2021

Innovative Statistical Models in Cancer Immunotherapy Trial Design

Jing Wei
University of Kentucky, wj0514@gmail.com

Digital Object Identifier: https://doi.org/10.13023/etd.2021.279

Innovative Statistical Models in Cancer Immunotherapy Trial Design

DISSERTATION

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Arts and Sciences at the University of Kentucky

By
Jing Wei

Lexington, Kentucky

Co-Directors: Dr. Arnold Stromberg, Professor of Statistics
and Dr. Jianrong Wu, Professor of Markey Cancer Center

Lexington, Kentucky

2021

ABSTRACT OF DISSERTATION

Innovative Statistical Models in Cancer Immunotherapy Trial Design

A challenge arising in cancer immunotherapy trial design is the presence of non-proportional hazards (NPH) patterns in survival curves. We considered three different NPH patterns caused by delayed treatment effect, cure rate and responder rate of treatment group in this dissertation. These three NPH patterns would violate the proportional hazard model assumption and ignoring any of them in an immunotherapy trial design will result in substantial loss of statistical power.

In this dissertation, four models to deal with NPH patterns are discussed. First, a piecewise proportional hazards model is proposed to incorporate delayed treatment effect into the trial design consideration. Second, we consider a piecewise proportional hazard model with cure rate to deal with both delayed treatment effect and cure rate. Third, we extended the second model as a general random delayed cure rate model in cancer immunotherapy trials design. Fourth, we proposed a piecewise proportional hazard responder rate model to deal with both delayed treatment effect and responder rate. Sample size formulas are derived for weighted log-rank tests under a fixed alternative hypothesis under various models. The accuracy of sample size calculation using the new formulas are assessed and compared with the existing methods via simulation studies. The sensitivities for mis-specifying the random delay time are also studied through simulations. What is more, a real immunotherapy trial is used to illustrate the study design along with practical consideration of balance between sample size and follow-up time in second model.

KEYWORDS: clinical trial, non-proportional hazards, delayed treatment effect, cure rate, responder and non-responder, weighted log-rank test

Jing Wei

July 26, 2021

Date

Innovative Statistical Models in Cancer Immunotherapy Trial Design

By

Jing Wei

Arnold Stromberg
Co-Director of Dissertation

Jianrong Wu
Co-Director of Dissertation

Katherine Thompson
Director of Graduate Studies

July 26, 2021
Date

This work is dedicated to my father Zhaorong Wei, my mother Yufeng Sun, my husband Shu Gu, my son Winston Juntong Gu, my brother Peng Wei, my two chinchillas Fishball and Lucky, my two cats Xiaohuihui and Xiaobaibai.

ACKNOWLEDGMENTS

During my Ph.D. journey, I have had incredible help and support from many faculties and friends. Here, it is my great pleasure to express my appreciation for their assistance in obtaining my degree and living in a happy life in Lexington.

First and foremost, to my amazing advisor Dr. Jianrong Wu, thank you for introducing me to the field of clinical trial, which opened a new window and changed my entire career path. This work would not be possible without your enormous patience, guidance, suggestions and encouragement. Your enthusiasm and attitude influenced me remarkably and will continue to guide me throughout my life.

Also, I am truly thankful to other members in my dissertation committee, Dr. Arnold J. Stromberg, Dr. Chi Wang, Dr. Derek S. Young, Dr. William S. Griffith and Dr. Brent Shelton, for your helpful comments and serving on my thesis committee.

In particular, I would like to thank Jing Liu, Jing Wu, Anqi Guo, Ding Zhao, Di Liang and Yucong Sang for being my ”family” in US, you've always been my rocks and ready to hug me whenever I need. To my friends Shuyi Zou and Kejia Ji from overseas, although apart from you, our friendship will never fade out.

I would also love to thank all my friends in Statistics Department, Xiaoli Kong, Tingting Zhai, Yan Xu, Zhang Xu, Baoying Yang, Pei Wang, Ting Zeng and Yue Cui, thank you all for helping me in statistical courses, planing career, and getting over difficulties. You all are fantastic and awesome. I will miss comprehensive exam solutions explaining, delicious home cooked foods, fun play date time and the days of chatting and drinking Starbucks together.

Last, but by no means least, I would like to express my sincere appreciation to all my family. My husband Shu Gu, thank you for your tremendous love and unconditional

support, especially in taking care of our son Winston Juntong Gu. My parents and brother, thank you for your constant encouragement and supporting in all my decision making.

Thank you all for your substantial part in my success.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

1.1 Immuno-oncology and immunotherapy

Cancer immunotherapy, also known as immuno-oncology, is a form of cancer treatment that use the power of the body′s own immune system to prevent, control, and eliminate cancer. Instead of poisoning a tumor or destroying it with radiation, the immune system is educated to attack ′foreign′ cells but at the same time leave healthy, self-tissues alone. Based on this characteristic, immunotherapy may have fewer side effects compared with chemotherapy. Also, the immune system learns to go after cancer cells if they return, which means some patients can get long term survival after treatment. What is more, some cancers, such as skin cancer, do not respond well to chemotherapy or radiation, but may respond well to immunotherapy. These benefits make immunotherapy a powerful tool in oncology during recent years.

There are several types of cancer immunotherapy (Smith, Smith), some types of immunotherapy boost your disease-fighting powers overall. Others teach it to attack specific kinds of cells found in tumors.

Checkpoint Inhibitors

Immune system usually uses checkpoints, which is a system of ”brakes”, to stop it from attacking your own healthy cells when attacked by invaders like bacteria and viruses. Cancer cells sometimes turn these checkpoints on or off so they can hide themselves. Checkpoint inhibitors are drugs that release the brakes on your immune system. In general, they stop the proteins on the cancer cells from pushing the stop button. This turns the immune system back on and the T cells are able to find and attack the cancer cells.

Seven of these drugs are approved by FDA, like PD-1 inhibitors included Pembrolizumab (Keytruda), Nivolumab (Opdivo) and Cemiplimab (Libtayo); PD-L1 inhibitors inclded Atezolizumab (Tecentriq), Avelumab (Bavencio) and Durvalumab (Imfinzi); CTLA-4 inhibitor included Ipilimumab (Yervoy). These drugs block the proteins PD-1, PD-L1, and

CTLA-4 on the surface of immune cells, to let these cells go after the cancerous growth.

Adoptive T cell therapies

Adoptive T cell therapies include a number of different types of immunotherapy treatments. They all use immune cells that are grown in the lab to large numbers followed by administering them to the body to fight the cancer. Sometimes, immune cells that naturally recognize cancer cells are used, while other times they are modified to make them recognize and kill the cancer cells.

There are several types of Adoptive T cell therapies such as Tumor-Infiltrating Lymphocytes (TIL) therapy, Engineered T-cell (TCR) therapy, CAR T-cell therapy and Natural Killer (NK) cell therapy. TIL therapy is the treatment that T-cells are grown from the tumor itself, TCR therapy is the treatment that tumor-specific T-cells are grown from the blood. CAR T-cell therapy is the treatment that a chimeric antibody/T-cell receptor gene is put into peripheral T-cells and NK cell therapy is the treatment that add CARs to NK cells helps them target the cancer better.

Monoclonal Antibodies

Antibodies actually are proteins made by immune system. They find and stick to other proteins called antigens on cancer cells and then recruit other parts of your immune system to destroy the cancer. Monoclonal antibodies is the antibodies made in the lab. In general, they are engineered versions of immune system proteins designed to attack specific parts of cancer cells.

Naked monoclonal antibodies are the most common type used in cancer treatment. These antibodies are unattached to anything and boost your immune system's response against the cancer, or block antigens that help the cancer grow and spread.

Conjugated monoclonal antibodies usually attach a chemotherapy drug or radioactive particle and effect directly to cancerous cells. It reduces side effects and helps chemotherapy and radiation treatments work better.

Bispecific monoclonal antibodies can attach to two proteins at once, for example can attach to both a cancer cell and an immune cell, which helps the immune system attack the cancer.

Cancer Vaccines

Cancer vaccines are made from dead cancer cells, proteins or pieces of proteins from cancer cells, or immune system cell, these substances put in the body to activate an immune response against certain types of cancer.

FDA has approved three vaccines to treat cancer. Sipuleucel-T (Provenge) treats advanced prostate cancer when hormone therapy doesn't work; Talimogene laherparepvec (T-VEC) treats melanoma skin cancer that has spread and Bacillus Calmette-Guérin, or BCG, treats early-stage bladder cancer.

1.2 Proportional hazards model

Most traditional time-to-event clinical trials are designed and analyzed using proportional hazards assumption and log-rank test. Let $S_1(t)$, $S_2(t)$, and $\lambda_1(t)$, $\lambda_2(t)$ be the hazard functions and survival functions for the control and treatment groups, respectively. The hazard functions of two groups satisfy the proportional hazards model which can be written as

$$\lambda_2(t) = \delta\lambda_1(t),$$

or equivalently, the survival distributions of the two groups satisfy

$$S_2(t) = [S_1(t)]^{\delta},$$

where δ is a constant hazard ratio over time, which is a measurement of treatment effect in survival curve between treatment group and control group. Figure 1.1 shows survival function and hazard function between control and treatment groups, the survival function of control group follows Weibull distribution with shape parameter $\kappa = 1.2$ and hazard ratio $\delta = 0.7$.

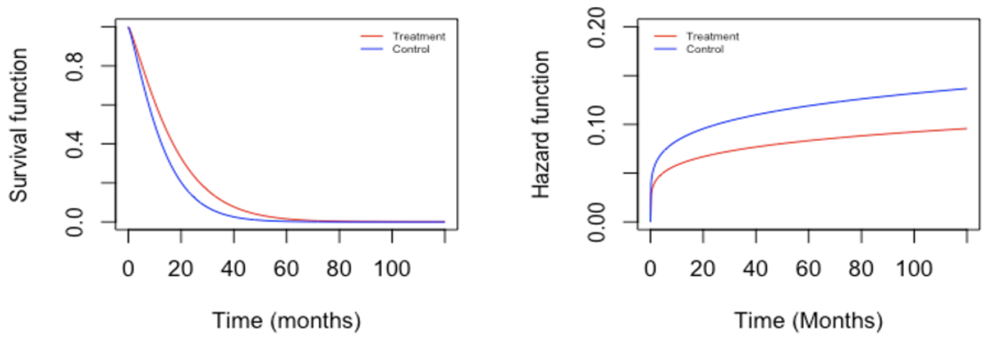

Figure 1.1: Survival function and hazard function for two groups

1.3 Log-rank test

A two-sided hypothesis for testing the difference between survival distributions of the experimental treatment group and control group is represented by

$$H_0 : S_2(t) = S_1(t) \quad \text{vs.} \quad H_1 : S_2(t) \neq S_1(t).$$

Under the PH model $\lambda_2(t) = \delta\lambda_1(t)$, this hypothesis is equivalent to the following hypothesis for the hazard ratio:

$$H_0 : \delta = 1 \quad \text{vs.} \quad H_1 : \delta \neq 1. \tag{1.1}$$

The log-rank test is a well-known optimal statistic to test the above hypothesis. To introduce the log-rank test, we assume that the unique and ordered failure times for two groups are denoted by $t_1 < t_2 < \cdots < t_k$, let d_{1j} be the number of failures and n_{1j} be the number at risk in control group at time t_j. Let d_{2j} and n_{2j} be the corresponding numbers for treatment group. Thus, there are $d_j = d_{1j} + d_{2j}$ failure in both groups at t_j and a total of $n_j = n_{1j} + n_{2j}$ is the number at risk in both groups at t_j, and $e_{1j} = n_{1j}d_j/n_j$ is the expected number of failure at t_j for the control group. It is well known that the log-rank score statistic

$$U = \sum_{j=1}^{k}(d_{1j} - e_{1j})$$

is an asymptotically normally distributed with mean zero under the null hypothesis and its asymptotic variance can be estimated by

$$V = \sum_{j=1}^{k} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}.$$

The log-rank test is then given by

$$L = \frac{U}{\sqrt{V}}. \tag{1.2}$$

Log-rank test is most commonly used for survival endpoint as well as sample size and power calculation, since log-rank test is asymptotically the most powerful test when the proportional hazards assumption holds. Based on log-rank test formula (1.2), Schoenfeld (Schoenfeld, 1981) proposed a sample size calculation method under a local alternative assumption. Given a two-side type I error of α, the study power of $1 - \beta$, ω_1 and ω_2 are the proportions of subjects assigned to treatment and control groups, and P is the overall failure probability of two groups, then the total sample size n for the two groups is given by

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\omega_1 \omega_2 [\log(\delta)]^2 P},$$

and the total number of events is given by

$$d = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\omega_1 \omega_2 [\log(\delta)]^2}. \tag{1.3}$$

This number of events formula (1.3) is widely used in trial design since it is robustness against the design parameters. This means there is no need to specify any assumption such as censoring distribution or accrual distribution. Power of the study only depends on the number of events observed, it is also called event-driven in trial design.

1.4 Non-proportional hazards pattens in immunotherapy trial design

In this thesis, we focus on cancer immunotherapy trial design and now the question is that can proportional hazards assumption and log-rank test also be used and performed well for cancer immunotherapy trail design?

Figure 1.2 is the study of Sipuleucel-T (Kantoff et al., 2010), the first therapeutic cancer vaccine approved by FDA. This study shows a delayed separation of survival curves in

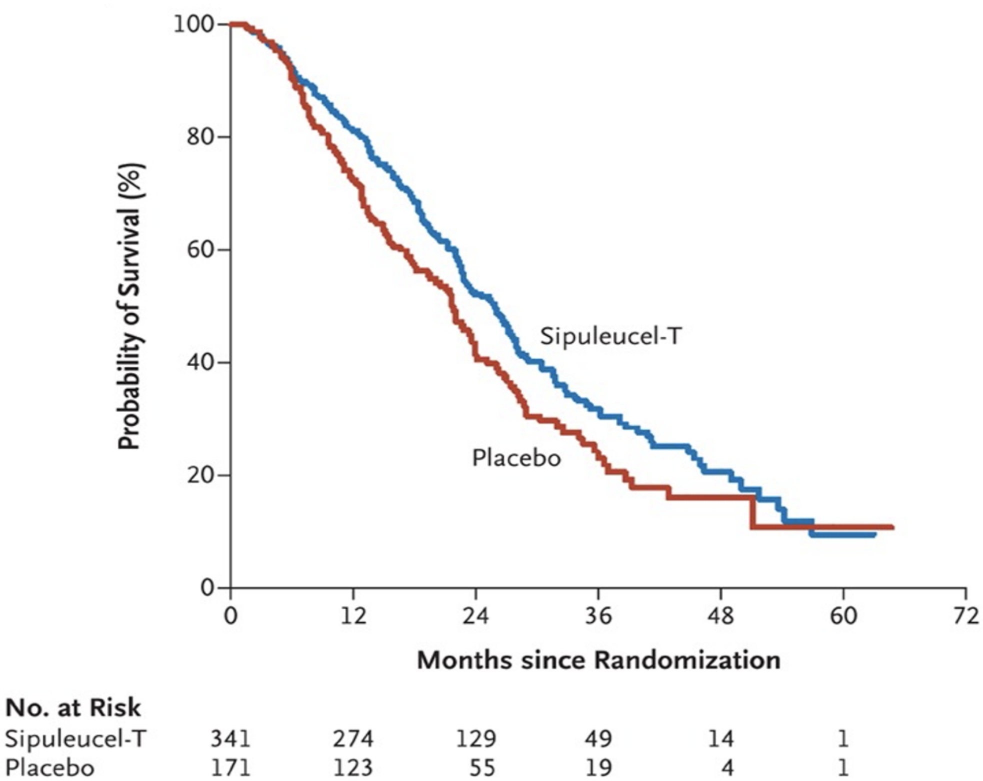

Figure 1.2: Kaplan Meier estimates of overall survival curve for Sipuleucel-T study

Kaplan Meier plot, which means there is no difference between placebo and Sipuleucel-T treatment group after randomization, then the curves separate after around 6 months. This delayed patten is largely caused by the indirect mechanism of action of the vaccine, which requires time to mount an effective immune response and time for that response to be translated into an observable clinical response.

Since immunotherapies are very effective, a proportion of patients will have long term survival, some patients will be cured after treatment. This is another typical feature in immunotherapy trial. Coiffier conducted a study of a randomized trial to compare CHOP plus Rituximab with CHOP alone in elderly patients with diffuse large-B-cell lymphoma (Coiffier et al., 2002). Figure 1.3 is the survival curves of control and treatment groups have a plateau at the end of the study.

Robert conducted a randomized Phase III immunotherapy trial for untreated metastatic melanoma (Robert et al., 2011) in figure 1.4. Patients were randomly assigned to receive either Ipilimumab plus dacarbazine or dacarbazine plus placebo. Delayed treatment effect

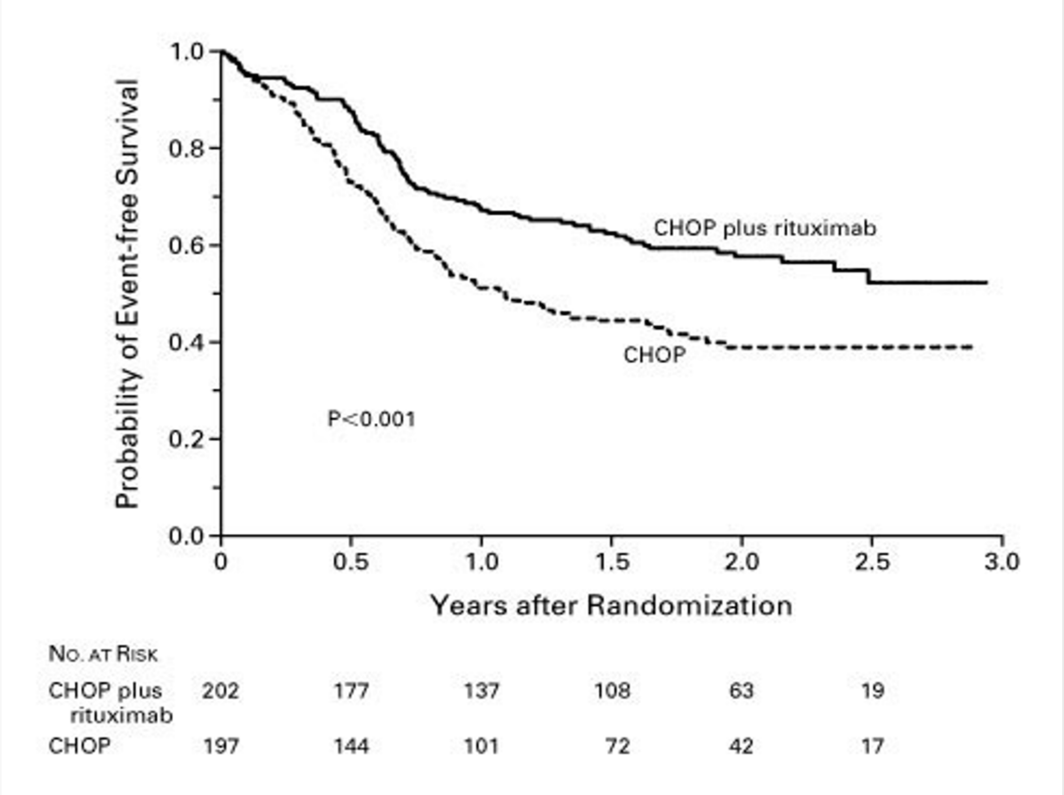

Figure 1.3: Event-free Survival among Patients assigned to Chemotherapy with Cyclophosphamide, Doxorubicin, Vincristine, and Prednisone (CHOP) or with CHOP plus Rituximab.

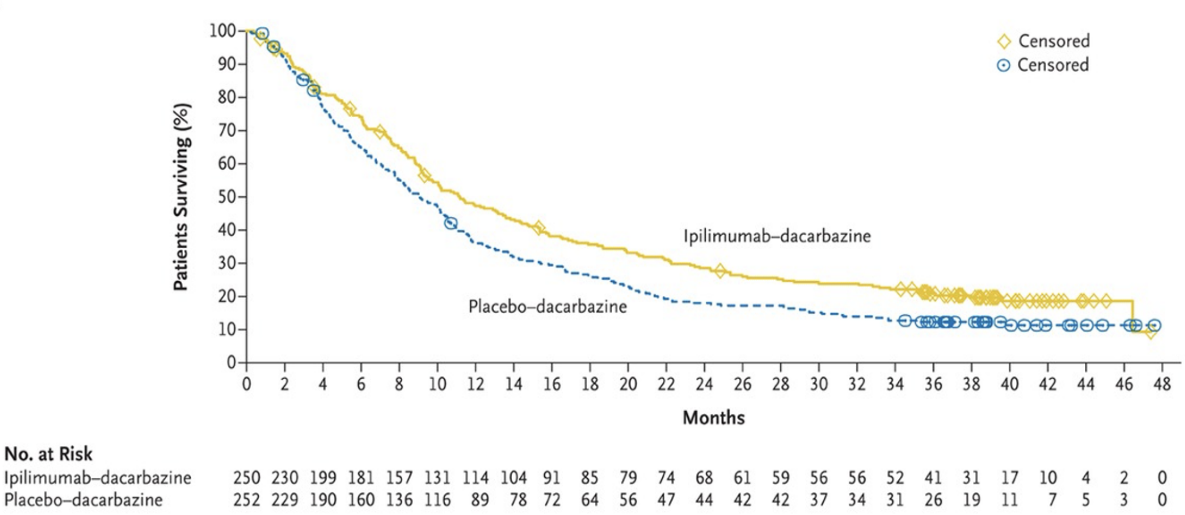

Figure 1.4: Overall survival of Robert's study for previously treated metastatic melanoma.

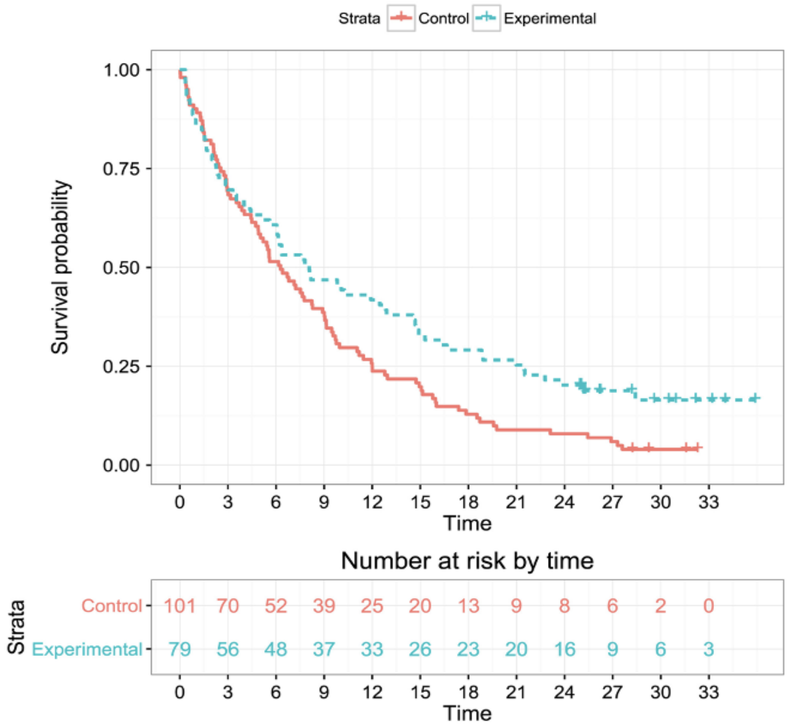

Figure 1.5: K-M plot of the random delayed treatment effect scenario with random time lag.

and cure rate appear in this study together.

The delayed treatment effect in figure 1.2 and figure 1.4 happened at a fixed point, there is no difference between survival distributions of control and treatment groups. After fixed delayed time point, the survival distribution curves seperated. However, in practice each patient may get different response to the same therapy based on individual biological manner, and the duration of treatment effect time may vary heterogeneously from subject to subject rather than fix at a constant. Xu illustrated a random delayed treatment effect model (Xu et al., 2018) in which the treatment effect time follows a random variable on an interval instead of a fixed time point. Figure 1.5 is the Kaplan-Meier plot in Xu's paper, which is generated using a synthetic dataset simulated based on a confidential real study. This figure illustrates that random delayed pattern follows uniform distribution between 3 and 12 months, which means the two survival curves will not separate until 3 months, then gradually separate at an increasing hazard ratio until 12 months, and remain at a constant hazard ratio after 12 months.

Delayed treatment effect and cure rate are two features in cancer immunotherapy trial design, both of these two features imply violate proportional hazards assumption, using standard sample size and power calculation methods based on log-rank test that would lead to a loss of power. So in this thesis we focus on how to deal with such kind of non-proportional hazards models, how to choose test statistics and how to calculate sample size during trial design.

1.5 Summary

The dissertation is organized in six chapters. In Chapter 2, we introduce a piecewise weighted log-rank test to incorporate the delayed treatment effect into the trial design and derive a new sample size under a fixed alternative hypothesis for the delayed treatment effect model.

Chapter 3 extends the delayed treatment effect model in Chapter 2. Here, we proposed a piecewise proportional hazard cure rate model to incorporate both delayed treatment effect and cure rate into the trial design consideration. Same as Chapter 2, the sample size formula is derived under a fixed alternative hypothesis. Chapter 3 also includes a real immunotherapy trial to illustrate the study design along with practical consideration to balance between sample size and follow-up time.

Chapter 4 is concerned with general random delayed cure rate model to design cancer immunotherapy trials. This kind of model considers the case when delayed treatment effect is not happened at a fixed point and illustrates that duration of lag is more suitable to be treated as an interval rather than a fixed constant. The sensitivity for mis-specifying random delayed time is also studied through simulations.

A novel design is proposed in Chapter 5 to deal with the dichotomized response incurred from non-responders in treatment group. How to find the weight function is the key point for such kind of NPH pattern. Sample size and empirical power are compared with existing weight functions via simulation studies.

Chapter 6 is the summary of four models from Chapters 2-5 and also discuss the future work. Appendices containing the proofs and other technical details are included after

Chapters 6.

Chapter 2 Delayed Treatment Effect

2.1 Introduction

In recent years, immunotherapies have been increasingly used for treating relapse or advanced-stage cancer patients. Because of the indirect mechanism of action of immunotherapy, it takes time for an immune outcome to be elicited and translated into a clinical outcome. Hence, a delayed treatment effect is often seen in immunotherapy trials wherein survival curves show no effect during the initial part of the study and evidence appears only later in the study. For example, the cancer vaccine trial of sipuleucel-T showed delayed separation of survival curves by 6 months (Kantoff et al., 2010). These findings suggest that the proportional hazard (PH) assumption no longer holds true in such cases, and using conventional sample size and power calculation methods (Schoenfeld, 1981; Freedman, 1982) based on the standard log-rank test will lead to substantial loss of statistical power. Various methods based on weighted log-rank tests have been proposed to increase the efficiency of designing clinical trials with a delayed treatment effect. For example, Lakatos (Lakatos, 1988) considered the Tarone-Ware class of weights to design clinical trials with a delayed treatment effect. Fine (Fine, 2007) and Hasegawa (Hasegawa, 2014) presented similar methods for calculating sample sizes with the Fleming-Harrington $G^{\rho,\gamma}$ class of weights(Fleming and Harrington, 1991), however they were not optimal to maximize statistical power under the delayed treatment effect model.

Recently, Xu et al. (Xu et al., 2016) showed that the piecewise weighted log-rank test was optimal for cases of delayed onset of treatment effect and derived sample size and power calculations for the piecewise weighted log-rank test under a sequence of local alternative hypotheses. However, in practice, the alternative hypothesis is always fixed and does not change as sample size increases. Thus, the accuracy of the sample size formula derived under the local alternative needs to be carefully assessed by simulations.

This Chapter is organized as follows. Section 2.2 introduces a piecewise weighted log-rank test and section 2.3 derives a new sample size formula under a fixed alternative hypothesis for the delayed treatment effect model based on section 2.2. The accuracy of the formula derived under local vs fixed alternative is compared for both balanced and unbalanced designs showed in section 2.4 and a real example is in section 2.5. Section 2.6 contains discussions and conclusions.

2.2 Piecewise weighted log-rank test

A two-sided hypothesis for testing the difference between survival distributions of the experimental treatment group and control group is represented by

$$H_0 : S_2(t) = S_1(t) \quad \text{vs.} \quad H_1 : S_2(t) \neq S_1(t),$$

where labels 1 and 2 represent control and treatment groups, respectively. Under the PH model $\lambda_2(t) = \delta\lambda_1(t)$, where $\lambda_1(t)$ and $\lambda_2(t)$ are the hazard functions of the control and treatment groups, respectively, and δ is the hazard ratio between the treatment and control groups, this hypothesis is equivalent to the following hypothesis for the hazard ratio:

$$H_0 : \delta = 1 \quad \text{vs.} \quad H_1 : \delta \neq 1. \tag{2.1}$$

The log-rank test is a well-known optimal statistic to test the above hypothesis. To introduce the weighted log-rank test, consider a study that compares survival curves with n subjects randomly allocated to the control or treatment group, with probability ω_1 and ω_2 ($\omega_1 + \omega_2 = 1$), respectively. Let D be the set of indices of subjects who experience the event of interest. At each distinct event time $t_j, j \in D$, let d_{1j} and d_{2j} be the number of events occurring at time t_j for the control and treatment groups, respectively, with n_{1j} and n_{2j} subjects being at risk in the two groups just before t_j, for $j \in D$. Thus, there are $d_j = d_{1j} + d_{2j}$ events at t_j among a total of $n_j = n_{1j} + n_{2j}$ subjects, and $e_{1j} = n_{1j}d_j/n_j$ is the expected number of events at t_j for the control group. Let w_j be the weight at each distinct event time t_j and all w_j are nonnegative weights, it is well known that the weighted log-rank score statistic

$$U = \sum_{j \in D} w_j(d_{1j} - e_{1j}),$$

is an asymptotically normally distributed with mean zero under the null hypothesis and its asymptotic variance can be estimated by

$$V = \sum_{j \in D} w_j^2 \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}. \tag{2.2}$$

In cases where a delayed treatment effect occurs, let t_0 denote the hazard ratio changing time point, which measures the duration of the delayed treatment effect since randomization. This delayed treatment effect model can be represented as follows:

$$\lambda_2(t) = \begin{cases} \lambda_1(t), & 0 \leq t \leq t_0, \\ \delta \lambda_1(t), & t > t_0, \end{cases} \tag{2.3}$$

which is referred to as the piecewise proportional hazard (PWPH) model. In practice, the delayed treatment effect often arises when there are no detectable effects of the treatment during the period $[0,\ t_0]$ but the treatment becomes fully effective afterward, as demonstrated in the sipuleucel-T trial. In this case, the optimal weight function for the log-rank test is proportional to log hazard ratio (Schoenfeld, 1981; Xu et al., 2016). Thus, we can set optimal weights to be $w_1 = 0$ for $j \in D \setminus D_2$ and $w_2 = 1$ for $j \in D_2$ which results a piecewise optimal weighted log-rank test given as follows:

$$L = \frac{\sum\limits_{j \in D_2} (d_{1j} - e_{1j})}{\left\{ \sum\limits_{j \in D_2} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \right\}^{1/2}}, \tag{2.4}$$

where D_2 is the set of indices of subjects who had the event after t_0. It is essentially similar to the standard log-rank test when only the events accumulated after the delayed onset are taken into account in the test statistics. This result makes intuitive sense, because if the treatment effect is not revealed until t_0, the events before t_0 do not contribute to detecting the treatment effects.

2.3 Sample size calculation

Xu et al. (Xu et al., 2016) showed that the total number of events required after the treatment effect onset calculated by the optimal piecewise weighted log-rank test of (2.4) is given as follows:

$$d = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\omega_1 \omega_2 \{\log(\delta)\}^2}, \tag{2.5}$$

where α and β are the type I and II errors, respectively, and ω_1 is the sample size allocation ratio of control group, $\omega_2 = 1 - \omega_1$. It is clear that the power in (2.5) is driven by the number of events after the delayed phase. Under the PWPH exponential model, Xu et al. (Xu et al., 2016) further derived an analytic power calculation method based on a piecewise weighted log-rank test (APPLE) which has been implemented in an R package 'DelayedEffect.Design'.

However, the exponential distribution assumption is strong and may be invalid for long-term survival studies. In the following section, the APPLE method is extended to a general class of PWPH models for flexibility of trial design. A new sample size formula is derived under a fixed alternative hypothesis to improve the accuracy of sample size estimation, and performance of the APPLE method and new formula are compared via simulation studies.

Let p_1 and p_2 be the failure probabilities of the control and treatment groups, respectively, after the delayed phase and $P = \omega_1 p_1 + \omega_2 p_2$ be the overall failure probability of two groups after the delayed phase. Then, the sample size required for the study is given by

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\omega_1 \omega_2 [\log(\delta)]^2 P}, \tag{2.6}$$

which has the same form as the Schoenfeld formula (Schoenfeld, 1981), however calculations of the failure probabilities are different. To calculate sample size, it is assumed that subjects are uniformly accrued over a time period t_a, an additional follow-up time t_f, with a study duration $\tau = t_a + t_f$, and no subject drops out or is lost to follow-up. Then, the censoring distribution is a uniform distribution on interval $[t_f, t_a + t_f]$. As shown in Appendix A, the probability of failure after the delayed phase for the control group can be calculated as

$$p_1 = S_1(t_0) - \frac{1}{t_a} \int_{t_f}^{t_a + t_f} S_1(t) dt, \tag{2.7}$$

and the probability of failure after the delayed phase for the treatment group can be calculated as

$$\begin{aligned} p_2 &= S_2(t_0) - \frac{1}{t_a}\int_{t_f}^{t_a+t_f} S_2(t)dt \\ &= \{S_1(t_0)\}^{1-\delta}\left[\{S_1(t_0)\}^{\delta} - \frac{1}{t_a}\int_{t_f}^{t_a+t_f}\{S_1(t)\}^{\delta}dt\right]. \end{aligned} \tag{2.8}$$

Under the PWPH exponential model, formula (2.6) reduces to the APPLE method derived by Xu et al. (Xu et al., 2016) under a sequence of local alternatives (Schoenfeld, 1981) which assume that the log hazard ratio is order of $O(n^{-1/2})$, that is the hazard ratio $\delta \to 1$ as $n \to \infty$. Thus, when the hazard ratio is small or effect size is large, the sample size calculated by formula (2.6) may be inaccurate.

To provide accurate sample size calculation, we have shown that the piecewise weighted log-rank test L under a fixed alternative $H_1 : \delta < 1$ is asymptotically normally distributed with mean $\sqrt{n}e$ and variance $\tilde{\sigma}^2/\sigma^2$, where $e = \mu/\sigma$, and μ, σ^2 and $\tilde{\sigma}^2$ are given by equations (2.10-2.12), respectively (see Appendix B for the derivation). Thus, given a two-sided type I error of α, the study power of $1-\beta$ satisfies the following:

$$\begin{aligned} 1-\beta &= P(|L| > z_{1-\alpha/2}|H_1) \\ &\simeq P\left\{\frac{\sigma(L-\sqrt{n}e)}{\tilde{\sigma}} > \frac{\sigma(z_{1-\alpha/2}-\sqrt{n}e)}{\tilde{\sigma}}\Big|H_1\right\} \\ &= \Phi\left(\frac{\sqrt{n}\mu - \sigma z_{1-\alpha/2}}{\tilde{\sigma}}\right), \end{aligned}$$

and it follows that

$$\sqrt{n}\mu - \sigma z_{1-\alpha/2} = \tilde{\sigma} z_{1-\beta}.$$

Solving for n, we obtain the following sample size formula

$$n = \frac{(\sigma z_{1-\alpha/2} + \tilde{\sigma} z_{1-\beta})^2}{\mu^2}, \tag{2.9}$$

where μ, σ^2, and $\tilde{\sigma}^2$ are given as follows:

$$\mu = \omega_1\omega_2(1-\delta)c(\delta)\int_{t_0}^{\infty}\frac{\{S_1(t)\}^{\delta}G(t)\lambda_1(t)}{[\omega_1 + \omega_2 c(\delta)\{S_1(t)\}^{\delta-1}]}dt, \tag{2.10}$$

$$\sigma^2 = \omega_1\omega_2 c(\delta)\int_{t_0}^{\infty}\frac{\{S_1(t)\}^{\delta}[\omega_1 + \omega_2\delta c(\delta)\{S_1(t)\}^{\delta-1}]G(t)\lambda_1(t)}{[\omega_1 + \omega_2 c(\delta)\{S_1(t)\}^{\delta-1}]^2}dt, \tag{2.11}$$

$$\tilde{\sigma}^2 = \omega_1\omega_2\delta c(\delta)\int_{t_0}^{\infty}\frac{\{S_1(t)\}^{\delta}G(t)\lambda_1(t)}{[\omega_1 + \omega_2\delta c(\delta)\{S_1(t)\}^{\delta-1}]}dt. \tag{2.12}$$

with $c(\delta) = \{S_1(t_0)\}^{1-\delta}$ and $G(t)$ is the common survival distribution of censoring time for both control and treatment groups. The total number of events after delayed phase can be calculated by $d = nP$, where $P = \omega_1 p_1 + \omega_2 p_2$ is the overall failure probability of two groups after the delayed phase.

2.4 Simulation

To evaluate the accuracy of the APPLE method and formula (2.9), sample sizes were calculated under a PWPH Weibull model for the following parameter settings: The Weibull distribution of the control group was $S(t) = e^{-\lambda t^{\kappa}}$; hazard ratio changing time point was set to $t_0 = 0.5$ and the proportion of control patients who could survive beyond t_0 was set to $S_1(t_0) = 90\%$; hazard ratio δ was set between 0.4 and 0.7; assuming a uniform accrual with accrual duration $t_a = 1$ and follow-up time $t_f = 2$; the shape parameter of the Weibull was set at $\kappa = 0.5, 1$, and 1.5 to represent the decreasing, constant and increasing hazard functions, respectively; sample size allocation ratio was set to $\omega_1 = 1/2$ (1:1 allocation for control and treatment group), 1/3 (1:2 allocation and more subjects assigned to the treatment group) and 2/3 (2:1 allocation and more subjects assigned to the control group). Random samples for the PWPH Weibull model were generated according to the method given in Appendix C. Assuming no loss to follow up, sample sizes were calculated with a two-sided type I error of 5% and a power of 80%. Empirical powers were estimated by performing 10,000 simulation runs. The simulation results for Xu's formula (2.6) are shown in Table 2.1 and for the new formula proposed by us (2.9) are shown in Table 2.2.

Table 2.1: Sample sizes (n) were calculated using Xu's formula (2.6) under the Weibull delayed treatment effect model with $S_1(t_0) = 90\%$, the proportion of subjects who could survive beyond the delay time $t_0 = 0.5$, a two-sided type I error of 5%, power of 80%. The corresponding empirical type I errors ($\hat{\alpha}$) and powers ($1-\hat{\beta}$) were estimated by performing 10,000 simulation runs.

		$\kappa = 0.5$			$\kappa = 1$			$\kappa = 1.5$		
ω_1	δ	n	$\hat{\alpha}$	$1-\hat{\beta}$	n	$\hat{\alpha}$	$1-\hat{\beta}$	n	$\hat{\alpha}$	$1-\hat{\beta}$
1/2	.40	483	.053	.760	168	.051	.773	84	.048	.785
(1:1)	.45	614	.051	.774	213	.049	.782	106	.052	.788
	.50	787	.052	.785	273	.051	.790	137	.051	.798
	.55	1025	.049	.792	356	.047	.788	178	.050	.797
	.60	1361	.052	.798	473	.051	.792	238	.050	.802
	.65	1856	.051	.799	647	.054	.789	327	.051	.804
	.70	2631	.048	.792	918	.049	.799	466	.046	.803
1/3	.40	630	.052	.846	215	.053	.842	104	.054	.845
(1:2)	.45	786	.048	.845	269	.050	.845	131	.051	.847
	.50	991	.049	.840	340	.052	.844	166	.049	.838
	.55	1270	.052	.833	436	.051	.830	214	.051	.835
	.60	1662	.049	.840	573	.049	.833	283	.047	.835
	.65	2239	.050	.832	773	.050	.826	385	.050	.834
	.70	3134	.047	.829	1086	.052	.827	544	.050	.819
2/3	.40	478	.051	.684	167	.050	.696	85	.053	.730
(2:1)	.45	616	.052	.708	216	.048	.716	110	.049	.739
	.50	801	.052	.723	281	.051	.733	143	.053	.747
	.55	1055	.052	.730	370	.049	.741	189	.052	.763
	.60	1418	.050	.742	497	.052	.759	254	.051	.764
	.65	1957	.052	.753	687	.047	.763	352	.048	.767
	.70	2803	.049	.767	985	.046	.771	506	.050	.785

Table 2.2: Sample sizes (n) were calculated using new formula (2.9) under the Weibull delayed treatment effect model with $S_1(t_0) = 90\%$, the proportion of subjects who could survive beyond $t_0 = 0.5$, a two-sided type I error of 5%, power of 80%. The corresponding empirical type I errors ($\hat{\alpha}$) and powers ($1 - \hat{\beta}$) were estimated by performing 10,000 simulation runs.

		$\kappa = 0.5$			$\kappa = 1$			$\kappa = 1.5$		
ω_1	δ	n	$\hat{\alpha}$	$1 - \hat{\beta}$	n	$\hat{\alpha}$	$1 - \hat{\beta}$	n	$\hat{\alpha}$	$1 - \hat{\beta}$
1/2	.40	514	.050	.793	175	.052	.797	85	.048	.799
(1:1)	.45	644	.054	.793	220	.052	.793	107	.052	.795
	.50	816	.054	.796	280	.054	.797	138	.051	.798
	.55	1052	.052	.794	362	.049	.808	179	.050	.801
	.60	1387	.049	.798	479	.048	.789	239	.050	.797
	.65	1882	.053	.804	652	.051	.800	328	.051	.804
	.70	2655	.050	.807	923	.047	.798	467	.046	.805
1/3	.40	547	.051	.801	188	.054	.797	94	.051	.819
(1:2)	.45	690	.051	.802	239	.052	.795	120	.057	.809
	.50	881	.051	.803	305	.050	.800	154	.053	.815
	.55	1143	.050	.799	397	.052	.799	201	.049	.813
	.60	1514	.046	.803	528	.046	.806	268	.051	.815
	.65	2065	.046	.798	721	.052	.809	368	.050	.813
	.70	2926	.054	.803	1025	.051	.803	525	.045	.802
2/3	.40	612	.048	.794	205	.049	.789	97	.052	.789
(2:1)	.45	760	.050	.793	256	.051	.795	122	.051	.784
	.50	957	.051	.793	324	.051	.792	156	.052	.791
	.55	1227	.051	.792	418	.052	.790	203	.049	.790
	.60	1608	.049	.799	550	.052	.798	270	.048	.795
	.65	2171	.050	.799	746	.051	.797	369	.048	.791
	.70	3050	.051	.790	1053	.047	.792	525	.049	.793

The results from sample size calculations and simulations can be summarized as follows. For balanced designs, the APPLE estimates the sample size accurately for a large hazard ratio ($\delta \geq 0.5$) (a small effect size), which is consistent with the simulation results reported by Xu et al. (Xu et al., 2016) under the exponential distribution ($\kappa = 1$). However, for a small hazard ratio ($\delta < 0.5$) (a large effect size), the sample sizes calculated by the formula (2.6) were underestimated. Because the APPLE was derived under local alternatives, it would be expected to perform well when the hazard ratio was close to 1 and perform worse when hazard ratio departs away from 1. New formula (2.9) was derived under the fixed alternative, so it would be expected to perform well no matter hazard ratio was close to 1 or not. The simulation results shown in Table 2.2 confirmed this conclusion.

For unbalanced designs, the APPLE method overestimated the sample size when more subjects were allocated to the treatment group (1:2 allocation ratio) and underestimated the sample size when more subjects were allocated to the control group (2:1 allocation ratio). The empirical power could be as low as 63% (17% lower than the nominal level), and could be as high as 86% (6% higher than the nominal level) while the new formula (2.9) performed very well for both cases of the unbalanced designs. The empirical powers simulated from new formula (2.9) were all close to the nominal level of 80%. Thus, overall the new formula (2.9) outperformed the APPLE method under both balanced and unbalanced design.

2.5 Example

Eggermont et al. (Eggermont et al., 2016) conducted a phase III, placebo-controlled immunotherapy trial for advanced melanoma. Patients who had undergone complete resection of stage III melanoma were randomly assigned in a 1:1 ratio to receive either placebo or Ipilimumab (checkpoint inhibitor), and the primary endpoint for the trial is recurrence-free survival and overall survival (OS) is a secondary endpoint. Visual separation of Kaplan-Meier curves for OS occurred approximately 6 months after randomization. The original trial design didn't consider delayed treatment effect. Here, we illustrate sample size calculation to incorporate delayed treatment effect. It is assumed that the OS times for patients receiving placebo follow an exponential distribution, whereas the OS times for patients receiving Ipilimumab follow a piecewise exponential distribution with a delay time $t_0 = 6$ months as follows:

$$\begin{aligned} S_1(t) &= e^{-\lambda t}, \\ S_2(t) &= \begin{cases} e^{-\lambda t}, & 0 \leq t < t_0, \\ ce^{-\delta\lambda t}, & t \geq t_0, \end{cases} \end{aligned}$$

where $c = e^{-\lambda t_0(1-\delta)}$ is a normalizing constant, λ is the hazard rate of the placebo group and $\delta\lambda$ is the hazard rate of the Ipilimumab group after time t_0. Thus, the hazard ratio can be expressed as

$$\frac{\lambda_2(t)}{\lambda_1(t)} = \begin{cases} 1, & 0 \leq t < t_0, \\ \delta, & t \geq t_0. \end{cases}$$

From the trial report (Eggermont et al., 2016), a 5-year OS for placebo group is 54.4%, or hazard rate of the placebo group $\lambda = -\log(0.544)/(5 \times 12) = 0.01$, and hazard ratio δ after the delay time $t_0 = 6$ (months) is 0.72 (Figure 2.1).

Further, assuming patients are accrued to the trial for $t_a = 30$ months at a constant rate (uniform accrual), followed for $t_f = 50$ months, and the study duration $\tau = t_a + t_f = 80$ months. Using the new formula, the number of events after delayed phase and sample size are 391 and 1051, respectively, to achieve 90% power with a two-sided type I error of 5%. The R code for the sample size calculation is provided in Appendix D.

A major concern to use the proposed new formula (2.9) is its robustness against the design parameters. To address this concern, we explored the relationship between sample size/number of events after the delayed phase and different design parameters. Specifically, we set up the length of accrual to 20 and 30 months; accrual duration to 40, 50 and 60 months; underlying distribution is the Weibull distribution $e^{-\lambda t^{\kappa}}$ with shape parameter $\kappa = 0.7, 1$ and 1.3, and hazard rate of the placebo group $\lambda = 0.005$ and 0.01. Sample size and number of events after the delay time $t_0 = 6$ months were calculated under the combination of these design parameters. The results (Table 2.3) showed that sample size changed from as small as 492 to as large as 7982. In contrast, the number of events after delayed phase changed from 390 to 393, which is almost a constant and very robust against the length of accrual, length of follow-up and underlying survival distribution. Therefore, to avoid

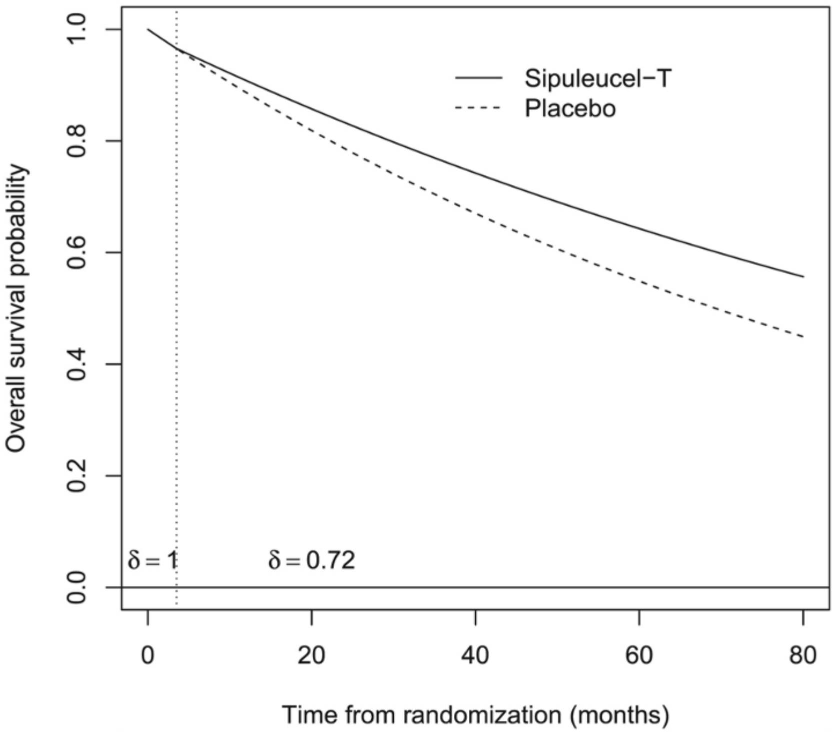

Figure 2.1: Survival Curves for the ipilimumab and placebo groups

potential power loss due to misspecification of the design parameters, it is wise to design the trial by an event-driven. That is the trial cutoff time point is based on the observed number of events after the delayed phase rather than number of patients enrolled on the study. Thus, if an event-driven design is used for this example, we will stop the trial accrual after observing 393 deaths occurred after the delayed phase to guarantee the desired statistical power for detecting the treatment effect.

Table 2.3: Sample size (n) and number of events after delayed phase $t_0 = 6$ months (d) were calculated using new formula (2.9) Weibull model with shape $\kappa = 0.7$, 1 and 1.3; hazard rate of control $\lambda = 0.01$ and 0.005; accrual duration $t_a = 20$ or 30 months; follow-up time $t_f = 40$, 50 and 60; hazard ratio $\delta = 0.72$; equal allocation ratio 1:1; two-sided type I error of 5% and power of 90%. The corresponding empirical powers (%) were estimated by performing 10,000 simulation runs.

			$\kappa = 0.7$			$\kappa = 1$			$\kappa = 1.3$		
λ	t_a	t_f	d	n	$1-\hat{\beta}$	d	n	$1-\hat{\beta}$	d	n	$1-\hat{\beta}$
.01	20	40	392	4167	90.1	391	1326	90.2	391	603	89.5
		50	392	3572	90.0	391	1124	89.8	391	541	90.0
		60	392	3152	90.2	391	986	89.8	392	503	89.9
	30	40	392	3850	89.6	391	1218	89.9	391	572	90.1
		50	392	3351	89.8	391	1051	89.6	392	522	90.0
		60	392	2988	90.0	390	934	90.3	393	492	90.5
.005	20	40	393	7982	89.8	392	2348	89.9	390	867	90.4
		50	392	6814	90.2	392	1953	90.2	390	732	89.7
		60	392	5988	89.8	391	1682	89.9	390	645	90.5
	30	40	392	7358	89.8	392	2134	89.9	390	795	90.1
		50	392	6378	89.8	391	1808	89.8	390	686	90.2
		60	392	5664	90.0	391	1578	90.3	390	614	90.2

2.6 Discussion

A challenge arising in cancer immunotherapy trials design is the presence of a delayed treatment effect which violates the proportional hazards assumption. As a result, the traditional survival trial design based on the standard log-rank test that ignores the delayed treatment effect will lead to substantial loss of statistical power. Xu et al. (Xu et al., 2016) proposed using the piecewise weighted log-rank test to incorporate the delayed treatment effect into the study design. However, their method was derived under the local alternative hypothesis and may result in an underestimated sample size when the hazard ratio is small ($\delta < 0.5$). Their formula could also overestimate or underestimate the sample size for unbalanced designs even when the hazard ratio is relatively large. This is because Xu's formula used the Schoenfeld's approach which makes assumption that the at-risk ratio is constant throughout the trial. However, the actually at-risk ratio changes as the trial progresses, particularly for an unbalanced design.

To provide accurate sample size estimation, we derived a new sample size formula under the fixed alternative hypothesis without making the constant at-risk ratio assumption. The new formula is not limited to the exponential PWPH model. It can be applied to other distribution as well. We conducted extensive simulation studies which show that the new formula provides accurate sample size estimation not only for balanced design but also for unbalanced design. Extraordinary, the number of events after delayed phase calculated using new formula (2.9) is very robust against the length of accrual, length of follow-up and underlying survival distribution. Thus, the widely used event-driven trial design is applicable to the new formula to avoid potential power loss due to misspecification of the design parameters.

The PWPH model discussed in this paper assumes that the delayed treatment effect is homogeneous across the individual subjects. It is however more natural to assume that the effect may vary heterogeneously across individuals, in which case a random delayed effect model would be more appropriate. Both Xu et al. (Xu et al., 2018) and Liu et al. (Liu et al., 2018) proposed a generalized weighted log-rank test to accommodate for the random delayed effect model. Our proposed method can be extended to the random delayed effect model as well. Further extension the proposed method is possible to a general delayed treatment effect model with random lag time by using generalized weighted log-rank test which is an optimal test. We will discuss this extension in chapter 4.

Chapter 3 Delayed Treatment Effect with Cure Rate

3.1 Introduction

Cancer immunotherapy trials have two special features. First, a delayed treatment effect discussed in Chapter 2 is common to see in survival distributions between the control and treatment groups. Second, because immunotherapies are very effective, a proportion of patients may be cured. These two features suggest that the standard proportional hazards (PH) model (Cox, 1972) no longer holds true, and using conventional sample size and power calculation methods based on the standard log-rank test will lead to substantial loss of statistical power (Schoenfeld, 1981; Freedman, 1982).

To include these features in trial design, Wang et al. (Wang et al., 2012) proposed a proportional hazards cure rate (PHCR) model while Xu et al. (Xu et al., 2016) proposed a piecewise proportional hazards (PWPH) model. However, currently no model exits to incorporate both features in the trial design. Recently, Liu et al. (Liu et al., 2018) proposed a model to incorporate both cure rate and delayed treatment effect. However, the cure rate is a nuisance parameter in their model and the trial can not be designed to testing the hypothesis for the cure rate. Furthermore, the sample size calculations from all existing methods were derived under a local alternative hypothesis. In practice, the alternative hypothesis is always fixed and does not change as sample size increase. Thus, accuracy of the sample size formula derived under a local alternative hypothesis may not be guaranteed. So we proposed a new method to properly design an immunotherapy trial.

This Chapter is organized as follows. Section 3.2 proposed a piecewise proportional hazards cure rate (PWPHCR) model to incorporate both delayed treatment effect and cure rate. A sample size formula is derived for a weighted log-rank test under a fixed alternative hypothesis in section 3.3. Section 3.4 is the simulation studies to access the accuracy of sample size calculation using the new formula and compared with the existing methods. Section 3.5 includes a real immunotherapy trial to illustrate the study design along with practical consideration of balance between sample size and follow-up time for the long-term survivors.

3.2 Piecewise proportional hazards cure rate model

For a two-arm randomized survival trial, let $S_i(t)$ denote the overall survival distribution (or latency survival distribution) and let $\lambda_i(t)$ and $f_i(t)$ denote its corresponding hazard function and density function for group i, where $i = 1$ and 2 represents control group and treatment group, respectively. Similarly, let $S_i^*(t)$ denote the continuous conditional survival function (or latency survival function) of uncured patients and let $\lambda_i^*(t)$ and $f_i^*(t)$ denote its hazard function and density function for group i. The cure rate in group i is defined by π_i, where $0 \leq \pi_i < 1$. Then, overall survival distribution of the control group is a mixture cure model (Farewell, 1982)

$$S_i(t) = \pi_i + (1 - \pi_i)S_i^*(t). \tag{3.1}$$

To incorporate a delayed treatment effect discussed in chapter 2 into the design consideration, we assume no treatment effect within period up to a fixed time point t_0 (> 0) and then full treatment effect after time t_0. Thus, the survival model can be described by a PWPH model with the latency hazard function of the treatment group is given by

$$\lambda_2^*(t) = \begin{cases} \lambda_1^*(t), & t \leq t_0, \\ \delta\lambda_1^*(t), & t > t_0, \end{cases}$$

where δ is the hazard ratio of uncured patients after a fixed delay time t_0. We assume that t_0 is known from pilot data or preclinical study, then for $t > t_0$ we can get

$$\begin{aligned} S_2^*(t) &= e^{-\Lambda_2^*(t)} \\ &= e^{-\int_0^t \lambda_2^*(\mu)d\mu} \\ &= e^{-\int_0^{t_0} \lambda_1^*(\mu)d\mu - \int_{t_0}^t \delta\lambda_1^*(\mu)d\mu} \\ &= e^{-\int_0^{t_0} \lambda_1^*(\mu)d\mu} e^{-\int_0^t \delta\lambda_1^*(\mu)d\mu} e^{\int_0^{t_0} \delta\lambda_1^*(\mu)d\mu} \\ &= S_1^*(t_0)[S_1^*(t)]^\delta [S_1^*(t_0)]^{-\delta}. \end{aligned}$$

Hence, the latency survival distribution of the treatment group is given by

$$S_2^*(t) = \begin{cases} S_1^*(t), & t \le t_0, \\ [S_1^*(t_0)]^{1-\delta} [S_1^*(t)]^{\delta}, & t > t_0. \end{cases} \tag{3.2}$$

Combining mixture cure model (3.1) and PWPH model (3.2) we can define following PWPHCR model. A mixture cure model for the control group is

$$S_1(t) = \pi_1 + (1 - \pi_1) S_1^*(t)$$

with density function $f_1(t) = (1 - \pi_1) f_1^*(t)$ and hazard function $\lambda_1(t) = f_1(t)/S_1(t)$, and a mixture cure rate model with a delayed effect for the treatment group

$$S_2(t) = \begin{cases} \pi_1 + (1 - \pi_1) S_1^*(t), & t \le t_0, \\ \tilde{\pi}_2 + (1 - \tilde{\pi}_2) [S_1^*(t_0)]^{1-\delta} [S_1^*(t)]^{\delta}, & t > t_0. \end{cases}$$

However, due to the delayed treatment effect and difference of cure rates between the control arm and treatment arm, this mixture distribution $S_2(t)$ has a discontinuous point at t_0 with a jump size of $(\tilde{\pi}_2 - \pi_1)(1 - S_1^*(t_0))$. To smooth $S_2(t)$ at t_0, we multiple a constant $c = \{\pi_1 + (1 - \pi_1) S_1^*(t_0)\}/\{\tilde{\pi}_2 + (1 - \tilde{\pi}_2) S_1^*(t_0)\}$ to rescaling of the $S_2(t)$ when $t \ge t_0$ and resulting following PWPHCR model

$$S_2(t) = \begin{cases} \pi_1 + (1 - \pi_1) S_1^*(t), & t \le t_0, \\ \pi_2 + (1 - \pi_2) \tilde{c} [S_1^*(t_0)]^{1-\delta} [S_1^*(t)]^{\delta}, & t > t_0, \end{cases} \tag{3.3}$$

where $\pi_2 = c\tilde{\pi}_2$ and $\tilde{c} = c(1 - \tilde{\pi}_2)/(1 - c\tilde{\pi}_2)$. It can be verified that $S_2(t)$ is a continuous survival function of a mixture cure model with cure rate of π_2. The density function for treatment group is

$$f_2(t) = \begin{cases} (1 - \pi_1) f_1^*(t), & t \le t_0, \\ (1 - \pi_2) \tilde{c} \delta [S_1^*(t_0)]^{1-\delta} [S_1^*(t)]^{\delta - 1} f_1^*(t), & t > t_0, \end{cases}$$

and the corresponding hazard function is $\lambda_2(t) = f_2(t)/S_2(t)$.

If $\pi_1 = \tilde{\pi}_2$ and $\delta = 1$, we have $c = 1$ and $\pi_1 = \pi_2$ and $S_2(t) = S_1(t)$. The PWPHCR model (3.3) is a general model which includes, as special case, the following:

- $\pi_1 = \pi_2 = 0$ (no cure) and $t_0 = 0$ (no delay), the PWPHCR model reduces to the standard PH model (Schoenfeld, 1981);

- $\pi_1 = \pi_2 = 0$ (no cure) and $t_0 \neq 0$ (with delay), the PWPHCR model reduces to the PWPH model(Xu et al., 2016);
- $\pi_1 \leq \pi_2 \neq 0$ (with cure) and $t_0 = 0$ (no delay), the PWPHCR model reduces to the PHCR model (Wang et al., 2012).

Under the PWPHCR model, testing the null hypothesis of no treatment effect

$$H_0 : S_1(t) = S_2(t),$$

is equivalent to testing the following null hypothesis:

$$H_0 : \delta = 1 \ \text{ and } \ \pi_1 = \pi_2.$$

Various alternative hypotheses are of interest: $H_{1a} : \delta \neq 1, \pi_1 \neq \pi_2$, with differences in both the short-term survival and the cure fraction; $H_{1b} : \delta \neq 1, \pi_1 = \pi_2$, with a difference in the short-term survival but not in the cure fraction; and $H_{1c} : \delta = 1, \pi_1 \neq \pi_2$, with difference in the cure fraction but not in the short-term survival.

3.3 Sample size calculation

Assume that there are n patients allocated between the control and treatment groups. Let D be the set of identifiers in the two groups who died, and let t_j be the death time of the j^{th} patient in either group. We assume that the $\{t_j\}$ are distinct. Let y_j be an indicator variable of the control group of j^{th} patient; that is, $y_j = 1$ if the j^{th} patient belongs to the control (group 1) and $y_j = 0$ if the j^{th} patient belongs to the treatment (group 2). If we define $n_i(t)$ to be the number at risk just before time t in group i, then the weighted log-rank test L is given by

$$L = \frac{\sum\limits_{j \in D} w_j \{y_j - p(t_j)\}}{\left[\sum\limits_{j \in D} w_j^2 p(t_j)\{1 - p(t_j)\}\right]^{1/2}}.$$

where $p(t_j) = n_1(t_j)/\{n_1(t_j) + n_2(t_j)\}$ and $w_j = W(t_j)$, and $W(t)$ is a weight function converging to a deterministic function $w(t)$. As shown in Appendix E, under the PWPHCR model and a fixed alternative hypothesis, the weighted log-rank test L is asymptotically normally distributed with mean $\sqrt{n}e$, where $e = \mu_w/\sigma_w$ and variance $\tilde{\sigma}_w^2/\sigma_w^2$, where μ_w, σ_w^2 and $\tilde{\sigma}_w^2$ are given in following equations (3.5-3.7). With a two-sided type I error of α, the power of $1-\beta$ satisfies the following:

$$\begin{aligned} 1-\beta &= P(|L| > z_{1-\alpha/2}|H_1) \\ &\simeq P\left(\frac{\sigma_w(L-\sqrt{n}e)}{\tilde{\sigma}_w} > \frac{\sigma_w(z_{1-\alpha/2}-\sqrt{n}e)}{\tilde{\sigma}_w}\Big| H_1\right) \\ &= \Phi\left(\frac{\sqrt{n}\mu_w - \sigma_w z_{1-\alpha/2}}{\tilde{\sigma}_w}\right). \end{aligned}$$

Therefore it follows that

$$\sqrt{n}\mu_w - \sigma_w z_{1-\alpha/2} = \tilde{\sigma}_w z_{1-\beta}.$$

Solving for n, we obtain the following sample size formula for the weighted log-rank test

$$n = \frac{(\sigma_w z_{1-\alpha/2} + \tilde{\sigma}_w z_{1-\beta})^2}{\mu_w^2}, \tag{3.4}$$

where

$$\begin{aligned} \mu_w &= \int_0^\infty w(t)\frac{\pi(t)(1-\pi(t))\{\lambda_1(t)-\lambda_2(t)\}}{\pi(t)\lambda_1(t)+\{1-\pi(t)\}\lambda_2(t)}V(t)dt, && (3.5) \\ \sigma_w^2 &= \int_0^\infty w^2(t)\pi(t)\{1-\pi(t)\}V(t)dt, && (3.6) \\ \tilde{\sigma}_w^2 &= \int_0^\infty w^2(t)\frac{\pi(t)(1-\pi(t))\lambda_1(t)\lambda_2(t)}{[\pi(t)\lambda_1(t)+\{1-\pi(t)\}\lambda_2(t)]^2}V(t)dt, && (3.7) \end{aligned}$$

and function $V(t)$ is an incomplete density function of failure and $\pi(t)$ is a ratio of probability at risk of a subject belong to the control group vs. the overall probability at risk of the two groups. It can be shown that

$$\begin{aligned} V(t) &= \{\omega_1\lambda_1(t)S_1(t) + \omega_2\lambda_2(t)S_2(t)\}G(t), \\ \pi(t) &= \frac{\omega_1 S_1(t)G(t)}{\omega_1 S_1(t)G(t) + \omega_2 S_2(t)G(t)}. \end{aligned}$$

where ω_1 and ω_2 are the allocation ratio to the control and treatment groups, respectively. This new formula (3.4) can be applied to the following special cases:

- $\pi_1 = \pi_2 = 0$ and $t_0 = 0$, sample size calculation under the standard PH model was derived by Schoenfeld (Schoenfeld, 1981);

- $\pi_1 = \pi_2 = 0$ and $t_0 > 0$, sample size calculation under the PWPH model was derived by Xu et al. (Xu et al., 2016); and
- $\pi_1 \leq \pi_2 \neq 0$ and $t_0 = 0$, sample size calculations under the PHCR model were derived by Wang et al. (Wang et al., 2012) and Xiong and Wu (Xiong and Wu, 2017).

Because an optimal weight function for the log-rank test under the PWPHCR model is unknown, we simply use the piecewise weighted log-rank test (i.e., $w(t) = 0, t \leq t_0$ and $w(t) = 1, t > t_0$) for sample size calculation in the following sections.

3.4 Simulation

To evaluate the accuracy of the proposed new sample size formula (3.4) and compare to the existing methods, sample sizes were calculated under the PWPHCR model where the latency distribution of the control group is the Weibull distribution $S_1^*(t) = e^{-\lambda t^\kappa}$, cure rate of the control group is set to $\pi_1 = 0.1$ and fixed delay time is set to $t_0 = 0.5$, with other design parameters set as follows: Hazard ratio δ is set between 0.3 and 0.7; accrual duration $t_a = 2$ and follow-up time $t_f = 10$ for the PWPHCR model and $t_a = 1$ and $t_f = 2$ for other models; the shape parameter of the Weibull distribution is set to $\kappa = 0.5, 1,$ and 1.5 to represent the decreasing, constant and increasing hazard functions, respectively; the hazard parameter λ is determined by the proportion of uncured control patients who could survive beyond t_0 is $S_1^*(t_0) = 90\%$ or set to $\lambda = 0.1$ for the model without a delayed treatment effect; and sample size allocation ratio is set to $\omega_1 = 1/2$ (1:1 equal allocation). Random samples for the PWPHCR Weibull model were generated according to the method given in Appendix F. Assuming uniform accrual and no loss to follow up, sample sizes were calculated with a two-sided type I error of 5% and power of 80% or 90%. Empirical powers were estimated by performing 10,000 simulation runs.

First, sample sizes and total number of events were calculated under the general PWPHCR model and the corresponding empirical type I errors and powers were simulated and shown in Table 3.1. Results showed that the simulated empirical powers are all close to the nominal level. Thus, the new formula gives the accurate sample size estimation in all three hypothesis testing scenarios under the PWPHCR model.

Second, by setting $\pi_1 = \pi_2 = 0$ and $t_0 = 0$, the PWPHCR model reduces to a standard PH model. We compared our new formula to the Schoenfeld formula. The simulation results (Table 3.2) showed that the new formula provides more accurate sample size estimation than the Schoenfeld formula, particular when the hazard ratio is small $(\delta < 0.5)$ whereas the Schoenfeld formula underestimated the sample size and number of events.

Third, by setting $\pi_1 = \pi_2 = 0$, and $t_0 > 0$, the PWPHCR model reduces to a PWPH model with a fixed delay time t_0. Therefore, we compared the new formula to the Xu's formula. Again, the simulation results (Table 3.3) showed that the new formula is more accurate than the Xu's formula, particular when the hazard ratio is small $(\delta < 0.5)$ whereas the Xu's formula underestimated the sample size and number of events.

Fourth, by setting $\pi_1 \leq \pi_2 \neq 0$, and $t_0 = 0$, the PWPHCR model reduces to PHCR model. Thus, we compared the new formula to the Wang's formula. Simulation results (Table 3.4) showed that Wang's formula did not provide the correct sample size estimation. However, the new formula provides more accurate sample size and number of events estimation in all scenarios.

Overall, the new formula is general and applicable to many different survival models to accommodate for the cancer immunotherapy trial design. The new formula provides more accurate sample size and number of events estimation than exiting methods in the literature.

Table 3.1: Sample sizes (n) were calculated by the new formula (3.4) under the PWPHCR Weibull model with $S_1^*(t_0) = 90\%$ for three hypothesis scenarios. Uniform accrual with accrual period $t_a = 2$ and follow-up duration $t_f = 10$, no loss to follow-up, cure rate of the control group $\pi_1 = 0.1$, a two-sided type I error of 5% and power of 90%. The corresponding empirical type I errors ($\hat{\alpha}$) and powers ($1-\hat{\beta}$) were estimated by performing 10,000 simulation runs.

PWPHCR		$\kappa = 0.5$			$\kappa = 1$			$\kappa = 1.5$		
Test	δ/π_2	$n(d)$	$\hat{\alpha}$	$1-\hat{\beta}$	$n(d)$	$\hat{\alpha}$	$1-\hat{\beta}$	$n(d)$	$\hat{\alpha}$	$1-\hat{\beta}$
H_{1a}	.70/.12	1324(418)	.049	89.8	561(427)	.053	90.1	870(775)	.050	89.7
	.65/.13	906(280)	.049	90.2	368(276)	.051	89.6	551(488)	.052	89.4
	.60/.14	653(197)	.050	90.0	257(190)	.050	90.6	377(332)	.049	89.3
	.55/.15	489(144)	.047	89.6	188(135)	.054	89.7	272(239)	.052	89.2
	.50/.16	376(108)	.051	89.4	141(100)	.056	90.6	204(178)	.050	88.5
	.45/.17	295(83)	.049	89.6	109(75)	.047	90.3	156(136)	.055	88.7
H_{1b}	.70/.1	1525(484)	.047	89.6	718(553)	.053	90.0	1401(1261)	.050	89.7
	.65/.1	1075(335)	.050	89.9	491(374)	.051	89.7	964(867)	.049	89.8
	.60/.1	787(240)	.050	89.7	349(262)	.051	90.0	689(619)	.050	89.3
	.55/.1	593(178)	.048	89.9	256(188)	.050	90.3	505(455)	.049	89.2
	.50/.1	457(135)	.049	89.9	191(139)	.050	90.2	377(339)	.049	88.9
	.45/.1	358(103)	.048	89.6	145(145)	.047	90.0	385(256)	.053	88.8
H_{1c}	1/.30	1857(595)	.051	90.0	301(219)	.050	89.7	205(165)	.051	90.2
	1/.32	1525(484)	.047	90.4	252(181)	.047	90.3	173(138)	.054	90.5
	1/.35	1168(366)	.055	90.0	198(140)	.052	90.4	138(108)	.054	89.8
	1/.38	921(284)	.051	89.8	160(112)	.051	90.0	113(88)	.055	90.5
	1/.40	795(243)	.050	89.3	141(97)	.056	90.5	100(76)	.050	90.6
	1/.42	693(210)	.052	89.3	125(85)	.051	90.2	89(67)	.054	90.0

Table 3.2: Sample sizes (n) were calculated using Schoenfeld's formula (SF) under the standard PH Weibull model with hazard parameter of control $\lambda = 0.1$; uniform accrual with accrual period $t_a = 1$ and follow-up duration $t_f = 2$; no loss to follow-up; a two-sided type I error of 5%, power of 80%. The corresponding empirical type I errors ($\hat{\alpha}$) and powers ($1 - \hat{\beta}$) were estimated by performing 10,000 simulation runs.

PH		$\kappa = 0.5$			$\kappa = 1$			$\kappa = 1.5$		
Method	δ	$n(d)$	$\hat{\alpha}$	$1-\hat{\beta}$	$n(d)$	$\hat{\alpha}$	$1-\hat{\beta}$	$n(d)$	$\hat{\alpha}$	$1-\hat{\beta}$
SF	.30	226(22)	.053	.743	148(22)	.056	.753	99(22)	.048	.751
	.35	286(29)	.047	.759	188(29)	.050	.759	125(29)	.047	.771
	.40	362(38)	.050	.770	237(38)	.051	.770	159(38)	.048	.784
	.45	460(50)	.050	.775	301(50)	.049	.778	202(50)	.052	.787
	.50	590(66)	.046	.784	387(66)	.051	.789	259(66)	.052	.789
	.55	767(88)	.051	.786	504(88)	.047	.793	337(88)	.050	.790
	.60	1019(121)	.051	.790	669(121)	.051	.796	448(121)	.047	.797
	.65	1390(170)	.054	.796	913(170)	.054	.795	612(170)	.047	.798
	.70	1970(247)	.051	.793	1295(247)	.049	.798	869(247)	.049	.795
New	.30	251(25)	.049	.793	163(24)	.053	.788	107(24)	.048	.799
	.35	310(31)	.051	.795	201(31)	.047	.788	133(31)	.049	.806
	.40	384(40)	.049	.788	250(40)	.052	.799	166(40)	.050	.796
	.45	481(52)	.049	.801	314(52)	.052	.794	208(51)	.051	.798
	.50	610(68)	.054	.793	399(68)	.054	.805	265(68)	.049	.793
	.55	787(91)	.046	.802	515(90)	.049	.800	343(90)	.049	.801
	.60	1038(123)	.048	.797	680(123)	.048	.802	453(122)	.043	.800
	.65	1408(172)	.051	.792	923(172)	.051	.798	617(171)	.052	.796
	.70	1988(250)	.049	.804	1305(249)	.047	.799	874(249)	.051	.799

Table 3.3: Sample sizes (n) were calculated using Xu's formula under the PWPH Weibull model with $S_1^*(t_0) = 90\%$, the proportion of subjects who could survival beyond $t_0 = 0.5$ of fixed delay time. Assuming uniform accrual with a accrual period $t_a = 1$ and follow-up duration $t_f = 2$; no loss to follow-up; a two-sided type I error of 5%, power of 80%. The corresponding empirical type I errors $(\hat{\alpha})$ and powers $(1-\hat{\beta})$ were estimated by performing 10,000 simulation runs.

PWPH		$\kappa = 0.5$			$\kappa = 1$			$\kappa = 1.5$		
Method	δ	$n(d)$	$\hat{\alpha}$	$1-\hat{\beta}$	$n(d)$	$\hat{\alpha}$	$1-\hat{\beta}$	$n(d)$	$\hat{\alpha}$	$1-\hat{\beta}$
Xu	.30	302(52)	.050	.741	105(33)	.055	.756	52(27)	.053	.768
	.35	382(67)	.049	.759	132(42)	.050	.763	66(36)	.058	.779
	.40	483(86)	.049	.772	168(55)	.053	.777	84(46)	.052	.785
	.45	614(111)	.049	.775	213(71)	.053	.779	106(60)	.049	.788
	.50	787(145)	.050	.780	273(93)	.051	.785	137(80)	.051	.798
	.55	1025(191)	.052	.791	356(124)	.048	.791	178(106)	.053	.797
	.60	1361(257)	.051	.791	473(168)	.048	.792	238(145)	.049	.802
	.65	1856(355)	.054	.797	647(234)	.048	.802	327(202)	.051	.804
	.70	2631(510)	.049	.796	918(339)	.050	.800	466(294)	.048	.803
New	.30	336(59)	.050	.792	113(36)	.059	.788	54(29)	.054	.791
	.35	415(73)	.047	.787	140(45)	.050	.787	68(37)	.052	.789
	.40	514(92)	.052	.782	175(57)	.053	.802	85(47)	.048	.799
	.45	644(117)	.047	.789	220(73)	.051	.798	107(61)	.052	.795
	.50	816(150)	.050	.796	280(95)	.049	.796	138(80)	.051	.798
	.55	1052(197)	.049	.795	362(127)	.048	.794	179(107)	.050	.801
	.60	1387(262)	.049	.801	479(170)	.050	.799	239(145)	.050	.797
	.65	1882(361)	.052	.793	652(237)	.049	.801	328(203)	.051	.804
	.70	2655(516)	.049	.800	923(342)	.051	.799	467(295)	.046	.805

Table 3.4: Sample sizes (n) were calculated using Wang's formula under the PHCR Weibull model with hazard parameter $\lambda = 0.1$ and cure rate of $\pi_1 = 0.1$ for the control group; uniform accrual with accrual period $t_a = 1$ and follow-up duration $t_f = 2$; no loss to follow-up; a two-sided type I error of 5%, power of 80%. The corresponding empirical type I errors $(\hat{\alpha})$ and powers $(1-\hat{\beta})$ were estimated by performing 10,000 simulation runs.

PHCR		$\kappa = 0.5$			$\kappa = 1$			$\kappa = 1.5$		
Method	δ/π_2	$n(d)$	$\hat{\alpha}$	$1-\hat{\beta}$	$n(d)$	$\hat{\alpha}$	$1-\hat{\beta}$	$n(d)$	$\hat{\alpha}$	$1-\hat{\beta}$
Wang	.30/.12	162(14)	.047	.559	108(15)	.051	.562	75(15)	.054	.594
	.35/.13	208(19)	.047	.601	139(19)	.052	.606	96(20)	.053	.608
	.40/.14	266(25)	.046	.618	177(25)	.052	.626	122(26)	.048	.635
	.45/.15	338(33)	.046	.646	225(33)	.049	.651	154(34)	.048	.658
	.50/.16	431(42)	.043	.676	286(43)	.050	.672	195(44)	.054	.690
	.55/.17	552(56)	.049	.699	366(56)	.048	.695	249(57)	.051	.701
	.60/.18	712(74)	.053	.722	471(74)	.049	.727	319(75)	.048	.730
	.65/.19	929(98)	.052	.747	612(98)	.049	.745	413(99)	.048	.755
	.70/.20	1231(133)	.049	.772	808(133)	.046	.779	541(132)	.049	.781
New	.30/.12	274(24)	.048	.784	179(24)	.052	.787	119(24)	.054	.790
	.35/.13	330(30)	.048	.795	216(30)	.050	.795	144(30)	.048	.795
	.40/.14	398(37)	.046	.790	261(37)	.049	.789	174(37)	.052	.798
	.45/.15	481(46)	.046	.794	315(46)	.049	.791	211(46)	.049	.803
	.50/.16	581(57)	.052	.797	381(57)	.047	.799	255(57)	.047	.796
	.55/.17	705(71)	.052	.806	462(71)	.047	.803	309(71)	.052	.797
	.60/.18	860(89)	.046	.793	563(88)	.048	.795	376(88)	.051	.804
	.65/.19	1054(111)	.051	.794	689(111)	.050	.800	459(110)	.047	.798
	.70/.20	1302(140)	.050	.789	849(139)	.045	.788	562(137)	.050	.797

3.5 Example

Robert et al. (Robert et al., 2011) conducted a randomized Phase III immunotherapy trial for previously untreated metastatic melanoma. Patients were randomly assigned in a 1:1 ratio to receive either Ipilimumab plus dacarbazine (treatment arm) or dacarbazine plus placebo (control arm), and primary endpoint of the trial is overall survival (OS). Visual separation of the Kaplan-Meier curves occurred approximately 3.5 months after randomization and plateaus in survival curves for both groups (Figure 3.1). The hazard functions of the two groups approach to zero after the study duration beyond 50-60 months (Figure 3.2).

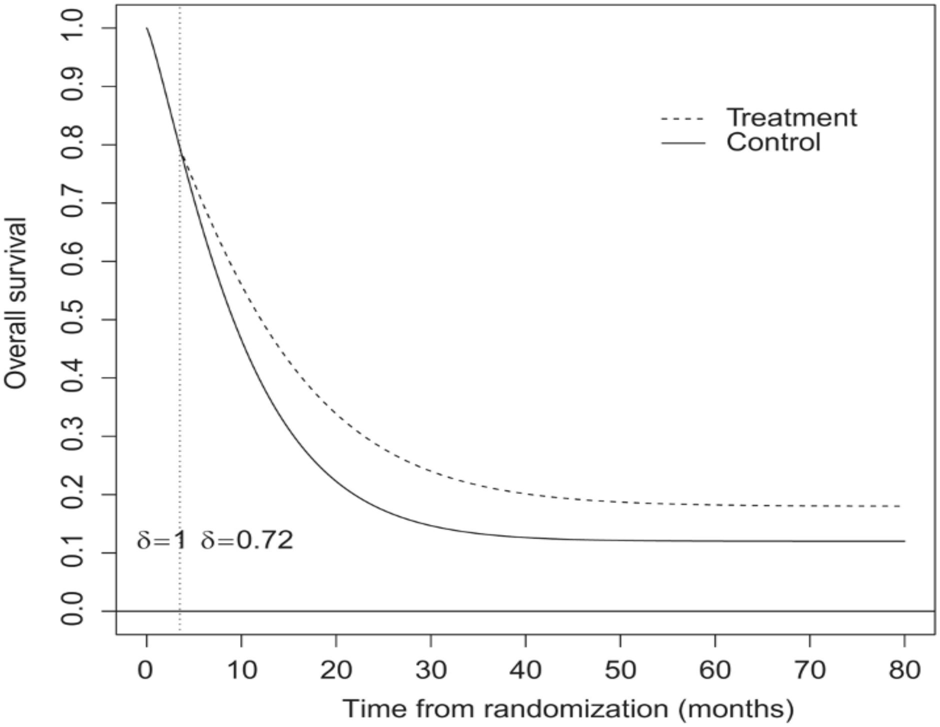

Figure 3.1: Survival distributions of the control and treatment groups for the example

The original trial design however didn't consider either delayed treatment effect or cure rate. With a two-sided type I error 0.05, and power of 90% to detect a hazard ratio 0.727, the total number of events calculated by the Schoenfeld formula is $d = 414$. Assuming accrual time 17 months and follow-up time 17 months, the total sample size calculated under exponential distribution is $n = 496$. Actually, a total of 502 patients were randomly assigned to the study and it took 37 months follow-up to observe 414 deaths for the final analysis and the total study duration was about 54 months.

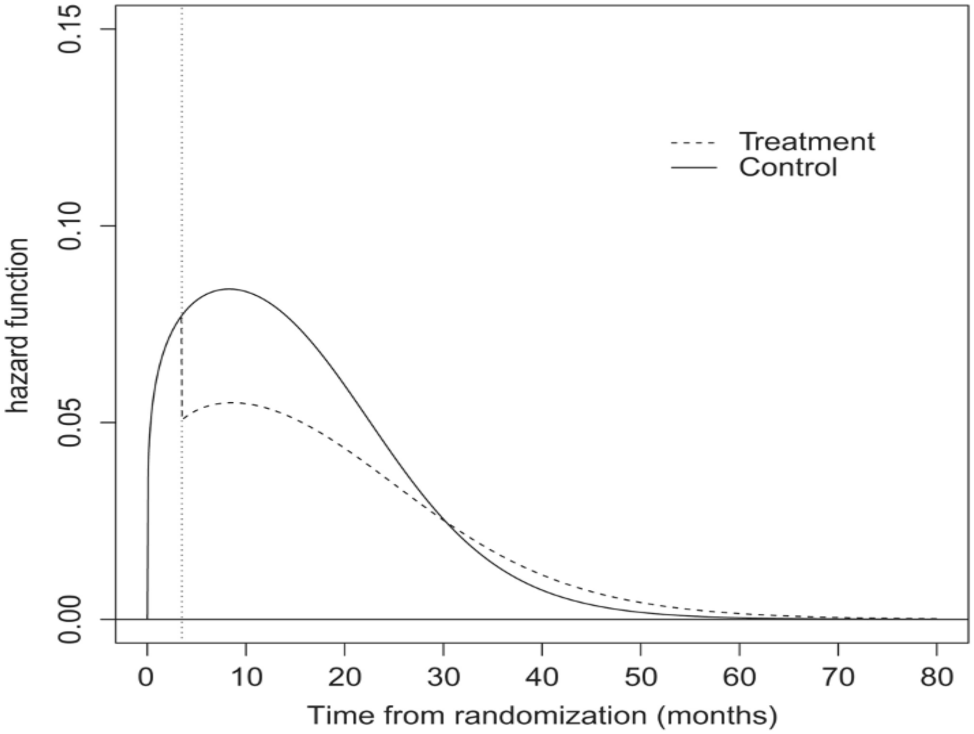

Figure 3.2: Hazard functions of the control and treatment groups for the example

Here, we illustrate the trial design using proposed PWPHCR model to incorporate both delayed treatment effect and cure rate. From the trial report (Robert et al., 2011), the medians OS are $m_1 = 9.1$ and $m_2 = 11.2$ months, and cure rates are approximately $\pi_1 = 12\%$ and $\pi_2 = 18\%$ for the control arm and treatment arm, respectively. We use the Weibull distribution $S_1^*(t) = e^{-\lambda_1 t^\kappa}$ to model the OS survival for uncured patients of the control arm, where $\kappa = 1.2$ and $\lambda_1 = 0.059$ are fitted shape and scale parameters. Thus, the mixture cure model for the control group is

$$S_1(t) = \pi_1 + (1 - \pi_1)S_1^*(t)$$

and the mixture cure model with a delayed treatment effect for the treatment group is given by

$$S_2(t) = \begin{cases} \pi_1 + (1-\pi_1)S_1^*(t), & t \le t_0 \\ \pi_2 + (1-\pi_2)\tilde{c}\left[S_1^*(t_0)\right]^{1-\delta}\left[S_1^*(t)\right]^{\delta}, & t > t_0 \end{cases}$$

where δ is the hazard ratio after delayed time t_0. The hypothesis of interest for this trial could be

$$H_0 : \delta = 1 \text{ and } \pi_1 = \pi_2 \quad vs. \quad H_{1a} : \delta \neq 1 \text{ and } \pi_1 \neq \pi_2$$

Based on the trial results, we calculate the sample size under the alternatives: $\delta = 0.72$ and $\pi_1 = 12\%$ and $\pi_2 = 18\%$ for the PWPHCR model with a delayed treatment effect time $t_0 = 3.5$ months. Additional, we assume accrual $t_a = 17$ months and follow-up $t_f = 37$ months, the total study duration $\tau = t_a + t_f = 54$ months, with a two-sided type I error 5% and power of 90%, the total number of events and sample size required for the trial are $d = 466$ and $n = 553$, respectively. This design requires more number of events or sample size because of the delayed treatment effect. The R code for the sample size calculation is provided in Appendix G.

3.6 Discussion

It is common that cancer immunotherapy trials present a delayed treatment effect and cure fraction. Ignoring the delayed treatment effect and/or cure rate in the trial design will result in substantial power loss. In this chapter, we proposed a PWPHCR model to incorporate both delayed treatment effect and cure rate in cancer immunotherapy trial design and derived a general sample size formula under a fixed alternative hypothesis. Simulation results showed that the new formula provides more accurate sample size estimation than existing methods.

However, a question for the trial design with long-term survivors is how to balance between sample size and length of follow-up so that the trial is practically feasible and data are also mature enough. To address this question, we use the example in section 3.5 for illustration. Figure 3.2 shows that hazard functions of both groups approach to zero after 50-60 months from the time of randomization. Therefore, the study duration should exceed 50-60 months so the data are mature enough and cure rates are identifiable.

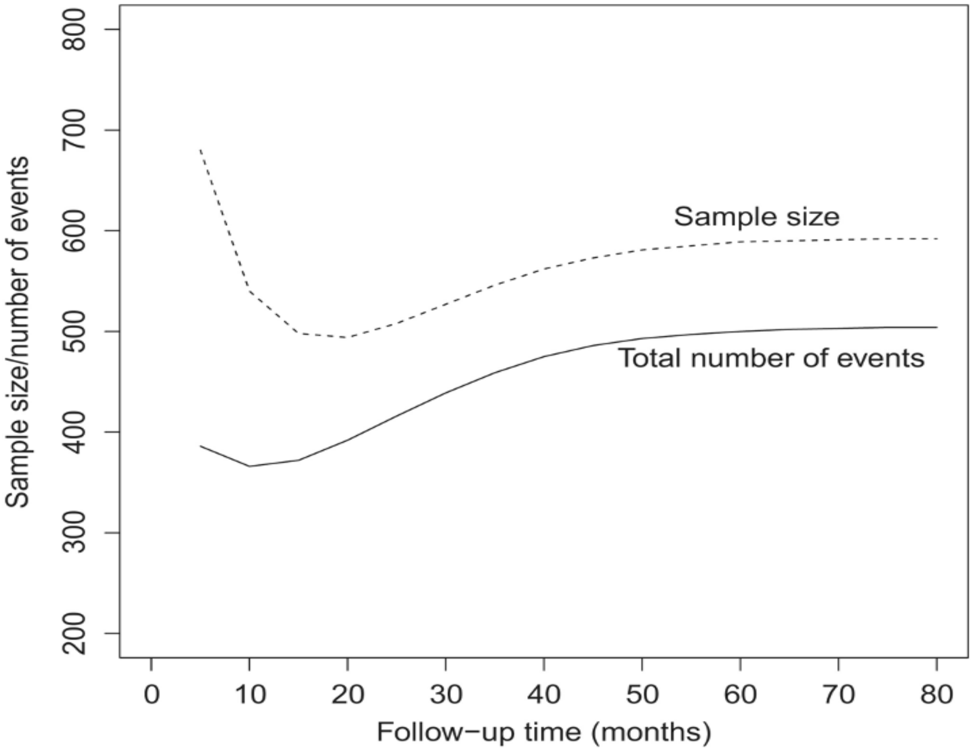

Figure 3.3: Relationship between sample size/number of events and length of follow-up for the example

The relationships between follow-up time and total number of events and sample size in Figure 3.3 shows that as the follow-up time increases, sample size and total number of events decrease first and then gradually increase to approach reasonable levels after hazard functions approach zero. To optimize the study design, we will choose the study cutoff date at the follow-up time $t_f = 20$ months (or study duration $\tau = 37$ months) at which time the study has a relative small sample size and large power (Figure 3.3). However, the long-term survival cannot be observed. Therefore, the choice between detecting a short-term risk reduction and identifying a long-term survival should be made in advance for the trial design.

Planning an interim analysis is difficult for the trial with both delayed treatment effect and long-term survival. We do not want to perform an interim analysis too early to stop futility because it could result a high false negative rate due to the delayed treatment effect. We also do not want to perform an interim analysis early to stop efficacy because it could result unobservable for the cure rate. Furthermore, event-driven trial design is no longer applied for the PWPHCR model due to the non-proportionality. Additional research is needed for group sequential design under the PWPHCR model.

Chapter 4 Random Delayed Treatment Effect with Cure Rate

4.1 Introduction

Immunotherapies have been increasingly used for treating patients with advanced-stage cancers. Because of the indirect mechanism of action of immunotherapy, a delayed treatment effect is often seen in immunotherapy trials. When patients are homogeneous across the individual subjects, such delayed treatment effect occurs in a fixed time period which results a threshold delayed effect model. We discussed such kind of delayed treatment effect in chapter 2. Various weighted log-rank tests have been proposed to increase the efficiency of trial design with a threshold delayed effect model. Hasegawa (Hasegawa, 2014) considered to use the Fleming-Harrington $G^{\rho,\gamma}$ class of weighted log-rank test. Xu et al. (Xu et al., 2016) recommended a piecewise weighted log-rank test. Magirr and Burman (Magirr and Burman, 2019) developed a modestly-weighted log-rank test. Zucker and Lakatos (Zucker and Lakatos, 1990) proposed a general class treatment lag model and derived a maximin efficiency robust test for the trial design. Ye and Yu (Ye and Yu, 2018) extended Zucker and Lakatos' results to a generalized linear lag model. The maximin efficiency robust test is also a weighted log-rank test which put less weight on early events and full weight after the delayed period. Recently, Ding and Wu (Ding and Wu, 2020) considered a simple robust test for designing cancer immunotherapy trials.

When patients enrolled on a immunotherapy trial are heterogeneous, the duration of delayed effect is more suitable as a random variable rather than a fixed time period. Immunotherapy trial designs with a random delay time have also been studied in the literature (Xu et al., 2018; Liu et al., 2018). Suppose the random delay time τ follows a distribution $F_\tau(t)$, both Xu et al (Xu et al., 2018) and Liu at al (Liu et al., 2018) proposed to use the $F_\tau(t)$-weighted log-rank test and showed it is nearly optimal test for a random delayed proportional hazards (PH) model. Furthermore, it is also uncommon to see a proportion of patients had long-term survival or cure from immunotherapy trials. Liu et al (Liu et al., 2018) included a cure rate in the random delayed model but limited to their study design under the PH model assumption, which results the difference of cure rates between treatment groups can't not be tested.

In this chapter, we extend Liu et al model to a general random delayed cure rate model and derived a sample size formula for designing cancer immunotherapy trials which provides testing on the hypotheses for both short-term and long-term survival. The rest of this chapter is organized as follows. In section 4.2, we describe the random delayed effect cure rate model. Section 4.3 presents a sample size formula for the $F_\tau(t)$-weighted log-rank test. In section 4.4, simulations are conducted to study the performance of the proposed the $F_\tau(t)$-weighted log-rank test and sample size formula, the robustness of misspecification is also considered in section 4.4. Discussions are given in Section 4.5.

4.2 Generalized piecewise proportional hazards cure rate model

Let $\lambda_k^*(t)$ be the hazard function of uncured patients for group $k = 1, 2$ which represents the control and treatment groups, respectively, and τ be the random delay time. The survival model with a random delay time for uncured patients can be described by a piecewise proportional hazards model which is given by

$$\lambda_2^*(t) = \begin{cases} \lambda_1^*(t), & t \le \tau, \\ \delta\lambda_1^*(t), & t > \tau, \end{cases}$$

where δ is the hazard ratio of uncured patients after the random delay time τ. The survival function of the treatment group for uncured patients is given by

$$S_2^*(t) = \begin{cases} S_1^*(t), & t \le \tau, \\ \left[S_1^*(\tau)\right]^{1-\delta}\left[S_1^*(t)\right]^{\delta}, & t > \tau. \end{cases} \tag{4.1}$$

Similar as Chapter 3, combine the cure rate model and piecewise random delayed treatment effect model, We define a random delayed cure rate model as follows. A mixture cure rate model for the control arm is

$$S_1(t) = \pi_1 + (1 - \pi_1)S_1^*(t),$$

where $0 \leq \pi_1 < 1$ is the cure rate of control group, and a mixture cure model for the experimental treatment arm is given by

$$S_2(t,\tau) = \begin{cases} \pi_1 + (1-\pi_1)S_1^*(t), & t \leq \tau, \\ \tilde{\pi}_2 + (1-\tilde{\pi}_2)\left[S_1^*(\tau)\right]^{1-\delta}\left[S_1^*(t)\right]^{\delta}, & t > \tau, \end{cases}$$

where $0 \leq \tilde{\pi}_2 < 1$. The survival distribution $S_2(t,\tau)$ also has a single jump time point τ. To smooth the function $S_2(t,\tau)$, we define following smooth factor

$$A(\tau) = \frac{\pi_1 + (1-\pi_1)S_1^*(\tau)}{\tilde{\pi}_2 + (1-\tilde{\pi}_2)S_1^*(\tau)}$$

and multiple it to $S_2(t,\tau)$, we obtain

$$S_2(t,\tau) = \begin{cases} \pi_1 + (1-\pi_1)S_1^*(t), & t \leq \tau, \\ A(\tau)\{\tilde{\pi}_2 + (1-\tilde{\pi}_2)\left[S_1^*(\tau)\right]^{1-\delta}\left[S_1^*(t)\right]^{\delta}\}, & t > \tau. \end{cases}$$

Since the random delay time τ is not observed, we integrate respect to the distribution of τ to obtain the marginal survival function

$$\begin{aligned} S_2(t) &= E(S_2(t,\tau)) \\ &= \{\pi_1 + (1-\pi_1)S_1^*(t)\}P(\tau > t) + \int_0^t A(\mu)\{\tilde{\pi}_2 + (1-\tilde{\pi}_2)\left[S_1^*(\mu)\right]^{1-\delta}\left[S_1^*(t)\right]^{\delta}\}dF_\tau(u) \\ &= \{\pi_1 + (1-\pi_1)S_1^*(t)\}S_\tau(t) \\ &+ \tilde{\pi}_2\int_0^t A(u)dF_\tau(u) + (1-\tilde{\pi}_2)[S_1^*(t)]^{\delta}\int_0^t A(u)\left[S_1^*(u)\right]^{1-\delta}dF_\tau(u) \end{aligned}$$

and marginal density

$$\begin{aligned} f_2(t) &= \frac{-dS_2(t)}{dt} \\ &= -\{(\pi_1 + (1-\pi_1)S_1^*(t))\frac{dS_\tau(t)}{dt} + (1-\pi_1)\frac{dS_1^*(t)}{dt}S_\tau(t)\} \\ &- \tilde{\pi}_2 A(t)f_\tau(t) - (1-\tilde{\pi}_2)A(t)[S_1^*(t)]^{\delta}\left[S_1^*(t)\right]^{1-\delta}f_\tau(t) \\ &- \delta[S_1^*(t)]^{\delta-1}\frac{dS_1^*(t)}{dt}\int_0^t A(u)(1-\tilde{\pi}_2)[S_1^*(u)]^{1-\delta}dF_\tau(u) \\ &= f_1^*(t)\left\{(1-\pi_1)S_\tau(t) + (1-\tilde{\pi}_2)\delta[S_1^*(t)]^{\delta-1}\int_0^t A(u)[S_1^*(u)]^{1-\delta}dF_\tau(u)\right\}, \end{aligned}$$

where $F_\tau(t)$ are the survival function of the random delay time τ and $S_\tau(t) = 1 - F_\tau(t)$, and $f_1^*(t)$ is the density function of uncured patients for the control group.

Assume that random variable τ has support on domain $[t_1, t_2]$, let

$$\pi_2 = \tilde{\pi}_2 \int_{t_1}^{t_2} A(u) dF_\tau(u)$$

be the cure rate of the treatment group, then, $\tilde{\pi}_2$ can be solved from above equation. It is easy to verify that the marginal survival function $S_2(t) = \pi_1 + (1 - \pi_1) S_1^*(t)$ when $t \leq t_1$ and $S_2(t) = \pi_2 + (1 - \pi_2)\tilde{c}[S_1^*(t)]^\delta$ when $t > t_2$, where

$$\tilde{c} = \frac{(1 - \tilde{\pi}_2)}{(1 - \pi_2)} \int_{t_1}^{t_2} A(u) \left[S_1^*(u)\right]^{1-\delta} dF_\tau(u).$$

Thus, the survival distribution of the treatment group is also a mixture cure model with cure rate π_2. Between t_1 and t_2, the hazard ratio changes from 1 to δ gradually, instead of a sudden jump as in the fixed delay effect model.

When $\pi_1 = \pi_2 = 0$ (no cure), the random delayed cure rate model reduces to a generalized piecewise proportional hazards (GPWPH) model with marginal survival function

$$S_2(t) = S_1^*(t) S_\tau(t) + [S_1^*(t)]^\delta \int_0^t [S_1^*(u)]^{1-\delta} dF_\tau(u) \tag{4.2}$$

and the marginal density

$$f_2(t) = f_1^*(t) \left\{ S_\tau(t) + \delta [S_1^*(t)]^{\delta - 1} \int_0^t [S_1^*(u)]^{1-\delta} dF_\tau(u) \right\}.$$

It has been shown that $F_\tau(t)$-weighted log-rank test is a nearly optimal test under the GPWPH model (Xu et al., 2018; Liu et al., 2018).

4.3 Sample size calculation

In this section, we present a sample size formula for the $F_\tau(t)$-weighted log-rank test for trial designs under the random delayed cure rate models.

Consider a two-sided hypothesis for testing the difference of survival distributions between the experimental and control groups

$$H_0 : S_2(t) = S_1(t) \quad \text{vs} \quad H_1 : S_2(t) \neq S_1(t).$$

The log-rank test is a well-known optimal test statistic under the PH model. However, it could loss the power when PH assumption is invalid. To increase the power to detect the treatment effect, a weighted log-rank test can be used. Let T_i and C_i denote, respectively, the failure time and censoring time of the i^{th} subject. We assume that T_i and C_i are continuous random variables. The observed failure time and failure indicator are $X_i = T_i \wedge C_i$ and $\Delta_i = I(T_i \leq C_i)$, respectively, $i = 1, \cdots, n$, and $Z_i = 0, 1$ for group 1 and 2. Define $N_i(t) = \Delta_i I(X_i \leq t)$ and $Y_i(t) = I(X_i \geq t)$ be the failure process and at risk process, and $\overline{Y}_1(t) = \sum_{i=1}^n (1 - Z_i) Y_i(t)$, $\overline{Y}_2(t) = \sum_{i=1}^n Z_i Y_i(t)$, then the weighted log-rank score test

$$U = n^{-1/2} \sum_{i=1}^n \int_0^\infty W_n(t) \left\{ Z_i - \frac{\overline{Y}_2(t)}{\overline{Y}_1(t) + \overline{Y}_2(t)} \right\} dN_i(t)$$

is asymptotically normal distributed and its asymptotic variance can be estimated by

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n \int_0^\infty W_n^2(t) \frac{\overline{Y}_1(t)\overline{Y}_2(t)}{\{\overline{Y}_1(t) + \overline{Y}_2(t)\}^2} dN_i(t),$$

where $W_n(t)$ is a bounded nonnegative weight function that converges in probability to $w(t)$. By martingale central limited theorem (Fleming and Harrington, 1991), the weighted log-rank test $L = U/\hat{\sigma}$ is asymptotically standard normal distributed under the null hypothesis H_0. Thus, given a two-sided type I error rate α, we reject null hypothesis if $|L| > z_{1-\alpha/2}$.

Under a general fixed alternative hypothesis, same as discussed in chapter 3, we derived (Wei and Wu, 2020) an asymptotic distribution of the weighted log-rank test L, which is normally distributed with mean $\sqrt{n}\mu/\sigma$ and variance $\sigma^2/\tilde{\sigma}^2$, where μ, σ^2 and $\tilde{\sigma}^2$ are given in following equations (4.4), (4.5) and (4.6), respectively. Thus, sample size can be calculated using following formula

$$n = \frac{(\sigma z_{1-\alpha/2} + \tilde{\sigma} z_{1-\beta})^2}{\mu^2}, \tag{4.3}$$

where μ, σ^2, and $\tilde{\sigma}^2$ are given as follows:

$$\mu = \int_0^\infty w(t) \frac{\pi(t)(1-\pi(t))\{\lambda_2(t) - \lambda_1(t)\}}{\pi(t)\lambda_1(t) + \{1-\pi(t)\}\lambda_2(t)} V(t) dt, \tag{4.4}$$

$$\sigma^2 = \int_0^\infty w^2(t)\pi(t)\{1-\pi(t)\}V(t)dt, \tag{4.5}$$

$$\tilde{\sigma}^2 = \int_0^\infty w^2(t) \frac{\pi(t)(1-\pi(t))\lambda_1(t)\lambda_2(t)}{[\pi(t)\lambda_1(t) + \{1-\pi(t)\}\lambda_2(t)]^2} V(t) dt, \tag{4.6}$$

and the functions $\pi(t)$ and $V(t)$ are given by

$$\begin{aligned}\pi(t) &= \frac{\omega_1 S_1(t)}{\omega_1 S_1(t) + \omega_2 S_2(t)}, \\ V(t) &= \{\omega_1 \lambda_1(t) S_1(t) + \omega_2 \lambda_2(t) S_2(t)\} G(t),\end{aligned}$$

where $\lambda_1(t)$ and $\lambda_2(t)$ are the hazard functions, ω_1 and $\omega_2 = 1 - \omega_1$ are the allocation ratios of the control and treatment groups, and $G(t)$ is the common censoring distribution function of two groups. We will use weight function $w(t) = F_\tau(t)$ for the sample size calculation under the random delayed cure rate model proposed in section 4.2.

When $\pi_1 = \pi_2 = 0$, the random delayed cure rate model reduces to GPWPH model. By using $F_\tau(t)$-weighted log-rank test, Xu et al (Xu et al., 2018) propose to use following sample size formula

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\omega_1 \omega_2 [\log(\delta)]^2 \int_0^\infty F_\tau^2(t) V(t) dt}. \tag{4.7}$$

4.4 Simulation

In this section, we conduct simulations to study the performance of the proposed a $F_\tau(t)$-weighted log-rank test and sample size formula and the impact of misspecifying the random delayed effect on the sample size and study power in following two sub-sections.

Performance of new sample size formula

To evaluate the accuracy of the proposed sample size formula (4.3) and compare to the existing methods, sample sizes were calculated under the random delayed cure rate models where the distribution of uncured patients for the control group is the Weibull distribution $S_1^*(t) = e^{-\lambda t^\kappa}$ and cure rate of the control group is set to $\pi_1 = 0.1$ and cure rate of the treatment group π_2 is set as given in table 4.1, with other design parameters are set as follows: the scale parameter is set to $\lambda = 0.01$; the shape parameter is set to $\kappa = 0.7, 1$, and 1.3 to represent the decreasing, constant and increasing hazard functions, respectively; hazard ratio δ is set between 0.45 and 0.7; uniform accrual with accrual duration $t_a = 2$ and follow-up time $t_f = 10$; sample size allocation ratio is set to $\omega_1 = 0.5$ (1:1 equal allocation). Assuming the random delay time τ follows an uniform distribution on interval

$[2, 6]$, sample sizes were calculated with a two-sided type I error rate 5% and power of 90%. Empirical type I error rate and power were estimated from 10,000 simulated trials.

Results recorded in Table 4.1 showed that the simulated empirical type I error rates and powers were all close to the nominal levels. Thus, the proposed $F_\tau(t)$-weighted log-rank test preserved type I error rate and sample size formula provided accurate sample size estimation in all three hypothesis testing scenarios: H_{1a}: differences in both the short-term survival and the cure fraction; H_{1b}: difference in the short-term survival but not in the cure fraction; and H_{1c}: difference in the cure fraction but not in the short-term survival. Results recorded in Table 4.2 showed that sample sizes were quite robust against the random delay time τ distributions: either an uniform distribution or a Beta(a, b) distribution with different parameters on domain $[2, 10]$, that is Beta$(\frac{t-2}{10-2}, a, b)$, where a, b are the parameters. Results also showed that the simulated empirical powers were all close to the nominal level.

By setting $\pi_1 = \pi_2 = 0$, and random delay time τ follows an uniform distribution on domain $[1, 6]$ or $[2, 10]$, the random delayed cure rate model reduces to a GPWPH model. Therefore, we compared the new formula to Xu's formula. The simulation results recorded in Table 4.3 showed that the new formula is more accurate than the Xu's formula, particular when the hazard ratio is small ($\delta \leq 0.5$) whereas the Xu's formula underestimated the sample size.

Impact of misspecifying delayed effect

To explore impact of misspecifying delayed effect for the proposed methods, we first consider scenarios where the true underlying delayed effect is fixed but misspecified to be a random delay or vice verse. We compared empirical powers by simulations under each misspecification scenario. Under the fixed delay scenario where the true fixed time is $t_0 = 6$ months, results recorded in Table 4.4 showed that the power loss was nearly 10% when under-specified the fixed time point less than 5 months whereas the power gain was nearly 11% when over-specified the fixed time point more than 5 months. Similar results were observed when misspecifying random delayed effect on domain $[3, 9]$ months as fixed time points. Therefore, we can conclude that misspecifying a random delay to a fixed delay could result a relative a big loss or gain on the study power. In contrast, misspecifying to a fixed delay to a random delay led only 1% to 3% power loss or gain no matter the true scenario is fixed or random delay effect.

We also assessed the robustness of the proposed methods when the distribution of random delay time τ was misspecified in the study design. Two scenarios of misspecifications are considered.

First, we assumed that the true random delay time τ follows an uniform distribution on domain $[3, 9]$ months whereas the misspecified domains are $[3, 7]$, $[3, 11]$, $[1, 7]$ and $[1, 11]$ months. Table 4.5 illustrated the impact of misspecifiying the random delay time domain on the sample size and empirical power. Sample sizes did not change much and empirical powers were close to the nominal level and misspecifying domains led to only 1% to 2% power loss or gain.

Second, we assumed that the true random delay time τ follows an uniform distribution on domain $[2, 10]$ months whereas the misspecified random delay time τ follows Beta(2,3), Beta(2,2) and Beta(3,2) distributions on domain [2, 10]. From results recorded in Table 4.6, we can make the conclusion that sample size and empirical power were not sensitive to the distributions of random delay time.

Overall, new formula under random delayed cure rate model is not sensitive to the distribution or lag time of the random delay, which means the new formula is more robust when compared with fixed delay effect. Also, the new formula provides more accurate sample size estimation than the exiting methods in the literature.

Table 4.1: Sample sizes (n) were calculated by proposed formula under the Weibull random delayed cure rate model with uniform random delayed treatment effect on interval $[2, 6]$ for three hypothesis scenarios. Uniform accrual with accrual period $t_a = 2$ and follow-up duration $t_f = 10$, baseline $\lambda = 0.01$, no loss to follow-up, cure rate of the control group $\pi_1 = 0.1$, a two-sided type I error of 5% and power of 90%. The corresponding empirical type I errors ($\hat{\alpha}$) and powers (EP) were estimated by performing 10,000 simulation runs.

		$\kappa = 0.7$			$\kappa = 1$			$\kappa = 1.3$		
Test	δ/π_2	n	$\hat{\alpha}$	EP	n	$\hat{\alpha}$	EP	n	$\hat{\alpha}$	EP
H_{1a}	.70/.12	1668	.050	89.7	590	.054	90.4	759	.051	90.1
	.65/.13	1147	.049	89.9	396	.051	90.0	478	.052	89.8
	.60/.14	829	.049	90.2	282	.050	89.9	325	.050	89.6
	.55/.15	622	.048	89.9	208	.051	90.4	232	.047	89.9
	.50/.16	479	.051	89.6	158	.049	89.5	171	.051	89.8
	.45/.17	376	.046	89.0	123	.048	90.0	129	.054	89.7
H_{1b}	.70/.1	1895	.046	90.1	706	.052	89.9	1155	.047	90.3
	.65/.1	1338	.047	90.3	491	.052	90.6	778	.049	89.6
	.60/.1	982	.053	90.3	355	.052	90.1	542	.050	89.4
	.55/.1	742	.048	89.5	264	.049	90.4	387	.050	89.9
	.50/.1	573	.048	90.2	200	.048	89.8	281	.053	89.8
	.45/.1	449	.052	89.6	155	.049	90.3	207	.047	89.5
H_{1c}	1/.30	2755	.046	88.9	508	.048	90.1	202	.054	90.2
	1/.32	2251	.050	88.9	419	.048	89.9	170	.055	90.1
	1/.35	1712	.049	88.6	324	.049	90.2	135	.050	90.7
	1/.38	1339	.052	89.1	258	.055	89.8	110	.048	90.5
	1/.40	1152	.053	89.8	224	.049	89.7	97	.053	91.3
	1/.42	998	.051	89.4	196	.051	89.5	86	.052	90.0

Table 4.2: Sample sizes (n) were calculated using different random delayed effect distributions (Uniform and Beta) on domain [2, 10] under the Weibull random delayed cure rate model with hazard parameter of control $\lambda = 0.01$; uniform accrual with accrual period $t_a = 1$ and follow-up duration $t_f = 2$; no loss to follow-up; cure rate of the control group $\pi_1 = 0.1$; a two-sided type I error of 5% and power of 90%. The corresponding empirical type I errors ($\hat{\alpha}$) and powers (EP) were estimated by performing 10,000 simulation runs.

		$\kappa = 0.7$			$\kappa = 1$			$\kappa = 1.3$		
Dist	δ/π_2	n	$\hat{\alpha}$	EP	n	$\hat{\alpha}$	EP	n	$\hat{\alpha}$	EP
Unif	.70/.12	1737	.050	89.2	608	.501	90.4	816	.051	89.7
$[2, 10]$	.65/.13	1195	.048	89.9	409	.048	89.9	512	.049	89.4
	.60/.14	864	.053	89.8	291	.051	89.3	347	.052	89.6
	.55/.15	648	.051	89.5	215	.050	90.2	247	.048	89.5
	.50/.16	499	.051	89.2	163	.052	89.7	182	.050	89.1
	.45/.17	393	.051	88.9	127	.045	89.6	137	.050	89.2
Beta	.70/.12	1721	.048	91.1	603	.050	90.5	804	.051	89.8
$(2, 2)$	.65/.13	1183	.050	90.7	405	.050	90.5	505	.050	90.4
	.60/.14	856	.049	91.2	288	.051	90.8	342	.050	90.0
	.55/.15	642	.053	90.0	213	.051	90.7	244	.054	90.6
	.50/.16	494	.047	90.9	162	.054	90.7	180	.052	90.4
	.45/.17	389	.053	90.1	126	.051	90.3	136	.053	90.7
Beta	.70/.12	1683	.051	91.6	595	.052	91.1	768	.048	90.3
$(1, 3)$	.65/.13	1157	.048	91.4	400	.046	91.4	484	.050	90.1
	.60/.14	837	.052	91.1	284	.051	90.3	329	.05 0	90.9
	.55/.15	628	.050	91.8	210	.052	90.8	235	.054	91.1
	.50/.16	484	.053	91.2	160	.048	90.8	173	.051	90.5
	.45/.17	380	.049	90.8	124	.053	90.9	131	.050	90.9
Beta	.70/.12	1805	.051	92.8	631	.053	92.1	847	.047	91.9
$(.5, .5)$	.65/.13	1241	.048	92.5	424	.048	91.5	531	.050	91.3
	.60/.14	898	.051	92.0	301	.053	92.1	360	.048	91.8
	.55/.15	674	.051	92.2	223	.053	91.6	256	.053	91.7
	.50/.16	520	.053	92.4	169	.052	92.0	189	.054	91.7
	.45/.17	409	.052	92.1	132	.046	92.1	143	.054	92.2

Table 4.3: Sample sizes (n) were calculated using Xu's formula under the Weibull random delayed effect model with baseline hazard parameter of control group is $\lambda = 0.01$. Assuming uniform accrual with a accrual period $t_a = 2$ and follow-up duration $t_f = 10$; no loss to follow-up; a two-sided type I error rate 5% and power of 90%. The corresponding empirical powers (EP) were estimated by performing 10,000 simulation runs.

	New Method				Xu's Method			
	Unif$[1,6]$		Unif$[2,10]$		Unif$[1,6]$		Unif$[2,10]$	
δ	n	EP	n	EP	n	EP	n	EP
.70	516	90.5	690	90.9	500	89.6	676	89.9
.65	358	89.6	479	91.3	342	89.1	460	89.0
.60	259	90.8	346	91.0	242	88.7	325	88.8
.55	192	90.0	257	90.8	176	88.0	236	88.4
.50	146	90.4	195	91.7	131	87.9	175	88.3
.45	113	91.5	151	91.0	98	86.6	131	87.7

Table 4.4: The empirical power comparison when the delayed effect scenarios misspecified, the fixed delay time point $t_0 = 6$ months and random delay τ follows an uniform on interval [3, 9] months. Uniform accrual with accrual period $t_a = 1$ and follow-up duration $t_f = 2$, no loss to follow-up, cure rate of the control group $\pi_1 = 0.1$ and of the treatment group $\pi_2 = 0.12$, hazard parameter of control $\lambda = 0.2$ and hazard ratio $\delta = 0.7$; a two-sided type I error rate 5% and power of 80%. The corresponding empirical powers (EP) under misspecified scenarios were estimated by 10,000 simulation runs.

Misspecified Setting	True Setting	
	Fixed delay $t_0 = 6$	Random delay Unif$[3, 9]$
PWPHCR	EP	EP
1 month	70.5%	67.1%
3 months	73.6%	70.8%
6 months	81.0%	76.2%
9 months	85.7%	83.6%
12 months	91.4%	89.1%
New Model	EP	EP
$[1, 11]$ months	82.4%	80.4%
$[1, 9]$ months	81.2%	78.1%
$[3, 11]$ months	83.9%	81.2%
$[3, 9]$ months	82.6%	79.5%

Table 4.5: Sample sizes (n) were calculated under the Weibull random delayed cure rate model model with misspecified random delayed effect domain. The true random delay is uniform on interval [3,9]. Hazard parameter of control $\lambda = 0.01$ and Uniform accrual with accrual period $t_a = 1$ and follow-up duration $t_f = 2$; no loss to follow-up;cure rate of the control group $\pi_1 = 0.1$; a two-sided type I error rate 5% and power of 90%.

Dist		Unif$[3, 9]$		Unif$[3, 7]$		Unif$[3, 11]$		Unif$[1, 7]$		Unif$[1, 11]$	
κ	δ/π_2	n	EP	n	EP	n	EP	n	EP	n	EP
0.7	.70/.12	1739	90.4	1697	88.8	1745	90.2	1659	88.3	1745	90.1
	.65/.13	1196	89.9	1167	88.0	1200	89.9	1141	88.3	1200	89.7
	.60/.14	865	90.2	844	88.2	868	90.0	825	88.4	868	89.4
	.55/.15	649	89.3	633	88.7	651	89.7	619	87.6	651	89.2
	.50/.16	500	89.4	487	88.9	501	89.3	476	87.2	502	89.6
	.45/.17	393	89.3	383	88.9	394	89.4	375	87.5	395	89.1
1	.70/.12	609	89.9	597	89.4	609	90.0	586	89.4	611	90.5
	.65/.13	409	90.1	401	89.4	409	89.8	394	88.7	411	90.8
	.60/.14	291	90.2	285	89.4	291	89.5	280	88.2	292	89.7
	.55/.15	215	89.7	211	88.6	215	89.6	207	88.8	216	89.4
	.50/.16	164	90.4	160	89.1	164	90.2	158	88.5	164	89.5
	.45/.17	127	89.4	125	89.1	127	90.5	122	87.9	128	89.9
1.3	.70/.12	811	89.8	779	88.8	832	91.0	761	88.3	827	91.1
	.65/.13	510	89.3	490	89.3	521	90.8	480	88.2	519	90.6
	.60/.14	345	89.8	333	88.8	353	90.2	326	87.6	351	90.1
	.55/.15	246	89.4	237	88.4	251	90.8	232	87.9	250	90.4
	.50/.16	181	89.5	175	88.8	185	90.3	171	88.0	164	89.7
	.45/.17	137	89.7	132	88.6	140	90.1	129	88.4	139	90.2

Table 4.6: Sample sizes (n) were calculated under the Weibull random delayed cure rate model by mis-specified Beta distributions of random delayed effect on domain [2, 10]. The true random delay time is Uniform on interval [2,10]. Hazard parameter of control $\lambda = 0.01$ and Uniform accrual with accrual period $t_a = 1$ and follow-up duration $t_f = 2$; no loss to follow-up; cure rate of the control group $\pi_1 = 0.1$; a two-sided type I error of 5% and power of 90%.

Dist		Unif$[2, 10]$		Beta$(2, 3)$		Beta$(2, 2)$		Beta$(3, 2)$	
κ	δ/π_2	n	EP	n	EP	n	EP	n	EP
0.7	.70/.12	1737	89.2	1695	89.2	1721	89.6	1740	90.3
	.65/.13	1195	89.9	1165	89.1	1183	89.5	1197	89.6
	.60/.14	864	89.8	843	88.7	856	89.2	865	89.2
	.55/.15	648	89.5	632	88.7	642	89.0	649	89.4
	.50/.16	499	89.2	487	88.2	494	89.0	500	89.7
	.45/.17	393	88.9	382	88.3	389	89.2	393	88.9
1	.70/.12	608	90.4	596	88.9	603	89.3	608	89.8
	.65/.13	409	89.9	400	89.2	405	89.5	409	90.1
	.60/.14	291	89.3	285	89.2	288	89.3	290	89.7
	.55/.15	215	90.1	210	89.5	213	89.6	215	89.9
	.50/.16	163	89.7	160	88.9	162	89.0	163	90.2
	.45/.17	127	89.6	124	88.9	126	89.3	127	89.3
1.3	.70/.12	816	89.7	783	88.5	804	89.8	818	89.7
	.65/.13	512	89.4	493	88.8	505	89.5	513	90.0
	.60/.14	347	89.6	334	88.6	342	88.8	347	90.1
	.55/.15	247	89.5	238	87.7	244	88.8	247	89.7
	.50/.16	182	89.2	133	89.0	180	88.6	182	89.3
	.45/.17	137	89.2	133	89.0	136	88.9	138	89.4

4.5 Example

In this section, we use data from a two-arm phase III Eastern Cooperative Oncology Group (ECOG) trial for melanoma to illustrate the trial design. There were 92 deaths among 146 patients in the treatment group. The treatment arm (high-dose interferon alpha-2b) relapse-free survival (RFS) data was fitted using SAS macro PSPMCM (Corbière and Joly, 2007) and got the Weibull cure rate model with an estimated shape parameter $\kappa = 1.018$ (take as 1, ie, exponential distribution), scale parameter $\lambda = 0.836$ (years) (ie, median RFS 10 months for uncured patients) and cure rate of 35%. Thus, for designing a new immunotherapy trial, the RFS for the control arm could be appropriately assumed to be

$$S_1(t) = 0.35 + 0.65e^{-\frac{\log(2)}{10}t}.$$

Further assuming that new immunotherapy has a random delay effect which follows an uniform distribution on interval $[0, 6]$ months. Three scenarios are considered here: (1) improve the short-term survival by increasing the median RFS to 14.28 months for uncured patients but not the cure rate; (2) increase the cure rate to 0.45 but not the short-term survival; and (3) increase the cure rate to 0.45 and improve the median short-term RSF to 14.28 months for uncured patients. With a two-sided type I error rate of 0.05, power of 80% at the alternative, 24 months accrual period, and 12 months follow up, the total sample sizes for two groups are 1641, 1281, and 423 for scenarios (1), (2) and (3), respectively. The corresponding simulated empirical type I error and power are 0.05 and 80%, 0.049 and 77%, and 0.051 and 79% for scenarios (1), (2) and (3), respectively. Thus, the proposed methods preserved the type I error rate and provided adequate power for the trial designs. Figure 4.1 shows the RFS survival functions for three different hypotheses scenarios. The vertical dot line indicates a uniform random delay on interval $[0, 6]$ (months). The R code for the sample size calculation is provided in Appendix H.

4.6 Discussion

How to deal with delayed treatment effect in cancer immunotherapy trial design is a typical challenge since the duration of lag time can be considered as a fixed time period or a ran-

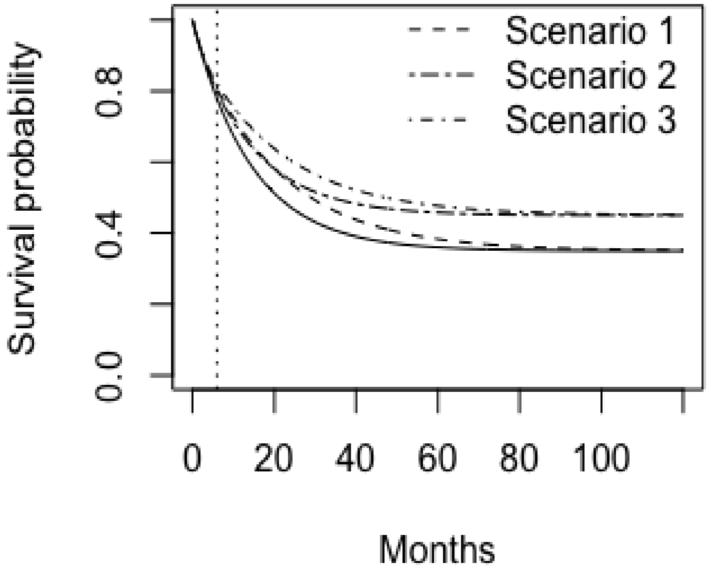

Figure 4.1: Hypothetical random delayed cure rate model for three scenarios

dom interval by different enrollment types of patients. Xu et al (Xu et al., 2016) proposed a fixed delayed effect model whereas both Xu et al (Xu et al., 2018) and Liu at al (Liu et al., 2018) proposed a random delayed effect model. However, Xu et al did not include a cure rate in their random effect model and Liu et al's model included a cure rate but limited to their study design under the PH model assumption. In this chapter, we proposed a random delayed cure rate model to incorporate both random delayed effect and cure rate for cancer immunotherapy trial designs. Simulation results showed that the new formula provides an accurate sample size estimation under the random delayed cure rate model.

In real trial design, a fixed delayed effect or random delayed effect need to be prespecified. Usually we make assumption for the time domain of delayed effect from pilot data during the trial design. However, the true time domain is unknown in advance, the misspecification is inevitable when doing trial design. Our simulation results showed that misspecifying a random delayed effect to a fixed delayed effect could result a relative a larger loss or gain on the study power while random delayed effect model is less sensitive to the lag time domain and distribution compared to the fixed delayed effect model.

Chapter 5 Delayed Treatment Effect with Non-responders

5.1 Introduction

We discussed fixed or random delayed treatment effect model with cure rate in Chapters 3 and 4 repectively and summarized that fixed/random delayed treatment effect and cure rate are the two underlying causes behind non-proportional hazards (NPH) patterns in cancer immunotherapy trial design. Since proportional hazard assumption no longer holds under NPH patterns, using standard sample size and power calculation methods based on log-rank test would lead to a loss of power. Various weighted log-rank tests have been proposed to improve the efficiency of trial design. As we discussed in chapter 2, 3 and 4, Xu et al. (Xu et al., 2016) considered a piecewise weighted log-rank test since piecewise weight is an optimal weight for fixed delayed treatment effect model. Xu et al. (Xu et al., 2018) also recommended a weighted log-rank test and proved that $F_\tau(t)$-weight is a nearly optimal weight for a random delayed model (Xu et al., 2018). Magirr and Burman (Magirr and Burman, 2019) developed a modestly-weighted log-rank test (MWLRT) and used $1/\max(\hat{S}(t_j-), \hat{S}(t_0))$ as a weight function, which is entirely analogous to implementing the Fleming-Harrington-(0,1) test. To avoid pre-specifying the delay changed time t_0, a milestone weight function was also included in Magirr's paper and performed well in delayed-effect scenario with reasonable mature data.

On the other hand, compared with other oncology trials of traditional cancer treatments, only a limited percentage of patients would respond to the treatment in reality since immunotherapy-sensitive of tumors are heterogeneous (Schlom and Gulley, 2018) in immunotherapy trials. It is more suitable to treat patients as non-responders and responders in treatment group. Immunotherapy trial designs with such kind of dichotomized response incurred by treating responders and non-responders in treatment group have also been studied in literature (Xu et al., 2020). Xu et al. showed responders and non-responders in treatment group of inadequate size would give rise to a variety of NPH patterns and present a novel P%-responder information embedded (PRIME) method to deal with dichotomized response in treatment group. However, sample size calculation based on PRIME method is complex and the corresponding R package (Immunotherapy.Design) is not efficient.

In this chapter, we follow the assumption of Xu et al. (Xu et al., 2020) to consider responders and non-responders in treatment group and derive a sample size formula under weighted log-rank test for canner immunotherapy trials design. The rest of this chapter is organized as follows. In section 5.2, we describe the responder model with delayed treatment effect. Section 5.3 presents how to calculate weight function $w_R(t)$ and derives a sample size formula under proposed weighted log-rank test. In section 5.4, simulations are conducted to study the performance of the proposed sample size formula under various weight functions. Discussions are given in Section 5.5.

5.2 Piecewise proportional hazards responder rate model

For a two-arm randomized survival trial, let $S_C(t)$ and $S_T(t)$ denote the overall survival distributions for control and treatment groups. Let $\lambda_C(t)$, $f_C(t)$, $\lambda_T(t)$ and $f_T(t)$ denote the corresponding hazard functions and density functions for two groups. Similarly, let $S_R(t)$ and $S_{NR}(t)$ denote the continuous conditional survival functions of responder patients and non-responder patients in treatment group. Let $\lambda_R(t)$, $f_R(t)$, $\lambda_{NR}(t)$ and $f_{NR}(t)$ denote its hazard functions and density functions for corresponding responder and non-responder patients. The response rate in treatment group is defined by p, where $0 \leq p \leq 1$. Then, overall survival distribution of the treatment group is a mixture model

$$S_T(t) = pS_R(t) + (1-p)S_{NR}(t) \tag{5.1}$$

To incorporate a delayed treatment effect discussed in Chapter 2 into the design consideration, we assume no treatment effect within period up to a fixed time point t_0 (> 0) and then full treatment effect after time t_0. Thus, the survival model can be described by a PWPH model with the hazard function of the treatment group for responders. It is can be written in the form of

$$\lambda_R(t) = \begin{cases} \lambda_C(t), & t \leq t_0, \\ \delta\lambda_C(t), & t > t_0, \end{cases}$$

where δ is the hazard ratio between responder patients in treatment group and patients in control group after a fixed delay time t_0. We assume that t_0 is known from pilot data or preclinical study and $S_{NR}(t) = S_C(t)$, then survival distribution of the responders in treatment group is given by

$$S_R(t) = \begin{cases} S_C(t), & t \leq t_0, \\ \left[S_C(t_0)\right]^{1-\delta} \left[S_C(t)\right]^{\delta}, & t > t_0. \end{cases} \tag{5.2}$$

Combining mixture cure model (5.1) and PWPH model (5.2), we can define the following model: The piecewise proportional hazards responder rate (PWPHRR) model for the treatment group is

$$S_T(t) = \begin{cases} S_C(t), & t \leq t_0, \\ p\left[S_C(t_0)\right]^{1-\delta} \left[S_C(t)\right]^{\delta} + (1-p)S_C(t), & t > t_0. \end{cases} \tag{5.3}$$

The density function for treatment group when $t > t_0$ can be written as

$$\begin{aligned} f_T(t) &= \frac{dF_T(t)}{dt} \\ &= \frac{d(1 - S_T(t))}{dt} \\ &= -p[S_C(t_0)]^{1-\delta}\delta[S_C(t)]^{\delta-1}\frac{dS_C(t)}{dt} - \frac{dS_C(t)}{dt} + p\frac{dS_C(t)}{dt} \\ &= p[S_C(t_0)]^{1-\delta}\delta[S_C(t)]^{\delta}\lambda_C(t) + S_C(t)\lambda_C(t) - pS_C(t)\lambda_C(t), \end{aligned}$$

where $\frac{dS_C(t)}{dt} = f_C(t) = \lambda_C(t)S_C(t)$. Hence the density function for treatment group can be written as

$$f_T(t) = \begin{cases} f_C(t), & t \leq t_0, \\ \{p\delta[S_C(t_0)]^{1-\delta}[S_C(t)]^{\delta} + (1-p)S_C(t)\}\lambda_C(t), & t > t_0 \end{cases}$$

and the corresponding hazard function is $\lambda_2(t) = f_2(t)/S_2(t)$ can be written as

$$\lambda_T(t) = \begin{cases} \lambda_C(t), & t \leq t_0, \\ \frac{\{p\delta[S_C(t_0)]^{1-\delta}[S_C(t)]^{\delta} + (1-p)S_C(t)\}\lambda_C(t)}{p[S_C(t_0)]^{1-\delta}[S_C(t)]^{\delta} + (1-p)S_C(t)}, & t > t_0. \end{cases}$$

The PWPHRR model (5.3) is a general model which includes special cases as the following

- $p = 1$ (fully response) and $t_0 = 0$ (no delay), the PWPHRR model reduces to the standard PH model (Schoenfeld, 1981);
- $p = 1$ (fully response) and $t_0 \neq 0$ (with delay), the PWPHRR model reduces to the the PWPH model (Xu et al., 2016).

Under the PWPHRR model, a two-sided hypothesis for testing the difference between survival distributions of the experimental treatment group and control group is represented by

$$H_0 : S_T(t) = S_C(t) \quad \text{vs.} \quad H_1 : S_2(t) \neq S_1(t),$$

and this hypothesis is equivalent to the following hypothesis for the hazards ratio and responder rate for treatment group:

$$H_0 : \delta = 1 \;\; \text{vs.} \quad H_1 : \delta \neq 1.$$

5.3 Sample size calculation

As we discussed in Chapter 3 and Chapter 4, the weighted log-rank test L is asymptotically standard normal distributed under the null hypothesis H_0. Thus, given a two-sided type I error rate α, we reject null hypothesis if $|L| > z_{1-\alpha/2}$.

Under a general fixed alternative hypothesis, same as we discussed in Chapters 3 and 4, we derived (Wei and Wu, 2020) an asymptotic distribution of the weighted log-rank test L, which is normally distributed with mean $\sqrt{n}\mu_w/\sigma_w$ and variance $\sigma^2/\tilde{\sigma}_w^2$, where μ_w, σ_w^2 and $\tilde{\sigma}_w^2$ are given in following equations (5.5), (5.6) and (5.7), respectively. Thus, sample size can be calculated using following formula

$$n = \frac{(\sigma_w z_{1-\alpha/2} + \tilde{\sigma}_w z_{1-\beta})^2}{\mu_w^2}, \tag{5.4}$$

where

$$\mu_w = \int_0^\infty w_R(t) \frac{\pi(t)(1-\pi(t))\{\lambda_1(t) - \lambda_2(t)\}}{\pi(t)\lambda_1(t) + \{1-\pi(t)\}\lambda_2(t)} V(t)dt, \tag{5.5}$$

$$\sigma_w^2 = \int_0^\infty w_R^2(t)\pi(t)\{1-\pi(t)\}V(t)dt, \tag{5.6}$$

$$\tilde{\sigma}_w^2 = \int_0^\infty w_R^2(t) \frac{\pi(t)(1-\pi(t))\lambda_1(t)\lambda_2(t)}{[\pi(t)\lambda_1(t) + \{1-\pi(t)\}\lambda_2(t)]^2} V(t)dt, \tag{5.7}$$

and $w_R(t)$ is the weight function for proposed model, function $V(t)$ is an incomplete density function of failure and $\pi(t)$ is a ratio of probability at risk of a subject belong to the control group versus the overall probability at risk of the two groups. It can be shown that

$$\begin{aligned} V(t) &= \{\omega_1\lambda_1(t)S_1(t) + \omega_2\lambda_2(t)S_2(t)\}G(t), \\ \pi(t) &= \frac{\omega_1 S_1(t)G(t)}{\omega_1 S_1(t)G(t) + \omega_2 S_2(t)G(t)}. \end{aligned}$$

where ω_1 and ω_2 are the allocation ratio to the control and treatment groups, respectively. This new formula (5.4) can be applied to the following special cases:

- $p = 1$ and $t_0 = 0$, sample size calculation under the standard PH model was derived by Schoenfeld (Schoenfeld, 1981);

- $p = 1$ and $t_0 > 0$, sample size calculation under the PWPH model was derived by Xu et al. (Xu et al., 2016).

Schoenfeld (Schoenfeld, 1981) showed that the optimal weighting function is given basically by the log hazards ratio function, that is weight function $w(t_j) \approx \log(\frac{\lambda_T(t_j)}{\lambda_C(t_j)})$, an optimal weight function for the log-rank test under the PWPHRR model when $t > t_0$ can be write as following by using Taylor expansion.

$$\begin{aligned} \log\left(\frac{\lambda_T(t)}{\lambda_C(t)}\right) &= \log\left(\frac{p\delta[S_C(t_0)]^{1-\delta}[S_C(t)]^{\delta} + (1-p)S_C(t)}{p\left[S_C(t_0)\right]^{1-\delta}\left[S_C(t)\right]^{\delta} + (1-p)S_C(t)}\right) \\ &= \log\left(1 - \frac{(1-\delta)p[S_C(t_0)]^{1-\delta}[S_C(t)]^{\delta}}{p\left[S_C(t_0)\right]^{1-\delta}\left[S_C(t)\right]^{\delta} + (1-p)S_C(t)}\right) \\ &\approx \frac{(1-\delta)p[S_C(t_0)]^{1-\delta}[S_C(t)]^{\delta}}{p\left[S_C(t_0)\right]^{1-\delta}\left[S_C(t)\right]^{\delta} + (1-p)S_C(t)}. \end{aligned}$$

Hence, we will use the following weight function

$$w_R(t) = \begin{cases} 0, & t \leq t_0, \\ \dfrac{1}{p\left[S_C(t_0)\right]^{1-\delta} + (1-p)[S_C(t)]^{1-\delta}}, & t > t_0 \end{cases} \tag{5.8}$$

for the sample size calculation under the piecewise proportional hazard responder rate model.

5.4 Simulation

In this section, we conduct simulations to study the performance of the proposed sample size formula for responder rate model and compare proposed weight function with other existing weight functions.

Performance of new sample size formula

To evaluate the accuracy of the proposed sample size formula (5.4), sample sizes were calculated under a PWPHRR Weibull model for the following parameter settings: The Weibull distribution of the control group was $S(t) = e^{-\lambda t^{\kappa}}$; hazard ratio changing time point was set to $t_0 = 6$ months and the proportion of control patients who could survive beyond t_0 was set to $S_1(t_0) = 90\%$; the responder rate in treatment group was set as $p = 0.2, 0.4$ and 0.6; hazard ratio δ between responders in treatment group and control groups was set as 0.01, 0.05 and 0.1; assuming a uniform accrual with accrual duration $t_a = 12$ months and follow-up time $t_f = 24$ months; the shape parameter of the Weibull was set at $\kappa = 0.7, 1$, and 1.3 to represent the decreasing, constant and increasing hazard functions, respectively; sample size allocation ratio was set to $\omega_1 = 1/2$ (1:1 allocation for control and treatment group), 1/3 (1:2 allocation and more subjects assigned to the treatment group) and 2/3 (2:1 allocation and more subjects assigned to the control group). Random samples for the PWPHRR Weibull model were generated according to the method given in Appendix I. Assuming no loss to follow up, sample sizes were calculated with a two-sided type I error of 5% and a power of 80%. Empirical powers were estimated by performing 10,000 simulation runs. The simulation results for the new formula (5.4) are shown in Table 5.1.

Table 5.1: Sample sizes (n) were calculated using formula (5.4) under the Weibull delayed treatment effect model with $S_1(t_0) = 90\%$, the proportion of subjects who could survive beyond the delay time $t_0 = 6$ months, a two-sided type I error of 5%, power of 80%. The corresponding empirical type I errors ($\hat{\alpha}$) and powers ($1-\hat{\beta}$) were estimated by performing 10,000 simulation runs.

		$\kappa = 0.7$			$\kappa = 1$			$\kappa = 1.3$		
ω_1	δ/p	n	$\hat{\alpha}$	$1-\hat{\beta}$	n	$\hat{\alpha}$	$1-\hat{\beta}$	n	$\hat{\alpha}$	$1-\hat{\beta}$
1/2	.01/.2	2727	.049	.804	1605	.050	.804	1010	.051	.798
(1:1)	.01/.4	616	.047	.805	370	.050	.798	239	.048	.799
	.01/.6	238	.050	.795	146	.053	.804	96	.049	.802
	.05/.2	3007	.048	.800	1781	.049	.802	1129	.047	.802
	.05/.4	684	.053	.798	413	.053	.796	269	.051	.799
	.05/.6	267	.052	.802	164	.053	.801	110	.049	.801
	.1/.2	3415	.049	.800	2038	.045	.797	1305	.049	.804
	.1/.4	783	.050	.799	476	.053	.798	313	.050	.799
	.1/.6	310	.048	.801	192	.048	.798	129	.049	.802
1/3	.01/.2	3032	.052	.804	1789	.051	.803	1129	.049	.805
(1:2)	.01/.4	674	.054	.804	406	.051	.808	264	.052	.814
	.01/.6	254	.054	.802	156	.047	.802	104	.054	.802
	.05/.2	3346	.054	.806	1986	.051	.804	1263	.050	.803
	.05/.4	750	.049	.798	454	.048	.808	298	.054	.811
	.05/.6	287	.050	.794	177	.047	.803	119	.050	.808
	.1/.2	3803	.048	.793	2274	.051	.797	1461	.050	.808
	.1/.4	860	.052	.799	525	.056	.803	347	.051	.808
	.1/.6	334	.051	.805	208	.047	.798	141	.053	.805
2/3	.01/.2	3104	.050	.808	1823	.044	.796	1143	.048	.793
(2:1)	.01/.4	713	.047	.796	426	.048	.801	274	.051	.808
	.01/.6	282	.047	.789	172	.046	.793	113	.051	.800
	.05/.2	3421	.051	.800	2022	.049	.800	1278	.053	.798
	.05/.4	790	.053	.802	475	.050	.803	308	.054	.812
	.05/.6	316	.049	.785	193	.051	.801	128	.049	.795
	.1/.2	3882	.051	.793	2313	.049	.802	1477	.049	.791
	.1/.4	902	.046	.806	547	.048	.801	357	.048	.805
	.1/.6	365	.049	.795	225	.051	.795	150	.051	.795

The results in table 5.1 showed that the simulated empirical type I error rates and powers were all close to the nominal levels. Thus, the proposed weighted log-rank test preserved type I error rate and sample size formula provided accurate sample size estimation for either balance design or unbalance design.

Evaluation of study efficiency by parameters setting

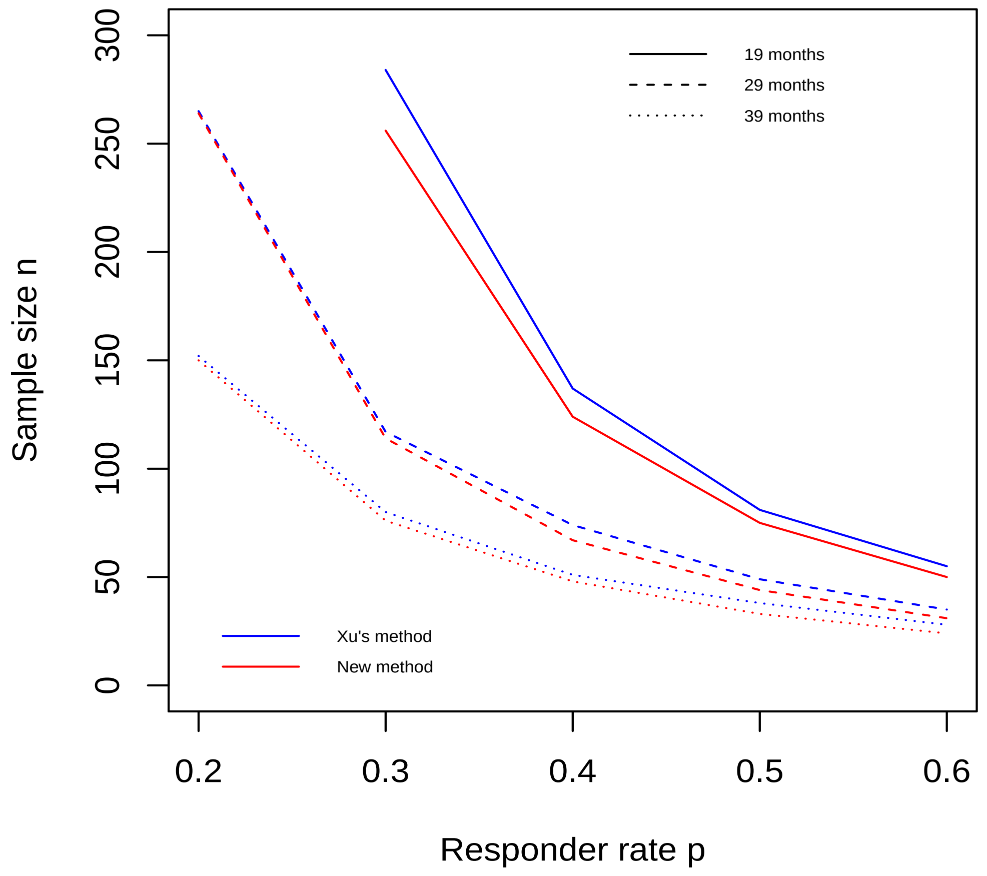

Figure 5.1: The relationship between sample size and responder rate under different trial durations. Hazard ratio for responding patients is 0.01 and $t_0 = 2$ months.

Three figures have similar tends between responder rate p and sample size n, that the sample size decreases as responder rate increases. What is more, there is no too much difference in sample size among three scenarios when responder rate is high (p=0.6), but significant differences are present when responder rate is low (p=0.2). Our proposed method performs better than Xu's method under three scenarios in the same setting, in the other words, our method need less sample size in order to achieve the target power compared with Xu's PRIME design.

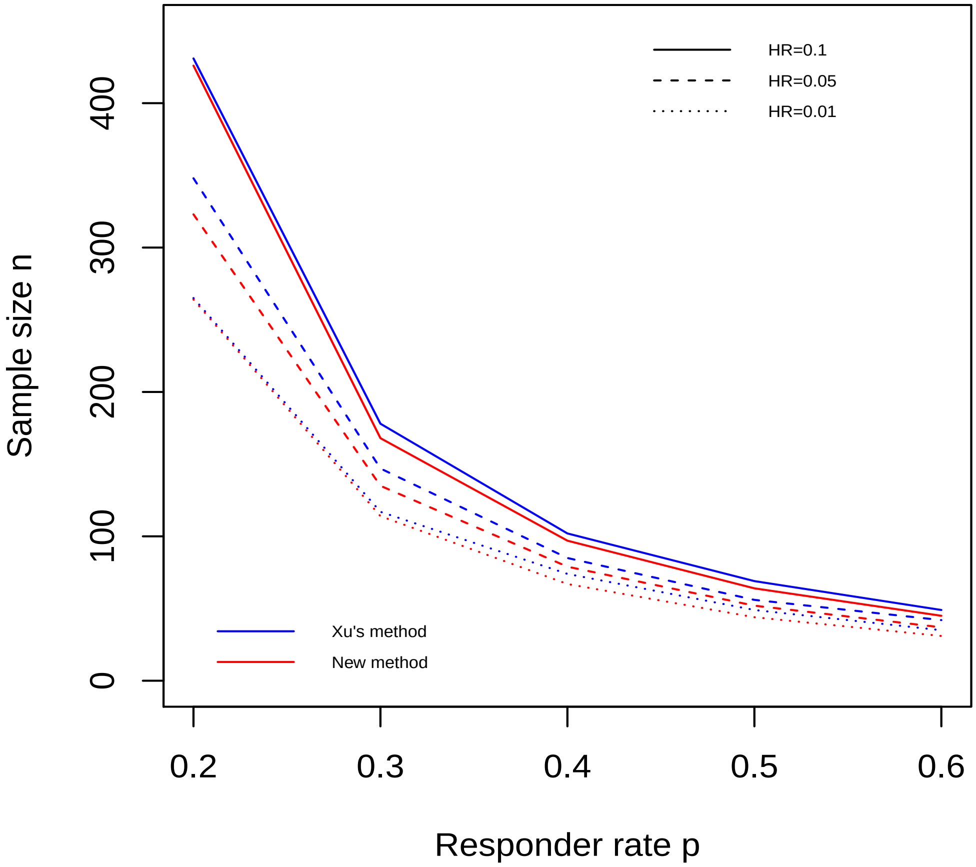

Figure 5.2: The relationship between sample size and responder rate under different hazard ratios of responding patients. Study duration is 29 months and $t_0 = 2$ months.

We explored the relationship between the responder rate in treatment group and the sample size under different scenarios of parameters setting based on our proposed new formula and Xu's PRIME design (Xu et al., 2020). Figure 5.1, 5.2 and 5.3 include the trial

Table 5.2: Formulas for different weight functions.

Weight	P-W	Responder	MWLRT	Milestone
$t \leq t_0$	0	0	0	0
$t > t_0$	1	$\frac{1}{p[S_C(t_0)]^{1-\delta}+(1-p)[S_C(t)]^{1-\delta}}$	$1/\max\{\hat{S}(t_j-), S(\hat{t_0})\}$	$1/\max\{\hat{S}(t_j-), 0.5\}$

parameters of interest such as trial durations, hazard ratios between responder in treatment group and control group and delayed change points.

Figure 5.1 illustrates the relationship between sampler size and responder rate under various trial duration based on two methods. A larger sample size (n= 256) is requested with less study duration time (19 months) at the same responder rate (p= 0.3) and if the study duration is longer (39 months), the trial needs fewer subjects (n=76) for target power in proposed new methods. Hazards ratio between responders in treatment group and control group also affect the the sample size when response rate p is fixed in figure 5.2. For example, when the hazards ratio is 0.1, sample size n decreases from 426 to 45 as response rate p increases from 0.2 to 0.6. Similar results when hazards ratio is 0.05 and 0.01 can be obtained, sample size changed from 426 to 323, then from 323 to 264 as the hazards ratio changed from 0.1 to 0.05, then 0.05 to 0.01 at fixed responder rate (p=0.2). A larger subjects (n=391) is required to achieve the targeted power in trial design when the pre-specified delayed change point is larger ($t_0 = 4$) in figure 5.3.

Weight functions comparison

We compared proposed weight function with other existing weight functions we discussed in introduction part and all weight function formulas are shown in table 5.2. All sample sizes in table 5.3 were calculated under the PWPHRR model using weight function in table 5.2, where the distribution of the control group is the Weibull distribution $S_C(t) = e^{-\lambda t^{\kappa}}$, response rate of the treatment group is set between 0.2 and 0.6, and fixed delay time is set to $t_0 = 2$ months, with other design parameters set as follows: Hazard ratio δ is set as 0.01, 0.05 and 0.1; accrual rate is 36.8 subjects per month and total study duration is 29 months; the shape parameter of the Weibull distribution is set to $\kappa = 1$; the hazard parameter is set as $\lambda = 0.0737$ of the control group; and sample size allocation ratio is set to $\omega_1 = 1/2$

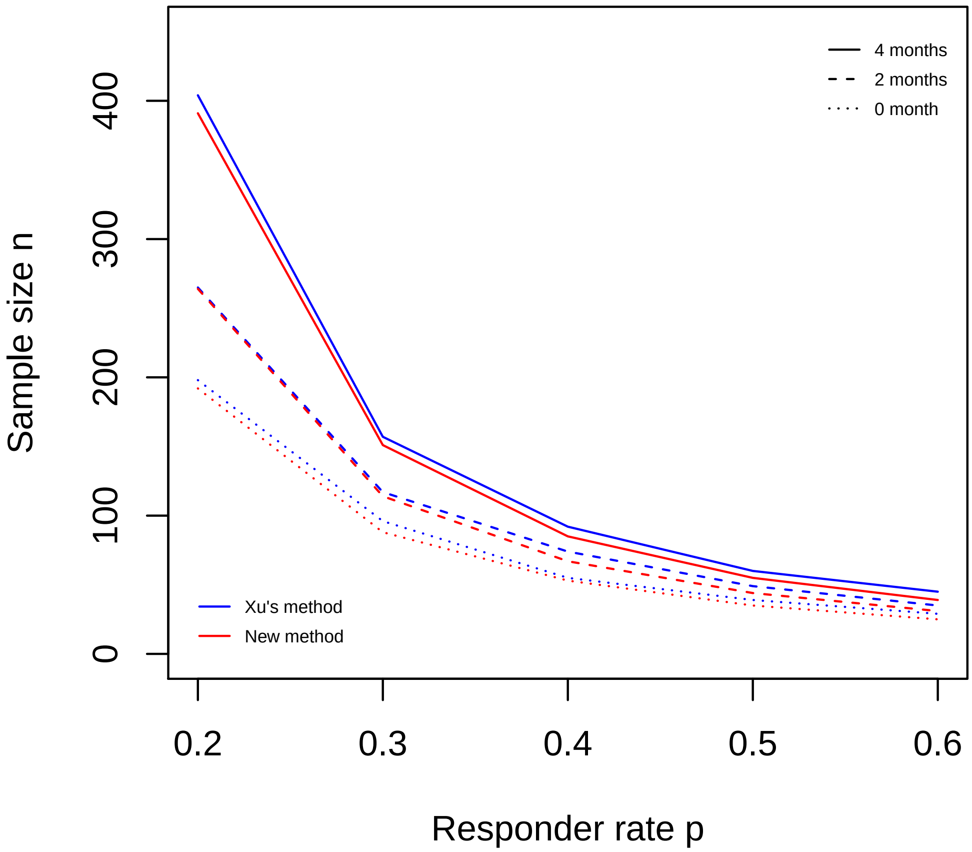

Figure 5.3: The relationship between sample size and responder rate under different delayed time points. Hazard ratio for responding patients is 0.01 and study duration is 29 months.

(1:1 equal allocation). Assuming no loss to follow up, sample sizes were calculated with a two-sided type I error of 5% and power of 80%. Empirical powers were estimated by performing 10,000 simulation runs.

Table 5.3 shows that sample size derived using responder weight function is more efficient than other weight functions for the targeted power. Milestone weight function enroll fewer subjects than piecewise weight function when HR=0.01 or vice versa when HR=0.1. MWLRT function performs worst when compared with other weights function under responder rate model. Therefore, the choice among weight functions should be made care-

Table 5.3: Sample size (n) were calculate by the new formula under different weight functions.

p	P-W			Responder			MWLRT			Milestone		
	n	$1-\hat{\beta}$	$\hat{\alpha}$	n	$1-\hat{\beta}$	$\hat{\alpha}$	n	$1-\hat{\beta}$	$\hat{\alpha}$	n	$1-\hat{\beta}$	$\hat{\alpha}$
$HR = 0.01$												
0.2	321	.804	.048	264	.809	.050	398	.793	.048	315	.783	.049
0.3	132	.800	.048	114	.808	.048	158	.798	.051	130	.775	.051
0.4	75	.804	.054	67	.820	.051	89	.800	.052	75	.790	.052
0.5	48	.798	.052	44	.812	.049	58	.792	.057	49	.792	.055
0.6	33	.804	.057	31	.822	.050	41	.815	.062	34	.774	.058
$HR = 0.05$												
0.2	388	.798	.053	323	.806	.051	495	.801	.053	386	.789	.050
0.3	155	.797	.051	135	.809	.051	184	.794	.052	154	.784	.052
0.4	87	.804	.054	79	.816	.051	104	.801	.053	87	.787	.054
0.5	56	.802	.051	52	.814	.049	67	.800	.053	57	.799	.056
0.6	39	.810	.052	37	.821	.052	47	.806	.052	40	.780	.056
$HR = 0.1$												
0.2	507	.803	.047	426	.803	.046	726	.794	.050	517	.785	.049)
0.3	189	.809	.053	168	.804	.053	226	.792	.054	190	.782	.054)
0.4	106	.800	.053	97	.820	.050	125	.801	.051	107	.796	.057)
0.5	68	.808	.051	64	.810	.046	81	79.6	.052	70	.794	.056)
0.6	48	.801	.053	45	.822	.051	57	80.9	.056	50	.793	.057)

fully for non-proportional hazards model in cancer immunotherapy trial design.

5.5 Example

Borghaei et al. (Borghaei et al., 2015) conducted a phase III, immunotherapy vs. chemotherapy trial for non-squamous non-small cell lung cancer (NSCLC) whose disease progresses after first-line chemotherapy are limited. Patients after failure of platinum double were randomly assigned in a 1:1 ratio to receive either Docetaxel (chemotherapy) or Nivolumab (PD-1), and the primary endpoint for the trial is overall survival (OS). The observed median OS for Docetaxel group is 9.4 months (baseline hazard rate is 0.074 under exponential distribution) and the overall hazard ratio between Nivolumab and Docetaxel group is 0.73. Consider a total study duration is 29 months and enrollment rate for patients is 36.8 subjects/months, the sample size required for the study is 582 under 90% power and two side 5% type I error setting.

However, visual separation of Kaplan-Meier curves for OS has been observed approximately 2 months after randomization and the responder rate in Nivolumab group is approximately 20% in Borghaei's study (Borghaei et al., 2015). Since the original trial design didn't consider delayed treatment effect and the responder rate in treatment group, we illustrate sample size recalculation to incorporate both delayed treatment effect and responder rate in treatment group.

It is assumed that the OS times for patients receiving Docetaxel follow an exponential distribution, whereas the OS times for patients receiving Nivolumab follow a piecewise exponential distribution with a delay time $t_0 = 2$ months and responder rate $p = 0.2$ as follows

$$\begin{aligned} S_1(t) &= e^{-0.074t} \\ S_2(t) &= \begin{cases} e^{-0.074t} & 0 \leq t < 2 \\ 0.2ce^{-\delta\lambda t} + 0.8e^{-0.074t} & t \geq 2, \end{cases} \end{aligned}$$

where $c = e^{-0.074*2*(1-\delta)}$ is a normalizing constant, and δ is the hazard ratio between the responders in Nivolumab group and patients in Docetaxel group after 2 months. Xu et al. used a simulation-based grid searching algorithm (Xu et al., 2020) to explore responder hazard ratio and get hazard ratio $\delta_2 = 0.01$ when overall hazard ratio δ_1 is 0.73 and responder rate is 0.2 (Figure 5.4). Thus, assuming patients are accrued to the trial with enrollment rate 36.8 subjects/months and the study duration is 29 months. Using the new formula, the sample size is 392, to achieve 90% power with a two-sided type I error of 5%. The R code for the sample size calculation is provided in Appendix J.

5.6 Discussion

Delayed treatment effect and long term survival are two challenges in cancer immunotherapy trials design which violate the proportional hazards assumption. Other causes of non-proportional hazards patten such as responder rate in treatment group are discussed in this section. Xu et al. (Xu et al., 2020) illustrated this kind of responder rate in treatment group in immunotherapy trial design and proposed a PRIME approach to incorporate the dichotomized response incurred from nonresponders in treatment group. However, their method used PRIME likelihood test and more complex in sample size and power calculation compared with our methods.

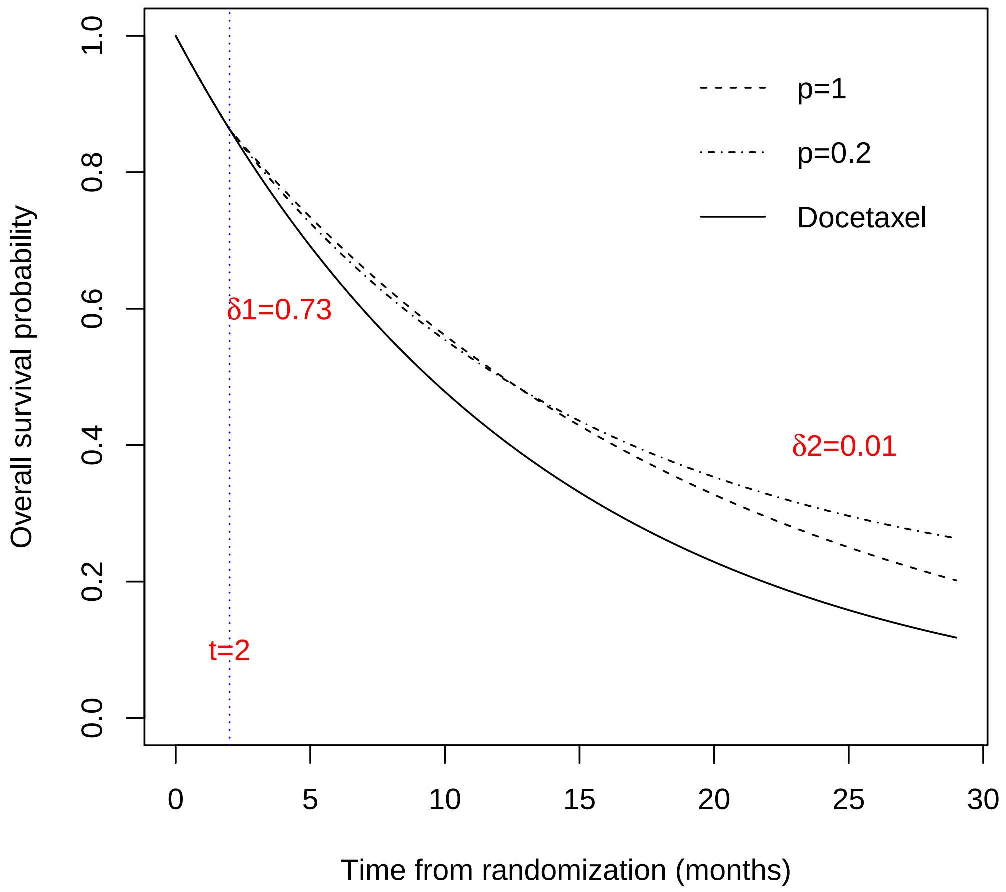

Figure 5.4: Survival curves for the Docetaxel and Nivolumab groups.

Same as PWPHCR model discussed in Chapter 3, the PWPHRR model discussed in this chapter assumes that the delayed treatment effect is homogeneous across the individual subjects. It is more natural to assume that the effect may vary heterogeneously across individuals, in which case a random delayed effect model would be more appropriate. Our proposed method can be extended to the random delayed effect model with responder rate as well. It is also possible to extend the proposed method to a general delayed treatment effect model with random lag time by using weighted log-rank test. How to choose the weight function is also needed to be considered in the extended model.

In real trial design, the responder rate in treatment group needs to be pre-specified.

Usually we make assumption for the responder rate from pilot data during the trial design. However, the true responder is unknown in advance, the mis-specification is inevitable when doing trial design. So how to develop a robust method to choose responder rate is another extension in the future.

Chapter 6 Summary

6.1 Summary and conclusion

In this dissertation, new statistical models used to design and analysis cancer immunotherapy trials were introduced. Delayed treatment effect, long term survivors, responders and non-responders in treatment groups are underlying causes of non-proportional hazards patterns in cancer immunotherapy trials. As a result, a traditional survival trial design based on standard log-rank test will lead to substantial loss of power.

A piecewise weighted log-rank test is proposed to incorporate the delayed treatment effect into consideration of the trial design and derive sample size under a fixed alternative hypothesis for the proposed piecewise proportional hazards (PWPH) model. This new sample size formula provides accurate sample size estimation for both balance and unbalance design regardless of the size of hazard ratio.

A piecewise proportional hazard cure rate (PWPHCR) model is proposed to incorporate both delayed treatment effect and cure rate into the trial design consideration. Sample size formula also is derived under a fixed alternative hypothesis. The accuracy of sample size calculation using this new formula is assessed and compared with existing methods via simulation studies.

A more general and suitable random delayed cure rate model was proposed to design cancer immunotherapy trials. F_τ weighted log-rank test is used to do sample size calculation. The sensitivity for mis-specifying the random delay lag time duration and distributions is also studied via simulation.

A limited percentage of patients would response to the treatment in reality. In light of this, we need to treat patients as non-responders and responders in treatment group. A piecewise proportional hazard responder rate (PWPHRR) model considering responders and non-responders in treatment group is proposed and a sample size formula is derived under w_R weighted log-rank test for canner immunotherapy trials design. Simulations are conducted to study the performance of the proposed sample size formula under various weight functions.

6.2 Future work

We proposed several statistical models, weighted log-rank tests and sample size formulas to deal with NPH patterns in cancer immunotherapy trial design. An R package included all discussed models in this thesis will be developed for implementation later.

Other problems in cancer immunotherapy trial design such as how to do interim analysis or sample size calculation in adaptive design are also need to be considered in the future.

At last, more general models combining fixed/random delayed effect, cure rate or response and non-response rate together will be proposed to satisfy more complex scenarios in cancer immunotherapy trial design.

Appendices

Appendix A: Derivation of the probability of failure

Assume that patients are accrued over a time period t_a, with an additional follow-up time t_f, so that the study duration $\tau = t_a + t_f$, and the entry time Y is uniformly distributed over $[0, t_a]$ with distribution $H(t)$. If no patient drops out or is lost to follow-up, the administrative censoring time $\tau - Y$ follows survival distribution $G(t) = H(\tau - t)$ which is uniform over the interval $[t_f, t_a + t_f]$. Let T be the event time with survival distribution $S_1(t)$ for the control group. The probability of a participant in the control group having an event before calendar time $t(> t_0)$ but after the delayed phase can be calculated by

$$\begin{aligned} p_1 &= P\{(t_0 < T) \cap (T \leq \tau - Y)\} \\ &= \int_0^\infty P\{(t_0 < T) \cap (T \leq \tau - Y) | Y = x\} dH(x) \\ &= \int_0^\infty P\{(t_0 < T \leq \tau - x)\} dH(x) \\ &= \frac{1}{t_a} \int_0^{t_a} \{S_1(t_0) - S_1(\tau - x)\} dx \\ &= S_1(t_0) - \frac{1}{t_a} \int_{t_f}^{t_a + t_f} S_1(t) dt, \quad t > t_0, \end{aligned}$$

under the delayed treatment effect model, it is easy to show

$$S_2(t) = \{S_1(t_0)\}^{1-\delta} \{S_1(t)\}^\delta, \quad t > t_0.$$

Thus, the probability of a participant in the treatment group having an event before calendar time t but after the delayed phase can be calculated by

$$\begin{aligned} p_2 &= S_2(t_0) - \frac{1}{t_a} \int_{t_f}^{t_a + t_f} S_2(t) dt \\ &= \{S_1(t_0)\}^{1-\delta} \left[\{S_1(t_0)\}^\delta - \frac{1}{t_a} \int_{t_f}^{t_a + t_f} \{S_1(t)\}^\delta dt \right], \quad t > t_0. \end{aligned}$$

which are formulae p_1 and p_2 given by equations (2.7) and (2.8), respectively.

Appendix B: Derivation of asymptotic distribution of the piecewise weighted log-rank test

Assume that n patients are allocated between the control and treatment groups, which are designated groups 1 and 2, respectively. Let D be the set of identifiers in the two groups who died, and let t_j be the death time of the j^{th} patient in either group. We assume that the $\{t_j\}$ are distinct. Let y_j be an indicator variable of the control group, that is, $y_j = 1$ if the j^{th} event belongs to the control group and $y_j = 0$ if the j^{th} event belongs to the treatment group. If we define $n_i(t)$ to be the number at risk just before time t in group i, then, the weighted log-rank score can be expressed as

$$U = \sum_{j \in D} w_j \{y_j - p(t_j)\},$$

where $p(t_j) = n_1(t_j)/\{n_1(t_j) + n_2(t_j)\}$ and $\{w_j\}$ are a set of predetermined weights. The weighted log-rank test is given by

$$L = \frac{\sum\limits_{j \in D} w_j \{y_j - p(t_j)\}}{\left[\sum\limits_{j \in D} w_j^2 p(t_j)\{1 - p(t_j)\}\right]^{1/2}}.$$

Conditional on $n_1(t_j)$ and $n_2(t_j)$, the $\{y_j\}$ are a sequence of Bernoulli random variables with means

$$\mu_j = \frac{n_1(t_j)\lambda_1(t_j)}{n_1(t_j)\lambda_1(t_j) + n_2(t_j)\lambda_2(t_j)}$$

and variances $\mu_j(1 - \mu_j)$, where $\lambda_i(t)$ is the hazard function of group i. To derive the asymptotic distribution, we define function $\pi(t)$ be the ratio of probability a subject in group 1 being at risk at time t vs. overall probability of the subject at risk at time t and $V(t)$ be the incomplete density function of failure at time t, given as

$$\pi(t) = \frac{\omega_1 S_1(t) G(t)}{\omega_1 S_1(t) G(t) + \omega_2 S_2(t) G(t)} \tag{B.1}$$

and

$$V(t) = \{\omega_1 \lambda_1(t) S_1(t) + \omega_2 \lambda_2(t) S_2(t)\} G(t), \tag{B.2}$$

where $S_i(t)$ is the survival distribution of group i, ω_i is the proportion of subjects assigned to group i, and $G(t)$ is the common survival distribution of the censoring time for the two groups (Schoenfeld, 1981). The log-rank test L can be written as

$$
\begin{aligned}
L &= \frac{\sum\limits_{j\in D} w_j\{y_j - p(t_j)\}}{\left[\sum\limits_{j\in D} w_j^2 p(t_j)\{1-p(t_j)\}\right]^{1/2}} \\
&= \frac{\sum\limits_{j\in D} w_j^2\{y_j - \mu_j\}}{\left[\sum\limits_{j\in D} w_j^2 \mu_j(1-\mu_j)\right]^{1/2}} \times \frac{\left[\sum\limits_{j\in D} w_j^2 \mu_j(1-\mu_j)\right]^{1/2}}{\left[\sum\limits_{j\in D} w_j^2 p(t_j)\{1-p(t_j)\}\right]^{1/2}} \\
&\qquad + \frac{\sum\limits_{j\in D} w_j\{\mu_j - p(t_j)\}}{\left[\sum\limits_{j\in D} w_j^2 p(t_j)\{1-p(t_j)\}\right]^{1/2}} \\
&= I_1 \times I_2 + I_3.
\end{aligned}
$$

Using the martingale central limit theorem (Fleming and Harrington, 1991), we can show that the first term I_1 has a limiting standard normal distribution. As

$$
\begin{aligned}
\mu_j - p(t_j) &= \frac{n_1(t_j)}{n_1(t_j) + n_2(t_j)\delta(t_j)} - \frac{n_1(t_j)}{n_1(t_j) + n_2(t_j)} \\
&= \frac{n_1(t_j)n_2(t_j)\{1-\delta(t_j)\}}{\Big\{n_1(t_j) + n_2(t_j)\Big\}\Big\{n_1(t_j) + n_2(t_j)\delta(t_j)\Big\}} \\
&= \frac{p(t_j)\{1-p(t_j)\}(1-\delta(t_j))}{\Big[p(t_j) + \{1-p(t_j)\}\delta(t_j)\Big]},
\end{aligned}
$$

where $\delta(t) = \lambda_2(t)/\lambda_1(t)$, replacing $p(t_j)$ by its limit $\pi(t_j)$, we have

$$
\begin{aligned}
& n^{-1}\sum_{j\in D} w_j\{\mu_j - p(t_j)\} \\
& \xrightarrow{P} \int_0^\infty w(t)\frac{\pi(t)\{1-\pi(t)\}\{1-\delta(t)\}}{\Big[\pi(t) + \{1-\pi(t)\}\delta(t)\Big]}V(t)dt \\
& = \mu,
\end{aligned}
$$

and

$$\begin{aligned} & n^{-1}\sum_{j\in D} w_j^2 p(t_j)\{1-p(t_j)\} \\ & \quad \xrightarrow{P} \int_0^\infty w(t)^2\pi(t)\{1-\pi(t)\}V(t)dt = \sigma^2. \end{aligned}$$

Thus, the third term, I_3, converges to

$$\frac{\sum\limits_{j\in D} w_j\{\mu_j - p(t_j)\}}{\left[\sum\limits_{j\in D} w_j^2 p(t_j)\{1-p(t_j)\}\right]^{1/2}} - \sqrt{n}e \xrightarrow{P} 0,$$

where $e = \mu/\sigma$. We can further show

$$\begin{aligned} & n^{-1}\sum_{j\in D} w_j^2\mu_j(1-\mu_j) \\ & = n^{-1}\sum_{j\in D} w_j^2 \frac{n_1(t_j)n_2(t_j)\delta(t_j)}{\Big\{n_1(t_j)+n_2(t_j)\delta(t_j)\Big\}^2} \\ & = n^{-1}\sum_{j\in D_2} w_j^2 \frac{p(t_j)(1-p(t_j))\delta(t_j)}{\Big[p(t_j)+\{1-p(t_j)\}\delta(t_j)\Big]^2} \\ & \xrightarrow{P} \int_0^\infty w^2(t)\frac{\pi(t)\{1-\pi(t)\}\delta(t)}{\Big[\pi(t)+\{1-\pi(t)\}\delta(t)\Big]^2}V(t)dt = \tilde{\sigma}^2. \end{aligned}$$

and it follows that

$$I_2 = \frac{\left\{\sum\limits_{j\in D} w_j^2\mu_j(1-\mu_j)\right\}^{1/2}}{\left[\sum\limits_{j\in D} w_j^2 p(t_j)\{1-p(t_j)\}\right]^{1/2}} \xrightarrow{P} \frac{\tilde{\sigma}}{\sigma}.$$

Combining these results, we have shown that the log-rank test L is asymptotically normally distributed with a variance $\tilde{\sigma}^2/\sigma^2$ and mean $\sqrt{n}e$, where $e = \mu/\sigma$.

We now consider the delayed treatment effect model (2.3), and using the piecewise weight function $w(t) = 0$ when $t \le t_0$ and $w(t) = 1$ when $t > t_0$, and hazard ratio $\delta(t) = 1$ when $t \le t_0$ and $\delta(t) = \delta$ when $t > t_0$ and substituting $\pi(t)$ of equation (B.1), $V(t)$ of equation (B.2) and $S_2(t) = [S_1(t_0)]^{1-\delta}[S_1(t)]^\delta$ into μ, σ^2 and $\tilde{\sigma}^2$, we obtain the following expressions

$$
\begin{aligned}
\mu &= \omega_1\omega_2(1-\delta)c(\delta)\int_{t_0}^{\infty}\frac{\{S_1(t)\}^{\delta}G(t)\lambda_1(t)}{\big[\omega_1+\omega_2 c(\delta)\{S_1(t)\}^{\delta-1}\big]}dt,\\
\sigma^2 &= \omega_1\omega_2 c(\delta)\int_{t_0}^{\infty}\frac{\{S_1(t)\}^{\delta}[\omega_1+\omega_2\delta c(\delta)\{S_1(t)\}^{\delta-1}]G(t)\lambda_1(t)}{[\omega_1+\omega_2 c(\delta)\{S_1(t)\}^{\delta-1}]^2}dt,\\
\tilde{\sigma}^2 &= \omega_1\omega_2\delta c(\delta)\int_{t_0}^{\infty}\frac{\{S_1(t)\}^{\delta}G(t)\lambda_1(t)}{[\omega_1+\omega_2\delta c(\delta)\{S_1(t)\}^{\delta-1}]}dt,
\end{aligned}
$$

where $c(\delta) = \{S_1(t_0)\}^{1-\delta}$.

Appendix C: Generating random number under the delayed treatment effect model

Under the PWPH model (2.3), we have

$$S_2(t) = \begin{cases} S_1(t), & t \le t_0, \\ \{S_1(t_0)\}^{1-\delta}\{S_1(t)\}^{\delta}, & t > t_0. \end{cases}$$

Assume that T is a random variable with survival distribution $S_2(t)$. Then, $U = S_2(T)$ is a uniform random variable on interval $[0, 1]$. If $U \ge S_1(t_0)$, then $U = S_1(T)$, and thus $T = S_1^{-1}(U)$. If $U < S_1(t_0)$, then $U = c\{S_1(T)\}^{\delta}$, where $c = \{S_1(t_0)\}^{1-\delta}$, and thus $T = S_1^{-1}\{(U/c)^{1/\delta}\}$. Therefore, a random variable T can be generated, which follows survival distribution $S_2(t)$ as follows:

$$T = S_2^{-1}(U) = \begin{cases} S_1^{-1}(U), & U \ge S_1(t_0), \\ S_1^{-1}\{(U/c)^{1/\delta}\}, & U < S_1(t_0). \end{cases}$$

For the Weibull distribution $S_1(t) = e^{-\lambda t^{\kappa}}$, solving t, its inverse function, is given by $t = S_1^{-1}(u) = \{-\log(u)/\lambda\}^{1/\kappa}$.

Appendix D: R code used for sample size calculation in Chapter 2

The R function ‘Size’ used for the sample size calculation in Section 5 is given below. ‘Size’ has implemented the sample size calculation using formulae (2.6) and (2.9) with the Weibull distribution. It can be modified to accommodate other parametric or non-parametric logspline survival distribution. The input parameters in the R function ‘Size’: kappa is the Weibull shape parameter; lambda is the hazard parameter of the control group; delta is the hazard ratio; alpha and beta are the type I and II errors; ta and tf are accrual duration and follow-up period; t0 is the fixed delay time; omega is the sample size allocation ratio.

```
####################################################################
### kappa is the shape parameter of the Weibull distribution; ######
### lambda is the hazard parameter of control group; delta is ######
### the hazard ratio; alpha and beta are type I and II error; ######
### ta and tf are accrual and follow-up durations; t0 is the  ######
### lag time, omega is the sample size allocation ratio.      ######
####################################################################
Size=function(kappa, lambda, delta, alpha, beta, ta, tf, t0, omega)
{S1=function(t){exp(-lambda*t^kappa)}
 S2=function(t){S1(t)^delta}
 h1=function(t){lambda*kappa*t^(kappa-1)}
 G=function(t){1-punif(t, tf, ta+tf)}
 c=S1(t0)^(1-delta)
 m1=function(t){(S1(t)^delta)/(omega+(1-omega)*c*S1(t)^(delta-1))}
 m2=function(t){(S1(t)^delta)*(omega+(1-omega)*delta*c*S1(t)^(delta-1))/
                (omega+(1-omega)*c*S1(t)^(delta-1))^2}
 m3=function(t){(S1(t)^delta)/(omega+(1-omega)*delta*c*S1(t)^(delta-1))}
 f1=function(t){m1(t)*G(t)*h1(t)}
 f2=function(t){m2(t)*G(t)*h1(t)}
 f3=function(t){m3(t)*G(t)*h1(t)}
 I1=integrate(f1, t0, ta+tf)$value
 I2=integrate(f2, t0, ta+tf)$value
 I3=integrate(f3, t0, ta+tf)$value
 mu=(1-delta)*omega*(1-omega)*c*I1
 var1=omega*(1-omega)*c*I2
 var2=omega*(1-omega)*delta*c*I3
 z0=qnorm(1-alpha/2); z1=qnorm(1-beta)
 nW=(sqrt(var1)*z0+sqrt(var2)*z1)^2/mu^2  ## new formula (8)
 p1=S1(t0)-integrate(S1, tf, ta+tf)$value/ta
 p2=c*(S1(t0)^delta-integrate(S2, tf, ta+tf)$value/ta)
 P=omega*p1+(1-omega)*p2
 dW=ceiling(nW*P)
 dX=(z0+z1)^2/(omega*(1-omega)*log(delta)^2)
 nX=ceiling(dX/P) ## Xu's formula (3.5)
 ans=list(c(dX=ceiling(dX),nX=nX,dW=dW, nW=ceiling(nW)));
     return(ans)}
Size(kappa=1,lambda=0.01,delta=0.72,alpha=0.05,beta=0.1,ta=30,tf=50,
     t0=6,omega=1/2)
 dX   nX   dW   nW   # X and W refer to Xu and New method #
390 1050  391 1051   # d and n refer to events and sample size #
```

Appendix E: Derivation of the asymptotic distribution of the weighted log-rank test under PWPHCR model

The weighted log-rank test L_w is given by

$$L = \frac{\sum\limits_{j\in D} w_j\{y_j - p(t_j)\}}{\left[\sum\limits_{j\in D} w_j^2 p(t_j)\{1-p(t_j)\}\right]^{1/2}},$$

where $p(t_j) = n_1(t_j)/\{n_1(t_j)+n_2(t_j)\}$ and $w_j = W(t_j)$. Conditionally on $n_1(t)$ and $n_2(t)$, the $\{y_j\}$ are a sequence of Bernoulli random variables with means

$$\mu_j = \frac{n_1(t_j)\lambda_1(t_j)}{n_1(t_j)\lambda_1(t_j) + n_2(t_j)\lambda_2(t_j)}$$

and variances $\mu_j(1-\mu_j)$, where $\lambda_i(t)$ is the hazard function of group i. To derive the asymptotic distribution, we define the functions

$$\begin{aligned} V(t) &= \{\omega_1\lambda_1(t)S_1(t) + \omega_2\lambda_2(t)S_2(t)\}G(t), \\ \pi(t) &= \frac{\omega_1 S_1(t)G(t)}{\omega_1 S_1(t)G(t) + \omega_2 S_2(t)G(t)}. \end{aligned}$$

Under the PWPHCR model, we have

$$\begin{aligned} \mu_j - p(t_j) &= \frac{n_1(t_j)\lambda_1(t_j)}{n_1(t_j)\lambda_1(t_j) + n_2(t_j)\lambda_2(t_j)} - \frac{n_1(t_j)}{n_1(t_j)+n_2(t_j)} \\ &= \frac{p(t_j)\{1-p(t_j)\}\{\lambda_1(t_j) - \lambda_2(t_j)\}}{p(t_j)\lambda_1(t_j) + \{1-p(t_j)\}\lambda_2(t_j)}. \end{aligned}$$

Replacing $w_j = W(t_j)$ and $p(t_j)$ by their limits $w(t_j)$ and $\pi(t_j)$, we obtain

$$\begin{aligned} & n^{-1}\sum_{j\in D} w_j\{\mu_j - p(t_j)\} \\ & \quad \to \int_0^\infty w(t)\frac{\pi(t)(1-\pi(t))\{\lambda_1(t)-\lambda_2(t)\}}{\pi(t)\lambda_1(t) + \{1-\pi(t)\}\lambda_2(t)} V(t)dt = \mu_w \end{aligned} \tag{E.1}$$

and

$$\begin{aligned} & n^{-1}\sum_{j\in D} w_j^2 p(t_j)\{1-p(t_j)\} \\ & \quad \to \int_0^\infty w^2(t)\pi(t)\{1-\pi(t)\}V(t)dt = \sigma_w^2. \end{aligned} \tag{E.2}$$

The weighted log-rank test L_w can be written as

$$\begin{aligned}
L &= \frac{\sum\limits_{j\in D} w_j\{y_j - p(t_j)\}}{\left[\sum_{j\in D} w_j^2 p(t_j)\{1-p(t_j)\}\right]^{1/2}} \\
&= \frac{\sum\limits_{j\in D} w_j\{y_j - \mu_j\}}{\left[\sum\limits_{j\in D} w_j^2 \mu_j(1-\mu_j)\right]^{1/2}} \times \frac{\left[\sum\limits_{j\in D} w_j^2 \mu_j(1-\mu_j)\right]^{1/2}}{\left[\sum\limits_{j\in D} w_j^2 p(t_j)\{1-p(t_j)\}\right]^{1/2}} \\
&\quad + \frac{\sum\limits_{j\in D} w_j\{\mu_j - p(t_j)\}}{\left[\sum\limits_{j\in D} w_j^2 p(t_j)\{1-p(t_j)\}\right]^{1/2}} \\
&= I_1 \times I_2 + I_3.
\end{aligned}$$

By martingale central limiting theorem (Fleming and Harrington, 1991), we can show that the first term I_1 has a limiting standard normal distribution. From equations (E.1) and (E.2), the third term I_3 converges in probability to

$$\frac{n^{-1/2}\sum\limits_{j\in D} w_j\{\mu_j - p(t_j)\}}{\left[n^{-1}\sum\limits_{j\in D} w_j^2 p(t_j)\{1-p(t_j)\}\right]^{1/2}} - \sqrt{n}\frac{\mu_w}{\sigma_w} \xrightarrow{P} 0$$

and

$$\sum_{j\in D} w_j^2 \mu_j(1-\mu_j)]^{1/2} \xrightarrow{P} \int_0^\infty w^2(t)\frac{\pi(t)(1-\pi(t))\lambda_1(t)\lambda_2(t)}{[\pi(t)\lambda_1(t) + \{1-\pi(t)\}\lambda_2(t)]^2}V(t)dt = \tilde{\sigma}_w^2.$$

Thus, the weighted log-rank test L_w is asymptotically normal distributed with mean $\sqrt{n}\mu_w/\sigma_w$ and variance $\sigma_w^2/\tilde{\sigma}_w^2$, where

$$\begin{aligned}
\mu_w &= \int_0^\infty w(t)\frac{\pi(t)(1-\pi(t))\{\lambda_1(t)-\lambda_2(t)\}}{\pi(t)\lambda_1(t) + \{1-\pi(t)\}\lambda_2(t)}V(t)dt, \\
\sigma_w^2 &= \int_0^\infty w^2(t)\pi(t)\{1-\pi(t)\}V(t)dt, \\
\tilde{\sigma}_w^2 &= \int_0^\infty w^2(t)\frac{\pi(t)(1-\pi(t))\lambda_1(t)\lambda_2(t)}{[\pi(t)\lambda_1(t) + \{1-\pi(t)\}\lambda_2(t)]^2}V(t)dt.
\end{aligned}$$

Appendix F: Generating random number under the PWPHCR model in chapter 3

Under the PWPHCR model (3.3), we have

$$S_1(t) = \pi_1 + (1 - \pi_1)S_1^*(t)$$

and

$$S_2(t) = \begin{cases} \pi_1 + (1 - \pi_1)S_1^*(t), & t \leq t_0, \\ \pi_2 + (1 - \pi_2)\tilde{c}\left[S_1^*(t_0)\right]^{1-\delta}\left[S_1^*(t)\right]^{\delta}, & t > t_0, \end{cases}$$

where $\pi_2 = c\tilde{\pi}_2$ and $\tilde{c} = c(1 - \tilde{\pi}_2)/(1 - c\tilde{\pi}_2)$

Assume that T_1 is a random variable with survival distribution $S_1(t)$ and separate $S_1(t)$ as cured patients and uncured patients. Setting $t = \inf$ for cured patients and using the same inverse method discussed in Appendix C to get $T = S_1^{-1}(U)$ for uncured patients. For the Weibull distribution $S_1(t) = e^{-\lambda t^{\kappa}}$, solving t, its inverse function, is given by $t = S_1^{-1}(u) = \{-\log(u)/\lambda\}^{1/\kappa}$.

Similarity, Assume that T_2 is a random variable with survival distribution $S_2(t)$ and separate $S_2(t)$ as cured patients and uncured patients. Setting $t = \inf$ for cured patients and using the same inverse method discussed in Appendix C to get

$$T = S_2^{-1}(U) = \begin{cases} S_1^{-1}(U), & U \geq S_1(t_0), \\ S_1^{-1}\{(U/c)^{1/\delta}\}, & U < S_1(t_0) \end{cases}$$

for uncured patients where $c = \{S_1(t_0)\}^{1-\delta}\tilde{c}$. For the Weibull distribution $S_1(t) = e^{-\lambda t^{\kappa}}$, solving t, its inverse function, is given by $t = S_1^{-1}(u) = \{-\log(u)/\lambda\}^{1/\kappa}$.

Appendix G: R code for the sample size calculations of Chapter 3

Below is the R function ‘Size’ used for the sample size calculation in Chapter 3. ‘Size’ has implemented sample size calculation for several different models. By setting $\pi_1 = \pi_2 = 0$ and $t_0 = 0$, it does the sample size calculation under the standard PH model; by setting $\pi_1 = \pi_2 = 0$ and $t_0 > 0$, it does the sample size calculation under the PWPH model; by setting $\pi_1 < \pi_2 \neq 0$ and $t_0 = 0$, it does the sample size calculation under the PHCR model; by setting $\pi_1 \leq \pi_2 \neq 0$ and $t_0 \neq 0$, it does the sample size calculation under the PWPHCR model. For the PWPH model, the optimal piecewise weighted log-rank test is implemented. However, for the PHCR model or PWPHCR model, we used the standard log-rank test because the optimal weight function for the log-rank test remains unknown. The R function ‘Size’ can be modified to accommodate other parametric survival distribution and non-parametric logspline distribution.

```
####################################################################
### kappa and lambda are the Weibull shape and hazard parameter; ###
### pi1 and pi2 are the cure rates of two groups;                ###
### p is the allocation ratio of control group;                  ###
### ta and tf are accrual duration and follow-up period;         ###
### alpha and beta are type I and II errors;                     ###
### delta is the hazard ratio; t0 is the delay time;             ###
####################################################################
Size=function(kappa,lambda,pi1,pi2,p,ta,tf,delta,alpha,power,t0)
{ z0=qnorm(1-alpha/2); z1=qnorm(power)
  S1=function(t){exp(-lambda*t^kappa)}
  h1=function(t){kappa*lambda*t^(kappa-1)}
  tau=ta+tf; c0=S1(t0)^(1-delta)
  St0=S1(t0)
  pi2.tilde=1/(1+((pi1+(1-pi1)*St0)/pi2-1)/St0)
  c=(pi1+(1-pi1)*St0)/(pi2.tilde+(1-pi2.tilde)*St0)
  c.tilde=c*(1-pi2.tilde)/(1-c*pi2.tilde)
  G=function(t){1-punif(t, tf, tau)}
  S2=function(t){c0*S1(t)^delta}
  h2=function(t){delta*h1(t)}
  S11=function(t){pi1+(1-pi1)*S1(t)}
  S21=function(t){pi2+(1-pi2)*c.tilde*S2(t)}
  h11=function(t){(1-pi1)*S1(t)*h1(t)/S11(t)}
  h21=function(t){(1-pi2)*c.tilde*S2(t)*h2(t)/S21(t)}
  pi=function(t){p*S11(t)/(p*S11(t)+(1-p)*S21(t))}
  V=function(t){(p*h11(t)*S11(t)+(1-p)*h21(t)*S21(t))*G(t)}
  f1=function(t){pi(t)*(1-pi(t))*(h11(t)-h21(t))*V(t)/
                 (pi(t)*h11(t)+(1-pi(t))*h21(t))}
  f2=function(t){pi(t)*(1-pi(t))*V(t)}
  f3=function(t){pi(t)*(1-pi(t))*h11(t)*h21(t)*V(t)/
                (pi(t)*h11(t)+(1-pi(t))*h21(t))^2}
  mu=integrate(f1, t0, tau)$value
  sigma1=integrate(f2, t0, tau)$value
  sigma2=integrate(f3, t0, tau)$value
  P=integrate(V, t0, tau)$value
  n=(sqrt(sigma1)*z0+sqrt(sigma2)*z1)^2/mu^2
  dt0=ceiling(n*P)
  d1=(1-S11(t0))*n
  d=ceiling(d1+dt0)
  ans=c(dt0=dt0,d=d,n=ceiling(n)); return(ans)
}
Size(kappa=1.2,lambda=0.059,pi1=0.12,pi2=0.18,p=0.5,ta=17,tf=37,
       delta=0.72,alpha=0.05,power=0.9,t0=3.5)
dt0   d   n  # dt0 and d are number of events after delay and
352 466 553  # total number of events; n is sample size
```

Appendix H: R code used for sample size calculation of chapter 4

Below is the R function ‘Size’ used for the sample size calculation in chapter 4. ‘Size’ has implemented the sample size calculation using formulae (4.3) and (4.7) with the Weibull random delayed cure rate model.

```
####################################################################
### kappa and lambda are the Weibull shape and hazard parameter; ###
### pi1 and pi2 are the cure rates of two groups;                ###
### omega is the allocation ratio of control group;              ###
### ta and tf are accrual duration and follow-up period;         ###
### alpha and beta are type I and II errors;                     ###
### delta is the hazard ratio;                                   ###
### T1 and T2 are the lag domain for the random delay time;      ###
####################################################################
library (statmod)
GQ<-gauss.quad(n=50,kind="legendre")
GQ.int<-function (g, limits=c(0,t)){
  upp=limits[2];low=limits[1];
  sum(sapply(GQ$nodes, function(x){g((upp-low)*x/2+
                (upp+low)/2 )*(upp-low)/2})*GQ$weights)}
Size<-function(alpha,beta,kappa,lambda,ta,tf,pi1,pi2,delta,T1,T2,omega)
{   total<-ta+tf
    lambda1star<-function(x){kappa*lambda*x^(kappa-1)}
    G<-function(x){1-punif(x,min=tf,max=total)}
    s1star<-function(x){exp(-lambda*x^kappa)}
    s1<-function(x){pi1+(1-pi1)*s1star(x)}
    f1<-function(x){(1-pi1)*s1star(x)*lambda1star(x)}
    lambda1<-function(x){f1(x)/s1(x)}
    g<-function(pi2tu){
    f_tau<-function(mu){return(dunif(mu,T1,T2))}
    s1star<-function(mu){exp(-lambda*mu^kappa)}
    A_tau<-function(mu){(pi1+(1-pi1)*s1star(mu))/
    (pi2tu+(1-pi2tu)*s1star(mu))}
    intepart<-function(mu){A_tau(mu)*f_tau(mu)}
    return(pi2tu*integrate(intepart,lower=T1,upper=T2)
   $value-pi2)}
    pi2tu<-uniroot(g,c(0,0.99))$root
    A_tau<-function(tau){(pi1+(1-pi1)*s1star(tau))
    /(pi2tu+(1-pi2tu)*s1star(tau))}
    s_tau<-function(x){1*I(x<T1)+((T2-x)/(T2-T1))*I((T1<= x)&(x<= T2))
    +0*I(x>T2)}
    weight<-function(x){1-s_tau(x)}
    f_tau<-function(x){return(dunif(x,T1,T2))}
    intepart1<-function(mu){A_tau(mu)*f_tau(mu)}
    intepart2<-function(mu){A_tau(mu)*(s1star(mu))^(1-delta)*f_tau(mu)}
    intepart3<-function(mu){A_tau(mu)*((s1star(mu))^(1-delta)*f_tau(mu))}
    s2<-function(x){(pi1+(1-pi1)*s1star(x))*s_tau(x)+
      pi2tu*GQ.int(intepart1,limits=c(0,min(T2,x)))+
      (s1star(x))^delta*(1-pi2tu)*
      GQ.int(intepart2,limits=c(0,min(T2,x)))}
    f2<-function(x){s1star(x)*lambda1star(x)*((1-pi1)*s_tau(x)
    +(1-pi2tu)*delta*s1star(x)^(delta-1)
    *GQ.int(intepart3,limits=c(0,min(T2,x))))}
    lambda2<-function(x){f2(x)/s2(x)}
    pifunction<-function(x){omega*s1(x)/(omega*s1(x)+(1-omega)*s2(x))}
    v<-function(x){omega*f1(x)*G(x)+(1-omega)*f2(x)*G(x)}
    bndry.mat=matrix(c(0,total),nrow = 1,ncol = 2)
    integrand1 <- function(x) {pifunction(x)*(1-pifunction(x))*
    (lambda1(x)-lambda2(x))*v(x)*weight(x)/(pifunction(x)*lambda1(x)+
       (1-pifunction(x))*lambda2(x))}
    mu0=sum(apply(bndry.mat,1,function(x) GQ.int(integrand1,limits=x)))
    integrand2 <- function(x) {pifunction(x)*(1-pifunction(x))
    *lambda1(x)*lambda2(x)*v(x)*weight(x)^2/
    ((pifunction(x)*lambda1(x)+
    (1-pifunction(x))*lambda2(x))^2)}
    sigma1=sum(apply(bndry.mat,1,function(x) GQ.int(integrand2,limits=x)))
    integrand3<-function(x) {pifunction(x)*
    (1-pifunction(x))*v(x)*weight(x)^2}
    sigma0=sum(apply(bndry.mat,1,function(x)
    GQ.int(integrand3,limits=x)))
    n=(qnorm(1-alpha/2)*sqrt(sigma0)+
    qnorm(1-beta)*sqrt(sigma1))^2/((mu0)^2)
    return(ceiling(n))}
Size(alpha=0.05,beta=0.2,kappa=1,lambda=log(2)/(10/12),
ta=2,tf=1,pi1=0.35,pi2=0.35,
delta=0.7,T1=0/12,T2=6/12,omega = 0.5)
1641
```

Appendix I: Generating random number under the PWPHRR model in chapter 5

Under the PWPHCR model (5.3), we have

$$S_T(t) = \begin{cases} S_C(t), & t \leq t_0, \\ p\left[S_C(t_0)\right]^{1-\delta}\left[S_C(t)\right]^{\delta} + (1-p)S_C(t), & t > t_0. \end{cases}$$

Assume that T is a random variable with survival distribution $S_T(t)$ and can generate $T = S_C^{-1}(U)$ for non-responder patients. For the Weibull distribution $S_C(t) = e^{-\lambda t^{\kappa}}$, solving t, its inverse function, is given by $t = S_C^{-1}(u) = \{-\log(u)/\lambda\}^{1/\kappa}$. For responder patients, can get

$$T = S_2^{-1}(U) = \begin{cases} S_1^{-1}(U), & U \geq S_1(t_0), \\ S_1^{-1}\{(U/c)^{1/\delta}\}, & U < S_1(t_0), \end{cases}$$

where $c = \{S_C(t_0)\}^{1-\delta}$. For the Weibull distribution $S_1(t) = e^{-\lambda t^{\kappa}}$, solving t, its inverse function, is given by $t = S_C^{-1}(u) = \{-\log(u)/\lambda\}^{1/\kappa}$.

Appendix J: R code used for sample size calculation of chapter 5

Below is the R function ‘Size’ used for the sample size calculation in chapter 5. ‘Size’ has implemented the sample size calculation using formula (5.4) with the Weibull delayed responder rate model.

```
####################################################################
### kappa and lambda are the Weibull shape and hazard parameter; ###
### p is the responder rates of two groups;                      ###
### omega1 and omega2 are the allocation ratio of two groups;    ###
### total is the study period;                                   ###
### alpha and beta are type I and II errors;                     ###
### delta is the hazard ratio;                                   ###
### t0 is the fixed delay time;                                  ###
### r is the enrollment rate for patients                        ###
####################################################################

Size<-function(alpha,beta,r,total,k,lambda,delta,w1,w2,p,t0){
  root<-function(ta){
    tf<-total-ta
    G<-function(x){
      1-punif(x,min=tf,max=total)
    }

    ##control group survival
    s1<-function(x){
      exp(-lambda*x^k)
    }
    ##control hazard
    lambda1<-function(x){
      k*lambda*x^(k-1)
    }

    ###treatment survival after t0
    s2<-function(x){
      p*s1(t0)^(1-delta)*s1(x)^delta+(1-p)*s1(x)
    }

    ###treatment hazard after t0
    lambda2<-function(x){
      (p*delta*s1(t0)^(1-delta)*s1(x)^(delta-1)+1-p)/(p*s1(t0)^(1-
      delta)*s1(x)^(delta-1)+1-p)*lambda1(x)
    }

    pifunction<-function(x){
      w1*s1(x)/( w1*s1(x)+ w2*s2(x))
    }

    v<-function(x){
      w1*s1(x)*lambda1(x)*G(x)+w2*s2(x)*lambda2(x)*G(x)
    }

    weight<-function(x){
      (p*s1(t0)^(1-delta))/(p*s1(t0)^(1-delta)+(1-p)*s1(x)^(1-delta))
    }

    integrand1 <- function(x) {pifunction(x)*(1-pifunction(x))
    *(lambda1(x)-lambda2(x))*v(x)*weight(x)/
    (pifunction(x)*lambda1(x)+(1-pifunction(x))*lambda2(x))}
    mu0=(integrate(integrand1, lower = t0, upper = total)$value)

    integrand2 <- function(x) {pifunction(x)*(1-pifunction(x))*lambda1(x)
    *lambda2(x)*v(x)*weight(x)^2
    /((pifunction(x)*lambda1(x)+(1-pifunction(x))*lambda2(x))^2)}
    sigma1=(integrate(integrand2, lower = t0, upper = total)$value)

    integrand3<- function(x) {pifunction(x)*(1-pifunction(x))*v(x)*
    weight(x)^2}
    sigma0=(integrate(integrand3, lower = t0, upper =total)$value)

    n=(qnorm(1-alpha/2)*sqrt(sigma0)+qnorm(1-beta)*sqrt(sigma1))^2
    /((mu0)^2)
```

```
    ans<-r*ta-n
  }
  ta<-uniroot(root,lower=0.1,upper=200)$root
  n<-ceiling(r*ta)
  ta<-round(ta,2)
  ans<-list(c(ta=ta,n=n))
  return(ans)
}

Size(alpha=0.05,beta=0.1,r=36.8,total=29,k=1,
lambda=0.074,delta=0.01,w1=0.5,w2=0.5,p=0.2,t0=2)
    ta      n
 10.65 392.00
```

Bibliography

Borghaei, H., L. Paz-Ares, L. Horn, D. R. Spigel, M. Steins, and et al. (2015). Nivolumab versus docetaxel in advanced non-squamous non-small cell lung cancer. *The New England Journal of Medicine 273*(17), 1627–1639.

Coiffier, B., E. Lepage, J. Brière, R. Herbrecht, E. Tilly, and et al. (2002). Chop chemotherapy plus rituximab compared with chop alone in elderly patients with diffuse large-b-cell lymphoma. *The New England Journal of Medicine 346*(4), 235–242.

Corbière, F. and P. Joly (2007). A sas macro for parametric and semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine 85*(2), 173–180.

Cox, D. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B 34*(2), 187–202.

Ding, X. and J. Wu (2020). Designing cancer immunotherapy trials with delayed treatment effect using maximin efficiency robust statistics. *Pharmaceutical Statistics 19*(4), 424–435.

Eggermont, A., V. Chiarion-Sileni, J. Grob, R. Dummer, and et al. (2016). Prolonged survival in stage iii melanoma with ipilimumab adjuvant therapy. *The New England Journal of Medicine 375*(19), 1845–1855.

Farewell, V. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics 38*(4), 1041–1046.

Fine, G. (2007). Consequences of delayed treatment effects on analysis of time-to-event endpoints. *Drug Information Journal 41*(4), 535–539.

Fleming, T. and D. Harrington (1991). *Counting processes and survival analysis*. John Wiley and Sons: New York.

Freedman, L. (1982). Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine 1*(2), 121–129.

Hasegawa, T. (2014). Sample size determination for the weighted log-rank test with the fleming-harrington class of weights in cancer vaccine studies. *Pharmaceutical Statis-*

tics 13(2), 128–135.

Kantoff, P., C. Higano, N. Shore, E. Berger, E. Small, and et al. (2010). Sipuleucel-t immunotherapy for castration-resistant prostate cancer. *The New England Journal of Medicine 363*(5), 411–422.

Lakatos, E. (1988). Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics 44*(1), 229–241.

Liu, S., C. Chu, and A. Rong (2018). Weighted log-rank test for time-to-event data in immunothearpy trials with random delayed treatment effect and cure rate. *Pharmaceutical Statistics 17*(5), 541–554.

Magirr, D. and C. Burman (2019). Modestly weighted logrank tests. *Statistics in Medicine 38*(20), 3782–3790.

Robert, C., L. Thomas, I. Bondarenko, and et al. (2011). Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. *The New England Journal of Medicine 364*(26), 2517–2526.

Schlom, J. and J. L. . Gulley (2018). Vaccines as an integral component of cancer immunotherapy. *JAMA 320*(21), 2195–2196.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika 68*(1), 316–319.

Smith, M. W. Types of immunotherapy. `https://www.webmd.com/cancer/immunotherapy-treatment-types.html`.

Wang, S., J. Zhang, and W. Lu (2012). Sample size calculation for the proportional hazards cure mode. *Statistics in Medicine 31*(29), 3959–3971.

Wei, J. and J. Wu (2020). Cancer immunotherapy trial design with cure rate and delayed treatment effect. *Statistics in Medicine 39*(6), 698–708.

Xiong, X. and J. Wu (2017). A novel sample size formula for the weighted log-rank test under the proportional hazards cure model. *Pharmaceutical Statistics 16*(29), 87–94.

Xu, Z., B. Zhen, Y. Park, and B. Zhu (2016). Designing therapeutic cancer vaccine trials with delayed treatment effect. *Statistics in Medicine 36*(4), 592–605.

Xu, Z., B. Zhen, Y. Park, and B. Zhu (2018). Designing cancer immunotherapy trials with random treatment time-lag effect. *Statistics in Medicine 37*(30), 4589–4609.

Xu, Z., B. Zhu, and Y. Park (2020). Design for immuno-oncology clinical trials enrolling both responders and nonresponders. *Statistics in Medicine 39*(27), 3914–3936.

Ye, T. and M. Yu (2018). A robust approach to sample size calculation in cancer immunotherapy trials with delayed treatment effect. *Biometrics 74*(4), 1292–1300.

Zucker, D. and E. Lakatos (1990). Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika 77*(4), 853–864.

Vita

Jing Wei

EDUCATION

- Ph. D. in Statistics, University of Kentucky, 2021 (Expected)
- M. S. in Mathematics, University of Kentucky, 2017
- M. S. in Public Health, University of Kentucky, 2015
- B. S. in Electronic Engineering, Nanjing University of Posts & Telecommunications, 2010

WORKING EXPERIENCE

- Research Assistant, University of Kentucky, 2018-2021
- Teaching Assistant, University of Kentucky, 2015-2017

PUBLICATIONS

- Wei, J., Wu, J., Cancer Immunotherapy Trial Design with Cure Rate and Delayed Treatment Effect. STATISTICS IN MEDICINE, 2020; 39:698-708.
- Wu, J., Wei, J., Cancer Immunotherapy Trial Design with Delayed Treatment Effect. PHARMACEUTICAL STATISTICS, 2020; 19(3):202-213.
- Wu, J., Chen, L., Wei, J, et al., Phase II Trial Design with Growth Modulation Index as the Primary Endpoint. PHARMACEUTICAL STATISTICS, 2019; 18(2):212-222.
- Wu, J., Chen, L., Wei, J, et al., Optimal Two-Stage Phase II Survival Trial Design. PHARMACEUTICAL STATISTICS, 2020; 19(3):214-229.
- Chauhan, A., Kabir, T., Wu, J., Wei, J., et al., Prognostic and predictive factors associated with ipilimumab related adverse events: A retrospective analysis of 11 NCI sponsored ipilimumab phase I clinical trials. ONCOTARGET, 2020; 11: 1427-1434.
- Jacob, A., Raj, R., Alagusundaramoorthy, S., Wei, J., Wu, J., Impact of Patient Load on the Quality of Electronic Medical Record Documentation. JOURNAL OF MEDICAL EDUCATION AND CURRICULAR DEVELOPMENT, 2021; 8: 2382120520988597.
- Wei, J., Wu, J., Random treatment time-lag effect with cure rate in Cancer Immunotherapy Trial Design, submited.