A Sensitivity Analysis for Clinical Trials with Informatively Censored Survival Endpoints

Eric Norbert Meier

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2012

Committee:

Scott Emerson

Gary Chan

Program Authorized to Offer Degree:

School of Public Health, Department of Biostatistics

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Table Number Page

ACKNOWLEDGEMENTS

I would like to express my gratitude for the support and guidance of my advisor, Professor Scott Emerson, and the encouragement and patience of my family, without which, I could not have completed this thesis.

DEDICATION

To Jean and Henry.

Introduction

Analyses of clinical trials with time-to-event endpoints typically employ the assumption of non-informative censoring. While this assumption is usually appropriate for end-of-study (EOS) censoring, its applicability to lost-to-follow-up (LTFU) censoring is often suspect and may result in biased estimates of the treatment effect. To assess the robustness of estimates to departures from non-informative censoring, authors [1][2] have proposed sensitivity analyses that assume a semiparametric model for the censoring mechanism, with the parameters representing associations between censoring and increased or decreased rates of survival. The parameters are varied over a plausible range resulting in a corresponding range of estimates for the treatment effect. We consider such an approach for two-arm trials in which the sensitivity parameters represent hazard ratios within a proportional hazards model with a time varying covariate comparing subjects who have been lost to follow-up to all other subjects. Using hypothesized hazard ratios for each arm separately, we multiply impute the unobserved data as it might have been observed in the absence of informative censoring. The treatment effect estimates computed using the imputed data are then summarized in a graphical display.

Of particular interest in this research is the robustness of our approach to violations of the proportional hazards assumptions used when imputing the missing data. On the basis of extensive simulation studies, we find that the accuracy of the sensitivity analyses are relatively unaffected by departures from the semiparametric assumptions.

The remainder of the thesis is outlined as follows. In Chapter 1, we discuss censoring, the problem of non-identifiability, and methods authors have proposed to address informative censoring. In Chapter 2, we describe an imputation-based sensitivity analysis for two-arm trials. In the next two chapters, we explore the performance of the imputation method when the assumption regarding the proportionality of hazards between LTFU-censored subjects and all other subjects is correct (Chapter 3) and incorrect (Chapter 4). We conclude with a discussion of the results, possible extensions to the method, and some thoughts about the use of sensitivity analyses in the evaluation of clinical trials.

Chapter 1

Background

Clinical trial investigators are often interested in estimating the difference in the time to some event, such as death, between subjects randomly assigned to different treatments. The *time-to-event* is defined as the difference between a well-defined start time, often the time of randomization, and a well-defined end time (e.g. death). If all the subjects were to experience the event, analysts could simply test for differences in the mean time-to-event between treatment groups using a standard method such as the *t* test. However, it is rarely feasible to wait until all participants have experienced the event to carry out the analysis. Instead, analysis typically occurs after a planned number of subjects have experienced the event or a planned number of subjects have been followed up for a specified period of time. In either case, there are nearly always subjects for whom the time-to-event is not known at analysis time. Furthermore, there are often subjects who drop out of the trial before experiencing the event of interest and it may not be possible to learn anything more about their time-to-event.

Subjects for whom time-to-event is not fully known are said to be *censored*. Censoring can be divided into two mechanisms, *point* and *interval* [3]. Point censoring occurs when either the start or end time of the time-to-event definition is not known precisely but rather can only be bounded in one direction. In clinical trials, the start time is almost always known for all subjects, thus point censoring usually arises because the end time is known only to exceed some value.

This type of point censoring is referred to as *right censoring* since when plotted on a horizontal timeline a subject's follow-up time begins on the left and ends on the right. Subjects whose start time is known only to have occurred by a certain time are said to be *left censored* and those with both left and right censoring are said to be *doubly censored*.

Interval censoring occurs when the time-to-event is known to be within an interval. This may occur, for example, in trials in which periodic examinations are conducted to assess disease progression. If progression is first noted at a given examination, it is only known that progression occurred between the last examination and the present one.

In the remainder of this paper we will restrict our attention to right censoring, the most common type in clinical trials. Unless otherwise specified, any discussion of censoring will refer to this type.

Censoring can be broadly separated into two types: end-of-study (EOS) and lost-to-follow-up (LTFU). EOS censoring, also referred to as administrative censoring, occurs when it is known that a subject has not experienced the event by the time planned follow-up ends.

LTFU censoring, in our broad categorization, includes all other types of censoring. The causes of it for a given subject may or may not be known. Subjects may drop out because they perceive the trial to be too burdensome or because they feel better, perhaps due to the treatment. They may move out of the area and be unreachable or be unable to continue treatment because it is not available in their new location. They may experience treatment-related events such as drug toxicity or intolerance or they may find that the study medication is ineffective.

An important consideration, particularly when there is significant censoring of outcomes, is whether the censoring is likely to be *informative* for survival. In other words, does the event time for subjects who have not experienced the event by time *t* depends on when follow-up ends? If it does not, we say that censoring is *non-informative*. Unfortunately, there is often nothing in our data to help us determine whether censoring is informative, because the data we need to determine dependence, the event time, is unknown for censored subjects.

If we consider it unlikely that the timing of study recruitment is related to time-to-event, then it is reasonable to assume that EOS censoring is not informative. Examples of situations where this assumption may not hold include trials in which entry criteria are altered over time (so a different mix of patients are enrolled at different points in time) and trials for which the efficacy of the treatment is subject to variation over time (e.g. if practitioners become more skilled with a new surgical procedure the more they perform it).

Many types of LTFU censoring are likely to be informative. For example, a subject's health may deteriorate, and the subject may conclude that the study medication is not working. The subject may then decide to drop out of the trial. We might expect such subjects to have shorter residual survival times than those subjects who remain in the trial. On the other hand, a subject may feel better, decide she no longer needs the study medication, and then drop out of the trial. We might expect such subjects to have longer residual survival times than those who remain under observation.

LTFU rates can vary widely. In cancer trials, for example, rates are commonly in the 5 to 10 percent range, whereas in some HIV/AIDS trials the rate has been significantly higher. It is not uncommon for trials of antipsychotic medication to have a dropout rate exceeding 50 percent.

The techniques used to analyze censored time-to-event data are generally referred to as survival analysis methods. In the typical analytical framework, we define random variables T and C as the event time and censoring time, respectively. We are interested in estimating the distribution of T, but what we observe is $Y = min(T, C)$ and $\delta = I(T \leq C)$ where $I(.)$ is a function that indicates whether the event was observed or not. The data (Y, δ) are insufficient to determine the joint distribution of (T, C). Tsiatis [4] proved that in addition to a model in which T and C are independent, there exist one or more models where T and C are dependent that would tend to yield the same observed data. This has come to be known as the non-identifiability problem.

The typical solution to this problem is to assume censoring is non-informative. Under this assumption, the likelihood contribution of an event is:

$$Pr[T = y_i, C > y_i] = (1-G(y_i))\, f(y_i) \qquad (1)$$

where G is the censoring distribution function and f is the time-to-event density. The likelihood contribution of a censored observation is:

$$Pr[C = y_i, T > y_i] = (1-F(y_i))\, g(y_i) \qquad (2)$$

where F is the time-to-event distribution function and g is the censoring density. The *survival* *function S* is often used in place of *1-F*. The likelihood can be written as:

$$L = \prod [(1-G(y_i))\, f(y_i)]^{\delta_i} [S(y_i)\, g(y_i)]^{1-\delta_i} \tag{3}$$

If the distribution of *C* does not depend on the parameters of the distribution of *T*, the terms $(1-G(y_i))^{\delta_i}$ and $g(y_i)^{1-\delta_i}$ can be factored out such that:

$$L \propto \prod f(y_i)^{\delta_i}\, S(y_i)^{1-\delta_i}\, . \tag{4}$$

This is the usual likelihood employed for the estimation of the survival function.

Partial likelihood-based approaches are the mainstay of the analysis of clinical trial time-to-event data. The typical analysis involves estimation of treatment-specific survival functions using the method of Kaplan and Meier [5], testing the hypothesis of no difference in the functions with the log rank test [6], and estimating the treatment effect, or hazard ratio, using the Cox proportional hazards model [7]. The assumption of non-informative censoring is invariably invoked for all these methods in the primary analyses of trial data. However, the assumption is often suspect for LTFU censoring.

In their 1958 paper introducing the product-limit estimator, Kaplan and Meier noted that the assumption of independence between censoring and survival times "deserves special scrutiny". Many authors since then have showed that informatively censored observations can bias estimates of the survival function [8]. Peterson [9] gave sharp bounds on the marginal survival function without making further assumptions about the association between survival and censoring. However, in most situations these bounds are considered too wide for practical use.

We can conceptualize informative censoring as being composed of a part that can be explained by measured factors that are prognostic for both censoring and survival and the remaining part that is not explained by measured factors. When all the factors prognostic for both censoring and survival are available, Robins [10] proved that the marginal survival function is identifiable and Robins and colleagues [11][12][13] proposed methods for estimating the function.

The general approach taken when there are no measured prognostic factors is to model the association between censoring and survival. The parameters of the models, which correspond to the degree of association, are varied over a plausible range and the resulting estimates are used to place bounds on the survival function. Fisher and Kanarek [14] considered such a model. They assumed that being lost to follow-up occurs simultaneously with an event that alters survival by an amount associated with a scale parameter α. In their model, when $\alpha = 1$, censoring has no effect on survival, while $\alpha < 1$ contracts survival and $\alpha > 1$ stretches it by $\alpha(t\text{-}c)$, where $t\text{-}c$ is the survival time following the censoring event. Lagakos and Williams [15] considered a model with an exponential survival function, an unspecified function $c(y)$ that measures the relative odds of observing a failure at $y = min(t,c)$, and a parameter θ that corresponds with the degree of association between censoring and survival with 0 indicating death immediately following censoring and 1 indicating non-informative censoring. Slud and Rubinstein [16] introduced a known function $\rho(t)$ specifying the hazard ratio between censored and uncensored subjects over time and showed how to calculate bounds on the survival function given bounds on $\rho(t)$. Klein and Moeschberger [17] took a similar approach with a fixed parameter (θ) representing the hazard ratio comparing the censored to uncensored. They demonstrated the relationship between this parameter and Kendall’s coefficient of concordance (τ) and discussed the use of that measure to specify the plausible range of association between censoring and survival. Zheng and Klein [18] showed that a known copula defining the dependence between censoring and survival is sufficient to identify the marginal survival function. The copula is chosen to be monotone in a given parameter and bounds for the survival function are estimated by varying the parameter over a range.

Scharfstein, Robins, Rotnitzky, and colleagues [1][2][19] developed methods to adjust for measured factors prognostic for both censoring and survival while simultaneously measuring the sensitivity of estimates to assumptions about the residual dependency between censoring and survival due to unmeasured factors. The most recent of these papers accommodates multiple competing censoring mechanisms with differing degrees of association with survival. The cause-specific censoring mechanisms are modeled with *censoring bias functions* which require the specification of two parameters. The parameter α is interpreted as the log hazard ratio of dropping out at time t between subjects who are at risk at time t and have the same covariate history, but who differ by one unit in their ultimate time-to-event. The parameter β represents the mean time to event for subjects who do not experience the event prior to the maximum potential follow-up time.

Siannis and colleagues [20] evaluated the sensitivity of inferences to small departures from non-informative censoring in what can be called a *local* sensitivity analysis. Their method assumes a parametric model which allows for dependence between censoring and survival in terms of a parameter δ, which can be interpreted in terms of a correlation coefficient between the two mechanisms, and a bias function. Along similar lines, Zhang and Heitjan [21] derived an index of local sensitivity to nonignorability (INSI) for a general parametric model of survival and censoring. Liu and Heitjan [22] extended this work to a nonparametric survival model.

Despite the extensive literature addressing the problem of informative censoring, sensitivity analyses evaluating the sensitivity of inference to departures from non-informative censoring are rarely presented in regulatory submissions or publications. This may be because, in a setting where analysis methods need to be pre-specified, analysts are concerned that the simpler methods may not provide accurate inference and the more complex methods are too difficult to specify without knowledge of the data. In addition, few, if any, of the methods are implemented in readily available software.

Chapter 2

A Sensitivity Analysis

In this section, we consider a two-parameter sensitivity analysis for two-arm trials. Its end product is a single figure that presents estimated treatment effect under a wide range of assumptions regarding treatment-specific hazard ratios comparing informatively censored subjects to all other subjects.

Our method pertains to the situation in which all censoring can be divided into one of two types: end-of-study (EOS) or lost-to-follow-up (LTFU). For each treatment arm, we further assume that EOS censoring is non-informative and that LTFU censoring is potentially informative.

Our approach is similar to the method of Fisher and Kanarek [14] in that we assume a survival-altering event simultaneous with censoring, but our α parameter represents the hazard ratio comparing LTFU-censored subjects to all other subjects. Thus, $\alpha > 1$ tends to decrease survival and $\alpha < 1$ tends to increase survival, relative to the assumption of non-informative censoring, while $\alpha = 1$ represents the assumption of non-informative censoring.

As described in the previous section, there are many different causes that result in subjects being lost to follow-up. Each may have a different degree of association with subsequent survival. A key assumption of our approach is that all the different causes can be summarized by some sort of an “average” association. Furthermore, we assume that decision makers can specify a plausible range of values for this association.

The “average” association, which we call the censoring adjustment factor (CAF), is treatment-specific. For simplicity of exposition, we assume a trial with “treatment” and “control” arms and refer to their CAFs as α_t and α_c. When not referring to a particular treatment arm, α is used without subscript.

For a given *scenario* (*i.e.* combination of α_t and α_c), we estimate the treatment effect and its variance through multiple imputation. We do not impute event times for every censored subject, but rather we seek to effectively “remove” informative censoring. The resulting dataset can be viewed as the answer to the question: “What would the data look like if there were no patients lost to follow-up in this trial?”

For each iteration of a scenario, we “remove” informative censoring from the data by assigning to each LTFU-censored subject, the minimum of (1) the time the subject would have been EOS censored and (2) the LTFU censoring time plus a residual survival time that is imputed using the relevant CAF. It is important to note that by taking the minimum here we do not impute survival times beyond the support of the observed data. Thus the method can be applied to situations in which the whole survival curve is not fully identified.

The residual survival time is imputed from a distribution derived from the Nelson-Aalen estimate of the cumulative hazard function. The cumulative hazard function under the assumption of informative censoring is as follows:

$$\Lambda_I(t) = I(t \leq c)\, \Lambda_N(t) + I(t > c)\, \{\Lambda_N(c) + \alpha\, [\Lambda_N(t) - \Lambda_N(c)]\} \tag{5}$$

where t is time, c is the LTFU censoring time, $I(.)$ is an indicator function that equals 1 if its argument is true and 0 otherwise, α is the censoring adjustment factor, and $\Lambda_N(.)$ is the estimated cumulative hazard function under the assumption of no informative censoring. The subscript N is used here and elsewhere to indicate a function or coefficient that is estimated under the assumption of no informative censoring. We will sometimes refer to such estimates as naïve estimates.

The cumulative failure distribution under informative censoring is

$$F_I(t) = 1 - exp[-\Lambda_I(t)] \tag{6}$$

and the survival time for an informatively censored subject is imputed as follows:

$$t_I = F_I^{-1}(U[F_I(c), 1]) \tag{7}$$

where $F_I^{-1}(.)$ is the inverse of $F_I(.)$ and $U[a, b]$ is a randomly generated number from the uniform distribution with range $[a, b]$.

After all subjects with LTFU censoring have been assigned the minimum of their imputed survival time and EOS censoring time, we estimate the treatment effect θ_i and its variance v_i for the i-th imputation *iteration* using Cox proportional hazards regression. The process is repeated I

times for a given scenario. The estimated scenario-specific treatment effect and its variance based on the multiple imputations are listed below [23].

$$\theta_s = mean(\theta_i), \;\; v_s = mean(v_i) + (I+1)/I \; var(\theta_i) \qquad (8)$$

The treatment effect and variance are estimated under a wide range of scenarios covering plausible values of α_t and α_c.

We now illustrate the method with the hypothetical example of a randomized trial with "treatment" and "control" arms, each with 200 enrolled subjects, and a primary endpoint of overall survival. The recruitment rate is assumed to be constant throughout 24 months and follow-up continues for 12 months following the end of recruitment. Thus, with the exception of early deaths and dropouts, subjects are followed up for a minimum of 12 months. Analysis is assumed to occur 36 months after the beginning of the trial.

Trial data are generated by simulating three times for each subject: lost-to-follow up, end-of-study, and survival. For the exploration of the robustness of our methods we find it convenient to generate the two censoring times independently and to specify the survival time as conditional on the LTFU censoring time. The observed data are the minimum of the three times and an indication of which time was the minimum. LTFU times are generated from a Weibull distribution with treatment-specific scale and shape parameters σ and w. Survival times are conditioned on the LTFU times and are generated by

$$s = F_E^{-1}(U[0,1]) \qquad (9)$$

where $F_E^{-1}(.)$ is the inverse of an estimated cumulative failure distribution specific to the subject's LTFU time, and $U[0,1]$ is a randomly generated number from the uniform distribution with range $[0,1]$. We estimate the cumulative failure distribution by:

$$F_E = 1\text{-}exp[-\Lambda_E(t)] \tag{10}$$

where $\Lambda_E(t)$ is the cumulative hazard function arrived at by numerically integrating from 0 to t over the hazard function:

$$h_E(t) = I(t \leq C)\ h_w(\beta,\ k,\ t) + I(t > C)\ [\alpha\ h_w(\beta,\ k,\ t)] \tag{11}$$

where $I(.)$ is the indicator function, t is time, C is the random variable for LTFU time, α is the true censoring adjustment factor, and $h_w(\beta,\ k,\ t)$ is the Weibull hazard function with treatment-specific scale (β) and shape (k) parameters. EOS times are independent of both LTFU and survival times and are generated from the uniform distribution where a is the minimum follow-up time and b is the maximum follow-up time.

Our example uses the following parameter values for the control and treatment arms.

Table 1. Parameters of example survival and LTFU censoring distributions

Control	Treatment
σ_c = 130 months	σ_t = 190 months
w_c = 1.1	$\mathrm{w_t}$ = 0.7
β_c = 15 months	β_t = 20 months
k_c = 1.0	k_t = 1.0

The EOS censoring parameters are the same for each arm: a =12 months and b = 36 months.

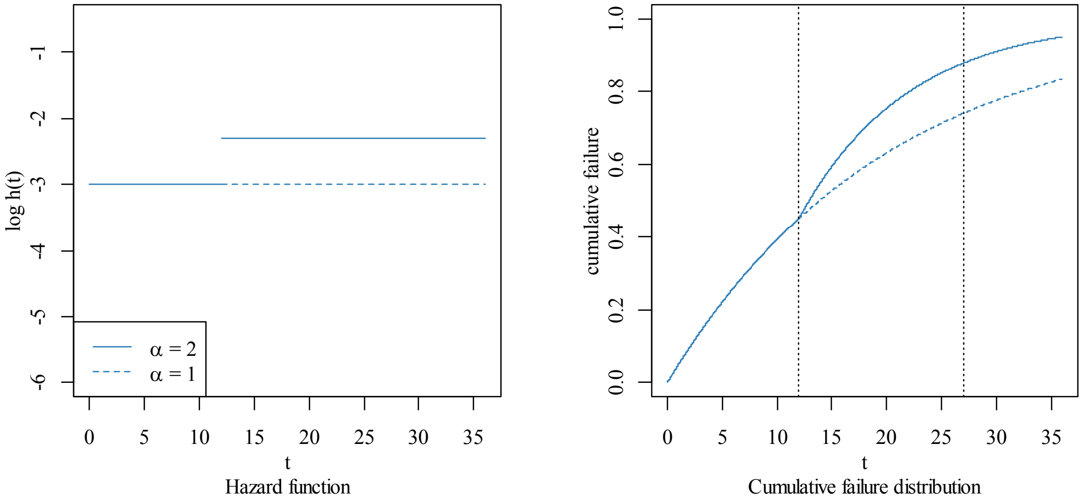

Figure 1. Theoretical hazards and cumulative failure distributions for hypothetical subject

An example of the data generation for one subject may help to further clarify the process. Let us suppose that in generating the trial data, one subject had an LTFU censoring time of 12 months and an EOS censoring time of 27 months[1]. The theoretical hazard and the related cumulative failure distribution used to generate the survival time for such a subject are presented in Figure 1. Curves are shown for two different assumptions: no informative censoring ($\alpha = 1$) and a doubling of the hazard after LTFU censoring ($\alpha = 2$). Note that the hazards are the same under both assumptions prior to the point of LTFU censoring. Assuming that the informative censoring assumption ($\alpha = 2$) is in fact true, the survival time generated for the subject depends upon the LTFU time and a randomly generated number from the *U[0, 1]* distribution, which we will call *u*. The simulated survival time is the time that corresponds to a cumulative failure probability of *u*. When $u \leq 0.41$, an event time is observed; otherwise, the subject is observed to be lost to follow-up at 12 months. The cutoff of 0.41 corresponds with the LTFU censoring time of

[1] The subject was simulated to have a maximum follow-up of 27 months. In other words the subject was recruited 9 months after the beginning of the study at which time there were 27 months before analysis time.

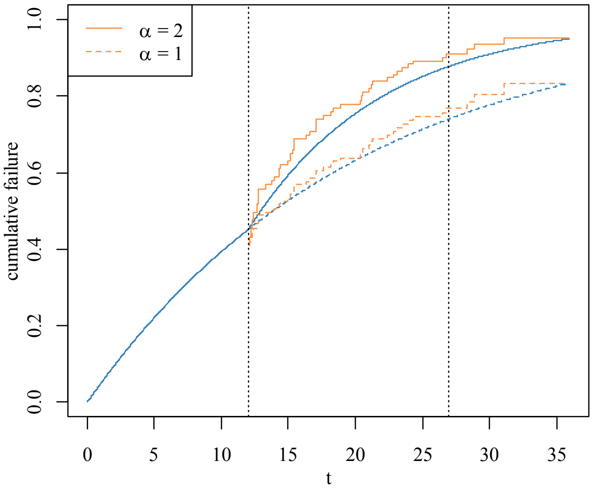

Figure 2. Cumulative failure distributions used for imputing survival

12 months in the cumulative failure distribution.

If the observed data turned out to be LTFU censoring at 12 months, the survival time for this subject would be imputed for the sensitivity analysis. Figure 2 shows an example of the cumulative failure plots on which this imputation would be based overlaid on the theoretical plots. The dashed line is the Kaplan-Meier estimated cumulative failure and the solid line is the cumulative failure that results from equations *(5)* and *(6)* where $\alpha = 2$ and $c = 12$. The imputed survival time is derived by generating a number, x, from $U[0.41, 1]$ and finding the time that corresponds with the cumulative failure of x. When $x \leq 0.91$, an event is observed for the imputed analysis; otherwise the subject is EOS censored at 27 months. In either case, the informative censoring has been effectively "removed" by the imputation process.

As noted above, when $u > 0.41$, the subject is LTFU censored. However, a survival time is still generated for the subject. This is the survival time that would be observed if there were no censoring at all, neither LTFU nor EOS. If we allow this survival time to be EOS censored when it is greater than 27 months, we arrive at the data we would have observed for the subject if there were no LTFU censoring. We will refer to such data as the “true” data for the subject. The table below summarizes the example subject’s observed data for the “true”, naïve, and imputed analyses under different ranges of *u* and *x*.

Table 2. Example subject’s observed data under different ranges of *u* and *x*

u ~ U[0, 1]	**Observed data**			***x ~ U[0.41, 1]***
	“True”	**Naïve**	**Imputed**	
$u \leq 0.41$	Event $s = F_E^{-1}(u)$	Event $s = F_E^{-1}(u)$	Event $s = F_E^{-1}(u)$	*n/a*
$0.41 < u \leq 0.77$	Event $s = F_E^{-1}(u)$	Censored 12 mos.	Event $s = F_I^{-1}(x)$	$0.41 < x \leq 0.91$
			Censored 27 mos.	$x > 0.91$
$u > 0.77$	Censored 27 mos.	Censored 12 mos.	Event $s = F_I^{-1}(x)$	$0.41 < x \leq 0.91$
			Censored 27 mos.	$x > 0.91$

We generate all the subject data for one instance of the example trial outlined above in this manner. Figure 3 presents a summary of the so-called naïve analysis, which assumes all censoring is non-informative. The results indicate a statistically significant treatment effect (p = 0.032) suggesting the treatment is beneficial. However, Figure 4 shows considerable LTFU censoring in both treatment arms. At analysis time, 7 percent of the control subjects and 12 percent of the treatment subjects had been lost to follow-up. This is a situation where we would like to assess the robustness of the results to alternative assumptions about the association between LTFU censoring and survival.

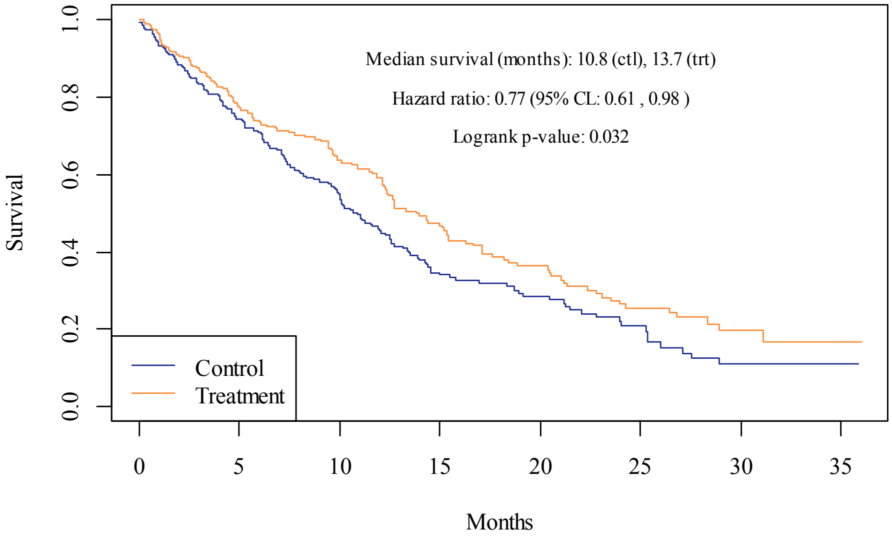

Figure 3. Kaplan-Meier survival under assumption of noninformative censoring

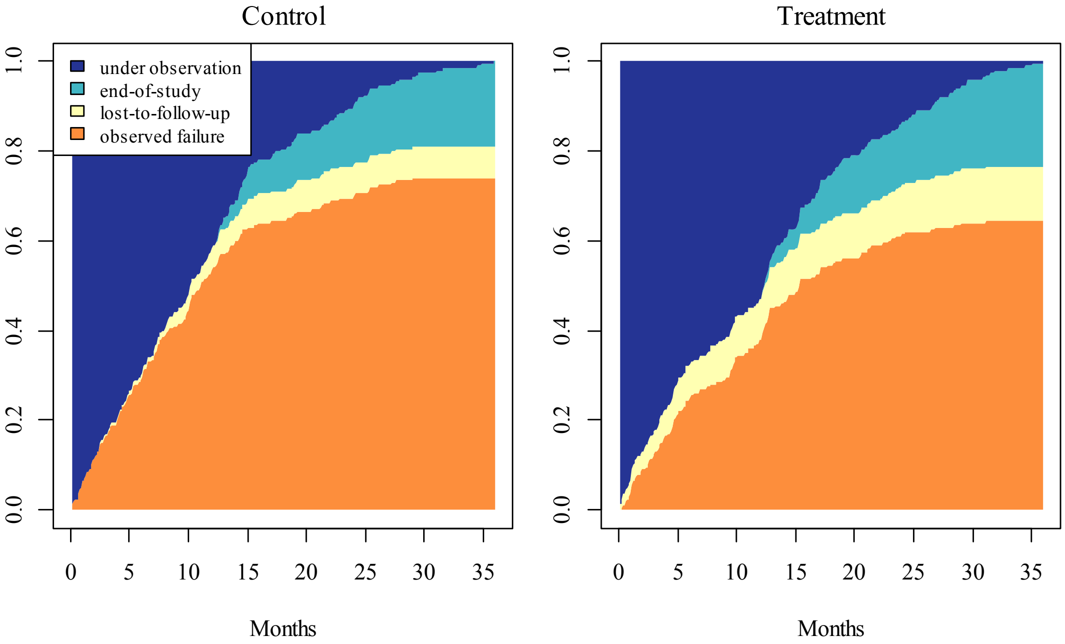

Figure 4. Distribution of censoring status over time

Using the imputation method outlined above, we estimate the treatment effect over a range of assumptions for α_t and α_c and present the results in two contour plots in Figure 5. The estimated hazard ratio is on the left, and its upper confidence limit is on the right. We vary each parameter on the log scale over a range of -1.1 to 1.1 in increments of 0.025. The range corresponds with hazard ratios from 0.33 to 3.00. Estimates are made for the 7,921 combinations of the two parameters. For each scenario, the estimated treatment effect is based on 50 imputed datasets (or *iterations*).

Contour lines are displayed for the estimated p values of 0.10, 0.05, and 0.01. These are plotted using a LOESS smoother on the parameter coordinates that correspond with scenarios that have estimated p values in the range of [0.09, 0.11), [0.045, 0.055), and [0.009, 0.011), respectively. The p value for a scenario is estimated by:

$$2\,[1-T(|\theta_S| / \sqrt{v_S}\,,\, \upsilon)] \qquad (12)$$

where $T(.)$ is the cumulative distribution function for the t distribution with υ degrees of freedom calculated from the following equation based on a Satterthwaite approximation [23].

$$\upsilon = (I-1)\,[1 + 1/(I+1)\; v_S / var(\theta_i)]^2 \qquad (13)$$

We now consider the plot of the estimated hazard ratios. The center corresponds with the assumption of non-informative censoring. We can see that the statistical significance of the treatment effect is robust to LTFU censoring assumptions that are non-differential ($\alpha_t = \alpha_c$) when $\alpha_t < 2$. We would expect such a turning point for trials in which the distributions of LTFU times

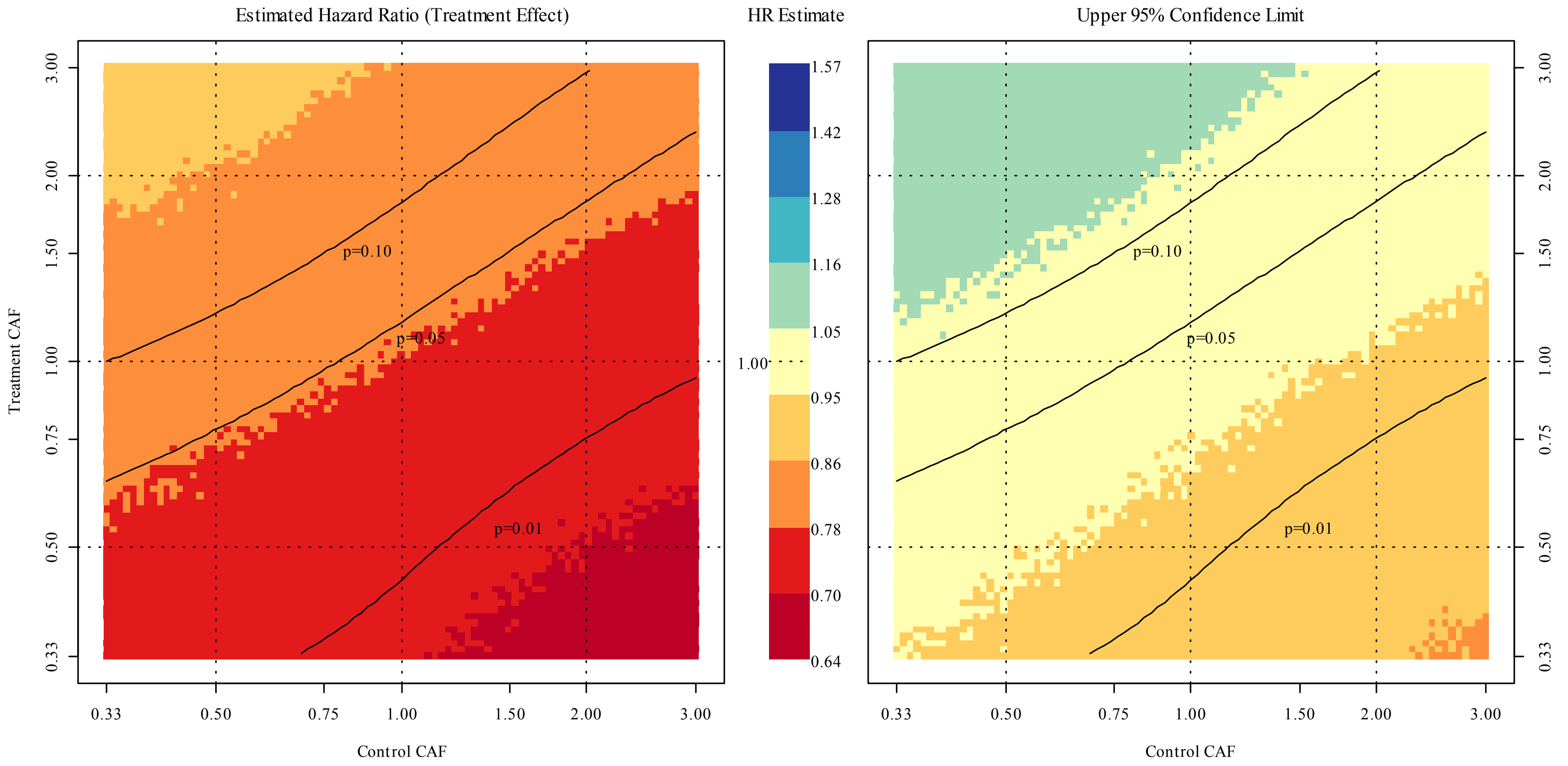

Figure 5. Graphical sensitivity analysis

differed between treatment arms as is the case with our example. When the distribution of LTFU censoring is similar between the two arms, the contours tend to run parallel to $\alpha_t = \alpha_c$.

The plots in Figure 5 are intended for decision makers who presumably will have some sense of plausible ranges for the parameters. This sense would likely be informed by data describing the likely reasons for a trial subject being lost to follow-up and the relative frequency of each reason.

The data for our example were simulated with $\alpha_c = 1.1$ and $\alpha_t = 2.0$. This is approximately the point in Figure 5 where the p = 0.10 contour crosses the $\alpha_t = 2$ line. The dashed lines in Figure 6 show what we refer to as the “true” estimates of the survival functions. The “true” estimate is obtained by changing the definition of the observed data to the minimum of the EOS time and survival time, effectively “removing” informative censoring. These data can be thought of as the data we would have observed had there been no lost to follow-up.

In Figure 6, we see that for much of the follow-up period the “true” survival estimate of the treatment arm is lower than the naïve estimate as we would expect with a doubling of the hazard ($\alpha_t = 2$) for subjects who are lost to follow-up. The “true” estimate of the control arm is actually slightly higher than the naïve estimate for most time points after 10 months despite the slightly increased hazard modeled for lost to follow-up subjects ($\alpha_c = 1.1$). This particular result can be ascribed to random variation in the generation of the survival times. In the next section we demonstrate that if we were to run the trial many times under the same assumptions, the “true” survival estimate of the control arm would, on average, be lower than the naïve estimate.

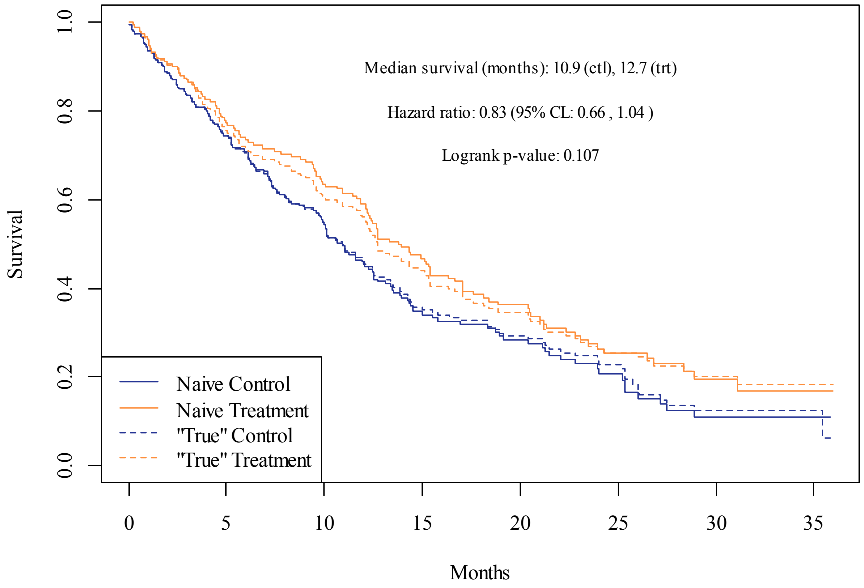

Figure 6. Kaplan-Meier survival after “removing” informative censoring

Chapter 3

Performance Under Proportional Hazards

Our approach assumes that, upon being lost to follow-up, subjects experience an increase or decrease in hazard proportional to subjects who have not been lost to follow-up. This increase or decrease is specified through treatment-specific parameters called censoring adjustment factors (CAFs). In this section, we assess the performance of our estimates when the assumptions regarding the values of the CAFs and the proportionality of hazards are indeed correct.

Taking up the above example again, but focusing only on the treatment arm where LTFU censoring is strongly associated with survival, we may ask how well the imputed estimate of the survival function compares to the "true" estimate when the correct value of α_t is used for the imputation. Figure 7 shows a point-wise average imputed survival function based on 50 imputed datasets using $\alpha_t = 2$. We see that the average imputed survival follows the "true" estimate more closely than the naïve estimate does, as we would expect, up until about 20 months. Thereafter the naïve estimate actually tracks closer to the "true" one.

However, this is only one instance of a trial. We would like to do a similar comparison over many such trials. Using the same parameters as our example, we simulate 5,000 trials and compare the imputed estimate of the survival function for the treatment arm to the naïve estimate. Given the number of simulations, the precision of our confidence interval coverage assuming a correct 0.05-level test is +/- 0.006. Figure 8 shows the average difference in survival

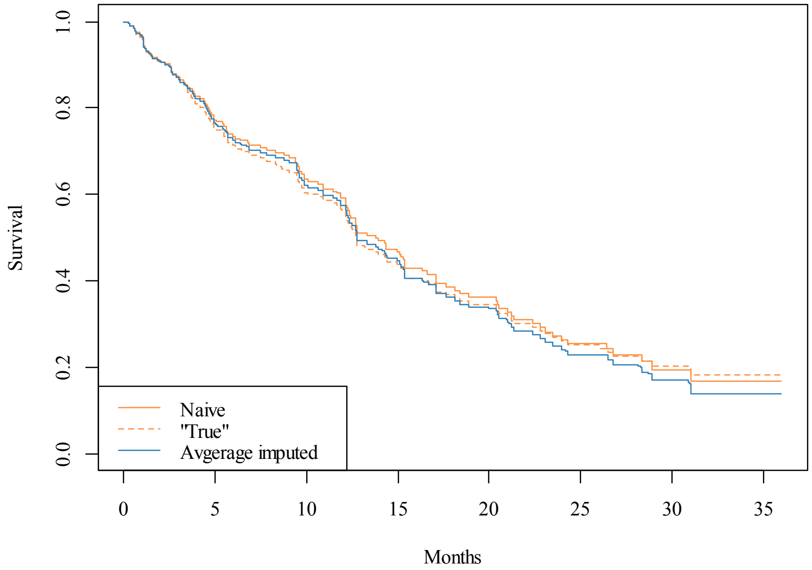

Figure 7. Average imputed survival compared to “true” and naïve estimates

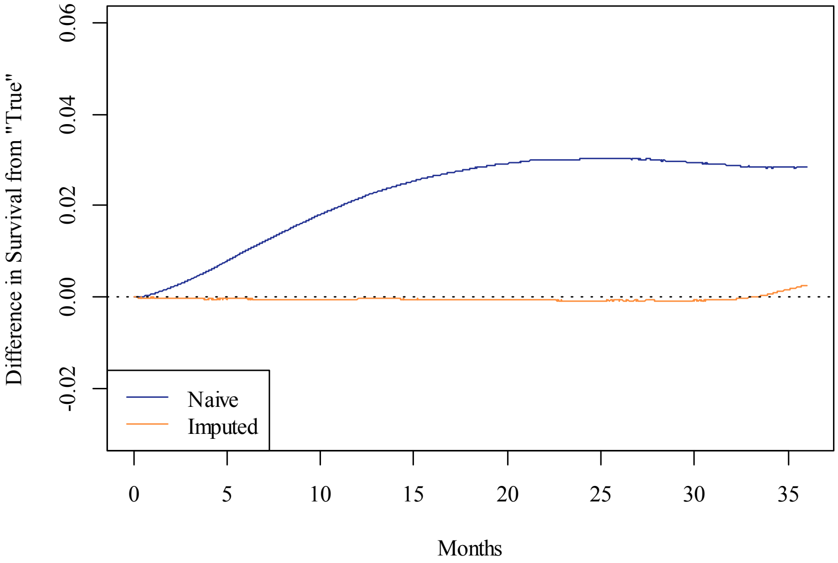

Figure 8. Average difference in survival from “true” estimate

from the “true” estimate across follow-up time for the two methods. We see that, on average, the imputed estimate hews to the “true” estimate throughout most of the follow-up period while the naïve estimate increasingly overestimates survival since it does not account for informative censoring.

There is a slight bias in the imputed survival estimate near the maximum follow-up time. In general, this occurs because we have relatively little data with which to estimate survival toward the end of follow-up. Figure 9 shows the naïve cumulative hazard for the treatment arm of our example along with the cumulative hazard used to impute survival time for a hypothetical subject who is lost to follow-up at 15 months. We can see that the cumulative hazard is flat from 32 to 36 months because there were no events during that time period. Since the naïve hazard, which is used as the baseline hazard that is either increased or decreased by the CAF for imputations, is zero from the last event time until the maximum follow-up time, subjects with LTFU censoring during this time period are effectively imputed to be EOS censored at the time they would have reached their maximum follow-up had they not been LTFU censored. Thus, during this period, there would be no simulated change to their hazard with respect to non-LTFU subjects.

Figure 10 shows the average root mean squared error (RMSE) from “true” survival across follow-up time for the naïve and imputed methods. For the first six months the imputed and naïve RMSEs are similar, but thereafter the imputed RMSE is considerably lower. The gap narrows somewhat toward the end of the follow-up period due to the bias noted above.

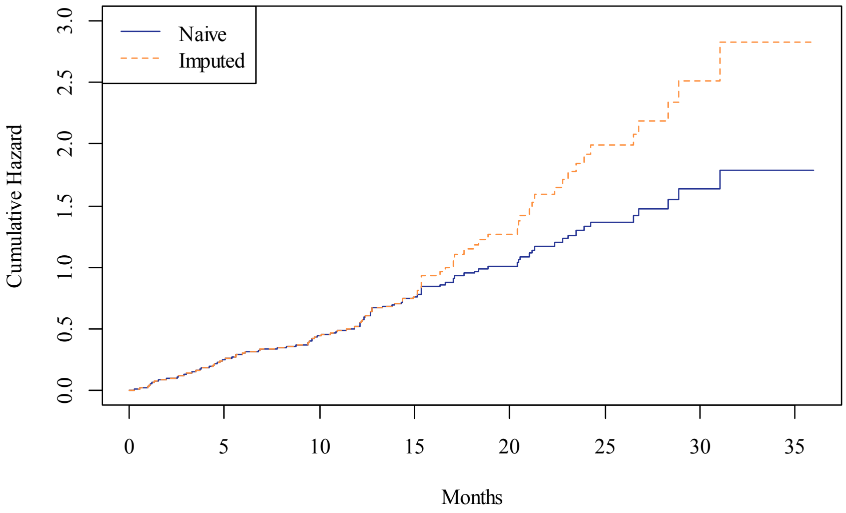

Figure 9. Cumulative hazard for hypothetical subject lost to follow-up at 15 months

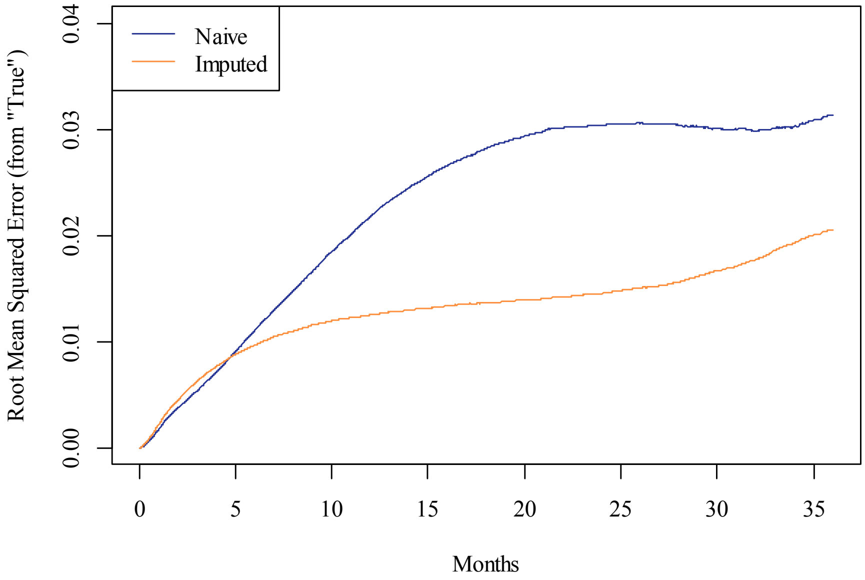

Figure 10. Average root mean squared error from “true” for imputed and naïve methods

Our goal in using the imputed method is to generate the data as it would have been observed in the absence of informative censoring. Thus, in our assessment of the method, we have taken the "true" estimate to be a sort of gold standard. We continue to use this standard as we consider estimation of the treatment effect under different trial scenarios and different censoring adjustment factors. For each trial scenario that we consider, we run 5,000 simulations. The average of the "true" treatment effect over these simulations is considered to be the best estimate we have of the treatment effect and we calculate how often the confidence intervals estimated from the imputed and naïve methods cover this estimate.

Our base trial scenario is the example trial we have used throughout. (See Table 1 for the specifications of the survival and censoring distributions.) The simulation results for this scenario are presented as scenario A1 in Table 4. Variations on the base scenario are presented in scenarios A2 through A8.

The second column of the table indicates the number of subjects per treatment arm. For most scenarios this is 200, but for A4, we double the sample size.

The third column indicates the type of end-of-study censoring. For most scenarios, we assume uniform study recruitment through 24 months and a subsequent 12 month period of follow-up without any new recruitment. Under scenario A5, there is also a 12 month period of follow-up without new recruitment, but for the first 24 months of the study we assume that recruitment starts out slow and increases or "ramps up". In this scenario we would expect higher overall rates of EOS censoring, lower rates of LTFU censoring, and fewer events since more patients would have shorter maximum follow-up times.

In all the scenarios of Table 4, the censoring adjustment factors used for imputing survival time for LTFU subjects are the same as those used in the simulation of the data. The base scenario assumes a slightly increased hazard for subjects lost to follow-up in the control group ($\alpha_c = 1.1$) and a doubling of the hazard for such subjects in the treatment group ($\alpha_t = 2.0$). Scenario A3 assumes a tripling of the hazard in the treatment group; scenario A7 doubles the hazard for the control group and triples it for the treatment group; and scenario A8 reduces the hazard by a third for the control group.

Scenarios A2, A6, and A8 employ variations on the Weibull scale parameters used for the LTFU censoring distribution. These are presented in Table 3.

Table 3. Variations to the base LTFU censoring distribution parameters

Scenario	Control	Treatment
A2, A8	σ_c = 60 months $w_c = 1.1$	σ_t = 90 months $w_t = 0.7$
A6	σ_c = 30 months $w_c = 1.1$	σ_t = 30 months $w_t = 0.7$

The resulting rates of LTFU censoring for each scenario are presented in the columns under the heading “Mean LTFU censoring”. For the base scenario, approximately 7 percent of subjects on the control arm and 14 percent on the treatment arm are censored. In scenarios A2 and A8, the treatment arm rate is increased by about 50 percent and the control arm rate is approximately doubled. In scenario A6, the treatment arm rate is approximately triple and the control arm rate is approximately quadruple their base rates. Throughout this paper we will refer to LTFU censoring with rates below 15 percent as “light”, 15-25 percent as “moderate” and above 25 percent as “heavy”. Of course, censoring above 10 percent, for example, may be considered heavy in some study settings. The terminology is only meant to distinguish the censoring scenarios that we explore in our analysis.

In Table 4, we see that the mean treatment effect for the naïve estimate is fairly consistent throughout all scenarios as we would expect. The data are simulated such that the increase in hazard occurs only after the point at which a subject is lost to follow-up so differences in the censoring patterns or the hazard ratio between LTFU subjects and others should not affect the naïve estimate. We also see that the mean treatment effect for the imputed estimate is quite close to the “true” estimate for all scenarios with the largest difference being 0.004 or about 0.4 percent for scenario A6. Given the number of simulations, we would expect our confidence interval coverage rate to be within +/- 0.006 of 0.950. Six of the eight scenarios are within this range, scenario A2 with a coverage rate of 0.942 is quite close, and the coverage of scenario A6 is 0.928. Scenario A6 has overall LTFU censoring in the 30-35 percent range, which greatly exceeds the typical amount of most trial settings. While the coverage rate for A6 is lower than we would like, it is considerably better than the 70 percent coverage rate of the naïve CI.

The imputed method CI coverage exceeds that of the naïve method for all scenarios. In addition, the widths of the imputed method CIs are also narrower than those of the naïve method. The difference ranges from 0.003 (A4) to 0.040 (A6) on the log hazard ratio scale for A1-A8.

Table 5 compares CI coverage rates for early versus late LTFU censoring. Scenario A9 uses the moderate LTFU censoring distribution of scenario A2, but increases the sample size to 400 per arm and increases the control arm CAF to 1.5. Scenario A10 differs from A9 only in the LTFU censoring distributions. The distributions for A10 were chosen to result in approximately the same amount of LTFU for each arm as scenario A9, but to have more of it occur later in follow-up. Figure 11 shows the distributions of observed LTFU censoring for the control and treatment arms for A9 (early) and A10 (late). The imputed method CI coverage for the late LTFU censoring scenario is worse than for the early, 92.4 percent versus 94.6 percent. Figure 12 shows bias and average root mean squared error by trial arm and scenario for the imputed and naïve methods. Toward the end of follow-up there is significantly more bias in the late LTFU censoring scenario than the early one for the imputed method. Again this is likely driven by the relatively fewer number of events following LTFU censoring in this scenario. In Figure 13 we can see that a significant portion of LTFU occurs beyond 27 months after which there are hardly any observed events.

Scenarios A11 and A12 have early censoring for one arm and later censoring for the other. The CI coverage is worse when the late censoring is in the control arm, which is consistent with the results presented in Figure 12, which show more bias in the control arm.

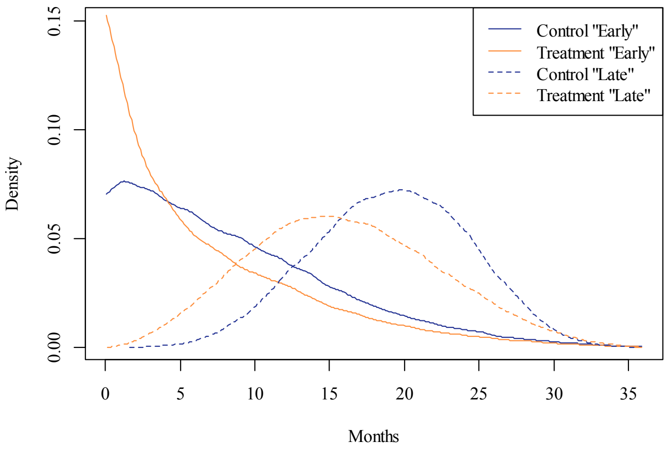

Figure 11. Early (A9) and late (A10) observed LTFU censoring distributions

Table 4. Simulation scenarios with exponential survival times

Scenario	N (per arm)	EOS censoring	Censoring adjustment factor		Mean LTFU censoring		Mean treatment effect (HR) estimate			95% CI coverage of mean "true"		
			C	T	C	T	"True"	Naïve	Imputed	"True"	Naïve	Imputed
A1	200	uniform	1.1	2.0	7%	14%	0.804	0.749	0.803	0.953	0.916	0.952
A2	200	uniform	1.1	2.0	**15%**	**22%**	0.835	0.749	0.833	0.948	0.871	0.943
A3	200	uniform	1.1	**3.0**	7%	14%	0.833	0.748	0.831	0.950	0.863	0.949
A4	**400**	uniform	1.1	2.0	7%	14%	0.802	0.748	0.802	0.952	0.875	0.950
A5	200	**ramp up**	1.1	2.0	6%	13%	0.802	0.750	0.802	0.947	0.921	0.948
A6	200	uniform	1.1	2.0	**28%**	**40%**	0.919	0.749	0.915	0.954	0.700	0.928
A7	200	uniform	**2.0**	**3.0**	7%	14%	0.807	0.748	0.807	0.951	0.910	0.945
A8	200	uniform	**0.67**	2.0	**15%**	**22%**	0.884	0.749	0.881	0.950	0.752	0.945

Abbreviations

C = control arm   T = treatment arm   EOS = end-of-study   LTFU = lost-to-follow-up   HR = hazard ratio

CI = confidence interval

Table 5. Simulation scenarios comparing early to late LTFU censoring

Scenario	N (per arm)	EOS censoring	Censoring adjustment factor		Timing of LTFU censoring		Mean LTFU censoring		Mean treatment effect (HR) estimate			95% CI coverage of mean "true"		
			C	T	C	T	C	T	"True"	Naïve	Imputed	"True"	Naïve	Imputed
A9	400	uniform	1.5	2.0	E	E	15%	22%	0.807	0.751	0.807	0.950	0.878	0.946
A10	400	uniform	1.5	2.0	L	L	14%	24%	0.797	0.750	0.799	0.945	0.895	0.924
A11	400	uniform	1.5	2.0	L	E	14%	22%	0.819	0.751	0.824	0.951	0.840	0.931
A12	400	uniform	1.5	2.0	E	L	15%	24%	0.785	0.750	0.784	0.952	0.922	0.940

Abbreviations

C = control arm T = treatment arm E = early L = late EOS = end-of-study LTFU = lost-to-follow-up HR = hazard ratio

CI = confidence interval

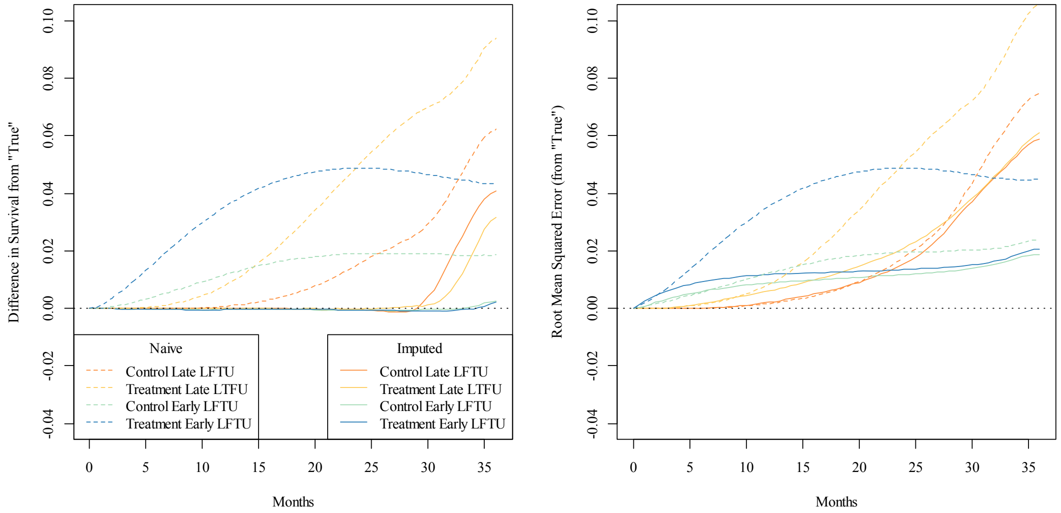

Figure 12. Bias and average RMSE for early and late LTFU censoring

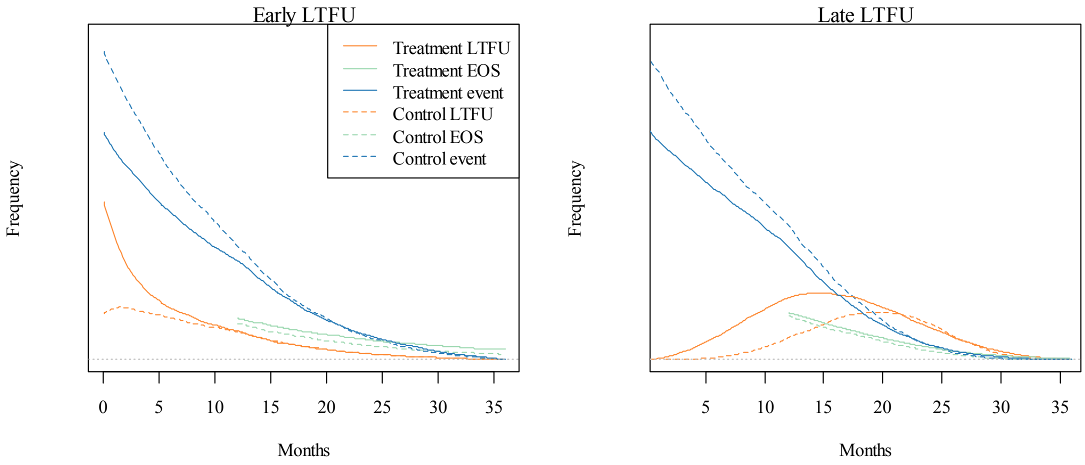

Figure 13. Distributions of events and observed censoring times by early and late LTFU

Up until this point we have considered survival times from Weibull distributions with a shape parameter of 1 (*i.e.* exponential survival). Table 6 shows the results of a number of scenarios where we use other shape parameters. There are two base scenarios: increasing hazard (WI1) and decreasing hazard (WD1). The hazards for the control and treatment arms for each of these are presented in Figure 14. All other parameters for these scenarios are the same as those of A1.

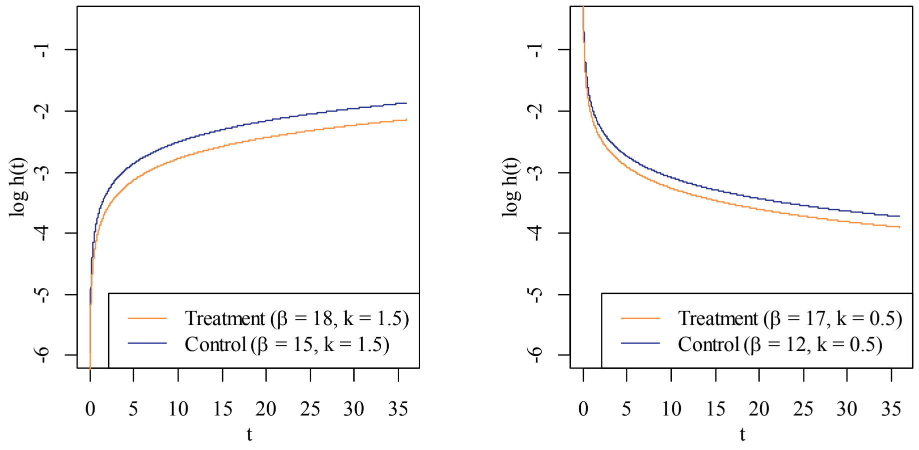

Figure 14. Weibull hazards for scenarios WI1-WI5 and WD1-WD5

The variations to the base Weibull scenarios (WI2-WI5 and WD2-WD5) are similar to the variations to the base exponential scenario. The varying characteristics of each scenario are underlined in Table 6. As with the exponential survival scenarios, the mean treatment effect for the imputed method is within a percent of the "true" mean treatment effect for all scenarios. The imputed CI coverage is also good for all but the two scenarios that have heavy LTFU censoring, and even for those the coverage is above 92 percent. The CI coverage under the imputed method is always better than that of the naïve method and in most cases considerably so.

Figure 15 shows the average difference in survival from the “true” estimate over follow-up time for the imputed and naïve methods for the treatment arm of scenarios WI3 and WD3, respectively. As with Figure 8, which presented an exponential survival scenario, imputed survival for the increasing- and decreasing-hazard Weibull scenarios with moderate LTFU censoring, on average, only departs from the “true” estimate near the maximum follow-up time, and even then the departure is slight.

Figure 16 shows average root mean squared error from “true” survival over follow-up time for four Weibull survival scenarios. The left plot shows increasing hazard scenarios WI2 (moderate LTFU) and WI3 (heavy LTFU) and the right plot shows decreasing hazard scenarios WD2 (moderate LTFU) and WD3 (heavy LTFU). RMSE increases with greater LTFU censoring for both increasing and decreasing hazard survival. This greater variability with increases in LTFU censoring likely contributes to the decrease in CI coverage that is observed in the heavy LTFU scenarios.

Table 6. Simulation scenarios with Weibull survival times

Scenario	EOS censoring	Censoring adjustment factor		Mean LTFU censoring		Mean treatment effect (HR) estimate			95% CI coverage of mean "true"		
		C	**T**	**C**	**T**	**"True"**	**Naïve**	**Imputed**	**"True"**	**Naïve**	**Imputed**
WI1	uniform	1.1	2.0	7%	14%	0.820	0.761	0.820	0.953	0.906	0.949
WI2	uniform	1.1	2.0	**16%**	**22%**	0.852	0.759	0.850	0.952	0.848	0.947
WI3	uniform	1.1	2.0	**30%**	**41%**	0.919	0.758	0.917	0.956	0.719	0.924
WI4	**ramp up**	**2.0**	**3.0**	**15%**	**22%**	0.847	0.761	0.846	0.951	0.871	0.942
WI5	uniform	**0.67**	2.0	**16%**	**22%**	0.915	0.759	0.909	0.953	0.681	0.949
	-										
WD1	uniform	1.1	2.0	6%	11%	0.877	0.835	0.877	0.954	0.938	0.951
WD2	uniform	1.1	2.0	**13%**	**18%**	0.907	0.839	0.905	0.951	0.915	0.947
WD3	uniform	1.1	2.0	**23%**	**33%**	0.971	0.840	0.968	0.953	0.796	0.935
WD4	**ramp up**	**2.0**	**3.0**	**11%**	**17%**	0.908	0.842	0.907	0.954	0.915	0.947
WD5	uniform	**0.67**	2.0	**13%**	**18%**	0.939	0.840	0.938	0.951	0.854	0.951

Abbreviations

C = control arm T = treatment arm EOS = end-of-study LTFU = lost-to-follow-up HR = hazard ratio

CI = confidence interval

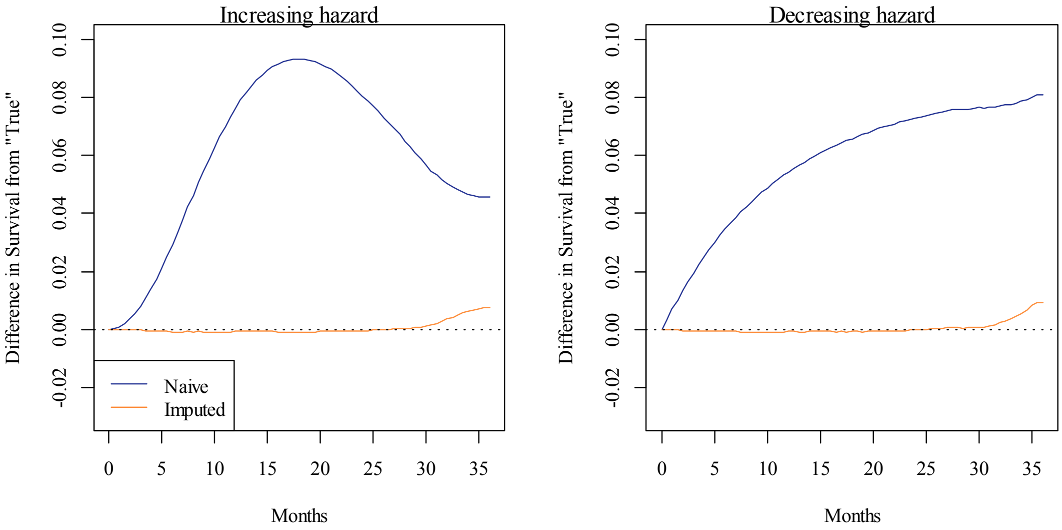

Figure 15. Average difference in survival from “true” estimate
Treatment arm (scenarios WI3 and WD3)

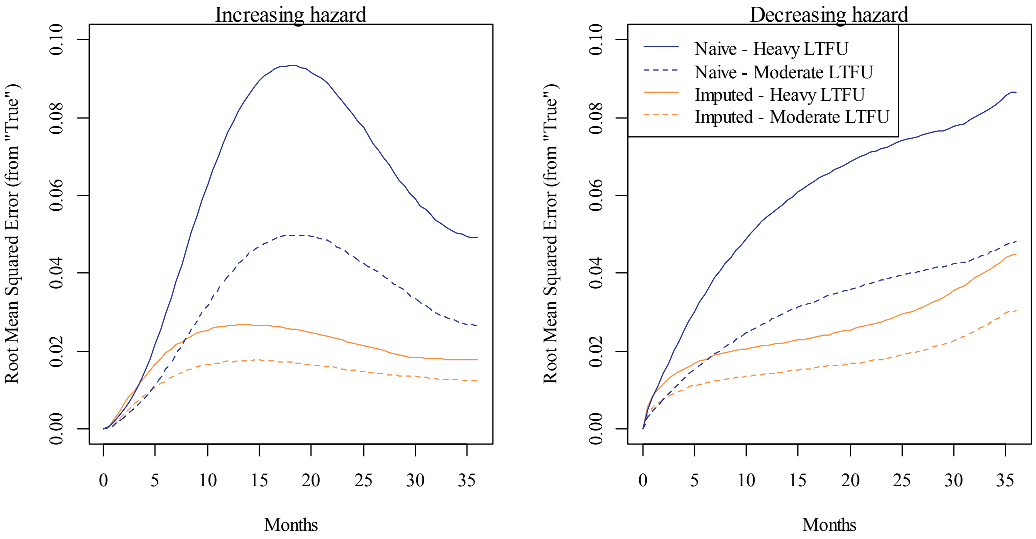

Figure 16. Average RMSE from “true” survival for imputed and naïve methods
Treatment arm (scenarios WI2, WI3, WD2 and WD3)

Chapter 4

Performance Under Non-Proportional Hazards

In the previous section, we examined the performance of the imputation method when our assumption regarding the proportionality of hazards between subjects who are lost to follow-up and those who are not holds true. In this section, we consider the performance when that assumption does not hold. We would like to know how the method performs when we know the true "average" effect of LTFU censoring (i.e. the overall hazard ratio), but not what the relationship between the hazards is at any given point in time.

To simulate non-proportional hazards we apply sinusoidal perturbations to the underlying hazard at the point of LTFU censoring. We choose this approach to investigate scenarios with varying levels of non-monotonic bounds and with differences in the association between censoring and survival for subjects who are censored at different points in time.

For a given treatment arm, we consider only sinusoidal perturbations that are "equivalent" to a given proportional hazards censoring adjustment factor (CAF) in the sense that in very large samples under the same trial parameters and when LTFU censoring is modeled as a time varying covariate, they will result in the same hazard ratio comparing LTFU subjects to other subjects. This concept is similar to the "average regression effect" discussed by Xu and O'Quigley [24]. Figure 17 shows four sinusoidal perturbations that are equivalent to the proportional hazards CAF used in the treatment arm of our example ($\alpha_t = 2$). The black line shows the underlying exponential hazard. This line becomes dashed after the point of LTFU censoring at six months. The CAF-adjusted hazard is the horizontal red line. The other curves are sinusoidal perturbations of the base exponential hazard with wavelengths of 9, 18, 36, and 72 months. Their amplitudes were chosen to achieve the above definition of equivalence. The corresponding cumulative failure distributions of the hazard functions are presented on the right hand plot of Figure 17.

Survival times for the sinusoidal perturbation scenarios are generated using the same method described in the previous section, but instead of numerically integrating over the hazard function in *(11)*, the following hazard function is used:

$$h_S(t) = I(t \leq c)\ h_W(\beta, k, t) + I(t > c)\ [\alpha\ \mathbf{exp\{Asin[2\pi/d(t\text{-}c)]\}}\ h_W(\beta, k, t)] \qquad (14)$$

where *A* and *d* are the amplitude and wavelength of the sine function and *c* is the LTFU censoring time.

Equivalent sinusoidal perturbation scenarios are identified for each treatment arm by specifying an amplitude and wavelength, generating data for a million subjects, and modeling an LTFU indicator as a time varying covariate in a Cox regression model. The "true" data, for which EOS is the only type of censoring, are the observed data for the model. The estimated coefficient for the time varying covariate is the equivalent CAF for the specified sinusoidal perturbation parameters.

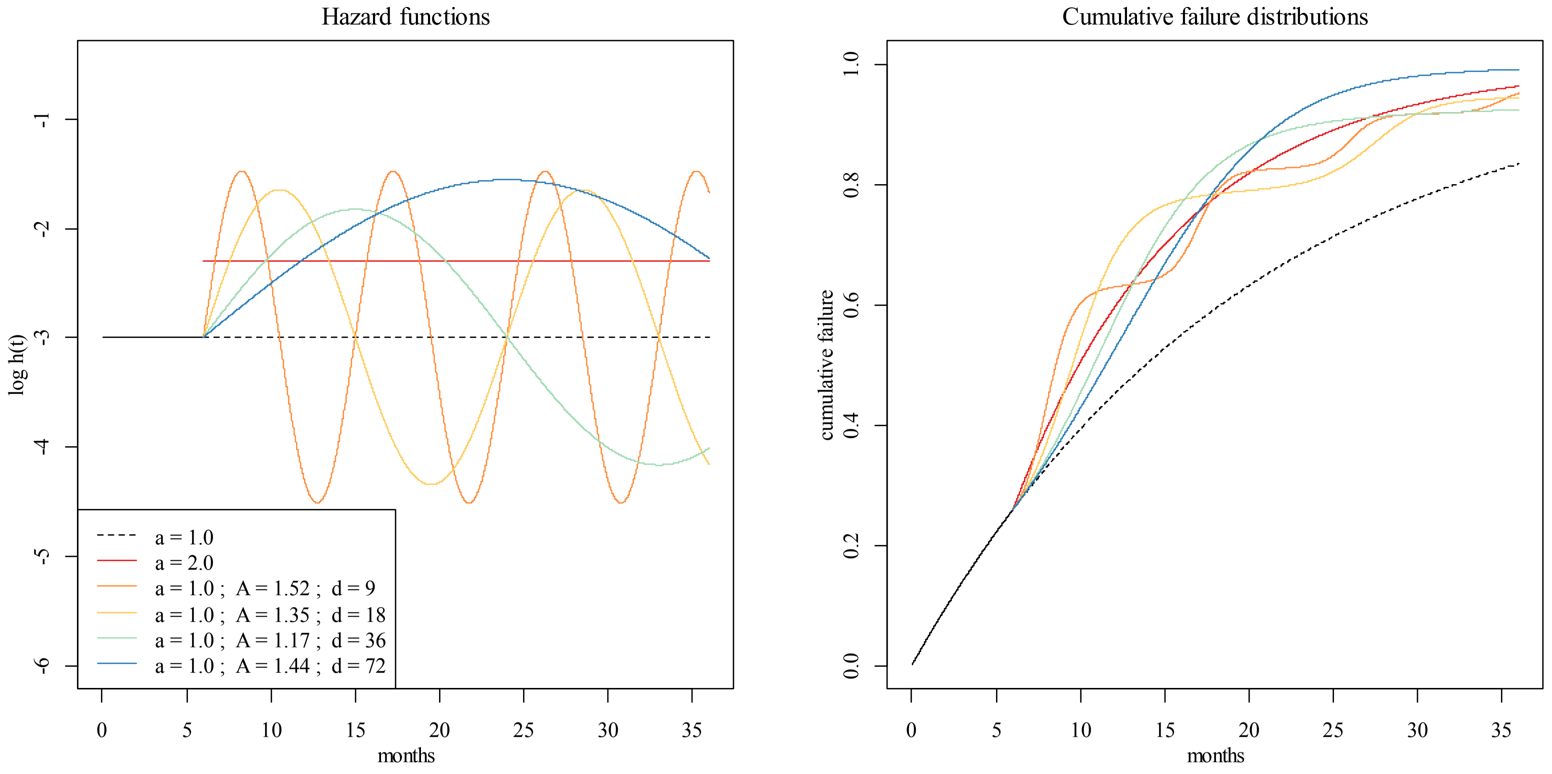

Figure 17. Sinusoidal perturbations equivalent to $\alpha_t = 2.0$

Table 7 presents the results of the sinusoidal perturbation scenarios. The trial specifications for scenarios S1-S4 vary from scenario A1 only in the amplitudes and wavelengths used to generate the treatment arm data. The control arm data generation does not employ the sinusoidal perturbations. Figure 18 shows the average difference from “true” survival over follow-up time for the treatment arm in scenarios S1-S4. In contrast to the proportional hazards scenarios of the previous section in which the average difference between imputed and “true” survival was negligible throughout most of the follow-up period, the average difference for these scenarios fluctuates between -1 and 1 percentage point. However, this variation is apparently not sufficient to affect CI coverage much. The imputed method coverage rates for the treatment effect are all within the expected precision of +/- 0.006.

Scenarios S5-S8 simulate moderate early censoring for both trial arms and various sinusoidal perturbations. For scenarios S6 and S7, the control arm has equivalent CAFs below 1 indicating that LTFU censored subjects have better survival prospects than other subjects on the control arm, while the treatment arm has equivalent CAFs greater than 2. In scenarios S5 and S8, the equivalent CAFs of the control and treatment arms are similar. The mean imputed estimates for these scenarios are within 1 percent of the “true” mean. The CI coverage rate ranges from 93.7 to 94.4 percent.

Scenarios S9 and S10 employ the late LTFU censoring used in the previous section for scenarios A10-A12. In S9, only the treatment arm has late LTFU censoring and in S10 both arms have it. The mean treatment effect of the imputed method is further from the mean “true” estimate in these scenarios than in S5-S8 and the CI coverage is somewhat worse. While the naïve CI

Table 7. Simulation scenarios with sinusoidal perturbations

	Control arm sinusoidal parameters		Treatment arm sinusoidal parameters		Censoring adjustment factor		Timing of LTFU censoring		Mean LTFU censoring		Mean treatment effect (HR) estimate			95% CI coverage of mean "true" estimate		
Scenario	**d**	**A**	**d**	**A**	**C**	**T**	**C**	**T**	**C**	**T**	**"True"**	**Naïve**	**Imputed**	**"True"**	**Naïve**	**Imputed**
S1	n/a	n/a	9	1.52	1.1	2.0	E	E	7%	14%	0.800	0.748	0.801	0.953	0.919	0.947
S2	n/a	n/a	18	1.35	1.1	2.0	E	E	7%	14%	0.801	0.748	0.801	0.952	0.918	0.947
S3	n/a	n/a	36	1.17	1.1	2.0	E	E	7%	14%	0.802	0.748	0.801	0.950	0.915	0.948
S4	n/a	n/a	72	1.45	1.1	2.0	E	E	7%	14%	0.803	0.748	0.801	0.951	0.915	0.947
S5	18	1.40	36	1.40	2.19	2.31	E	E	15%	22%	0.794	0.749	0.790	0.948	0.931	0.937
S6	18	-1.20	36	1.40	0.89	2.31	E	E	15%	22%	0.876	0.751	0.874	0.946	0.774	0.943
S7	36	-0.80	36	1.70	0.69	2.75	E	E	15%	22%	0.928	0.749	0.921	0.949	0.606	0.944
S8	36	1.20	72	1.60	2.00	2.11	E	E	15%	22%	0.788	0.749	0.786	0.948	0.934	0.938
S9	36	1.20	72	1.60	2.00	1.79	E	L	15%	24%	0.752	0.749	0.743	0.950	0.952	0.935
S10	72	0.90	18	1.20	1.29	2.00	L	L	14%	24%	0.798	0.751	0.809	0.947	0.919	0.927
S11	72	1.10	18	1.40	1.60	2.13	E	E	28%	40%	0.843	0.749	0.859	0.954	0.873	0.924
S12	72	1.10	18	1.40	1.48	2.33	L	L	14%	24%	0.808	0.750	0.817	0.953	0.911	0.935

Abbreviations

C = control arm T = treatment arm d = wave length (months) A = amplitude E = early L = late LTFU = lost-to-follow-up

HR = hazard ratio CI = confidence interval coverage is better than the imputed coverage in S9, this is likely due to the scenarios coincidentally generating “true” data that have the same relationship as the naïve data rather than evidence that the naïve consistently performs better under late LTFU censoring. In scenario S12 where the CAF and amplitude parameters are altered somewhat and there is late LTFU censoring on both arms, the CI coverage for the imputed method is similar to what it was for S9 and S10, but the naïve method coverage is significantly worse.

In scenario S11 we simulate heavy censoring and choose wavelengths for the control and treatment arms that result in opposite biases over time (see Figure 19). The mean imputed treatment effect estimate is about 2 percent higher than the mean “true” estimate suggesting some bias. The coverage rate of 92.4 percent is similar to other heavy censoring scenarios (*e.g.* A6, WI3).

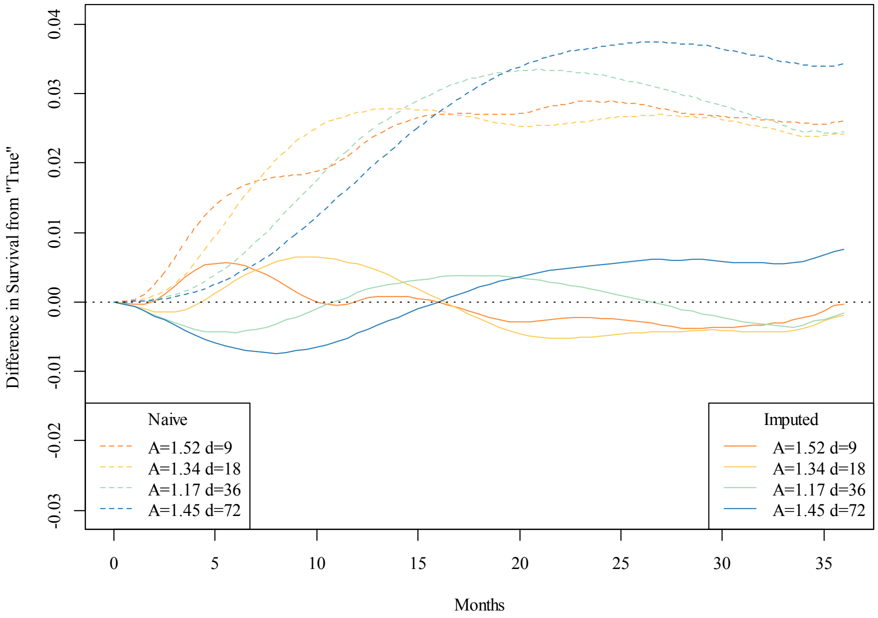

Figure 18. Average difference in survival from "true" estimate
Treatment arm (scenarios S1-S4)

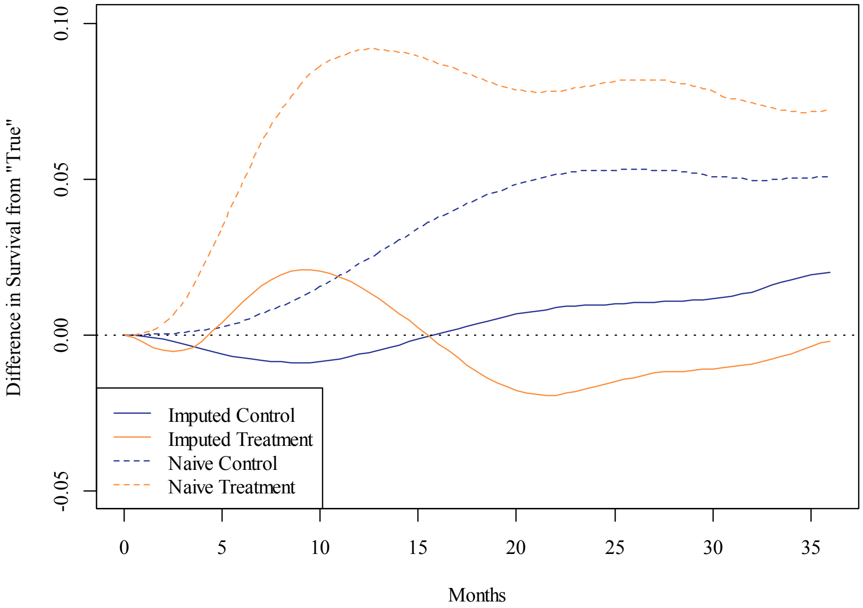

Figure 19. Average difference in survival from "true" estimate
Treatment and control arms (scenario S11)

Chapter 5

Discussion

We have considered the use of an imputation-based method to explore the sensitivity of treatment effect estimates to departures from non-informative censoring. On the basis of extensive simulations, we find the accuracy of the sensitivity analyses to be relatively unaffected by censoring rates, the degree of association between censoring and survival, and departures from the method's assumption of proportionality of hazards between LTFU and non-LTFU subjects. Under the low censoring scenarios we explored, confidence interval coverage is generally within the precision we would expect for a correct estimation procedure. Under heavy censoring and late censoring, the coverage was not ideal, but it did not fall below 92.4 percent in any of the scenarios explored.

The accuracy of the estimates appear to be most affected by late censoring, which results in the imputation of residual survival times over a period in which the estimated hazard may be zero for non-LTFU subjects. Further research could be done to explore whether adjustments to the non-LTFU hazard toward the end of follow-up can improve the accuracy of estimates. One possibility is to use the median unbiased estimate of an exponential model as the hazard beyond the last observed event. In this approach, we would use a hazard estimate on each arm such that there is 50 percent probability of observing 0 events in the cumulative residual observation time past the time of the last observed event on that arm. Specifically, letting M be the maximum time at which an event was observed for treatment arm *k*, and $PY = \Sigma\,(Y_i - M\,)\,I(Y_i > M),$ then we estimate the hazard beyond time *M* as $\lambda(t) = \; log(2) / PY,$ because under an exponential model, the probability of observing 0 events in time PY is e^{-hPY}.

The sensitivity analysis described in this paper could be readily extended to include CAF parameters for several types of LTFU censoring per treatment arm. However, as more parameters were added, the presentation and digestion of the results would become increasingly difficult and the virtue of the method's straightforwardness would erode. Simple variations such as using a few distinct values (e.g. low, medium, high) of an additional parameter could be accommodated without too much sacrifice to simplicity. LTFU types thought to be unassociated with survival could be treated in the same manner as EOS censoring and not be imputed.

The method could also be extended to incorporate covariate adjustment. The two parameter analysis could be retained with the assumption that the treatment-specific CAFs did not vary across covariate strata. Allowing for CAF variation across strata would increase complexity substantially and likely undermine the method's simplicity.

Although many authors have proposed methods for investigating the sensitivity of clinical trial results to departures from the assumption of non-informative censoring, to our knowledge, such analyses are rarely presented in regulatory submissions and publications. In addition, there is little advice in trial guidelines about how to address missing data in general or potentially informative time-to-event censoring in particular. In recognition of this deficiency, the U.S. Food and Drug Administration (FDA) has been seeking advice in developing guidelines for the treatment of missing data. As part of this effort, in 2010, the FDA created the Panel on the

Handling of Missing Data in Clinical Trials. Among other advice, the Panel recommended the following.

(1) Statistical methods for handling missing data should be specified by clinical trial sponsors in study protocols, and their associated assumptions stated in a way that can be understood by clinicians.

(2) Sensitivity analyses should be part of the primary reporting of findings from clinical trials. Examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting.

One of the virtues of the sensitivity analysis considered in this paper is its simplicity. The method can be boiled down to the following: *multiply impute survival times for LTFU-censored patients assuming an "average" relative increase or decrease in their hazard compared to the hazard of subjects that are not LTFU censored*. The sensitivity analysis can span a wide range of assumptions regarding the "average" proportion that the hazard is increased or decreased and its results can be easily summarized in a single graphical presentation. In addition, the method is relatively unaffected by departures from its proportional hazards assumption.

However, there does remain the challenge of determining a plausible range for the parameters. As noted in the background chapter, there are many possible reasons for a subject being LTFU, some of which may tend to be associated with longer survival, some shorter, and some not associated with survival at all. Investigators and regulators would have to have some sense of how the particular mix of reasons for LTFU for a trial translates into an "average effect" hazard ratio.

The interpretability of the sensitivity parameters and the selection of plausible ranges for the parameters are a challenge in any sensitivity analysis. The parameters used here would seem to be as easily interpreted as any of those noted in Chapter 1.

The lack of readily available software for executing sensitivity analyses is another likely barrier to exploring departures from non-informative censoring. The methods described in this paper are easy to implement and the required inputs vary little from those of standard survival methods. The user need only specify plausible ranges for the treatment specific CAFs, provide a maximum follow-up time for subjects who are LTFU, and redefine the usual censoring indicator such that LTFU censoring is distinguishable from EOS censoring. In our modeling we use $\delta = 0$ for an event, $\delta = 1$ for EOS censoring, and $\delta = 2$ for LTFU censoring.

In the estimation of a given scenario, we have the modest goal of imputing the data as it might have been observed in the absence of informative censoring. We do not attempt to impute administratively censored subjects or to impute survival beyond LTFU subjects' subject-specific maximum follow-up period. This allows us to minimize assumptions regarding the survival distribution and to employ standard survival analysis methods once the data have been imputed.

The limitations of the method's accuracy under heavy and late LTFU censoring must be weighed against the limitations of the status quo in which sensitivity analyses are not conducted at all. With this method we have a straightforward approach that is easy to implement, has reasonably accurate estimates under a wide range of conditions, and can provide valuable insight as to the robustness of treatment effect estimates under departures from non-informative censoring.

References

[1] D. O. Scharfstein and J. M. Robins, “Estimation of the failure time distribution in the presence of informative censoring,” *Biometrika*, vol. 89, no. 3, pp. 617–634, 2002.

[2] A. Rotnitzky, A. Farall, A. Bergesio, and D. Scharfstein, “Analysis of failure time data under competing censoring mechanisms,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 3, pp. 307–327, 2007.

[3] K. M. Leung, R. M. Elashoff, and A. A. Afifi, “Censoring issues in survival analysis,” *Annual Review of Public Health*, vol. 18, pp. 83–104, 1997.

[4] A. Tsiatis, “A nonidentifiability aspect of the problem of competing risks,” *Proceedings of the National Academy of Sciences*, vol. 72, no. 1, p. 20, 1975.

[5] E. L. Kaplan and P. Meier, “Nonparametric Estimation from Incomplete Observations,” *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[6] N. Mantel, “Evaluation of survival data and two new rank order statistics arising in its consideration.,” *Cancer Chemother Rep*, vol. 50, no. 3, pp. 163–70, Mar. 1966.

[7] D. R. Cox, “Regression Models and Life-Tables,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, Jan. 1972.

[8] S. W. Lagakos, “General right censoring and its impact on the analysis of survival data,” *Biometrics*, pp. 139–156, 1979.

[9] A. V. Peterson, “Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks,” *Proceedings of the National Academy of Sciences*, vol. 73, no. 1, pp. 11 –13, Jan. 1976.

[10] J. Robins, “A new approach to causal inference in mortality studies with a sustained exposure period--application to control of the healthy worker survivor effect,” *Mathematical Modelling*, vol. 7, no. 9–12, pp. 1393–1512, 1986.

[11] J. M. Robins and A. Rotnitzky, “Recovery of information and adjustment for dependent censoring using surrogate markers,” *Aids Epidemiology, Methodological issues*, pp. 297–331, 1992.

[12] J. M. Robins, “Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers,” in *Proceedings of the Biopharmaceutical section, American Statistical Association*, 1993, pp. 24–33.

[13] G. A. Satten, S. Datta, and J. Robins, “Estimating the marginal survival function in the presence of time dependent covariates,” *Statistics & probability letters*, vol. 54, no. 4, pp. 397–403, 2001.

[14] L. Fisher and P. Kanarek, “Presenting censored data when censoring and survival times may not be independent,” in *Reliability and Biometry, Statistical Analysis of Lifelength*, Philadelphia: SIAM, 1974, pp. 303–326.

[15] S. W. Lagakos and J. S. Williams, “Models for censored survival analysis: A cone class of variable-sum models,” *Biometrika*, vol. 65, no. 1, pp. 181 –189, Apr. 1978.

[16] E. V. Slud and L. V. Rubinstein, “Dependent competing risks and summary survival curves,” *Biometrika*, vol. 70, no. 3, pp. 643 –649, Dec. 1983.

[17] J. P. Klein and M. L. Moeschberger, “Bounds on net survival probabilities for dependent competing risks,” *Biometrics*, pp. 529–538, 1988.

[18] M. Zheng and J. P. Klein, “Estimates of marginal survival for dependent competing risks based on an assumed copula,” *Biometrika*, vol. 82, no. 1, pp. 127–138, 1995.

[19] D. Scharfstein, J. M. Robins, W. Eddings, and A. Rotnitzky, “Inference in Randomized Studies with Informative Censoring and Discrete Time-to-Event Endpoints,” *Biometrics*, vol. 57, no. 2, pp. 404–413, 2001.

[20] F. Siannis, J. Copas, and G. Lu, “Sensitivity analysis for informative censoring in parametric survival models,” *Biostatistics*, vol. 6, no. 1, pp. 77–91, 2005.

[21] J. Zhang and D. F. Heitjan, “Nonignorable censoring in randomized clinical trials,” *Clinical Trials*, vol. 2, no. 6, pp. 488–496, 2005.

[22] T. Liu and D. F. Heitjan, “Sensitivity of the discrete-time Kaplan-Meier estimate to nonignorable censoring: Application in a clinical trial,” *Statistics in Medicine*, vol. (In press), 2011.

[23] D. B. Rubin and R. J. A. Little, “Statistical analysis with missing data,” *Hoboken, NJ: J Wiley & Sons*, 2002.

[24] R. Xu and J. O’Quigley, “Estimating average regression effect under non-proportional hazards,” *Biostatistics*, vol. 1, no. 4, pp. 423–439, 2000.