series of matrix operations on $\boldsymbol{h}$ that is analogous to those for $\boldsymbol{h}$ from $\boldsymbol{a}$ in (A.2). Thus, by the manner described in (A.6), the covariance matrix $\boldsymbol{V}_{\boldsymbol{h}_\theta}$ for $\boldsymbol{h}_\theta$ may be estimated via

$$\boldsymbol{V}_{\boldsymbol{h}_\theta} = \boldsymbol{H_2 V_h H_2'} \text{ with } \boldsymbol{H_2} = \boldsymbol{D_4 A_2 D_3^{-1} A_3} \tag{A.23}$$

where $\boldsymbol{D_3}^{-1}$ is a diagonal matrix with the reciprocals of the elements of the $2t \times 1$ vector $\boldsymbol{A_3 h} + \boldsymbol{C_2} = [\theta \boldsymbol{h}', (\theta - 1)\boldsymbol{h}' + \boldsymbol{1}_t]'$ on the main diagonal and $\boldsymbol{D_4}$ is a diagonal matrix

with the elements of the vector $\boldsymbol{h}_\theta$ on the main diagonal. Moreover, substituting (A.2) into (A.20), the vector $\boldsymbol{h}_\theta$ can be expressed in terms of the observed categorical data counts contained in $\boldsymbol{a}$. As a result, $\boldsymbol{V}_{\boldsymbol{h}_\theta}$ is derived directly from $\boldsymbol{V}_{\boldsymbol{a}}$ as

$$\boldsymbol{V}_{\boldsymbol{h}_\theta} = \boldsymbol{H_3 V_a H_3'} \text{ with } \boldsymbol{H_3} = \boldsymbol{H_2 H_1}. \tag{A.24}$$

Alternatively, for any particular time interval k, one could view $h_{k\theta}$ as a univariate function of h_k . Applying the Taylor series approximation on the function (4.13), we easily show that

$$\text{Var}(h_{k\theta}) = \frac{\theta^2 h_k(1-h_k)}{[(\theta-1)h_k+1]^4 n_k} = \frac{h_{k\theta}^2(1-h_{k\theta})^2}{n_k h_k(1-h_k)}, \tag{A.25}$$

using the following equalities

$$(\theta-1)h_k + 1 = \frac{1-h_k}{1-h_{k\theta}} = \frac{\theta h_k}{h_{k\theta}}. \tag{A.26}$$

Finally, we note that the covariance matrix estimate for $\boldsymbol{h_\theta}$ is a diagonal matrix,

$$\boldsymbol{V}_{\boldsymbol{h_\theta}} = \text{Diag}\{\frac{h_{k\theta}^2(1-h_{k\theta})^2}{n_k h_k(1-h_k)}\}, \tag{A.27}$$

since the covariance matrix estimate for $\boldsymbol{h}$ is also diagonal.

Appendix 7:

Variance/covariance estimates for $\boldsymbol{q}$ and $\boldsymbol{q}_\theta$

The formulation of q_k and $q_{k\theta}$, for $k = 1, 2, \ldots, t, t{+}1$, involves the same functional structure, except that the h_k in (4.10) for q_k was replaced with the $h_{k\theta}$ in (4.14) for $q_{k\theta}$. Hence, we construct two concatenated vectors with similar structures as

$$\boldsymbol{g} = \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{h} \\ (\mathbf{1}_t - \boldsymbol{h}) \end{bmatrix} \text{ and } \boldsymbol{g}_\theta = \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{h}_\theta \\ (\mathbf{1}_t - \boldsymbol{h}_\theta) \end{bmatrix}, \tag{A.28}$$

so that the vectors $\boldsymbol{q} = (q_1, q_2, \ldots, q_t, q_{(t+1)})'$ and $\boldsymbol{q}_\theta = (q_{1\theta}, q_{2\theta}, \ldots, q_{t\theta}, q_{(t+1)\theta})'$ can be expressed in the compound functions of $\boldsymbol{g}$ and $\boldsymbol{g}_\theta$, respectively, via an identical series of matrix operations. As a consequence, the estimated covariance matrices of $\boldsymbol{q}$ and $\boldsymbol{q}_\theta$ may be computed in the same manner, by applying the linear Taylor series approximation. Here, we only describe the estimation for the covariance matrix of $\boldsymbol{q}$ in detail, and then the covariance matrix of $\boldsymbol{q}_\theta$ can be consistently estimated in a similar fashion.

On the basis of (A.6), the first partial derivative matrix of the concatenated vector $\boldsymbol{g}$ with respect to $\boldsymbol{a}$ can be written as

$$\boldsymbol{H_4} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{a}} = \begin{bmatrix} & \boldsymbol{I}_{(2t+1)} \\ & \boldsymbol{H_1} \\ - & \boldsymbol{H_1} \end{bmatrix} \tag{A.29}$$

It then follows that the consistent estimate of the covariance matrix for $\boldsymbol{g}$ takes the

form of (A.30).

$$\begin{aligned} \boldsymbol{V_g} &= \boldsymbol{H_4 V_a H_4'} \\ &= \begin{bmatrix} \boldsymbol{V_a} & \boldsymbol{V_a H_1'} & \boldsymbol{-V_a H_1'} \\ \boldsymbol{H_1 V_a} & \boldsymbol{H_1 V_a H_1'} & \boldsymbol{-H_1 V_a H_1'} \\ \boldsymbol{-H_1 V_a} & \boldsymbol{-H_1 V_a H_1'} & \boldsymbol{H_1 V_a H_1'} \end{bmatrix} \end{aligned} \tag{A.30}$$

By suitable choices of matrices $\boldsymbol{A_4}$ and $\boldsymbol{A_5}$ according to the specific value of t (i.e., the number of follow-up time intervals), we can express the vector $\boldsymbol{q}^* = (q_1, q_2, \ldots, q_t)'$ as

$$\boldsymbol{q}^* = \boldsymbol{A_5} \exp(\boldsymbol{A_4} \log_e \boldsymbol{g}). \tag{A.31}$$

Thus, the linear Taylor series based covariance estimate ($\boldsymbol{V_{q^*}}$) is

$$\boldsymbol{V_{q^*}} = \boldsymbol{H_5 V_g H_5'} \text{ with } \boldsymbol{H_5} = \boldsymbol{A_5 D_6 A_4 D_5^{-1}} \tag{A.32}$$

where $\boldsymbol{D_5}^{-1}$ is a diagonal matrix with the reciprocals of the elements of $\boldsymbol{g}$ on the main diagonal and $\boldsymbol{D_6}$ is a diagonal matrix with the elements of the $2t \times 1$ vector $[\exp(\boldsymbol{A_4} \log_e \boldsymbol{g})]$ on the main diagonal.

With $t^* = t(t-1)/2 = \sum_{j=1}^{t-1} j$, the matrix $\boldsymbol{A_4}$ in (A.33), generates the appropriate logarithms of product terms in (4.10) using the elements of $\boldsymbol{g}$. More specifically, the structure [$\boldsymbol{I}$ **0 0 0 0**] generates $\frac{1}{n}f_1$, $\frac{1}{n}f_2$, …, $\frac{1}{n}f_t$, the structure [**0** $\boldsymbol{I}$ **0** $\boldsymbol{I}$ **0**] generates $\frac{1}{n}h_1w_1$, $\frac{1}{n}h_2w_2$, …, $\frac{1}{n}h_tw_t$, and the structure [**0** $\boldsymbol{A_{4.1}}$ **0** $\boldsymbol{A_{4.2}}$ $\boldsymbol{A_{4.3}}$] generates the rest of the product terms involved in formulating $\boldsymbol{q}^*$. Then, the row of the matrix $\boldsymbol{A_5}$, defined in (A.34), sums the appropriate product terms into the corresponding elements

of the vector $\boldsymbol{q}^*$.

$$\boldsymbol{A_4} = \begin{bmatrix} \boldsymbol{I}_t & \boldsymbol{0}_{t,t} & \boldsymbol{0}_t & \boldsymbol{0}_{t,t} & \boldsymbol{0}_{t,t} \\ \boldsymbol{0}_{t,t} & \boldsymbol{I}_t & \boldsymbol{0}_t & \boldsymbol{I}_t & \boldsymbol{0}_{t,t} \\ \boldsymbol{0}_{t^*,t} & \boldsymbol{A_{4.1}} & \boldsymbol{0}_{t^*} & \boldsymbol{A_{4.2}} & \boldsymbol{A_{4.3}} \end{bmatrix}_{(2t+t^*)\times(4t+1)} \tag{A.33}$$

$$\boldsymbol{A_5} = \begin{bmatrix} \boldsymbol{I}_t & \boldsymbol{I}_t & \boldsymbol{A'_{4.2}} \end{bmatrix}_{t\times(2t+t^*)} \tag{A.34}$$

The forms of $\boldsymbol{A_{4.1}}$, $\boldsymbol{A_{4.2}}$, and $\boldsymbol{A_{4.3}}$ are described in (A.35), (A.36), and (A.37), respectively, where T' is a lower triangular matrix with all non-zero elements equal to 1. The first $(t-1)$ rows of the structure $[\ \boldsymbol{0}\ \boldsymbol{A_{4.1}}\ \boldsymbol{0}\ \boldsymbol{A_{4.2}}\ \boldsymbol{A_{4.3}}]$ generate the logarithms of the $(t-1)$ product terms involving w_1, then the next $(t-2)$ rows generate the logarithms of the $(t-2)$ product terms involving w_2, and so on, the last row generates the product term involving w_{t-1}. As an example, when $t=5$, the structures of matrices $\boldsymbol{A_{4.1}}$, $\boldsymbol{A_{4.2}}$, and $\boldsymbol{A_{4.3}}$ are shown in (A.38).

$$\boldsymbol{A_{4.1}} = \begin{bmatrix} \boldsymbol{1}_{(t-1)} & \boldsymbol{0}_{(t-1)} & \boldsymbol{0}_{(t-1)} & \cdots & \boldsymbol{0}_{(t-1)} & \boldsymbol{0}_{(t-1)} & \boldsymbol{0}_{(t-1)} \\ \boldsymbol{0}_{(t-2)} & \boldsymbol{1}_{(t-2)} & \boldsymbol{0}_{(t-2)} & \cdots & \boldsymbol{0}_{(t-2)} & \boldsymbol{0}_{(t-2)} & \boldsymbol{0}_{(t-2)} \\ \boldsymbol{0}_{(t-3)} & \boldsymbol{0}_{(t-3)} & \boldsymbol{1}_{(t-3)} & \cdots & \boldsymbol{0}_{(t-3)} & \boldsymbol{0}_{(t-3)} & \boldsymbol{0}_{(t-3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \boldsymbol{0}_2 & \boldsymbol{0}_2 & \boldsymbol{0}_2 & \cdots & \boldsymbol{1}_2 & \boldsymbol{0}_2 & \boldsymbol{0}_2 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix}_{t^*\times t} \tag{A.35}$$

$$\boldsymbol{A_{4.2}} = \begin{bmatrix} \mathbf{0}_{(t-1)} & \boldsymbol{I}_{(t-1)} \\ \mathbf{0}_{(t-2),2} & \boldsymbol{I}_{(t-2)} \\ \mathbf{0}_{(t-3),3} & \boldsymbol{I}_{(t-3)} \\ \vdots & \vdots \\ \mathbf{0}_{2,(t-2)} & \boldsymbol{I}_2 \\ \mathbf{0}'_{(t-1)} & 1 \end{bmatrix}_{t^* \times t} \tag{A.36}$$

$$\boldsymbol{A_{4.3}} = \begin{bmatrix} \boldsymbol{T}'_{(t-1)} & \mathbf{0}_{(t-1)} \\ [\mathbf{0}_{(t-2)} \quad \boldsymbol{T}'_{(t-2)}] & \mathbf{0}_{(t-2)} \\ [\mathbf{0}_{(t-3),2} \quad \boldsymbol{T}'_{(t-3)}] & \mathbf{0}_{(t-3)} \\ \vdots & \vdots \\ [\mathbf{0}_{2,(t-3)} \quad \boldsymbol{T}'_2] & \mathbf{0}_2 \\ [\mathbf{0}'_{(t-2)} \quad 1] & 0 \end{bmatrix}_{t^* \times t} \tag{A.37}$$

$$\boldsymbol{A_{4.1}} = \begin{bmatrix} \mathbf{1}_4 & \mathbf{0}_4 & \mathbf{0}_4 & \mathbf{0}_4 & \mathbf{0}_4 \\ \mathbf{0}_3 & \mathbf{1}_3 & \mathbf{0}_3 & \mathbf{0}_3 & \mathbf{0}_3 \\ \mathbf{0}_2 & \mathbf{0}_2 & \mathbf{1}_2 & \mathbf{0}_2 & \mathbf{0}_2 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}_{10 \times 5} \tag{A.38}$$

$$\boldsymbol{A_{4.2}} = \begin{bmatrix} \mathbf{0}_4 & \boldsymbol{I}_4 \\ \mathbf{0}_{3,2} & \boldsymbol{I}_3 \\ \mathbf{0}_{2,3} & \boldsymbol{I}_2 \\ \mathbf{0}'_4 & 1 \end{bmatrix}_{10 \times 5}$$

$$\boldsymbol{A_{4.3}} = \begin{bmatrix} \boldsymbol{T}_4' & \mathbf{0}_4 \\ [\mathbf{0}_3 \quad \boldsymbol{T}_3'] & \mathbf{0}_3 \\ [\mathbf{0}_{2,2} \quad \boldsymbol{T}_2'] & \mathbf{0}_2 \\ [\mathbf{0}_3' \quad 1] & 0 \end{bmatrix}_{10\times 5}$$

Finally, the vector $\boldsymbol{q} = (q_1, q_2, \ldots, q_t, q_{(t+1)})'$ can be written as

$$\boldsymbol{q} = \boldsymbol{A_6 q^*} + \boldsymbol{C_3} \tag{A.39}$$

with

$$\boldsymbol{A_6} = \begin{bmatrix} \boldsymbol{I}_t \\ -\mathbf{1}_t' \end{bmatrix} \text{ and } \boldsymbol{C_3} = \begin{bmatrix} \mathbf{0}_t \\ 1 \end{bmatrix}.$$

The corresponding variance-covariance matrix estimate is then given by

$$\boldsymbol{V_q} = \boldsymbol{A_6 V_{q^*} A_6}' \tag{A.40}$$

To estimate the covariance matrix for vector $\boldsymbol{q}_\theta = (q_{1\theta}, q_{2\theta}, \ldots, q_{t\theta}, q_{(t+1)\theta})'$, one could replace $\boldsymbol{g}$ and $\boldsymbol{V_g}$ with $\boldsymbol{g}_\theta$ and $\boldsymbol{V}_{\boldsymbol{g}_\theta}$ in the derivations from (A.29) to (A.40). $\boldsymbol{V}_{\boldsymbol{g}_\theta}$ takes the form of

$$\boldsymbol{V}_{\boldsymbol{g}_\theta} = \begin{bmatrix} \boldsymbol{V_a} & \boldsymbol{V_a H_3}' & -\boldsymbol{V_a H_3}' \\ \boldsymbol{H_3 V_a} & \boldsymbol{H_3 V_a H_3}' & -\boldsymbol{H_3 V_a H_3}' \\ -\boldsymbol{H_3 V_a} & -\boldsymbol{H_3 V_a H_3}' & \boldsymbol{H_3 V_a H_3}', \end{bmatrix}, \tag{A.41}$$

since the first partial derivative matrix of $\boldsymbol{h}_\theta$ with respect to $\boldsymbol{a}$ is $\boldsymbol{H_3}$.

Appendix 8:

Covariance estimates for $\log_e$ IDR ($\hat{\boldsymbol{\eta}}$) and $\log_e$ OR ($\hat{\boldsymbol{\psi}}$)

Let $i = 1, 2$ index the test and control treatments. And let $\boldsymbol{q}_{i\theta_i} = (q_{i1\theta_i}, q_{i2\theta_i} \ldots, q_{it\theta_i}, q_{i(t+1)\theta_i})'$ contains the distribution estimated via imputation under θ_i. One can construct a concatenated vector

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{q}_{1\theta_1} \\ \boldsymbol{q}_{2\theta_2} \end{bmatrix} \tag{A.42}$$

with the consistent covariance matrix estimator

$$\boldsymbol{V}_{\boldsymbol{F}} = \begin{bmatrix} \boldsymbol{V}_{\boldsymbol{q}_{1\theta_1}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{\boldsymbol{q}_{2\theta_2}} \end{bmatrix} \tag{A.43}$$

Applying the principle of linear and logarithmic transformation described in **??**, the logarithms of the interval specific incidence density ratios (IDRs) $\hat{\boldsymbol{\eta}}$=($\hat{\eta}_1$, $\hat{\eta}_2$, …, $\hat{\eta}_t$)', can be formulated for the comparison between treatments as the compound function of the vector $\boldsymbol{F}$ via

$$\hat{\boldsymbol{\eta}} = \boldsymbol{A_8} \log_e(\boldsymbol{A_7 F}), \tag{A.44}$$

where

$$\boldsymbol{A_7} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{I}_t & \boldsymbol{0}_t \\ \boldsymbol{T}_t & \boldsymbol{1}_t \end{bmatrix}$$

with the upper triangular matrix $\boldsymbol{T}_t$ shown in (A.5), and

$$\boldsymbol{A_8} = [\ \boldsymbol{I}_t \quad -\boldsymbol{I}_t \quad -\boldsymbol{I}_t \quad \boldsymbol{I}_t\ ].$$

Similarly, the interval specific $\log_e$ ORs, $\hat{\boldsymbol{\psi}}=(\hat{\psi}_1, \hat{\psi}_2, \ldots, \hat{\psi}_t)$' can be expressed as

$$\hat{\boldsymbol{\psi}} = \boldsymbol{A_8} \log_e(\boldsymbol{A_9 F}), \tag{A.45}$$

where

$$\boldsymbol{A_9} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{I}_t & \boldsymbol{0}_t \\ (\boldsymbol{T}_t - \boldsymbol{I}_t) & \boldsymbol{1}_t \end{bmatrix}$$

It then follows from linear Taylor series methods that the corresponding covariance matrix estimators are

$$\boldsymbol{V}_{\hat{\boldsymbol{\eta}}} = \boldsymbol{A_8 D_7^{-1} A_7 V_F A_7' D_7^{-1} A_8'} \tag{A.46}$$

and

$$\boldsymbol{V}_{\hat{\boldsymbol{\psi}}} = \boldsymbol{A_8 D_8^{-1} A_9 V_F A_9' D_8^{-1} A_8'}, \tag{A.47}$$

respectively, where $\boldsymbol{D_7^{-1}}$ is a diagonal matrix with the reciprocals of the elements of the $4t \times 1$ vector $\boldsymbol{A_7 F}$ on the main diagonal and $\boldsymbol{D_8^{-1}}$ is a diagonal matrix with the reciprocals of the elements of the $4t \times 1$ vector $\boldsymbol{A_9 F}$ on the main diagonal.

Appendix 9:

Variance estimates for Mann-Whitney probability $\hat{\xi}$

With the concatenated vector $\boldsymbol{F}$ and the corresponding covariance matrix estimator $\boldsymbol{V_F}$ shown in (A.42 and A.43, respectively), the $\hat{\xi}$ shown in (4.23) can be written in the general log-linear vector form

$$\hat{\xi} = \mathbf{1}'_{(t+1)} \exp[\boldsymbol{A_{12}} \log_e(\boldsymbol{A_{11}} \boldsymbol{F})], \tag{A.48}$$

where

$$\boldsymbol{A}_{11} = \begin{bmatrix} \boldsymbol{I}_{(t+1)} & \mathbf{0}_{(t+1)} \\ \mathbf{0}_{(t+1),(t+1)} & \boldsymbol{T}'_{(t+1)} - 0.5\boldsymbol{I}_{(t+1)} \end{bmatrix}$$

with $\boldsymbol{T}'_{t+1}$ being a lower triangular matrix with all non-zero elements equal to 1, and

$$\boldsymbol{A}_{12} = [\ \boldsymbol{I}_{(t+1)} \quad \boldsymbol{I}_{(t+1)}\ ].$$

It then follows that the estimated variance for $\hat{\xi}$ can be computed as

$$V_{\hat{\xi}} = \mathbf{1}'_{(t+1)} \boldsymbol{D_{10}} \boldsymbol{A_{12}} \boldsymbol{D_9^{-1}} \boldsymbol{A_{11}} \boldsymbol{V_F} \boldsymbol{A'_{11}} \boldsymbol{D_9^{-1}} \boldsymbol{A'_{12}} \boldsymbol{D_{10}} \mathbf{1}_{(t+1)}, \tag{A.49}$$

where $\boldsymbol{D_9^{-1}}$ is a diagonal matrix with the reciprocals of the elements of the $2(t+1) \times 1$ vector $\boldsymbol{A_{11}F}$ on the main diagonal, and $\boldsymbol{D_{10}}$ is a diagonal matrix with the elements of the $(t+1) \times 1$ vector $\exp[\boldsymbol{A_{12}} \log_e(\boldsymbol{A_{11}F})]$ on the main diagonal.

Appendix 10:

Calculation of Q_{MH}

Define the vector $\tilde{\boldsymbol{F}}$ as

$$\tilde{\boldsymbol{F}} = \begin{bmatrix} \boldsymbol{f}_{1\theta_1} \\ \boldsymbol{f}_{2\theta_2} \end{bmatrix} = \begin{bmatrix} n_1 \boldsymbol{q}_{1\theta_1} \\ n_2 \boldsymbol{q}_{2\theta_2} \end{bmatrix} \tag{A.50}$$

with the corresponding covariance matrix estimator

$$\boldsymbol{V}_{\tilde{\boldsymbol{F}}} = \begin{bmatrix} n_1^2 \boldsymbol{V}_{\boldsymbol{q}_{1\theta_1}} & \boldsymbol{0} \\ \boldsymbol{0} & n_2^2 \boldsymbol{V}_{\boldsymbol{q}_{2\theta_2}} \end{bmatrix}. \tag{A.51}$$

The Mantel-Haenszel criterion in (4.24) is calculated as

$$\text{MHE} = \boldsymbol{A_{16}} \exp[\boldsymbol{A_{15}} \log_e(\boldsymbol{A_{14}} \boldsymbol{F})], \tag{A.52}$$

where

$$\boldsymbol{A}_{14} = \begin{bmatrix} \boldsymbol{I}_t & \boldsymbol{0}_t & \boldsymbol{0}_{t,(t+1)} \\ \boldsymbol{T}_t & \boldsymbol{1}_t & \boldsymbol{0}_{t,(t+1)} \\ \boldsymbol{I}_t & \boldsymbol{0}_t & \boldsymbol{I}_t \quad \boldsymbol{0}_t \\ \boldsymbol{T}_t & \boldsymbol{1}_t & \boldsymbol{T}_t \quad \boldsymbol{1}_t \end{bmatrix},$$

$$\boldsymbol{A_{15}} = \begin{bmatrix} \boldsymbol{I}_t & \boldsymbol{0}_{t,t} & \boldsymbol{0}_{t,t} & \boldsymbol{0}_{t,t} \\ \boldsymbol{0}_{t,t} & \boldsymbol{I}_t & \boldsymbol{I}_t & -\boldsymbol{I}_t \end{bmatrix},$$

and

$$\boldsymbol{A_{16}} = [\ \mathbf{1}_t' \quad -\mathbf{1}_t'\ ].$$

With the linear Taylor series method, the variance of MHE in (A.52), i.e., the denominator of Mantel-Haenszel statistics Q_{MH}, is then given by

$$\text{Var(MHE)} = \boldsymbol{A_{16}D_{12}A_{15}D_{11}^{-1}A_{14}V_F A_{14}'D_{11}^{-1}A_{15}'D_{12}A_{16}'}, \tag{A.53}$$

where $\boldsymbol{D_{11}^{-1}}$ is a diagonal matrix with the reciprocals of the elements of the $4t \times 1$ vector $\boldsymbol{A_{14}\tilde{F}}$ on the main diagonal, and $\boldsymbol{D_{12}}$ is a diagonal matrix with the elements of the $2t \times 1$ vector $\exp[\boldsymbol{A_{15}} \log_e(\boldsymbol{A_{14}\tilde{F}})]$ on the main diagonal.

Bibliography

Acion, L.; Peterson, J.; Temple, S., and Arndt, S. Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*, 25:591–602, 2006.

Ali, M.W. and Siddiqui, O. Multiple imputation compared with some informative dropout procedures in the estimation and comparison of rates of changes in longitudinal clinical trials with dropouts. *Journal of Biopharmaceutical Statistics*, 10(2): 165–181, 2000.

Austin, P.C.; Chiu, M.; Ko, D.T.; Geoeree, Ron, and Tu, J.V. Propensity score matching for estimating treatment effects. In Faries, D.E.; Leon, A.C.; Haro, J.M., and Obenchain, R.L., editors, *Analysis of Observational Health Care Data Using SAS*, chapter 3, pages 51–84. SAS Institute Inc., Cary, NC, 2010.

Brown, J.B.W.; Hollander, M., and Korwar, R.M. Nonparametric tests of independence for censored data, with applications to heart transplant studies. In Proschan, F. and Serfling, R. J., editors, *In Reliability and Biometry: Statistical Analysis of Lifelength*, pages 327–354. SIAM, Philadelphia, 1974.

Calabrese, J.R.; Bowden, C.L.; Sachs, G.; Yatham, L.N.; Behnke, K.; Mehtonen, O.P.; Montgomery, P.; Ascher, J.; Paska, W.; Earl, N.; DeVeaugh-Geiss, J., and Group, Lamictal 605 Study. A placebo-controlled 18-month trial of lamotrigine and lithium maintenance treatment in recently depressed patients with bipolar i disorder. *Journal of Clinical Psychiatry*, 64(9):1013–1024, 2003.

Carpenter, J.R. and Kenward, M.G. *Missing Data in Randomised Controlled Trials - A Practical Guide.* Birmingham: National Institute for Health Research, Publication RM03/JH17/MK, available at http:\\www.hta.ac.uk\mihrmethodology\methodology \reports\1598.pdf, 2008.

CHMP, . *Guideline on Missing data in Confirmatory Clinical Trials (EMA /CPMP /EWP /1776 /99).* Committee fo Medicinal Products for Human Use (CHMP), 2010.

Cole, S.R. and Hernán, M.A. Adjusted survival curves with inverse probaility weights. *Computer Methods and Programs in Biomedicine*, 75:45–49, 2004.

Collett, D. *Modelling Survival Data in Medical Research (2nd edn).* Chapman & Hall/CRC, Boca Raton, Florida, 2003.

Collins, L.M.; Schafer, J.L., and Kam, C.M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Phsychological Methods*, 6(4):330–351, 2001.

Cox, D.R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

Curtis, L.H.; Hammill, B.G.; Eisensein, E.L.; Kramer, J.M., and Anstrom, K.J. Using inverse probability-weighted estimators in comparative effectiveness analysis with observational databases. *Medical Care*, 45(10 Suppl 2):S103–S107, 2007.

Deddens, J.A. and Koch, G.G. Survival analysis, grouped data. In Kotz, S. and Johnson, N.L., editors, *Encyclopedia of Statistical Sciences (9)*, pages 129–134. John Wiley & Sons, Inc., New York, 1988.

DeSouza, C.M.; A.TLegedza, R., and Sankoh, A.J. An overview of practical approaches for handling missing data in clinical trials. *Journal of Biopharmaceutical Statistics*, 19(6):1055–1073, 2009.

Diggle, P.J. and Kenward, M.G. Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics*, 43(1):49–93, 1994.

Elashoff, J.D. and Koch, G.G. Statistical methods in trials of anti-ulcer drugs. In Swabb, E.A. and Szabo, S., editors, *Ulcer Disease: Investigation and Basis for Therapy*, chapter 18, pages 375–406. Marcel Dekker, Inc., New York, 1991.

Fitzmaurice, G.M. Methods for handling dropouts in longitudinal clinical trials. *statistica Neerlandica*, 57(1):75–99, 2003.

Flyer, P. and Hirman, J. Missing data in confirmatory clinical trials. *Journal of Biopharmaceutical Statistics*, 19(6):969–979, 2009.

Flyer, P.A. Discussion: "incompete data in clinical studies: Analysis, sensitivity, and sensitivity analysis" by geert molenberghs. *Drug Information Journal*, 43(4):437–439, 2009.

Gehan, E.A. A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52:203–223, 1965.

Glynn, R.J.; Schneeweiss, S., and Sturmer, T. Indications for propensity scores and review of their use in pharmacolepidemiology. *Basic & clinical Phaumacology & Toxocology*, 98:253–259, 2006.

Grizzle, J.E.; Starter, C.F., and Koch, G.G. Analysis of categorical data by linear models. *Biometrics*, 25:489–504, 1969.

Hauck, W.W.; Anderson, S., and Marcus, S.M. Should we adjust for covariates in nonlinear regression analyses of randomized trials. *Controlled Clinical Trials*, 19: 249–256, 1998.

Heitjan, D.F. Ignorability in general incomplete-data models. *Biometrika*, 81(4):701–708, 1994.

Horton, N.J. and Lipsitz, S.R. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician*, 55 (3):244–254, 2001.

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

Johnson, W.D. and Koch, G.G. Linear models analysis of competing risks for grouped survival times. *International Statistical Review*, 46:21–51, 1978.

Kalbfleisch, J.D. and Prentice, R.L. *The Statistitical Analysis of Failure Time Data (2nd edn)*. John Wiley & Sons, Inc., New York, 2002.

Kaplan, E. L and Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

Koch, G.G.; Johnson, W.D., and Tolley, H.D. A linear models approach to the analsis of survival and extent of disease in multidimensional conteingency tables. *Journal of the American Statistical Association*, 67(340):783–796, 1972.

Koch, G.G.; Amara, I.A.; Davis, G.W., and Gillings, D.B. A review of some statistical methods for covariance analysis of categorical data. *Biometrics*, 38:563–595, 1982.

Koch, G.G.; McCanless, I., and Ward Jr., J.F. Interpretation of statistical methodology associated with maintenance trials. *The American Journal of Medicine*, 77(suppl 5B): 43–50, 1984.

Koch, G.G.; Sen, P.K., and Amara, I.A. Log-rank scores, statistics, and tests. In *Encyclopedia of Statistics*, volume 5, pages 136–141. John Wiley & Sons, Inc., New York, 1985.

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

Lachin, J.M. Statistical considerations in the intent-to-treat principle. *Controlled Clinical Trials*, 21(5):167–189, 2000.

Lagakos, S.W. General right censoring and its impact on the analysis of survival data. *Biometrics*, 35(1):139–156, 1979.

Laird, N. and Olivier, D. Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76(374): 231–240, 1981.

Laird, N.M. and Ware, J.H. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

Lavange, L.M.; Keyes, L.L.; Koch, G.G., and Margolis, P.A. Application of sample survey methods for modelling ratios to incidence densities. *Statistics in Medicine*, 13:343–355, 1994.

Leung, K.; Elashoff, R.M., and Afifi, A.A. Censoring issues in survival analysis. *Annual Review of Public Health*, 18:83–104, 1997.

Lin, D.Y. and Wei, L.J. The robust inference for the cox proportional hazard model. *Journal of the American Statistical Association*, 84(408):1074–1078, 1989.

Little, R.J.A. Pattern-mixture models for multivariate incomplete data. *Journal of American Statistical Association*, 88(421):125–134, 1993.

Little, R.J.A. A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3):471–483, 1994.

Little, R.J.A. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121, 1995.

Little, R.J.A. and Rubin, D.B. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York, 2002.

Little, R.J.A. and Yau, L. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, 52(4):11324–1333, 1996.

Liu, G. and Gould, L.A. Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *Journal of Biopharmaceutical Statistics*, 12(2):207–226, 2002.

Mallinckrodt, C. H.; Sanger, T.M.; Dube, S.; DeBrota, D.J.; Molenberghs, G.; Carroll, R.J.; Potter, W.Z., and Tollefson, G.D. Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry*, 53(8):754–760, 2003.

Mallinckrodt, C. H.; Lane, P.W.; Schnell, D.; Peng, Y., and Mancuso, J.P. Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Information journal*, 42(4):303–319, 2008.

Mallinckrodt, C.H. and Kenward, M.G. Conceptual considerations regarding endpoints, hypotheses, and analyses for incomplete longtitudinal clinical trial data. *Drug Information journal*, 43(4):449–458, 2009.

Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50:163–170, 1966.

Mantel, N. and Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22:719–748, 1959.

Molenberghs, G. Incomplete data in clinical studies: Analysis, sensitivity, and sensitivity analysis. *Drug Information journal*, 43(4):409–429, 2009.

Molenberghs, G. and Kenward, M.G. *Missind Data in Clinical Studies*. John Wiley & Sons, Inc., New York, 2007.

Moodie, P.F.; Saville, B.R.; Koch, G.G., and Tangen, C.M. Estimating covariate-adjusted log hazard ratios for multiple time intervals in clinical trials using nonparametric randomization based ancova. *Statistics in Biopharmaceutical Research*, 3(2): 232–241, 2011.

Newcombe, R.G. Confidence interval for an effect size measure based on the mann-whitney statistic. part 1: General issues and tail-are-based methods. *Statistics in Medicine*, 25:543–557, 2006.

NRC, . *The Preventing and treatment of Missing Data in Clinical Trial*. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education, National Research Council of the National Academies, Washington, DC: The National Academies Press., 2010.

Permutt, T. and Pinheiro, J. Editorial: Dealing with the missing data challenges in clinical trials. *Drug Information Journal*, 43(4):403–408, 2009.

Robins, J.M.; Hernán, M.A., and Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

Roger, J.H. *Sensitivity Analysis for Longitudinal Studies with Withdrawals in Practice*. GlaxoSmithKline 2008 US Biostatistics Annual Conference, 2008.

Rosenbaum, P.R. Model-based direct adjustment. *Epidemiology*, 11(5):550–560, 2000.

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

Rothmann, M.D.; Koti, K.; Lee, K.Y.; Lu, H.L., and Shen, Y.L. Missing data in biologic oncology products. *Journal of Biopharmaceutical Statistics*, 19(6):1074–1084, 2009.

Ruan, P.K. and Gray, R.J. Sensitivity analysis of progression-free survival with dependent withdrawal. *Statistics in Medicine*, 27(8):1180–1198, 2008.

Rubin, D.B. *Multiple Imputation for Nonresponse in Survey*. John Wiley & Sons, Inc., New York, 1987.

Rubin, D.B. and Schenker, N. Multiple imputations in health-care database: an overview and some applications. *Statistics in Medicine*, 10(4):585–598, 1991.

Saville, B.R. and Koch, G.G. Estimating covariate-adjusted log hazard ratios in randomized clinical trials using cox proportional hazards models and nonparametric randomization based analysis of covariance. *Journal of Biopharmaceutical Statistics*, (in press), 2012.

Schafer, J.L. Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8(1):3–15, 1999.

Scharfstein, D.O. and Robins, J.M. Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89(3):617–634, 2002.

Siannis, F.; Copas, J., and G., Lu. Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics*, 6(1):77–91, 2005.

Siddiqui, O.; Hung, H.M., and O'Neill, R. Mmrm vs. locf: A comprehensive comparison based on simulation study and 25 nda datasets. *Journal of Biopharmaceutical Statistics*, 19(2):227–246, 2009.

Somerville, M.; Shannon, J., and T., Wilson. Design, summerization, alaysis, and interpretation of cancer prevention trials. In Peace, K.E., editor, *Design and Analysis of Clinical Trials with Time-to-Event Endpoints*, chapter 17, pages 387–420. Chapman & Hall/CRC, Boca Raton, Florida, 2009.

Stokes, M.E.; Davis, C.S., and Koch, G.G. Weighted least squares. In *Categorical Data Analysis Using the SAS System, Third Edition*, chapter 14, pages 427–486. SAS Institute Inc., Cary, NC, 2012.

Tangen, C.M. and Koch, G.G. Non-parametric covariance methods for incidence density analysis of time-to-event data from a randomized clinical trial and their complementary roles to proportional hazards regression. *Statistics in Medicine*, 19:1039–1058, 2000.

Taylor, J.M.G.; Murray, S., and Hsu, C-H. Survival estimation and testing via multiple imputation. *Statistics and Probability Letters*, 58(3):221–232, 2002.

Tsiatis, A. A nonidentifiability aspect of the problem of competing risks. *proceedings of the National Academy of Sciences*, 72(1):20–22, 1972.

Walton, M.K. Addressing and advancing the problem of missing data. *Journal of Biopharmaceutical Statistics*, 19(6):945–956, 2009.

Wei, L.J.; Lin, D.Y., and Weissfeld, L. Regression analysis of multivariate incomplete failure time data by modeling marginal distribution. *Journal of the American Statistical Association*, 84(408):1065–1073, 1989.

Wittes, J. Missing inaction: Preventing missing outcome data in randomized clinical clinical trials. *Journal of Biopharmaceutical Statistics*, 19(6):957–968, 2009.

Yan, X.; Lee, S., and Li, N. Missing data handling methods in medical device clinical trials. *Journal of Biopharmaceutical Statistics*, 19(6):1085–1098, 2009.

Zhang, J. Sensitivity analysis of missing data: Case studies using model-based mulitple imputation. *Drug Information Journal*, 43:475–484, 2009.

Zhao, Y.; Herring, A.H.; Zhou, H; Ali, M.W., and Koch, G.G. A multiple imputation strategy for sensitivity analyses of time-to-event data with possibly informative censoring. *Journal of Biopharmaceutical Statistics (Accepted)*, 2012.