POWER AND SAMPLE SIZE DETERMINATIONS IN DYNAMIC RISK PREDICTION

by

Zhaowen Sun

M.S., University of Pittsburgh, 2012

B.S.N., Wuhan University, China, 2010

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Zhaowen Sun

It was defended on

August 21, 2017

and approved by

Chung-Chou H. Chang, Ph.D., Professor, Departments of Medicine and Biostatistics, School of Medicine and Graduate School of Public Health, University of Pittsburgh

Stewart J. Anderson, Ph.D., M.A., Professor, Departments of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Chaeryon Kang, Ph.D., M.A., Assistant Professor, Departments of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Beth E. Snitz, Ph.D., Associate Professor, Department of Neurology, University of Pittsburgh

Dissertation Director: Chung-Chou H. Chang, Ph.D., Professor, Departments of Medicine and Biostatistics, School of Medicine and Graduate School of Public Health, University of Pittsburgh

POWER AND SAMPLE SIZE DETERMINATIONS IN DYNAMIC RISK PREDICTION

Zhaowen Sun, PhD

University of Pittsburgh, 2017

ABSTRACT

Dynamic risk prediction is a powerful tool to estimate the future risk of study subjects with data involves time-dependent information, including repeatedly measured covariates, intermediate events, and time-varying covariate effects. The quantity of interest for dynamic risk prediction is the probability of failure at the prediction horizon time conditional on the status at the prediction baseline (*aka landmark time*). For a clinical study, a series of horizon and landmark time points are usually planned in the design stage. This conditional probability can be estimated from a standard Cox proportional hazards model (for data without competing risks) or a Fine and Gray subdistributional hazards model (for data with competing risks) by appropriately setting up a landmark dataset. In this dissertation, I propose test statistics for testing the equal conditional probability between two patient groups according to their response to treatment at the prediction baseline under the scenarios of data with and without competing risks, respectively. The dissertation provides three different methods for estimating the variance of risk difference. In designing a randomized clinical trial for comparing risk difference between the two study arms, I derived formulas for power, the number of events, and the total sample size required with respect to the aforementioned hypothesis tests. Simulations were conducted to evaluate the impact of each design parameter on the power and sample size calculations.

Public health significance: This study aims to introduce new risk prediction methods that can incorporate time-dependent information and update risk estimation during the course of study follow-up, also provide researchers with references on the power and sample size requirements at the planning phase of studies involving dynamic risk prediction.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Models for risk prediction are useful tools for disease prognosis, treatment assignment and treatment. The probability of failure, defined as an individual's absolute risk of experiencing an event of interest at a given time point, is often used by clinical researchers because it is quantifiable, accessible, and easy to interpret.

Risk prediction can be classified as either static or dynamic. For static prediction, the prediction baseline (*aka* analysis time 0) is usually the study baseline, and the predictors contain only the current or historical values with respect to the prediction baseline. The quantity of interest for static prediction is simply the absolute risk at the prediction horizon time, that is, the marginal probability of failure at the prediction horizon time. Methods to estimate this probability are well established; among which Kaplan-Meier nonparametric estimation method and Cox proportional hazards regression model are most commonly used.

For dynamic risk prediction, each prediction baseline is chosen at a future time point relative to the study baseline before inspecting the data and often with clinical interests (e.g., initiation of chemotherapy, one year after surgery). The quantity of interest is the risk at the prediction horizon time *conditional on* the status at the prediction baseline. Note that values of predictors may change over time. Time-dependent information includes time-dependent covariates (e.g., repeatedly measured covariates and intermediate events) and time-varying covariate effect. For example, a risk prediction model for cancer metastasis among breast cancer patients with ER positive and node negative who underwent surgery is expected to incorporate time-fixed covariates (e.g., age, gender, primary tumor size, and surgery type), time-dependent covariates (e.g., white blood cell counts), and intermediate event status if applicable (e.g., presence of local-regional recurrence and time to local-regional recurrence).

Existing methods for handling time-dependent information include Cox model with time-dependent covariates [1, 2], multi-state modeling [3], joint modeling [4] among others. Estimation of conditional probability of failure involving time-dependent information often requires several steps and/or additional assumptions on the survival and covariate processes, which could render the above-mentioned methods computationally intensive. Furthermore, when intermediate event is involved, immortal-time bias could be introduced. One classic case is to use the response to chemotherapy as an intermediate event in cancer survival, where those who responded to the treatment prior to the survey had a "survival advantage" over those who did not respond to the chemotherapy.[5]

Landmarking technique, proposed by van Houwelingen in 2007, bypasses the aforementioned issues and achieves dynamic risk prediction in one step. The method provides a valid approximation of the conditional probability of failure at the prediction horizon time and is robust to possible violation of the proportional hazards (PH) assumption. [6, 7, 8] Following the works of van Houwelingen, several researchers have extended landmark methods for dynamic prediction in various settings. [9, 10, 11, 12]

In order to estimate and test the risk differences between treatment groups, statisticians need to provide methods to estimate the sample size and the power at the study design stage. Power and sample size calculation procedure is constructed based on the hypothesis to be tested and the corresponding test statistic. After the null and alternative hypotheses are determined and an appropriate test statistic is developed with its exact or asymptotic distributions under both the null and alternative hypotheses identified, power and sample size can be calculated using the pre-specified type I error rate, desired power level, and other design parameters. The general procedures for power and sample size calculations in clinical studies have been illustrated by many researchers yet currently there is little systematic work done on this subject in the context of dynamic risk prediction. In this study, we developed formulas to estimate the sample size and the power for comparing conditional probabilities of failure between two treatment groups in dynamic risk prediction.

In Chapter 2, we introduce methods for dynamic risk prediction under both single event and competing risks settings. In Chapter 3 we present our proposed test statistic for determining risk difference and the power function for the proposed test; and illustrate the

procedure of sample size calculation. In Chapter 4 we present simulation results assessing the factors that influence power and sample size using the proposed test. We conclude the work with discussions and the future work plan in Chapter 5.

2.0 DYNAMIC RISK PREDICTION USING LANDMARKING

2.1 LANDMARK COX MODEL

Let T_L and C be the main (long term) event and censoring time, respectively; with T_S as the intermediate (short term) event time. We define the intermediate event status $\delta_{Si} = I(T_s \leq t_s)$ and main event status $\delta_{Li} = I(T_L \leq t_l)$ where $I(A)$ is the indicator function that takes value of 1 when the condition A is true and 0 otherwise. Let $\boldsymbol{Z}_i$ represent the vector of covariates, for subject i we observe the independently and identically distributed data

$$\{X_{Li} = T_{Li} \wedge C_i, \delta_{Li} = I(T_{Li} < C_i), X_{Si} = T_{Si} \wedge (T_{Li} \wedge C_i), \boldsymbol{Z}_i\}, i = 1, ..., n.$$

We assume that C_i is independent of T_{Si}, T_{Li} and $\boldsymbol{Z}_i$. Probability of failure for the main event at time t is defined as:

$$F(t; \boldsymbol{Z}) = Pr(T_L \leq t | \boldsymbol{Z}).$$

For data containing no competing risks, the Cox proportional hazard model takes the form

$$\lambda(t_l | \boldsymbol{Z}) = \lambda_0(t_l) \exp(\boldsymbol{\beta}^T \boldsymbol{Z})$$

with

$$F(t; \boldsymbol{Z}) = 1 - \exp\{\int_0^t \lambda_0(u) \exp(\boldsymbol{\beta}^T \boldsymbol{Z})\ du\}.$$

The covariate effects, $\boldsymbol{\beta}$, are estimated via maximizing a partial log-likelihood function, then we have:

$$\hat{F}(t; \boldsymbol{Z}) = 1 - \exp\{-\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}) \hat{\Lambda}_0(t)\},$$

where $\hat{\Lambda}_0(t)$ is the Breslow-type estimator.

As stated in the introduction, the quantity of interest in dynamic risk prediction is the probability of failure from the main event conditional on that the individual has not failed from the main event by landmark time, t_{LM}, with information accumulated up until t_{LM}, $\boldsymbol{Z}_{LM}$, resulting in the conditional probability of failure at t_{hor}:

$$\begin{aligned} F\{t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}\} &= P\{T_L \leq t_{hor}|T_L > t_{LM}, \boldsymbol{Z}_{LM}\} \\ &= \frac{P(T_L \leq t_{hor}) - P(T_L \leq t_{LM})}{P(T_L > t_{LM})} \\ &= \frac{F\{t_{hor}; \boldsymbol{Z}_{LM}\} - F\{t_{LM}; \boldsymbol{Z}_{LM}\}}{1 - F\{t_{LM}; \boldsymbol{Z}_{LM}\}}. \end{aligned} \tag{2.1}$$

Instead of estimating each component in equation (2.1) separately, van Houwelingen introduced the *Landmark Cox model* that can be used to estimate the conditional probability in one step. The hazard function has the form [6]:

$$\lambda\{t|\boldsymbol{Z}_{LM}, t_{LM}\} = \lambda_0(t|t_{LM}) \exp\{\boldsymbol{\beta}_{LM}^T \boldsymbol{Z}_{LM}\}, t_{LM} \leq t \leq t_{hor}, \tag{2.2}$$

where $\lambda_0(t|t_{LM})$ is the unspecified, non-negative conditional baseline hazard function.

By landmarking we are selecting subjects that are still at risk for the main event at landmark time t_{LM} and enforcing administrative censoring at prediction horizon t_{hor}. As a result, the modified risk set (landmark data set) will differ with choices of t_{LM} and t_{hor}, as well as the information that could be used in parameter estimation. The notations $\boldsymbol{\beta}_{LM}$ and $\boldsymbol{Z}_{LM}$ indicate that the coefficient estimates and covariate vector are specific to the choice of prediction landmark time. $\hat{\boldsymbol{\beta}}_{LM}$ can be obtained by maximizing the *partial log-likelihood*

$$pl_s(\boldsymbol{\beta}_{LM}) = \sum_{t_i \geq t_{LM}} \delta_{Li}[\boldsymbol{\beta}_{LM}^T \boldsymbol{Z}_{i;LM} - \log\{\sum_{t_j \geq t_i} \exp(\boldsymbol{\beta}_{LM}^T \boldsymbol{Z}_{j;LM})\}].$$

The baseline hazard at time t_i can be estimated by a Breslow-type estimator:

$$\hat{\lambda}_0(t_i|t_{LM}) = \frac{1}{\sum_{t_i \leq t_j} \exp(\hat{\boldsymbol{\beta}}_{LM}^T \boldsymbol{Z}_{j;LM})}.$$

The conditional probability of failure can be obtained as:

$$\hat{F}\{t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}\} = 1 - \exp[\exp(-\hat{\boldsymbol{\beta}}_{LM}^T \boldsymbol{Z}_{LM}\{\hat{\Lambda}_0(t_{hor}) - \hat{\Lambda}_0(t_{LM-})\}], \tag{2.3}$$

where $\hat{\Lambda}_0(t_{LM-})$ and $\hat{\Lambda}_0(t_{hor})$ are cumulative baseline hazards with

$$\hat{\Lambda}_0(t_{LM-}) = \sum_{t_i \leq t, \delta_{li}=1} \hat{\lambda}_0(t_i|t_{LM}),$$

$$\hat{\Lambda}_0(t_{hor}) = \sum_{t_i \leq t, \delta_{li}=1} \hat{\lambda}_0(t_i|t_{hor}).$$

Previous work by Struther and Kalbfleisch [13] and Xu and O'Quigley [14] showed that when the true covariate effect is time-varying such that $\lambda(t|\boldsymbol{Z}) = \lambda_0(t)\exp(\boldsymbol{\beta}(t)^T\boldsymbol{Z})$, the limiting value of the maximum partial likelihood estimator from a Cox proportional hazards models converges to a weighted average of the underlying time-varying covariate effect. And it has been shown that the landmark Cox model preserves such a property [6], $\hat{\boldsymbol{\beta}}_{LM}$ is a consistent estimator of $\boldsymbol{\beta}^*$, which is the solution to:

$$\int_0^\infty \{\frac{s^{(1)}(\boldsymbol{\beta}(t),t)}{s^{(0)}(\boldsymbol{\beta}(t),t)} - \frac{s^{(1)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)}\} s^{(0)}(\boldsymbol{\beta}(t),t)\lambda_0(t)\ dt = 0. \tag{2.4}$$

Define

$$S^{(r)}(\boldsymbol{\beta},t) = n^{-1}\sum_{i=1}^{n} Y_i(t)\boldsymbol{Z}_i^r \exp(\boldsymbol{\beta}^T\boldsymbol{Z}_i)$$

with $Y_i(t)$ as the at-risk indicator from counting process, and

$$s^{(r)}(\boldsymbol{\beta},t) = ES^{(r)}(\boldsymbol{\beta},t),$$

thus

$$\frac{S^{(1)}(\boldsymbol{\beta}(t),t)}{S^{(0)}(\boldsymbol{\beta}(t),t)} = E[\boldsymbol{Z}|T=t]. \tag{2.5}$$

$$\frac{\partial}{\partial\boldsymbol{\beta}}(\frac{s^{(1)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)})|_{\boldsymbol{\beta}=\boldsymbol{\beta}(t)} = var(\boldsymbol{Z}|T=t) \tag{2.6}$$

Combine (2.5) and (2.6) we have:

$$\frac{S^{(1)}(\boldsymbol{\beta}(t),t)}{S^{(0)}(\boldsymbol{\beta}(t),t)} - \frac{s^{(1)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)} \approx \{\boldsymbol{\beta} - \boldsymbol{\beta}(t)\}var(\boldsymbol{Z}|T=t) \tag{2.7}$$

In the presence of random censoring:

$$\begin{aligned} s^{(0)}(\boldsymbol{\beta}(t),t)\lambda_0(t) &= E[Y(t)\exp(\boldsymbol{\beta}(t)^T)\boldsymbol{Z}\lambda_0(t)] \\ &= E[Y(t)\lambda(t|\boldsymbol{Z})] \\ &= E[Y(t)]E[\lambda(t|\boldsymbol{Z})|T \geq t] \\ &= S(t)C(t)\lambda(t), \end{aligned} \tag{2.8}$$

where $\lambda(t|\boldsymbol{Z})$ is the marginal hazard, $S(t)$ is the marginal survival function and $C(t)$ is the survival function of non-informative censoring time.

Using (2.7) and (2.8), equation (2.4) can be approximated by:

$$\int_0^\infty var(\boldsymbol{Z}|T=t)(\boldsymbol{\beta}-\boldsymbol{\beta}(t))S(t)C(t)\lambda(t)\ dt = 0. \tag{2.9}$$

With additional administrative censoring at t_{hor}, the original $\boldsymbol{\beta}^*$ becomes $\boldsymbol{\beta}^*_{hor}$ and

$$\hat{\boldsymbol{\beta}}_{pl} \xrightarrow{p} \boldsymbol{\beta}^*_{hor} \approx \frac{\int_0^{t_{hor}} S(t)C(t)h(t)var(\boldsymbol{Z}|T=t)\beta(t)\ dt}{\int_0^{t_{hor}} S(t)C(t)h(t)var(\boldsymbol{Z}|T=t)\ dt}. \tag{2.10}$$

If the marginal survival function $S(t)$ does not get too small, the covariate effect is not too large and does not vary too much overtime,in the absence of heavy censoring, we can further write:

$$\boldsymbol{\beta}^*_{hor} \approx \frac{\int_0^{t_{hor}} \lambda_0(t)\boldsymbol{\beta}(t)\ dt}{\int_0^{t_{hor}} \lambda_0(t)\ dt}. \tag{2.11}$$

Similarly, the limiting value of the baseline hazard can be approximated by:

$$\lambda^*_0(t) \approx \lambda_0(t)\exp E[\boldsymbol{Z}|T=t](\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*_{hor}), \tag{2.12}$$

and it follows that

$$\Lambda_0(t) \xrightarrow{p} \Lambda^*_0(t) \approx \exp(\boldsymbol{\beta}(t)-\boldsymbol{\beta}^*_{hor})\int_0^{t_{hor}} \lambda^*_0(t). \tag{2.13}$$

Dynamic risk prediction with the landmark Cox model takes fewer steps and can fit a more sparse model as compared to multi-state models or joint modeling. Landmark Cox model can be fitted using standard statistical software such as `coxph()` in R once the landmark data set is correctly created. The conditional probability of failure being estimated

is often clinically meaningful. It has also been shown that landmark Cox model provides a valid approximation of the conditional probability of failure $F\{t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}\}$, even in the presence of time-dependent covariates or time-varying covariate effects; given that $\boldsymbol{\beta}(t)$ does not vary dramatically overtime, is not too big and the follow-up time is not too long.

2.2 LANDMARK PROPORTIONAL SUB-DISTRIBUTION HAZARDS MODEL

In the presence of competing risk events, the cause-specific cumulative incidence function (CIF) is an appropriate measurement of one's absolute probability of failure from a certain type of event without any dependency assumptions among competing events. The proportional sub-distribution hazards (PSH) models proposed by Fine and Gray (1999) [15] is a popular tool for estimating cause-specific CIFs; which is easy to implement with the ability to incorporate multi-dimensional covariates and provide readily interpretable results.

Let T and C be the event and censoring times, respectively; with $\epsilon \in \{1, ..., k\}$ as the event types and $\boldsymbol{Z}$ representing the vector of covariates. We assume that C_i is independent of T_i and $\boldsymbol{Z}$. For each subject we observe the independently and identically distributed data $\{X_i = T_i \wedge C_i, \Delta_i = I(T_i < C_i), \Delta_i\epsilon_i, \boldsymbol{Z}_i\}$, $i = 1, ..., n$. Without loss of generality we will refer to the event of interest as type 1 event with corresponding CIF defined as:

$$F_1(t; \boldsymbol{Z}) = Pr(T \leq t, \epsilon = 1|\boldsymbol{Z}).$$

The *Fine-Gray PSH model* takes the form

$$\lambda_1(t|\boldsymbol{Z}) = \lambda_{10}(t) \exp(\boldsymbol{\beta}^T \boldsymbol{Z})$$

The sub-distribution hazard $\lambda_1(t)$ is the hazard for an improper failure time T^*, defined as $T \times I(\epsilon = 1) + \{1 - I(\epsilon = 1)\}$, possessing a cumulative distribution function $F_1(t)$ for $t \leq \infty$ and a point mass at $t = \infty$ with an unspecified, non-negative baseline hazard $\lambda_{10}(t) = -dlog\{1 - F_1(t; \boldsymbol{Z} = \boldsymbol{0})\}/dt$ and modified risk set $R(T_i) = \{j : (T_j \geq T_i) \cup (T_j < T_i \cap \epsilon_i \neq 1\}$.

The PSH model for failure from type 1 event can be viewed as a Cox PH model over the support of T^*.

The cause-specific CIF can be calculated as:

$$F_1(t; \boldsymbol{Z}) = 1 - \exp\{\int_0^t \lambda_{10}(u) \exp(\boldsymbol{\beta}^T \boldsymbol{Z})\ du\}.$$

When random right censoring is present, the covariate effects $\boldsymbol{\beta}$ are estimated via maximizing an inverse probability of censoring weighted (IPCW) partial log-likelihood function:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\infty} \{\boldsymbol{Z}_i - \frac{\sum_j \omega_j(t) Y_j(t) \boldsymbol{Z}_j \exp(\boldsymbol{\beta}^T \boldsymbol{Z}_j)}{\sum_j \omega_j(t) Y_j(t) \exp(\boldsymbol{\beta}^T \boldsymbol{Z}_j)}\} \omega_j(t)\ dN_i(t).$$

In terms of the counting process $N_i(t) = I(T_i \leq t, \epsilon_1 = 1)$, $Y_i(t) = I(T_i \geq t) + I(T_i < t, \epsilon_1 \neq 1)$ and $\omega_i(t) = I(C_i \geq T_i) \cap \hat{G}(t)/\hat{G}(X_i \wedge t)$ and $\hat{G}(t)$ is the Kaplan-Meier estimator of the censoring survival distribution $G(t) = Pr(C \geq t)$.

It follows that the baseline cumulative sub-distribution hazards $\Lambda_{10}(t) = \int_0^t \lambda_{10}(u)\ du$ can be obtained using a weighed version of the Breslow estimator:

$$\hat{\Lambda}_{10}(t) = \frac{1}{n} \sum_{i=1}^{n} \int_0^t \frac{1}{\frac{1}{n} \sum_j \omega_j(u) Y_j(u) \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_j)} \omega_i(u)\ dN_i(u).$$

Combing $\hat{\boldsymbol{\beta}}$ and $\hat{\Lambda}_{10}(t)$, we have the predicted cause-specific CIF as

$$\hat{F}_1(t; \boldsymbol{z}) = 1 - \exp\{-\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{z}) \hat{\Lambda}_{10}(t)\}.$$

In clinical practice it is often desirable to make use of time-dependent information, including repeatedly measured covariates, potentially time-varying covariate effects and intermediate event status; yet such time-dependent information may result in the violation of the PSH assumption and bias the CIF estimates. For instance, the effect of a treatment regimen could be diminishing over time or altered by some intermediate clinical event, such as local-regional recurrence or post-surgery complications in cancer patients.

A realistic prognostic model is one that can be used at the beginning of the study and is equally relevant and updatable during follow-up as time-dependent information accumulates; and is robust against non-proportional sub-distributions.

To achieve dynamic risk prediction, the landmarking technique can also be adapted to the competing risks scenario where the primary interest becomes the dynamic prediction of the conditional cause-specific cumulative incidence function, which quantifies subject level probability of survival beyond the pre-defined prediction horizon (t_{hor}) given that the subject is still at risk for the event of interest at a specified time point during follow-up (prediction landmark time, t_{LM}):

$$\begin{aligned}
F_1\{t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}\} &= \frac{F_1\{t_{hor}; \boldsymbol{Z}_{LM}\} - F_1\{t_{LM}; \boldsymbol{Z}_{LM}\}}{1 - \sum_{j=1}^{k} F_j\{t_{LM}; \boldsymbol{Z}_{LM}\}} \\
&= Pr(t_{LM} < t \leq t_{hor}, \epsilon = 1|\boldsymbol{Z}_{LM})/Pr(T > t_{LM}|\boldsymbol{Z}_{LM}) \\
&= 1 - Pr\{(T > t_{hor}) \cup (t_{LM} < T \leq t_{hor} \cap \epsilon \neq 1)|\boldsymbol{Z}\}/Pr(T > t_{LM}|\boldsymbol{Z}_{LM}) \\
&= 1 - \exp[-\{\Lambda_1(t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}) - \Lambda_1(t_{LM} - |\boldsymbol{Z}_{LM}, t_{LM})\}] \\
&= 1 - \exp\{-\int_{t_{LM}}^{t_{hor}} \lambda_1(t|\boldsymbol{Z}_{LM}, t_{LM})\ dt\}
\end{aligned}$$

The *Landmark PSH model* proposed by Liu et al. (2016) [12] extended the Fine-Gray PSH model to the landmark setting, which directly predicts the conditional CIFs in one step, bypassing the need to include time-varying covariate effects under non-proportional sub-distribution hazards. The model can be regarded as fitting the Fine-Gray PSH model to the landmark data set framed by $(t_{LM}, t_{hor}]$, which includes subjects who have not failed from any type of events by t_{LM} and ignoring all events after t_{hor}.

The landmark PSH model takes the form:

$$\lambda_1(t|\boldsymbol{Z}_{LM}, t_{LM}) = \lambda_{10}(t|t_{LM}) \exp\{\boldsymbol{\beta}_{LM}^T \boldsymbol{Z}_{LM}\}, t_{LM} \leq t \leq t_{hor}.$$

Thus the conditional CIF can be calculated from

$$\hat{F}_1(t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}) = 1 - \exp\{-\int_{t_{LM}}^{t_{hor}} \hat{\lambda}_1(t|\boldsymbol{Z}_{LM}, t_{LM})\ dt\}$$

Assuming random right censoring, following the arguments by Struther and Kalbfleisch [13] and Xu and O'Quigley [14], when the underlying covariate effects are time-varying, similar to the approximation of $\boldsymbol{\beta}^*$ shown by van Houweiligen[6], which was elaborated in the previous chapter. We have:

$$\boldsymbol{\beta}^*_{hor} \approx \frac{\int_0^{t_{hor}} \lambda_{10}(t)\boldsymbol{\beta}(t)\ dt}{\int_0^{t_{hor}} \lambda_{10}(t)\ dt}$$

as a weighted average of time-varying covariate effects $\boldsymbol{\beta}(t)$ over time, provided that the cumulative incidence function $F_1(t)$ does not get too large; the censoring rate is not too high prior to t_{hor}; and the covariate effect is not too large and does not vary drastically over time.

3.0 POWER AND SAMPLE SIZE CALCULATIONS

In the previous chapter we have presented estimation procedures for dynamic risk estimation. Power analysis and sample size determination are crucial so that researchers will be able to address some scientific questions with adequate evidence and confidence. Procedures of power and sample size determination are based on two important components: the null and alternative hypotheses and the corresponding test statistic. Survival studies focus on either risk estimation for one or more groups at some fixed time point(s) or the modeling of a complete survival curve and dynamic risk prediction falls in the first category.

When type I error, effect size and certain design effects specified, power level and sample size can be calculated from the other. Yet in survival analysis the study design effects can be complex and inconsistent over time thus simplifying assumptions are often needed. In dynamic risk prediction, the application of landmark technique would further complicate the situation as landmarking essentially subset the original data in a non-random manner. Dynamic prediction technique has been proofed to give valid approximation of the risk at prediction horizon times; when there are two comparison groups, such as randomized clinical trial (RCT), dynamic risk prediction technique can be used to quantify the risk for each group and compare the risks via hypothesis testing. In this chapter, we shall: (1) specify the null and alternative hypotheses in studies using dynamic risk prediction; (2) give the form of the test statistic; (3) derive the asymptotic distribution for the test statistic under both null and alternative hypotheses and (4) provide explicit form of the power function.

3.1 HYPOTHESIS TEST OF RISK DIFFERENCE

In a two-arm random clinical trial (RCT), researchers often aim to test for treatment effect in terms of risk difference and prediction baseline is chosen at some time point after the study baseline. For example, compare 3-year overall survival (OS) between breast cancer patients undergoing two different maintenance regimens, given that these patients were still alive 1 year after treatment initiation. In other words, the difference between two conditional probabilities of failure is being tested with $t_{LM} = 1$ and $t_{hor} = 4$. Assuming no competing risks, for this type of research questions, the hypotheses being tested can be expressed as:

$$H_0 : F(t_{hor}|\boldsymbol{Z}^1_{LM}, t_{LM}) - F(t_{hor}|\boldsymbol{Z}^2_{LM}, t_{LM}) = 0$$

$$H_1 : F(t_{hor}|\boldsymbol{Z}^1_{LM}, t_{LM}) - F(t_{hor}|\boldsymbol{Z}^2_{LM}, t_{LM}) = \delta_1 \neq 0$$

The hypotheses are specific to the choice of *prediction window* $(t_{LM}, t_{hor}]$. We use different vectors of covariates evaluated at t_{LM}, $\boldsymbol{Z}^1_{LM}$ and $\boldsymbol{Z}^2_{LM}$ to represent two comparison groups. These risk profiles can include treatment group indicator, applicable intermediate event status and other relevant covariates.

Under certain scenarios, it is also desirable to compare risk within the same treatment group. such as between subjects who responded to the treatment and those who did not; or subjects with early response to the treatment against those with late response. In a study with treatment and control groups, besides main event time t one could also observe the time of response or the time of some complications or adverse event. Use t_1 and t_2 to represent early and late intermediate event times such that $0 < t_1 < t_2 < t_{LM} < t_{hor}$. The following comparison schemes list examples of meaningful comparisons between two risk profiles:

Table 1: Examples of comparing two risk profiles in the presence of beneficial intermediate event

Comparison groups		Patients from the treatment group		
		No response	$t_s = t_1$ (Early response)	$t_s = t_2$ (Late response)
Patients from the treatment group	No response	—	$H_0: F_{trt}(t_{hor}\|t_s = t_1) = F_{trt}(t_{hor})$ $H_1: F_{trt}(t_{hor}\|t_s = t_1) \neq F_{trt}(t_{hor})$ Compare early responders to non-responders in the treatment group	$H_0: F_{trt}(t_{hor}\|t_s = t_2) = F_{trt}(t_{hor})$ $H_1: F_{trt}(t_{hor}\|t_s = t_2) \neq F_{trt}(t_{hor})$ Compare late responders to non-responders in the treatment group
	$t_s = t_1$ (Early response)	—	—	$H_0: F_{trt}(t_{hor}\|t_s = t_1) = F_{trt}(t_{hor}\|t_s = t_2)$ $H_1: F_{trt}(t_{hor}\|t_s = t_1) \neq F_{trt}(t_{hor}\|t_s = t_2)$ Compare early responders to late responders in the treatment group
Patients from the control group	No response	$H_0: F_{trt}(t_{hor}) = F_c(t_{hor})$ $H_1: F_{trt}(t_{hor}) \neq F_c(t_{hor})$ Among non-responders, compare the treatment to the control groups	$H_0: F_{trt}(t_{hor}\|t_s = t_1) = F_c(t_{hor})$ $H_1: F_{trt}(t_{hor}\|t_s = t_1) \neq F_c(t_{hor})$ Compare early responders in the treatment group to the non-responders in the control group	$H_0: F_{trt}(t_{hor}\|t_s = t_2) = F_c(t_{hor})$ $H_1: F_{trt}(t_{hor}\|t_s = t_2) \neq F_c(t_{hor})$ Compare late responders in the treatment group to the non-responders in the control group
	$t_s = t_1$ (Early response)	—	$H_0: F_{trt}(t_{hor}\|t_s = t_1) = F_c(t_{hor}\|t_s = t_1)$ $H_1: F_{trt}(t_{hor}\|t_s = t_1) \neq F_c(t_{hor}\|t_s = t_1)$ Among early responders, compare the treatment to the control groups	—
	$t_s = t_2$ (Late response)	—	$H_0: F_{trt}(t_{hor}\|t_s = t_1) = F_c(t_{hor}\|t_s = t_2)$ $H_1: F_{trt}(t_{hor}\|t_s = t_1) \neq F_c(t_{hor}\|t_s = t_2)$ Compare early responders in the treatment group to the late responders in the control group	$H_0: F_{trt}(t_{hor}\|t_s = t_2) = F_c(t_{hor}\|t_s = t_2)$ $H_1: F_{trt}(t_{hor}\|t_s = t_2) \neq F_c(t_{hor}\|t_s = t_2)$ Among late responders, compare the treatment and the control groups

t: main event time; t_s: intermediate event time; t_{LM}: prediction landmark; t_{hor} = prediction horizon time, where $0 < t_1 < t_2 < t_{LM} < t_{hor}$.
F_{trt}: probability of failure for subjects in the treatment group; F_c: probability of failure for subjects in the control group.

Table 2: Examples of comparing two risk profiles in the presence of adverse intermediate event

Comparison groups		Patients from the treatment group		
		No adverse event	$t_s = t_1$ (Early adverse event)	$t_s = t_2$ (Late adverse event)
Patients from the treatment group	No adverse event	—	$H_0: F_{trt}(t_{hor}\|t_s = t_1) = F_{trt}(t_{hor})$ $H_1: F_{trt}(t_{hor}\|t_s = t_1) \neq F_{trt}(t_{hor})$ Compare early to no adverse events in the treatment group	$H_0: F_{trt}(t_{hor}\|t_s = t_2) = F_{trt}(t_{hor})$ $H_1: F_{trt}(t_{hor}\|t_s = t_2) \neq F_{trt}(t_{hor})$ Compare late to no adverse events in the treatment group
	$t_s = t_1$ (Early adverse event)	—	—	$H_0: F_{trt}(t_{hor}\|t_s = t_1) = F_{trt}(t_{hor}\|t_s = t_2)$ $H_1: F_{trt}(t_{hor}\|t_s = t_1) \neq F_{trt}(t_{hor}\|t_s = t_2)$ Compare early to late adverse events in the treatment group
Patients from the control group	No adverse event	$H_0: F_{trt}(t_{hor}) = F_c(t_{hor})$ $H_1: F_{trt}(t_{hor}) \neq F_c(t_{hor})$ Among no adverse events, compare treatment to the control groups	—	—
	$t_s = t_1$ (Early adverse event)	$H_0: F_{trt}(t_{hor}) = F_c(t_{hor}\|t_s = t_1)$ $H_1: F_{trt}(t_{hor}) \neq F_c(t_{hor}\|t_s = t_1)$ Compare no adverse events in the treatment group to the early adverse events in the control group	$H_0: F_{trt}(t_{hor}\|t_s = t_1) = F_c(t_{hor}\|t_s = t_1)$ $H_1: F_{trt}(t_{hor}\|t_s = t_1) \neq F_c(t_{hor}\|t_s = t_1)$ Among early adverse events, compare treatment to the control groups	$H_0: F_{trt}(t_{hor}\|t_s = t_2) = F_c(t_{hor}\|t_s = t_1)$ $H_1: F_{trt}(t_{hor}\|t_s = t_2) \neq F_c(t_{hor}\|t_s = t_1)$ Compare late adverse events in the treatment group to the early adverse events in the control group
	$t_s = t_2$ (Late adverse event)	$H_0: F_{trt}(t_{hor}) = F_c(t_{hor}\|t_s = t_2)$ $H_1: F_{trt}(t_{hor}) \neq F_c(t_{hor}\|t_s = t_2)$ Compare no adverse events in the treatment group to the late adverse event in the control group	—	$H_0: F_{trt}(t_{hor}\|t_s = t_2) = F_c(t_{hor}\|t_s = t_2)$ $H_1: F_{trt}(t_{hor}\|t_s = t_2) \neq F_c(t_{hor}\|t_s = t_2)$ Among late adverse events, compare treatment to the control groups

t: main event time; t_s: intermediate event time; t_{LM}: prediction landmark; t_{hor} = prediction horizon time, where $0 < t_1 < t_2 < t_{LM} < t_{hor}$.
F_{trt}: probability of failure for subjects in the treatment group; F_c: probability of failure for subjects in the control group.

For competing risk data, the above hypotheses need to be rewritten in terms of the difference between two conditional cause-specific cumulative incidence functions. Modifying the above example, if researchers are interested in comparing 3-year progression-free survival (PFS) between breast cancer patients undergoing two different maintaining regimens, given that these patients were still recurrence-free one year after treatment initiation. With the event of interest marked as type 1, the hypotheses being tested are:

$$H_0 : F_1(t_{hor}|\boldsymbol{Z}^1_{LM}, t_{LM}) - F_1(t_{hor}|\boldsymbol{Z}^2_{LM}, t_{LM}) = 0$$
$$H_1 : F_1(t_{hor}|\boldsymbol{Z}^1_{LM}, t_{LM}) - F_1(t_{hor}|\boldsymbol{Z}^2_{LM}, t_{LM}) = \delta_1 \neq 0$$

Following from Chapter 2, the estimated risk difference

$$\hat{\delta} = \hat{F}\{t_{hor}|\boldsymbol{Z}^1_{LM}, t_{LM}\} - \hat{F}\{t_{hor}|\boldsymbol{Z}^2_{LM}, t_{LM}\}$$

or

$$\hat{\delta} = \hat{F}_1\{t_{hor}|\boldsymbol{Z}^1_{LM}, t_{LM}\} - \hat{F}_1\{t_{hor}|\boldsymbol{Z}^2_{LM}, t_{LM}\}.$$

The consistency of $\hat{F}(t_{hor}|\boldsymbol{Z}_{LM}, t_{LM})$ for the conditional probability of failure from landmark Cox model and the consistency of $F_1(t_{hor}|\boldsymbol{Z}_{LM}, t_{LM})$ for the conditional cause-specific cumulative incidence function from the landmark PSH model has been discussed in Chapter 2. As the two treatment groups are independent, applying Slutsky's theorem, $\hat{\delta}$ serves as a consistent estimator of the true risk difference δ.

The Wald-type test statistic for testing risk difference is:

$$d_0 = \frac{\hat{\delta}}{\hat{\sigma}(\hat{\delta})}, \tag{3.1}$$

where $\hat{\sigma}(\hat{\delta})$ is the standard error estimate of $\hat{\delta}$.

Under $H_0 : \delta = 0$, $\frac{\hat{\delta}}{\sigma(\delta)} \xrightarrow{d} N(0, 1)$, we reject the null hypothesis when $d_0 \leq z_{\alpha/2}$ or $d_0 \geq z_{1-\alpha/2}$ where z is the critical value from the standard normal distribution with type I error rate α. In addition, a point-wise $100(1-\alpha)\%$ confidence interval (CI) for δ can be constructed as $\hat{\delta} \pm z_{1-\alpha/2}\hat{\sigma}(\hat{\delta})$. It should be noted that the proposed hypothesis test for risk difference can only be appropriately interpreted under given choices of prediction landmark and horizon times. The lengths of prediction windows for 2 comparison groups needed to be identical for the two conditional probabilities of failure to be comparable.

3.2 POWER CALCULATIONS

The power function for the above test, which is the probability of H_0 being rejected when the true risk difference is δ_1, takes the form:

$$\begin{aligned}\pi(\delta_1) &= 1 - Pr\{z_{\alpha/2} \leq \frac{\hat{\delta}}{\sigma(\delta_1)} \leq z_{1-\alpha/2}|\delta_1\} \\ &= 1 - P\{z_{\alpha/2} - \frac{\delta_1}{\sigma(\delta_1)} \leq \frac{\hat{\delta} - \delta_1}{\sigma(\delta_1)} \leq t_{1-\alpha/2} - \frac{\delta_1}{\sigma(\delta_1)}|\delta_1\} \\ &= 1 - \Phi\{z_{1-\alpha/2} - \frac{\delta_1}{\sigma(\delta_1)}\} + \Phi\{z_{\alpha/2} - \frac{\delta_1}{\sigma(\delta_1)}\}.\end{aligned} \tag{3.2}$$

where Φ is the cumulative distribution function of the standard normal distribution.

3.3 VARIANCE ESTIMATION OF RISK DIFFERENCE

The dynamic risk prediction procedures facilitate the estimation of $\hat{\delta}$ yet in the same time complexes its standard error estimation. The variance estimation needs to take into account three sources of variation: $\hat{\boldsymbol{\beta}}$, $\hat{\Lambda}_0(t_{hor})$ and $\hat{\Lambda}_0(t_{LM-})$.

3.3.1 Empirical standard error

The empirical distribution of the failure times can be used as an alternative to non-parametric estimators of survival function or cumulative incidence function. When simulating survival data, by taking a large number (n_e) of realizations, the empirical survival functions as well as the empirical standard error estimate for the risk difference can be obtained. Especially in the cases where time-dependent covariates and (or) time-varying covariate effects are present and there are no closed forms for $F(t_{hor}|\boldsymbol{Z}_{LM}, t_{LM})$ and $\sigma(\delta)$, the empirical distribution can serve as a reliable approximation of the true risk difference and its standard error.

Recall the conditional probability of failure $F\{t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}\}$ defined in equation (2.1):

$$\begin{aligned} F_n\{t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}\} &= P_n\{T_L \leq t_{hor}|T_L > t_{LM}, \boldsymbol{Z}_{LM}\} \\ &= \frac{P_n(T_L \leq t_{hor}) - P_n(T_L \leq t_{LM})}{P_n(T_L > t_{LM})} \\ &= \frac{F_n\{t_{hor}; \boldsymbol{Z}_{LM}\} - F_n\{t_{LM}; \boldsymbol{Z}_{LM}\}}{1 - F_n\{t_{LM}; \boldsymbol{Z}_{LM}\}}. \end{aligned}$$

thus

$$E_n[F\{t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}\}] = \frac{1}{n_e}\sum_{i=1}^{n_e} F_{i;n}\{t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}\},$$

and

$$var_n[F\{t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}\}] = \frac{1}{n_e}\sum_{i=1}^{n_e}(F_{i;n}\{t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}\} - E_n[F\{t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}\}])^2$$

In the presence of competing risk events, the empirical conditional cause-specific cumulative incidence function is:

$$\begin{aligned} F_{n;1}\{t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}, \epsilon = 1)\} &= P_n\{T \leq t_{hor}|T > t_{LM}, \boldsymbol{Z}_{LM}, \epsilon = 1\} \\ &= \frac{F_n\{t_{hor}; \boldsymbol{Z}_{LM}, \epsilon = 1\} - F_n\{t_{LM}; \boldsymbol{Z}_{LM}, \epsilon = 1\}}{1 - \sum_i^k F_n\{t_{LM}; \boldsymbol{Z}_{LM}\}}. \end{aligned}$$

3.3.2 Bootstrap resampling method

Given the complexity of the variance estimation one can resort to the bootstrap resampling method (Efron, 1979)[16]. Start by drawing a bootstrap sample $T_1^*, ...T_N^*$ from the original data set with 100% sampling rate, compute $\hat{\delta}_N^*$. Repeat the previous step B times, yielding estimators $\hat{\delta}_{N,1}^*, ..., \hat{\delta}_{N,B}^*$. The bootstrap standard error can be computed as:

$$\hat{\sigma}(\hat{\delta})_B = \sqrt{\frac{1}{B}\sum_{i=1}^{B}(\hat{\delta}_{N,i}^* - \overline{\delta})^2}$$

The bootstrap $100(1-\alpha)\%$ confidence interval for $\hat{\delta}$ can be constructed accordingly.

3.3.3 Perturbation resampling method

Perturbation resampling [17, 18, 19] is another popular method in survival analysis to overcome the complexity in variance estimation and approximate the distribution of survival estimators. Let $\{V_j^{(b)} : j = 1, ..., N^*, b = 1, ..., B\}$ be $N^* \times B$ independent random samples from a strictly positive distribution with mean and variance equal to one where N^* is the sample size of the landmark data set. Let $pl_s^*(\boldsymbol{\beta}_{LM})$ be the perturbed version of the partial log-likelihood $pl_s(\boldsymbol{\beta}_{LM})$ for the Landmark Cox model from Chapter 2.1 with:

$$pl_s^*\{\boldsymbol{\beta}_{LM}^{(b)}\} = \sum_{t_i \geq t_{LM}} \Delta_{Li} V_j^{(b)} [\boldsymbol{\beta}_{LM}^{(b)T} \boldsymbol{Z}_{i;LM} - \log\{\sum_{t_j \geq t_i} V_j^{(b)} \exp(\boldsymbol{\beta}_{LM}^{(b)T} \boldsymbol{Z}_{j;LM})\}]$$

Solve for $\hat{\boldsymbol{\beta}}_{LM}^{T(b)}$ and it follows that:

$$\hat{\Lambda}_0^{(b)}(t) = \sum_{t_i \leq t, \delta_{li}=1} \hat{\lambda}_0^{(b)}(t_i|t_{LM}) = \sum_{t_i \leq t, \delta_{li}=1} \frac{1}{\sum_{t_i \leq t_j} V_j^{(b)} \exp(\hat{\boldsymbol{\beta}}_{LM}^{T(b)} \boldsymbol{Z}_{j;LM})};$$

$$\hat{F}^{(b)}\{t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}\} = 1 - \exp[\exp(-\hat{\boldsymbol{\beta}}_{LM}^{(b)} \boldsymbol{Z}_{LM}\{\hat{\Lambda}_0^{(b)}(t_{hor}) - \hat{\Lambda}_0^{(b)}(t_{LM-})\}].$$

B replicates of $\hat{F}^{(b)}\{t_{hor}|\boldsymbol{Z}_{LM}, t_{LM}\}$ can be used to obtain $\hat{\delta}^{(b)}$. To construct CIs, one can either use the empirical quantile of the perturbed sample or a normal approximation. The validity of the perturbation resampling procedure can be established following the arguments in Cai et al. [20] and Zhao et al. [21] since the distribution of $\sqrt{N^*}\{\hat{\delta} - \delta\}$ can be approximated by the distribution of $\sqrt{N^*}\{\hat{\delta}^{(b)} - \hat{\delta}\}$

3.3.4 Functional delta method

Another candidate method to derive $\hat{\sigma}(\hat{\delta})$ is via the functional delta method using the robust standard errors of the cumulative hazards using the arguments by Nicolaie et al. [11]. Let $\hat{\Sigma}_{(\hat{\boldsymbol{\beta}})}$ be the robust variance estimator of $\hat{\boldsymbol{\beta}}$ from the landmark Cox model. For simplicity, let $t_{LM} = s, t_{hor} = s + w$. We start by obtaining the asymptotic variance of $\hat{\Lambda}\{s+w|\boldsymbol{Z}(s), s)\} = \exp(-\hat{\boldsymbol{\beta}}^T(s)\boldsymbol{Z}(s)\{\hat{\Lambda}_0(s+w) - \hat{\Lambda}_0(t_{LM-})\}$, which can be estimated by:

$$\sum_{s\le t_i\le s+w,\delta_{li}=1}[\frac{\exp\{\hat{\boldsymbol{\beta}}(s)^T\boldsymbol{Z}_i(s)\}}{\sum_{t_k:s\le t_i\le t_k<s+w}\exp\{\hat{\boldsymbol{\beta}}(s)^T\boldsymbol{Z}_k(s)\}}]^2+\hat{q}(s+w|\boldsymbol{Z}(s),s)^T\hat{\Sigma}_{(\hat{\boldsymbol{\beta}})}\hat{q}(s+w|\boldsymbol{Z}(s),s),$$

and $\hat{q}$ given by

$$\hat{q}(s+w|\boldsymbol{Z}(s),s)=\sum_{s\le t_i\le s+w,\delta_{li}=1}(\omega(\boldsymbol{Z}_i(s))-\bar{\omega}_i)\frac{\exp(-\hat{\boldsymbol{\beta}}(s)^T\boldsymbol{Z}(s)}{\sum_{t_k:s\le t_i\le t_k<s+w}\exp\{\hat{\boldsymbol{\beta}}(s)^T\boldsymbol{Z}_k(s)\}},$$

$$\omega(\boldsymbol{Z}_i(s))=\frac{\partial(\hat{\boldsymbol{\beta}}(s)^T\boldsymbol{Z}_i(s))}{\partial\boldsymbol{\beta}},$$

and $\bar{\omega}_i$ as the weighted average of $\omega(\boldsymbol{Z}_i(s))$:

$$\bar{\omega}_i=\frac{\sum_{t_k:s\le t_i\le t_k<s+w}\omega(\boldsymbol{Z}_i(s))\exp\{\hat{\boldsymbol{\beta}}(s)^T\boldsymbol{Z}_k(s)\}}{\sum_{t_k:s\le t_i\le t_k<s+w}\exp\{\hat{\boldsymbol{\beta}}(s)^T\boldsymbol{Z}_k(s)\}}.$$

The estimate of $\hat{\Lambda}\{s+w|\boldsymbol{Z}(s),s)\}$ and its variance using function delta method will yield the targeted standard deviation (SD) of the conditional probability of failure within each treatment group. In the two groups comparison setting, we assume the covariance between different treatment groups is zero and end up with $\hat{\sigma}(\hat{\delta})=\sqrt{\frac{SD_1}{n_1}+\frac{SD_2}{n_2}}$. This variance estimation procedure borrows the idea of Aalen-Johanssen estimator from the multi-state modeling and can be implemented in the `mstate` package in R.[22][23]

3.4 SAMPLE SIZE DETERMINATION

Sample size calculation in survival analysis have been a extensively studied topic, among which one of the most widely used one is the Schoenfeld formula derived based on the Cox proportional hazards model [24]. In his review on sample size calculations in survival studies, Collett [25] mentioned that most existing methods focus on the two-sample comparison problem and depend on proportional hazards assumption; some with more restricted assumption on the actual distribution of the event times. Many researches have been conducted either to generalize to more than two-sample scenarios or relax the PH assumption.

The sample size calculation for dynamic risk prediction is based on the landmark Cox model and landmark sub-distribution proportional hazards model. These models give valid approximation of conditional probability of failure for the main event at prediction horizon; the corresponding sample size determination procedure can still be used when PH or PSH assumption is violated. The sample size calculation follows the same procedure as that for power analysis; and are similar for single event and competing risks setting. The major difference lies in the set-up of probability of main event.

The proposed hypothesis test for risk difference can only be appropriately interpreted with given choices of prediction landmark and horizon times. The prediction windows for 2 comparison groups needed to be identical for the two conditional probabilities of failure or CIFs to be comparable.

The type I error rate α and desired power $1-\beta$ in the power function are only relevant within the prediction window $w : (t_{LM}, t_{hor}]$. The interest in inference lies in the comparison of risk difference at the prediction horizon time instead of at all time points or for the entire survival curves.

Besides censoring, the data would be manually subset via landmarking, namely not all the subjects (events) will be made use of in risk prediction. As Simon [26] pointed out, it is the number of events rather than the number of subjects that is most important in sample size determination for survival analysis. To derive the number of events/subjects needed when planning such studies, one strategy is to calculate the number of events/subjects needed in

the prediction window to reach the desired power level for the hypothesis testing and restore the subjects/events not included in dynamic risk prediction due to landmarking.

In this chapter we use N_i to represent the sample size for group i for the entire study and E_i as the corresponding number of events; N_i^* and E_i^* to differentiate the parameters for the landmark data set.

Using the Wald-type test statistic in Chapter 3.1.1, let

$$\pi(\delta) = 1 - \Phi\{z_{1-\alpha/2} - \frac{\delta}{\sigma(\delta)}\} + \Phi\{z_{\alpha/2} - \frac{\delta}{\sigma(\delta)}\} = 1 - \beta,$$

assuming allocation ratio $N_1^*/N_2^* = r^*$, solve for N_i^*.

The sample sizes needed in prediction window for each group are :

$$N_1^* = r^* N_2^*, \;\; N_2^* = (1 + 1/r^*)\{\frac{(z_{1-\alpha/2} + z_{1-\beta})SD_\delta}{\delta}\}^2, \tag{3.3}$$

where SD_δ is the standard deviation of the true risk difference δ.

The number of events in the prediction window will be:

$$E^* = \sum_{i=1,2} E_i^* = \sum_{i=1,2} N_i^* F_i(t_{hor}|t_{LM}) = r^* N_2^* F_1(t_{hor}|t_{LM}) + N_2^* F_2(t_{hor}|t_{LM}), \tag{3.4}$$

$F_i(t_{hor}|t_{LM})$ corresponds to the conditional probability of failure estimated from the landmark Cox model or the conditional cause-specific cumulative incidence function estimated from the landmark PSH model for risk profile i.

Generalize to the entire study:

$$E = \sum_{i=1,2} \frac{E_i^*}{P_i(t_{LM} < T_i \leq t_{hor})/F_i}, \tag{3.5}$$

where $F_i(t) = Pr\{E_i(t)\} = 1 - S_i(t), i = 1, 2$ is the probability of failure for each group in the entire study and $P_i(t_{LM} < T_i \leq t_{hor}) = F_i(t_{hor}|t_{LM})\{1 - F_i(t_{LM})\}, i = 1, 2$ is the *unconditional* probability of failure for each group within the prediction window, which is the same as the numerator of equation (2.1). The denominator $P_i(t_{LM} < T_i \leq t_{hor})/F_i$ stands for the proportion of events that fall into the prediction window.

Thus the required number of subjects in the entire study for each group is:

$$N_i = \frac{E_i}{Pr(E_i)} = \frac{E_i}{F_i}, = \frac{E_i^*}{P_i(t_{LM} < T_i \leq t_{hor})}, \tag{3.6}$$

The allocation ratio in the landmark data set, r^*, should not be set up a priori. Even with fixed choices of prediction landmark and horizon times, r^* could only be observed after applying landmarking and could be affected by numerous factors. It is not practical or obtainable to enforce a certain allocation ratio for the landmark data set. On the other hand, subject allocation for the entire study is a crucial study design parameter that requires compliance.

Following from above steps:

$$\begin{aligned} N_1/N_2 &= \frac{E_1^* P_2(t_{LM} < T_i \leq t_{hor})}{E_2^* P_1(t_{LM} < T_i \leq t_{hor})} \\ &= \frac{r^* N_2^*(t_{hor}|t_{LM}) P_2(t_{LM} < T_i \leq t_{hor})}{N_2^*(t_{hor}|t_{LM}) P_1(t_{LM} < T_i \leq t_{hor})} \\ &= \frac{r^* F_1(t_{hor}|t_{LM}) P_2(t_{LM} < T_i \leq t_{hor})}{F_2(t_{hor}|t_{LM}) P_1(t_{LM} < T_i \leq t_{hor})}, \end{aligned}$$

Let $N_1/N_2 = r$, solve for r^*:

$$r^* = \frac{r F_2(t_{hor}|t_{LM}) P_1(t_{LM} < T_i \leq t_{hor})}{F_1(t_{hor}|t_{LM}) P_2(t_{LM} < T_i \leq t_{hor})}$$

Plug in the value of r^* to calculate N_i^*, E_i^*, E_i and N_i using equations $4.1 - 4.4$.

The value of r^* does not only depend on the choice of r but also on the specifications of the probabilities of failure within the prediction window and over the entire time course of the study; which is the reason that some prior knowledge of the pattern of event occurrence is considerably helpful in study planning. For instance, if the main event of interest is some acute disease one would expect a plunge in survival curve during early phase of the study; whereas when studying chronic conditions researchers would expect the events to be more scattered over a longer time span and censoring could also increase as study time elapses.

Although the sample size determination process takes more steps as compared to the regular survival analysis setting, it would result in useful and non-duplicated sample size and number of events information both within the prediction window and over the entire time course of the study. Researchers will be able to examine the sample size calculation result at each step and make relevant and timely correction and adjustment as find justifiable.

3.5 DESIGN PARAMETERS

Design parameters in dynamic risk prediction include:

1. Research question: two-sided or one-sided alternative hypothesis; specifications of t_{LM} and t_{hor};
2. Type I error rate α;
3. Desired power for the test $1 - \beta$;
4. True risk difference δ:
    a. For data containing single event: conditional probabilities of failure within the prediction window $F_i(t_{hor}|t_{LM})$,
    b. For data containing competing event(s): conditional cause-specific cumulative incidence functions within the prediction window $F_i(t_{hor}|t_{LM}, \epsilon = 1)$;
5. Standard deviation of the risk difference SD_δ;
6. Probabilities of failure in the entire study:
    a. For data containing single event: probabilities of event $F_i(t)$,
    b. For data containing competing event(s): cause-specific cumulative incidence functions $F_i(t|\epsilon = 1)$;
7. Subject allocation ratio r;

To set up the above study design parameters, especially items 4, 5, and 6, one needs to specify meaningful expected measurements of the conditional and overall risks using knowledge about the survival pattern of the trial population, frequently some results from earlier investigations.

4.0 SIMULATION STUDIES

4.1 SIMULATION SET-UP

In the first part of simulation studies, we evaluated the performance of landmark Cox model in estimating the conditional probabilities of failure and risk difference using mean squared error (MSE) and examined the effect of sample size, prediction landmark time and effect size on power and coverage probability of the proposed test under two different settings. We also evaluated the empirical type I error rate of the proposed test under the null hypothesis where there was no risk difference.

In the first non-PH setting, the main event times T were generated from a two-parameter Weibull distribution,

$$h(t|\boldsymbol{Z}) = \lambda\kappa t^{\kappa-1}\exp(\beta_1 X + \beta_2 X \ln t),$$

where $(\lambda, \kappa) = (0.12, 1.2)$. The treatment effect was set to be diminishing over time with a small positive value for β_2 and the effect size values were adjusted with varying β_1's.

In the second non-PH setting with intermediate event status as a time-dependent covariate, intermediate (short term) event times T_s were generated from an Exponential distribution

$$h(t_s|\boldsymbol{Z}) = \kappa_1 {t_s}^{\kappa_1}\exp(\beta_1 X),$$

and the main (long term) event times T_l were generated from a two parameter Weibull distribution

$$h(t_l|\boldsymbol{Z}, t_s) = \lambda_2\kappa_2 {t_l}^{\kappa_2-1}\exp(\beta_2 X + \beta_3(t_s)\delta_S),$$

where $\delta_S = I(T_s \leq t_s)$, and $(\kappa_1, \lambda_2, \kappa_2) = (0.8, 0.1, 1.5)$. $(\beta_1, \beta_2, \beta_3)$ could assume different values.

In both setting we let the total sample size N vary from 100 to 1,000 with an increment of 100. The treatment indicator was generated from a binomial distribution with equal allocation $X \sim BIN(N, 0.5)$. Censoring times were generated from an independent uniform distribution resulting in about 20% censoring. After the original data were generated, we created the landmark data sets with different choices of t_{LM} and t_{hor}.

Figure 1 shows how intermediate event status was incorporated in the risk prediction as a time-dependent covariate. Take response to chemotherapy as the intermediate event and cancer recurrence as the main event and assuming earlier response as beneficial. The treatment group was set up to have earlier response times and high overall response rate as compared to the control group. When we chose an early prediction landmark time (the blue line on the *left*), only early responders would be picked up to have the intermediate event; while as we postponed the prediction landmark time to the blue line on the *right*, more and more responders would be identified, including the late responders such that the values of the intermediate event indicator would vary over different t_{LM}'s. Incorporating intermediate event status that varies with different prediction landmark times we could make use of both the intermediate event status and also the time to intermediate event information.

In the second part of the simulation studies the same were repeated the first non-PH setting and generated data under non-PSH setting by further including competing risk events. For simplicity, only two failure types were considered with Type 1 failure as the main event of interest and Type 2 failure as the competing event. The Type 2 event times were generated from an Exponential distribution with $Pr(\epsilon_1 = 2|\boldsymbol{Z}_i) = 1 - Pr(\epsilon_i = 1|\boldsymbol{Z}_i) = 1 - p$. For data containing competing risk events, besides sample size, effect size and prediction landmark times, we also allowed the probability of experiencing the competing events, $1 - p$, to vary.

Each setting was repeated for a total of 1,000 data sets and 500 resampling samples were used to estimate $\hat{\sigma}(\hat{\delta})$.

In the third part of simulation studies we focused on the sample size calculations with different combinations of study design parameters.

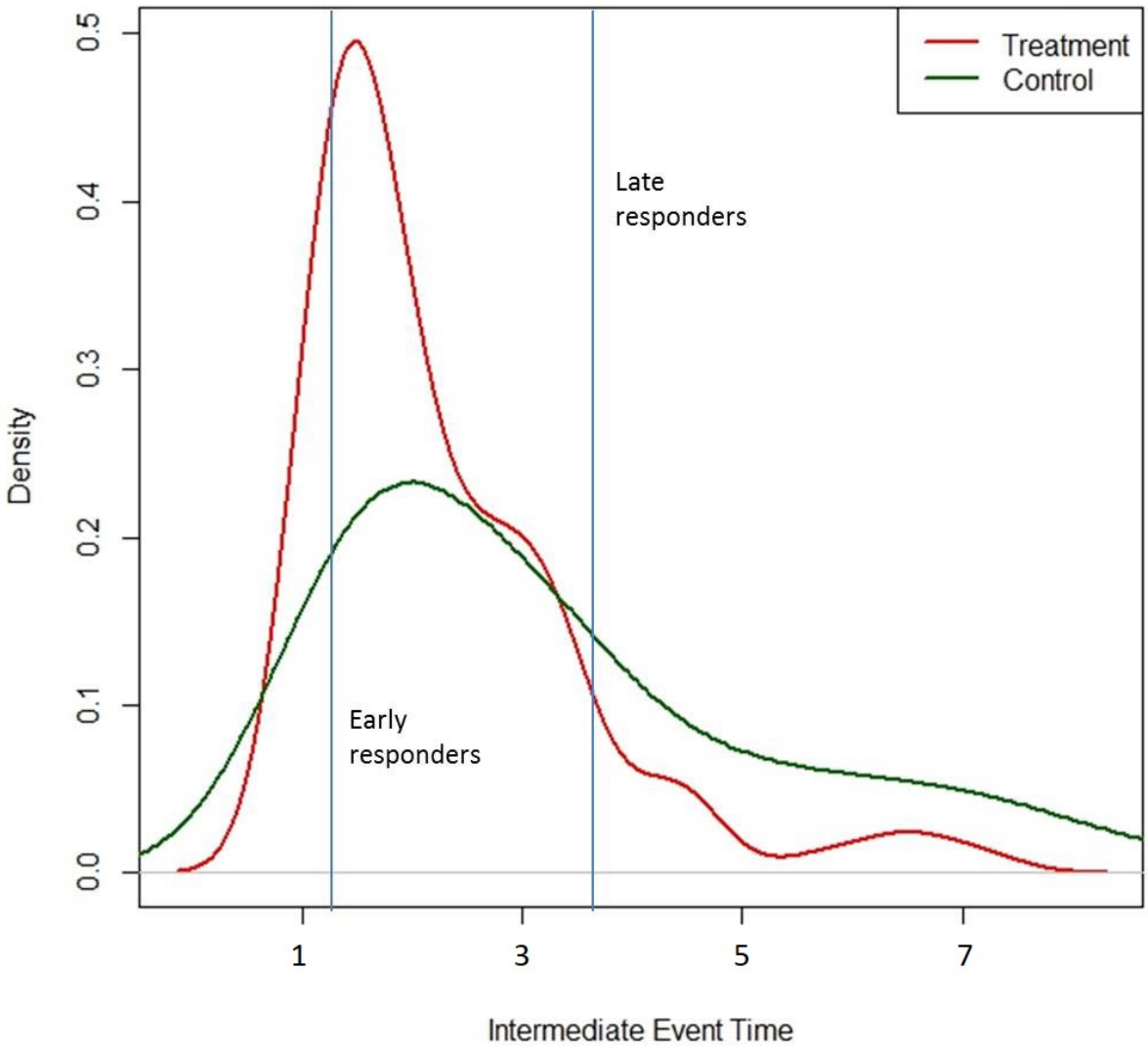

Figure 1: Intermediate event status as a time-dependent covariate

4.2 SIMULATION RESULTS

4.2.1 Power analysis under landmark Cox model

In both settings, the landmark Cox model gave reliable estimates of conditional probabilities of failure and risk differences. The MSE and all types of standard error estimates decreased as total sample size increased. When the sample size was held constant, the higher the effect size (true risk difference) the lower the MSE and standard error estimates, but only to a limited extent.

Total sample size and effect size are the two factors that demonstrated major impacts on the power and coverage probability (CP) of the proposed test. The power of the Wald-type test showed a trend of steady increase when sample size increased. The coverage probability was more stable against changes in sample size and effect size as compared to power; but could be somewhat off the nominal level (95%) either with small sample size and/or effect size and eventually stabilized and maintained the targeted Type I error rate for the test (Tables 3-4, Figure 2).

In the first setting with time-varying treatment effect, we also evaluated the influence of random right censoring on power and CP (Table 3). Although the pre-set censoring rate was 20% in the complete data set, the actual decrease in the number of events observed was not as large since landmarking was applied regardless of censoring. The non-informative censoring would slightly inflate the MSE and standard error estimates and lower the coverage probability and power by a small extent. The simulated data set was not heavily censored. We expect the impact of censoring to be more prominent as the censoring rate increases.

The choice of prediction landmark time is another component in risk prediction that often varies in practice. There were little changes in CP under different t_{LM}'s. We noticed that the coverage probability was unstable for later landmark times especially when paired with a small-to-moderate sample size. Yet the test could achieve and maintain the nominal level as sample size increased (Figure 3, *top* panel). As we postponed the prediction landmark time from 0.25 to 2 with other factors unchanged, the power of the proposed test showed a clear trend of decrease. There is a trade-off between event observations and the sample size in

choosing the landmark time, i.e., the later the landmark time we set, the more information would be available for risk prediction, but the smaller the sample size would become due to truncation of subjects (events) Note that a satisfactory power level (0.80 or higher) for the test requires a sufficient sample size. Examples of sample size requirement under later prediction landmark times are shown in Figure 3. (Figure 3, *bottom* panel).

In the second setting with intermediate event as a time-dependent covariate, we compared the performance of the test using different sample and effect sizes under random right censoring. Even with the effect size doubled, not much change was observed in the total number of events, which could be because that the number of events decreased in the treatment group or increased in the control group. We also found that the power of the test is higher with larger effect sizes and it reaches a high and stable level faster than that with smaller effect sizes. No trend was observed for coverage probability under large or small effect sizes. When the sample size was 600 or larger, the coverage probability stabilized around the pre-set nominal level (Table 4).

With both t_{LM} and t_{hor} fixed, the actual value of the time-dependent covariate had a small impact on the performance of the proposed test statistic, and the performance of the test fluctuated slightly more than that in the first non-PH setting; this coincides with the aforementioned robustness of landmark Cox model against time-dependent covariate or time-varying covariate effect. When the sample size was large enough, the power and coverage probability did not vary much regardless of different effect sizes, choices of prediction landmark times, prediction window width or presence of random right censoring.

Table 3: Power and coverage probability under single event, non-PH setting ($t_{LM} = 0.25, t_{hor} = 5, \delta \sim 0.14$)

Censoring	N	**MSE**	n(event)	Bootstrap resampling			Empirical			Perturbation resampling		
				Ave.$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power
0%	100	0.006	48	0.100	0.960	0.429	0.100	0.974	0.356	0.092	0.757	0.503
	200	0.005	95	0.070	0.954	0.683	0.071	0.939	0.644	0.067	0.902	0.644
	300	0.004	142	0.056	0.944	0.811	0.058	0.952	0.816	0.055	0.945	0.841
	400	0.003	190	0.049	0.947	0.901	0.050	0.957	0.886	0.048	0.947	0.893
	500	0.002	237	0.044	0.947	0.971	0.044	0.959	0.966	0.043	0.952	0.968
	600	0.002	285	0.040	0.946	0.973	0.041	0.951	0.971	0.039	0.948	0.974
	700	0.001	332	0.037	0.957	0.994	0.038	0.950	0.994	0.036	0.950	0.994
	800	0.001	386	0.034	0.953	0.997	0.035	0.959	0.998	0.034	0.962	0.997
	900	0.001	427	0.032	0.950	0.996	0.033	0.961	0.997	0.032	0.949	0.997
	1000	0.001	474	0.030	0.948	0.999	0.031	0.954	0.999	0.031	0.946	0.999
20%	100	0.011	43	0.100	0.868	0.439	0.100	0.959	0.405	0.097	0.864	0.453
	200	0.005	87	0.072	0.934	0.656	0.070	0.937	0.605	0.070	0.928	0.668
	300	0.003	130	0.059	0.955	0.817	0.058	0.954	0.824	0.058	0.952	0.821
	400	0.003	173	0.051	0.953	0.912	0.050	0.951	0.919	0.050	0.954	0.914
	500	0.002	217	0.045	0.943	0.947	0.045	0.939	0.953	0.045	0.944	0.949
	600	0.002	260	0.041	0.954	0.974	0.041	0.951	0.976	0.041	0.952	0.976
	700	0.002	303	0.038	0.943	0.991	0.037	0.947	0.992	0.038	0.946	0.990
	800	0.001	346	0.036	0.953	0.993	0.035	0.955	0.993	0.036	0.951	0.993
	900	0.001	389	0.034	0.953	0.998	0.033	0.951	0.998	0.034	0.949	0.998
	1000	0.001	432	0.032	0.955	0.999	0.032	0.953	0.999	0.032	0.956	0.999

MSE: Mean Squared Error; **CP**: Coverage Probability.

Table 4: Power and coverage probability under single event, non-PH setting with time-dependent covariate ($t_{LM} = 0.25, t_{hor} = 5$, varying δ)

δ	N	**MSE**	n(event)	Bootstrap resampling			Empirical			Perturbation resampling		
				Ave.$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power
0.00	100	0.011	44	0.103	0.952	0.047	0.100	0.940	0.056	0.099	0.936	0.067
	200	0.006	94	0.073	0.942	0.050	0.071	0.945	0.050	0.074	0.952	0.050
	300	0.003	142	0.060	0.955	0.053	0.056	0.945	0.053	0.060	0.948	0.053
	400	0.003	190	0.052	0.954	0.048	0.050	0.949	0.048	0.051	0.954	0.048
	500	0.002	237	0.047	0.941	0.053	0.045	0.940	0.053	0.046	0.942	0.053
	600	0.002	284	0.042	0.933	0.065	0.041	0.934	0.065	0.042	0.933	0.065
	700	0.002	331	0.040	0.947	0.063	0.038	0.941	0.063	0.040	0.948	0.063
	800	0.001	378	0.037	0.955	0.044	0.035	0.945	0.044	0.037	0.950	0.044
	900	0.001	426	0.035	0.953	0.047	0.034	0.947	0.047	0.034	0.949	0.047
	1000	0.001	473	0.033	0.949	0.051	0.032	0.949	0.051	0.033	0.946	0.051
0.12	100	0.011	41	0.106	0.935	0.336	0.103	0.947	0.341	0.080	0.849	0.475
	200	0.006	84	0.075	0.947	0.495	0.073	0.947	0.554	0.059	0.871	0.655
	300	0.004	125	0.061	0.946	0.662	0.058	0.956	0.737	0.049	0.865	0.772
	400	0.003	171	0.053	0.944	0.765	0.051	0.947	0.793	0.042	0.889	0.832
	500	0.002	208	0.047	0.930	0.862	0.045	0.954	0.860	0.038	0.893	0.901
	600	0.002	251	0.043	0.947	0.897	0.042	0.949	0.907	0.035	0.891	0.937
	700	0.002	292	0.040	0.948	0.922	0.039	0.942	0.942	0.032	0.882	0.968
	800	0.001	333	0.037	0.957	0.960	0.035	0.939	0.962	0.030	0.881	0.979
	900	0.001	376	0.035	0.948	0.978	0.034	0.946	0.987	0.029	0.882	0.987
	1000	0.001	417	0.033	0.947	0.984	0.032	0.948	0.986	0.027	0.893	0.993
0.24	100	0.011	43	0.105	0.940	0.729	0.100	0.949	0.752	0.075	0.845	0.870
	200	0.005	85	0.073	0.955	0.942	0.070	0.939	0.958	0.055	0.839	0.973
	300	0.004	128	0.060	0.944	0.986	0.058	0.950	0.996	0.046	0.852	0.997
	400	0.003	171	0.052	0.955	0.990	0.050	0.941	0.997	0.040	0.855	0.998
	500	0.002	213	0.046	0.955	0.999	0.044	0.937	0.999	0.036	0.843	0.999
	600	0.002	256	0.042	0.946	0.999	0.040	0.927	0.999	0.033	0.863	0.999
	700	0.002	299	0.039	0.950	0.999	0.038	0.948	0.999	0.030	0.866	0.999
	800	0.001	341	0.037	0.945	0.999	0.035	0.942	0.999	0.028	0.877	0.999
	900	0.001	383	0.035	0.948	0.999	0.033	0.944	0.999	0.027	0.852	0.999
	1000	0.001	426	0.033	0.948	0.999	0.031	0.945	0.999	0.025	0.879	0.999

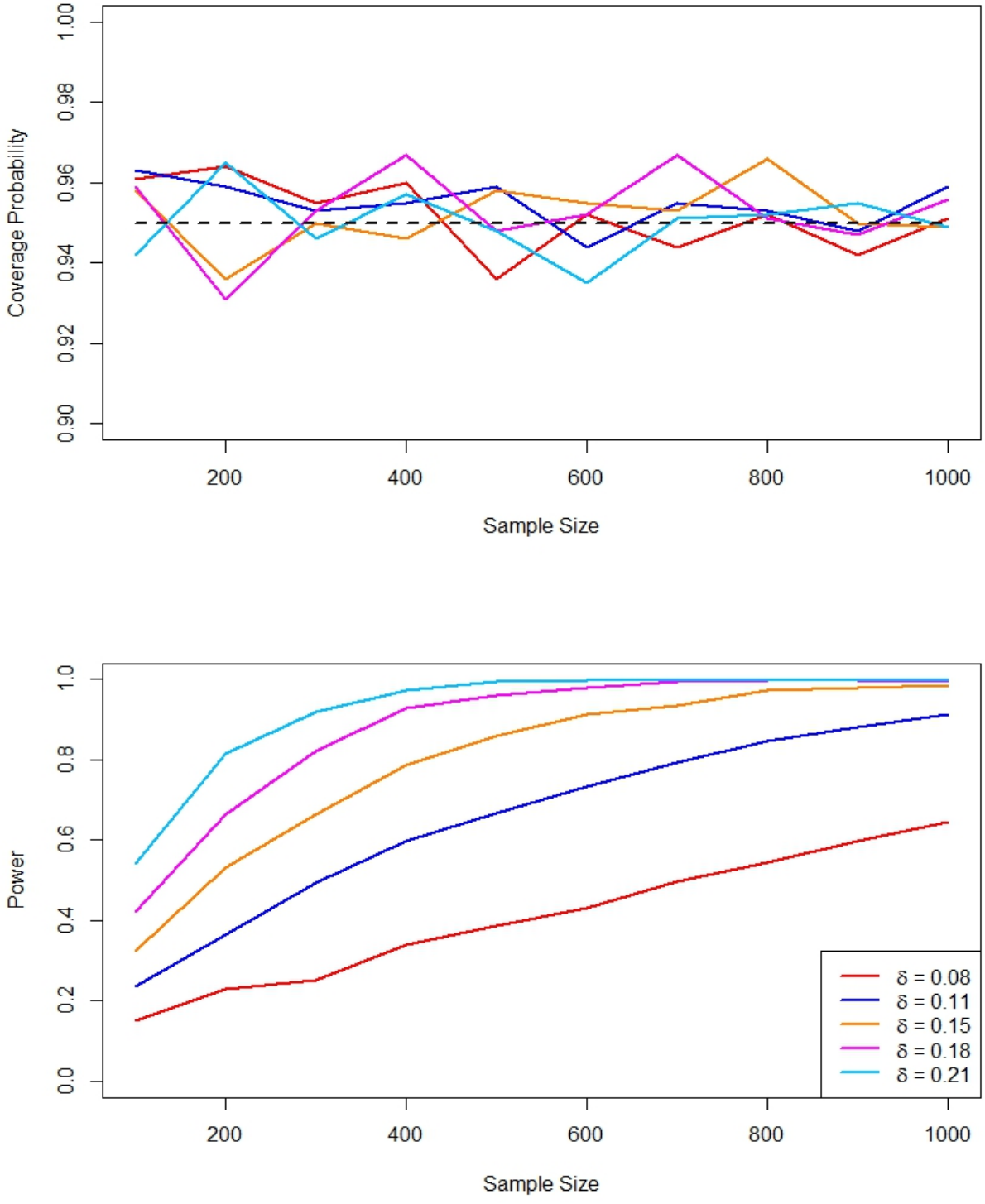

Figure 2: Single Event: Coverage probability and power under different effect sizes

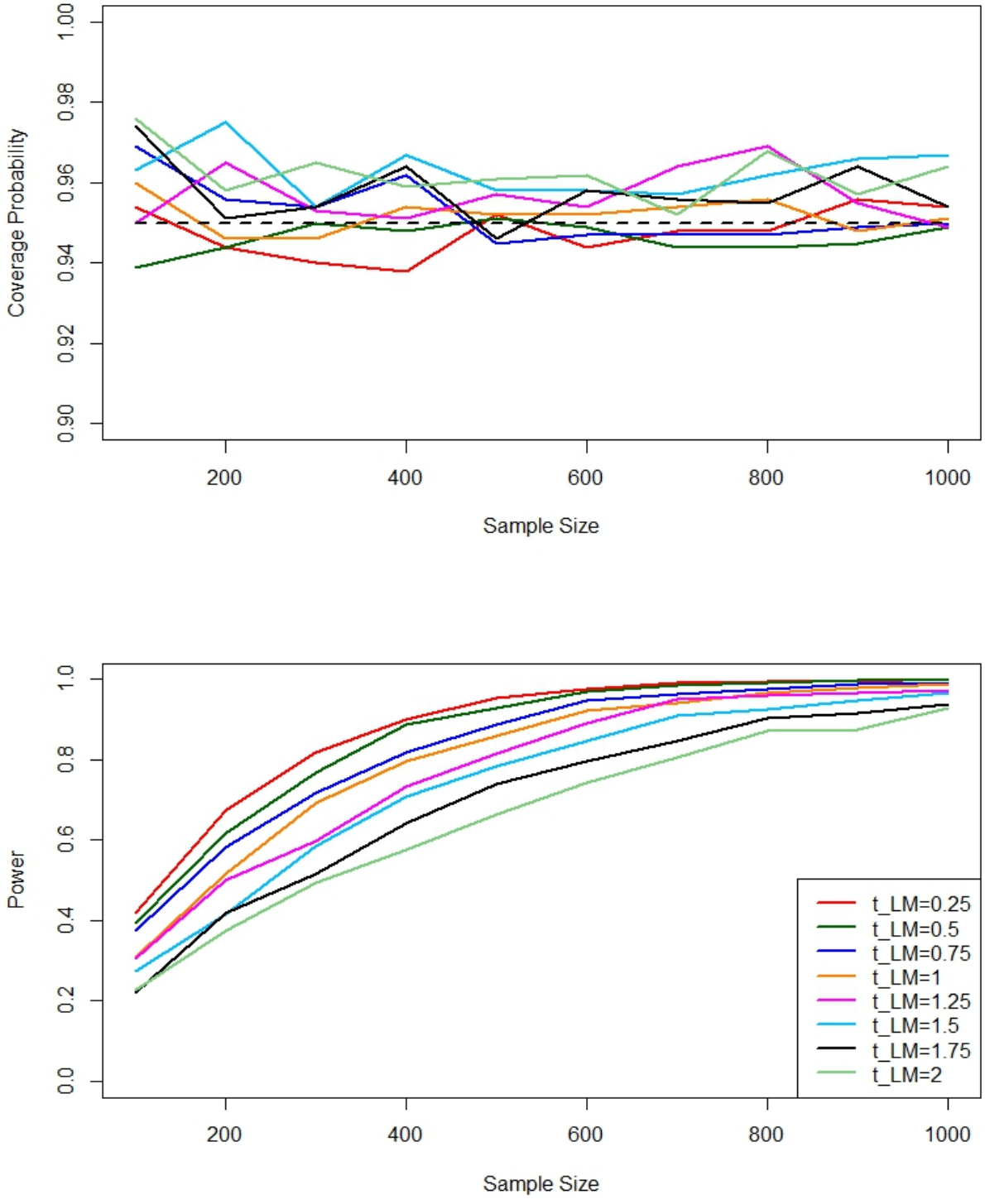

Figure 3: Single Event: Coverage probability and power under different prediction landmark times

4.2.2 Power analysis under landmark PSH model

The landmark PSH model showed satisfying prediction performance for conditional cause-specific cumulative incidence functions and corresponding risk difference under different sample sizes and effect sizes; as well as varying prediction landmark times and probabilities of the competing event and was robust to the violation the PSH assumption. Similar to the single event setting, MSE and all types of standard error estimates decreased as the total sample size or the underlying effect size increased.

The power of the Wald-type test showed a monotonic increasing trend with the increase in sample size or effect size. However, when the effect size we wished to detect was small (Table 6, left), a sample size of 500 per arm would not be large enough to guarantee a high probability of statistically significant test results. The coverage probability but could be somewhat off the nominal level (95%) with small sample size and/or effect size, then temporarily decreased under moderate sample sizes and eventually stabilized and maintained the targeted Type I error rate. (Table 6)

Table 7 included the results for varying prediction landmark times with prediction window width fixed at 3. Likewise, changes in prediction landmark time did not result in major changes in the coverage probability. It was possible for the proposed test to be a little bit conservative under small sample sizes ($N = 100, 200$). The power of the proposed test showed a trend of decrease with the postponing of landmark time; but the amount of decrease was not as much as compared to the single event scenario.

In the competing risks setting, there are three reasons for the smaller number of main events being observed: presence of competing event(s), random right censoring and additional left truncation and administrative censoring by landmarking. As the last two factors impacted both treatment and control groups, we took another step to evaluate the influence of the competing event's probability on power and CP (Table 8). With higher probability of experiencing the competing events, the modified risk set for the main event got inflated, which could in a way introducing diminishing treatment effect. Higher proportion of the competing event(s) would slightly inflate the standard error estimates and lower the power, yet the MSE and coverage probability remained relatively stable.

When the sample size was large enough, the power and coverage probability became stabilized regardless of different effect sizes, choices of prediction landmark times, prediction window width or presence of random right censoring or higher probability of competing event(s).

Table 5: Power and coverage probability under competing risks non-PSH setting ($t_{LM} = 1, t_{hor} = 4$, varying δ)

δ	N	**MSE**	n(event)	Bootstrap resampling			Empirical			Perturbation resampling		
				Ave.$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power
0.07	100	0.012	26	0.108	0.945	0.201	0.112	0.952	0.179	0.102	0.922	0.228
	200	0.006	52	0.077	0.935	0.258	0.079	0.946	0.240	0.074	0.927	0.277
	300	0.004	77	0.063	0.936	0.322	0.065	0.947	0.314	0.061	0.928	0.340
	400	0.003	104	0.054	0.947	0.371	0.056	0.948	0.372	0.054	0.941	0.382
	500	0.003	130	0.049	0.943	0.428	0.050	0.952	0.427	0.048	0.939	0.436
	600	0.002	155	0.044	0.942	0.504	0.049	0.945	0.515	0.044	0.934	0.507
	700	0.002	182	0.041	0.956	0.578	0.041	0.947	0.554	0.041	0.938	0.577
	800	0.002	208	0.038	0.948	0.618	0.039	0.947	0.568	0.038	0.951	0.631
	900	0.001	234	0.036	0.947	0.628	0.037	0.941	0.625	0.036	0.949	0.630
	1000	0.001	259	0.034	0.949	0.694	0.035	0.946	0.680	0.034	0.943	0.701
0.12	100	0.012	25	0.106	0.942	0.327	0.110	0.942	0.334	0.100	0.920	0.362
	200	0.006	49	0.075	0.942	0.515	0.076	0.941	0.504	0.073	0.933	0.535
	300	0.004	73	0.062	0.938	0.654	0.062	0.936	0.648	0.060	0.935	0.656
	400	0.003	97	0.053	0.937	0.766	0.054	0.950	0.764	0.053	0.938	0.771
	500	0.002	121	0.048	0.946	0.861	0.049	0.946	0.828	0.047	0.941	0.871
	600	0.002	146	0.044	0.955	0.906	0.044	0.943	0.847	0.043	0.955	0.907
	700	0.002	170	0.040	0.940	0.924	0.040	0.948	0.927	0.040	0.936	0.923
	800	0.002	194	0.038	0.935	0.960	0.038	0.958	0.947	0.037	0.932	0.957
	900	0.001	219	0.036	0.936	0.962	0.036	0.952	0.953	0.036	0.936	0.971
	1000	0.001	243	0.034	0.941	0.987	0.034	0.952	0.950	0.034	0.945	0.986
0.17	100	0.012	22	0.105	0.921	0.499	0.107	0.946	0.430	0.099	0.920	0.499
	200	0.006	45	0.074	0.945	0.735	0.076	0.949	0.707	0.072	0.940	0.745
	300	0.004	68	0.061	0.950	0.874	0.061	0.956	0.858	0.060	0.933	0.872
	400	0.003	91	0.053	0.946	0.943	0.053	0.946	0.927	0.052	0.934	0.923
	500	0.002	114	0.047	0.945	0.972	0.047	0.945	0.977	0.046	0.938	0.978
	600	0.002	137	0.043	0.941	0.987	0.043	0.940	0.983	0.042	0.934	0.990
	700	0.002	160	0.040	0.940	0.995	0.040	0.942	0.995	0.039	0.935	0.995
	800	0.002	182	0.037	0.947	0.997	0.038	0.948	0.998	0.037	0.943	0.997
	900	0.001	206	0.035	0.952	0.999	0.035	0.950	0.999	0.035	0.944	0.999
	1000	0.001	229	0.033	0.949	0.999	0.033	0.947	0.999	0.033	0.949	0.999

Table 6: Power and coverage probability under competing risks non-PSH setting ($t_{LM} = 1, t_{hor} = 4$, varying δ)

N	$\delta = 0$					$\delta = 0.07$					$\delta = 0.12$					$\delta = 0.17$				
	MSE	n(event)	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**MSE**	n(event)	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**MSE**	n(event)	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**MSE**	n(event)	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power
100	0.013	23	0.120	0.959	0.044	0.013	26	0.112	0.952	0.179	0.013	24	0.110	0.942	0.334	0.012	22	0.107	0.946	0.430
200	0.006	47	0.084	0.963	0.043	0.006	52	0.079	0.946	0.240	0.006	49	0.076	0.941	0.504	0.006	45	0.076	0.949	0.707
300	0.004	71	0.067	0.961	0.036	0.004	78	0.065	0.947	0.314	0.004	73	0.062	0.936	0.648	0.004	68	0.061	0.956	0.858
400	0.003	94	0.061	0.967	0.037	0.003	104	0.056	0.948	0.372	0.003	97	0.054	0.950	0.764	0.003	91	0.053	0.946	0.927
500	0.003	117	0.053	0.957	0.050	0.002	130	0.050	0.952	0.427	0.002	122	0.049	0.946	0.828	0.002	114	0.047	0.945	0.977
600	0.002	141	0.050	0.951	0.057	0.002	156	0.049	0.945	0.515	0.002	146	0.044	0.943	0.874	0.002	137	0.043	0.940	0.983
700	0.002	166	0.045	0.954	0.050	0.002	181	0.041	0.947	0.554	0.002	170	0.040	0.948	0.927	0.002	160	0.040	0.942	0.995
800	0.002	189	0.042	0.963	0.038	0.002	207	0.039	0.947	0.568	0.002	195	0.038	0.958	0.947	0.002	182	0.038	0.948	0.998
900	0.002	212	0.041	0.955	0.053	0.002	233	0.037	0.941	0.625	0.001	219	0.036	0.952	0.953	0.001	206	0.035	0.950	0.999
1000	0.001	235	0.037	0.956	0.044	0.001	259	0.035	0.946	0.680	0.001	244	0.034	0.952	0.950	0.001	229	0.033	0.947	0.999

Table 7: Power and coverage probability under competing risks non-PSH setting (varying δ & $t_{LM}, t_{hor} = t_{LM} + 3$)

δ	N	$t_{LM} = 1$					$t_{LM} = 1.25$					$t_{LM} = 1.5$					$t_{LM} = 1.75$					$t_{LM} = 2$				
		MSE	n(event)	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**MSE**	n(event)	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**MSE**	n(event)	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**MSE**	n(event)	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**MSE**	n(event)	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power
0.00	100	0.013	23	0.120	0.959	0.044	0.015	23	0.126	0.956	0.057	0.016	22	0.136	0.966	0.046	0.016	21	0.143	0.976	0.043	0.018	20	0.150	0.976	0.038
	200	0.006	47	0.084	0.963	0.043	0.007	46	0.080	0.957	0.054	0.009	45	0.096	0.961	0.050	0.009	43	0.097	0.967	0.050	0.009	41	0.110	0.965	0.048
	300	0.004	71	0.067	0.961	0.036	0.005	69	0.072	0.956	0.051	0.006	67	0.079	0.957	0.058	0.006	64	0.084	0.963	0.057	0.006	61	0.085	0.965	0.050
	400	0.003	94	0.061	0.967	0.037	0.004	93	0.063	0.957	0.050	0.004	89	0.065	0.961	0.040	0.004	86	0.068	0.967	0.038	0.005	82	0.075	0.977	0.046
	500	0.003	117	0.053	0.957	0.050	0.003	115	0.056	0.957	0.048	0.004	111	0.059	0.952	0.059	0.003	107	0.062	0.967	0.045	0.004	103	0.067	0.974	0.047
	600	0.002	141	0.050	0.951	0.057	0.002	138	0.052	0.964	0.043	0.003	134	0.055	0.966	0.044	0.003	129	0.059	0.963	0.048	0.003	123	0.060	0.964	0.048
	700	0.002	166	0.045	0.954	0.050	0.002	161	0.049	0.960	0.055	0.002	156	0.049	0.955	0.049	0.002	150	0.054	0.967	0.035	0.003	144	0.056	0.965	0.047
	800	0.002	189	0.042	0.963	0.038	0.002	184	0.044	0.953	0.053	0.002	178	0.048	0.961	0.053	0.002	171	0.049	0.957	0.053	0.002	165	0.053	0.965	0.046
	900	0.002	212	0.041	0.955	0.053	0.002	207	0.041	0.955	0.050	0.002	201	0.045	0.960	0.044	0.002	192	0.048	0.966	0.051	0.002	184	0.051	0.969	0.043
	1000	0.001	235	0.037	0.956	0.044	0.002	231	0.040	0.955	0.047	0.002	223	0.042	0.958	0.049	0.002	214	0.045	0.967	0.047	0.002	204	0.047	0.963	0.444
0.07	100	0.013	26	0.112	0.952	0.179	0.013	22	0.121	0.960	0.140	0.015	22	0.128	0.964	0.135	0.017	21	0.134	0.958	0.132	0.018	20	0.142	0.974	0.120
	200	0.006	52	0.079	0.946	0.240	0.007	44	0.084	0.948	0.236	0.008	43	0.089	0.953	0.196	0.008	42	0.095	0.967	0.179	0.009	41	0.097	0.961	0.146
	300	0.004	78	0.065	0.947	0.314	0.005	67	0.069	0.945	0.297	0.005	65	0.073	0.948	0.259	0.006	63	0.077	0.947	0.207	0.006	61	0.081	0.952	0.209
	400	0.003	104	0.056	0.948	0.372	0.004	88	0.059	0.944	0.340	0.004	87	0.062	0.953	0.299	0.004	84	0.066	0.965	0.271	0.004	82	0.070	0.968	0.212
	500	0.002	130	0.050	0.952	0.427	0.003	110	0.054	0.956	0.364	0.003	108	0.056	0.954	0.342	0.003	105	0.060	0.952	0.280	0.004	101	0.062	0.952	0.267
	600	0.002	156	0.049	0.945	0.515	0.002	133	0.050	0.959	0.436	0.003	130	0.051	0.955	0.391	0.003	126	0.054	0.956	0.349	0.003	122	0.056	0.954	0.310
	700	0.002	181	0.041	0.947	0.554	0.002	153	0.045	0.950	0.495	0.002	152	0.047	0.951	0.438	0.002	147	0.050	0.964	0.407	0.003	142	0.053	0.960	0.349
	800	0.002	207	0.039	0.947	0.568	0.002	176	0.043	0.953	0.520	0.002	173	0.045	0.962	0.453	0.002	169	0.047	0.959	0.420	0.002	163	0.050	0.959	0.380
	900	0.002	233	0.037	0.941	0.625	0.002	198	0.040	0.946	0.576	0.002	195	0.042	0.958	0.509	0.002	190	0.044	0.960	0.451	0.002	183	0.047	0.961	0.415
	1000	0.001	259	0.035	0.946	0.680	0.001	222	0.038	0.955	0.618	0.002	217	0.040	0.951	0.595	0.002	210	0.042	0.959	0.496	0.002	204	0.044	0.964	0.448
0.12	100	0.013	24	0.110	0.942	0.334	0.013	21	0.116	0.953	0.297	0.014	20	0.123	0.953	0.265	0.016	20	0.129	0.951	0.263	0.016	19	0.135	0.963	0.218
	200	0.006	49	0.076	0.941	0.504	0.007	41	0.082	0.958	0.437	0.008	41	0.086	0.944	0.422	0.008	40	0.090	0.953	0.419	0.008	39	0.096	0.962	0.365
	300	0.004	73	0.062	0.936	0.648	0.005	61	0.066	0.947	0.596	0.005	61	0.070	0.958	0.546	0.005	60	0.073	0.950	0.547	0.006	58	0.078	0.969	0.509
	400	0.003	97	0.054	0.950	0.764	0.003	83	0.058	0.950	0.684	0.004	82	0.060	0.946	0.682	0.004	80	0.063	0.964	0.672	0.004	77	0.067	0.958	0.607
	500	0.002	122	0.049	0.946	0.828	0.003	103	0.051	0.943	0.801	0.003	102	0.054	0.954	0.785	0.003	99	0.057	0.960	0.710	0.004	91	0.060	0.956	0.695
	600	0.002	146	0.044	0.943	0.874	0.002	123	0.048	0.951	0.843	0.002	122	0.049	0.957	0.816	0.003	119	0.053	0.966	0.791	0.003	116	0.055	0.951	0.751
	700	0.002	170	0.040	0.948	0.927	0.002	145	0.043	0.945	0.895	0.002	142	0.046	0.945	0.862	0.002	140	0.049	0.961	0.862	0.002	135	0.051	0.952	0.824
	800	0.002	195	0.038	0.958	0.947	0.002	165	0.040	0.945	0.927	0.002	162	0.043	0.940	0.902	0.002	160	0.045	0.954	0.892	0.002	154	0.048	0.958	0.856
	900	0.001	219	0.036	0.952	0.953	0.002	186	0.039	0.944	0.940	0.002	184	0.041	0.944	0.945	0.002	179	0.042	0.955	0.921	0.002	173	0.045	0.947	0.896
	1000	0.001	244	0.034	0.952	0.950	0.002	207	0.036	0.946	0.968	0.002	204	0.038	0.958	0.953	0.002	198	0.040	0.947	0.933	0.002	193	0.042	0.945	0.924
0.17	100	0.011	22	0.107	0.946	0.430	0.013	23	0.110	0.955	0.452	0.014	22	0.115	0.948	0.431	0.014	23	0.122	0.970	0.363	0.016	21	0.128	0.953	0.367
	200	0.006	45	0.076	0.949	0.707	0.006	45	0.078	0.956	0.708	0.007	44	0.082	0.949	0.668	0.007	44	0.084	0.950	0.627	0.008	42	0.090	0.945	0.590
	300	0.004	68	0.061	0.956	0.858	0.004	68	0.064	0.950	0.845	0.004	67	0.067	0.955	0.810	0.005	66	0.070	0.951	0.768	0.006	63	0.072	0.953	0.717
	400	0.003	91	0.053	0.946	0.927	0.003	91	0.056	0.949	0.922	0.003	89	0.057	0.949	0.903	0.004	87	0.059	0.949	0.893	0.004	85	0.063	0.951	0.845
	500	0.002	114	0.047	0.945	0.977	0.002	113	0.050	0.949	0.961	0.003	112	0.052	0.951	0.946	0.003	110	0.054	0.952	0.925	0.003	106	0.057	0.954	0.912
	600	0.002	137	0.043	0.940	0.983	0.002	136	0.045	0.947	0.985	0.002	134	0.047	0.945	0.980	0.002	131	0.049	0.955	0.978	0.003	127	0.051	0.955	0.956
	700	0.002	160	0.040	0.942	0.995	0.002	159	0.041	0.947	0.990	0.002	157	0.043	0.945	0.993	0.002	153	0.046	0.953	0.981	0.002	149	0.048	0.948	0.974
	800	0.002	182	0.038	0.948	0.998	0.002	186	0.039	0.941	0.993	0.002	178	0.041	0.960	0.995	0.002	175	0.042	0.948	0.992	0.002	170	0.045	0.952	0.987
	900	0.001	206	0.035	0.950	0.999	0.001	204	0.036	0.951	0.997	0.002	201	0.038	0.946	0.998	0.002	196	0.040	0.943	0.992	0.002	191	0.042	0.949	0.984
	1000	0.001	229	0.033	0.947	0.999	0.001	227	0.035	0.951	0.999	0.001	223	0.037	0.959	0.999	0.002	219	0.038	0.947	0.999	0.002	213	0.040	0.963	0.998

Table 8: Power and coverage probability under competing risks non-PSH setting ($\delta \sim 0.17, t_{LM} = 1, t_{hor} = 4$, varying $Pr(\epsilon = 2)$)

N	$Pr(\epsilon = 2) = 0.2$					$Pr(\epsilon = 2) = 0.25$					$Pr(\epsilon = 2) = 0.3$					$Pr(\epsilon = 2) = 0.35$					$Pr(\epsilon = 2) = 0.4$				
	MSE	n(event)	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**MSE**	n(event)	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**MSE**	n(event)	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**MSE**	n(event)	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power	**MSE**	n(event)	**Ave.**$\hat{\sigma}(\hat{\delta})$	**CP**	Power
100	0.012	26	0.104	0.946	0.487	0.010	25	0.105	0.967	0.564	0.011	22	0.107	0.946	0.430	0.011	21	0.107	0.947	0.443	0.012	20	0.109	0.949	0.411
200	0.006	52	0.074	0.946	0.759	0.006	49	0.073	0.947	0.725	0.006	45	0.076	0.949	0.707	0.007	43	0.074	0.936	0.701	0.006	39	0.075	0.945	0.664
300	0.004	78	0.059	0.943	0.888	0.004	73	0.060	0.950	0.878	0.004	68	0.061	0.956	0.858	0.004	63	0.060	0.948	0.861	0.004	58	0.061	0.931	0.803
400	0.003	104	0.051	0.926	0.947	0.003	98	0.052	0.944	0.941	0.003	91	0.053	0.946	0.927	0.003	85	0.053	0.947	0.927	0.003	79	0.053	0.945	0.905
500	0.002	130	0.046	0.952	0.983	0.002	122	0.047	0.944	0.976	0.002	114	0.047	0.945	0.977	0.002	106	0.047	0.945	0.968	0.002	97	0.048	0.949	0.957
600	0.002	157	0.043	0.938	0.987	0.002	147	0.042	0.961	0.994	0.002	137	0.043	0.940	0.983	0.002	127	0.043	0.935	0.986	0.002	118	0.043	0.942	0.979
700	0.002	183	0.039	0.950	0.998	0.002	172	0.039	0.955	0.999	0.002	160	0.040	0.942	0.995	0.002	149	0.040	0.941	0.992	0.002	137	0.040	0.945	0.989
800	0.001	208	0.037	0.949	0.999	0.002	195	0.036	0.945	0.999	0.002	182	0.038	0.948	0.998	0.002	170	0.038	0.943	0.995	0.002	156	0.038	0.945	0.990
900	0.001	234	0.035	0.956	0.999	0.001	220	0.034	0.945	0.999	0.001	206	0.035	0.950	0.999	0.001	191	0.035	0.941	0.995	0.001	176	0.036	0.947	0.998
1000	0.001	261	0.033	0.948	0.999	0.001	244	0.033	0.945	0.999	0.001	229	0.033	0.947	0.999	0.001	211	0.033	0.947	0.997	0.001	195	0.034	0.945	0.999

4.2.3 Comparison of variance estimation techniques

The standard error estimates obtained using empirical distribution can be a reference, especially in cases where there is no closed form expression of the standard error. Although the bootstrap method does not alter the estimation procedure, it could underestimate the variance when large sample sizes are used, as pointed out by Minnier et al.[27]. Therefore, we used perturbation resampling to generate perturbed covariates, risk difference estimates, and their weighted average and variance instead of using the ones from the original model.

As expected, performing the proposed test using an empirical standard error (ESE) gave the most stable coverage probability; but the test could be somewhat conservative with small N. The bootstrap resampling standard error (BSE) is very close to the ESE, although it could be slightly higher when smaller sample sizes are used.

By using the perturbation resampling standard error (PSE) we can get the smallest standard error estimates, the narrowest confidence interval for the estimated risk difference and lower coverage probability, although having an inflated chance to detect a statistically significant result. As shown in the right panel of Table 3, under small sample sizes, the narrowest CI had the lowest coverage probability yet the highest power for the test. As the MSE decreased, the coverage probability began to improve and the difference between PSE and ESE/BSE became smaller. When a time-dependent covariate was incorporated in the landmark Cox model (Table 4), the difference between PSE and BSE/ESE became more pronounced and the corresponding PSE coverage probability was far from the nominal level despite having a large sample size or a large effect size.

Perturbation resampling method showed its advantage in the competing risks setting as the risk estimation procedure was more complex and is a weighted one by itself. However, in the single event of interest setting, incorporating additional perturbed weight did not improve the precision of the risk estimation but produced narrower point-wise confidence interval estimates as compared to the bootstrap and empirical methods; thus amplifying the affect of point estimate on the coverage probability. On the other hand, smaller standard error estimates, which could well be underestimated even under small sample sizes, resulted in the highest power among all scenarios for the risk difference test (Table 5).

ESE and PSE tended to produce a tighter band of standard error estimates as compared with BSE. As for the computation time, it took longer time to compute BSE or PSE than did ESE, especially with large sample sizes.

The validity of functional delta method replies upon correct specification of the model covariates as it was derived in the context of landmark cause-specific hazards model[11]. As when the structural part of the risk prediction models changes, the dimension and the values of the variance/covariance matrix for the coefficient estimates could be quite different. One key feature of dynamic prediction technique discussed so far is that it guarantees valid risk estimates at the prediction horizon in the place of the unbiasedness of $\hat{\boldsymbol{\beta}}$.

4.2.4 Sample size and study design

For sample size determination, the later the proposed prediction landmark time, the larger the sample size needed to reach the desired power level, as stated earlier later prediction landmark time would lower the power of the proposed test. Under the same desired power level and prediction landmark time, a smaller sample would be needed if the true risk difference one wished to detect was larger. When the desired power level increased from 0.8 (Figure 4, *left*) to 0.9 (Figure 4, *right*), the sample size needed would be uniformly higher for all prediction landmark time and effect size combinations.

Besides power level, effect size and prediction landmark time, we need to make some assumptions on the underlying distribution of the main event time. For example, when the majority of the events happen in early stage of the study, the influence of postponing the prediction landmark could become more obvious as more events will be discarded via landmarking.

In the competing risks setting, additional assumptions regarding the distribution(s) of competing events will be needed, or at least some knowledge on the probability of observing the competing event(s) in the place of the main event. Previous work has shown that when sub-distribution hazards for both the main and competing events are related to the treatment, the sample size needed when using cause-specific cumulative incidence function as the measure of effect size will be larger; yet it is the more realistic case. [28]

In practice, there would be limited choices of prediction landmark time, effect size and desired power so appropriate assumptions and specifications on other design parameters, such as the standard deviation of the risk difference and overall pattern of events and probability of failure should not be ignored in planning a study with sufficient power that is also cost-efficient.

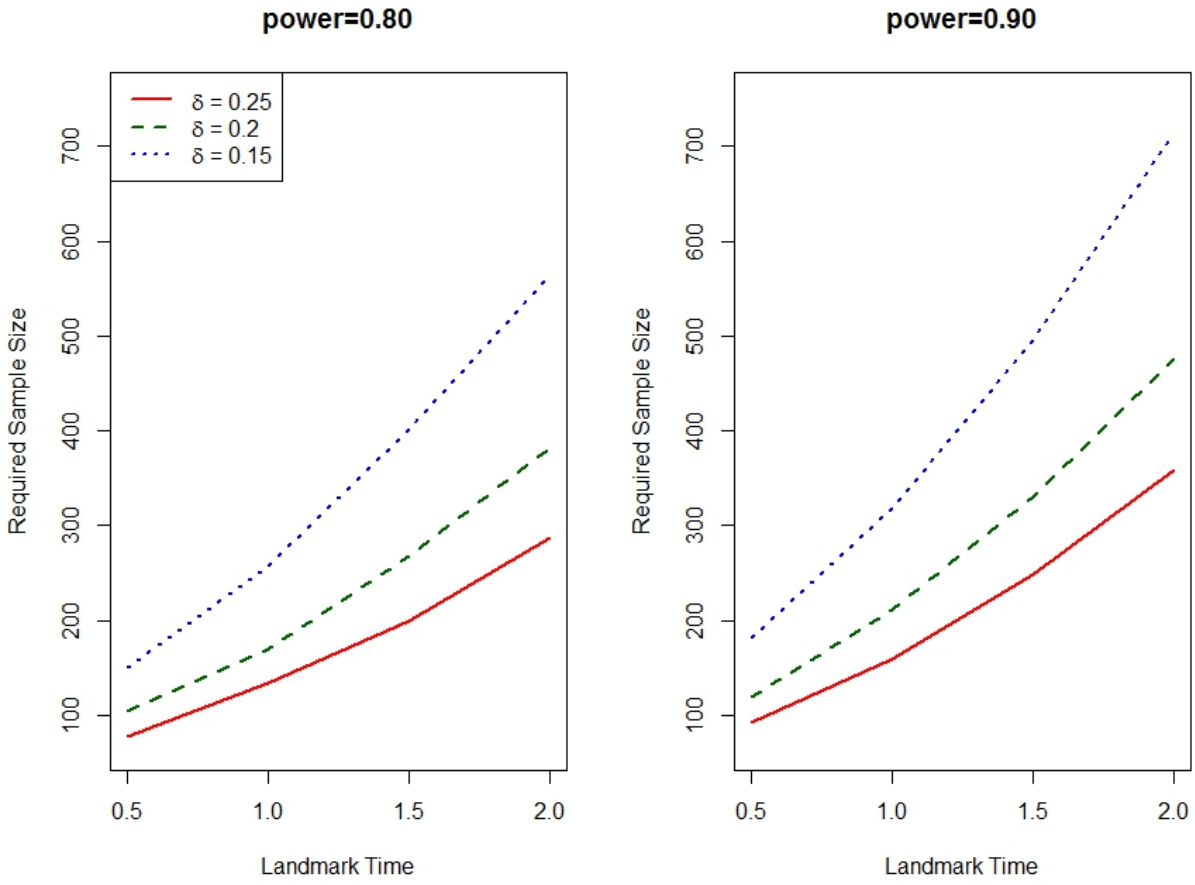

Figure 4: Effects of prediction landmark time, power and effect size on sample size

5.0 DISCUSSION

Landmark Cox and Proportional Sub-Distribution Hazards models provide direct ways to dynamically estimate the conditional probability of failure and conditional cause-specific cumulative incidence function when time-dependent covariates and/or time-varying covariate effects are present. The risk difference produced by taking the difference between the conditional probabilities of failure of conditional CIFs, is a ready-to-use measurement for quantifying treatment effect widely accepted by the clinical researchers. Note that even if correct estimation of the regression coefficients is not the ultimate goal, correct specification of the model is still crucial in predicting probability of failure at the prediction horizon time, especially when time-dependent covariates are involved.

In this dissertation we developed a test statistic for detecting risk difference between two treatment groups in dynamic prediction under both single event and competing risks settings. We then derived a closed form of power function for the test and the corresponding procedure for sample size calculation given desired power, effect size and other study design parameters. The performances of landmark Cox model, landmark proportional sub-distribution hazards model, and the proposed test of risk difference were evaluated using simulation studies with various settings. We also compared the statistical power and coverage probability with different sample sizes, effect sizes as well as prediction landmark time points.

The power of the proposed test was more subject to the influence of sample size and effect size, generally showing a monotone increasing trend with the increase in either of the two. The choices of prediction landmark and prediction horizon times; or the choices of standard error estimation technique was not as obvious since theses changes actually affect the power of the test via the differences in effect sizes. Choice of prediction landmark time points along with the width of the prediction window play an important role in dynamic risk prediction,

which affects power of the test for risk differences. Therefore, although information may increase by including a large set of landmark time points, effective sample size decreases due to landmarking. This trade-off will be explicitly reflected in the power of the test. Still, later landmark times can also be used as long as it answers a scientifically reasonable research question; turning down the idea of later landmark times because of the reduced power as compared to earlier landmark times concerns one single aspect of the problem.

Similar to the scenario with varying landmark times, the effect of effect size on coverage probability is not uniform, especially in small to moderate sample size, which comes from the fact that the estimated $100(1-\alpha)\%$ confidence interval for $\hat{\delta}$ was wider with larger standard error estimate and could result in a more conservative test.

Coverage probability is driven by two factors, the precision of the risk estimation (represented by the magnitude of MSE) and the estimated standard error of the risk difference. With smaller sample size, the biases in risk difference estimation is more pronounced. The decrease in MSE becomes visibly slower as the sample size reaches the moderate stage and higher ($N = 600$ or more). The influence of underlying effect size on standard error estimation is not as large as that of the sample size. There are heuristically three "stages" with respect to how coverage probability behaves with increasing sample size. Under small sample size ($N = 100 - 300$), larger standard error estimates masks part of the effect of the bias in risk estimation; even if the point estimation might not be satisfactory, the 95% point-wise confidence interval could still possess a nominal level coverage probability with its width. When it comes to moderate sample size, there is often fluctuations in coverage depending on the speeds that the biases in risk estimation and standard error estimates decrease, whichever could be slightly slower could be the cause of fluctuation and deviation from the nominal level. Within this stage, the coverage probabilities could be rendered lower than those for the smaller sample sizes and even demonstrate a trend of slight decrease. Finally with larger sample size that leads to both smaller standard errors and estimation biases, the coverage probability starts to increase and stabilize around nominal level. There also exists certain possibility that the test would be conservative with large sample sizes.

The effect of sample size on power and coverage probability of the test is indeed that of the number of events of main interest. Factors that could alter the number of main events observed, including total sample size, censoring rate, probabilities of failure from competing events, as well as choices of prediction landmark and horizon times, should not be considered in isolation. Factors that could contribute to higher power of the test for risk difference include: larger total sample size, higher main event rate, less censoring, larger effect size, wider prediction window.

When performing the test, we recommend that the two risk profiles being compared assume the same prediction window width as the differences in the widths of the prediction window would cause to the conditional risk to differ thus make it impossible to isolate the treatment effect from the prediction window effect. On the other hand, this test can be extended to compare risk profiles within the same treatment group. For example, among breast cancer patients that underwent the same chemotherapy regimen we can test for the risk difference of early responders against late responders; or responders against non-responders by correctly specifying $\boldsymbol{Z}_{LM}^1$ and $\boldsymbol{Z}_{LM}^2$.

Assumptions made when planning the study can sometimes be too optimistic regarding subject recruitment and or pessimistic with respect the survival of subjects. In clinical trials we may experience the latter more often than the former; the selected cohort may be more "healthier" then expected thus we were not able to observe as many events as assumed in the trial design stage, especially with limited time of follow-up and more realistic impact of censoring. Prior knowledge of the attributes of the event of interest, appropriate choices of prediction landmark and horizon times as well as certain assumptions on event and censoring times are needed to make the sample size calculations for dynamic risk prediction reliable yet less complicated.

Also, accrual pattern and follow-up scheme are very important in study planning. The incorporations of accrual and follow-up times can be realized by linking them to the distribution of censoring times. There are various possible accrual patterns such as uniform, increasing or decreasing and can be modeled using certain distribution with examples in Maki [29] and Wang et al.[30]. The length of accrual period has been shown to affect the sample size, but only by a single digit even with possible violation of PH assumption. [31]

The test we developed thus far is a two-sided one. Some future works include extending the method to a one-sided test to accommodate scenarios such as superiority trials and incorporating the effect of accrual into the sample size determination and study planning for dynamic risk prediction. Also, it would be desirable to relax the non-informative censoring to a conditional independent censoring one as in practice the absolute independent censoring assumption is often violated and difficult to verify; also this would help generalize our work to observational studies.

BIBLIOGRAPHY

[1] L. D. Fisher, D. Y. Lin. Time-dependent covariates in the Cox proportional-hazards regression model. *Annual Review of Public Health*. 20: 135-157, 1999.

[2] T. M. Therneau, P. M. Grambsch. Modeling survival data: extending the Cox model. *Statistics for biology and health*, Springer: New York, 2000.

[3] P. K. Andersen, N. Keiding. Multi-state models for event history analysis. *Statistical Methods in Medical Research*,11: 91115, 2002.

[4] A. A. Tsiatis, M. Davidian. Joint modeling of longitudinal and time-to-event data: an overview *Statistics Sinica*, 14, 793818, 2004.

[5] J.R. Anderson, K.C. Cain, R.D. Gelber. Analysis by tumor response. *Journal of Clinical Oncology*, 1(11): 710-719, 1983.

[6] H.C. van Houwelingen, H. Putter. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1):70-85, 2007

[7] H.C. van Houwelingen, H. Putter. Dynamic prediction by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime Data Analysis*, 14(4):447-463, 2008

[8] H.C. van Houwelingen, H. Putter. *Dynamic prediction in clinical survival analysis*, Chapman & Hall/CRC: Boca Raton, 2012.

[9] L. Parast, S-C. Cheng, C. Tian. Incorporating short-term outcome information to predict long-term survival with discrete markers. *Biometrical Journal*, 53(2): 294-307, 2011.

[10] L. Parast, S-C. Cheng, C. Tian. Landmark prediction of long-term survival incorporating short-term event time information. *Journal of American Statistical Association*, 107(500): 1492-1501, 2012.

[11] M.A. Nicolaie, J.C. van Houwelingen, T.M. de Witte, H. Putter. Dynamic prediction by landmarking in competing risks. *Statistics in Medicine*,32: 2013-2047, 2013.

[12] Q. Liu, G. Tang, J.P. Constantino, C-C. H. Chang. Robust prediction of the cumulative incidence function under on-proportional hazards. *The Canadian Journal of Statistics*, 44(2): 127-141, 2016.

[13] C.A. Struther, J.D. Kalbfleisch. Misspecified proportional hazard models. *Biometrika*, 73: 363-369, 1986

[14] R. Xu, J. O'Quigley Estimating average regression effect under non-proportional hazards. *Biostatistics*, 1: 423-439, 2000.

[15] J. P. Fine, R. J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446): 496-509, 1999.

[16] B. Efron. Bootstrap methods: another look at the Jackknife. *The Annals of Statistics*, 7(1):1-26, 1979.

[17] Y. Park, L. J. Wei. Estimating subject-specific survival functions under th accelerated failure time model. *Biometrika*, 90(3):717-723, 2003.

[18] T. Cai, L. Tian, L.J. Wei. Semiparametric Box-Cox power transformation models for censored survival observations. *Biometrika*, 92(3):619-632, 2005.

[19] L. Tian, T. Cai, E. Goetghebeur, L.J. Wei.Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika*, 94(2):297-311, 2007.

[20] T. Cai, L. Tian, H. Uno, S. Solomon, L. Wei. Calibrating parametric subject specific risk estimation. *Biometrika* 97: 389-404, 2010.

[21] L. Zhao, T. Cai, L. Tian, H. Uno, S. Solomon, L. Wei. Stratifying subjects for treatment selection with censored event time data from a comparative study. *Harvard University Biostatistics Working Paper Series*,122, 2010.

[22] L. dr Wreede L, M. Fiocco, H. Putter. The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine*, 99(3): 261-274, 2010.

[23] L. dr Wreede L, M. Fiocco, H. Putter. mstate: an R package for the analysis of competing risks and multi-state models.*Journal of Statistical Software*, 38(7), 2011.

[24] D. A. Schoenfeld. Sample-size formula for the propotional-hazards regression model. *Biometrics*, 39: 499-503, 1983.

[25] D. Collett. *Modeling survival data in medical research (2nd edition)*, Chapman & Hall/CRC: USA, 2003.

[26] S. Simon. Sample size for a survival date model. *www.pmean.com/04/survival/html*, 2004.

[27] J. Minnier, L. Tian, T. Cai A Perturbation Method for Inference on Regularized Regression Estimates. *Journal of American Statistical Association*, 106(496):13711382, 2011.

[28] B. C. Tai, Z. J. Chen, D. Machin. Estimating sample size in the presence od competing risks - cause-specific hazard or cumulative incidence approach? *Statistical Methods in Medical Research*, 0(0)1-14, 2015.

[29] E. Maki. Power and sample size considerations in clinical trials with competing risk endpoints. *Pharmaceutical Statistics*, 5: 159-171, 2006.

[30] S. Wang, J. Zhang, W. Lu. Sample size calculation from the proportional hazards cure model. *Statistics in Medicine*, 31(29): 3959-3917, 2012.

[31] Y. Wang. Sample size calculation based on the semiparametric analysis of short-term and long-term hazard ratios. *Columbia University Academic Commons*, https://doi.org/10.7916/D8ST7X25, 2013.