UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXIV

Model-based prediction on cumulative incidence functions and sample size calculation for competing risks survival data

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Giuliana Cortese

Dottorando: Mohammad Anamul Haque

30 April 2022

Abstract

The primary objective was to estimate the cumulative incidence function (CIF), defined as the probability of occurrence of the main event of interest over time, allowing patients to be censored or to fail from competing events. The CIF is often of great interest in medical research and can be estimated by different regression models and inferential approaches. The performance among cause-specific hazard (CSH), sub-distribution hazard (SDH), pseudo-value, and binomial regression approaches were compared using a simulation study in the presence of competing risks survival data. The empirical bias was found higher under some of these approaches. However, no substantial differences between the estimated and empirical standard errors of the estimators, were reported among the regression approaches, and this is essential in clinical studies to establish a treatment effect with precision. Meanwhile, a slight under-estimation was observed only for the pseudovalue approach. It was found that time-varying regression coefficients improve the coverage probability under the binomial approach. Furthermore, the binomial and pseudo-value approaches showed a gain in efficiency compared to the CSH approach. Additionally, a real data application was illustrated for estimating the CIF of dying from Covid-19 as well as for other causes. Several risk factors and patient characteristics such as sex, age, and race, were found to increase significantly the cumulative risk of death due to Covid. SDH and CSH approaches showed very similar model-based predictions of CIF. Another objective of the thesis was to give guidelines to a new user for estimating the sample size under a fixed design and a group sequential (Gs) design, following the CSH and SDH approaches. For this scope, several simulation studies were performed. The Weibull, exponential, and Gompertz time-to-event distributions were studied under fixed design. When there was a positive treatment effect on the competing event, CSH provided a smaller required sample size than the SDH approach, given a fixed power for all these distributions. Under Gs design, the contribution of a new

treatment was studied by analyzing interim stage clinical data under various competing risks scenarios. Within this scope, efficacy and futility boundaries were computed, and the decision to continue or stop a trial was taken by calculating the conditional power. It was concluded that the SDH approach could be preferred when the main attention is devoted to increasing conditional power, and on the other hand, CSH is the best choice when the main focus is to reduce the required number of events.

Sommario

L'obiettivo principale della tesi è stato quello di stimare la funzione d'incidenza cumulata (CIF), definita come la probabilità cumulata che accada l'evento di interesse nel tempo, sotto la condizione che i soggetti possano essere censurati a destra o sperimentare altri eventi competitivi. La CIF è spesso di grande interesse nella ricerca medica e può essere stimata con diversi modelli di regressione ed approcci inferenziali. È stato condotto uno studio di simulazione per confrontare le prestazioni tra diversi approcci in presenza di dati di sopravvivenza con rischi competitivi: modelli per i rischi causa-specifici (CSH), modelli per i rischi della sotto-distribuzione (SDH), modelli basati sui pseudovalori e modelli di regressione binomiale. Si è trovato che la distorsione empirica è maggiore in alcuni degli approcci considerati. Tuttavia, non è stata rilevata alcuna differenza sostanziale tra gli errori standard stimati e quelli empirici, e ciò è essenziale negli studi clinici che hanno lo scopo di stabilire con precisione l'effetto di un intervento. Il solo approccio basato sui pseudovalori ha mostrato una lieve sottostima. È stato trovato che l'uso di coefficienti di regressione tempo-dipendenti nell'approccio binomiale, migliora le probabilità di copertura. Inoltre, si è constatato che l'approccio binomiale e quello basato sui pseudovalori conducono ad una maggiore efficienza statistica degli stimatori, rispetto al metodo CSH. Infine, è stata riportata un'applicazione su dati reali con lo scopo di stimare la CIF per la mortalità dovuta al Covid-19 come anche la mortalità dovuta ad altre cause. È stato trovato che diversi fattori di rischio e caratteristiche dei pazienti, quali sesso, età e razza, aumentano significativamente il rischio cumulato di morte per il Covid. Gli approcci SDH e CSH hanno mostrato risultati molto simili per quanto riguarda le previsioni della CIF basate sul modello di regressione. Un secondo obiettivo della tesi è stato quello di fornire linee guida per stimare la dimensione campionaria sotto un disegno fisso e sotto un disegno sequenziale a gruppi (Gs) seguendo gli approcci CSH e SDH. Per questo scopo, sono stati condotti diversi studi di simulazione. Abbiamo studiato le distribuzioni esponenziale, Weibull e Gompertz

per il tempo all'evento, sotto un disegno fisso. Sotto l'ipotesi che si sia verificato un effetto positivo del trattamento sull'evento competitivo, l'approccio CSH ha fornito una dimensione campionaria richiesta inferiore rispetto all'approccio SDH, data una certa potenza fissata per tutte le distribuzioni. Nell'ambito dei disegni Gs, abbiamo studiato il contributo di un nuovo trattamento analizzando i dati clinici nei successivi stadi intermedii (interim), sotto diversi scenari di rischi competitivi. In questo ambito, sono stati calcolati i limiti di efficacia e futilità, e la decisione di continuare o interrompere lo studio è stata presa sulla base della funzione di potenza condizionata, calcolata ai vari stadi sequenziali. Abbiamo concluso che l'approccio SDH potrebbe essere preferito quando l'attenzione principale è rivolta all'aumento della potenza condizionata e, allo stesso tempo, si ha preferenza per il metodo CSH quando si mira principalmente a ridurre il numero richiesto di eventi.

to Mother and Father, for the unconditional love

to Annie, for the patient support

to all of my family members and above all
to the kindness of Almighty

Acknowledgements

It had always been on my mind to begin a PhD program, but by the grace of almighty, I was finally able to start in 2018. This widely desired goal is surprisingly very close, so it is hard to put into words all the emotions I feel and the things I want to say about it. Whatever the case may be, I did my best. First of all, I thank my supervisor Giuliana Cortese for guiding me into the Ph.D. journey with patience, precious pieces of advice, the opportunity to work with her. I am grateful for the mates of the cycle XXXIV: Anna, Emmanuel, Pino, Mattia, Fabio, Lorenzo, Silvia, Jacopo, Dung, Laura, and Ovielt. I have enjoyed the last three years spent with you, not only in a professional sense, but also on an individual level. Thanks to the community in Padova, the people in my country specifically from B.T.R.I, Ramnagar, Srimangal, and the faculty and students of SUST, who helped me out. And last but not least, let me thank my precious family members, who are always so kindly ready to assist me. *“So, surely with hardship comes ease. Surely with that hardship comes more ease” (holy Qur’an:94:5–6).*

Contents

List of Figures

List of Tables

Introduction

Overview

Competing Risks (CR) survival analysis answers questions about the time to occurrence of events with the extension of multiple causes of failure. Many studies have investigated identifying appropriate statistical models for a given clinical data by either ignoring the competing events or using inappropriate regression-based statistical methods to analyze complex clinical information. Thus, one of the objectives in this thesis is to consider the competing risks settings to estimate the probability of the event of interest among the many possible events over time using the Cumulative Incidence Function (CIF). The quantity CIF estimates the marginal probability of patients who actually developed the event of interest, regardless of whether they were censored or failed in other competing events. The CIF can be estimated using a conventional survival technique, which is called the Kaplan-Meier (K-M) estimate. When there is only one event of interest, then the CIF equals the complementary of the K-M estimate. However, in the presence of CR data, the K-M method for estimating cumulative incidence, the log-rank test for comparing cumulative incidence curves, and the standard Cox model for assessing covariates lead to incorrect and biased results (Kim, 2007). This bias arises because these conventional techniques assume that all events are independent, which means they censor events other than the event of interest. In this context, four regression approaches were studied to estimate the CIF in the presence of competing events. These are Cause-Specific Hazard (CSH), Sub-Distribution Hazard (SDH), pseudo-value, and binomial regression approaches as proposed by Andersen *et al.* (2012), Fine and Gray (1999), Andersen *et al.* (2003), and Scheike *et al.* (2008) respectively. The idea was to compare all the regression approaches and provide guidelines to users to help them choose the best method. Although the interpretation of the regression parameters for all the regression approaches is not straightforward, but the graphical representation of CIF curves between treatment and control groups are always appealing and thus popular in medical research.

Another objective of the thesis is to compute the sample sizes for fixed and Group sequential (Gs) design under CR survival data. When a randomized clinical trial is designed, sample size computations are important to detect the efficacy of treatments with sufficient power. In a survival study, this size is determined not by the number of patients accrued but by the number of events observed during a specific follow-up period. If the follow-up continues until all patients enrolled in the trial have experienced the event of interest, the required sample size coincides with the number of patients. However, clinical trials often have to be completed within a relatively short period and only part of the trial population experience the event of interest, allowing patients to be censored or fail in competing events. In the fixed design, the CSH and SDH approaches were considered for the Weibull, exponential, and Gompertz time-to-event distributions to estimate the main event's probability over time and determine the necessary number of patients.

In the sample size under the Gs design settings, the objective is to perform analysis at any time points (interim stages) after having some data information. Here, the analysis can be conducted after each patient is accrued (sequential design), but this is unrealistic as it is common for data about efficacy to be available only at discrete times (once or twice each year) while multi-center clinical trials are being monitored. Thus, the motivation for using the Gs design is to allow for the study to be stopped early for efficacy or for futility. Here, the former means faster access to the new treatment during an interim analysis, while the latter refers to the actual unknown effect is far from anticipated under the alternative hypothesis. Additionally, interim analyses are also a requirement of data monitoring committees as it is highly dependent on patients' ethical issues. Specifically, it may be inappropriate to ask patients to continue to participate in trials in which the high-level outcome already seems clear (Gallo *et al.*, 2014). This design helps reduce the number of randomized patients per treatment group, thus saving the resources that can be reallocated later to more productive undertakings. However, Gs design is ubiquitous in modern clinical research, and hence, computation of the decision boundaries and the sample size is not trivial. Here, the sample size fixed design results are extended to the Gs design settings under the CSH and SDH approaches.

Main contributions of the thesis

Regression approaches for the cumulative incidence function

Chapter 1 discusses the computational aspects of the CIF in the non-parametric and semi-parametric regression approaches for competing risks data by reviewing the literature. In particular, the CSH, SDH, pseudo-value, and binomial regression approaches are compared through a simulation study, which was conducted based on the inverse probability method. In the context, Bender *et al.* (2005) explained the general settings for survival data. The survival function is a probability distribution function that maps a number in the domain $[0, \infty]$ to a probability between 0 and 1. Here, random numbers are generated from a standard uniform distribution in the interval [0,1] and then inverted the survival function to transform uniformly distributed random numbers into event times. It is assumed that the hazards for the event of interest and competing events follow Weibull distributions. Specifically, the idea to choose the Weibull distribution is based on the Bone Marrow Transplant (BMT) data as it was found that the survival probability in the Weibull distribution fits the BMT data well (Figure 1.1). The BMT data are available in the `timereg` package in R as well as in the International Blood and Marrow Transplant Research (CIBMTR) study (Sierra *et al.*, 2002). Next, the latent failure time approach is applied to simulate the event-time distribution. The simulation settings have been described in Section 1.4.

Additionally, helpful insights and representations of a complex phenomenon are provided through simulation in Section 1.4.6. In particular, the following are computed: bias, ratios of estimated to empirical standard errors, coverage probabilities (CP), and relative efficiencies of the CIF for both events for the experimental and control groups in all the regression approaches by assuming four discrete time points $\{10, 50, 90, 130\}$ with $3,000$ replications of size 500. Furthermore, the following are shown: the computational techniques of the true CIF, the regression coefficients in the four regression approaches, and the variance and confidence interval calculations of CIF using the pseudo-value regression approach beacuse the estimated variance of CIF using this approach was not straightforward to obtain.

Application to real data

Chapter 2 contains a real data application to validate the simulation results obtained in Chapter 1. Here, we provide practical guidelines for using all the regression approaches and interpreting the results using the R statistical software. In particular, the non-parametric techniques (without covariates) for the CSH and SDH approaches are discussed in Section 2.2.1. Next, Section 2.2.2 describes all the regression approaches (CSH, SDH, binomial, and pseudo-value approaches) to explore the covariates effect. Moreover, in Section 2.3, the results from an application to COVID-19 data are reported to study the competing risks of dying from Covid and other causes. Finally, results from the different approaches are compared and some relevant aspects that have not been explained elsewhere, are highlighted.

Sample size computation using fixed design

Chapter 3 provides guidelines to a new user to estimate the sample size using a fixed design. Here, two of the most popular approaches for regression modeling of competing risks data are investigated: CSH and SDH approaches for the Weibull, exponential, and Gompertz distributions. Specifically, Tai *et al.* (2018) compared sample size calculation for these two different approaches with some limitations. Additionally, the following novel aspects are discussed: the computational methods employed and the comparison between CSH and SDH approaches for Weibull and Gompertz distributions, which have not been explored elsewhere. Next, in Section 3.1, the importance of sample size under fixed design as well as the novel contributions of this method are discussed. Here, the inferential algorithms under the CSH, and SDH approaches are implemented. In particular, Section 3.2 explains the theoretical aspects, while Section 3.3 describes practical guidelines for computing sample size for CSH and SDH approaches with a simulation study. Further, different shape parameter values are studied for the Weibull and Gompertz distributions to observe how sample size behaves when the shape parameter values are changed in Section 3.4. Finally, Section 3.5 explained the summary results. The guidelines for computing sample size when using the fixed design and the results of our simulations provide better insights than simulation studies in the existing literature.

Sample size computation using group sequential design

Chapter 4 focuses on the Gs design used for a new treatment to justify the continuation or interruption of a clinical trial in interim analyses when there are competing risks data. One of the main scopes in Gs design is to calculate the boundary values after adjusting for type I and type II errors. Among several functional forms proposed by Gordon Lan and DeMets (1983), a flexible approach with Wang-Tsiatis bounds. This approach has the possibility of applying the most popularly used methods in terms of O'Brien-Fleming and Pocock bounds in the computations. An efficacy boundary offers the possibility of early stopping and savings in sample size if the alternative hypothesis is true. However, if the null hypothesis is true, then a futility stopping rule has to be added. Within this scope, efficacy and futility boundaries are computed, and finally, conditional power (following the findings of Jennison and Turnbull (1999)) was calculated. The motivation of using this conditional power is that it potentially stops a trial early due to poor or disappointing efficacy results. In the Gs design, the main contributions are as follows: the guidelines to compute the SDH ratios, events size, and conditional power in each interim stage in the CSH and SDH approaches are explained through simulation studies. Moreover, in Section 4.2, the theoretical aspects to compute boundary values are examined. Error spending function and conditional power formulations are then described in Sections 4.3 and 4.4, respectively. A simulation study is undertaken in Section 4.5. Next, the guidelines for deriving sample size using the direct modeling approach is provided in Section 4.6.

Finally, the necessary future directions to create a valid and feasible user-friendly interface to compute fixed and Gs design using R Shiny for the daily use of a practicing statistician are discussed in Chapter 5. Specifically, the functions and the naming of the arguments in existing software are not straightforward or even confusing. Thus, for daily use of a friendly user interface would be of much help. Furthermore, the guidelines to compute sample size under the binomial and pseudo-value regression approaches are also provided. Overall, this thesis will help a new user to compare all competing risks regression approaches and to compute sample size under a fixed or Gs design for competing risks survival data.

Chapter 1

Regression approaches for the cumulative incidence function

1.1 Introduction

Competing-Risks (CR) analysis is a technique that extends the conventional survival analysis such as the Kaplan-Meier (K-M) estimate, the log-rank test, and the Cox regression to handle data that have multiple event types. The Cause-Specific Hazard (CSH) function and the Cumulative Incidence Function (CIF) are two important quantities of interest in medical research. The CIF can be modeled either by the Fine-Gray method (Fine and Gray, 1999) for one particular event of interest, or by performing the cause-specific survival analysis by modeling the CSHs which models the CSH of all causes. The Fine-Gray method is also known as Sub-Distribution Hazard (SDH) approach. When there is only one event of interest, the CIF can be correctly estimated using the K-M estimator, which equals the complementary of the survival function. However, in the presence of CR data, the K-M method for estimation of cumulative incidence, the log-rank test for comparison of cumulative incidence curves, and the standard Cox model for assessing covariates lead to incorrect and biased results (Kim, 2007). This bias arises because the aforementioned conventional techniques assume that all events are independent, which means they censor events other than the event of interest. Under the CR settings, the log-rank test and the Cox regression do not automatically lead to a correct analysis of the CIF, although they can be adapted with minimal effort to make inferences about the CSH function (Guo and So, 2018). This can be overcome with the use of the Fine and Gray methodology, which is specialized to estimate the CIF (Gray, 1988; Fine and Gray, 1999). The Fine-Gray method provides an important contribution to modeling the CIF. However, this method has limitations in terms of interpretation.

The model covariates, although they can be interpreted as having an effect on the CIF, do not directly link to an underlying event rate, which is interpretable in the real world (Andersen *et al.*, 2012).

Then to simulate data in a CR context, the latent failure time model (David and Moeschberger, 1978) has been used in the literature, where survival time and cause are modeled as arising from the minimum of latent failure times corresponding to the different causes (Moriña and Navarro, 2017). In the existing literature, see, for instance, Andersen *et al.* (2002), Kalbfleisch and Prentice (2002), Prentice *et al.* (1978), and Tsiatis (1975), the dependence structure between the postulated latent failure times cannot be identified from the observable data. Moreover, the literature search reported in Beyersmann *et al.* (2009) found that most of the published papers using simulations in a competing risks setting, used a latent failure time model (Moriña and Navarro, 2017).

It is convenient to model survival times through the hazard function because of censoring (Bender *et al.*, 2005). It may be completely specified the joint distribution of event time and event cause through the CSH. In this thesis, the CSH and SDH approaches are discussed to estimate the CIF in presence of competing events. Moreover, modeling these two functions leads to different types of regression models with the presence of covariates. With the CSH approach, the CIF for event k has no direct relation to the CSH rate (λ_k) since this cause-specific CIF is determined by all CSHs. On the contrary, with the SDH approach, CIF can be modeled by its direct relation with the SDH rate (λ_k^*) under the assumption that only one event is possible at a given time t. Furthermore, the CSH and SDH approaches differ in the definition of the risk set. The risk set decreases when there is an event with a competing cause or censoring with the former, whereas with the latter, patients who failed from an event other than the one of interest before t, remain in the risk set. The SDH approach is similar to a Cox proportional regression model, but the cumulative incidence is associated with the SDH rate. The motivation for this model is that the effect of a covariate on the CSH function may be quite different from the effect on CIF. This means that a covariate may have a strong influence on the CSH function, but have no effect on the CIF (Fine and Gray, 1999). Thus, the difference between CSH and SDH is that the CR events are treated differently. The former considers CR events as non-informative censoring, whereas the latter takes into account the informative censoring nature of the CR events (Satagopan *et al.*, 2004).

Meanwhile, this thesis further discussed two alternative regression approaches available in the literature: pseudo-value and binomial. Nevertheless, the interpretation of the regression parameters in all the approaches is not straightforward, depending on the relationship between the CIF and link function. However, the graphical representation of CIF curves between treatment and control groups is always possible to help make clinical decisions.

This chapter is organized as follows. In Section 1.2, the non-parametric (without covariate) estimation technique, and parametric (with covariate) estimation techniques are discussed for all the approaches in Section 1.3. The settings of the simulation studies are mentioned in Section 1.4. In Subsection 1.4.1, the censoring parameter formulation is described. Next, the comparison of link functions is described in Subsection 1.4.2 and the computational techniques to compute true CIF are explained in Subsection 1.4.3. The computation of regression coefficients using the four regression approaches is explained in Section 1.4.4. Then, the variance of the CIF estimators is shown in Subsection 1.4.5. Finally, the simulation settings are described in Section 1.4 and the simulation results based on the four regression approaches are in Section 1.4.6.

1.2 Non-parametric estimation technique

Consider a CR setting with an event (i.e. cause) of interest (type 1; k=1) and a competing event (type 2; k=2). Here, the indicator would be $\varepsilon \in k = \{1, 2\}$. Then assume that, T_1, T_2 are the potential unobservable event times of type $k = 1$ and $k = 2$, respectively. For the CR data, $T = \min(T_1, T_2)$ are observed, and the indicators of the type of event are, $\varepsilon = 1$, if $T = T_1$ and $\varepsilon = 2$, if $T = T_2$.

Denote the observed data on the $i-$th individual (i.e. patient) by (T_i, C_i), $i = 1, \ldots, n$, where T_i and C_i are the event time and censoring time for the $i-$th patient. Then for right-censored CR data, $t_i = \min(T_i, C_i)$ for each patient is observed. The event indicator $\delta_i = \mathbb{1}(T_i \le C_i)$, where $\mathbb{1}(.)$ is an indicator function, $\delta_i = 1$ if $\{T_i \le C_i\}$ and $\delta_i = 0$ if $\{C_i < T_i\}$, and $k_\mathrm{i} \in \{1, 2\}$, for the causes of event types 1 and 2. The CIF for event tpe 1 is the probability that an event of type 1 occurs at or before time t, i.e., $CIF_1(t) = P(T \le t, k = 1)$. In this context, the CIF in clinical trial settings can be defined as follows: assume that, a is the patients accrued time and $\tilde{f}$ is the follow-up time. Then, the probability of a patient who has the event of death within the time interval $[t, a + \tilde{f}]$ can be observed, given that they entered the study at time t. This is a conditional CIF that can be rewritten as $CIF_1(a + \tilde{f} - t) = P(\tilde{T} \le a + \tilde{f} - t, k = 1)$ where $\tilde{T} = T - t$ is the survival time given that the patient enters at time t without having an event before t.

1.2.1 The CSH approach

The advantage of using a non-parametric estimator in the CSH approach is that it provides a template for predicting the CIF in regression models for the CSHs. This approach is used in this study to replace the Nelson-Aalen estimator with its model-based counterparts (Klein *et al.*, 2014). The CSH function of event type k is defined as follows:

$$\lambda_k(t) = \lim_{\Delta t \downarrow 0} \frac{\mathrm{P}(t \leq T < t + \Delta t, \varepsilon = k | T \geq t)}{\Delta t}$$

For simplicity, event type 1 (main event of interest) and event type 2 (competing event) have considered in this thesis. The CIF for main event of interest (type 1) is then determined by considering all other events:

$$CIF_1\left(t\right) = \int_0^t \lambda_1(u) e^{-\{\Lambda_1(u)+\Lambda_2(u)\}} du$$

where, $\Lambda_k(u) = \int_0^u \lambda_k(v)dv$ is the cumulative CSH function for event k and $k = 1, 2$. It is clear that $CIF_1(t)$ involves not only the hazard function but also all the competing CSH functions when $k > 1$. When $k = 1$, the sub-distribution function degenerates to $CIF_1(t) = 1 - \exp(-\Lambda_1(t))$ and becomes a function of only $\lambda_1(t)$.

To estimate $CIF_1(t)$ non-parametrically, let us assume D distinct event time points, $0 = t_0 < t_1 < \ldots < t_D$. Then, at a particular event time t_i, let d_1 and d_2 be the number of patients who experienced event types 1 and 2, respectively, and assume, $\mathcal{R}\left(t_i\right)$ denotes the risk set at event times t_i and includes individuals who did not fail due to any causes or are not censored before t_i. Here, it should be noted that under the CSH approach, a patient is no longer at risk for having the event of interest if he/she experiences a competing event and thus leaves the risk set. Therefore, the CSH rate λ_k is estimated by counting the number of events of type k, divided by the observed number at risk:

$$\widehat{\lambda_k}\left(t_i\right) = \frac{d_k\left(t_i\right)}{\mathcal{R}\left(t_i\right)}.$$

Subsequently, the variance can be estimated by, $\mathrm{V}\widehat{[\Lambda_k(t)]} = \sum_{t_i \leq t} \hat{\lambda}_k(t_i) = \sum_{t_i \leq t} \frac{d_k(t_i)}{(\mathcal{R}(t_i))^2}$ for k$= 1, 2$.

The overall survival function for T can be obtained by using the Kaplan-Meier estimate (Kaplan and Meier, 1958):

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d\left(t_i\right)}{\mathcal{R}\left(t_i\right)}\right)$$

where $d(t_i) = d_1(t_i) + d_2(t_i)$. Alternatively, $S(t)$ can be obtained through $\hat{S}(t) = \exp\left[-\sum_{k=1}^{2} \hat{\Lambda}_k(t)\right]$. Here, $\hat{\Lambda}_k(t)$ is the Nelson-Aalen estimator for the cumulative CSH function for the event type k.

Finally, the CIF function for event type k can be obtained from the CSHs through $CIF_k(t) = \int_0^t \lambda_k(u)S(u)du$, and a natural non-parametric estimate of $CIF_k(t)$ is

$$\widehat{CIF}_k(t) = \int_0^t \hat{\lambda}_k(u)\hat{S}(u)du = \sum_{t_i \leq t} \frac{d_k\left(t_i\right)}{\mathcal{R}\left(t_i\right)} \hat{S}\left(t_i^-\right) \quad \text{for, } k = 1, 2.$$

A step function is returned by the estimator for the CIF, with jumps at time points of observed events of type k, and constant values at times where no events or an observed competing event is observed (Haller, 2014). That estimator for the CIF in a CR setting is a special case of the Aalen-Johansen estimator for transition probabilities in multi-state models (Aalen, 1978). Aalen-Johansen estimator can be obtained as the product–integral of the Nelson–Aalen estimators for the cumulative transition intensities (Borgan, 2014).

1.2.2 The SDH approach

Contrary to the CSH approach, patients who experienced an earlier competing event remain included in the risk set. Thus, in the SDH, the risk set at time t is,

$$\mathcal{R}^*\left(t_i\right) = \left\{i : \left(t \leq T_i\right) \cup \left(t \geq T_i \cap \varepsilon_i \neq 1\right), i = 1, \ldots, N\right\}.$$

A patient who has not failed due to the event of interest by time t is at risk. This includes two distinct groups: those who have not failed due to any cause and those who have previously failed due to another cause. Here, the hazard of the subdistribution can be interpreted as the probability of observing an event of interest in the next time interval while knowing that either the event of interest did not happen until then or that the CR event was observed. It is called a subdistribution function because it is not a proper distribution: as time progresses, the value does not increase from zero to one because a competing event can prevent the event of interest from happening. Gray (1988) described SDH for cause 1 as:

$$\begin{aligned} \lambda_1^*(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t, \varepsilon = 1 | T \geq t \cup \{T \leq t \cap \varepsilon \neq 1\})}{\Delta t} \\ &= \frac{-\partial \log\{1 - CIF_1(t)\}}{\partial t} \end{aligned}$$

The cumulative SDH is defined as, $\Lambda_1(t) = \int_0^t \lambda_1^*(s)ds$. For SDH approach, a direct relationship exists between the $CIF_1(t)$ and SDH rates (λ_1^*) (Fine and Gray, 1999):

$$CIF_1(t) = 1 - S_1^*(t) = 1 - e^{-\Lambda_1^*(t)} = 1 - e^{-\int_0^t \lambda_1^*(u)du}.$$

where λ_1^* is the SDH rate for both groups. This implies

$$\int_0^t \lambda_1^*(u)du = -\log(1 - CIF_1(t)) = g(CIF_1(t)) \tag{1.1}$$

where $g(\cdot)$ is the log link function. In terms of estimation, this means that the occurrence of a competing event is ignored and such patients remain in the risk set until the time at which they are censored for another reason than the competing event. This suggests the existence of the following estimator: $\widehat{\lambda_k^*}(t_i) = \frac{d_k(t_i)}{\mathcal{R}^*(t_i)}$, where $\mathcal{R}^*(t_i)$ is never smaller than $\mathcal{R}(t_i)$. Therefore, the classical K-M is always at least as steep as the estimator of the cause-specific cumulative incidence due to overestimation.

1.3 Modeling using regression approaches

The difference in the cumulative incidence curves between treatment groups is identified either using a Cox PH model for the main event of interest (considering other CR as censored) or with the direct regression model of the effect of covariates on the CIF (without censoring CR events).

1.3.1 The CSH regression approach

The PH model assumed that hazards are proportional in the follow-up period, and a separate model can be fit for each event type. However, the analysis is more powerful when all competing events are combined. Specifically, Prentice *et al.* (1978) and Cheng *et al.* (1998) considered the following Cox proportional hazards models for all causes:

$$\lambda_k(t|x) = \lambda_{k0}(t) \exp\left\{\beta_k^{\mathrm{T}} x\right\} \tag{1.2}$$

where $\lambda_{\mathrm{k0}}(\mathrm{t})$ is the baseline hazard function for cause k, x is a binary covariate and assumed to be equal among events and β_k is the vector of regression coefficients.

The regression coefficients β_k can be estimated for cause k, by maximizing the modified Cox partial likelihood and log partial likelihood as,

$$L(\boldsymbol{\beta_k}) = \prod_{i=1}^{N} \left[\frac{\exp\left(\beta_k x_i\right)}{\sum_{j \in \mathcal{R}(t_i)} \exp\left(\beta_k x_j\right)} \right]^{\delta_i}$$
$$\Rightarrow \ln L(\beta_k) = \sum_{i=1}^{N} \delta_i \left[\beta_k x_i - \ln \left(\sum_{j \in \mathcal{R}(t_i)} \exp\left(\beta_k x_j\right) \right) \right]$$

where $x_i = 0$ and $x_i = 1$ indicate that the i^{th} patient is assigned to the control and experimental groups, respectively. The score statistic is,

$$s = \frac{\partial \ln L(\beta_k)}{\partial \beta_k} = \sum_{i=1}^{N} \delta_i \left[x_i - \frac{\sum_{j \in \mathcal{R}(t_i)} x_j \exp\left(\beta_k x_j\right)}{\sum_{j \in \mathcal{R}(t_i)} \exp\left(\beta_k x_j\right)} \right]$$

Asymptotically, when $N \to \infty$, the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ is normally distributed as, $\sqrt{N}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \simeq \mathcal{N}(\mathbf{0}, \mathbf{V})$, where $\mathbf{V} = \mathbf{I}_{\beta}^{-1}$ is the asymptotic variance-covariance matrix of the $\sqrt{N}\widehat{\beta}$.

Here, let us assume the asymptotic variance of $\sqrt{N}\hat{\beta}$ is obtained from the diagonal elements of information matrix, thus $V = V_{jj}$. According to the Wald test, the test statistic for the null hypothesis $\beta = \beta_0$ is

$$Z = \frac{\sqrt{N}\left(\widehat{\beta} - \beta_0\right)}{\sqrt{\widehat{V}}} \simeq \mathscr{N}(0, 1)$$

However, in practice, the variance is evaluated under the alternative hypothesis (H_{A} : $\beta \neq \beta_0$). So, let V be the variance of $\sqrt{N}\hat{\beta}$ for the alternative. Theoretically, it is proved by Slutsky's theorem that, the distributions of Z and $\sqrt{N}\left(\widehat{\beta} - \beta_0\right)/\sqrt{V}$ are equivalent for large N (Demidenko, 2013, 2007; Bickel and Doksum, 2001).

Now, let us consider the time-varying effect in the model. When the covariate effects are strongly time-varying and when there is an extended follow-up or the data at hand are sufficiently large, then the Cox model will often be unable to describe sufficiently noticeable and essential features of the data (Cortese *et al.*, 2010). Then, there is a need for alternative and more flexible regression models to extract key data elements. Moreover, the Cox proportional hazard model and the additive model by Lin and Ying (1994) do not allow the covariates to have a time-varying effect. However, one extension of the former that allows some effects to be non-proportional assumes that the hazard is of the following form:

$$\lambda_{k,i}(t|\mathbf{x_i}, \mathbf{z_i}) = \lambda_0(t) \exp[\beta^{\mathrm{T}}(t)\mathbf{x_i} + \gamma^{\mathrm{T}}\mathbf{z_i}]$$

This extension allows some covariates to have time-varying effects while that of other covariates is of the standard relative risk form. The parameters of this model can be estimated by considering the partial likelihood and choice of smoothing parameters (Cortese *et al.*, 2010). A flexible additive-multiplicative model that combines the Cox proportional model and Aalen's additive model (Cox-Aalen model) is proposed by Scheike and Zhang (2002, 2003):

$$\lambda_{k,i}(t|\mathbf{x_i}, \mathbf{z_i}) = \left\{\alpha_k^{\mathrm{T}}(t)\mathbf{x_i}\right\} \exp\left\{\gamma_k^{\mathrm{T}}\mathbf{z_i}\right\}$$

Here, some covariates, $\mathbf{x_i}$, have additive and time-varying effects and others $\mathbf{z_i}$, have constant multiplicative effects. This model is a special form of the Cox model when $\mathbf{x_i} = 1$. Moreover, when $\mathbf{z_i} = 0$, it leads to Aalen's additive model. Here, it is suggested to include those covariates that have a time-varying effect in the additive part of the model.

Next to predict the CIF is not straightforward when using the CSH approach. To do so for a particular event type, the fitted cause-specific Cox model has to be used for each event type. Here, if we assume that, the goal is to fit separate models to each of the k events for the given covariates $\mathbf{x}$, then the cause-specific Cox model leads to,

$$\hat{\Lambda}_k\left(t|\mathbf{x}\right) = \exp\left(\hat{\boldsymbol{\beta}}_k^{\mathrm{T}}\mathbf{x}\right)\hat{\Lambda}_{k0}(t),$$

where $\hat{\beta}_k$ is the maximum partial likelihood estimate and $\hat{\Lambda}_{k0}(t)$ is the Breslow estimator of the baseline cumulative CSH function.

Then, the predicted CIF is,

$$\widehat{CIF}_k\left(t|\mathbf{x}\right) = \int_0^t \hat{S}\left(s^{-}|\mathbf{z}\right) d\hat{\Lambda}_k\left(s|\mathbf{x}\right)$$

where the predicted survival function is,

$$\hat{S}\left(t|\mathbf{x}\right) = \prod_{s:t_i=s\leq t}\left[1 - \hat{\Lambda}_0\left(s|\mathbf{x}\right)\right]$$

where $\hat{\Lambda}_0\left(t|\mathbf{x}\right) = \sum_{k=1}^{2}\hat{\Lambda}_k\left(t|\mathbf{x}\right)$ is the predicted cumulative baseline function estimate for a patient with the covariates $\mathbf{x}$.

1.3.2 The SDH regression approach

For the SDH approach, the likelihood function differs from that of the CSH approach in terms of the definition of risk set. Although the risk set is unconventional, it leads to a proper partial likelihoood (Fine and Gray, 1999), which can be expressed as:

$$\tilde{L}(\boldsymbol{\beta_k}) = \prod_{i=1}^{N} \left[\frac{\exp{(\beta_k x_i)}}{\sum_{j \in \mathcal{R}_i^*} \exp{(\beta_k x_j)}} \right]^{\delta_i}$$

To allow the SDH model to have time-dependent covariates or to test proportionality by adding a time-dependent covariate, the following model can be considered:

$$\lambda_1^*(t|\mathbf{z}) = \lambda_{10}^*(t) \exp[\beta^{\mathrm{T}}(t)\mathbf{z_i} + \gamma^{\mathrm{T}}\mathbf{z_i}].$$

Alternatively, a flexible Cox-Aalen subdistribution hazards model can be used:

$$\lambda_1^*(t|\mathbf{x},\mathbf{z}) = \left\{\alpha_k^{\mathrm{T}}(t)\mathbf{x_i}\right\} \exp\left\{\gamma_k^{\mathrm{T}}\mathbf{z_i}\right\}$$

where $\mathbf{x}$ is a $(p+1)-$ dimensional covariate with the first element as 1 for all patients, and $\mathbf{z}$ is a $q-$ dimensional covariate.

After fitting the CR models, one can use the models to make predictions about future CIFs. For the Fine and Gray model, predicting them for the event of interest is a straightforward task because the subdistribution hazard is modeled directly, and the CIF is only one transformation away. The Cox-type proportional sub-distributional hazard model can be written as (Fine and Gray, 1999):

$$-\log\{1 - CIF_k(t \mid \mathbf{X})\} = \int_0^t \lambda_{10}^*(u) \exp(\mathbf{X}^T \boldsymbol{\beta_k}) du = \exp(\mathbf{X}^T \boldsymbol{\beta_k}) \int_0^t \lambda_{10}^*(u) du$$

Then, the predicted CIF with the time-invariant covariates, $\mathbf{X} = \mathbf{x}$, can be estimated by,

$$\widehat{CIF}_k\,(t|\mathbf{x}) = 1 - \exp\left[-\hat{\Lambda}_{k0}(t) \exp\left(\hat{\boldsymbol{\beta}}_k \mathbf{x}\right)\right]$$

where $\hat{\Lambda}_{k0}(t)$ is the baseline cumulative subdistribution hazard function.

Although the estimation technique by Fine and Gray (1999) is efficient to estimate the proportional SDHs, alternative approaches such as pseudo-value and binomial regression approaches have more flexibility to model CIF directly through different link functions.

1.3.3 The pseudo-value regression approach

The pseudo-value approach is used to make inferences with incomplete survival data based on pseudo-values or pseudo-observations obtained from jackknife statistics constructed from non-parametric estimators for the quantity of interest (Klein *et al.*, 2014). These pseudo-values are then used as outcome variables in a generalized linear model, and model parameters are estimated using generalized estimating equations (GEE) (Liang and Zeger, 1986). Most importantly, this method can handle the assumption of proportional hazard being violated by the cost of minimal loss of efficiency, and can be very easily applied to a broad class of multi-state models for which standard regression analysis is often not available. Furthermore, pseudo-value regression can be easily implemented using existing software packages once the pseudo-values have been obtained (Klein *et al.*, 2014).

When using the pseudo-value approach, a set of fixed time points have to be selected to perform the analysis because using all event times is not practical for large datasets due to the complexity of algorithm. Although five to ten time points spaced equally on the event scale works well in most cases, the number of points can be chosen based on the objectives of interest by the researchers (Zhang *et al.*, 2008; Klein and Andersen, 2005; Klein, 2006). Here, consider a prefixed grid of time points, $t_1, \ldots, t_M$. At any grid time point, t_j, the CIF can be estimated by a standard non-parametric estimator based on the complete data set, $\widehat{CIF}_1(t_j)$, and based on a sample of size $n-1$, which is obtained by deleting the i^{th} observation, $\widehat{CIF}_1^{(i)}(t_j)$. The pseudo-value of the i^{th} patient at time t_i is defined as,

$$\widehat{\theta}_{ij} = n\widehat{CIF}_1(t_j) - (n-1)\widehat{CIF}_1^{(i)}(t_j)$$

where, $i = 1, 2, \cdots, n; j = 1, 2, \cdots, m$. Once a pseudo-value for each patient is obtained, these are used to estimate the parameters of a generalized linear model. Pseudo-value regression works when the pseudo-values come from a consistent and an approximately unbiased estimator of the parameter of interest being modeled. The regression parameters can be estimated by solving a pseudo-score equation and its covariance matrix can be estimated by a sandwich variance estimator (Klein and Andersen, 2005). Here, it should be noted that the GEE approach requires the selection of a working covariance matrix, and with that selection, one may improve efficiency.

1.3.4 The binomial regression approach

The key to using the binomial regression for survival outcomes is that at any time horizon after the time origin the event status is binary and takes the value 1 if the event has occurred and 0 otherwise (Klein *et al.*, 2014). The inference is based on weighting the observed status by the so-called inverse of the probability of censoring weights (IPCW), the details of which can be found in the work of Van der Laan *et al.* (2003). This is a very general regression model that allows covariates to have time-varying effects (Zhang *et al.*, 2008), which is in contrast to the pseudo-value approach that assumes that the censoring mechanism is independent of observed covariates. Here, one can specify a working regression model for the censoring times to increase efficiency and to reduce bias. Moreover, the advantages of binomial regression are feasibility and direct interpretation of regression parameters, while the disadvantages are that the censoring mechanism needs to be modeled and the estimates of the regression coefficients depend on that of the baseline risk (Klein *et al.*, 2014). By assuming different known link functions, the proposed model ends up with different known models. For instance, a fully non-parametric additive model as explained by Scheike and Zhang (2011) can be expressed as,

$$-\log\{1 - CIF_1(t; \mathbf{x}, \mathbf{z})\} = \alpha(t)^\top \mathbf{x} + \gamma^\top \mathbf{z}$$

Although some goodness-of-fit tests have been studied to test the time-varying effects, in practice, it is sufficient to plot and visually examine the estimated regression function with the confidence bands (Scheike *et al.*, 2008).

1.4 Simulation studies

Simulations have been conducted based on the inverse probability method. The general settings under survival data is explained by Bender *et al.* (2005) whereas the CR survival data for generating event time distribution are discussed by Beyersmann *et al.* (2009, 2011), and Moriña and Navarro (2017). The main consideration in the process of simulation of event times, is that since the survival function $S(t)$ is a probability distribution function that maps a number in the domain $[0, \infty]$ to a probability between 0 and 1, random numbers can be generated from a standard uniform distribution in the interval [0,1]. Then one can invert the survival function $S(t)$ and transform uniformly distributed random numbers into event times (Wan, 2017). The survival function with a marginal cumulative baseline hazard function $\Lambda_0(t)$ and relative risk component

$\exp(\beta'\mathbf{X})$ can be written as,

$$S(t \mid \mathbf{X}) = \exp\left(-\Lambda_0(t)\exp(\mathbf{X}'\beta)\right)$$

where for two competing causes, $\Lambda_0(t) := \int_0^t \sum_{k=1}^2 \lambda_{0k}(u)\mathrm{d}u = \Lambda_{01}(t) + \Lambda_{02}(t)$ and $\Lambda_{0k}(t)$ are the CSH functions. Then, with $\mathrm{U} \sim Uniform(0,1)$, the random variable

$$T = S^{-1}(\mathrm{U} \mid \mathbf{X}) = \Lambda_0^{-1}\left(-\frac{\log(1-\mathrm{U})}{\exp(\mathbf{X}'\beta)}\right)$$

has the survival function $S(\cdot \mid \mathbf{X})$. This means that to generate a survival time $T \sim S(\cdot \mid \mathbf{x})$, given the covariate vector, it suffices to draw u from $\mathrm{U} \sim Uniform(0,1)$ and to make the inverse transformation $t = S^{-1}(\mathrm{u} \mid \mathbf{x})$.

Moreover, the computation of the inverse of the cumulative baseline hazard function, $\Lambda_0^{-1}(\cdot)$ is not straightforward and thus requires numerical inversion. For the complex cumulative hazard function, one can use the R function *uniroot*. It searches for a root such that,

$$\Lambda_0(t) + \left(\frac{\log(1-\mathrm{u})}{\exp(\mathbf{x}'\beta)}\right) = 0.$$

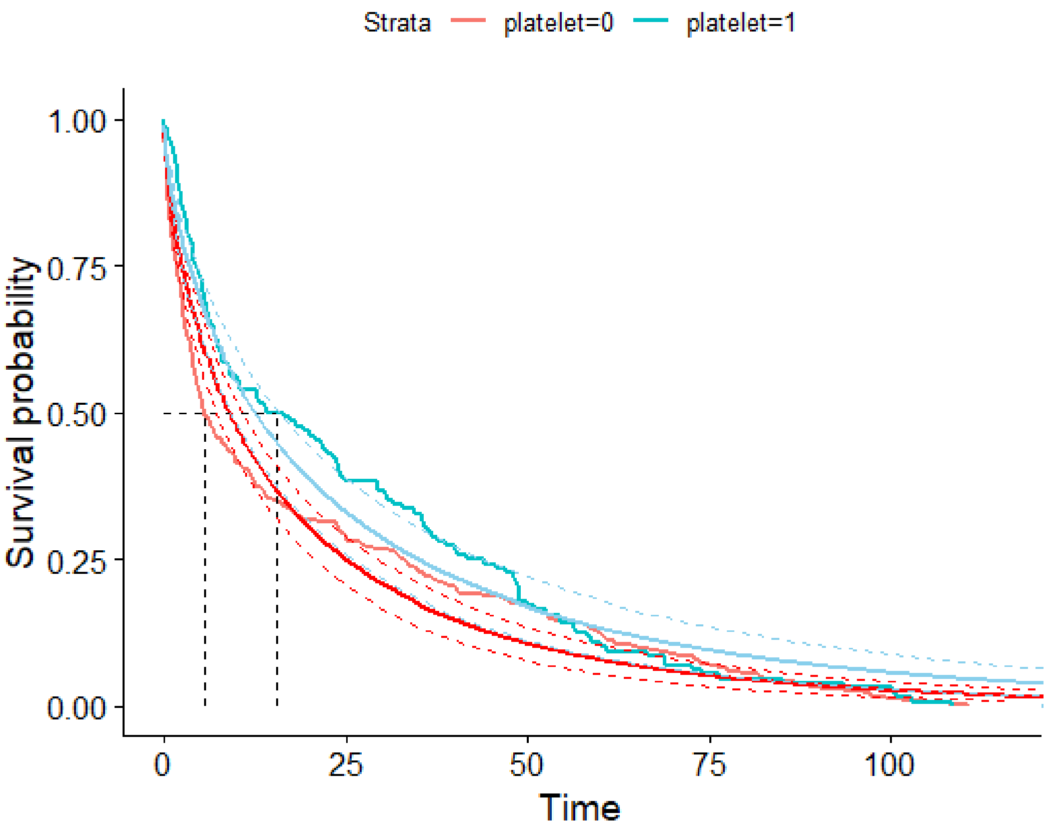
The design of the simulations is as follows. It is assumed that the hazards for event types 1 and 2 follow Weibull distribution. The idea to choose this distribution is based on BMT data (see Section 2.2 for the explanation) since the associated survival probability fits the data well (Figure 1.1). Further, one independent categorical covariate $X = \{x_1\} \sim Ber\,(p_1)$ is assumed for both event types. The CSH for event types 1 and 2 are,

$$\begin{cases} \lambda_1(t) & = a_1 b_1 t^{a_1-1} e^{x_1\beta_1} \\ \lambda_2(t) & = a_2 b_2 t^{a_2-1} e^{x_1\beta_2} \end{cases}$$

Then, the cumulative hazard functions for event types 1 and 2 are,

$$\begin{cases} \Lambda_1(t) & = b_1 t^{a_1} e^{x_1\beta_1} = \Lambda_{01}(t) e^{x_1\beta_1} \\ \Lambda_2(t) & = b_2 t^{a_2} e^{x_1\beta_2} = \Lambda_{02}(t) e^{x_1\beta_2} \end{cases}$$

FIGURE 1.1: Fitting a model using BMT data with Weibull time-to-event distribution (smoothed lines) and K-M (step-wise lines) to choose true parameter values of $(a_1, b_1, a_2, b_2, \beta_1, \beta_2)$ for the simulation study.



The overall survival function is, $S(t) = e^{-\Lambda(t)} = e^{-[\Lambda_1(t)+\Lambda_2(t)]} = e^{-[\Lambda_{01}(t)e^{x_1\beta_1}+\Lambda_{02}(t)e^{x_1\beta_2}]}$. Now, a system of equations are required to generate survival times. Here, it is assumed that $U = F(y) \sim Uniform[0,1]$. Thus, $S(y) = 1 - U \sim Uniform[0,1]$. Now, the following is known,

$$\begin{aligned}
F(t) &= 1 - S(t) \\
\Rightarrow 1 - F(t) &= S(t) \\
\Rightarrow 1 - U &= e^{-\Lambda(t)} = e^{-(\Lambda_1(t)+\Lambda_2(t))} \\
\Rightarrow \log(1-U) &= -(\Lambda_1(t) + \Lambda_2(t)) \\
\Rightarrow -\log(1-U) &= \Lambda_1(t) + \Lambda_2(t) \\
&= b_1 t^{a_1} e^{x_1\beta_1} + b_2 t^{a_2} e^{x_1\beta_2}
\end{aligned}$$

Finally, the following equation is obtained,

$$\Rightarrow [b_1 t^{a_1} e^{x_1 \beta_1} + b_2 t^{a_2} e^{x_1 \beta_2} + \log(1 - U)] = 0.$$

So, once the distribution of event-time T is found, then the latent failure time approach (David and Moeschberger, 1978; Prentice *et al.*, 1978) can be applied. Finally, a follow-up time is assumed and right-censoring times C is generated. Then, the observed time is computed as the minimum between the event and censoring times as well as the observed event indicator. In addition, the individual is considered as censored if the generated survival time is over the maximum follow-up time.

1.4.1 Choosing a predefined censoring rate

The value of the censoring parameter τ can be selected in such a way that the desired nominal censoring proportions (p) is drawn in the simulated data. Here, it is assumed that censoring time is $C \sim$ Uniform $(0, \tau)$ with censoring distribution as $g(C)$ and time to event as $T \sim$ Weibull $(\gamma, \tilde{\lambda})$ with density $f(T)$ and $T \perp C$ with $\tilde{\delta} = \mathbb{1}(T \geq C)$. To derive a value of τ in the censoring distribution $g(c \mid \tau)$ that yields censoring proportion p, a function $\eta(\tau)$ is set up using individual censoring probabilities, $\mathbb{P}\left(\tilde{\delta} = 1 \mid \tilde{\lambda}, \gamma, \tau\right)$, and density functions for covariates, $f_X(u)$:

$$\eta(\tau \mid p) = \mathbb{P}(\tilde{\delta} = 1 \mid \gamma, \tilde{\lambda}, \tau) - p = \int_0^{+\infty} \mathbb{P}(\tilde{\delta} = 1 \mid \tilde{\lambda}, \gamma, \tau) f_X(u) du - p$$

Then, for each possible combination of individual censoring probability and density function for the covariate, $\eta(\tau \mid p) = 0$ is solved for censoring parameter τ. However, the equation cannot be solved explicitly and thus requires numerical integration. In this context, Wan (2017) explained the techniques to identify the censoring proportions.

$$\begin{aligned}
\mathbb{P}(\tilde{\delta}=1 \mid \tilde{\lambda},\gamma,\tau) &= \mathbb{P}(C \leqslant T \leqslant \infty, 0 \leqslant C \leqslant \tau) \\
&= \int_0^{\tau}\int_c^{\infty} g(c \mid \tau) f(t \mid \tilde{\lambda},\gamma) dtdc \\
&= \int_0^{\tau}\int_c^{\infty} \frac{1}{\tau} f(t \mid \tilde{\lambda},\gamma) dtdc \\
&= \int_0^{\tau} \frac{1}{\tau} e^{-\Lambda_0(c|\tilde{\lambda},\gamma)} dc \\
&= \int_0^{\tau} \frac{1}{\tau} (e^{-\tilde{\lambda}c^{\gamma}}) dc \\
&= \int_0^{\tau} \frac{1}{\tau} e^{-\left(\tilde{\lambda}^{1/\gamma}c\right)^{\gamma}} dc \\
&= \frac{1}{\tau\lambda^{1/\gamma}} \int_0^{\tau\tilde{\lambda}^{1/\gamma}} \exp\left(-z^{\gamma}\right) dz, \quad z = c\tilde{\lambda}^{1/\gamma} \\
&= \frac{1}{\gamma\tau\tilde{\lambda}^{1/\gamma}} \int_0^{(\tau^{\gamma}\tilde{\lambda})} s^{\frac{1}{\gamma}-1}\exp(-s)ds \quad , \quad s = z^{\gamma} \\
&= \frac{1}{\gamma\tau\tilde{\lambda}^{1/\gamma}} \Gamma\left(\frac{1}{\gamma}, \tau^{\gamma}\tilde{\lambda}\right)
\end{aligned}$$

where $\Gamma(1/\gamma, u) = \int_0^u x^{1/\gamma-1}\exp(-x)/\Gamma(1/\gamma)dx$ is the incomplete gamma function. If it is assumed that $\tilde{\lambda} = e^{(\alpha_0+\beta_1 x_1)}$, where covariate $x_1 \sim Ber\left(p_1\right)$, the following equation is derived:

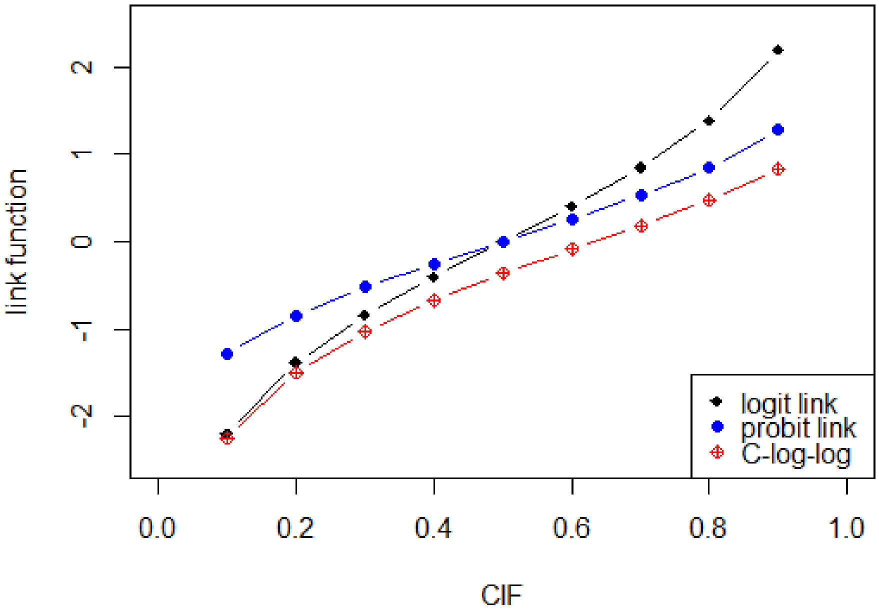
$$\eta(\tau \mid p) = \sum_{u_c=0}^{1} \frac{1}{\gamma\tau\left(e^{\alpha_0+\beta_1 * u_c}\right)^{1/\gamma}} \Gamma\left(\frac{1}{\gamma}, \tau^{\gamma} e^{(\alpha_0+\beta_1 * u_c)}\right) * p_1^{u_c}\left(1-p_1\right)^{1-u_c} - p$$

Here, it should be noted that to solve $\eta(\tau) = 0$ for a numerical solution of τ, one needs to use a statistical software.

1.4.2 Comparison of link functions

Figure 1.2 shows that when the CIF is low (< 0.3), the link functions between logistic and complementary log-log are very close, which indicates that with lower CIF, the coefficients in the SDH approach can be interpreted as odds ratios (Austin and Fine, 2017). Furthermore, when the CIF increases, the differences are higher among all the link functions. In the simulations, a complementary log-log link function was used.

FIGURE 1.2: Comparison of logit, probit, and complementary log-log link functions



1.4.3 Computation of the true CIF

In this section, the true CIF is computed using the CSH approach, and the same value are then used for the SDH, binomial, and pseudo-value regression approaches. Here, it is assumed that the time-to-event follows Weibull distribution with one independent categorical covariate $X = \{x_1\} \sim Ber\,(p_1)$ for both event types 1 and 2. Thus, for a fixed values of $(a_1, b_1, a_2, b_2, \beta_1, \beta_2)$, the assumed CSH rates for event types 1 and 2 in the CSH approach are

$$\begin{cases} \lambda_1(t \mid x) & = a_1 b_1 t^{a_1 - 1} e^{x_1 \beta_1} \\ \lambda_2(t \mid x) & = a_2 b_2 t^{a_2 - 1} e^{x_1 \beta_2}. \end{cases}$$

Moreover, the CIF for the main event of interest (event type 1) can be written as,

$$\begin{aligned} \mathrm{CIF}_1(t \mid x) &= \int_0^t \lambda_1(u \mid x) S(u \mid x) du \\ &= \int_0^t \left(a_1 b_1 u^{a_1-1} e^{x_1\beta_1}\right) \exp\left[-\left\{\Lambda_1(u \mid x) + \Lambda_2(u \mid x)\right\}\right] du \\ &= \int_0^t \left(a_1 b_1 u^{a_1-1} e^{x_1\beta_1}\right) \exp\left[-\left\{b_1 e^{x_1\beta_1} u^{a_1} + b_2 e^{x_1\beta_2} u^{a_2}\right\}\right] du \end{aligned}$$

Then, the equation for the experimental and control groups can be written as

$$\begin{cases} \mathrm{CIF}_1(t \mid x=1) & = \int_0^t \left(a_1 b_1 u^{a_1-1} e^{\beta_1}\right) \exp\left[-\left\{b_1 e^{\beta_1} u^{a_1} + b_2 e^{\beta_2} u^{a_2}\right\}\right] du \\ \mathrm{CIF}_1(t \mid x=0) & = \int_0^t \left(a_1 b_1 u^{a_1-1}\right) \exp\left[-\left\{b_1 u^{a_1} + b_2 u^{a_2}\right\}\right] du \end{cases}$$

Similar steps can be followed to compute the CIF for event type 2.

1.4.4 Computation of regression coefficients using the four regression approaches

While the interpretation of the regression coefficient is not straightforward, it provides a positive or negative sign, which gives information about increases in the covariate that are associated with an increase or decrease in the probability of the occurrence of events (CIF, taken in the c-log-log scale). For instance, assuming that when using the SDH approach for event type 1, β shows more negative values, this indicates that over time, a 1-unit increase in the covariate is associated with a reduction in the CIF.

The regression coefficient β varies based on the approach. In the CSH approach, the β provides the effect of covariates on the cause-specific hazards and it is computed as follows:

$$\log\left[\int_0^t \lambda_1(u \mid x) du\right] - \log\left[\int_0^t \lambda_{01}(u) du\right] = \beta x,$$

where, λ_{01} is the baseline hazard for cause 1. For the binomial or pseudo-value regression approaches, β is computed through the complementary log-log link function with the CIF. However, using the SDH approach, the computational technique of the regression coefficients is not straightforward. Here, the β^* has a direct relationship with the SDH rate λ^*.

β^* coefficient in the SDH approach

The CIF for event type 1 can be written as,

$$1\text{-CIF}_1(t \mid x) = S_1^*(t \mid x) = e^{-\Lambda_1^*(t|x)} = \exp(-\int_0^t \lambda_1^*(u \mid x) du).$$

Now, if the complementary log-log link function is assumed, the following can be derived:

$$\log\{-\log\{1\text{-CIF}_1(t \mid x)\}\} = \log\{-\log\{\exp(-\int_0^t \lambda_1^*(u \mid x) du)\}\} = \log \int_0^t \lambda_1^*(u \mid x) du.$$

The covariate effect on CIF with the Cox regression model for event type 1 can be computed as follows:

$$\begin{aligned}
\lambda_1^*(t \mid x) &= \lambda_{01}^*(t) e^{\beta^* x} \\
\Rightarrow \int_0^t \lambda_1^*(u \mid x) du &= \int_0^t \lambda_{01}^*(u) e^{\beta^* x} du = e^{\beta^* x} \int_0^t \lambda_{01}^*(u) du \\
\Rightarrow \log \left[\int_0^t \lambda_1^*(u \mid x) du \right] &= \beta^* x + \log \left[\int_0^t \lambda_{01}^*(u) du \right] \\
\Rightarrow \log \left[\int_0^t \lambda_1^*(u \mid x) du \right] - \log \left[\int_0^t \lambda_{01}^*(u) du \right] &= \beta^* x
\end{aligned}$$

To compute $\lambda_1^*(t \mid x)$ and $\lambda_{01}^*(t)$, the CSH and SDH relationship (Gray, 1988) was used:

$$\begin{cases}
\lambda_1^*(t \mid x = 1) & = \left[\frac{S(t|x=1)}{(1-CIF_1(t|x=1)} \right] \lambda_1(t \mid x = 1) \\
\lambda_1^*(t \mid x = 0) & = \left[\frac{S(t|x=0)}{(1-CIF_1(t|x=0)} \right] \lambda_1(t \mid x = 0)
\end{cases}$$

Here, we can find the values of $S(t), \text{CIF}_1(t), \lambda_1(t)$ can be found using the CSH approach along with the SDH rate:

$$\begin{cases}
\lambda_1^*(t) & = \left[\frac{\exp\left[-\left\{b_1 e^{\beta_1} t^{a_1} + b_2 e^{\beta_2} t^{a_2}\right\}\right]}{1-\int_0^t \left(a_1 b_1 u^{a_1-1} e^{\beta_1}\right) \exp\left[-\left\{b_1 e^{\beta_1} u^{a_1} + b_2 e^{\beta_2} u^{a_2}\right\}\right] du} \right] (a_1 b_1 t^{a_1-1} e^{\beta_1}) \\
\lambda_{01}^*(t) & = \left[\frac{\exp[-\{b_1 t^{a_1} + b_2 t^{a_2}\}]}{1-\int_0^t \left(a_1 b_1 u^{a_1-1} e^{\beta_1}\right) \exp[-\{b_1 u^{a_1} + b_2 u^{a_2}\}] du} \right] (a_1 b_1 t^{a_1-1}).
\end{cases}$$

β coefficient in the binomial or pseudo-value regression approaches

In the binomial and pesudo-value approaches, the same link function clog-log(x) is chosen to model the CIF:

$$\Phi(CIF_1(t \mid x) = \log\{-\log(1\text{-CIF}_1(t \mid x))\} = \alpha_{01}(t) + \beta(t) x \tag{1.3}$$

where, $\alpha_{01}(t)$ and $\beta(t)$ are the possibly time-varying regression coefficients. Equation (1.3) corresponds to a Cox regression model type for the fixed value of t.

In a simpler case of a time-constant regression coefficient β, equation (1.3) can be derived for the experimental and control groups, and then the difference can be taken as follows:

$$\begin{aligned}\log\{-\log(\text{1-CIF}_1(t \mid x=1))\} - \log\{-\log(\text{1-CIF}_1(t \mid x=0))\} &= \beta \\ \Rightarrow \log\left[\frac{-\log(\text{1-CIF}_1(t \mid x=1))}{-\log(\text{1-CIF}_1(t \mid x=0))}\right] &= \beta\end{aligned}$$

$$\Rightarrow \log\left[\frac{-\log(\text{1-} \int_0^{\text{t}} \left(\text{a}_1\text{b}_1\text{u}^{\text{a}_1\text{-1}}\text{e}^{\beta_1}\right) \exp\left[\text{-} \left\{\text{b}_1\text{e}^{\beta_1}\text{u}^{\text{a}_1} + \text{b}_2\text{e}^{\beta_2}\text{u}^{\text{a}_2}\right\}\right] \text{du})}{-\log(\text{1-} \int_0^{\text{t}} (\text{a}_1\text{b}_1\text{u}^{\text{a}_1\text{-1}}) \exp[\text{-} \{\text{b}_1\text{u}^{\text{a}_1} + \text{b}_2\text{u}^{\text{a}_2}\}] \text{du})}\right] = \beta \tag{1.4}$$

Here, $\log(-\log(1 - \text{CIF}_1(t \mid x=0))) = \alpha_{01}(t)$

1.4.5 Variance and confidence interval calculations of the CIF in the pseudo-value regression approach

In the simulation, four discrete time points, i.e., $j = \{10, 50, 90, 130\}$, were assumed along with a time-constant regression coefficient β. Then, at each time point, the CIF values were computed. Here, it is noted that the R package `pseudo` did not provide the observed CIF values, but gave the estimated regression parameter $\hat{\beta}$ and $s.e.(\hat{\beta})$. Thus, to compute the variance of the estimated CIF, it is further needed to compute $v(\hat{\alpha}_{01(j)} + \hat{\beta})$ as follows:

$$\begin{cases}\hat{\alpha}_{01(j)} \mathrel{\dot{\sim}} N(\alpha_{01(j)}, v(\hat{\alpha}_{01(j)})) \\ \hat{\beta} \mathrel{\dot{\sim}} N(\beta, v(\hat{\beta}))\end{cases}$$

If it is assumed that $\hat{\alpha}_{01(j)} + \hat{\beta} = \hat{\hat{\alpha}}_{01(j)}$, then the following can be stated,

$$\hat{\hat{\alpha}}_{01(j)} \mathrel{\dot{\sim}} N(\alpha_{01(j)} + \beta, v(\hat{\alpha}_{01(j)}) + v(\hat{\beta}) + 2\text{cov}(\hat{\alpha}_{01(j)}, \hat{\beta}))$$

where, $v(\hat{\alpha}_{01(j)}) + v(\hat{\beta}) + 2\text{cov}(\hat{\alpha}_{01(j)}, \hat{\beta})$ can be estimated as, $\hat{v}(\hat{\hat{\alpha}}_{01(j)}) = \hat{v}(\hat{\alpha}_{01(j)}) + \hat{v}(\hat{\beta}) + 2\widehat{\text{cov}}(\hat{\alpha}_{01(j)}, \hat{\beta})$, and the s.e. is, $s.e.(\hat{\hat{\alpha}}_{01(j)}) = \sqrt{\hat{v}(\hat{\hat{\alpha}}_{01(j)})}$.

Variance computation of CIF: First, it is assumed that the relation between the regression coefficients and CIF are established by the complementary-log-log link function.

When the event type 1 is studied, and $x = 0$:

$$\begin{aligned}
\hat{\alpha}_{01(j)} &= g\left(\widehat{\text{CIF}}_{1j}^{(0)}\right) \\
\Rightarrow \widehat{\text{CIF}}_{1j}^{(0)} &= 1 - e^{-e^{\hat{\alpha}_{01(j)}}}, \quad j = 1, 2, 3, 4
\end{aligned}$$

When the event type 1 is studied, and $x = 1$:

$$\begin{aligned}
\widehat{\text{CIF}}_{1j}^{(1)} &= 1 - e^{-e^{\alpha_{0\hat{1}(j)}+\hat{\beta}}} \\
&= 1 - e^{-e^{\hat{\tilde{\alpha}}_{01(j)}}} \quad \text{where,} \quad \hat{\tilde{\alpha}}_{01(j)} = \hat{\alpha}_{01(j)} + \hat{\beta}.
\end{aligned}$$

Then the variance for CIF can be found using the delta method as follows:

$$\begin{aligned}
v\left(\widehat{\text{CIF}}_{1j}^{(0)}\right) &= v\left(1 - e^{-e^{\widehat{\alpha}_{01(j)}}}\right) \\
&= \hat{v}\left(\widehat{\alpha}_{01(j)}\right)\left[\frac{\partial}{\partial \widehat{\alpha}_{01(j)}}\left(1 - e^{-e^{\widehat{\alpha}_{01(j)}}}\right)\right]^2 \\
&= \hat{v}\left(\widehat{\alpha}_{01(j)}\right)\left[-\frac{\partial}{\partial \widehat{\alpha}_{01(j)}}\left(e^{-e^{\widehat{\alpha}_{01(j)}}}\right)\right]^2 \\
&= \hat{v}\left(\widehat{\alpha}_{01(j)}\right)\left[e^{-e^{\widehat{\alpha}_{01(j)}}} \cdot e^{\widehat{\alpha}_{01(j)}}\right]^2 \\
&= \hat{v}\left(\widehat{\alpha}_{01(j)}\right)\left[e^{-e^{\widehat{\alpha}_{01(j)}}+\widehat{\alpha}_{01(j)}}\right]^2
\end{aligned}$$

Similarly,

$$v\left(\widehat{\text{CIF}}_{1j}^{(1)}\right) = \hat{v}\left(\widehat{\tilde{\alpha}}_{01(j)}\right)\left[e^{-e^{\widehat{\tilde{\alpha}}_{01(j)}}+\widehat{\tilde{\alpha}}_{01(j)}}\right]^2 \quad \text{where,} \quad \widehat{\tilde{\alpha}}_{01(j)} = \hat{\alpha}_{01(j)} + \hat{\beta}.$$

Next, the confidence interval of the CIF can be estimated in two different ways:

(1) The delta method:

$$\widehat{\text{CIF}}_{1j}^{(0)} \pm z_\alpha \times \sqrt{v\left(\widehat{\text{CIF}}_{1j}^{(0)}\right)} \quad \text{and} \quad \widehat{\text{CIF}}_{1j}^{(1)} \pm z_\alpha \times \sqrt{v\left(\widehat{\text{CIF}}_{1j}^{(1)}\right)}$$

(2) Transforming β: An alternative way is to compute the CI for the $\alpha_{01(j)}, \widetilde{\alpha}_{01(j)}$ parameters and then transform them using the link function in relation with the CIF as follows:

$$
\begin{aligned}
P\left(\hat{\alpha}_{01(j)} - Z_\alpha \cdot \text{se}\left(\hat{\alpha}^{(0)}_{01(j)}\right) \le \alpha_{01(j)} \le \hat{\alpha}_{01(j)} + Z_\alpha \cdot \text{se}\left(\hat{\alpha}^{(0)}_{01(j)}\right)\right) &= 0.95 \\
\Rightarrow P\left(e^{\hat{\alpha}_{01(j)} - Z_\alpha \cdot \text{se}\left(\hat{\alpha}^{(0)}_{01(j)}\right)} \le e^{\alpha_{01(j)}} \le e^{\hat{\alpha}_{01(j)} + Z_\alpha \cdot \text{se}\left(\hat{\alpha}^{(0)}_{01(j)}\right)}\right) &= 0.95 \\
\Rightarrow P\left(-e^{\hat{\alpha}_{01(j)} + Z_\alpha \cdot \text{se}\left(\hat{\alpha}^{(0)}_{01(j)}\right)} \le -e^{\alpha_{01(j)}} \le -e^{\hat{\alpha}_{01(j)} - Z_\alpha \cdot \text{se}\left(\hat{\alpha}^{(0)}_{01(j)}\right)}\right) &= 0.95 \\
\Rightarrow P\left(e^{-e^{\hat{\alpha}_{01(j)} + Z_\alpha \cdot \text{se}\left(\hat{\alpha}^{(0)}_{01(j)}\right)}} \le e^{-e^{\alpha_{01(j)}}} \le e^{-e^{\hat{\alpha}_{01(j)} - Z_\alpha \cdot \text{se}\left(\hat{\alpha}^{(0)}_{01(j)}\right)}}\right) &= 0.95 \\
\Rightarrow P\left(-e^{-e^{\hat{\alpha}_{01(j)} - Z_\alpha \cdot \text{se}\left(\hat{\alpha}^{(0)}_{01(j)}\right)}} \le -e^{-e^{\alpha_{01(j)}}} \le -e^{-e^{\hat{\alpha}_{01(j)} + Z_\alpha \cdot \text{se}\left(\hat{\alpha}^{(0)}_{01(j)}\right)}}\right) &= 0.95 \\
\Rightarrow P\left(1 - e^{-e^{\hat{\alpha}_{01(j)} - Z_\alpha \cdot \text{se}\left(\hat{\alpha}^{(0)}_{01(j)}\right)}} \le 1 - e^{-e^{\alpha_{01(j)}}} \le 1 - e^{-e^{\hat{\alpha}_{01(j)} + Z_\alpha \cdot \text{se}\left(\hat{\alpha}^{(0)}_{01(j)}\right)}}\right) &= 0.95 \\
\Rightarrow P\left(1 - e^{-e^{\hat{\alpha}_{01(j)} - Z_\alpha \cdot \text{se}\left(\hat{\alpha}^{(0)}_{01(j)}\right)}} \le \text{CIF}^{(0)}_{1j} \le 1 - e^{-e^{\hat{\alpha}_{01(j)} + Z_\alpha \cdot \text{se}\left(\hat{\alpha}^{(0)}_{01(j)}\right)}}\right) &= 0.95.
\end{aligned}
$$

Similarly, the C.I. for $\text{CIF}^{(1)}_{1j}$ can be calculated as follows:

$$
P\left(1 - e^{-e^{\hat{\tilde{\alpha}}_{01(j)} - Z_\alpha \cdot \text{se}\left(\hat{\tilde{\alpha}}^{(0)}_{01(j)}\right)}} \le \text{CIF}^{(1)}_{1j} \le 1 - e^{-e^{\hat{\tilde{\alpha}}_{01(j)} + Z_\alpha \cdot \text{se}\left(\hat{\tilde{\alpha}}^{(0)}_{01(j)}\right)}}\right) = 0.95
$$

where,

$$
\hat{\tilde{\alpha}}_{01(j)} = \hat{\alpha}_{01(j)} + \hat{\beta}
$$

The first approach, where the CI is computed using the s.e. from the delta method, provides different results as compared to the second approach, where the CI is directly transformed. In the simulation, the second approach was used because an investigator never gets values above 1 or lower than 0 for the CIF (as this is a probability). On the contrary, in the first approach, one can also get a lower bound smaller than 0 or an upper bound greater than 1.

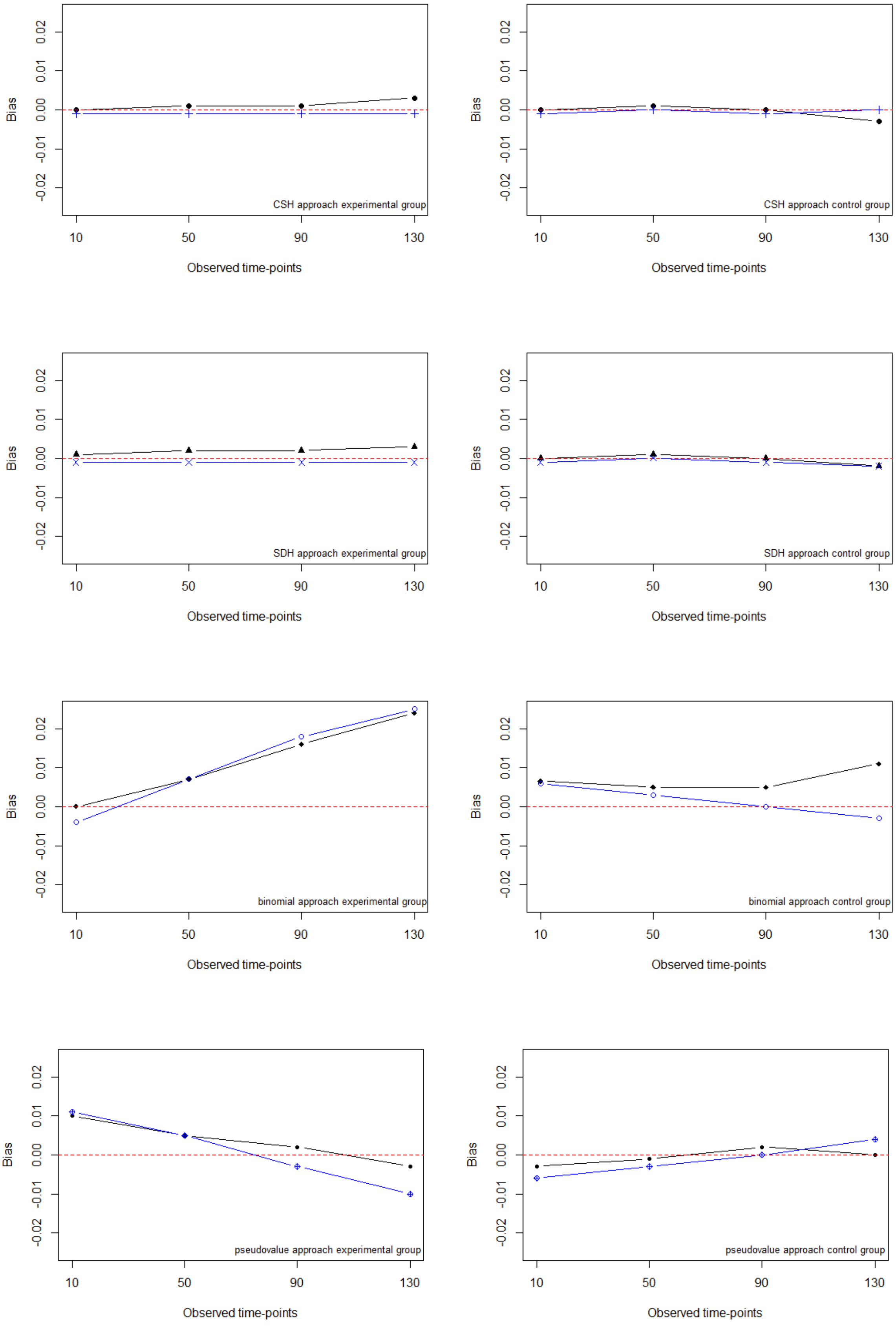
1.4.6 Simulation results among the four regression approaches

A simulation study with $3,000$ replications is conducted for generating datasets of size $n = 500$. Here, it is assumed that the following are known in advance: shape parameters $a_1 = 0.45, a_2 = 0.5$; scale parameters $b_1 = 0.15, b_2 = 0.06$; regression coefficient $\beta_1 = -0.6, \beta_2 = -0.17$; censoring parameter $\tau = 180$ with censoring proportion 0.01. Further, one categorical covariate was assumed for both event types as, x $\sim$ Ber (0.32).

The analysis has been implemented in the freeware statistical software R (http://cran.r-project.org) using the libraries `Survival` (Lumley and Therneau, 2003) and `prodlim` (Gerds, 2019) for the CSH approach, `cmprsk` (Gray *et al.*, 2004) for the SDH approach, `timereg` (Scheike and Martinussen, 2006) for the binomial regression approach, and `pseudo` (Pohar Perme and Gerster, 2017) for the pseudo-value approach. Then, the bias and other specific characteristics among all the regression approaches were computed. Under binomial approach, for computing coverage probabilities (CP), both time-fixed and time-varying coefficients were applied, whereas in all the other computations, only the time-fixed coefficient was used.

Figure 1.3 (and Table 1.1) shows the bias between the true and simulated CIF values for event types 1 and 2 for both treatment groups in the CSH, SDH, binomial and pseudo-value regression approaches for 3000 simulations at time, $t = \{10, 50, 90, 130\}$. Overall, it was found that the bias was lower in the CSH and SDH approaches for both causes. Furthermore, at the beginning of the study (time point 10), the biases were very close among the CSH, SDH, and pseudo-value approaches for both the experimental and control groups. However, the biases were higher in the binomial approach as compared to all other approaches. Specifically, the maximum bias for the main event of interest for the experimental group was 0.024 for the binomial approach at time 130. For this approach, the bias was always higher for the main event of interest at time points 90 and 130. Meanwhile in the pseudo-value approach, the bias was higher at the beginning of the study for event type 1 and there was a substantial reduction over time for both groups. On the contrary, for event type 2, the bias was higher at the beginning and ending of the study (at time 10 and 130) for both groups.

FIGURE 1.3: Comparison of the biases between the true and simulated CIF values among all the regression approaches for event types 1 (Black lines) and 2 (Blue lines). All the figures on the left show the experimental groups while those on the right show the control groups



Figures 1.4 and 1.5 show the ratios of estimated (estimated mean of all $3,000$ replications standard errors at time t) to empirical (standard deviation of all $3,000$ replications of the estimated CIF at time t) standard errors of CIF$_1$ over 3000 simulations with $t = \{10, 50, 90, 130\}$ for the experimental and control groups in the CSH, SDH, binomial, and pseudo-value regression approaches. Most of the ratios were near 1, indicating no substantial differences between the estimated and empirical standard errors of the CIF for event type 1. Thus, the estimators of the variance of CIF were in good agreement with the simulated sample variance in all the approaches. However, the binomial and pseudo-value approaches showed a slight underestimation of the variance at all time points in the experimental groups.

Figure 1.4: Ratio of estimated to empirical standard errors for the CIF_1 for experimental and control groups in the CSH, SDH, binomial, and pseudo-value regression approaches

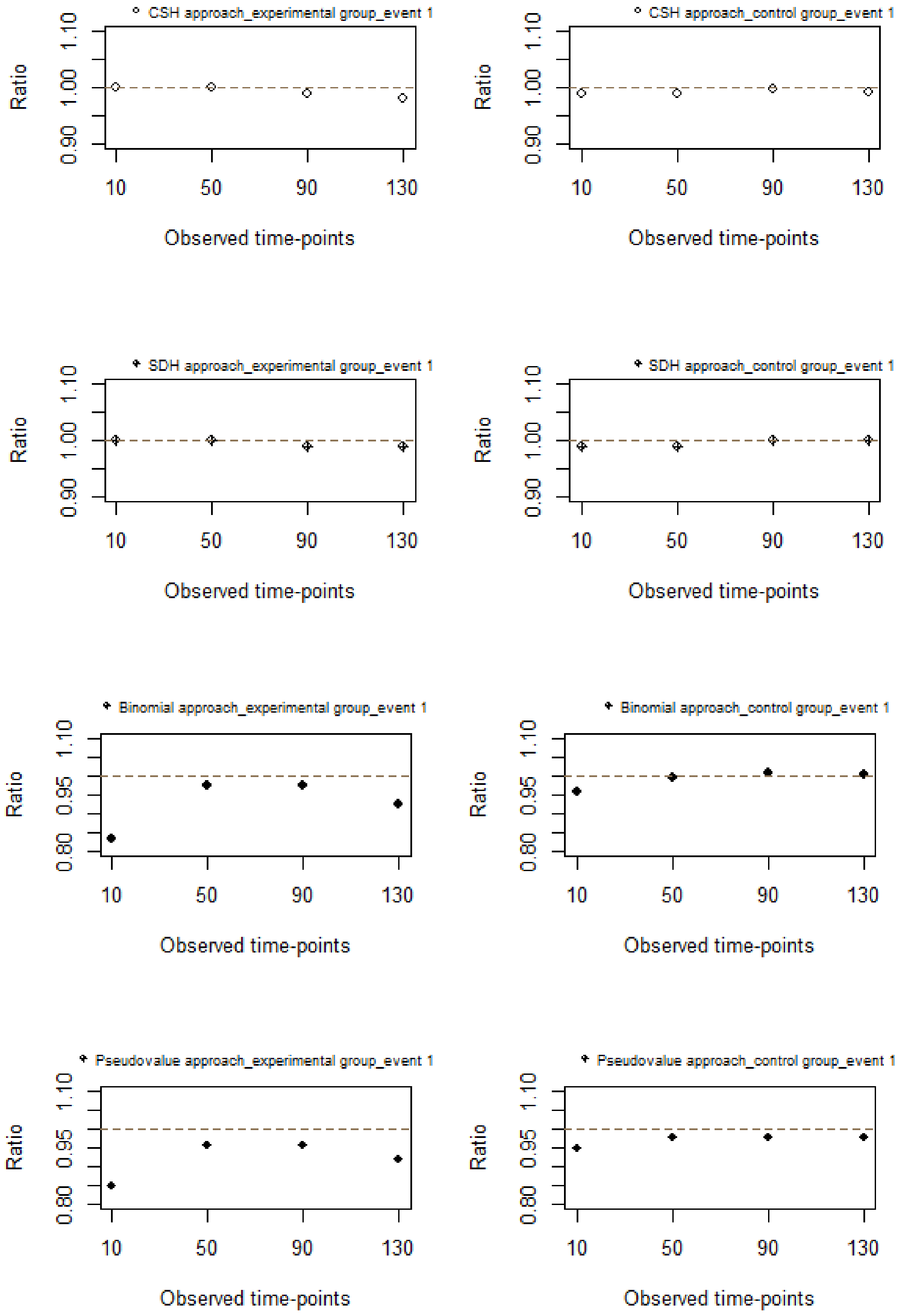
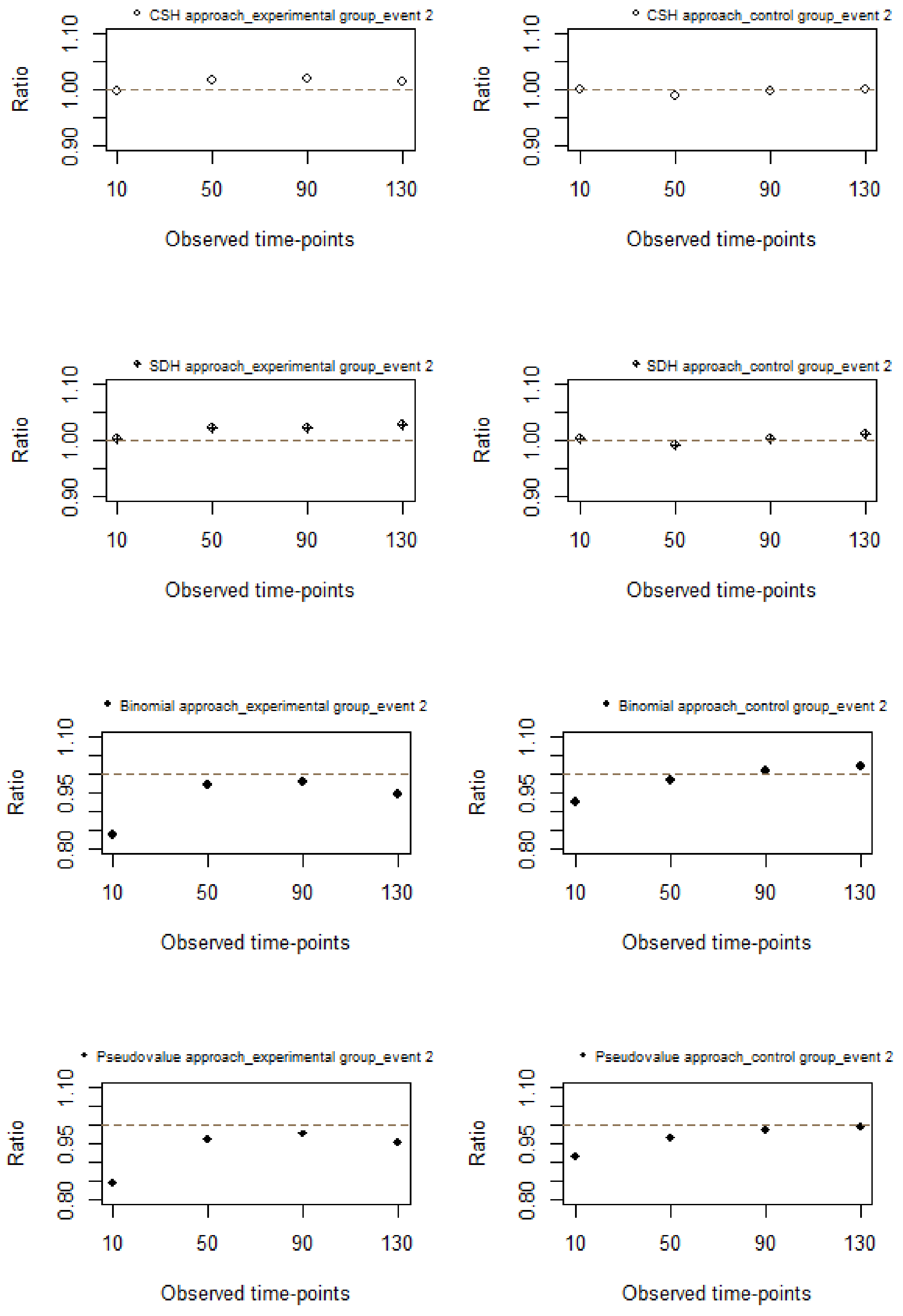


FIGURE 1.5: Ratio of estimated to empirical standard errors for CIF_2 for experimental and control groups in the CSH, SDH, binomial, and pseudo-value regression approaches



Figures 1.6, 1.7, 1.8, 1.9, and 1.10 depict the coverage probabilities (CP) of CIF_1 and CIF_2 for the experimental and control groups in the CSH, SDH, binomial, and pseudo-value regression approaches. In the binomial approach, when the time-fixed coefficient was assumed, the CP performed worse in the experimental group for event type 1 at later time points and was the worst for event type 2. However, when the time-varying coefficient was assumed, the CP improved considerably for the experimental group as compared to the time-fixed coefficient scenario. Moreover, under the pseudo-value approach (Figure 1.10), it was found that the CP performed well at each of the time points for event type 1 for experimental group whereas for event type 2, it was below 95% level at later time points. On the contrary, the CP performance was worst for event type 2 for control group.

Under the direct binomial regression approach, the time-varying coefficients are used to improve the CP and thus it is important to focus on the general guidelines related to that approach. The direct binomial regression approach is based on the inverse probability of censoring weighting (IPCW) technique, which can be used to improve efficiency. However, the weights need to be estimated without bias so that the estimates of the CIF are also unbiased. It was found that the censoring distribution depends significantly on the covariates and this dependence is partially captured by a Cox's regression model for the censoring times. Therefore, the option cens.model="cox" was added in the function call to specify Cox models for the IPCW.

Figure 1.6: Coverage probability of CIF for event types 1 and 2 over 3,000 simulations with $t = \{10, 50, 90, 130\}$ for the experimental and control groups in the CSH approach

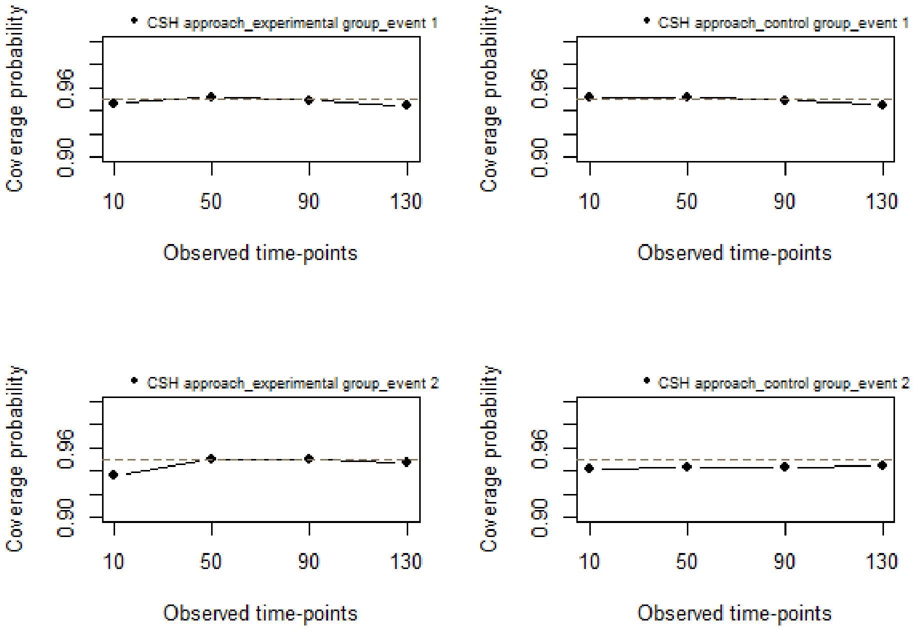


Figure 1.7: Coverage probability of CIF for event types 1 and 2 for the experimental and control groups in the SDH approach

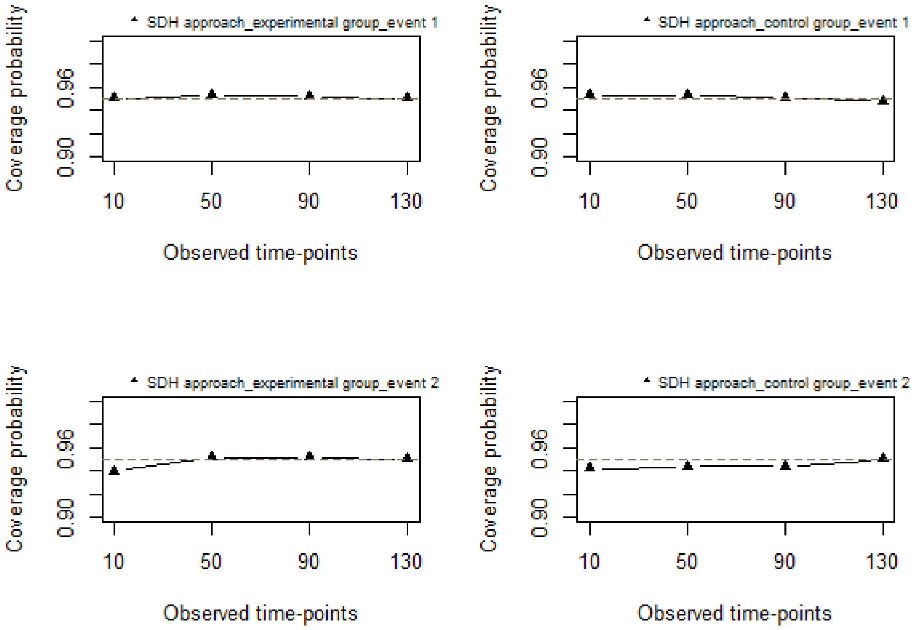


FIGURE 1.8: Coverage probability of CIF for event types 1 and 2 for the experimental and control groups in the binomial approach with time-fixed coefficient

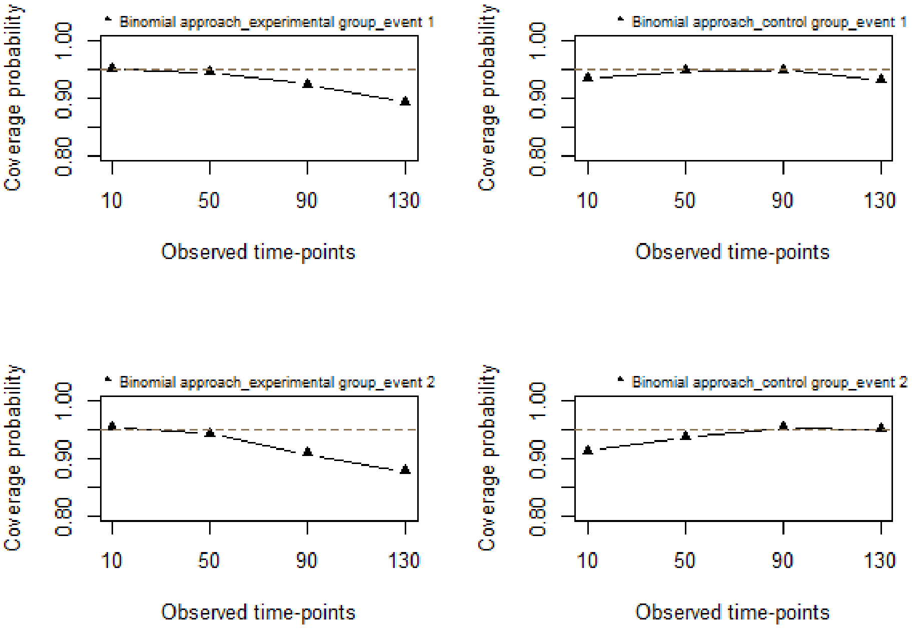


FIGURE 1.9: Coverage probability of CIF for event types 1 and 2 for the experimental and control groups in the binomial approach with time-varying coefficient

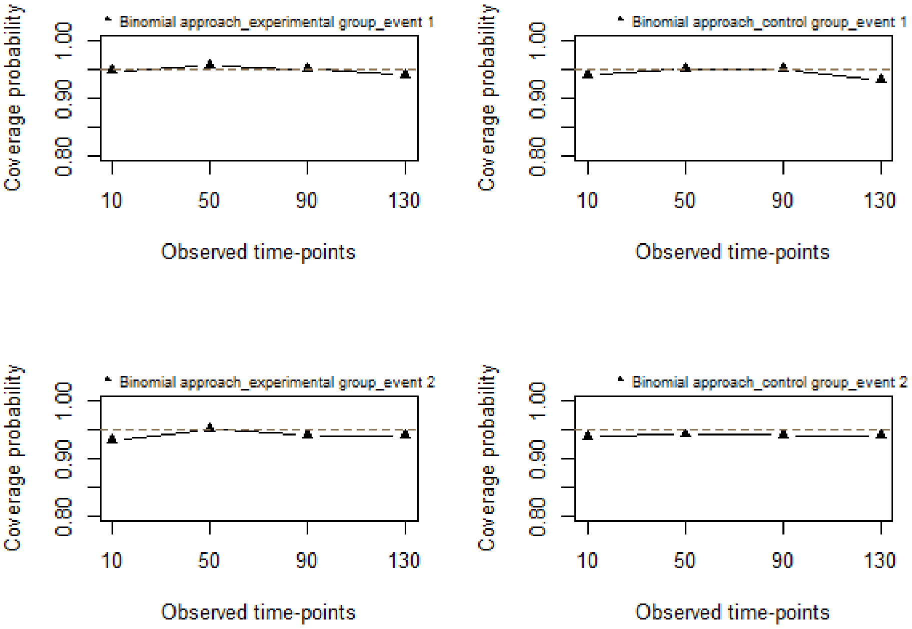
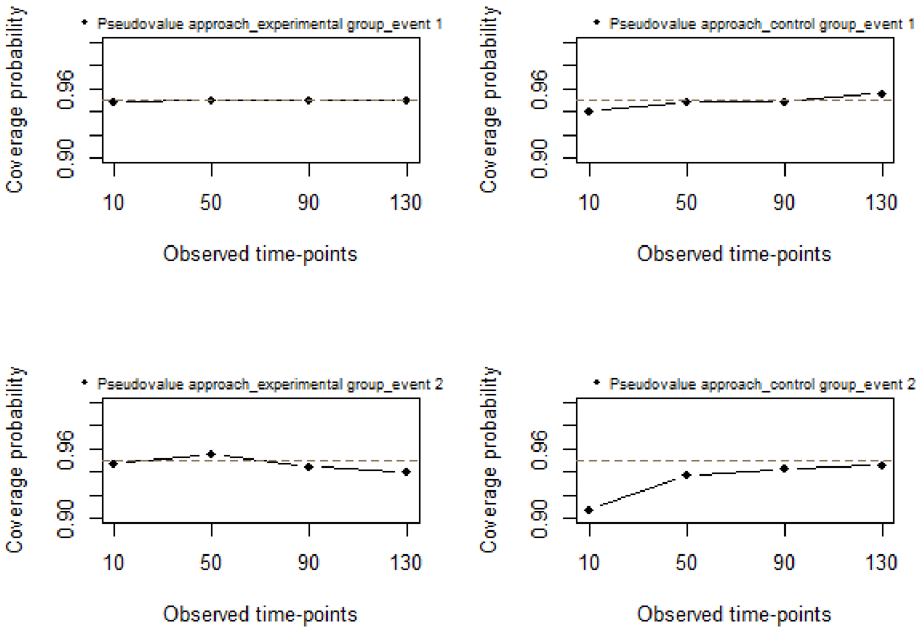


FIGURE 1.10: Coverage probability of CIF for event types 1 and 2 for the experimental and control groups in the pseudo-value approach



To check the performance of the estimated variance when using the binomial and pesudo-value approaches in comparison to the CSH approach, the relative efficiencies (RE) of CIF were computed over 3000 replications as

$$\begin{cases} RE_{\text{binomial vs. CSH}} = \dfrac{\text{empirical variance of CIF for the binomial approach}}{\text{empirical variance of CIF for the CSH approach}} \\ RE_{\text{pseudo-value vs. CSH}} = \dfrac{\text{empirical variance of CIF for the pseudo-value approach}}{\text{empirical variance of CIF for the CSH approach}} \end{cases}$$

Figure 1.11 depicts the RE performances. When using the binomial approach for both event types, the experimental group showed efficiency gain as compared to the CSH approach, while for the control group, there was a loss in efficiency gain after time point 90. Meanwhile, when using the pseudo-value approach, there was a gain in the effeciency as compared to the CSH approach for both events and in both groups.

FIGURE 1.11: Relative efficiency of estimators when using the binomial, and pseudo-value as compared to the CSH over 3000 simulations with $t = \{10, 50, 90, 130\}$ for the experimental and control groups

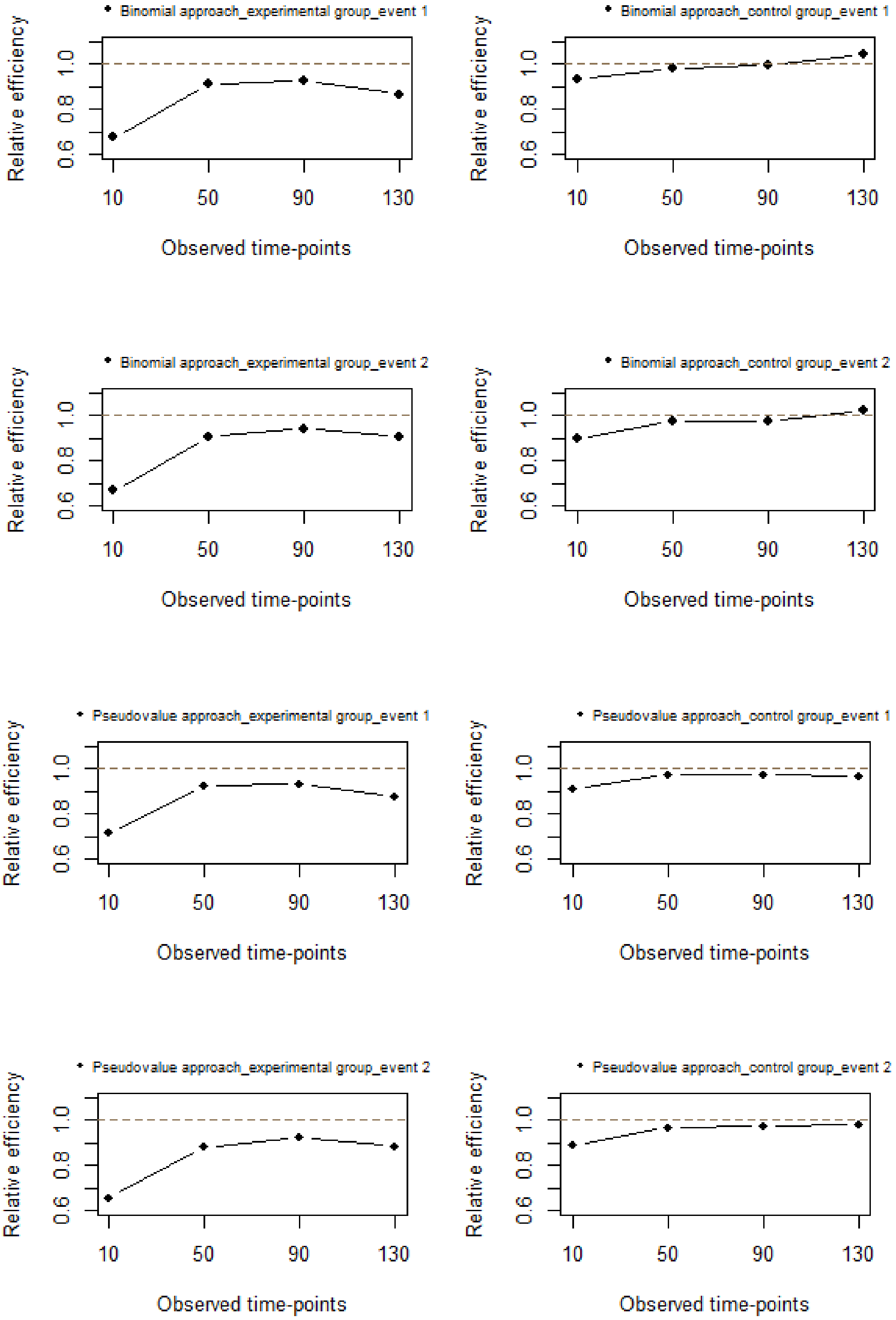


Table 1.1: Comparison of the true and simulated CIF values among all the regression approaches for event types 1 and 2. For the binomial approach, only the time-fixed coefficient scenario is reported. E indicates experimental group and C indicates control group.

Time points			10	50	90	130			10	50	90	130
	x			Cause 1			x			Cause 2		
CSH	E	True	0.193	0.327	0.383	0.416	E	True	0.132	0.239	0.286	0.315
		Sim	0.193	0.328	0.384	0.419		Sim	0.131	0.238	0.285	0.314
		Bias	**0.000**	**0.001**	**0.001**	**0.003**		**Bias**	**0.001**	**0.001**	**0.001**	**0.001**
	C	True	0.318	0.494	0.554	0.589	C	True	0.140	0.232	0.265	0.282
		Sim	0.318	0.495	0.554	0.586		Sim	0.139	0.232	0.264	0.282
		Bias	**0.000**	**0.001**	**0.000**	**0.003**		**Bias**	**0.001**	**0.000**	**0.001**	**0.000**
SDH	E	True	0.193	0.327	0.383	0.416	E	True	0.132	0.239	0.286	0.315
		Sim	0.194	0.329	0.385	0.419		Sim	0.131	0.238	0.285	0.314
		Bias	**0.001**	**0.002**	**0.002**	**0.003**		**Bias**	**0.001**	**0.001**	**0.001**	**0.001**
	C	True	0.318	0.494	0.554	0.589	C	True	0.140	0.232	0.265	0.282
		Sim	0.318	0.495	0.554	0.587		Sim	0.139	0.232	0.264	0.280
		Bias	**0.000**	**0.001**	**0.000**	**0.002**		**Bias**	**0.001**	**0.000**	**0.001**	**0.002**
Binomial	E	True	0.193	0.327	0.382	0.416	E	True	0.132	0.239	0.286	0.315
		Sim	0.193	0.320	0.367	0.391		Sim	0.136	0.232	0.268	0.290
		Bias	**0.000**	**0.007**	**0.015**	**0.025**		**Bias**	**-0.004**	**0.007**	**0.018**	**0.025**
	C	True	0.318	0.494	0.554	0.589	C	True	0.140	0.232	0.265	0.282
		Sim	0.312	0.489	0.549	0.578		Sim	0.134	0.229	0.265	0.285
		Bias	**0.006**	**0.005**	**0.005**	**0.011**		**Bias**	**0.006**	**0.003**	**0.000**	**-0.003**
Pseudovalue	E	True	0.193	0.327	0.382	0.416	E	True	0.132	0.239	0.286	0.315
		Sim	0.203	0.335	0.384	0.413		Sim	0.143	0.244	0.283	0.305
		Bias	**0.010**	**0.005**	**0.002**	**0.003**		**Bias**	**0.011**	**0.005**	**0.003**	**0.010**
	C	True	0.318	0.494	0.554	0.589	C	True	0.140	0.232	0.265	0.282
		Sim	0.315	0.493	0.555	0.589		Sim	0.134	0.229	0.265	0.286
		Bias	**0.003**	**0.001**	**0.002**	**0.000**		**Bias**	**0.006**	**0.003**	**0.000**	**0.004**

In conclusion, it can be said that the bias was lower when using the CSH and SDH approaches for both causes. The bias was higher for the binomial approach than all the other approaches, and the maximum bias for the main event of interest for the experimental group was 0.013 at time 130. Meanwhile, when using the pseudo-value approach, the bias was higher at the beginning of the study, and there was a substantial reduction over time. Moreover, the ratios of estimated to empirical standard errors of the

CIF for both event types and treatment groups for all approaches were near 1, indicating that there were no substantial differences between the observed and empirical standard errors. However, when using the pseudo-value, there was a slight underestimation of the variance over time. Further, with the binomial approach, the CP performed the worst for event type 2 when the time-fixed coefficient was assumed. However, when the time-varying coefficient was used, the CP improved considerably, but it decreased slightly at time point 10 for both events as compared to the time-fixed coefficient scenario. Furthermore, the efficiency measurements between the binomial and CSH as well as the pseudo-value and CSH approaches were studied. With the pseudo-value approach, there was a gain in effeciency in all the scenarios. However, with the binomial approach, there was a loss in efficiency for control group. In general, for a study on clinical trials, it is important to investigate the efficiency measurements as the objective here is to establish the effect of an intervention. On the contrary, in an observational study, heterogeneity exists as the inferences are based on individual preferences, thus making other measurement (for instance, bias and CP) more important.

Chapter 2

Application to real data

2.1 Introduction

The methods described in Chapter 1 will be illustrated using Bone Marrow Transplant (BMT) and Covid-19 data. Specifically, BMT data are described for the purpose of providing practical guidelines to a new user using R in Section 2.2. The Covid-19 data are used as an example of real data as described in Section 2.3. Further, BMT data are available under the `timereg` package in R and also in the International Blood and Marrow Transplant Research (CIBMTR) study (Sierra *et al.*, 2002). Covid-19 data are obtained from the Ministry of Health of Brazil database for all COVID-19 patients from January 01, 2020 to April 30, 2021 period.

Section 2.2.1 discusses the non-parametric techniques (without covariates) for the cause-specific hazard (CSH) and sub-distribution hazard (SDH) approaches, while Section 2.2.2 describes regression approaches. In Sections 2.3.1 and 2.3.2, we explored Covid data using non-parametric and regression approaches, respectively to estimate cumulative risk of dying from Covid and the effect of covariates on such risk.

2.2 Practical guidelines using BMT data in R

The study has two competing risks (CR): treatment-related mortality (TRM), defined as death in complete remission, and relapse, defined as the recurrence of myelodysplasia (MDS). Complete information from 408 patients was included in this example (161 patients died in complete remission, and 87 patients relapsed). The CIBMTR study indicated that the CIFs of TRM were different for patients with low and high platelet counts ($< 100 \times 10^9/\text{l}$ [n=280] versus $\geq 100 \times 10^9/\text{l}$ [n=128]). The covariates of the study were as follows: age (continuous variable, standardized and centered at a mean of 35 years old, ranging from 2 to 64 years); platelet counts (1 for high platelet counts and 0 for low platelet counts); and graft versus host disease (GVHD) prophylaxis (1 for T-cell depletion BMT and 0 for non-T-cell depletion BMT) for TRM and relapse.

The removal of T-cells from the donor graft (T-cell depletion) offers the possibility of preventing GvHD and thereby reducing transplant-related morbidity and mortality (Daniele *et al.*, 2012). Factors predictive of poor survival following chronic GVHD diagnosis includes low platelet count and a history of acute liver GVHD (Pavletic *et al.*, 2005).

In this thesis, the main interest is to compute the CIF. To this end, a conventional technique such as the K-M method, will first be applied. However, it should be noted that the KM method may give biased estimates because it considers the CR events as censored. Thus, the CIF will be computed non-parametrically by using the *cuminc()* function in the `cmprsk` package. This function also allows for group comparison and visualization of the estimated CIF. Subsequently, the results will be compared with the classical Aalen-Johansen estimator using the `prodlim` package. The model proposed by Fine and Gray (1999) can be fitted with the *crr()* function using the `cmprsk` package. Time-varying covariates are allowed in the *crr()* function as specified by *cov*2 and *tf* arguments. Then, predictions and visualizations of the CIF for patients can be undertaken with the given covariate values for the *crr()* object. An alternative method to fit the CR model is the `riskRegression` package with different link functions between covariates and outcomes.

Meanwhile, Scheike and Zhang (2011) developed the *comp.risk()* function in the binomial regression approach, which is available in the `timereg` package. It implements two classes of flexible models: proportional and additive models. These are special sub-models of the proportional regression model proposed by Fine and Gray (1999), special additive model by Lin and Ying (1994), and Aalen's full additive regression model (Zhang, 2017). In addition, a useful goodness-of-fit test was proposed by Scheike and Zhang (2011) to identify whether the time-varying effect is present for a specific covariate. In medical studies, it is often useful to estimate the predicted CIF for a given set of values of covariates. The *predict()* function in the `timereg` package computes the predicted CIF and an estimate of its variance at each fixed time point, after which it constructs (1-α)100% simultaneous confidence bands over a given time interval.

Finally, the pseudo-value approach can be analyzed using the `pseudo` package. Here, the pseudo-values will first be computed, after which the parameters can then be estimated using generalized estimating equations (GEE) with the function *geese* (Klein *et al.*, 2008).

2.2.1 Non-parametric techniques

Under this technique, the CIF is computed using K-M as well as with the CSH and SDH approaches without considering any covariates.

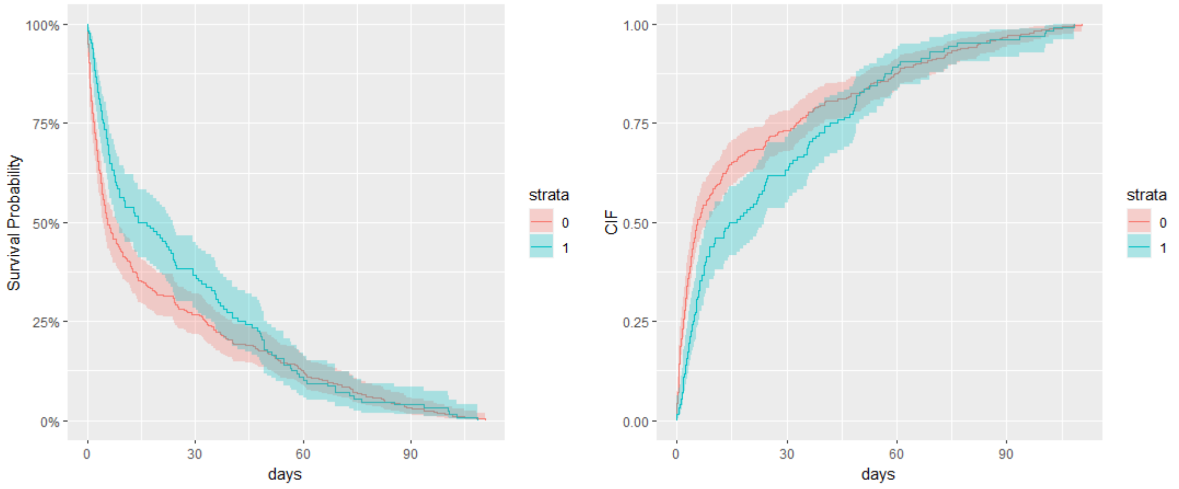
Naive approach: K-M

The K-M approach focuses on the non-occurrence of events. The outcome of interest indicates that no death and no relapse is important. Further, the K-M plot estimates the proportion of patients who have not experienced any of the endpoints (relapse or death). Figure 2.1 (left panel) shows that the patients with high platelet counts have a better chance of surviving without experiencing relapse or death. Here, the CIF can be obtained by plotting inversely (i.e., 1-KM estimate), which is the estimated proportions of patients undergoing one of the endpoints (relapse or death) over time (Figure 2.1, right panel).

The R code is explained below:

```
library(survival)
km=survfit(formula   = Surv(time, cause != "Censored") ~ platelet,
        data      = bmt, type       = "kaplan-meier",
        error     = "greenwood", conf.type = "log-log")
autoplot(km, xlab="days", ylab = "Survival Probability")
autoplot(km,fun = function(x) {1 - x},  xlab="days", ylab = "CIF") ##1-KM
```

FIGURE 2.1: left- K-M survival probability for event of interest (TRM) and CR event (Relapse), right-(1-event free survival probability). Here, 0 indicates “low platelet count”, while 1 indicates “high platelet count”



Sub-distribution function

The sub-distribution function can be estimated using the cmprsk package with *ftime* and *fstatus* arguments. Specifically, *ftime* defines the variable containing the observation time, while *fstatus* represents the event rate (by default, 0 denotes censored observations). The command is thus *cuminc(ftime, status).* The group argument takes a variable specifying distinct groups. The basic results and estimates contained within the *ci.cmprsk* object may be obtained with the option *print.cuminc().* In that option, the argument *ntp* indicates number of the periods (for instance, ntp=3 indiactes $3-$ time points) for which the estimates of the sub-distribution functions and their variances are needed. It is also possible to estimate the CIF values for a particular time point. Here, the CIF estimates are shown for the $10, 50$, and 90 time points.

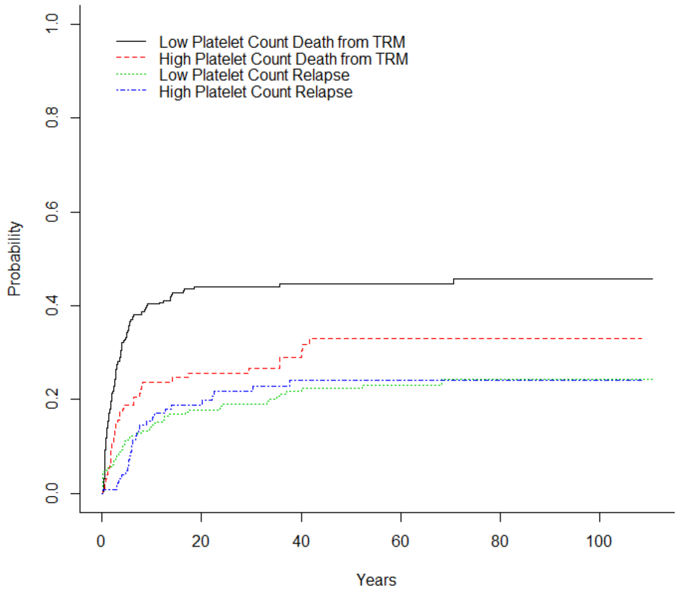
```
library(cmprsk)
ci.cmprsk=cuminc(ftime=bmt$time,fstatus=bmt$cause, group = bmt$platelet)
print(ci.cmprsk, ntp=3)

CIF=cuminc(bmt$times,bmt$cause, bmt$platelet)
CIF=timepoints(CIF, c(10,50,90))

Estimates and Variances are as follows:
$est
                                         10        50        90
Low Platelet Count Death from TRM 0.4035896 0.4457578 0.4582107
High Platelet Count Death from TRM 0.2377285 0.3310266 0.3310266
Low Platelet Count Relapse         0.1451157 0.2235970 0.2429186
High Platelet Count Relapse        0.1544569 0.2412607 0.2412607

$var
                                       10           50           90
Low Platelet Count Death from TRM 0.0008761345 0.0009252231 0.0010397042
High Platelet Count Death from TRM 0.0014547312 0.0021189199 0.0021189199
Low Platelet Count Relapse         0.0004534633 0.0007111760 0.0008672657
High Platelet Count Relapse         0.0010742456 0.0016472975 0.0016472975
```

The estimated CIFs can be visualized with generic function *plot().*

```
plot(ci.cmprsk, curvlab = c("Low Platelet Count, Death from TRM",
                            "High Platelet Count, Death from TRM",
                            "Low Platelet Count, Relapse",
                            "High Platelet Count, Relapse"),
     col=c(1:4))
```

Figure 2.2 shows that a low platelet count has a higher risk of death from TRM than a high platelet count. However, difference between low and high platelet counts was not prominent for the other causes (Relapse). The formal statistical test for the difference between groups can be performed using the modified χ^2 statistic (Gray, 1988).

FIGURE 2.2: CIF for both causes (TRM and Relapse) among different groups for "platelet" covariate



```
Tests:
                    stat           pv         df
Death from TRM    8.68527512  0.003207912   1
Relapse           0.02290726  0.879698496   1
```

The first column of the output shows the χ^2 statistic for the between-group test, and the second column shows the respective p values. The test results point to the statistically significant difference between the TRM sub-distribution functions for low and high platelet counts (p=0.003) and lack of difference for relapse (p=0.88).

Classical approach: Aalen-Johansen estimator

Now, the previous results can be compared with those from the classical Aalen-Johansen estimator. The package used is `prodlim`, and the CIF for time points $10, 50$, and 90 for event types 1 and 2 are computed separately. Here, similar results were obtained as from the previous function.

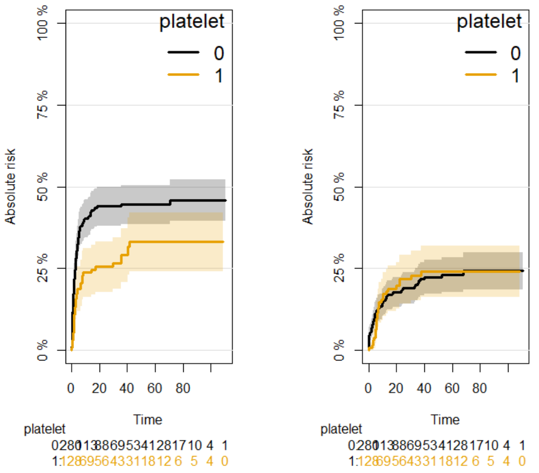
```
library (prodlim)
aj <- prodlim(Hist(time, cause) ~ platelet, data = bmt2)
res1=summary(aj,  cause=1, times=c(10, 50, 90))
res2=summary(aj,  cause=2, times=c(10, 50, 90))
```

```
----------> Cause:  1 (TRM)
platelet=0 (Low Platelet Count) :
  time n.risk n.event n.lost cuminc se.cuminc lower upper
1   10    116       0      0  0.404    0.0295 0.346 0.461
2   50     49       0      0  0.446    0.0303 0.386 0.505
3   90      9       0      0  0.458    0.0321 0.395 0.521

platelet=1 (High platelet count) :
  time n.risk n.event n.lost cuminc se.cuminc lower upper
1   10     72       0      0  0.238    0.0380 0.163 0.312
2   50     22       0      0  0.331    0.0456 0.242 0.420
3   90      5       0      0  0.331    0.0456 0.242 0.420

----------> Cause:  2 (Relapse)
platelet=0 (Low Platelet Count) :
time n.risk n.event n.lost cuminc se.cuminc lower upper
1   10    116       0      0  0.145    0.0212 0.103 0.187
2   50     49       0      0  0.224    0.0266 0.172 0.276
3   90      9       0      0  0.243    0.0292 0.186 0.300

platelet=1  (High Platelet Count) :
  time n.risk n.event n.lost cuminc se.cuminc  lower upper
1   10     72       0      0  0.154    0.0326 0.0906 0.218
2   50     22       0      0  0.241    0.0403 0.1623 0.320
3   90      5       0      0  0.241    0.0403 0.1623 0.320
```

The *plot* option gives the CIF for the grouping factor “platelet” (Figure 2.3).

```
par(mfrow=c(1,2))
plot(ajx)
plot(ajx,cause=2)
```

FIGURE 2.3: CIF for event type 1 (TRM, left panel) and CIF for event type 2 (relapse, right panel). 0 indicates the "low platelet count" while 1 indiactes the "high platelet count".



2.2.2 Regression approaches

The CSH approach

The CSH regression model can be fitted with Cox regression by considering only the event of interest and failure in other events as the censored observation. The effect of covariates on the CSH can be estimated using Cox proportional hazard regression. The model is fitted with *coxph()* function in the `survival` package.

```
summary(coxph.TRM<-coxph(Surv(time, cause == 1)~
                                age+ platelet + tcell, data=bmt))
summary(coxph.Relapse<-coxph(Surv(time, cause == 2)~
                              age+platelet+tcell, data=bmt))

Call:
coxph(formula = Surv(time, cause == 1) ~ age + platelet + tcell,
    data = bmt)


  n= 408, number of events= 161
```

```
             coef exp(coef) se(coef)      z Pr(>|z|)
age       0.40836   1.50435  0.08903  4.587 4.51e-06 ***
platelet -0.51987   0.59460  0.18721 -2.777  0.00549 **
tcell    -0.65169   0.52116  0.27634 -2.358  0.01836 *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

         exp(coef) exp(-coef) lower .95 upper .95
age         1.5043     0.6647    1.2635    1.7912
platelet    0.5946     1.6818    0.4120    0.8582
tcell       0.5212     1.9188    0.3032    0.8958

Concordance= 0.645  (se = 0.022 )
Likelihood ratio test= 37.35  on 3 df,   p=4e-08
Wald test            = 33.68  on 3 df,   p=2e-07
Score (logrank) test = 34.72  on 3 df,   p=1e-07
```

The first argument of the *coxph()* function takes an object of class *Surv*, where the *cause*== 1 indicates the event type 1 and other values are considered as censored. Similarly, event type 2 can be computed as *cause*== 2. In the latter case, the software assumes interest in computing event type 2 and thus considers it the main event of interest, which also allows it to assume that event type 1 is censored. The summary output shows the coefficients and corresponding hazard ratio ($e^{\hat{\beta}}$). The last section on the above code displays statistics for model's goodness of fit.

An alternative way to perform the analysis is to use the *CSC()* function contained in the `riskRegression` package. The summary output is similar to that of the *coxph()* function except that the *CSC()* function automatically produces the results of the CSH models for both types of events together.

```
install.packages ("riskRegression")
library (riskRegression)
fit3<-CSC(Hist (time, cause) ~ platelet + age+  tcell, data=bmt)
summary(fit3)

----------> Cause:  1
Call:
coxph(formula = survival::Surv(time, status) ~ age + platelet +
    tcell, x = TRUE, y = TRUE)
```

```
  n= 408, number of events= 161

             coef exp(coef) se(coef)      z Pr(>|z|)
age       0.40836   1.50435  0.08903  4.587 4.51e-06 ***
platelet -0.51987   0.59460  0.18721 -2.777  0.00549 **
tcell    -0.65169   0.52116  0.27634 -2.358  0.01836 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

         exp(coef) exp(-coef) lower .95 upper .95
age         1.5043     0.6647    1.2635    1.7912
platelet    0.5946     1.6818    0.4120    0.8582
tcell       0.5212     1.9188    0.3032    0.8958

Concordance= 0.645  (se = 0.022 )
Likelihood ratio test= 37.35  on 3 df,   p=4e-08
Wald test            = 33.68  on 3 df,   p=2e-07
Score (logrank) test = 34.72  on 3 df,   p=1e-07

----------> Cause:  2
Call:
coxph(formula = survival::Surv(time, status) ~ age + platelet +
    tcell, x = TRUE, y = TRUE)

  n= 408, number of events= 87

            coef exp(coef) se(coef)      z Pr(>|z|)
age       0.1425    1.1532   0.1118  1.275    0.202
platelet -0.2346    0.7909   0.2321 -1.011    0.312
tcell     0.3015    1.3519   0.2827  1.067    0.286

         exp(coef) exp(-coef) lower .95 upper .95
age         1.1532     0.8671    0.9263     1.436
platelet    0.7909     1.2644    0.5018     1.246
tcell       1.3519     0.7397    0.7769     2.353
```

```
Concordance= 0.584  (se = 0.033 )
Likelihood ratio test= 4.46  on 3 df,   p=0.2
Wald test            = 4.45  on 3 df,   p=0.2
Score (logrank) test = 4.48  on 3 df,   p=0.2
```

Predictions using the Cox PH model

Moreover, patients' risk can be predicted with the given covariates with the results from the fitted regression model. As an example, let us consider predicting the risk of a patient aged "36 years" old who has "non t-cell depleted BMT" and "high platelet count." The prediction can be made using the pec package. Then, a data frame is created for the predicted patients and then used in the *predict* object along with the fitted model for event type 1.

```
library (pec)
fit3<-CSC(Hist (time, cause) ~ age+platelet + tcell, data=bmt)
newdata=data.frame
        (age=c(36,36),
          platelet=c("High Platelet Count","Low Platelet Count"),
       tcell=c("T-cell Depleted BMT", "Non T-cell Depleted BMT"))
newdata
age            platelet                    tcell
1  36 High Platelet Count     T-cell Depleted bmt2
2  36  Low Platelet Count     Non T-cell Depleted bmt2
predictEventProb(fit3, cause=1 ,newdata=newdata, time=c(10, 20))
```

The CIF at time points of 10 and 20 days are:

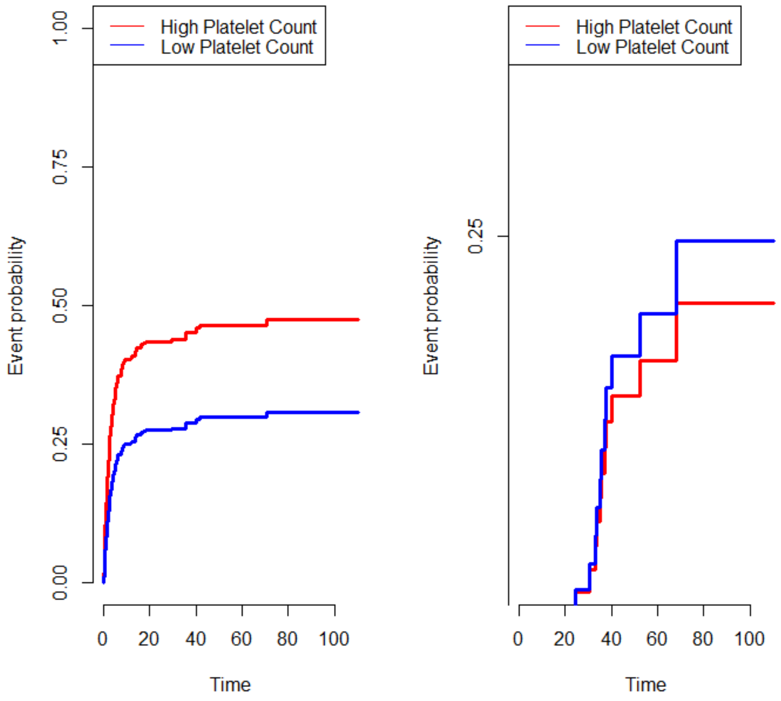
```
                   Times
          [,1]        [,2]
[1,] 0.4019179    0.4349749
[2,] 0.2502650    0.2753849
```

The first row represents CIF of death from TRM for the aforementioned patient at time points 10 and 20, whereas the second row represents the CIF for the same patient, but with "low platelet count" and "non t-cell depleted BMT" at the same time points. The visual representation for this patient can be done as shown in the figure below (Figure 2.4):

```
par(mfrow=c(1,2))
```

```
plotPredictEventProb(fit3,newdata=newdata,cause=1,col=c("red","blue"))
legend("topleft",legend=c("High Platelet Count","Low Platelet Count"),
          lty=1,col=c("red","blue"))

plotPredictEventProb(fit3,newdata=newdata,cause=2,
                    col=c("red","blue"), ylim=c(0.2,0.28))
legend("topleft",legend=c("High Platelet Count","Low Platelet Count"),
            lty=1,col=c("red","blue"))
```

Figure 2.4: CIF for event type 1 for patient aged 36 years (left panel); CIF for event type 2 for the patient aged 36 years (right panel)



The SDH approach

Here, the following formula can be used to obtain the CIF and it's predicted value at time points $\{10, 50, 90\}$ using the `cmprsk` package:

```
CIF=cuminc(bmt$times, bmt$cause, bmt$platelet)
CIF=timepoints(CIF, c(10,50,90))  ##timepoints is a function, so need object CIF.


$est
            10          50          90
```

```
Low Platelet Count TRM  0.4035896  0.4457578  0.4582107
High Platelet Count TRM   0.2377285  0.3310266  0.3310266
Low Platelet Count Relapse   0.1451157  0.2235970  0.2429186
High Platelet Count Relapse   0.1544569   0.2412607  0.2412607


$var
               10            50              90
Low Platelet Count TRM  0.0008761345  0.0009252231  0.0010397042
High Platelet Count TRM  0.0014547312  0.0021189199  0.0021189199
Low Platelet Count Relapse 0.0004534633  0.0007111760  0.0008672657
High Platelet Count Relapse 0.0010742456  0.0016472975  0.0016472975
```

However, the following alternatives need to be applied to obtain the covariate effects. The Fine-Gray model can be fitted using the *FGR()* function shipped with the `riskRegression` package.

```
SDH <- FGR(Hist(time,cause)~ platelet+tcell+age,data=bmt)
> SDH
Right-censored response of a competing.risks model
No.Observations: 408
Pattern:
Cause     event right.censored
  1         161              0
  2          87              0
  unknown     0            160
Fine-Gray model: analysis of cause 1
Competing Risks Regression
Call:
FGR(formula = Hist(time, cause) ~ age + platelet + tcell, data = bmt,
    cause = "1")

           coef exp(coef) se(coef)     z p-value
age       0.344     1.410   0.0803  4.28 1.8e-05
platelet -0.425     0.654   0.1806 -2.35 1.9e-02
tcell    -0.596     0.551   0.2704 -2.20 2.7e-02

         exp(coef) exp(-coef)  2.5% 97.5%
age          1.410      0.709 1.205 1.650
```

```
platelet     0.654      1.530 0.459 0.931
tcell        0.551      1.815 0.324 0.936

Num. cases = 408
Pseudo Log-likelihood = -909
Pseudo likelihood ratio test = 28.8  on 3 df,

Convergence: TRUE



##Fine-Gray model: analysis of cause 2



Call:
FGR(formula = Hist(time, cause) ~ age + platelet + tcell, data = bmt2,
        cause = 2)

            coef exp(coef) se(coef)       z p-value
age      -0.0113     0.989    0.122 -0.0929   0.930
platelet -0.0398     0.961    0.228 -0.1744   0.860
tcell     0.5304     1.700    0.280  1.8930   0.058

         exp(coef) exp(-coef)  2.5% 97.5%
age          0.989      1.011 0.779  1.26
platelet     0.961      1.041 0.615  1.50
tcell        1.700      0.588 0.981  2.94

Num. cases = 408
Pseudo Log-likelihood = -502
Pseudo likelihood ratio test = 3.3  on 3 df,

Convergence: TRUE
```

As observed, the estimated coefficient for cause 1 deviated a little from that obtained from the CSH model (Hazard ratio for age: 1.41 vs. 1.50), which reflected the different assumptions for the CR. Moreover, the numerical values derived from the Fine-Gray model have no simple interpretation, but it reflects the ordering of cumulative incidence curves (Fine and Gray, 1999). Here, it should be noted that the CSH is the rate of cause 1 failure per time unit for still alive patients. However, event type 1 when using the SDH is the rate of event type 1 failure per time unit for patients who are either alive or have already failed in terms of event type 2. This function then calls another function *crr()* from the `cmprsk` package.

Alternatively, a *crr* function in the `cmprsk` package developed by Gray *et al.* (2004) can be applied to fit the proportional SDH approach. This allows the model to have time-dependent covariates or test proportionality by adding a time-dependent covariate. The predicted CIF for a given set of covariate values can be calculated using the *predict.crr* function and then to produce a plot, the *plot.predict.crr* function can be used. Here, the function arguments are different from those in *FGR()* function. The *model.matrix* function can be used to generate suitable matrices of covariates from the *factors.cov1* argument that specifies the matrix of explanatory variables for which the sub-distribution functions are to be estimated. Meanwhile, the argument *failcode* indicates the event of interest in the specified variable in the *fstatus* argument for which the model is estimated. The results are similar, and therefore, were not reported here.

```
crr.mat <- model.matrix(~age+platelet+tcell, data=bmt)[,-1]
mod.trm<-crr(bmt$time,bmt$cause,cov1=crr.mat,failcode=1 )
mod.relapse<-crr(bmt$time,bmt$cause,cov1=crr.mat, failcode=2)
summary(mod.trm)
summary(mod.relapse)
```

Model prediction

The fitted Fine-Gray model can predict new observations with the given combinations of covariates. In the following example, a new dataset containing three patients is given.

```
newdata<-data.frame(platelet=c("Low Platelet Count",
                    "Low Platelet Count","High Platelet Count"),
           age=c(30,31,32),
           tcell=c("Non T-cell Depleted BMT",
                  "Non T-cell Depleted BMT","T-cell Depleted BMT"))
```

```
 newdata
             platelet  age                  tcell
```

```
1  Low Platelet Count  30  Non T-cell Depleted BMT
2  Low Platelet Count  31  Non T-cell Depleted BMT
3 High Platelet Count  32     T-cell Depleted BMT
```

However, the data frame had to be transformed to a matrix and the factor variables to dummy variables. Further, the *predict()* function applied to the *crr* object requires the columns of *cov* be in line with that in the original call to *crr()* function. This is because both platelet and t-cell are factor variables and need to be transformed to dummy variables using the *model.matrix()* function. Alternatively, a custom-made function called *factor2ind()* by Scrucca *et al.* (2010) can be useful.

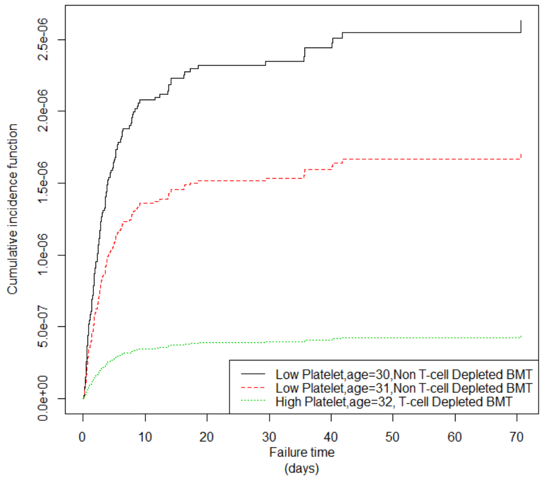
```
dummy.new<-model.matrix(~platelet+age+tcell,data=newdata)[,-1]
dummy.new
plateletLow Platelet Count  age  tcellT-cell Depleted BMT
1                          1   30                        0
2                          1   31                        0
3                          0   32                        1
```

Here, the variables “platelet” and “t-cell” are transformed to 0/1 variables, while the varaible “age” is continuous and thus unchanged. So, the prediction code will be as shown below:

```
pred<-predict(mod.trm,dummy.new)
plot(pred,lty=1:3,col=1:3,xlab="Failure time(days)",
                         ylab="Cumulative incidence function")
legend("topleft",
         c("Low Platelet,age=30,Non T-cell Depleted BMT",
            "Low Platelet,age=35, Non T-cell Depleted BMT",
            "High Platelet,age=50, T-cell Depleted BMT"),
         lty=1:3,col=1:3)
```

An object of *crr* class is passed to the *predict()* function, followed by a matrix containing the covariate combinations (Zhang, 2017). The *predict()* function returns a matrix (not shown) with the unique *cause =1* event times in the first column, and the other columns give the estimated subdistribution function corresponding to the covariate combinations at each event time. Figure 2.5 was produced using the generic function *plot()* to draw CIF for each patient.

FIGURE 2.5: CIF for three patients with given covariate values.



Here, prediction and plotting are more convenient using functions in the `riskRegression` package. Specifically, the function *riskRegression()* provides a variety of link functions for the survival regression model in the presence of CR (Gerds *et al.*, 2012). The link argument controls the link function to be used: *prop* for the regression model of Fine and Gray (1999), *relative* for the absolute risk regression model, and *logistic* for the logistic risk regression model.

```
reg<-riskRegression(Hist(time, cause) ~ age+platelet+tcell,
                              data = bmt, cause = 1,link="prop")
reg
plot(reg,newdata=newdata)
```

```
> reg
Competing risks regression model

IPCW weights: marginal Kaplan-Meier for the censoring distribution.
Link: 'cloglog' yielding sub-hazard ratios \citep{Fine}
No covariates with time-varying coefficient specified.

Time constant regression coefficients:
 Variable Levels  Coef Lower Upper  Pvalue
      age        1.383 1.155 1.656 0.00041
 platelet        0.571 0.382 0.852 0.00605
    tcell        0.499 0.276 0.903 0.02171
```

```
Note: The coefficients (Coef) are sub-hazard ratios
         (Fine & Gray 1999)
```

The binomial regression approach

The binomial regression approach depends on the censoring weighting technique. More generally, it has been established that regression modeling for inverse probability censoring weights (IPCW) can improve efficiency (Scheike *et al.*, 2008). Thus, it is important that before estimating the cumulative incidence curve when using this approach, the censoring weights need to be estimated without bias. For instance, if it is found that the censoring distribution depends significantly on the covariates X and is well described by Cox's regression model, using a simple Kaplan-Meier estimate for the censoring weights may lead to severely biased estimates. Therefore the option *cens.model = "cox"* would need to be selected in the function call. To analyze the data, the `timereg` package in R was used as previously mentioned. Here, the event time and censoring variable are specified in timereg's *comp.risk()* function as *Event (time, cause).* The cause variable gives the causes associated with the different events. Then, *causes* = 1 specifies that type 1 events be considered.

Assuming that the time-varying effect of a covariate has to be identified, the following fully non-parametric additive model can be first fit using *log link* function

$$-\log\left\{1 - CIF_1(t; \mathbf{x}, \mathbf{z})\right\} = \alpha(t)^\top \mathbf{x} + \gamma^\top \mathbf{x}$$

where, the regression coefficients $\alpha(t)$ and γ are estimated by a simple binomial regression approach. Then a proportional SDH model proposed by Fine and Gray (1999) is fit for comparison purposes. Finally, a flexible model that incorporates both time-varying and time fixed covariates is fit.

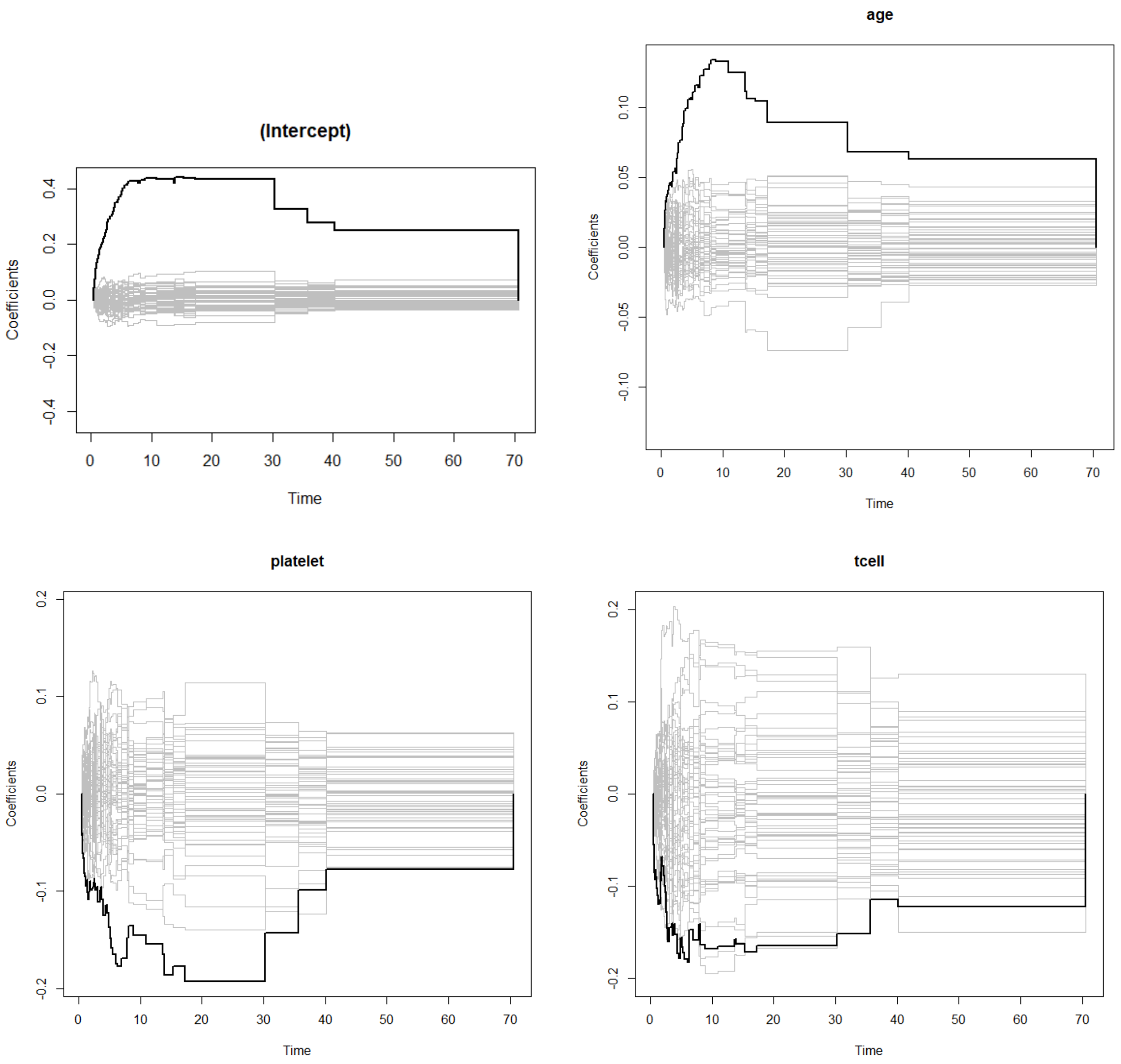
```
library ("timereg")
Aalen <- comp.risk(Event(time, cause) ~ age + platelet + tcell,
                           data = bmt, cause=1,  n.sim=5000,
                          model = "additive", cens.model = "cox")
> summary(Aalen)
Competing risks Model

Test for nonparametric terms

Test for non-significant effects
```

```
            Supremum-test of significance p-value H_0: B(t)=0
(Intercept)                         11.00              0.0000
age                                  5.64              0.0000
platelet                             3.23              0.0142
tcell                                4.99              0.0000

Test for time invariant effects
                  Kolmogorov-Smirnov test p-value H_0:constant effect
(Intercept)                         0.441                      0.0000
age                                 0.134                      0.0000
platelet                            0.193                      0.0034
tcell                               0.183                      0.0352
                    Cramer von Mises test p-value H_0:constant effect
(Intercept)                         4.710                      0.0000
age                                 0.291                      0.0000
platelet                            0.739                      0.0028
tcell                               0.844                      0.0162
```

The tests of significance based on the non-parametric tests show that all the covariates are significant. Then a figure can be created using the *sim.ci* and *score* options to plot the estimated regression coefficients with their 95% confidence bands and the observed test process for constant effects and simulated test processes under the null, respectively (Scheike and Zhang, 2011). R code was used to create the time-varying coefficient effects shown in Figure 2.6:

```
plot(Aalen, sim.ci=2)
plot(Aalen, score=1)
```

FIGURE 2.6: Observed (black line) and simulated test processes under the null (gray lines).



These effects are not constant over time. Here, there were 95% pointwise confidence intervals and bands (sim.ci=2 in the plot call, 2 for broken lines).

Alternatively, the model can be fit using the `survival` package, and then the estimates shown in Figure 2.7 can be created using `ggplot2` and `ggfortify` packages. The fitted model results are not reported here because of the similar results, however, the plot has been shown.

```
aa_fit <-aareg(Surv(time, cause==1) ~ age+platelet + tcell , data = bmt)
library(ggfortify)
library(ggplot2)
autoplot(aa_fit)
```

FIGURE 2.7: Estimates of time-varying effects with 95% confidence intervals

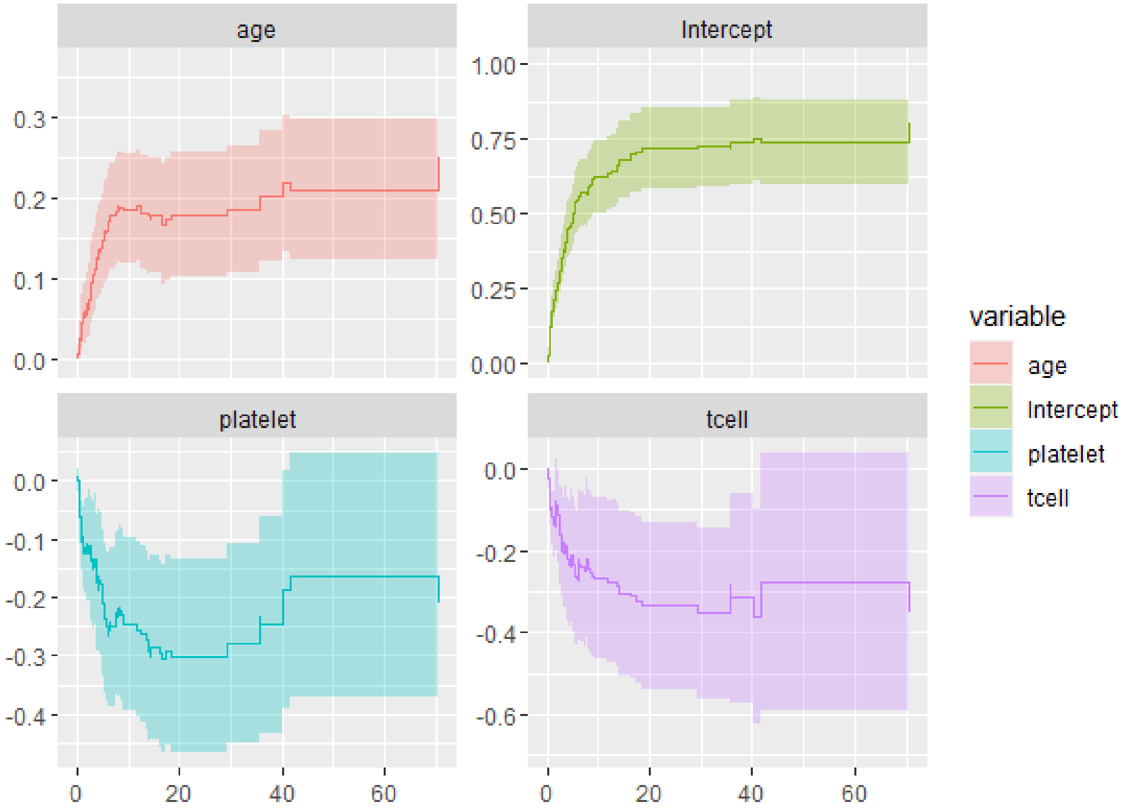


Figure 2.6 shows the related test processes under the null hypothesis of constant effect, together with the observed processes (black lines), for deciding whether covariate effects are significantly time-varying or whether a null hypothesis can be accepted. The summary of these graphs is given in the output, and it is observed that all of the covariates (age, platelet, t-cell) are time-varying, and thus not consistent with the model by Fine and Gray (1999) as shown below. The intercept is also time-varying and in Figure 2.7 it is increasing until time point 20 and then becomes flat after that time point. Moreover, the p values related to these plots are given in the above output, and it can be seen that the Kolmogorov-Smirnov (supremum) test leads to p values of 0.00, 0.003, and 0.035, for age, platelet, t-cell, respectively. Similarly, the Cramer von Mises test statistics based on the same score processes are 0.00, 0.003, and 0.016, respectively. These test statistics are described in detail in the paper by Scheike and Zhang (2011). Here, it should be noted that the two different summaries of the test processes using the Kolmogorov-Smirnov and Cramer von Mises tests statistics are consistent with the figures, and the overall conclusion is that none of the three variables have proportional Cox type effects. This means that, in each figure, the observed process of the data

(black line) is outside (far from) the process simulated under the null hypothesis under the model (gray lines).

Fine and Gray model for comparison:

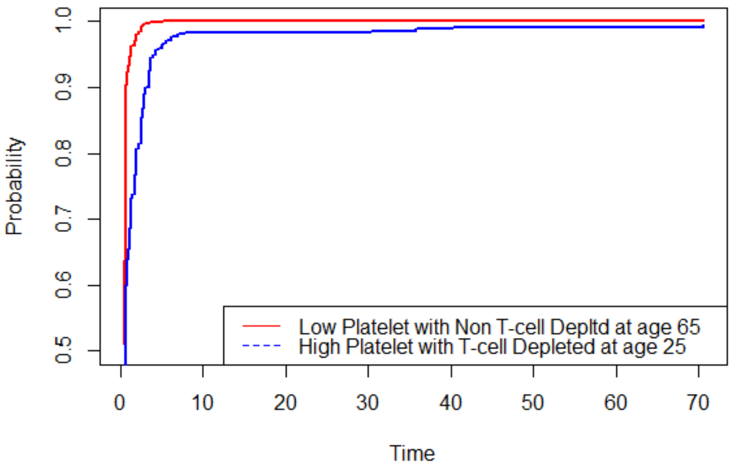
```
FG <- comp.risk(Event(time, cause) ~ const(age) + const(platelet) +
                     const(tcell), data = bmt, cause=1, n.sim=5000,
                     model = "fg", cens.model = "cox")
summary(FG)
> summary(FG)
Competing risks Model

Test for nonparametric terms

Test for non-significant effects
            Supremum-test of significance p-value H_0: B(t)=0
(Intercept)                          10.3                   0

Test for time invariant effects
                  Kolmogorov-Smirnov test p-value H_0:constant effect
(Intercept)                         0.496                           0
                    Cramer von Mises test p-value H_0:constant effect
(Intercept)                          13.4                           0

Parametric terms :
            Coef.    SE   Robust SE    z         P-val     lower  upper
                                                            2.5%  97.5%
const(age)  0.326  0.0918    0.0918   3.55    0.000388   0.146   0.506
const(plat)-0.555  0.2040    0.2040  -2.73    0.006420  -0.955  -0.155
const(tcell)-0.691 0.3020    0.3020  -2.29    0.022000  -1.280  -0.099
```

We can conclude from the above results that both the Fine-Gray and non-parametric additive models show non-proportionality. Thus, the Aalen model will be used as the final model for prediction. For instance, a new data set as assigned below for the ages 25 and 65 years old with the covariates of “platelet,” and “t-cell” are:

```
ndata<-data.frame(platelet=c("Low Platelet Count",
                              "High Platelet Count"),
                    age=c(65,25),
                        tcell=c("Non T-cell Depleted BMT",
                                "T-cell Depleted BMT"))

ndata

out<-predict(FG,ndata)

par(mfrow=c(2,2))
plot(out,multiple=1,uniform=0, ylim=c(0.5,1),
     col=c("red", "blue"),lty=1,se=0)
legend("bottomright",c("Low Plat with Non T-cell Depltd BMT
                       for age 65,",
                       "High Plat with T-cell Depleted BMT
                       for age 25"),lty=1:2,col=c("red", "blue"))
```

Moreover, the predictions can be plotted (Figure 2.8) without pointwise confidence intervals (*se = 0*) and confidence bands (*uniform = 0*).

Figure 2.8: Predictions of the cumulative incidence curves for different ages under Aalen model



Alternatively, these estimates can be obtained using the ARR object in the `riskRegression` package.

```
> library(riskRegression)
> fit.arr <- ARR(Hist(time,cause)~age+platelet+tcell, data=bmt, cause=1)
> print(fit.arr)
Competing risks regression model

IPCW weights: marginal Kaplan-Meier for the censoring distribution.
Link: ’log’ yielding absolute risk ratios
No covariates with time-varying coefficient specified.

Time constant regression coefficients:

   Factor   Coef exp(Coef) StandardError      z         CI_95    Pvalue
      age 0.2484    1.2820        0.0735 3.3809 [1.110;1.481] 0.0007224
 platelet -0.456     0.634         0.173 -2.635 [0.452;0.890] 0.0084249
    tcell -0.566     0.568         0.261 -2.169 [0.341;0.947] 0.0300456

Note: The values exp(Coef) are absolute risk ratios
```

The pseudo-value approach

The R function *pseudoci* in the `pseudo` package has three arguments: *time* (the event time variable), *event* (1, if there is occurrence of risk $1; 2$ if there is occurrence of risk 2; and 0 otherwise), and *tmax* (a list of time points at which the pseudo-values are to be computed). This routine produces an object containing the pseudo-values for both CR. In this context, a grid is presented of the three-time points from $10, 30$, and 50 days since the transplant and fitted with the regression model with three variables as follows: Z_1 = platelet counts (binary variable); Z_2 = patient age (continuous variable); and Z_3 = t-cell (binary variable). First, a regression model $\phi(\theta_{ij}) = \alpha_j + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3}$ with a complementary log-log link function is fitted, which is equivalent to the Fine and Gray (1999) proportional SDH model. Here, the probability of TRM is being modeled.

For each time cutoff, the *pseudoci* function generated several columns of pseudo-values (one column for each CR). Then, a generalized estimating equation (GEE) is run using the *geese* function. To use this function, only the TRM pseudo-values (if we are interested in TRM as a main event of interest) are needed and arranged in one column. In another column, the pseudo-value's time points are needed with the original BMT data, and these are prepared for analysis using the function *geese.* By default, this function uses a different "sandwich" estimator of the variance. An alternative to this estimator is the jackknife variance estimators (Yan and Fine, 2004). Here, the routine *geese* allows the user to decide between the fully iterated one-step and approximate jackknife (AJ) variance estimates. The AJ variance estimate is recommended by Klein *et al.* (2008).

The R code and output are given below:

```
library(pseudo)
data(bmt)
#calculate the pseudo-observations
cutoffs <- c(10,30,50)
pseudo <- pseudoci(time=bmt$time,event=bmt$cause,tmax=cutoffs)
##for each time cutoff the "pseudoci" function generated several
## columns of pseudovalues (one column for each competing risk)
#rearrange data into long format
#use only pseudo-observations for TRM (in the code: pseudo$pseudo[[1]]
#for cause 1 TRM)
b <- NULL
for(j in 1:length(pseudo$time)){
        b <- rbind(b,cbind(bmt,pseudo = pseudo$pseudo[[1]][,j],
                           tpseudo = pseudo$time[j],id=1:nrow(bmt)))
}
b <- b[order(b$id),]
> head(b)
     X time cause platelet     age tcell       pseudo tpseudo id
1    1 0.03     2        0 0.19566     0 0.000000e+00      12  1
1100 1 0.03     2        0 0.19566     0 0.000000e+00      36  1
1102 1 0.03     2        0 0.19566     0 2.842171e-14      60  1
2    2 0.03     2        0 0.63005     0 0.000000e+00      12  2
2100 2 0.03     2        0 0.63005     0 0.000000e+00      36  2
2102 2 0.03     2        0 0.63005     0 2.842171e-14      60  2

# fit the model for cause 1
fit <- geese(pseudo ~ age+platelet+tcell,data =b, jack = TRUE,
    scale.fix=TRUE, family=gaussian,  mean.link = "cloglog",
                corstr="independence")
#The results using the Approximate Jackknife (AJ) variance estimate
cbind(mean = round(fit$beta,4), SD = round(sqrt(diag(fit$vbeta.ajs)),4),
      Z = round(fit$beta/sqrt(diag(fit$vbeta.ajs)),4),
      PVal = round(2-2*pnorm(abs(fit$beta/sqrt(diag(fit$vbeta.ajs)))),4))
```

```
               mean     SD       Z   PVal
(Intercept) -0.5504 0.0959 -5.7420 0.0000
age          0.3035 0.0844  3.5972 0.0003
platelet    -0.3873 0.1906 -2.0318 0.0422
tcell       -0.5638 0.2926 -1.9267 0.0540
```

Similarly, results can be obtained for cause 2. Here, the b matrix has to be replaced as follows:

```
b2 <- NULL
for(j in 1:length(pseudo$time)){
        b2 <- rbind(b2,cbind(bmt,pseudo = pseudo$pseudo[[2]][,j],
                           tpseudo = pseudo$time[j],id=1:nrow(bmt)))
}
```

In this model, positive values of β for a covariate suggest a larger cumulative incidence for patients with Z = 1 or equivalently more TRM.

It should be noted that while the observed CIF value was not obtained from the R package `pseudo`, the estimated regression parameters $\hat{\beta}$ and $s.e.(\hat{\beta})$ were obtained. To compute the variance of the estimated CIF, the procedures explained in Subsection 1.4.5 were followed.

```
###first we have all parameters beta0 and beta1

pseudo_mean_1=cbind(mean =fit$beta[1:4],
                    VAR = diag(fit$vbeta.ajs[1:4,1:4]))

pseudo_mean_1
                  mean         VAR
(Intercept) -0.5503886 0.009187705
age          0.3035054 0.007118617
platelet    -0.3873073 0.036337434
tcell       -0.5637857 0.085625006

### only beta0 for 4 time points, so when x=0
pseudo_mean_10=cbind(mean =fit$beta[1:3],
                     VAR = diag(fit$vbeta.ajs[1:3,1:3]))
>pseudo_mean_10
            mean         VAR
```

```
(Intercept) -0.5503886 0.009187705
age          0.3035054 0.007118617
platelet    -0.3873073 0.036337434

###when x=1, so beta0+beta1
pseudo_mean_11=cbind(mean=pseudo_mean_1[1:3,1]+pseudo_mean_1[4,1],
                     VAR=pseudo_mean_1[1:3,2]+pseudo_mean_1[4,2]
                              + 2*fit_11$vbeta.ajs[1:3,4])
>pseudo_mean_11
                  mean       VAR
(Intercept) -1.1141743 0.1025666
age         -0.2602802 0.1026712
platelet    -0.9510930 0.1327096
```

The pseudo-value approach allows direct regression modeling of the CIF for CR data. The pseudo-values for each observation need to be computed to apply this technique. Next, a standard generalized estimating equation approach is used to obtain regression estimates. The interpretation under the pseudo-values approach is based on the modeling of $CIF_k(t \mid z) = 1 - \exp\left\{-\Lambda_0(t)e^{\beta z}\right\}$. Here, negative values of β for a covariate (say, platelet) suggest a smaller cumulative incidence for subjects with high platelet count ($z = 1$).

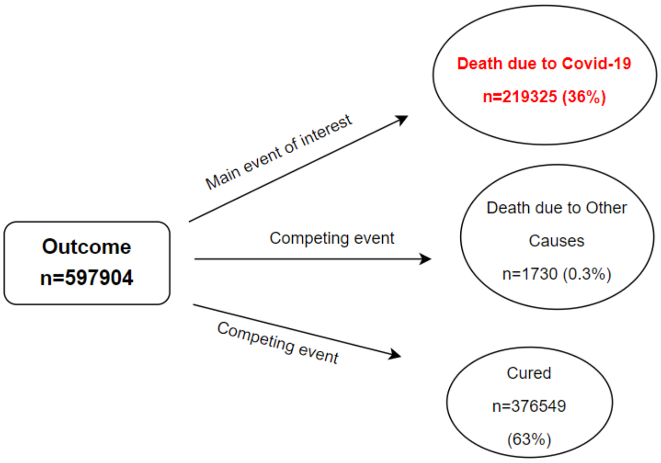
Under BMT data, the objective was to explore all the regression approaches for computing CIF using R software. To this end, we first introduced conventional techniques such as the K-M method. It was clear that the K-M technique showed biased results and thus appropriate techniques need to be addressed. Next, we applied nonparametric and semi-parametric regression approaches under CSH and SDH competing risks settings. The CIF under these approaches was found almost similar over time. Under the CSH approach for the main event of interest, if all variables are kept constant except age in two subjects, one-year older subject has a 50% higher risk of dying. Moreover, those who have high platelet counts were 40% less at risk of dying in comparison to low platelet counts. Next, similar results were found for the covariate t-cell. Under the SDH approach, the results were slightly different (for covariate age, the hazard is 1.41; for platelet 0.65, and for t-cell 0.55). Furthermore, we applied binomial and pseudo-value regression approaches which allow direct modeling of the CIF through an appropriate link function. Moreover, under the binomial regression approach, we used time-varying coefficients and found that the covariates age, platelet, and t-cell were not constant over time. It is important to note that while using the binomial approach, the censoring weights need to be estimated without bias before estimating correctly the CIF.

2.3 Real data example from subjects with Covid-19

A novel coronavirus was identified by the end of 2019 as the cause of a cluster of pneumonia cases in Wuhan, a city in China's Hubei Province. It rapidly spread, resulting in an epidemic throughout China, followed by several outbreaks in other countries worldwide (Ge *et al.*, 2020). In February 2020, as the situation worsened, the World Health Organization named the disease COVID-19, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Later on March 11^{th}, the COVID-19 was classified as a global pandemic. In COVID-19 subjects, the time between exposure and symptom onset is estimated to be around 5 days, but may range from 2 to 14 days. Among those who died from the disease, the time from development of symptoms to death is between 6 to 41 days, with a median of 14 days (Ghosh *et al.*, 2021).

Data analysis from COVID-19 subjects is required to study clinical prognostic exposures, generate possible treatment drugs, and design intervention strategies. Many studies have investigated on Covid-19 data (Zuccaro *et al.*, 2021; Salinas-Escudero *et al.*, 2020; Nijman *et al.*, 2021; Rathouz *et al.*, 2021) to identify important exposures for the occurrence of death or cured. However, statistical models were presented by either ignoring the competing events or using inappropriate regression-based statistical methods. We applied competing risk survival analyses for estimating the CIF of dying from Covid-19 and the CIF of dying from other causes in subjects with Covid-19, which have been under observation from the date of symptoms to the date of death or exit from the study because they are cured. Our time-to-event data are obtained from the Ministry of Health of Brazil for all COVID-19 patients from January 01, 2020 to April 30, 2021.

FIGURE 2.9: Outcome variable for one main event of interest and two competing events



The exposures that we considered are as follows: **asthma** (1:yes, 2:no), **cardio.dis** (chronic cardiovascular disease 1:yes, 2:no), **cough** (1:yes, 2:no), **diabetes** (1:yes, 2:no), **diarrhea** (1:yes, 2:no), **dyspnea** (1:yes, 2:no), **fatigue** (1:yes, 2:no), **fever** (1:yes, 2:no), **flu.vaccine** (Flu vaccine last campaign 1:yes, 2:no), **hepatic.dis** (chronic liver disease 1:yes, 2:no), **immuno** (Immunodepression which is decreased from immunological system function 1:yes, 2:no), **kidney** (Chronic kidney disease 1:yes, 2:no), **loss.smell** (1:yes, 2:no), **loss.taste** (1:yes, 2:no), **neuro** (Neurological diseases 1:yes, 2:no), **obesity** (1:yes, 2:no), **other.risk** (Other risk factors 1:yes, 2:no), **other.symp** (1:yes, 2:no), **parto** (Has the subject given birth less tan 45 days from the first symptoms? 1:yes, 2:no), **pneumo** (lung chronic disease 1:yes, 2:no), **pneumo.dis** (Other Chronic Pneumatopathy 1:yes, 2:no), **resp.disc** (Respiratory Discomfort 1:yes, 2:no) , **fatigue** (1:yes, 2:no), **sore.throat** (1:yes, 2:no), **vomit** (1:yes, 2:no), **saturation** (oxygen saturation $<$ 95%? 1:yes, 2:no), **abdom.pain** (Abdominal pain 1:yes, 2:no), **race** (1:white; 2:black; 3:yellow; 4:brown; 5:indigenous), **risk.factor** (the subject does present some risk factor? 1:yes, 2:no), **age** (age in years at first symptoms), **sex** (Male=1; Female=2), and **ICU** (admitted to Intensive care Unit 1:yes, 2:no).

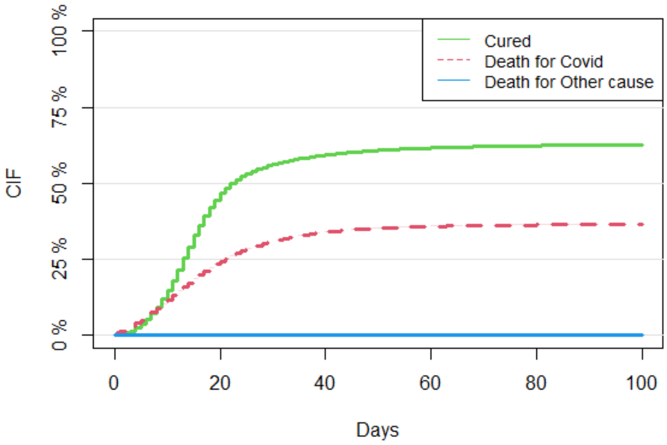
It was found in the recent literature (for instance, Ghosh *et al.* (2021)) that about 80% of deaths were in those over 60 years of age, and 75% had pre-existing health conditions. Thus, it was meaningful to study the effect of Covid-19 outcomes on different age groups. The variable **age** is categorized as follows: less than 40 years ("Young"); between $40-50$ years ("Young-Old"); between $50-60$ years ("Medium-Old"), between $60-70$ years ("Old"), and finally age greater than 70 years ("Old-old").

In our preliminary analysis (Sections 2.3.1 and 2.3.2), the time to become cured is considered the cause of interest and investigated on its own. However, in the competing risks regression analysis, this was considered as a censored time, since the main focus was on the causes of death. The latter violates the assumption of non-informative censoring, i.e., the cured patients are not representative of those who are still admitted to the hospital in terms of their risk of dying. However, when the approaches are analyzed based on the IPCW technique, these are particularly relevant because regression models can also account for dependent censoring.

2.3.1 Non-parametric estimation of the CIF

Figure 2.10 shows the CIF for cured subjects, death due to Covid, and death due to other causes, which is estimated non-parametrically by the Aalen Johansen estimator. Here, the estimated probability of death was 24% during the first 20 days from the day of symptoms and became 35% after 30 days. Meanwhile, the likelihood of cured was 50% during the first 20 days and around 60% after 40 days. The death due to other causes was found incomparable as the probability of death over time was slightly over 0%. The reason was that there were very few subjects who experienced death due to other causes (n=1, 730) compared to death due to covid (n=219, 325) and cured (n=376, 549) events.

FIGURE 2.10: CIF curve for cured, death due to Covid and other causes in the total population



2.3.2 Comparison between Kaplan-Meier (K-M) and CSH approach

The objective here was to compute the CIF based on conventional technique (K-M) and then compare the results with the competing risks CSH approach. The K-M plot for death from Covid estimates the survival probability of subjects who have not experienced the death from Covid. The CIF can be obtained by plotting inversely (1-KM), which estimates the cumulative risk of dying from Covid over time, in the absence of the competing events (here, treated all as right-censored times). Overall, the figures (2.11 to 2.20) satisfied the proportional hazards assumption since survival risk curves do not cross during the analyzed period. A clear overestimation of the CIF function over time was observed under the K-M estimation technique compared to the CSH approach. The overestimation gap between (1-KM) and CSH approaches was severe, mainly for death due to covid.

The subjects who developed the exposures of chronic liver disease (hepatic.dis), other symptoms, respiratory discomfort (resp.disc), oxygen saturation level and ICU admission, have lower probability of survival after 20 days of hospitalization than subjects who did not experience these characteristics. In particular, the most severe exposure group was the one who entered in the ICU and they have about 40% less probability of survival after 20 days than those subjects who did not admit in this unit (Figure 2.18). Furthermore, the gap between the CIF curves among the subjects who had a fever and those who had not experienced fever were negligible (Figure 2.12). Moreover, the CIF for the subjects who had been vaccinated was indistinguishable from the CIF of those who had not been vaccinated (Figure 2.13). Additionally, the probability of death due to Covid for male subjects was higher as compared to female subjects (Figure 2.11). However, flu vaccine and sex exposures were found statistically significant with the hazard ratio 0.94 and 1.06, respectively (Table 2.1). The probability of death due to covid was more severe for the age group greater than 70 years (it is $>$ 50% after 30 days, Figure 2.19). For the subjects from the race with indigenous was found a higher probability of dying as compared to other race (Figure 2.20).

FIGURE 2.11: CIF curve for exposure sex

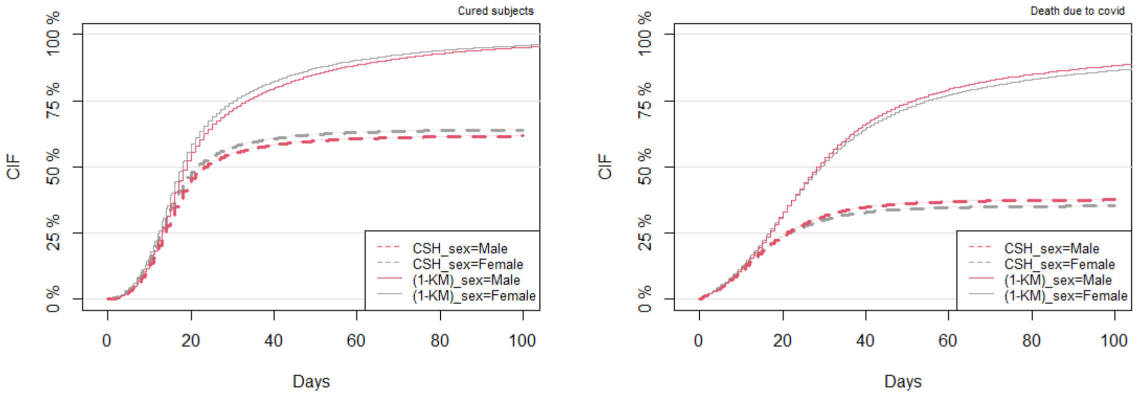


Figure 2.12: CIF curve for exposure fever

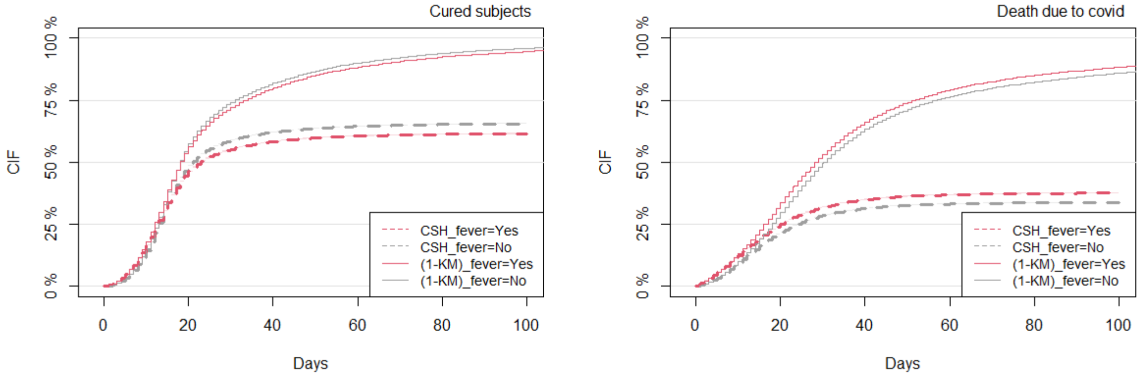


Figure 2.13: CIF curve for exposure flu vaccine

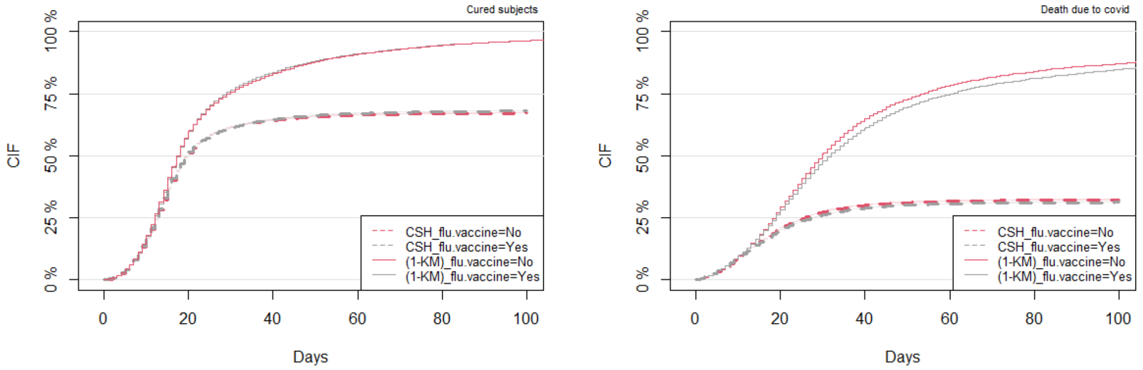


Figure 2.14: CIF curve for exposure hepatic.dis

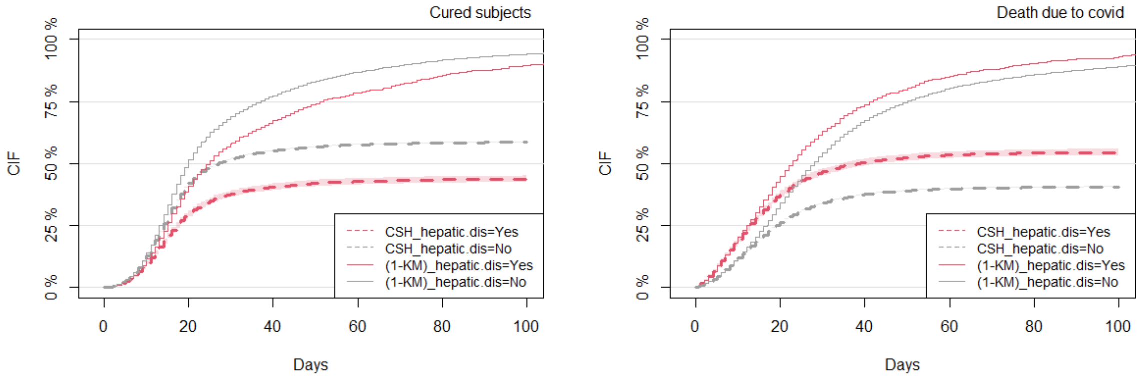


FIGURE 2.15: CIF curve for exposure other.symp

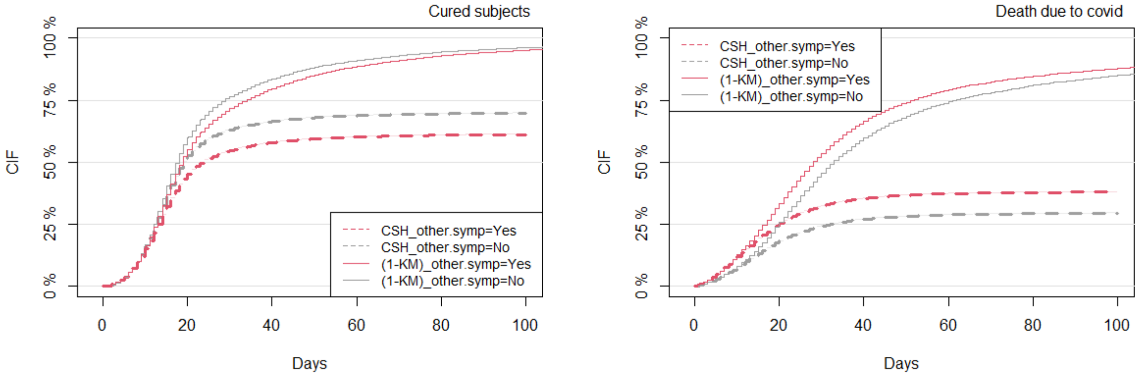


FIGURE 2.16: CIF curve for exposure resp.disc

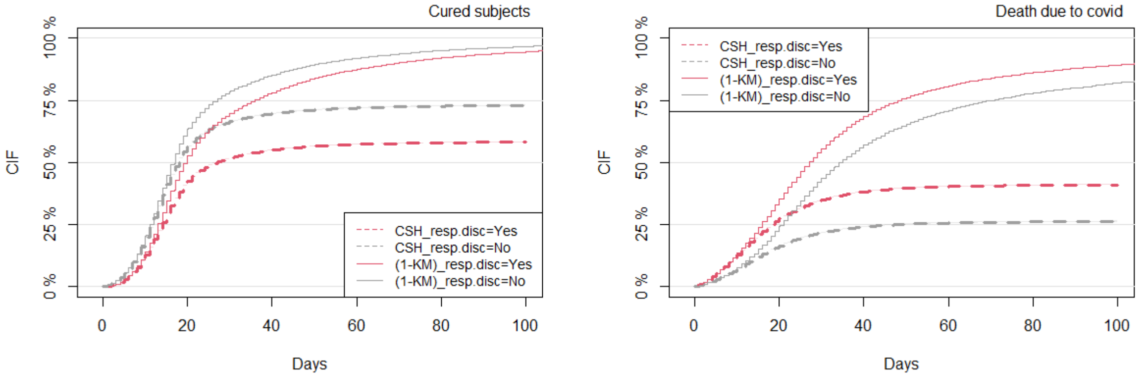


FIGURE 2.17: CIF curve for exposure saturation

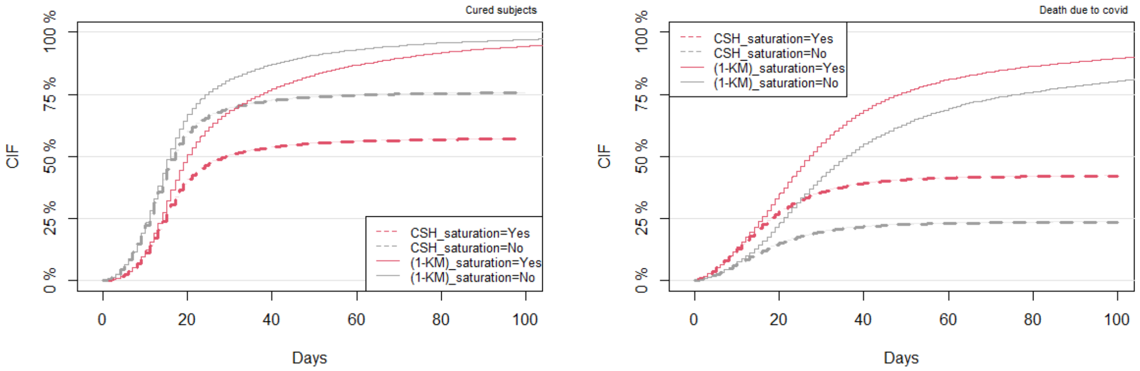


FIGURE 2.18: CIF curve for exposure ICU

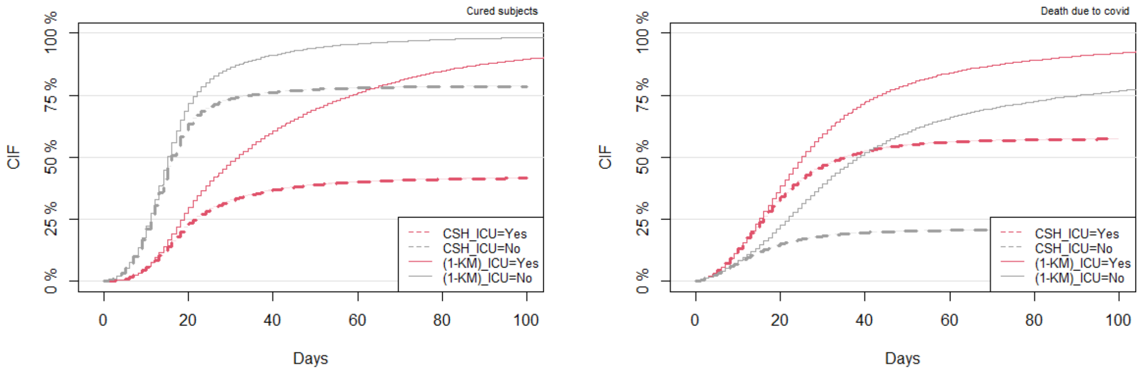


FIGURE 2.19: CIF curve for exposure age

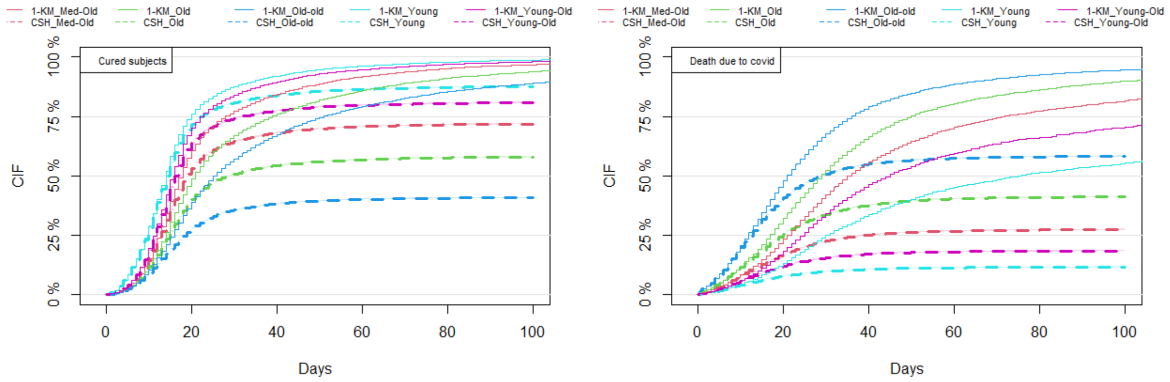
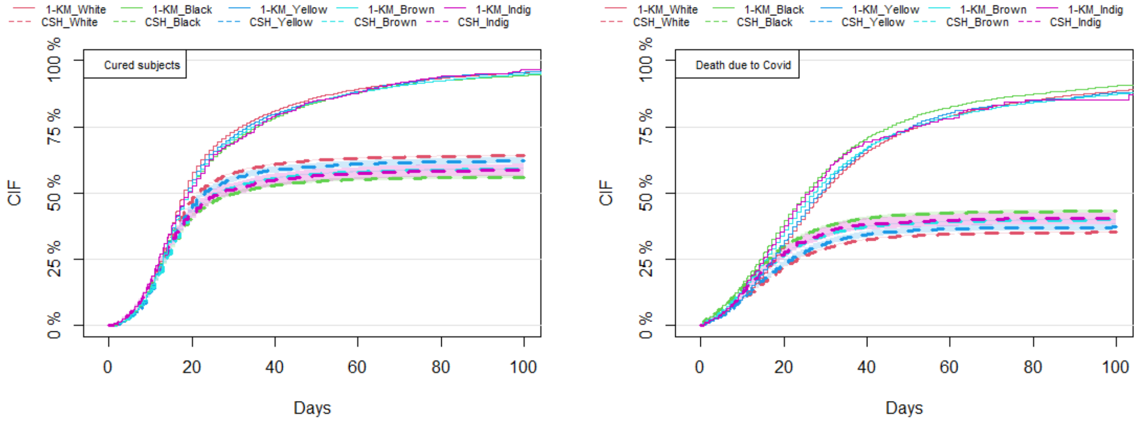


FIGURE 2.20: CIF curve for exposure race



2.3.3 Regression analysis to estimate the parameters

To analyze the effect of the exposures on the CIF, it was found that there were confounding effects among the symptoms and some of the patients' risk factors. Thus, we separated those confounding exposures and investigated the remaining risk factors on the cause of interest. In particular, the stepwise variable selection techniques were applied based on AIC and likelihood ratio test under the Cox proportional hazard assumptions for the CSH, and SDH approaches. The data were analyzed in R statistical software. The final model was as follows: asthma, diabetes, obesity, other.risk, immuno, kidney, neuro, flu.vaccine, hepatic.dis, age, sex, ICU, pneumo, and race.

Regression analysis for the CSH approach

From Table 2.1, the worst outcomes were observed for the age group Old-old (> 70 years) with a hazard ratio around two-folds (HR: 2.03784, CI: $1.9703 - 2.1048$) as compared to the reference group of medium-old age ($50 - 60$ years). Furthermore, the subjects who were admitted to the emergency unit (ICU) had a significantly higher Covid-mortality than those not admitted in the ICU (HR: 1.53741, CI: $1.5046 - 1.5709$). Moreover, the exposures diabetes, other risks, male sex, yellow race (category level 3), had approximately 1.06 hazard ratios. It was found that, the subjects who had been vaccinated were less risky than those who had not been vaccinated (HR: 0.93863, CI: $0.9180 - 0.9598$). Moreover, subjects with a state of decreased immunological system function (Immuno), chronic kidney disease, neurological disease, and chronic liver disease (hepatic.dis1) had approximately 28% higher risk of dying of Covid (HR= 1.28) as compared to those who had no such disease status. Additionally, the risk of dying was higher for the indigenous subjects (HR: 1.37049 with a larger confidence interval: $1.1361 - 1.6532$) than other race subjects. Subsequently, the mortality rate for the black and brown race subjects was also found significantly higher than white race subjects. Interestingly, the subject with the state of asthma was found to have a lower probability of dying (HR: 0.89099, CI: $0.8418 - 0.9430$) compared to those who had no asthma. This may be justified as the subjects with asthma received extra care in the hospital. Thus, the lesser the risk of dying undertaken.

Table 2.1: The estimated hazard ratio of the exposures for the main event of interest (death due to covid) under CSH approach

Exposures	Estimates	HR	S.E.	Lower CI	Upper CI	P-value
asthma1	-0.115	0.891	0.029	0.842	0.943	<0.001
diabetes1	0.077	1.080	0.011	1.058	1.104	<0.001
obesity1	0.034	1.034	0.018	0.998	1.072	0.06.
other.risk1	0.058	1.060	0.011	1.038	1.082	<0.001
immuno1	0.252	1.286	0.024	1.227	1.348	<0.001
kidney1	0.238	1.269	0.019	1.222	1.317	<0.001
neuro1	0.270	1.310	0.019	1.262	1.360	<0.001
flu.vaccine1	-0.063	0.939	0.011	0.918	0.960	<0.001
hepatic.dis1	0.247	1.280	0.040	1.184	1.384	<0.001
age2Old	0.276	1.318	0.018	1.272	1.366	<0.001
age2Old-old	0.712	2.038	0.017	1.973	2.104	<0.001
age2Young	-0.333	0.717	0.031	0.675	0.761	<0.001
age2Young-Old	-0.120	0.887	0.026	0.844	0.933	<0.001
sexM	0.060	1.062	0.011	1.040	1.085	<0.001
ICU1	0.430	1.537	0.011	1.504	1.571	<0.001
pneumo1	0.133	1.143	0.019	1.100	1.187	<0.001
race2 (black)	0.198	1.219	0.023	1.165	1.275	<0.001
race3 (yellow)	0.064	1.066	0.049	0.968	1.174	0.194
race4 (brown)	0.149	1.160	0.011	1.135	1.187	<0.001
race5 (indigenous)	0.315	1.370	0.096	1.136	1.653	<0.001

Regression analysis for the SDH approach

This section explores the performance of the SDH (Fine-Gray model) model. This approach makes it possible to obtain both the naive and model-based standard errors. Here, only model-based (robust) standard errors are reported. It is observed from Table 2.2 that the estimated coefficients for death due to covid deviate slightly from those obtained from the CSH model. The differences in the estimated parameters reflected the different underlying assumptions under competing risks survival data. The estimates derived from the Fine-Gray model have no simple interpretation, but they go in the same direction.

TABLE 2.2: The estimated hazard ratios of the exposures for the main event of interest (death due to covid) under SDH approach

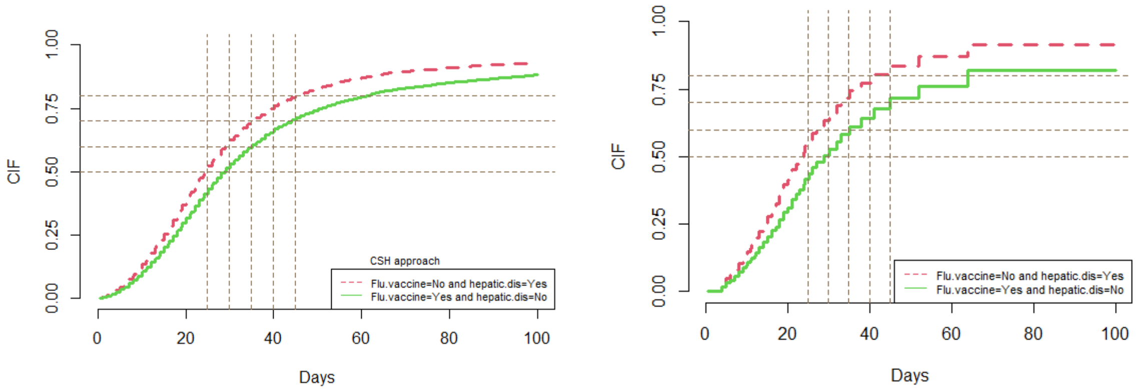
Exposures	Estimates	HR	Robust S.E.	Lower CI	Upper CI	P-value
asthma1	-0.115	0.891	0.029	0.842	0.943	<0.001
diabetes1	0.078	1.081	0.011	1.058	1.105	<0.001
obesity1	0.037	1.037	0.018	1.001	1.075	0.042 *
other.risk1	0.056	1.058	0.011	1.036	1.081	<0.001
immuno1	0.241	1.272	0.025	1.211	1.338	<0.001
kidney1	0.235	1.265	0.021	1.216	1.317	<0.001
neuro1	0.267	1.306	0.021	1.253	1.361	<0.001
flu.vaccine1	-0.062	0.940	0.011	0.919	0.961	<0.001
hepatic.dis1	0.244	1.276	0.044	1.171	1.391	<0.001
age2Old	0.278	1.321	0.017	1.277	1.367	<0.001
age2Old-old	0.710	2.035	0.016	1.971	2.101	<0.001
age2Young	-0.335	0.715	0.030	0.674	0.759	<0.001
age2Young-Old	-0.120	0.887	0.025	0.845	0.931	<0.001
sexM	0.061	1.063	0.011	1.041	1.086	<0.001
ICU1	0.434	1.543	0.011	1.510	1.577	<0.001
pneumo1	0.132	1.141	0.020	1.097	1.188	<0.001
race2 (black)	0.198	1.21932	0.02367	1.1640	1.277	<0.001
race3 (yellow)	0.053	1.05427	0.047	0.961	1.157	0.264
race4 (brown)	0.144	1.155	0.012	1.129	1.181	<0.001
race5 (indigenous)	0.323	1.381	0.106	1.121	1.701	0.002 **

Comparison of model prediction between CSH and SDH approaches

In this section, the model-based predictions are undertaken and compared by analyzing the subject's risk with the given exposures resulting from the fitted regression models. As an illustration, let us predict the risk of a subject with vaccination status and chronic liver disease under CSH, and SDH approaches. Figure 2.21 shows two groups based on the subjects' risk factors: group 1 is those who had been vaccinated and had no chronic liver disease, and group 2 is those who had not been vaccinated and had chronic liver disease. From the Figure 2.21 (left panel), it is observed that, at the beginning of the study, the CIF curves between the two groups seem to be similar until day 10. Then, the gap of the CIF probabilities is increased over time. In particular, in the CSH approach, the probability of death (CIF) due to Covid has reached 50% in

25 days for group 1, and in 30 days for group 2. Moreover, the CIF probability gap is almost similar from 30 days to more than 100 days. On the contrary, in the SDH approach, the CIF probability gap after 10 days increased slightly, and then this gap extends more after 25 days up to 70 days, and after that period, the CIF curves seem flat (see Figure 2.21, right panel).

FIGURE 2.21: Prediction of subjects with flu vaccine and saturation status under CSH (left panel) and SDH approaches (right panel)



In conclusion, it can be said that the exposures of asthma, diabetes, obesity, other.risk, immuno, kidney, neuro, flu.vaccine, hepatic.dis, age, sex, ICU, pneumo, and race, significantly increase the probability of death due to Covid. The highest hazard ratio (2.03) was observed for the subjects with age greater than 70 years compared to the age group $50 - 60$ years. SDH approach shows a slightly higher survival probability compared to the CSH approach. In this study, ICU is considered time-constant exposure, but one can consider ICU a time-dependent covariate. Moreover, it is of interest to study its time-varying effect on the CIF. The reason is that a significant proportion of COVID-19 cases develop pneumonia and acute severe respiratory failure, which commonly require hospitalization, ICU admission, and intubation (Lai *et al.*, 2020). Thus, the binomial regression approach can help to predict such an objective, allowing to model time-varying coefficients.

Chapter 3

Computing sample size using fixed design

3.1 Introduction

To design a randomized clinical trial, an essential step is the calculation of the sample size or the number of patients to be recruited to detect the efficacy of treatments with sufficient power. Many studies have investigated sample size calculations under clinical trials settings with time-to-event endpoints. In particular, Freedman (1982); Schoenfeld (1983); Lakatos (1988); Collett (2003) and others, were discussed the sample size computation under the proportional hazards model. Next, some methods were proposed, under a constant hazard function, i.e., exponential distribution (George and Desu, 1974; Bernstein and Lagakos, 1978; Lachin, 1981; Lachin and Foulkes, 1986) and many others. Very few considered the Weibull distribution (Heo *et al.*, 1998; Wu, 2015).

In a time-to-event study, the sample size is determined not by the number of patients accrued, but by the number of events observed during a specific follow-up period. If the follow-up continues until all patients enrolled in the trial have experienced the event of interest, the required sample size coincides with the number of patients. However, clinical trials often have to be completed within a relatively short time period. Furthermore, one of the main objectives is to study the main event of interest, given that patients can experience competing events. Moreover, in the presence of competing risks (CR), only part of the trial population will experience the event of interest, allowing patients to be censored or fail in competing events. Therefore, in determining the required number of patients, the probability of having the main event over time, denoted here with Ψ, should also be estimated.

In CR settings, Ψ can be calculated using the Cumulative Incidence Function (CIF), which is often of great interest in medical research and can be estimated by different alternative approaches. When there is only one event of interest, then the CIF can be easily calculated with the complementary of the Kaplan-Meier (K-M) estimator. The Cause-Specific Hazard (CSH) or the Sub-Distribution Hazard (SDH) approach can be employed to estimate the CIF in the presence of competing events. Meanwhile, Tai *et al.* (2018) compared sample size calculation for these two different approaches for a fixed design. With this in mind, in this chapter, novel aspects are discussed: the computational methods that can be employed are studied; the comparison between the CSH and SDH approaches are investigated in terms of Weibull and Gompertz distributions, which have not been explored elsewhere.

In Section 3.2, the theoretical aspects of the sample size is described. In particular, the derivation of N is shown in Appendix A.1, derivation of D is in Subsection 3.2.1, and Ψ is in Subsection 3.2.2. Meanwhile, Section 3.3 describes practical guidelines for computing sample size under the fixed design for the CSH and SDH approaches. Next, in Section 3.3.3, a simulation study is undertaken. Section 3.4 discusses a procedure to compute the sample size when changing the shape parameter values in the Weibull and Gompertz distributions. Finally, 3.5 explained the summary results under fixed design.

3.2 Sample size formulation

Let us consider, a simple survival setting with a single event of interest. Here, let D be the number of events required to be observed in the study, and $\tilde{t}$ be the duration of a study. Further, if it is assumed that a is the first time period during which patients are being accrued into the study and $\tilde{f}$ is the follow-up period, during which patients are under observation and no new patients enter the study, then $\tilde{t} = a + \tilde{f}$. Here, it should be noted that if $\tilde{f}$ is small, correspondingly more patients will need to be recruited to achieve a specific number of events. Here, $g(t)$ denotes the density of the entry distribution with $S(t)$ as the survival distribution and $L(t)$ as the distribution of the loss due to the follow-up pattern. $L(t)$ is required because patients entered the study at different calendar times and may be right censored, i.e., they may still be alive when data are collected at the end of the study or may fail to complete the course of the study for reasons unrelated to the event of interest. Then, the general formula for the required number of patients in a survival study is expressed as follows (Latouche *et al.*, 2004; Collett, 2003):

$$N = \frac{D}{\Psi} = \frac{f(\alpha, \tilde{\beta}, \theta)}{p\{S(t), g(t), L(t), a, \tilde{f}\}}, \tag{3.1}$$

where D is obtained as a function $f(\cdot)$ of the type I error probability α, the power $1-\tilde{\beta}$, and the effect size θ, while Ψ is a function $p(\cdot)$ that depends on $S(t), g(t), L(t), a$, and $\tilde{f}$. In this context, let us consider an experimental group E for instance, a group of patients where an experimental treatment is administrated and a control group C. Here, the effect size is usually expressed as either the hazard ratio, (HR), $\theta = \lambda_E(t)/\lambda_C(t), \forall t$, or $\log\theta$, the regression coefficient in a proportional hazards model. The derivation steps of N are shown in Appendix (A.1).

3.2.1 Derivation of D

In survival analysis, the Cox proportional hazards (PH) regression model assumes that the hazard function $\lambda_j(t)$ for the survival time T for patient j, given the predictors $X_{j1}, X_{j2}, \ldots, X_{jk}$, has the following regression formulation:

$$\log\left[\lambda_j(t \mid X)/\lambda_0(t)\right] = \beta_1 X_{j1} + \beta_2 X_{j2} + \ldots + \beta_k X_{jk}$$

where $\lambda_0(t)$ is the baseline hazard. The survival analysis allows the response, the survival time variable t, to be censored. In medical research, one may wish to test the effect of a specific predictor or covariate (say, treatment effect) possibly in the presence of other predictors on the response variable. To consider this, the model can be redefined by relabelling the covariates. For each patient j, the treatments are relabeled as 0 and 1 with x as the treatment label. For other covariates, it is assumed that there is a vector $Y_j = (Y_{j1}, \ldots, Y_{jk})'$ of covariates. The probability that Patient j receives treatment x is P_x, which is independent of $\mathbf{y}_j$. Then, the hazard function for the j^{th} patient is given by

$$\lambda_j(t \mid Z) = \lambda_0(t) \exp\left(\beta_0 X_j + \sum_{i=1}^{K} \beta_i Y_j\right)$$

The null hypothesis is $\mathrm{H}_0 : [\beta_0, \beta_1, \ldots, \beta_k] = [0, \beta_1, \ldots, \beta_k]$, which is tested against alternative $[\beta^*, \beta_1, \ldots, \beta_k]$.

Binary covariate: treatment

First, it is assumed that there is only one binary predictor (treatment variable) in the model. The PH model assumes that the hazard function has the following relationship:

$$\log\left[\lambda(t \mid X)/\lambda_0(t)\right] = \beta X$$

To test the null hypothesis, $\mathrm{H}_0 : \beta = \beta_0 \geq 0$, against $\mathrm{H}_1 : \beta = \beta^* < 0$, the score statistic for the Cox proportional hazards model can be written in a simple form:

$$S^2 = \frac{\left[\sum_{i \in D} (X_i - E_i)\right]^2}{\sum_{i \in D} V_i}$$

where i indexes the ordered death times; $D(i)$ identifies the death set at the i^{th} death time; $E_i = P_E, V_i = P_E(1 - P_E)$, and Σ_D is summations over $D(i)$. Under the null hypothesis, S^2 is treated as a chi-square distributed with one degree of freedom, or equivalently, S is the standard normal with a mean of 0 and variance of 1. To follow the derivation of Schoenfeld (1983), the following is defined:

$$e_i = \frac{\left[\sum_{j \in \mathscr{R}} X_j \exp\left(\beta X_j\right)\right]}{\left[\sum_{j \in \mathscr{R}} \exp\left(\beta X_j\right)\right]}$$

Here, $\mathscr{R}(i)$ identifies the risk set, and $\sum_{\mathscr{R}}$ is the summations over $\mathscr{R}(i)$. The numerator of S can be written as,

$$\sum_{i \in D} (X_i - E_i) = \sum_{i \in D} (X_i - e_i) + \sum_{i \in D} (e_i - E_i)$$

The first term is asymptotically normal with a mean of 0 and variance of $\sum e_j \left(1 - e_j\right)$, where the summation is over D (Cox, 1975; Tsiatis, 1981). Since $\beta \rightarrow 0$, e_j approaches P_E, and the variance $\sum e_j \left(1 - e_j\right)$ approaches $\sum_D P_E(1 - P_E)$, which is DP_CP_E.

Upon expanding the second term in a Taylor series to about $\beta = 0$, this term approaches $\beta \sum E_j(x) \left\{1 - E_j(x)\right\}$. With the same argument of $\beta \rightarrow 0$, E_j approaches P_E and the term $\beta \sum E_j(x) \left\{1 - E_j(x)\right\}$ approaches $\beta \sum_D P_E(1 - P_E)$, which in turns is βDP_CP_E.

By adding the two terms together and dividing by the denominator $\left[\sum_D V_i\right]^{1/2}$, S is asymptotically normal with the mean,

$$\begin{aligned} E(S) &= \frac{\Sigma_D E\left(X_i - e_i\right) + \Sigma_D \left(e_i - E_i\right)}{\sqrt{\Sigma_D V_i}} \\ &= \frac{0 + \beta DP_CP_E}{\sqrt{DP_CP_E}} \\ &= \beta \sqrt{DP_CP_E} \end{aligned}$$

and variance,

$$\begin{aligned} V(S) &= \frac{1}{(\sqrt{\Sigma_D V_i})^2} \Sigma_D \left[V(X_i - E_i)\right] \\ &= \frac{\Sigma_D V(X_i)}{\Sigma_D V_i} \\ &= \frac{D P_C P_E}{D P_C P_E} = 1. \end{aligned}$$

Thus, S is asymptotically normal with unit variance and mean equal to $\beta\,(P_C P_E)^{\frac{1}{2}}$ times the square root of the expected number of deaths on the trial.

So, the following is obtained:

$$S \sim \mathcal{N}(0,1) \quad \text{, under the null hypothesis}$$

and

$$S \sim \mathcal{N}(\beta\sqrt{P_C P_E D}, 1) \quad \text{, under the alternative hypothesis.}$$

From formula 3.1, it is seen that to calculate the number of events, the probability of type I error (significance level, α), power ($1 - \tilde{\beta}$ or the probability of type II error, $\tilde{\beta}$), and the effect size, β should be calculated.

Step (1)

The significance level, α, is set as equal to the probability of rejecting the null hypothesis when it is true:

$$\begin{aligned} \alpha &= P\left(\Sigma_D\left(X_i - E_i\right) < -c \mid H_0\right) \\ &= P\left(\frac{\Sigma_D\left(X_i - E_i\right)}{\sqrt{\Sigma_D V_i}} < -\frac{c}{\sqrt{\Sigma_D V_i}} \mid H_0\right) \\ &= P\left(Z < -\frac{c}{\sqrt{\Sigma_D V_i}}\right) = \Phi\left(-\frac{c}{\sqrt{\Sigma_D V_i}}\right) \\ \text{so } z_{1-\alpha} &= \frac{c}{\sqrt{\Sigma_D V_i}} \\ \text{or} \quad c &= z_{1-\alpha} * \sqrt{\Sigma_D V_i} \end{aligned}$$

Step (2)

The probability of rejecting the null hypothesis when H_1 is true is calculated here. First, the probability of a Type II error is considered as follows:

$$\begin{aligned}
\tilde{\beta} &= P\,(\text{ accept } H_0 \mid H_1) \\
\text{so } 1-\tilde{\beta} &= P\,(\text{ reject } H_0 \mid H_1) \\
&= P\,(\Sigma_D\,(X_i - E_i) < -c \mid H_1) \\
&= P\left(\frac{\Sigma_D\,(X_i - E_i) - \beta^*}{\sqrt{\Sigma_D V_i}} < \frac{-c-\beta^*}{\sqrt{\Sigma_D V_i}}\bigg|_{H_1}\right) \\
&= P\left(Z < \frac{-c-\beta^*}{\sqrt{\Sigma_D V_i}}\right).
\end{aligned}$$

So, it can be said that $z_{\tilde{\beta}} = -z_{1-\tilde{\beta}} = \frac{-c-\beta^*}{\sqrt{\Sigma_D V_i}}$. Now, c from Step (1) is substituted as follows:

$$\begin{aligned}
-z_{1-\tilde{\beta}} &= \frac{-z_{1-\alpha}\sqrt{\sum_D V_i} - \beta P_C P_E D}{\sqrt{\sum_D V_i}} \\
&= z_{1-\alpha} - \frac{\beta P_C P_E D}{\sqrt{P_C P_E D}} \\
&= z_{1-\alpha} - \beta\sqrt{P_C P_E D} \\
\Rightarrow \beta\sqrt{P_C P_E D} &= z_{1-\alpha} + z_{1-\tilde{\beta}} \\
\Rightarrow \sqrt{D} &= \frac{z_{1-\alpha} + z_{1-\tilde{\beta}}}{\beta\sqrt{P_C P_E}}
\end{aligned}$$

$$\Rightarrow D = \frac{\left(z_{1-\alpha} + z_{1-\tilde{\beta}}\right)^2}{(\log\theta)^2\, P_C P_E}. \tag{3.2}$$

where D is the total number of events, P_C and P_E are the proportions of the sample assigned to the control group and experimental group, respectively, and $Z_{1-\alpha}$ and $Z_{1-\tilde{\beta}}$ are standard normal deviates at the desired one-sided significance level α and power $1-\tilde{\beta}$, respectively.

Considering other covariates along with Treatment

Schoenfeld (1983) extended the multivariate model power calculation for binary x_1 to include additional covariates $y_1 \ldots y_k$. The assumption here is that x_1 is independent of $y_1 \ldots y_k$, as would occur if x_1 was randomly assigned in a controlled experiment.

Here, a function $E_j\{.\}$ is defined as:

$$E_j\{g(x,y)\} = \frac{\sum_{k\in R} g\left(x_k, y_k\right)\exp\left(\Sigma\beta_m y_{km}\right)}{\sum_{k\in R}\exp\left(\Sigma\beta_m y_{km}\right)}$$

Here, let $\hat{E}_j$ be E_j with maximum likelihood estimates (assuming $\beta = 0$) replacing the parameters $\{\beta_m\}$. Letting y_i be the i^{th} compound of y, the elements of the $k \times 1$ vector B are defined by

$$\hat{B}_i = N^{-1}\sum_{j\in D}\left(\hat{E}_j\left(xy_i\right) - \hat{E}_j(x)\hat{E}_j\left(y_i\right)\right\}$$

Meanwhile, the elements of the $p \times p$ matrix $\mathbf{M}$ are defined as

$$\hat{M}_{ik} = N^{-1}\sum_{j\in D}\left\{\hat{E}_j\left(y_i y_k\right) - \hat{E}_j\left(y_i\right)\hat{E}_j\left(y_k\right)\right\}$$

The score statistic can then be expressed as

$$\frac{N^{-1/2}\left[\sum_{j\in D}\left(x_j - \hat{E}_j(x)\right)\right]}{\left(N^{-1}\left[\sum_{j\in D}\hat{E}_j(x)\left(1 - \hat{E}_j(x)\right\}\right] - \hat{\mathbf{B}}'\hat{\mathbf{M}}^{-1}\hat{\mathbf{B}}\right)^{1/2}}$$

Here, the term $\hat{\mathbf{B}}'\hat{\mathbf{M}}^{-1}\hat{\mathbf{B}}$ is the effect of the estimation of $\beta_1, \beta_2, \ldots, \beta_k$ on the variance of $x_j - \hat{E}_j(x)$.

Moreover, it is assumed that β is $O\left(N^{-\frac{1}{2}}\right)$. At the start of the trial, the distribution of vectors y will be the same in the two treatment groups. Since $\beta \to 0$, this will remain true for any time, t, so $\hat{E}_j\left(xy_i\right) \to \hat{E}_j(x)\hat{E}_j\left(y_i\right)$. Thus, $\hat{\mathbf{B}} \xrightarrow{p} 0$ and the second term in the denominator of S can be ignored. The term $\hat{\mathbf{B}}$ appears in the Taylor expansion of $\hat{E}(x)$ about $\beta_1, \beta_2, \ldots, \beta_m$, which implies that

$$N^{-1/2}\sum_{j\in D}\left\{\hat{E}_j(x) - E_j(x)\right\} \xrightarrow{p} 0$$

Thus, S can be written as

$$S = \frac{N^{-1/2}\sum_{j\in D}\left(x_j - E_j(x)\right)}{\left[N^{-1}\sum_{j\in D}E_j(x)\left(1 - E_j(x)\right\}\right]^{1/2}}$$

Further, the following can be defined:

$$e_j = \left\{\sum_{k\in R} x_k \exp\left(\beta x_k + \sum_{m=1}^{p}\beta_m y_{km}\right)\right\} / \left\{\sum_{k\in R}\exp\left(\beta x_k + \sum_{m=1}^{p}\beta_m y_{km}\right)\right\}$$

The numerator of S can be written as

$$N^{-1/2}\sum_{j\in D}(x_j - E_j(x)) = N^{-1/2}\sum_{j\in D}(x_j - e_j) + N^{-1}\sum_{j=D}(e_j - E_j(x))$$

Similarly, as depicted before, it can be shown that S is asymptotically normal with unit variance and mean equal to $\beta\,(P_C P_E)^{1/2} \times \sqrt{D}$, which yields a formula to calculate the sample size for the survival endpoints. Additionally, it can be shown that Schoenfeld's argument also works when x_1 is not binary and is independent of other covarites. If the covariates are correlated with the main covariate of interest, x_1, then the formulation is different to adjust sample sizes to preserve power (for more details see Hsieh and Lavori (2000)).

3.2.2 Derivation of Ψ

The quantity Ψ, given that death is the single event of interest, can be formulated as (Collett, 2003) follows:

$$\Psi = \int_0^a P(\text{death} \mid \text{entry at time } t) \times P(\text{entry at time } t)dt, \tag{3.3}$$

where $g(t) = P(\text{entry at time } t)$ and $P(\text{death} \mid \text{entry at time } t)$ is the probability that a subject has the event of death within the time interval $[t, a+\tilde{f}]$, given that the patient entered the study at time t. This is a conditional CIF that can be rewritten as,

$$CIF_1(a+\tilde{f}-t) = P(\tilde{T} \le a+\tilde{f}-t, k=1)$$

where, $\tilde{T} = T - t$ is the survival time given that the patient enters at time t without having an event before t.

Now, let us consider a CR setting with an event of interest (type 1) and a competing event (type 2); then, the indicator $\varepsilon \in \{1,2\}$. Here, the scope is to derive the required sample size in (3.1) for event 1, $N_1 = D_1/\Psi_1$. Then, assuming that a patient enters in the study at time t with an uniform distribution $g(t)$ over the accrual period $[0, a]$ for both treatment and control groups, the cumulative probabilities for the event of interest in each group ($j = C, E$) reduce as follows (Collett, 2003; Schulgen *et al.*, 2005):

$$\Psi_{1\text{j}} = (1/a)\int_0^a CIF_{1j}(a+\tilde{f}-t)dt. \tag{3.4}$$

With a change of variable $a + \tilde{f} - t = u$ in the integration, the equation (3.4) becomes,

$$\Psi_{1\mathrm{j}} = (1/a) \int_{\tilde{f}}^{(a+\tilde{f})} CIF_{1j}(u)du. \tag{3.5}$$

Therefore, the results can be combined to obtain $\Psi_1 = P_{\mathrm{C}}\ \Psi_{1\mathrm{C}} + P_{\mathrm{E}}\ \Psi_{1\mathrm{E}}$.

When using the CSH approach, the CIF for cause 1 is

$$CIF_1\,(t) = \int_0^t \lambda_1(u)e^{-\{\Lambda_1(u)+\Lambda_2(u)\}}du \tag{3.6}$$

which depends on both the cause-specific rates $\lambda_1(t)$ and $\lambda_2(t)$, where $\Lambda_k(t) = \int_0^t \lambda_k(s)ds$, for $k = 1, 2$. Considering both events in the control and experimental groups, for $j = (C, E)$ and $k = (1, 2)$, the following are defined $\lambda_j(t) = \lambda_{1j}(t) + \lambda_{2j}(t)$, and $\tilde{\lambda}_j(t) = \lambda_{2j}(t)/\lambda_{1j}(t)$. To compute sample size, any parametric distribution can be assumed for the two CSHs; as an example, if an exponential distribution is assumed in the control and experimental groups, then since rates are constant over time, equation (3.6) becomes

$$CIF_{1j}(t) = \frac{1 - e^{-t\{\lambda_{1j}(1+\tilde{\lambda}_j)\}}}{1 + \tilde{\lambda}_j} \quad \text{and} \quad CIF_{2j}(t) = \frac{1 - e^{-\frac{t\{\lambda_{1j}[1+(1/\tilde{\lambda}_j)]\}}{[1/\tilde{\lambda}_j]}}}{1 + [1/\tilde{\lambda}_j]}. \tag{3.7}$$

Then, equation (3.5) is

$$\widetilde{\Psi}_{1j} = (\lambda_{1j}/\lambda_j)[1 - (e^{-\lambda_j \tilde{f}} - e^{-\lambda_j(a+\tilde{f})})/(a\lambda_j)]. \tag{3.8}$$

Under the SDH approach, there exists a direct relationship between the CIF and the subdistribution hazard (Fine and Gray, 1999):

$$CIF_{1j}(t) = 1 - S^*_{1j}(t) = 1 - e^{-\Lambda^*_{1j}(t)} = 1 - e^{-\int_0^t \lambda^*_{1j}(u)du}.$$

where λ^*_{1j} is the SDH rate for both groups. This implies

$$\int_0^t \lambda^*_{1j}(u)du = -\log(1 - CIF_{1j}(t)) = g(CIF_{kj}(t)) \tag{3.9}$$

where $g(\cdot)$ is the log link function.

However, the CSH and SDH approaches differ by the definition of the risk set. In the former, the risk set decreases when there is an event from competing cause or censoring, whereas in the latter, patients who failed from an event, other than the one of interest prior to t, remain in the risk set. This difference also affects the computation of sample size. Here, Ψ^*_{1j} can be computed in equation (3.5) based on the survival function at the different time points $\tilde{f}, (0.5a + \tilde{f})$, and $(a + \tilde{f})$, according to Simpson's approximation, as follows:

$$\Psi^*_{1j} = \frac{1}{6}\left[CIF_{1j}\left(\tilde{f}\right) + 4CIF_{1j}\left(0.5a + \tilde{f}\right) + CIF_{1j}\left(a + \tilde{f}\right)\right] \tag{3.10}$$

where $t = (\tilde{f}, (0.5a + \tilde{f}, a + \tilde{f}))$. Equation (3.10) is also applicable in the CSH approach when the exponential distribution is not a valid assumption (Pintilie, 2006).

However, the presence of right censoring, which has been ignored so far will now be considered. Therefore, it is now neceessary to specify the loss to follow-up parametric distribution $L(t)$. When using the CSH approach, for example, it can be assumed that both failure time and loss to follow-up distributions are exponential with the follow-up period $\left[0, a + \tilde{f}\right]$: $\Lambda_k(t) = \lambda_k t$ and $L(t) = 1 - e^{-\tau t}$, where τ is the overall censoring rate. This provides (Lachin and Foulkes, 1986) the followinh equation:

$$\widetilde{\Psi}^{\tau}_{1j} = (\lambda_{1j}/(\lambda_j + \tau))[1 - (e^{-(\lambda_j+\tau)\tilde{f}} - e^{-(\lambda_j+\tau)(a+\tilde{f})})/(a(\lambda_j + \tau))]. \tag{3.11}$$

Under the SDH approach, equation (3.10) can be modified as follows (Latouche *et al.*, 2004):

$$\Psi^{\tau *}_{1j} = \frac{(1-\tau)}{6}\left[CIF_{1j}\left(\tilde{f}\right) + 4CIF_{1j}\left(0.5a + \tilde{f}\right) + CIF_{1j}\left(a + \tilde{f}\right)\right] \tag{3.12}$$

3.3 Practical guidelines for computing sample size

The scope of this study is to derive the required sample size N_1 for the event of interest (type 1) within the CR settings. Under the CSH approach, it is necessary to calculate $\widetilde{N}_1 = \widetilde{D}_1/\widetilde{\Psi}_1$ where $\widetilde{\Psi}_1 = p_C\widetilde{\Psi}_{1C} + p_E\widetilde{\Psi}_{1E}$. $\widetilde{\Psi}_{1j}$, which is the probability of observing an event of type 1. If it is assumed that the time to event follows an exponential distribution, then the quantity $\widetilde{\Psi}_1$ can be derived from equations (3.7) and (3.8) as a function of $\lambda_{kj}(t)$, whereas when the Weibull and Gompertz distributions are assumed, then equation (3.10) can be employed.

Similarly, when using the SDH approach, it is necessary to calculate $N^*_1 = D^*_1/\Psi^*_1$, where $\Psi^*_1 = p_C\Psi^*_{1C} + p_E\Psi^*_{1E}$. The numerator D^*_1 can be computed based on the SDH ratio θ^*_1, and the probabilities Ψ^*_{1j} are obtained from equation (3.10). In the practical guidelines as well as simulations study (Section 3.3.3), it is assumed that the values for CIF_{1E} and CIF_{2E} are known at time point $t = \tilde{t}$, irrespective of any parametric distribution in the CSH and SDH approaches.

3.3.1 The CSH approach

In addition to knowing $CIF_{1E}(\tilde{t})$ and $CIF_{2E}(\tilde{t})$ values, it is necessary to assume the hazard ratio for both events (θ_1, θ_2) for the exponential time-to-event distribution as well as the shape parameters $\tilde{\gamma}$ and $\tilde{\eta}$ with the assumptions of $\tilde{\gamma}_{1E} = \tilde{\gamma}_{1C}, \tilde{\gamma}_{2E} = \tilde{\gamma}_{2C}$, $\tilde{\eta}_{1E} = \tilde{\eta}_{1C}, \tilde{\eta}_{2E} = \tilde{\eta}_{2C}$ in the Weibull and Gompertz distributions, respectively. Below, the steps used to calculate the exponential distribution are discussed:

1. For the known value of θ_1, the $\widetilde{D}_1$ applying formula (3.2) is first computed.

2. To compute $\widetilde{\Psi}_1$, the hazard rates, $\lambda_{1E}(t)$ and $\lambda_{2E}(t)$ are computed following the proposal by Pintilie (2002). Then, the probabilistic relation $CIF_{1j}(t)+CIF_{2j}(t)+S_j(t) = 1$ is applied along with the equation (3.7) to obtain the following:

$$\begin{cases} \lambda_{1E}(t) & = CIF_{1E}(t) \times \frac{-\log(1-CIF_{1E}(t)-CIF_{2E}(t))}{t(CIF_{1E}(t)+CIF_{2E}(t))}. \\ \lambda_{2E}(t) & = CIF_{2E}(t) \times \frac{-\log(1-CIF_{1E}(t)-CIF_{2E}(t))}{t(CIF_{1E}(t)+CIF_{2E}(t))} \end{cases} \tag{3.13}$$

   For given values of CIF_{1E} and CIF_{2E} at $t = \tilde{t}$, equations (3.13) can be computed.

3. Now, the CSH ratios formula, $\theta_1 = \lambda_{1E}(t)/\lambda_{1C}(t)$ and $\theta_2 = \lambda_{2E}(t)/\lambda_{2C}(t)$, is applied to obtain $\lambda_{1C}(t)$ and $\lambda_{2C}(t)$.

4. Plugg-in all the values of $\lambda_{kj}(t)$ for $j = E, C$ and $k = 1, 2$ in equation (3.8), $\widetilde{\Psi}_{1E}$ and $\widetilde{\Psi}_{1C}$ can be computed.

5. Finally, given the proportion of patients allocated to control (p_C) and experimental (p_E) groups, $\widetilde{\Psi}_1 = p_C\widetilde{\Psi}_{1C} + p_E\widetilde{\Psi}_{1E}$. Hence, the sample size is obtained from $\widetilde{N}_1 = \widetilde{D}_1/\widetilde{\Psi}_1$.

Now, the steps followed in the Weibull time-to-event distribution are discussed. Here, it is assumed that the CSH rates for event types 1 and 2 for $j = (E, C)$ are

$$\lambda_{1j}(t) = \lambda_{1j}\tilde{\gamma}_1 t^{\tilde{\gamma}_{1j}-1} \text{ and } \lambda_{2j}(t) = \lambda_{2j}\tilde{\gamma}_{2j} t^{\tilde{\gamma}_{2j}-1}$$

with the hazard ratios, $\theta_1(t) = \lambda_{1E}(t)/\lambda_{1C}(t) = \lambda_{1E}\tilde{\gamma}_{1E}t^{\tilde{\gamma}_{1E}-1}/\lambda_{1C}\tilde{\gamma}_{1C}t^{\tilde{\gamma}_{1C}-1}$ and $\theta_2(t) = \lambda_{2E}(t)/\lambda_{2C}(t) = \lambda_{2E}\tilde{\gamma}_{2E}t^{\tilde{\gamma}_{2E}-1}/\lambda_{2C}\tilde{\gamma}_{2C}t^{\tilde{\gamma}_{2C}-1}$, respectively. Assuming that $\tilde{\gamma}_{1E} = \tilde{\gamma}_{1C}$ and $\tilde{\gamma}_{2E} = \tilde{\gamma}_{2C}$, the hazard ratios become, $\theta_1(t) = \lambda_{1E}/\lambda_{1C} = \theta_1$ and $\theta_2(t) = \lambda_{2E}/\lambda_{2C} = \theta_2$. This implies that, the hazard ratios are constant over time (proportional hazard assumptions). Thus, the values of $\lambda_{1j}, \lambda_{2j}$, $j = (C, E)$ can be found from steps $(2-3)$ of the procedure from the exponential distribution.

1. $\widetilde{D}_1$ is computed using the value of θ_1 from exponential distribtuion, i.e., $\widetilde{D}_1$ is similar in the exponential and Weibull distributions.

2. To compute $\widetilde{\Psi}_1$, the $CIFs$ values for the three time points following equation (3.10) need to be computed. However, the values at time $t = \tilde{t}$ are only known. Thus, for the given values of $CIF_{1E}(\tilde{t}), CIF_{2E}(\tilde{t}), \lambda_{1E}$, and λ_{2E}, the unknown parameters at other time points can be computed. Let us find $\tilde{\gamma}_{1E}$ and $\tilde{\gamma}_{2E}$ solutions to the system of equations are as follows:

$$\begin{cases} CIF_{1E}(\tilde{t}) - \int_0^t (\lambda_{1E}\tilde{\gamma}_{1E}u^{\tilde{\gamma}_{1E}-1}) \exp\left[-\left(\lambda_{1E}u^{\tilde{\gamma}_{1E}} + \lambda_{2E}u^{\tilde{\gamma}_{2E}}\right)\right] du = 0. \\ CIF_{2E}(\tilde{t}) - \int_0^t (\lambda_{2E}\tilde{\gamma}_{2E}u^{\tilde{\gamma}_{2E}-1}) \exp\left[-\left(\lambda_{2E}u^{\tilde{\gamma}_{2E}} + \lambda_{2E}u^{\tilde{\gamma}_{2E}}\right)\right] du = 0. \end{cases} \tag{3.14}$$

This system has no analytical solution and thus requires the use of the numerical integration technique.

3. With the assumptions of $\tilde{\gamma}_{1E} = \tilde{\gamma}_{1C}, \tilde{\gamma}_{2E} = \tilde{\gamma}_{2C}$ and the values of $\lambda_{1C}, \lambda_{2C}$ found using the exponential distribution procedure, the unknown quantities of CIF_{1C} and CIF_{2C} can be computed at the desired t as follows:

$$CIF_{1C}(t) = \int_0^t (\lambda_{1C}\tilde{\gamma}_{1C}u^{\tilde{\gamma}_{1C}-1}) \exp\left[-\left(\lambda_{1C}u^{\tilde{\gamma}_{1C}} + \lambda_{2C}u^{\tilde{\gamma}_{2C}}\right)\right] du. \tag{3.15}$$

$$CIF_{2C}(t) = \int_0^t (\lambda_{2C}\tilde{\gamma}_{2C}u^{\tilde{\gamma}_{2C}-1}) \exp\left[-\left(\lambda_{1C}u^{\tilde{\gamma}_{1C}} + \lambda_{2C}u^{\tilde{\gamma}_{2C}}\right)\right] du. \tag{3.16}$$

4. CIF_{1E}, CIF_{1C} values at time points $t = \tilde{f}$, and $t = (0.5a + \tilde{f})$ can be computed by using equations (3.14) and (3.15). For instance, the CIF_{1E} at time point $(0.5a + \tilde{f})$ can be computed by rewriting the equation (3.14) as,

$$CIF_{1E}(0.5a + \tilde{f}) = \int_0^{0.5a+\tilde{f}} (\lambda_{1E}\tilde{\gamma}_{1E}u^{\tilde{\gamma}_{1E}-1}) \exp\left[-\left(\lambda_{1E}u^{\tilde{\gamma}_{1E}} + \lambda_{2E}u^{\tilde{\gamma}_{2E}}\right)\right] du.$$

5. Step 5 of the exponential distribution should be repeated here.

Finally, the steps in the Gompertz time-to-event distribution will be explained below. The CSH rates for event types 1 and 2 with the shape parameter $\tilde{\eta}$ for $j = (E, C)$ are

$$\lambda_{1j}(t) = \lambda_{1j}e^{\tilde{\eta}_{1j}t} \text{ and } \lambda_{2j}(t) = \lambda_{2j}e^{\tilde{\eta}_{2j}t}$$

with the hazard ratios, $\theta_1(t) = \lambda_{1E}e^{\tilde{\eta}_{1E}t}/\lambda_{1C}e^{\tilde{\eta}_{1C}t}$ and $\theta_2(t) = \lambda_{2E}e^{\tilde{\eta}_{2E}t}/\lambda_{2C}e^{\tilde{\eta}_{2C}t}$. These hazard ratios are constant over time with the assumptions of $\tilde{\eta}_{1E} = \tilde{\eta}_{1C}$ and $\tilde{\eta}_{2E} = \tilde{\eta}_{2C}$, i.e., $\theta_k(t) = \theta_k = \lambda_{kE}/\lambda_{kC}$ for $k = (1, 2)$. Thus, with the assumption of the proportional hazard, the values of λ_{1j} and λ_{2j} for $j = (E, C)$ can be found assuming that θ_1, θ_2 are the rates in an exponential distribution (see exponential procedure steps $2 - 3$).

1. As before, the values of $CIF_{1E}(\tilde{t})$ and $CIF_{2E}(\tilde{t})$ are already known.

2. $\widetilde{D}_1$ is computed by applying the value of θ_1 from the exponential distribution. Thus, under the CSH approach, the number of events are the same for the exponential, Weibull, and Gompertz time-to-event distributions. However, the $\tilde{\Psi}_1$ parameter differs among them.

3. Now, $\tilde{\Psi}_1$ is computed using the Gompertz time-to-event distribution. Given the values for $CIF_{1E}(\tilde{t}), CIF_{2E}(\tilde{t}), \lambda_{1E}$, and λ_{2E}, the unknown parameters of $\tilde{\eta}_{1E}, \tilde{\eta}_{2E}$ can be computed by applying the numerical integration technique to the following system of equations:

$$\begin{cases} CIF_{1E}(\tilde{t}) - \int_0^{\tilde{t}} (\lambda_{1E}e^{\tilde{\eta}_{1E}u}) \exp\left[-\left\{\frac{\lambda_{1E}}{\tilde{\eta}_{1E}}\left(e^{\tilde{\eta}_{1E}u} - 1\right) + \frac{\lambda_{2E}}{\tilde{\eta}_{2E}}\left(e^{\tilde{\eta}_{2E}u} - 1\right)\right\}\right] du = 0 \\ CIF_{2E}(\tilde{t}) - \int_0^{\tilde{t}} (\lambda_{2E}e^{\tilde{\eta}_{2E}u}) \exp\left[-\left\{\frac{\lambda_{1E}}{\tilde{\eta}_{1E}}\left(e^{\tilde{\eta}_{1E}u} - 1\right) + \frac{\lambda_{2E}}{\tilde{\eta}_{2E}}\left(e^{\tilde{\eta}_{2E}u} - 1\right)\right\}\right] du = 0. \end{cases} \tag{3.17}$$

4. Given the values of λ_{kj} that were found using the exponential distribution and $\tilde{\eta}_{kj}$, the unknown quantities of $CIF_{1C}(t)$ and $CIF_{2C}(t)$ can be computed as follows:

$$CIF_{1C}(t) = \int_0^t (\lambda_{1C}e^{\tilde{\eta}_{1C}u}) \exp\left[-\left\{\frac{\lambda_{1C}}{\tilde{\eta}_{1C}}\left(e^{\tilde{\eta}_{1C}u} - 1\right) + \frac{\lambda_{2C}}{\tilde{\eta}_{2C}}\left(e^{\tilde{\eta}_{2C}u} - 1\right)\right\}\right] du \tag{3.18}$$

$$CIF_{2C}(t) = \int_0^t (\lambda_{2C}e^{\tilde{\eta}_{2C}u}) \exp\left[-\left\{\frac{\lambda_{1C}}{\tilde{\eta}_{1C}}\left(e^{\tilde{\eta}_{1C}u} - 1\right) + \frac{\lambda_{2C}}{\tilde{\eta}_{2C}}\left(e^{\tilde{\eta}_{2C}u} - 1\right)\right\}\right] du \tag{3.19}$$

5. To compute $\widetilde{\psi}_{1j}$ for all the three time points, step 4 as per the Weibull distribution should be followed and thus equations (3.17), (3.18) should be recomputed to obtain $CIF_{1j}(0.5a + \tilde{f})$ and $CIF_{1j}(\tilde{f})$.

6. Step 5 of the exponential distribution should be repeated to compute sample size.

3.3.2 The SDH approach

As discussed in the Section 3.2 that the computation of hazard ratio is different under the SDH and the CSH approaches, the steps to compute D_1^* and Ψ_{1j}^* are also different. Further, the SDH ratio θ_1^* is computed using the following formula:

$$\theta_1^* = \frac{\int_0^t \lambda_{1E}^*(u)du}{\int_0^t \lambda_{1C}^*(u)du} = \frac{-\log\{1 - CIF_{1E}(t)\}}{-\log\{1 - CIF_{1C}(t)\}}. \tag{3.20}$$

Here, it should be noted that, the value for $CIF_{1E}(t)$ is known irrespective of any parametric distribution. However, $CIF_{1C}(t)$ has three possible values depending on three time-to-event distributions (exponential, Weibull and Gompertz), and thus three different θ_1^* values are obtained when using the SDH approach. Hence, there are three different D_1^*. This was not the case in the CSH approach because here a known value for θ_1 was assumed to directly obtain $\tilde{D}_1$. Meanwhile, to obtain Ψ_{1j}^*, equation (3.10) is applied and $CIF_{1j}(t)$ values are solved for at three time points $\tilde{f}, (0.5a + \tilde{f})$ and $\tilde{t}$ for all the distributions. As previously shown, equation (3.10) was also applied using the CSH approach for Weibull and Gompertz time-to-event distributions. However, now the $CIF_{1j}(t)$ values are different than those in that approach because of the relation in equation (3.9). Here, it should be noted that only the values of CIF_{1E} and CIF_{2E} are known at time point $t = \tilde{t}$. To compute the CIF_{1E} values at other time points, equation (3.9) is inverted as

$$CIF_{1j}(t) = 1 - exp[-\int_0^t \lambda_{1j}^*(u)du], \tag{3.21}$$

where $\lambda_{1j}^*(t)$ is the SDH rate, which has different values for different time-to-event distributions. Now, the steps to compute D_1^* and Ψ_{1j}^* in the exponential time-to-event distribution are explained.

1. Assuming that $CIF_{1E}(\tilde{t})$ is known as is $CIF_{1C}(\tilde{t})$. If not, the latter is found by applying the CSH procedure using exponential distribution. In this case, θ_1 and θ_2 using the CSH approach would need to be given in advance (known).

2. Suppose $CIF_{1E}(\tilde{t})$ and $CIF_{1C}(\tilde{t})$ are known, then by plugging-in these values in equation (3.20), θ_1^* is obtained. Thus, D_1^* can be computed by applying formula (3.2).

3. To compute Ψ_{1j}^*, first equation (3.21) has to be solved by assuming the exponential distribution, i.e., $\lambda_{1j}^*(t) = \lambda^*$. Then, $CIF_{1j}(t) = 1 - e^{-\lambda^* t}$, which implies $\lambda^* =$

$-\log(1-CIF_{1j}(t))/t$. Now, for given values of CIF_{1E}, CIF_{1C} at time point $t = \tilde{t}$, λ^*_{1E} and λ^*_{1C} are obtained.

4. CIF_{1j} is recomputed for time points $\tilde{f}$ and $(0.5a + \tilde{f})$,

$$CIF_{1j}(\tilde{f}) = 1 - e^{-\lambda^* \tilde{f}}, \qquad CIF_{1j}(0.5a + \tilde{f}) = 1 - e^{-\lambda^*(0.5a+\tilde{f})}$$

5. Then, all the $CIF_{1j}(t)$ values are plugged in in equation (3.10) to obtain Ψ^*_{1j}.

6. Finally, given the proportion of patients allocated to control (p_C) and experimental (p_E) groups, $\Psi^*_1 = p_C\Psi^*_{1C}+p_E\Psi^*_{1E}$ is computed. Hence, the sample size is obtained from $N^*_1 = D^*_1/\Psi^*_1$.

Now, the steps in the Weibull time-to-event distribution will be described.

1. Assuming that $CIF_{1E}(\tilde{t})$ is known as is $CIF_{1C}(\tilde{t})$. If the latter is not, it is found by applying the CSH procedure using the Weibull distribution. In this case, θ_1 and θ_2 as per the CSH approach would again need to be given. In addition, the shape parameters γ_{1E}, γ_{2E} have to be computed. Finally, it is assumed that $\gamma_{1E} = \gamma_{1C}$.

2. Plugging-in $CIF_{1E}(\tilde{t})$ and $CIF_{1C}(\tilde{t})$ values in equation (3.20), θ^*_1 is obtained and thus D^*_1 is computed by applying formula (3.2).

3. To compute Ψ^*_{1j}, the SDH rate is assumed as $\lambda^*_{1j}(t) = \lambda^*\gamma^*t^{\gamma^*-1}$, where λ^* and γ^* are scale and shape parameters, respectively. Then, upon integrating this function in equation (3.21), $CIF_{1j}(t) = 1 - e^{-\lambda^* t^{\gamma^*}}$, which implies $\lambda^* = -\log(1 - CIF_{1j}(t))/t^{\gamma^*}$. Here, it is necessary to assume a fixed value for γ^*. Thus, for a given value of γ^* and CIF_{1j} at time $t = \tilde{t}$, λ^* is obtained.

4. CIF_{1j} is then recomputed for time points $\tilde{f}$ and $(0.5a + \tilde{f})$,

$$CIF_{1j}(\tilde{f}) = 1 - e^{-\lambda^* \tilde{f}^{\gamma^*}}, \qquad CIF_{1j}(0.5a + \tilde{f}) = 1 - e^{-\lambda^*(0.5a+\tilde{f})^{\gamma^*}}$$

5. Then, all the $CIF_{1j}(t)$ values are plugged in equation (3.10) to obtain Ψ^*_{1j}.

6. Step 7 of the SDH approach for the exponential distribution is then repeated.

Finally, for the Gompertz distribution, the steps are as follows:

1. Assuming that $CIF_{1E}(\tilde{t})$ is known as is $CIF_{1C}(\tilde{t})$. If the latter is not, it is found by applying the CSH procedure using the Gompertz distribution. In this case, as per the CSH approach, θ_1 and θ_2 need to be given again. In addition, the shape parameters η_{1E}, η_{2E} require computing. Finally, it is assumed that $\eta_{1E} = \eta_{1C}$.

2. Then, the $CIF_{1E}(\tilde{t})$ and $CIF_{1C}(\tilde{t})$ values are plugged in equation (3.20) to obtain θ_1^*. D_1^* is computed by applying formula (3.2).

3. To compute Ψ_{1j}^*, it is assumed that $\lambda_{1j}^*(t) = \lambda^* e^{\eta^* t}$, where λ^* and η^* are the scale and shape parameters, respectively. Then, applying equation (3.21) is applied to obtain

$$CIF_{1j}(t) = 1 - exp\left[- \frac{\lambda^*(e^{\eta^* t} - 1)}{\eta^*} \right]. \tag{3.22}$$

Then,

$$\lambda^* = \frac{\eta^*[-\log(1 - CIF_{1j}(t))]}{e^{\eta^* t} - 1}.$$

Again, here, it is necessary to assume a fixed value for η^*. Thus, for given values of η^* and $CIF_{1j}(t)$, λ^* is obtained.

4. Then, the values of λ^*, η^* are plugged in equation (3.22) and considering the time points $t = \tilde{f}, t = (0.5a + \tilde{f})$ and $t = \tilde{t}$, $CIF_{1j}(\tilde{f}), CIF_{1j}(0.5a + \tilde{f})$, and $CIF_{1j}(\tilde{t})$ can be computed.

5. All the $CIF_{1j}(t)$ values are plugged in equation (3.10) to obtain Ψ_{1j}^*.

6. Step 7 of the SDH approach for exponential distribution is then repeated.

3.3.3 Simulation results

The simulation study was conducted for the CSH and SDH approaches using the exponential, Weibull and Gompertz time-to-event distributions following the guidelines in Section 3.3. An R package `rootSolve` was used to compute the shape parameter values. Here, the shape parameters for the Weibull and Gompertz distributions were denoted as $\tilde{\gamma}$, and $\tilde{\eta}$, respectively, as per the CSH approach. Similarly, for the SDH approach the shape parameters for aforementioned distributions were denoted as γ^* and η^*, respectively. To compute results in Table 3.1, the following values were assumed for event types 1 and 2: $CIF_{1E} = (0.1, 0.2, 0.3); CIF_{2E} = 0.1$; hazard ratios: $\theta_1 = (0.8, 0.6, 0.4), \theta_2 = (0.8, 1.0, 1.2)$; total study period: $\tilde{t} = (a + \tilde{f}) = (2 + 2) = 4$ years; type I and II error rates: $\alpha = 0.05, \tilde{\beta} = 0.20$. Then, the following were computed: CIF_{1C} at time points $t = (\tilde{f}, 0.5a + \tilde{f}, \tilde{t})$, CIF_{1E} at time points $(\tilde{f}, 0.5a + \tilde{f})$, and SDH ratio θ_1^*. Finally, $(\tilde{D}, \tilde{\psi})$ and (D_1^*, ψ_1^*) were obtained using the CSH and SDH approaches, respectively.

Comparing sample sizes between the CSH and SDH approaches

Using the general overview of all the results reported in Table 3.1, it can be noted that when a lower number of events occured in both event types and groups, (with consequently lower CIFs) and the hazard ratio for the main event was 0.80 (first three rows of the three scenarios in Table 3.1), then the required sample size is very high for all distributions. Further, when $CIF_{jk} < 0.13$, for $j = (C, E)$ and $k = (1, 2)$, then the required sample size increased substantially in comparison to that with a very high positive effect of treatment ($\theta_1 < 0.80$). Moreover, the computed sample size under CSH and SDH approaches are very similar for the three parametric distributions, with slightly systematically lower values for the Gompertz. For comparison purposes in Table 3.1, the same shape parameter values were used for the CSH and SDH approaches. Here, it should be noted that distribution is a special case of Weibull distribution when $\gamma = 1$. Thus, with a very small change in $\gamma = 0.95$, slight reductions were observed in the sample size for both approaches.

For a fixed $\theta_1 (e.g., \theta_1 = 0.80)$, assuming an increased number of type 1 events in the experimental group (i.e., a number that leads from $CIF_{1E} = 0.10$ to $CIF_{1E} = 0.20$), there is approximately a two-fold increase in the CIF_{1C} and thus, the required sample size is nearly halved. Likewise an even larger CIF_{1E} (e.g., $CIF_{1E} = 0.30$) provides even smaller sample sizes (from about 3600 to 2400). Similar conclusions can be stated by comparing all results for $\theta_1 = 0.60$ or all results for $\theta_1 = 0.40$ in Table 3.1.

For $\theta_2 \leq 1$, the CSH approach provides lower sample sizes as compared to the SDH, and is thus preferable. However, when $\theta_2 > 1$ (e.g., $\theta_2 = 1.20$), the sample size is lower in the SDH approach, which is instead preferable.

TABLE 3.1: Sample sizes using the CSH and SDH approaches with different event time distributions

Here, $(a, \tilde{f})$ indicates accrual and follow up duration; CIF$_{1E}$, CIF$_{2E}$, CIF$_{1C}$, CIF$_{2C}$ indicate CIF for main event and competing event for the experimental and control groups respectively; θ is the CSH ratio; θ^* is the SDH ratio; γ, η indicate shape parameter for the Weibull and Gompertz distributions respectively. $\tilde{N}_{1_{exp}}$ represents sample size when using the CSH approach whereas $N^*_{1_{exp}}$ indicates sample size under the SDH approach for the exponential distribution for event type 1 and so on. Meanwhile, $CIF_{1C_{Weib}} = CIF_{1C_{Gom}}$ indicate the value of CIF_{1C} at time point $t = \tilde{t}$ in the Weibull and Gompertz distributions, respectively.

			CSH Approach					SDH Approach				
$\tilde{t} = a + \tilde{f} = (2+2)$ $\alpha = 0.025$(one-sided), $\tilde{\beta} = 0.20$	$CIF_{1C_{exp}}$	$CIF_{1C_{Weib}}$ $= CIF_{1C_{Gom}}$	θ_1	θ_2	$\tilde{N}_{1_{exp}}$	$\tilde{N}_{1_{Weib}}$	$\tilde{N}_{1_{Gom}}$	θ^*_{1exp}	θ^*_{1Weib} $= \theta^*_{1Gom}$	$N^*_{1_{exp}}$	$N^*_{1_{Weib}}$	$N^*_{1_{Gom}}$
$CIF_{1E} = 0.10$	0.122	0.121	0.80	0.80	7379	7283	7270	0.81	0.82	8584	8473	8441
$CIF_{2E} = 0.10$	0.123	0.121	0.80	1.00	7337	7241	7228	0.80	0.82	7465	7368	7341
$\gamma_{1E} = 0.9475$	0.125	0.121	0.80	1.20	7309	7213	7200	0.79	0.82	6830	6741	6716
$\gamma_{2E} = 0.9475$	0.159	0.161	0.60	0.80	1203	1187	1185	0.61	0.60	1300	1283	1278
$\eta_{1E} = -0.04$	0.161	0.161	0.60	1.00	1195	1179	1177	0.60	0.60	1217	1201	1197
$\eta_{2E} = -0.04$	0.161	0.161	0.60	1.20	1190	1174	1172	0.60	0.60	1165	1150	1146
	0.227	0.232	0.40	0.80	293	289	289	0.41	0.40	309	305	304
	0.232	0.232	0.40	1.00	291	287	286	0.40	0.40	296	293	292
	0.232	0.232	0.40	1.20	289	286	285	0.40	0.40	288	285	284
$CIF_{1E} = 0.20$	0.241	0.241	0.80	0.80	3663	3657	3656	0.81	0.81	4366	4355	4357
$CIF_{2E} = 0.10$	0.243	0.243	0.80	1.00	3641	3635	3634	0.80	0.81	3734	3725	3726
$\gamma_{1E} = 0.9933$	0.246	0.249	0.80	1.20	3626	3621	3620	0.79	0.78	3381	3373	3375
$\gamma_{2E} = 0.9933$	0.306	0.311	0.60	0.80	604	603	603	0.61	0.60	665	663	663
$\eta_{1E} = -0.005$	0.311	0.311	0.60	1.00	600	599	599	0.60	0.60	617	616	616
$\eta_{2E} = -0.005$	0.311	0.311	0.60	1.20	598	597	597	0.60	0.60	588	586	586
	0.420	0.420	0.40	0.80	151	151	151	0.41	0.41	163	162	162
	0.428	0.428	0.40	1.00	150	150	150	0.40	0.40	155	155	155
	0.428	0.428	0.40	1.20	149	149	149	0.40	0.40	150	150	150
$CIF_{1E} = 0.30$	0.353	0.349	0.80	0.80	2422	2402	2400	0.82	0.83	2983	2962	2957
$CIF_{2E} = 0.10$	0.360	0.360	0.80	1.00	2407	2388	2386	0.80	0.80	2496	2478	2474
$\gamma_{1E} = 0.967$	0.363	0.360	0.80	1.20	2397	2379	2377	0.79	0.80	2231	2215	2211
$\gamma_{2E} = 1.05$	0.437	0.437	0.60	0.80	405	401	401	0.62	0.62	456	453	452
$\eta_{1E} = -0.023$	0.448	0.448	0.60	1.00	402	399	399	0.60	0.60	419	416	415
$\eta_{2E} = 0.034$	0.448	0.448	0.60	1.20	400	398	397	0.60	0.60	396	393	393
	0.572	0.581	0.40	0.80	104	103	103	0.42	0.41	115	114	114
	0.581	0.581	0.40	1.0	103	103	102	0.41	0.41	109	108	108
	0.590	0.590	0.40	1.20	103	102	102	0.40	0.40	105	104	104

Study of the power

In this section, the objective is to study and compare power under the CSH and SDH approaches for exponential, Weibull, and Gompertz distributions. To visualize results in Figures 3.1, 3.2, and 3.3, the hazard ratios were as assumed as follows- event type

1: $\theta_1 = (0.8, 0.6, 0.4)$; event type $2 : \theta_2 = (0.8, 1.2)$. Then, three scenarios of CIF were assumed for the experimental group: (a) $CIF_{1E} = 0.10, CIF_{2E} = 0.10$; (b) $CIF_{1E} = 0.20, CIF_{2E} = 0.10$; and (c) $CIF_{1E} = 0.30, CIF_{2E} = 0.10$. The total study duration was assumed to be 4 years ($a = 2, \tilde{f} = 2$). Next, 1000 sample sizes were simulated and the shape parameters for Weibull and Gompertz distributions as per the CSH approach were obtained for the aforementioned three scenarios: $\tilde{\gamma}_{1E} = \tilde{\gamma}_{2E} = 0.9475$ and $\tilde{\eta}_{1E} = \tilde{\eta}_{2E} = -0.04$ for the first scenario; $\tilde{\gamma}_{1E} = \tilde{\gamma}_{2E} = 0.9933$ and $\tilde{\eta}_{1E} = \tilde{\eta}_{2E} = -0.005$ for the second scenario; $\tilde{\gamma}_{1E} = 0.967, \tilde{\gamma}_{2E} = 1.05$, and $\tilde{\eta}_{1E} = -0.023, \tilde{\eta}_{2E} = 0.034$ for the third scenario. For the SDH approach, the same shape parameter values were assumed for event type $1 : \gamma^*_{1E} = 0.9475, \eta^*_{1E} = -0.04, \gamma^*_{1E} = 0.9933, \eta^*_{1E} = -0.005, \gamma^*_{1E} = 0.967, \eta^*_{1E} = -0.023$ for the three different scenarios, respectively.

Using the general overview of all the results depicted in Figures 3.1, 3.2, and 3.3, it was found that, when the positive treatment effect on the competing event, i.e., $\theta_2 = 0.8$ is considered, then most of the plots show that the CSH approach performs better than the SDH. However, when θ_2 was moved from the positive to adverse effect (i.e., a value that leads θ_2 from 0.8 to 1.2), then all the figures showed that the SDH approach is as good as CSH. Furthermore, for $\theta_1 = 0.60$, when the number of type 1 events was increased in the experimental group (i.e., CIF_{1E} values from 0.10 to 0.30), then the sample size requirements reduced substantially. For instance, in Figure 3.1, for $\theta_2 = 1.2$, the sample size reduced from 1000 to 400 to achieve 80% power.

Figure 3.1 shows the power computation for exponential distribution in both approaches. In the first scenario ($CIF_{1E} = 0.10, CIF_{2E} = 0.10$) when $\theta_2 = 0.8$, with a very large postive effect of treatment on event type 1 (i.e., $\theta_1 = 0.4$), the required sample size is 300 to keep 80% power for both approaches. This sample size increases considerably to more than 1000 when $\theta_1 = 0.6$. When moving from a positive effect ($\theta_2 = 0.8$) to an adverse effect ($\theta_2 = 1.2$) of treatment on the competing event, the required sample size is slightly changed for $\theta_1 = 0.4$ (i.e., 275). Meanwhile, for $\theta_1 = 0.6$, the sample size is about 1000. In the second scenario ($CIF_{1E} = 0.20, CIF_{2E} = 0.10$), when $\theta_2 = 0.8$ and then for $\theta_1 = 0.4$, the required sample size is less than 200 for both approaches; when $\theta_1 = 0.6$, then a sample size of 600 is required for the CSH approach and about 700 for the SDH; when $\theta_2 = 1.2$, then both approaches require a sample size of 600. In the third scenario ($CIF_{1E} = 0.30, CIF_{2E} = 0.10$), when $\theta_2 = 0.8$, then for $\theta_1 = 0.4$, the power shows about 100% with a very less sample size (< 200), and for $\theta_1 = 0.6$, the sample size is 400 with 80% power; when $\theta_2 = 1.2$, the SDH approach is as good as the CSH. For all the scenarios in Figure 3.1, the CSH approach performs better when $\theta_2 = 0.8$ and $\theta_1 = 0.6$.

FIGURE 3.1: Effect on power for exponential distribution under the CSH and SDH approaches by changing the effect of the competing event for the positive effect ($\theta_2 = 0.8$) and the adverse effect ($\theta_2 = 1.2$)

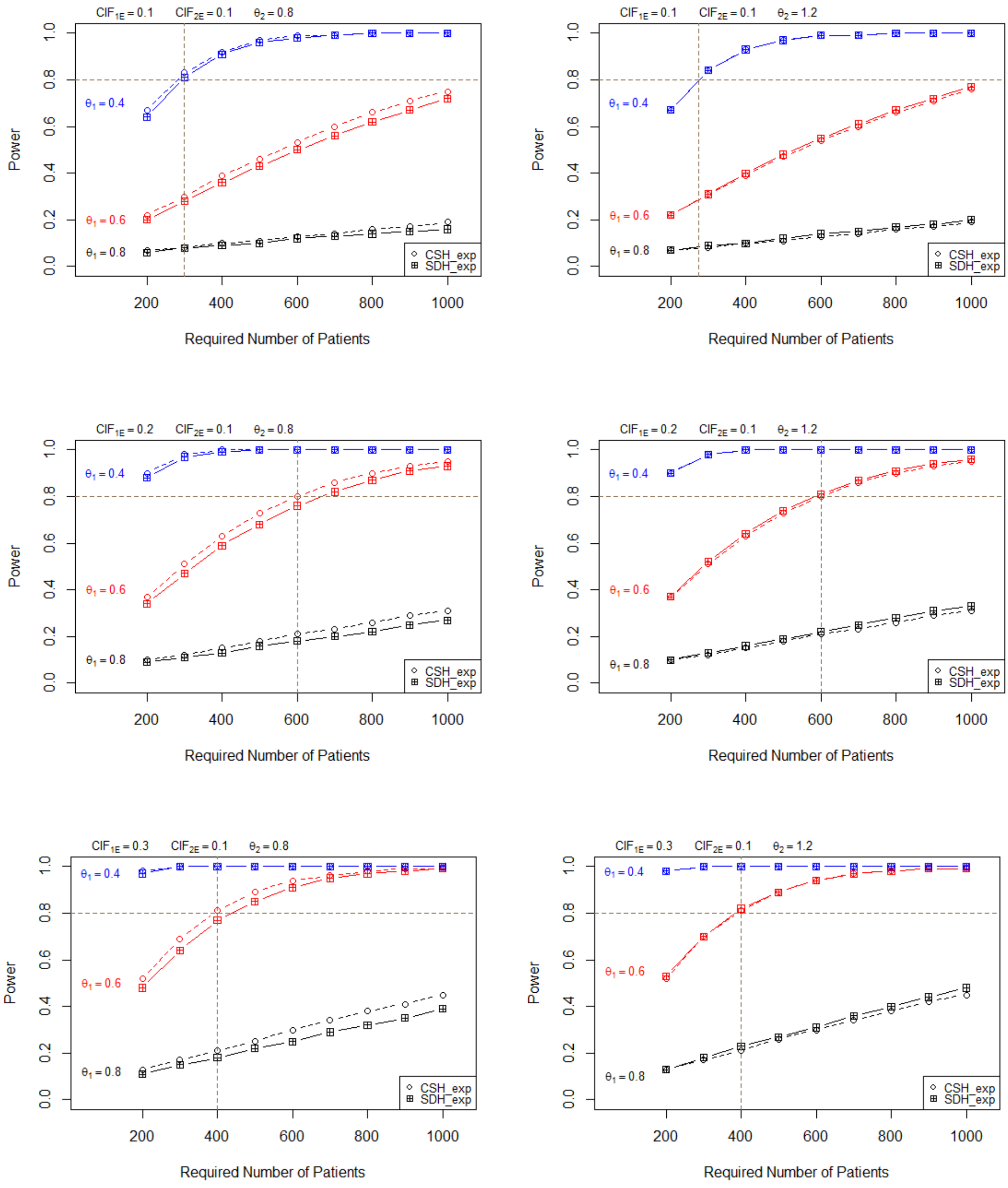


Figure 3.2 shows the power computations for Weibull distribution in both approaches. In the first scenario ($CIF_{1E} = 0.10, CIF_{2E} = 0.10$), in both cases of θ_2 (postitive or adverse effect of treatment), to attain 80% power, a very large treatment effect is required for event type 1, i.e., $\theta_1 = 0.4$. This is because, when $\theta_1 = 0.6$ is assumed

then, slightly more than 1000 is required for the sample size. This indicates that if an investigator wants to capture a positive effect of a treatment in the range $0.6 < \theta_1 < 1$, given a statistical power of 80%, then the required sample size is a value far above 1000. In the second scenario ($CIF_{1E} = 0.20, CIF_{2E} = 0.10$), the sample size requirement for $\theta_1 = 0.6$ reduces to 600 when $\theta_2 = 1.2$ in both approaches, and for $\theta_2 = 0.8$, the reduced value is 700 and 600 for the SDH and CSH approaches. In the third scenario ($CIF_{1E} = 0.30, CIF_{2E} = 0.10$), the sample size for $\theta_1 = 0.6$ is even less when $\theta_2 = 1.2$ (i.e., about 400 for both approaches) and when $\theta_2 = 0.8$, it is 400 or 450 for the CSH and SDH approaches, respectively.

FIGURE 3.2: Effect on power for Weibull distribution under the CSH and SDH approaches by changing the effect of the competing event for the positive effect ($\theta_2 = 0.8$), and adverse effect ($\theta_2 = 1.2$)

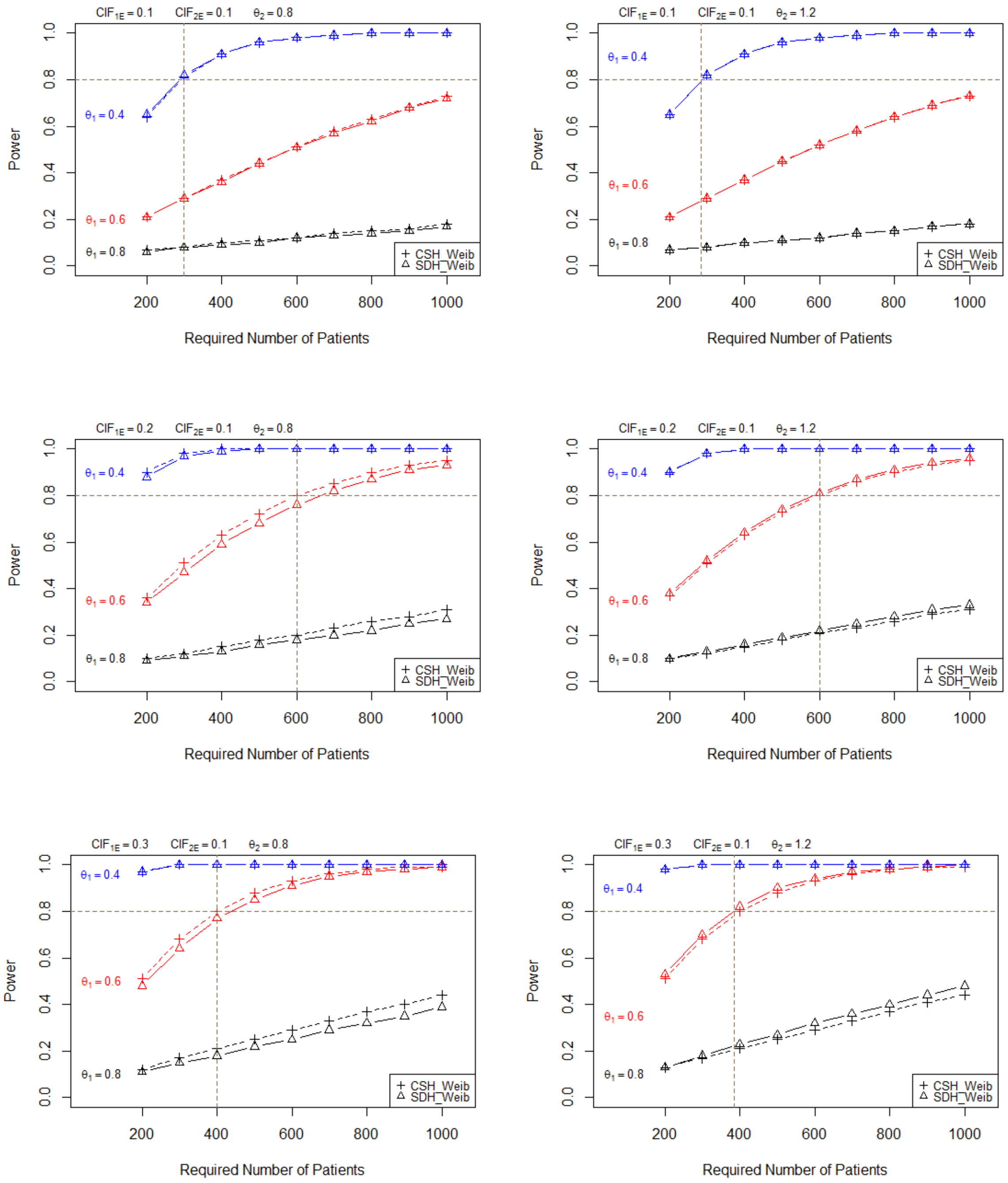
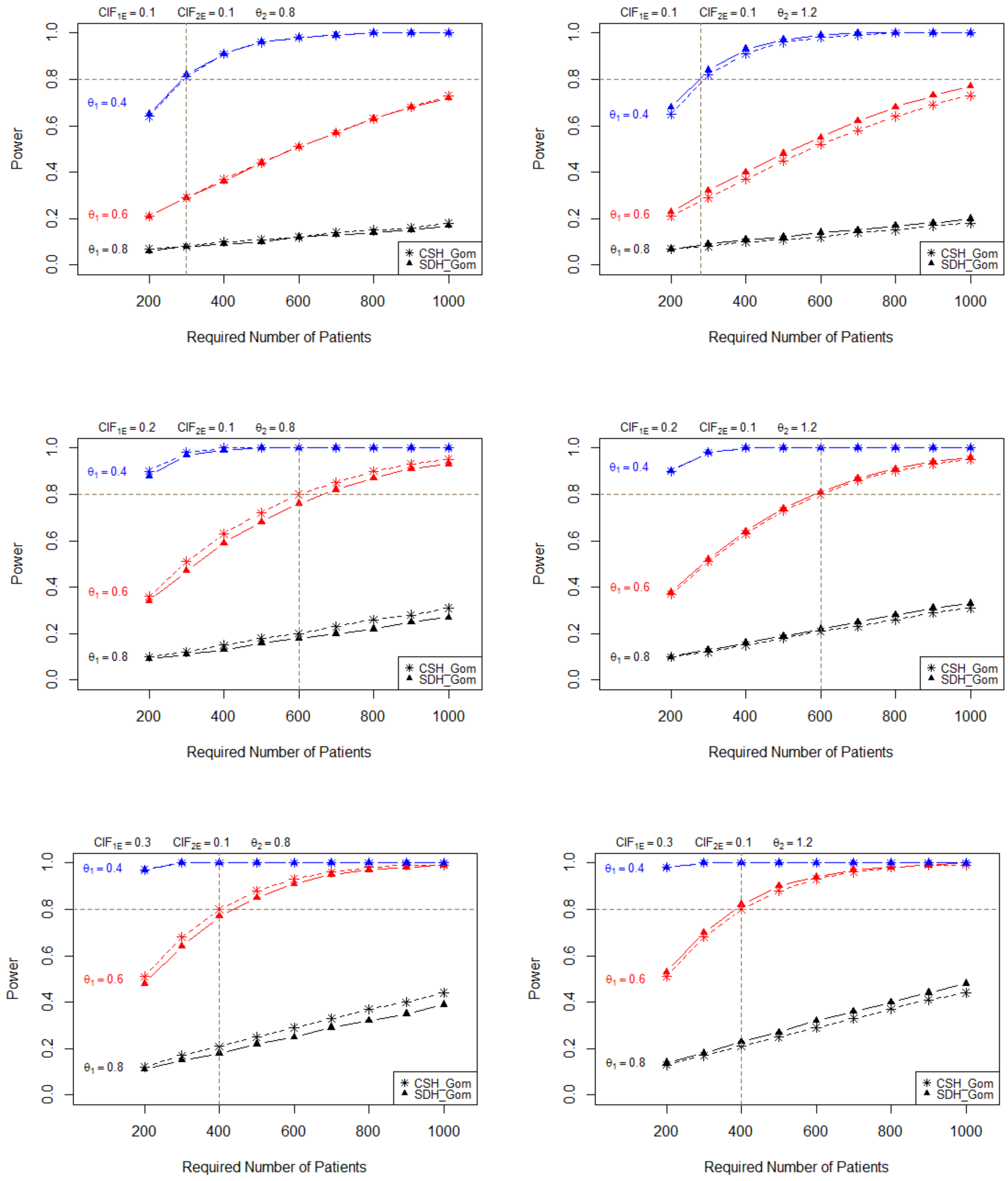


Figure 3.3 shows the power computation for Gompertz distribution for both approaches. In the first scenario ($CIF_{1E} = 0.10, CIF_{2E} = 0.10$), when $\theta_2 = 1.2$, the SDH approach performs slightly better for $\theta_1 = 0.4$ (275 for SDH and 300 for CSH) and for $\theta_1 = 0.6$, the sample size is about 1000 for the SDH approach and even more for the

CSH. In the second scenario ($CIF_{1E} = 0.20, CIF_{2E} = 0.10$), when $\theta_2 = 1.2$, the two approaches perform equally for all θ_1. In the third scenario ($CIF_{1E} = 0.30, CIF_{2E} = 0.10$), when $\theta_1 = 0.6$ and $\theta_1 = 0.8$, then for $\theta_2 = 0.8$, the CSH approach performs better and for $\theta_2 = 1.2$, the SDH approach performs slightly better.

FIGURE 3.3: Effect on power for the Gompertz distribution under the CSH and SDH approaches by changing the effect of the competing event for positive effect ($\theta_2 = 0.8$) and adverse effect ($\theta_2 = 1.2$)



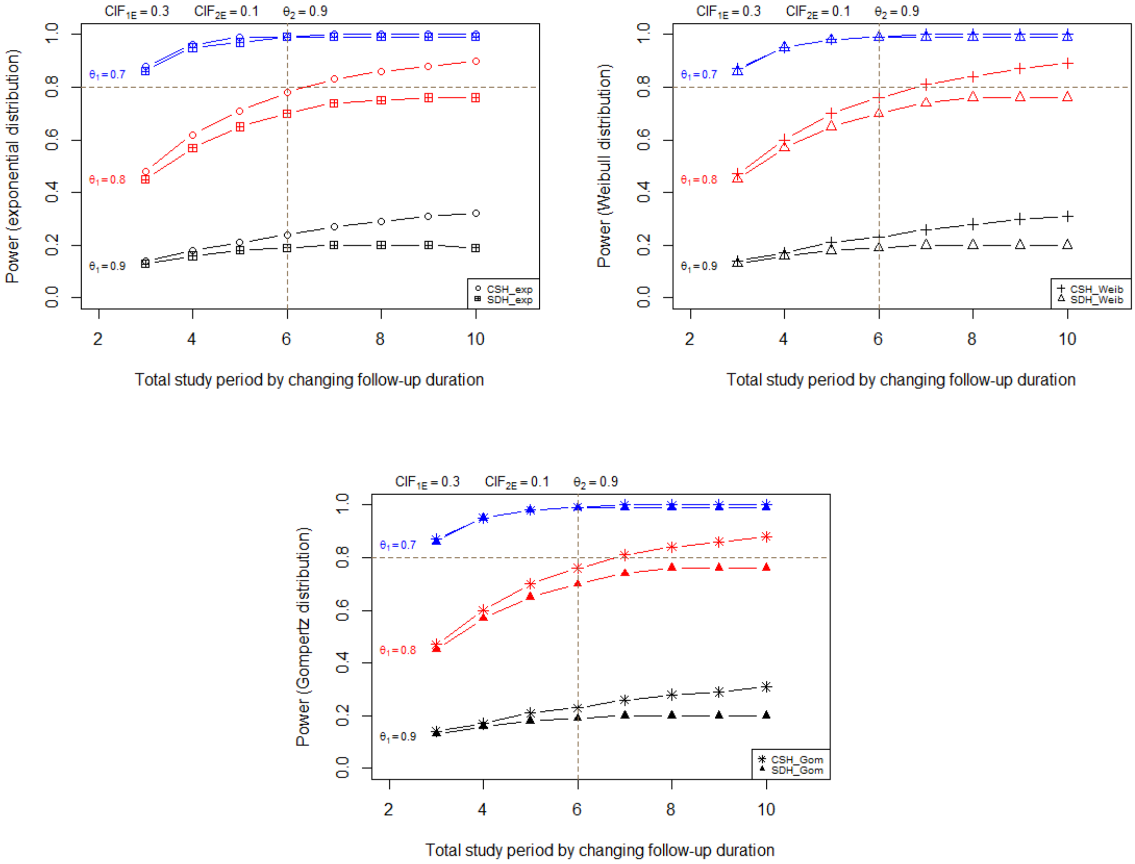
For a small change in the shape parameters in the two distributions ($\tilde{\gamma}_{1E} = \tilde{\gamma}_{2E} = 0.9475; \tilde{\gamma}_{1E} = \tilde{\gamma}_{2E} = 0.9933; \tilde{\gamma}_{1E} = 0.967, \tilde{\gamma}_{2E} = 1.05$ for Weibull in Figure 3.2 and $\tilde{\gamma}_{1E} = 1$ for the exponential in Figure 3.1), the sample size is almost unchanged. In all three types of distributions, when there is a positive treatment effect on the competing event ($\theta_2 = 0.8$), given a fixed power, CSH performs better in term's of proving that smaller sample size is required as compared to the SDH approach. This is true in particular for $\theta_1 > 0.4$. However, in the case of an adverse effect for the competing event ($\theta_2 = 1.2$), the two approaches seem to perform equally, with a negligible increase in sample size for the CSH approach, as compared to the SDH.

Study of power and sample size for different follow-up durations

Now, the objective is to conduct a study to observe the effect of increasing the follow-up duration on power and sample size. Here, the values of $\theta_1 = (0.9, 0.8, 0.7), \theta_2 = 0.9$, $a = 2$ years, $\tilde{f} = (1$ to 8 years $)$, $CIF_{1E} = 0.30$, and $CIF_{2E} = 0.10$ were assumed. Then, the shape parameters when using the CSH approach for the Weibull and Gompertz distributions were obtained: $\tilde{\gamma}_{1E} = 0.967, \tilde{\gamma}_{2E} = 1.05$ and $\tilde{\eta}_{1E} = -0.023, \tilde{\eta}_{2E} = 0.034$, respectively. For the SDH approach the same shape parameter values obtained in the CSH approach for event type $1 were assumed: \gamma^*_{1E} = 0.967$ and $\eta^*_{1E} = -0.023$.

Study the effect on power At the beginning, it was necessary to fix the sample size to 1500 for all the distributions under the CSH and SDH approaches. In Figure 3.4 (all scenarios), it can be observed that, in general, the power increases with longer study durations. In particular, when treatment is more effective on event type 1 (i.e., $\theta_1 = 0.7$), then all the distributions achieved 80% or more power with a study duration of less than 3 years. However, for $\theta_1 = 0.7$, this power always increases when the study period is extended upto 6 years. Meanwhile, when the treatment effect was slighty reduced (i.e., $\theta_1 = 0.8$), then the power for all distributions increases more for the CSH approach than the SDH for longer study periods. In contrast, the latter shows stability for all the distributions after 6 years with a slight reduction of power (about 75%).

FIGURE 3.4: Effect on power when using the CSH and SDH approaches for the exponential, Weibull, and Gompertz distributions. Here, $a = 2$ and $\tilde{f} = 1$ to 8 years, i.e., the total study duration ranges from 3 to 10 years.



Study of the effect on sample size In Figures 3.5 and 3.6 (i.e., all scenarios), it is observed that, with a minimal positive effect of treatment ($\theta_1 = 0.9$), there is a considerable decrease in sample size when the study period increased in length. Furthermore, when using the CSH approach for all the distributions, the sample size continues to decrease for study periods longer than 8 years. On the contrary, when using the SDH approach, this reduction in sample size stops after 8 years and remains constant after this period. When the treatment effect has a higher positive impact ($\theta_1 = 0.8$ and $\theta_1 = 0.7$), then the sample size has some changes until 7 years of the study period for all the distributions. After 7 years, the sample size remains approximately constant. Thus, the increment in the study duration for $\theta_1 < 0.9$ does not have a meaningful impact on the sample size after 7 years of the total study duration.

FIGURE 3.5: Effect on sample size when using the CSH and SDH approaches for the exponential distribution. Here, $a = 2$ and $\tilde{f} = 1$ to 8 years, i.e., the total study duration ranges from 3 to 10 years.

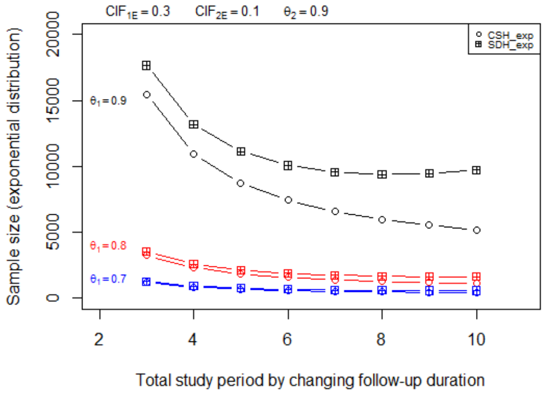
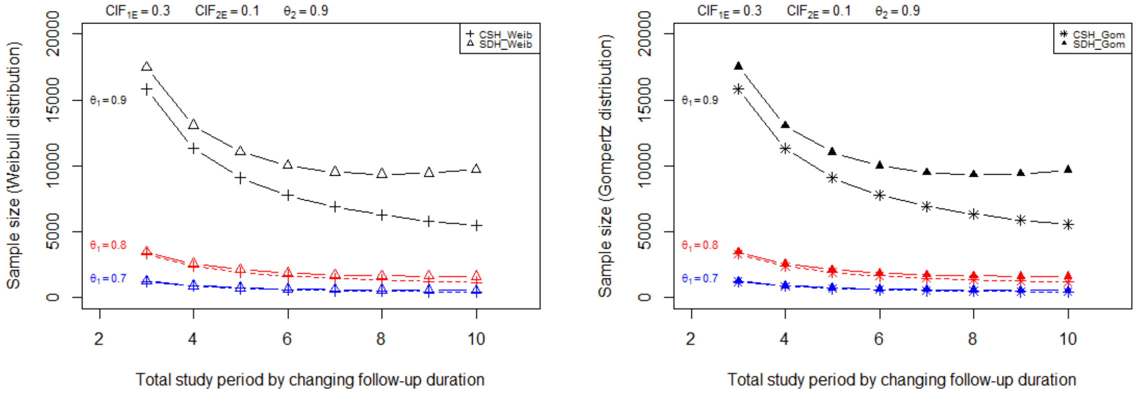


FIGURE 3.6: Effect on sample size when using the CSH and SDH approaches for the Weibull, and Gompertz distributions. Here, $a = 2$ and $\tilde{f} = 1$ to 8 years, i.e., the total study duration ranges from 3 to 10 years.



3.4 Simulation study with the change of shape parameter values

Choosing one best model from the exponential, Weibull, or Gompertz time-to-event distributions depends on a particular study design. If the study design suggests that the time-to-event follows the Weibull distribution, then assuming the other distributions indicate smaller sample sizes would be misleading. However, different shape parameter values for a particular distribution can be studied to observe how sample size behaves by changing the shape parameter values.

As seen in Section 3.3.3, shape parameters were computed for the Weibull and Gompertz distributions using the CSH approach with CIF_{1E}, CIF_{2E} at time $t = \tilde{t}$. Thus, when there are different shape parameter values, then comparison following the guidelines explained in Section 3.3.3 is not possible. In this context, to study the sample size with the change of shape parameter values for Weibull and Gompertz distributions, a simulation study (Table 3.2) is conducted using the CSH and SDH approaches, where the shape parameter values were already known. Furthermore, the CSH rates λ_{1E} and λ_{2E} for the main and competing events for the experimental group were assumed. Then, for fixed values of θ_1 and θ_2, the control group hazard rates, λ_{1C} and λ_{2C} were computed using the formula: $\theta_k = \lambda_{kE}/\lambda_{kC}$, for $k = (1,2)$. Finally, CIF_{1E}, CIF_{1C} values were computed at time points $t = (\tilde{f}, 0.5a + \tilde{f}, \tilde{t})$ for event type 1 (event of interest). Here, it is assumed that $\tilde{\gamma}_{1E} = \gamma^*_{1C} = (0.5, 1.5)$ and $\tilde{\eta}_{1E} = \eta^*_{1E} = (-0.5, 0.5)$ for the Weibull and Gompertz distributions, respectively, given that, $\tilde{\gamma}_{1E} = \tilde{\gamma}_{1C}$ and $\tilde{\eta}_{1E} = \tilde{\eta}_{1C}$. For instance, using CSH, the CIFs values for event type 1 at time point $t = \tilde{t}$ are

$$\begin{cases} \text{Weibull: } CIF_{1j}(\tilde{t}) & = \int_0^{\tilde{t}} (\lambda_{1j}\tilde{\gamma}_{1j}u^{\tilde{\gamma}_{1j}-1}) \exp\left[-\left(\lambda_{1j}u^{\tilde{\gamma}_{1j}} + \lambda_{2j}u^{\tilde{\gamma}_{2j}}\right)\right] du \\ \text{Gompertz: } CIF_{1j}(\tilde{t}) & = \int_0^{\tilde{t}} (\lambda_{1j}e^{\tilde{\eta}_{1j}u}) \exp\left[-\left\{\frac{\lambda_{1j}}{\tilde{\eta}_{1j}}\left(e^{\tilde{\eta}_{1j}u} - 1\right) + \frac{\lambda_{2j}}{\tilde{\eta}_{2j}}\left(e^{\tilde{\eta}_{2j}u} - 1\right)\right\}\right] du \end{cases}$$

In a similar manner, the CIF values were computed at time points $t = (\tilde{f}$, and $0.5a + \tilde{f})$.

Using the SDH approach, it is first necessary to compute the SDH ratio θ^*_1 and rate λ^*_{1j}. To compute θ^*_1, the following equation was used:

$$\theta^*_1 = \frac{\log\{1 - CIF_{1E}(\tilde{t})\}}{\log\{1 - CIF_{1C}(\tilde{t})\}}. \tag{3.23}$$

Here, either fixed values are assumed for $CIF_{1E}(\tilde{t})$ and $CIF_{1C}(\tilde{t})$, or these can be found using the CSH approach. Because it is here necessary to control the changes in the shape parameters using the second option and $CIF_{1E}(\tilde{t})$ and $CIF_{1C}(\tilde{t})$ values are obtained from the exponential distribution in the CSH approach by applying equation (3.7).

To compute λ^*_{1j}, the following formula is applied:

$$\begin{cases} \text{Weibull: } \lambda^*_{1j} = -\log(1 - CIF_{1j}(\tilde{t}))/\tilde{t}^{\gamma^*_{1j}} \\ \text{Gompertz: } \lambda^*_{1j} = \frac{\eta^*_{1j}[-\log(1-CIF_{1j}(\tilde{t}))]}{e^{\eta^*_{1j}\tilde{t}}-1}. \end{cases}$$

Here, for a given value of $CIF_{1j}(\tilde{t})$ and assuming $\gamma^*_{1j} = (0.5, 1.5)$, $\eta^*_{1j} = (-0.5, 0.5)$, λ^*_{1j} is obtained. Then, CIF_{1j} values can be recomputed at other time points using the following formula:

$$\begin{cases} \text{Weibull: } CIF_{1j}(t) & = 1 - e^{-\lambda^*_{1j} t^{\gamma^*_{1j}}} \\ \text{Gompertz: } CIF_{1j}(t) & = 1 - exp\Bigg[- \frac{\lambda^*_{1j}(e^{\eta^*_{1j}t} - 1)}{\eta^*_{1j}} \Bigg]. \end{cases}$$

Now, with all the CIFs values, $\tilde{\psi}_{1j}$ can be computed using the CSH approach and ψ^*_{1j} using the SDH approach following equation (3.10). To compute sample size, it is also necessary to know the number of events that occured for event type 1 in both approaches ($\tilde{D}_1$ for the CSH approach and D^*_1 for the SDH approach) . To compute $\tilde{D}_1$, a fixed value was assumed for θ_1, and the value was plugged in equation (3.2). On the contrary, for the SDH approach, θ^*_1 was computed from equation (3.23) and thus the value was plugged in equation (3.2), from which D^*_1 was obtained. Finally, given the proportion of patients allocated to the control (p_C) and experimental (p_E) groups, $\tilde{\Psi}_1 = p_C\tilde{\Psi}_{1C} + p_E\tilde{\Psi}_{1E}$ was computed using the CSH approach and $\tilde{\Psi}^*_1 = p_C\Psi^*_{1C} + p_E\tilde{\Psi}^*_{1E}$ using the SDH approach. Thus, $\tilde{N}_1 = \tilde{D}_1/\tilde{\Psi}_1$ and $N^*_1 = D^*_1/\Psi^*_1$ were obtained using the CSH and SDH approaches.

Study of sample size

Two scenarios are shown in Table 3.2. In both scenarios, it was assumed that $\theta_1 = (0.8, 0.6, 0.4)$ and $\theta_2 = (0.8, 1.2)$. Furthermore, the following were assumed- in Scenario (a), the hazarad rates for both event types are equal ($\lambda_{1E} = \lambda_{2E} = 0.03$); and in Scenario (b), an increased proportion in the hazard rate for event type $1 : \lambda_{1E} = 0.10$, and kept fixed for event type $2 : \lambda_{2E} = 0.03$. Overall from Table 3.2, it is observed that with $\tilde{\gamma} = \gamma^* = 0.5$, the SDH approach requires a less sample size, and with $\tilde{\gamma} = \gamma^* = 1.5$, the CSH requires a less sample size for Weibull distribution. Similar results were seen for the Gompertz distribution.

In Scenario (a), when the shape parameters $\tilde{\gamma}$ increased from 0.5 to 1.5, and $\tilde{\eta}$ increased from -0.5 to 0.5, the sample size reduced by about one-third when using the CSH approach. Contrarily, with the same shape parameter values, the changes in sample sizes for the SDH approach had the opposite results (i.e., sample size increased by about 30%).

In Scenario (b), an increment in the hazard rate in event type 1, yielded a substantial reduction in the sample sizes for all the distributions using both approaches compared to Scenario (a).

Table 3.2: Sample size for Weibull and Gompertz distributions using the CSH and SDH approaches with different shape parameter values to observe the effect of sample size.

Here, $a = 2, and\tilde{f} = 2$. CIF_{1E} at time point $\tilde{t}$ is only reported. However, CIF_{1E}, and CIF_{1C} at all the time points are considered to compute sample sizes. For the SDH approach, CIF at time $t = \tilde{t}$ is computed based on the exponential distribution when using the CSH approach.

		CSH approach					SDH approach			
θ_1	θ_2	$\tilde{N}_{Weib_1}$	$\tilde{N}_{Weib_1}$	$\tilde{N}_{Gom_1}$	$\tilde{N}_{Gom_1}$		$N^*_{Weib_1}$	$N^*_{Weib_1}$	$N^*_{Gom_1}$	$N^*_{Gom_1}$
		$\tilde{\gamma} = 0.5$	$\tilde{\gamma} = 1.5$	$\tilde{\eta} = -0.5$	$\tilde{\eta} = 0.5$	θ_1^*	$\gamma^* = 0.5$	$\gamma^* = 1.5$	$\eta^* = -0.5$	$\eta^* = 0.5$
Scenario (a):										
$\lambda_{1E} = 0.03$		$CIF_{1E} =$ 0.06	$CIF_{1E} = 0.19$	$CIF_{1E} = 0.05$	$CIF_{1E} = 0.27$		$\lambda^*_{1E} = 0.06$	$\lambda^*_{1E} = 0.02$	$\lambda^*_{1E} = 0.07$	$\lambda^*_{1E} = 0.01$
$\lambda_{2E} = 0.03$							$CIF_{1E} = 0.11$	$CIF_{1E} = 0.11$	$CIF_{1E} = 0.11$	$CIF_{1E} = 0.11$
0.8	0.8	11495	4272	12827	3319	0.81	7124	9230	6929	10538
0.8	1.2	11427	4196	12760	3238	0.79	5568	7213	5416	8235
0.6	0.8	1865	703	2079	551	0.61	1075	1389	1046	1585
0.6	1.2	1852	690	2067	536	0.60	956	1235	930	1409
0.4	0.8	449	175	500	140	0.41	257	330	250	376
0.4	1.2	446	171	497	136	0.40	238	306	232	348
Scenario (b):										
$\lambda_{1E} = 0.10$		$CIF_{1E} =$ 0.18	$CIF_{1E} = 0.50$	$CIF_{1E} = 0.15$	$CIF_{1E} = 0.62$		$\lambda^*_{1E} = 0.19$	$\lambda^*_{1E} = 0.05$	$\lambda^*_{1E} = 0.22$	$\lambda^*_{1E} = 0.03$
$\lambda_{2E} = 0.03$							$CIF_{1E} = 0.31$	$CIF_{1E} = 0.31$	$CIF_{1E} = 0.31$	$CIF_{1E} = 0.31$
0.8	0.8	3688	1562	4085	1300	0.82	2517	3189	2456	3611
0.8	1.2	3667	1538	4065	1274	0.79	1912	2422	1866	2742
0.6	0.8	607	268	671	226	0.62	389	490	380	553
0.6	1.2	604	263	668	222	0.60	340	428	332	483
0.4	0.8	152	72	167	63	0.41	99	123	97	138
0.4	1.2	151	71	166	62	0.40	91	113	89	126

3.5 Discussion

In this chapter, the objectives was to calculate the fixed design sample size for detecting a particular treatment effect for the Weibull, exponential and Gompertz time-to-event distributions using the CSH and SDH approaches. It was found that, in all three types of distributions, when there is a positive treatment effect on the competing event ($\theta_2 = 0.8$) given a fixed power, it was observed that CSH performs better in terms of providing a smaller required sample size than the SDH approach. This statement is true in particular for $\theta_1 > 0.4$. However, in case of an adverse effect for a competing event ($\theta_2 = 1.2$), the two approaches seem to perform equally, with a negligible increase in sample size for the CSH approach, as compared to the SDH.

Chapter 4

Computing sample size using group sequential design

4.1 Introduction

It is of common interest to justify a clinical trial about effectiveness of a new treatment by conducting interim analyses (the percentage of sample used from the fixed sample design) within the study period, which is referred to as the Group sequential (Gs) design. This design helps to reduce the number of allocated patients per treatment group, which in turns saves time and money. One of the main scopes in Gs design is to calculate the boundary values after having adjusted for type I and II errors using the error spending functions proposed in Gordon Lan and DeMets (1983) and Pampallona *et al.* (1995) as well as the conditional power proposed in Jennison and Turnbull (1999). These quantities help making an early decision to stop a trial at any time before the final stage analysis.

Moreover, the interim analysis is also a requirement by data monitoring committee. Most importantly, it is unethical to ask a patient to continue participation in a trial that should be stopped early either due to efficacy or futility. Specifically, the former means the drug was already found very effective during the interim period, while the latter occurs when the condition of a patient is futile due to the adverse effect of the treatment or alternatively, the true effect is far from the anticipated under the alternative hypothesis. For the Gs design, the main contributions are as follows: guidelines to compute the Sub-Distribution Hazard (SDH) ratios, events size, and conditional power at each interim stage for the Cause-Specific Hazard (CSH) and SDH approaches were explained through simulation studies.

In Section 4.1, the importance of sample size and the novel contributions in the

Gs design are discussed. Section 4.2 explains the theoretical aspects of computing boundary values. Next, error spending function and conditional power formulations are described in Sections 4.3 and 4.4, respectively. A simulation study is undertaken in Section 4.5. Finally, guidelines for deriving sample size using the direct modeling approach are provided in Section 4.6.

4.2 Calculation of boundary values

Let us suppose that the interest is to test the parameter β, obtained as $\beta = \log\theta$, where θ is the hazard ratio in a proportional hazard model. Let $\hat{\beta}$ be an estimate of β. The hypothesis then

$$H_0 : \beta \geq 0 \text{ against } H_1 : \beta < 0$$

Now, assuming that $\hat{\beta}$ is efficient, properly normalized and computed sequentially over time, it has an asymptotically normal independent increment process whose distribution depends only on the parameter β and Fisher's information $\mathscr{I}$ (Scharfstein *et al.*, 1997). Here, the total number of analyses is considered equal to K, indexed by $k = 1, \ldots, K$. Then, it is assumed that perform $(K - 1)$ interim analyses are performed, where K indicates the final stage analysis. Then, for each analysis, the estimator is

$$\widehat{\beta}_k \sim \mathcal{N}\left(\beta, \{\mathscr{I}_k\}^{-1}\right)$$

and the covariance of estimators from the two interim analyses, k_1 and k_2, with $k_1 < k_2$, is

$$\text{Cov}\left(\widehat{\beta}_{k_1}, \widehat{\beta}_{k_2}\right) = \text{Var}\left(\widehat{\beta}_{k_2}\right) = \{\mathscr{I}_{k_2}\}^{-1}$$

This is termed as the canonical joint distribution (Jennison and Turnbull, 1999). Moreover, the standardized statistic at analysis k is

$$Z_k = \frac{\widehat{\beta}_k}{\sqrt{\text{Var}\left(\widehat{\beta}_k\right)}} = \widehat{\beta}_k \sqrt{\mathscr{I}}_k$$

For this statistic, the canonical joint distribution of $\left(\widehat{\beta}_1, \ldots, \widehat{\beta}_K\right)$ implies that $(Z_1, \ldots, Z_K)$ has a multivariate normal distribution where,

$$Z_k \sim \mathcal{N}\left(\beta\sqrt{\mathscr{I}}_k, 1\right) \quad \text{and}$$
$$\text{Cov}\left(Z_{k_1}, Z_{k_2}\right) = \sqrt{\mathscr{I}_{k_1}/\mathscr{I}_{k_2}} \quad \text{for } k_1 < k_2.$$

Here, Fisher's information $\mathscr{I}$ for β can be written in terms of the number of observed events, $\mathscr{I} = D/4$ (Jennison and Turnbull, 1999). Proof in this context has been given in Appendix B.1. This information $\mathscr{I}$ is then applied in the group sequential design to compute the final stage information $\mathscr{I}_K$ as

$$\mathscr{I}_K = \mathscr{I} * \texttt{IF}$$

where, `IF` is called the inflation factor, which is a ratio of the group sequential test's maximum to fixed sample size. This `IF` needs to be computed and involves the Z-critical values (also called efficacy or futility boundary limits) at each interim stage and error spending functions. In this context, the derivation as well as the desired values for the interim stage and error rates can be found in Jennison and Turnbull (1999).

Once the final stage information ($\mathscr{I}_K$) is computed, then the interim stages information are straightforward to compute using the formula,

$$\mathscr{I}_k = (k/K)\mathscr{I}_K.$$

Here, it was assumed that the number of events occured equally over the interim stages, which are referred to as equally spaced information. Once $\mathscr{I}_k$ is computed, since it is known that $D_k = 4 * \mathscr{I}_K$, the interim event size D_k can be immediately derived.

To obtain the Z-critical values and number of events at each interim stage, an easy method is to use the R package `gsDesign` (Anderson, 2016). When the $Z-$ critical values which are called here (l_k, u_k) are obtained at each interim stage from either Jennison and Turnbull (1999) or R package `gsDesign`, then the decision to continue or terminate the trial at any stage can be made according to the following rule:

$$\begin{cases} Z_k \in \mathcal{M}_k = (l_k, u_k), \text{trial continue to the next stage} \\ Z_k \in \mathcal{U}_k = (u_k, +\infty), \text{stop the trial for efficacy and conclude } H_1 \\ Z_k \in \mathcal{L}_k = (-\infty, l_k), \text{stop the trial for futility and conclude } H_0 \end{cases}$$

where, Z_k is the observed z-values at interim stage k, $\{u_k\}$ is the upper limit or the efficacy boundary, and $\{l_k\}$ is the lower limit or the futility boundary. Further, $l_k \leq u_k$ until the final stage, and at this stage, $l_K = u_K$.

4.3 Calculation of the error spending function

Since the same sample data are used in the clinical trial over the study period, it is essential to adjust the error rate by using flexible approaches of *error spending functions* that do not require the number and exact timing of interim stages to be fixed in advance. Here, the proportion of events at interim k are defined as equal to the information fraction $t = k/K = \mathscr{I}_k/\mathscr{I}_K$. Then, the non-decreasing error spending functions for $\alpha(t)$ and $\tilde{\beta}(t)$ will be used to set α (under null, $\theta \geq 0$) and $\tilde{\beta}$ (under alternative, $\beta = \log\theta < 0$).

Among several functional forms proposed by Gordon Lan and DeMets (1983), the following are used O'Brien-Fleming (spends very little α at the beginning), with $\alpha(t) = 2 - 2\phi(z_{1-\alpha}/\sqrt{t})$; and Pocock (spends α more evenly across the stages), with $\alpha(t) = \alpha\log\{1 + (e-1)t\}$. The $\tilde{\beta}$-spending function is calculated in a similar manner.

However, a more flexible way is to consider Wang-Tsiatis bounds (Wang and Tsiatis, 1987) with both O'Brien-Fleming and Pocock approaches together. Here, let ω be a real-value shape parameter that characterizes the boundary shape, which generally ranges from 0 (O'Brien-Fleming design) to 0.5 (Pocock design). The value 0.25 yields intermediate rejection boundaries between those of O'Brien-Fleming and Pocock. The upper and lower boundaries are defined as $u_k = C_1(\omega, \alpha, k)\,(\mathscr{I}_K/\mathscr{I}_k)^{\omega-0.5}$ and $l_k = \theta_A\sqrt{\mathscr{I}_k} - C_2(\omega, \tilde{\beta}, k)\,(\mathscr{I}_K/\mathscr{I}_k)^{\omega-0.5}$, where $C(.)$ is a positive constant chosen in such a way that in the final stage, $u_k = l_k$. It is important to note that when the interim analysis was done at unplanned information fraction than initially planned, the values of boundaries at that stage is required to be recalculated.

4.4 Calculation of conditional power (CP)

In an interim analysis, the power depends on the interim stage and changes over time. Here, CP is defined as the probability of rejecting H_0 (when H_1 is true) given the observed interim stage data. When the trial starts, the CP is actually equal to the unconditional power. Once the CP is calculated at each stage, the decision can be made to stop the trial due to the futility or efficacy if CP is found to be very small or very large, respectively. Specifically, the CP is a function of the assumed hazard ratio θ for future patients entering the study at a follow-up time.

Following the findings of Jennison and Turnbull (1999), to reject a null hypothesis about $\beta = \log\theta < 0$ for a given value of observed Z_k (calculated using data collected up-to $k-1$), the general lower one-sided CP at interim stage k is,

$$P_k(\beta) = \Phi[(-Z_k\sqrt{\mathscr{I}_k} - Z_K\sqrt{\mathscr{I}_K} - (\mathscr{I}_K - \mathscr{I}_k)\beta)/\sqrt{\mathscr{I}_K - \mathscr{I}_k}]. \tag{4.1}$$

where, β is the log hazard ratio at the end of the trial.

When the values of β are changed for the CSH and SDH approaches, different CP are obtained at interim k.

4.5 Simulation results using Gs design

This simulation is conducted for 4 interim stages and 1 final stage (i.e., a total of 5 stages). Here, Wang-Tsiatis error spending function with $\omega = 0.25$ is considered for the adjustment of type I and type II error rates (see Section 4.3 for Wang-Tsiatis bound). Furthermore, it is assumed that $\alpha = 0.025$(one-sided test), $\tilde{\beta} = 0.20, \theta_1 = \theta_2 = 0.8, P_E = P_C = 0.5, CIF_{1E} = CIF_{2E} = 0.10$, and $t = \tilde{t} = (a + \tilde{f}) = 2 + 1 = 3$ years with the overall censoring rate $\tau = 0.01$. Using the CSH approach, the event size $\tilde{D}_1 = 632$ is obtained using equation (3.2) for fixed design. Next, for the SDH approach, the hazard ratio is not assumed and is instead computed (this was referred to SDH ratio, θ_1^*) using the equation (3.20); thus, the event size (D_1^*) differs from that of the CSH approach. Here, a short explanation of the computation of D_1^* is provided. To compute equation (3.20), the values of $CIF_{1E}(t)$ and $CIF_{1C}(t)$ are necessary. However, while $CIF_{1E}(t)$ is known, $CIF_{1C}(t)$ is unknown. This value can be found using exponential distribution with the CSH approach as follows:

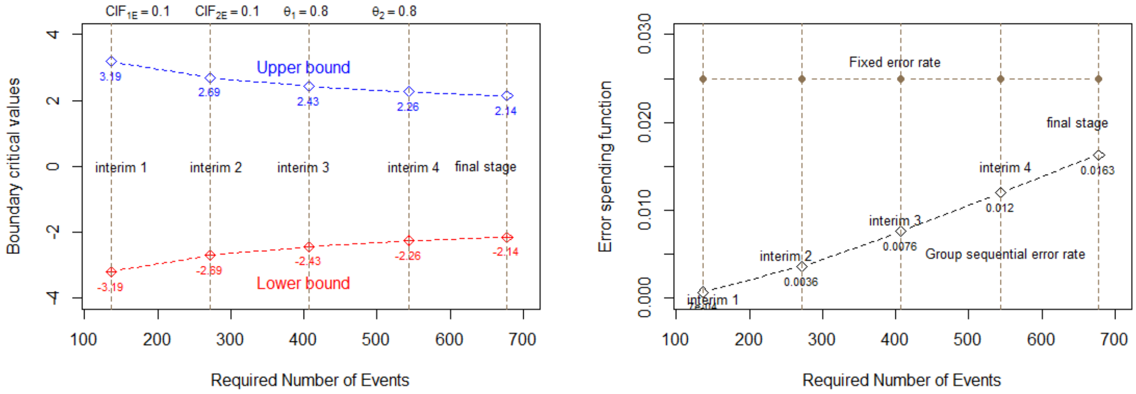
$$CIF_{1C}(t) = \frac{1 - e^{-t\{\lambda_{1C}(1+\tilde{\lambda}_C)\}}}{1 + \tilde{\lambda}_C}, \tag{4.2}$$

where, $\tilde{\lambda}_C = \lambda_{2C}/\lambda_{2C}$. Equation (4.2) is further required to find the values of $\lambda_{1C}, \lambda_{2C}$. These values are obtained using the equation, $\lambda_{1C}(t) = \lambda_{1E}(t)/\theta_1$ and $\lambda_{2C}(t) = \lambda_{2E}(t)/\theta_2$. Furthermore, the CSH rates $\lambda_{1E}(t)$ and $\lambda_{2E}(t)$ are computed using the system of equations (3.13). Finally, $CIF_{1C}(t) = 0.12$ and $\theta_1^* = 0.81$ are obtained. Then, keeping the values of α, β, P_E, P_C the same, the event size $D_1^* = 716$ is computed using equation (3.2). To compute Z-critical values and number of events at each interim stage, the R package `gsDesign` (Anderson, 2016) is used.

4.5.1 Computation of boundary critical values and error rates

From Figure 4.1 (left panel) it is observed that, boundary critical values (l_k, u_k) are larger in the first interim stages and have about one-third reduction in the final stage. When fixed design is used, the error rate is fixed, whereas in the Gs design, the error spending function increases with the stages and is always lower in comparison (Figure 4.1, right panel).

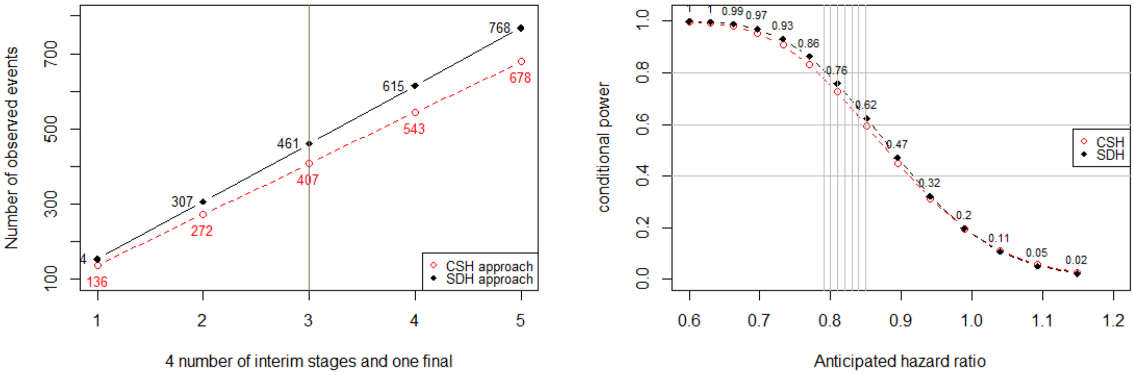
FIGURE 4.1: Boundary values (l_k, u_k) when using the CSH approach (left panel). The comparison of type I error rates in fixed and Gs designs (right panel).



4.5.2 Computation of conditional power

The information $\mathscr{I}$ is computed using the equation Appendix (B.1), $\mathscr{I} = [(1.97 + 0.84)/\log 0.8]^2 = 158$ from the CSH approach. Then, the fixed design event size is $\tilde{D}_1 = 158 * 4 = 632$. To compute information at final stage ($\mathscr{I}_5$), the inflation factor ($\texttt{IF}$)$= 1.072$ from statistical Table 2.10 in Jennison and Turnbull (1999) is used. Thus, the information at final stage is, $\mathscr{I}_5 = 158{*}1.072 = 169$. Then, the information at inetrim stage 3 is $\mathscr{I}_3 = (3/5) * 169 = 101.63$. Thus, the event size for stage 3 is $101.63 * 4 = 407$ using the CSH approach (Figure 4.2, left panel). When using the SDH approach, the information $\mathscr{I} = [(1.97 + 0.84)/\log 0.81]^2 = 179$. Then, the fixed design event size is, $D_1^* = 179 * 4 = 716$. With the inflation factor ($\texttt{IF}$)$= 1.072$, the final stage event size is, $179 * 1.072 = 192$ and at stage 3, it is $(3/5) * 192 = 115.20$. Thus, the event size for stage 3 is $115.20 * 4 = 461$ for the SDH approach (Figure 4.2, left panel).

FIGURE 4.2: Comparison of the CSH and SDH approaches with the Gs design. Left: Estimated number of events at each stage. Right: Conditional power as a function of the assumed hazard ratio at stage 3.



Here, the observed z values are assumed to be -2 and 2.14 at interim stage 3 and final stage 5, respectively. Now, suppose that an investigator wants to compute conditional power at interim stage 3 for a patient that enters the study in a later period with a CSH ratio of 0.80 and SDH ratio of 0.81. These hazard ratio values can then be directly used in the equation (4.1). However, the standard error of a given hazard ratio was computed in interim stage 3 using the CSH and SDH approaches. The standard error with the former is $se(\hat{\beta}_k) = \hat{\beta}_k / z = \log 0.8 / -2 = 0.112$ and with the latter is $se(\hat{\beta}_k) = \log 0.81 / -2 = 0.105$. Then, the assumed treatment effect (β) for interim stage 3 for both approaches are as follows:

$$\begin{cases} \text{CSH approach} : \beta_{\text{CSH}} = \log 0.80 / \sqrt{var(\hat{\beta}_k) * \mathscr{I}_3} = -0.22/(0.112 * \sqrt{102}) = -0.194. \\ \text{SDH approach} : \beta_{\text{SDH}} = \log 0.81 / \sqrt{var(\hat{\beta}_k) * \mathscr{I}_3} = -0.21/(0.105 * \sqrt{115}) = -0.186. \end{cases}$$

Then, the CP at stage 3 is

$$\begin{cases} \text{CSH approach} : P_3(\beta_{\text{CSH}}) = \Phi \left[\frac{\{-(-2)*\sqrt{102}\}-2.14*\sqrt{169}-\{(169-102)*(-0.194)\}}{\sqrt{(169-102)}} \right] = 0.74. \\ \text{SDH approach} : P_3(\beta_{\text{SDH}}) = \Phi \left[\frac{\{-(-2)*\sqrt{115}\}-2.14*\sqrt{192}-\{(192-115)*(-0.186)\}}{\sqrt{(192-115)}} \right] = 0.76. \end{cases}$$

This indicates that, under H_1, the probability of rejecting the null hypothesis if the experiment stops at stage 3 is reduced from 0.80 to 0.74 when using the CSH approach in comparison with trial completion. The first 407 of the planned 678 patients achieved 74% conditional power resulting in the detection of a hazard ratio of 0.80 at a significance level of 0.025 using a one-sided test. Moreover, the study was designed to have 50% of the patients in the experimental group. When the assumed $\log \theta$ is changed, given the other parameters fixed, then the CP changes accordingly, e.g., if we assume $\theta = 0.9$, the CP reduces to 0.45.

When using the SDH approach at stage 3, the computed SDH ratio is 0.81, which is a slightly higher value than the CSH ratio of 0.80. However, for such a small amount of change in the hazard ratio (0.01), the D_1^* has changed considerably with a slight increase in CP; the latter shows that 54 additional events are required as compared to that of the CSH approach (Figure 4.2, left panel). Additionally, for larger changes in the hazard ratio, e.g., from 0.79 to 0.85, the CP reduces from ≈ 0.80 to 0.60. For instance, if a patient that enters the study at an accrual period with assumed treatment effect, $\theta_1 = 0.85$, then the conditional power is reduced to 0.60 (Figure 4.2, right panel). Thus, the SDH approach yields a higher D_1^* and a slight gain in power as compared to the CSH approach (depending on the discrepancy between the two hazard ratios).

For an investigator, an easy way to decide on trial continuation or stoppage is to calculate the futility index, which is $1 -$ CP. For example, with a hazard ratio of 0.8, this index is 0.26 for the CSH approach (calculated as $1 - 0.74$). If the index is found greater than 0.8, which means the conditional power is lower than 0.20, then the study may be stopped because there is then a very small chance of achieving statistical significance.

4.6 Sample size when using the alternative direct modeling approach

This section details how sample size computations can be undertaken using the pseudo-value and binomial regression approaches and appropriate link functions. For simplicity, time-independent covariates will be considered, but the inference procedures for time-dependent covariate can be generalized. Moreover, the Cox-type proportional subdistributional hazard model can be written as

$$-\log\{1 - CIF_1(t \mid \mathbf{X})\} = \int_0^t \alpha_{01}(u) exp(\mathbf{X}^T \boldsymbol{\beta}) du = \exp(\mathbf{X}^T \boldsymbol{\beta}) \int_0^t \alpha_{01}(u) du \qquad (4.3)$$

where $\alpha_{01}(\cdot)$ is a completely unspecified, invertible, and monotone increasing function, and β is a regression parameter. By considering one binary covariate (treatment) for experimental group (E) and control group (C),

$$\begin{cases} -\log\{1 - CIF_E(t \mid x = 1)\} & = \exp(\beta) \int_0^t \alpha_{01}(u)du \\ -\log\{1 - CIF_C(t \mid x = 0)\} & = \int_0^t \alpha_{01}(u)du \end{cases}$$

Dividing this equation implies that

$$\Rightarrow \frac{-\log\{1 - CIF_E(t)\}}{-\log\{1 - CIF_C(t)\}} = \exp(\beta).$$

Further, using the log in equation (4.3) implies the following complimentary log-log link function:

$$\Rightarrow \log\{-\log\{1 - CIF_1(t \mid x)\}\} = x\boldsymbol{\beta} + \log \int_0^t \alpha_{01}(u)du = x\boldsymbol{\beta} + \alpha_{01}(t),$$

Now, assuming a proportional odds model, the CIF is,

$$\text{logit}\,(CIF_1(t; x, \beta)) = \log(\alpha_{01}(t)) + x\beta$$

where $\alpha_{01}(t)$ is an increasing positive function with $\alpha_{01}(0) = 0$. The cumulative incidence is thus linear on the logit scale with an intercept increasing over time and a time-constant log-odds ratio for failure from cause one. Then, the CIF is,

$$\begin{aligned} CIF_1(t; x) &= \frac{\exp\{\log \alpha_{01}(t)\} \exp(x\beta)}{1 + \exp\{\log \alpha_{01}(t)\} \exp(x\beta)} \\ &= \frac{\alpha_{01}(t) \exp{(x\beta)}}{1 + \alpha_{01}(t) \exp{(x\beta)}} \\ \Rightarrow \text{-log}(1 - CIF_1(t; x)) &= \log\{1 + \alpha_{01}(t-) \exp(x\beta)\}. \end{aligned}$$

Upon taking the first derivative for event type 1,

$$-\frac{\partial}{\partial t} \log{(1 - CIF_1(t; x))} = \frac{\exp(x\beta)\alpha_{01}{}'(t)}{1 + \alpha_{01}(t-) \exp(x\beta)} = \frac{1}{\exp(-x\beta) + \alpha_{01}(t-)} \alpha_{01}{}'(t)$$

where $\alpha'_{01}(t)$ is the derivative of $\alpha_{01}(t)$. Now, the distribution of

$$n^{1/2} \left\{ g\left(\widehat{CIF}_1(t; \boldsymbol{x})\right) - g\left(CIF_1(t; \boldsymbol{x})\right) \right\}$$

can be investigated by calculating the score function for regression parameters and then calculate the required sample size.

Conclusions

Discussion

Many studies have investigated the identification of appropriate statistical models for given clinical data by either ignoring the competing events or using inappropriate regression-based statistical methods to analyze complex clinical information. Hence, one of the objectives when using Competing Risks (CR) data was to estimate the probability of the main event among the many possible events over time, thus allowing the subjects to fail in competing events. This probability can be calculated by using the Cumulative Incidence Function (CIF), which is often of interest in medical research due to it's graphical representation. Moreover, there are four regression approaches available in the literature to estimate the CIF in the presence of competing events: Cause-Specific Hazard (CSH), Sub-Distribution Hazard (SDH), pseudo-value, and binomial regression approaches. However, the interpretation of the regression parameters for all the regression approaches are not straightforward and depend on the relationship with the CIF through link functions. However, patient's disease status can be examined using the CIF curves between treatment and control groups. In particular, the following objectives were studied: calculating the CIF using non-parametric and parametric regression approaches for CR data by reviewing the literature; providing a practical guideline using R data applications (BMT data); comparing the CIF among CSH, SDH, binomial, and pseudo-value regression approaches through simulation studies; validating the results with a medical application (Covid-19 data); computing sample sizes for fixed and Group sequential (Gs) designs.

The simulation study was conducted with $3,000$ replications for generating datasets of size $n = 500$. Here, the following values were assumed: shape parameters $a_1 = 0.45, a_2 = 0.5$; scale parameters $b_1 = 0.15, b_2 = 0.06$; regression coefficient: $\beta_1 = -0.6, \beta_2 = -0.17$; censoring parameter: $\tau = 180$ with censoring proportion 0.01. Additionally, one categorical covariate was assumed for both event types as x $\sim$ Ber (0.32). The analysis was then implemented in the freeware statistical package R (http://cran.r-project.org) and the bias was computed for all the regression approaches.

It was found that the bias is lower in the CSH and SDH approaches for both causes. Furthermore, at the beginning of the study (time point 10), the biases were very close among the CSH, SDH, and pseudo-value approaches for both the experimental and control groups. However, the biases are higher for the binomial approach in comparison to all other approaches. The maximum bias for the main event of interest for the experimental group is 0.025 for the binomial approach at time 130. For this approach, the bias is always higher for both events at time points 90 and 130. Meanwhile, for the pseudo-value approach, the bias was higher at the beginning of the study, and there was a substantial reduction over time. The ratios of observed to empirical standard errors of CIF for both event types and treatment groups for all approaches were near 1, indicating no substantial differences between the observed and empirical standard errors. However, for those approaches, there were underestimations of the variance over time. It was further compared time-fixed and time-varying coefficient effcts for the binomial approach. When the time-fixed coefficient was assumed, the CP performs worse after time point 40 for the experimental group for both event types. However, with the time-varying coefficient, the CP improved considerably for the experimental group, but also decreased slightly at time point 10 for both events, as compared to the time-fixed coefficient scenario. Furthermore, the efficiency measurements of binomial and pseudo-value with respect to the CSH were also studied and found that the pseudo-value approach showed a gain in effeciency for both events and in both groups. However, under the binomial approach, there was a loss in efficiency for the control group at later time points. In the clinical trial study, it is essential to study the efficiency as the objective of clinical trials is to establish the effect of an intervention. On the contrary, bias measurement is vital in the observational study as the inferences are individual preferences based.

Then, the regression methods using the Bone Marrow Transplant (BMT) and Covid-19 data were illustrated. Under BMT data, a practical guidelines using R for a new user was provided. Under Covid-19 data, it has been investigated the CR survival analyses for estimating the CIF of dying from Covid-19 and the CIF of dying from other causes in subjects with Covid-19. It was concluded that the exposures of asthma, diabetes, obesity, other.risk, immuno, kidney, neuro, flu.vaccine (who had not been vaccinated), hepatic.dis, age, sex, ICU, pneumo, and race, significantly increase the probability of death due to Covid. The highest hazard ratio (2.04) was observed for the subjects with age greater than 70 years compared to the age group $50 - 60$ years.

Then, the focus was on designing a sample size under fixed and Gs designs for competing risks survival data. To design a randomized clinical trial, an essential step is the calculation of the sample size or the number of patients to be recruited to detect the efficacy of treatments with sufficient power. In a time-to-event study, the sample size is determined not by the number of patients accrued, but instead by the number of events observed during a specific follow-up period. Here the objectives were as follows: to calculate the fixed design sample size for detecting a particular treatment effect for the Weibull, exponential and Gompertz time-to-event distributions using the CSH and SDH approaches and extend the fixed design sample size analysis into group sequential design by calculating conditional power. We first computed sample size and power under fixed design for exponential, Weibull, and Gompertz time-to-event distributions under CSH and SDH approaches. In all three types of distributions, when there is a positive treatment effect on the competing event ($\theta_2 = 0.8$) given a fixed power, it was observed that CSH performs better in terms of providing a smaller required sample size than the SDH approach. This statement is true in particular for $\theta_1 > 0.4$. However, in case of an adverse effect for a competing event ($\theta_2 = 1.2$), the two approaches seem to perform equally, with a negligible increase in sample size for the CSH approach, as compared to the SDH. To implement Gs design, interim stage information had to be computed to justify sample size in clinical trials as an ethical concern. Then simulation studies for this design were conducted assuming the same CIF for the experimental group using the CSH and SDH CR approaches. Even for a negligible increase in the hazard ratio (e.g. 0.01), it was found that the SDH model yields a higher number of events at each interim compared to CSH but with the advantage of a slight gain in conditional power. As a general recommendation, the SDH approach can be preferred when the main focus is to increase conditional power, while CSH is better at reducing the required number of events.

Future directions of research

A possible future work direction is to compute sample size using the binomial and pseudo-value regression approaches according to the theory given in Section 4.6. Then, a feasible user-friendly interface using R Shiny can be created for calculating the fixed and group-sequential design sample sizes in clinical trials. With this interface, a user can undertake design and interim monitoring without requiring any computational knowledge of R. Moreover, the estimates of a treatment effect can be biased when a clinical trial is terminated at an early stage regardless of whether a sequential or group sequential approaches are applied. Here, an alternative approach can be the use of an adaptive design that includes a prospectively planned opportunity to modify one or more specified aspects of the study design and interim data based on subjects in the study. In this manner, an investigator has the flexibility to identify the optimal clinical benefits of the new treatment under investigation without undermining the validity and integrity of the intended study.

Appendix A

Appendix for chapter 3

A.1 Derivation of N

The derivation of N is motivated from noninferiority clinical trial with time-to-event data in the presence of competing risks by the paper of Han *et al.* (2018). Consider the hypotheses of interest: $H_0 : \beta = \log\theta \geq 0$ against $H_1 : \beta = \log\theta < 0$. We can write the power function and the required sample size as,

$$Z_{1-\tilde{\beta}} = -Z_{1-\alpha} + \frac{(\beta - \beta_0)\sqrt{N}}{\sqrt{V}} \tag{A.1}$$

$$\Rightarrow N = \frac{\left(Z_{1-\alpha} + Z_{1-\tilde{\beta}}\right)^2 V}{(\beta - \beta_0)^2} \tag{A.2}$$

Now, we will calculate the variance V by deriving the information matrix $\mathscr{I}(\beta)$ for binary covariate x for Cox proportional hazard model,

$$\mathscr{I}(\beta) = \mathbb{E}_{\mathbf{x}\sim\mathscr{F}}\left[\sum_{i=1}^{N} I\left\{d_i = 1\right\} \frac{\left[\sum_{j\in\mathscr{R}} x_j^2 \exp\left(\beta x_j\right)\right]\left[\sum_{j\in\mathscr{R}} \exp\left(\beta x_j\right)\right] - \left[\sum_{j\in\mathscr{R}} x_j \exp\left(\beta x_j\right)\right]^2}{\left[\sum_{j\in\mathscr{R}} \exp\left(\beta x_j\right)^2\right]}\right].$$

Define, $\mathbb{E}_i(g(x)) = \sum_{j\in\mathscr{R}}\left[g(x)\exp\left(\beta x_j\right)\right]$. Since x is a binary covariate, we can write, $x_j^2 = x_j$ and rewrite the formula as,

$$\mathscr{I}(\beta) = \mathbb{E}_{\mathbf{x}\sim\mathscr{F}}\left[\sum_{i=1}^{N} I\left\{d_i = 1\right\} \frac{\mathbb{E}_i\left(x_j\right)\mathbb{E}_i(1) - \mathbb{E}_i^2\left(x_j\right)}{\mathbb{E}_i^2(1)}\right],$$

where, $\mathbb{E}_i(1) = N_{Ci} + N_{Ei}\exp(\beta)$ is a combination of patients with experimental (E) and control (C) group and $\mathbb{E}_i\left(x_j\right) = N_{Ei}\exp(\beta)$ is the patients with experimental group only. This implies,

$$\begin{aligned}
\mathscr{I}(\beta) &= \mathbb{E}_{\mathbf{x}\sim\mathscr{F}}\left[\frac{N_{Ei}\exp\beta * (N_{Ci}+N_{Ei}\exp(\beta)) - (N_{Ei}\exp(\beta))^2}{(N_{Ci}+N_{Ei}\exp(\beta))^2}\right] \\
&= \mathbb{E}_{\mathbf{x}\sim\mathscr{F}}\left[\frac{\prod_{x=0}^{1} N_{xi}\exp(\beta x)}{\left[\sum_{x=0}^{1} N_{xi}\exp(\beta x)\right]^2}\right] \\
&= \frac{N_{Ci}}{N_{Ci}+N_{Ei}\exp(\beta)} \times \frac{N_{Ei}\exp(\beta)}{N_{Ci}+N_{Ei}\exp(\beta)} \\
&= \frac{p_{Ci}}{p_{Ci}+p_{Ei}\exp(\beta)} \times \frac{p_{Ei}\exp(\beta)}{p_{Ci}+p_{Ei}\exp(\beta)} \\
\text{using Taylor series expansion} &= \frac{p_{Ci}p_{Ei}(1+\beta)}{(p_{Ci}+p_{Ei}(1+\beta))^2} \\
&= \frac{p_{Ci}p_{Ei}(1+\beta)}{(p_{Ci}+p_{Ei}+p_{Ei}\beta)^2} \\
&= \frac{p_{Ci}p_{Ei}(1+\beta)}{(1+p_{Ei}\beta)^2} \\
&\approx p_C p_E
\end{aligned}$$

where $p_{Ci} = N_{Ci}/N_i, p_{Ei} = N_{Ei}/N_i$ and $p_{Ci} + p_{Ei} = 1$. Note the Taylor expansion assumes that β is small. Therefore, the variance is expressed as follows:

$$V = \mathscr{I}_{\beta}^{-1} = (p_C p_E)^{-1}$$

So, the sample size formula A.1 becomes,

$$N = \frac{\left(Z_{1-\alpha/2} + Z_{1-\tilde{\beta}}\right)^2}{(\beta-\beta_0)^2\,(p_C p_E)} = D.$$

The derivation of $D = \frac{\left(z_{1-\alpha}+z_{1-\tilde{\beta}}\right)^2}{(\log\theta)^2 P_C P_E}$ is shown in Section 3.2.1.

The censoring mechanism and study duration are not considered here, thus it is assumed that the total number of deaths of all patients are observed and hence N is equal to D. However, in practice this is not feasible and we need to consider censoring mechanism and study duration. The following section explains on this issue.

A.1.1 Considering accrual period and loss to follow-up

Let CIF_x and f_x be the CIF and density function of the event of interest in group x and L_x be the cumulative incidence function of censoring. N denotes the total number of patients required in the two groups. p_x denotes the assignment proportion of subjects to group x. Therefore, N_{xi} can be rewritten as follows:

$$N_{xi} = Np_x \left(1 - CIF_x \left(T_i\right)\right) \left(1 - L_x \left(T_i\right)\right)$$

So, given $p_x, f(x),$ and H_x, we can define, $P(t) = \sum_{x=0}^{1} p_x f_x(t) \left(1 - L_x(t)\right).$ The information now can be expressed as,

$$\begin{aligned}
\mathscr{I}(b) &= E_{\mathbf{x}\sim\mathscr{F}} \left[\frac{\prod_{x=0}^{1} N_{xi} \exp(\beta x)}{\left[\sum_{x=0}^{1} N_{xi} \exp(\beta x)\right]^2} \right] \\
&= \int_0^{T_f} \left[\frac{\prod_{x=0}^{1} N_{xi} \exp(\beta x)}{\left[\sum_{x=0}^{1} N_{xi} \exp(\beta x)\right]^2} \right] P(t)dt \\
&= \int_0^{T_f} \frac{\prod_{x=0}^{1} \left[Np_x \left(1 - CIF_x(t)\right) \left(1 - L_x(t)\right)\right] \exp(bx)}{\left\{\sum_{x=0}^{1} \left[Np_x \left(1 - CIF_x(t)\right) \left(1 - L_x(t)\right)\right] \exp(bx)\right\}^2} P(t)dt
\end{aligned}$$

$$= \int_0^{T_f} \frac{N\left[p_C \left(1 - CIF_E(t)\right) \left(1 - L_C(t)\right)\right] \exp(0) \times \left[p_E \left(1 - CIF_E(t)\right) \left(1 - L_E(t)\right)\right] \exp(b)}{\{\left[Np_C \left(1 - CIF_C(t)\right) \left(1 - L_C(t)\right)\right] \exp(0) + \left[Np_E \left(1 - CIF_E(t)\right) \left(1 - L_E(t)\right)\right] \exp(b)\}^2} P(t)dt$$

In practice, $CIF_E(t) \approx CIF_C(t), L_E(t) \approx L_C(t)$. Then we have,

$$\begin{aligned}
\mathscr{I}(b) &= \int_0^{T_f} \frac{p_C p_E \exp(b)}{\left[p_C + p_E \exp(b)\right]^2} P(t)dt \\
&= \frac{p_C p_E \exp(b)}{\left[p_C + p_E \exp(b)\right]^2} \int_0^{T_f} P(t)dt \\
&= \frac{p_C p_E \exp(b)}{\left[p_C + p_E \exp(b)\right]^2} \Psi
\end{aligned}$$

where, $\quad \Psi = \int_0^{T_f} P(t)dt.$

Therefore, the variance is expressed as follows:

$$V(b) = \left(\frac{p_C p_E \exp(b)}{\left[p_C + p_E \exp(b)\right]^2} \Psi \right)^{-1}$$

As before, we can expand using Taylor series with $p_C + p_E = 1$ and obtain,

$$V(b) = \left(p_C p_E \Psi\right)^{-1}$$

So, plug-in the variance on sample size equation, we obtained,

$$\begin{aligned} N &= \frac{\left(Z_{1-\alpha/2} + Z_{1-\tilde{\beta}}\right)^2}{\left(\beta - \beta_0\right)^2} \left(p_C p_E \Psi\right)^{-1} \\ &= \frac{\left(Z_{1-\alpha/2} + Z_{1-\tilde{\beta}}\right)^2}{\left(\beta - \beta_0\right)^2 \left(p_C p_E\right)} * \frac{1}{\Psi} \\ &= \frac{D}{\Psi} \end{aligned}$$

Appendix B

Appendix for chapter 4

B.1 Fisher's information $\mathscr{I}$ for $\hat{\beta}$ can be written in terms of the number of observed events, $\mathscr{I} = D/4$ when $p_E = p_C = 1/2$.

Proof:

Recall the hypotheses of interest: $H_0 : \beta = \log\theta \geq 0$ against $H_1 : \beta = \log\theta < 0$

Step (1): Calculate the significance level, α (probability of Type I error):

$$\begin{aligned}
\alpha &= P\text{ (reject } H_0 \mid H_0 \text{ is true)} \\
&= P\left(\hat{\beta} < -c_\alpha \mid H_0\right) \\
&= P\left(\frac{\hat{\beta}-\beta}{\sqrt{V(\hat{\beta})}} < \frac{-c_\alpha - \beta}{\sqrt{V(\hat{\beta})}}\middle| H_0\right) \\
&= P\left(Z < \frac{-c_\alpha - 0}{\sqrt{1/\mathscr{I}}}\right) \\
&= P\left(Z < -c_\alpha\sqrt{\mathscr{I}}\right) \\
&= \Phi\left(-c_\alpha\sqrt{\mathscr{I}}\right) \\
&= 1 - \Phi\left(c_\alpha\sqrt{\mathscr{I}}\right) \\
\Rightarrow 1-\alpha &= \Phi\left(c_\alpha\sqrt{\mathscr{I}}\right) \\
\Rightarrow \Phi^{-1}(1-\alpha) &= c_\alpha\sqrt{\mathscr{I}} \\
\Rightarrow c_\alpha &= \Phi^{-1}(1-\alpha)/\sqrt{\mathscr{I}}.
\end{aligned}$$

Step (2):Calculate the significance level, $\tilde{\beta}$ (probability of Type II error). Assume $\beta < 0$.

$$
\begin{aligned}
\tilde{\beta} &= P(\text{ accept } H_0 \mid H_1 \text{ is true })\\
\Rightarrow 1-\tilde{\beta} &= P(\text{ reject } H_0 \mid H_1 \text{ is true })\\
&= P\left(\hat{\beta} < -c_\alpha \mid H_1\right)\\
&= P\left(\frac{\hat{\beta}-\beta}{\sqrt{V(\hat{\beta})}} < \frac{-c_\alpha-\beta}{\sqrt{V(\hat{\beta})}}\middle| H_1\right)\\
&= P\left(Z < \frac{-c_\alpha-\beta}{\sqrt{1/\mathscr{I}}}\right)\\
&= P\left(Z < (-c_\alpha-\beta)\sqrt{\mathscr{I}}\right)\\
&= P\left(Z < -c_\alpha\sqrt{\mathscr{I}}-\beta\sqrt{\mathscr{I}}\right)\\
&= P\left(Z < -\big(c_\alpha\sqrt{\mathscr{I}}+\beta\sqrt{\mathscr{I}}\big)\right)\\
&= \Phi\left(-\big(c_\alpha\sqrt{\mathscr{I}}+\beta\sqrt{\mathscr{I}}\big)\right)\\
\Rightarrow 1-\tilde{\beta} &= 1-\Phi\left(c_\alpha\sqrt{\mathscr{I}}+\beta\sqrt{\mathscr{I}}\right)\\
\Rightarrow \tilde{\beta} &= \Phi\left(c_\alpha\sqrt{\mathscr{I}}+\beta\sqrt{\mathscr{I}}\right)\\
\Rightarrow \Phi^{-1}(\tilde{\beta}) &= c_\alpha\sqrt{\mathscr{I}}+\beta\sqrt{\mathscr{I}}
\end{aligned}
$$

We can also write, $-\Phi^{-1}(1-\tilde{\beta}) = c_\alpha\sqrt{\mathscr{I}}+\beta\sqrt{\mathscr{I}}$. Thus, $\Phi^{-1}(\tilde{\beta}) = -\Phi^{-1}(1-\tilde{\beta}) = c_\alpha\sqrt{\mathscr{I}}+\beta\sqrt{\mathscr{I}}$. Now we substitute c_α from Step (1) :

$$
\begin{aligned}
-\Phi^{-1}(1-\tilde{\beta}) &= \left[\Phi^{-1}(1-\alpha)/\sqrt{\mathscr{I}}\right]\sqrt{\mathscr{I}}+\beta\sqrt{\mathscr{I}}\\
\Rightarrow -\Phi^{-1}(1-\tilde{\beta}) &= \Phi^{-1}(1-\alpha)+\beta\sqrt{\mathscr{I}}\\
\Rightarrow -\beta\sqrt{\mathscr{I}} &= \Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\tilde{\beta})\\
\Rightarrow \beta\sqrt{\mathscr{I}} &= -\left[\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\tilde{\beta})\right]\\
\Rightarrow \sqrt{\mathscr{I}} &= -\left[\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\tilde{\beta})\right]/\beta
\end{aligned}
$$

Thus, we have,

$$
\mathscr{I} = \left[\frac{\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\tilde{\beta})}{\beta}\right]^2 \tag{B.1}
$$

Now recalling formula (3.2), $D = [z_{1-\alpha} + z_{1-\tilde{\beta}}]^2/[(\log\theta)^2 P_E P_C]$ which is equivalent to,

$$\begin{aligned} D &= \left[\frac{\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\tilde{\beta})}{\beta}\right]^2 * (1/P_E P_C) \\ &= \mathscr{I} * (1/P_E P_C) = 4\mathscr{I}, \quad \text{when, } P_E = P_C = 1/2. \\ \Rightarrow \mathscr{I} &= D/4. \end{aligned}$$

It is clear from this equation that the information $\mathscr{I}$ is a function of fixed design events size.

Bibliography

Aalen, O. (1978) Nonparametric inference for a family of counting processes. *The Annals of Statistics* pp. 701–726.

Andersen, P. K., Abildstrom, S. Z. and Rosthøj, S. (2002) Competing risks as a multi-state model. *Statistical methods in medical research* **11**(2), 203–215.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (2012) *Statistical models based on counting processes.* Springer Science & Business Media.

Andersen, P. K., Klein, J. P. and Rosthøj, S. (2003) Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* **90**(1), 15–27.

Anderson, K. (2016) *R Package: gsDesign.* `https://cran.rproject.org/web/packages/gsDesign/gsDesign.pdf`.

Austin, P. C. and Fine, J. P. (2017) Practical recommendations for reporting fine-gray model analyses for competing risk data. *Statistics in medicine* **36**(27), 4391–4400.

Bender, R., Augustin, T. and Blettner, M. (2005) Generating survival times to simulate cox proportional hazards models. *Statistics in medicine* **24**(11), 1713–1723.

Bernstein, D. and Lagakos, S. (1978) Sample size and power determination for stratified clinical trials. *Journal of Statistical Computation and Simulation* **8**(1), 65–73.

Beyersmann, J., Allignol, A. and Schumacher, M. (2011) *Competing risks and multistate models with R.* Springer Science & Business Media.

Beyersmann, J., Latouche, A., Buchholz, A. and Schumacher, M. (2009) Simulating competing risks data in survival analysis. *Statistics in medicine* **28**(6), 956–971.

Bickel, P. J. and Doksum, T. (2001) *Mathematical Statistics (2nd edn).* Prentice-Hall.

Borgan, Ø. (2014) A alen–j ohansen estimator. *Wiley StatsRef: Statistics Reference Online* .

Cheng, S., Fine, J. P. and Wei, L. (1998) Prediction of cumulative incidence function under the proportional hazards model. *Biometrics* pp. 219–228.

Collett, D. (2003) *Modelling survival data in medical research.* CRC Press.

Cortese, G., Scheike, T. H. and Martinussen, T. (2010) Flexible survival regression modelling. *Statistical methods in medical research* **19**(1), 5–28.

Cox, D. R. (1975) Partial likelihood. *Biometrika* **62**, 269–276.

Daniele, N., Scerpa, M. C., Caniglia, M., Ciammetti, C., Rossi, C., Bernardo, M. E., Locatelli, F., Isacchi, G. and Zinno, F. (2012) Overview of t-cell depletion in haploidentical stem cell transplantation. *Blood Transfusion* **10**(3), 264.

David, H. A. and Moeschberger, M. L. (1978) *The Theory of Competing Risks: HA David, ML Moeschberger.* C. Griffin.

Demidenko, E. (2007) Sample size determination for logistic regression revisited. *Statistics in medicine* **26**(18), 3385–3397.

Demidenko, E. (2013) *Mixed models: theory and applications with R.* John Wiley & Sons.

Fine, J. P. and Gray, R. J. (1999) A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association* **94**(446), 496–509.

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

Gallo, P., Mao, L. and Shih, V. H. (2014) Alternative views on setting clinical trial futility criteria. *Journal of biopharmaceutical statistics* **24**(5), 976–993.

Ge, H., Wang, X., Yuan, X., Xiao, G., Wang, C., Deng, T., Yuan, Q. and Xiao, X. (2020) The epidemiology and clinical information about covid-19. *European Journal of Clinical Microbiology & Infectious Diseases* **39**(6), 1011–1019.

George, S. L. and Desu, M. (1974) Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of chronic diseases* **27**(1-2), 15–24.

Gerds, T. A. (2019) *prodlim: Product-Limit Estimation for Censored Event History Analysis.*
`https://cran.r-project.org/web/packages/prodlim/index.html`.

Gerds, T. A., Scheike, T. H. and Andersen, P. K. (2012) Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in medicine* **31**(29), 3921–3930.

Ghosh, S., Samanta, G. and Mubayi, A. (2021) Comparison of regression approaches for analyzing survival data in the presence of competing risks. *Letters in Biomathematics* **8**(1), 29–47.

Gordon Lan, K. and DeMets, D. L. (1983) Discrete sequential boundaries for clinical trials. *Biometrika* **70**(3), 659–663.

Gray, B., Gray, M. B. and Gray, R. (2004) *The cmprsk package, The comprehensive R Archive network.*
`http://cran.rproject.org/src/contrib/Descriptions/cmprsk.html`.

Gray, R. J. (1988) A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The annals of statistics* pp. 1141–1154.

Guo, C. and So, Y. (2018) Cause-specific analysis of competing risks using the phreg procedure. In *SAS Global Forum*, volume 2018, pp. 8–11.

Haller, B. (2014) *The analysis of competing risks data with a focus on estimation of cause-specific and subdistribution hazard ratios from a mixture model.* Ph.D. thesis, lmu.

Han, D., Chen, Z. and Hou, Y. (2018) Sample size for a noninferiority clinical trial with time-to-event data in the presence of competing risks. *Journal of Biopharmaceutical Statistics* **28**(4), 797–807.

Heo, M., Faith, M. S. and Allison, D. B. (1998) Power and sample size for survival analysis under the weibull distribution when the whole lifespan is of interest. *Mechanisms of ageing and development* **102**(1), 45–53.

Hsieh, F. and Lavori, P. W. (2000) Sample-size calculations for the cox proportional hazards regression model with nonbinary covariates. *Controlled clinical trials* **21**(6), 552–560.

Jennison, C. and Turnbull, B. W. (1999) *Group sequential methods with applications to clinical trials.* CRC Press.

Kalbfleisch, J. and Prentice, R. (2002) The statistical analysis of failure time data.

Kaplan, E. L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53**(282), 457–481.

Kim, H. T. (2007) Cumulative incidence in competing risks data and competing risks regression analysis. *Clinical cancer research* **13**(2), 559–565.

Klein, J. P. (2006) Modelling competing risks in cancer studies. *Statistics in medicine* **25**(6), 1015–1034.

Klein, J. P. and Andersen, P. K. (2005) Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* **61**(1), 223–229.

Klein, J. P., Gerster, M., Andersen, P. K., Tarima, S. and Perme, M. P. (2008) Sas and r functions to compute pseudo-values for censored data regression. *Computer methods and programs in biomedicine* **89**(3), 289–300.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G. and Scheike, T. H. (2014) *Handbook of survival analysis*. CRC Press.

Van der Laan, M. J., Laan, M. and Robins, J. M. (2003) *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

Lachin, J. M. (1981) Introduction to sample size determination and power analysis for clinical trials. *Controlled clinical trials* **2**(2), 93–113.

Lachin, J. M. and Foulkes, M. A. (1986) Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics* pp. 507–519.

Lai, C.-C., Shih, T.-P., Ko, W.-C., Tang, H.-J. and Hsueh, P.-R. (2020) Severe acute respiratory syndrome coronavirus 2 (sars-cov-2) and coronavirus disease-2019 (covid-19): The epidemic and the challenges. *International journal of antimicrobial agents* **55**(3), 105924.

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

Latouche, A., Porcher, R. and Chevret, S. (2004) Sample size formula for proportional hazards modelling of competing risks. *Statistics in medicine* **23**(21), 3263–3274.

Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* **73**(1), 13–22.

Lin, D. Y. and Ying, Z. (1994) Semiparametric analysis of the additive risk model. *Biometrika* **81**(1), 61–71.

Lumley, T. and Therneau, T. (2003) *The survival Package, The Comprehensive R Archive Network.*
`http://cran.r-project.org/src/contrib/Descriptions/survival.html`.

Moriña, D. and Navarro, A. (2017) Competing risks simulation with the survsim r package. *Communications in Statistics-Simulation and Computation* **46**(7), 5712–5722.

Nijman, G., Wientjes, M., Ramjith, J., Janssen, N., Hoogerwerf, J., Abbink, E., Blaauw, M., Dofferhoff, T., van Apeldoorn, M., Veerman, K. *et al.* (2021) Risk factors for in-hospital mortality in laboratory-confirmed covid-19 patients in the netherlands: A competing risk survival analysis. *PloS one* **16**(3), e0249231.

Pampallona, S., Tsiatis, A. and Kim, K. (1995) Spending functions for the type i and type ii error probabilities of group sequential tests. *J Statist Plan Inference* **42**(19), 1994–35.

Pavletic, S. Z., Smith, L. M., Bishop, M. R., Lynch, J. C., Tarantolo, S. R., Vose, J. M., Bierman, P. J., Hadi, A., Armitage, J. O. and Kessinger, A. (2005) Prognostic factors of chronic graft-versus-host disease after allogeneic blood stem-cell transplantation. *American journal of hematology* **78**(4), 265–274.

Pintilie, M. (2002) Dealing with competing risks: testing covariates and calculating sample size. *Statistics in medicine* **21**(22), 3317–3324.

Pintilie, M. (2006) *Competing risks: a practical perspective.* Volume 58. John Wiley & Sons.

Pohar Perme, M. and Gerster, M. (2017) pseudo: Computes pseudo-observations for modeling. *R package version* **1**(3).

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T. and Breslow, N. E. (1978) The analysis of failure times in the presence of competing risks. *Biometrics* pp. 541–554.

Rathouz, P. J., Valencia, V., Chang, P., Morton, D., Yang, H., Surer, O., Fox, S. J., Meyers, L. A., Matsui, E. C. and Haynes, A. B. (2021) Survival analysis methods for analysis of hospitalization data: Application to covid-19 patient hospitalization experience. *medRxiv* .

Salinas-Escudero, G., Carrillo-Vega, M. F., Granados-García, V., Martínez-Valverde, S., Toledano-Toledano, F. and Garduño-Espinosa, J. (2020) A survival analysis of covid-19 in the mexican population. *BMC public health* **20**(1), 1–8.

Satagopan, J., Ben-Porat, L., Berwick, M., Robson, M., Kutler, D. and Auerbach, A. (2004) A note on competing risks in survival data analysis. *British journal of cancer* **91**(7), 1229–1235.

Scharfstein, D. O., Tsiatis, A. A. and Robins, J. M. (1997) Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association* **92**(440), 1342–1350.

Scheike, T. H. and Martinussen, T. (2006) *Semi-parametric timevarying regression for R. Survival regression software.*
`http://stafi.pubhealth.ku.dk/ ts/timereg.html`.

Scheike, T. H. and Zhang, M.-j. (2002) An additive–multiplicative cox–aalen regression model. *Scandinavian Journal of Statistics* **29**(1), 75–88.

Scheike, T. H. and Zhang, M.-J. (2003) Extensions and applications of the cox-aalen survival model. *Biometrics* **59**(4), 1036–1045.

Scheike, T. H. and Zhang, M.-J. (2011) Analyzing competing risk data using the r timereg package. *Journal of statistical software* **38**(2).

Scheike, T. H., Zhang, M.-J. and Gerds, T. A. (2008) Predicting cumulative incidence probability by direct binomial regression. *Biometrika* **95**(1), 205–220.

Schoenfeld, D. A. (1983) Sample-size formula for the proportional-hazards regression model. *Biometrics* pp. 499–503.

Schulgen, G., Olschewski, M., Krane, V., Wanner, C., Ruf, G. and Schumacher, M. (2005) Sample sizes for clinical trials with time-to-event endpoints and competing risks. *Contemporary clinical trials* **26**(3), 386–396.

Scrucca, L., Santucci, A. and Aversa, F. (2010) Regression modeling of competing risk using r: an in depth guide for clinicians. *Bone marrow transplantation* **45**(9), 1388–1395.

Sierra, J., Pérez, W. S., Rozman, C., Carreras, E., Klein, J. P., Rizzo, J. D., Davies, S. M., Lazarus, H. M., Bredeson, C. N., Marks, D. I. *et al.* (2002) Bone marrow transplantation from hla-identical siblings as treatment for myelodysplasia. *Blood, The Journal of the American Society of Hematology* **100**(6), 1997–2004.

Tai, B., Chen, Z. and Machin, D. (2018) Estimating sample size in the presence of competing risks–cause-specific hazard or cumulative incidence approach? *Statistical methods in medical research* **27**(1), 114–125.

Tsiatis, A. (1975) A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences* **72**(1), 20–22.

Tsiatis, A. A. (1981) A large sample study of cox's regression model. *The Annals of Statistics* **9**(1), 93–108.

Wan, F. (2017) Simulating survival data with predefined censoring rates for proportional hazards models. *Statistics in medicine* **36**(5), 838–854.

Wang, S. K. and Tsiatis, A. A. (1987) Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* pp. 193–199.

Wu, J. (2015) Power and sample size for randomized phase iii survival trials under the weibull model. *Journal of biopharmaceutical statistics* **25**(1), 16–28.

Yan, J. and Fine, J. (2004) Estimating equations for association structures. *Statistics in medicine* **23**(6), 859–874.

Zhang, M.-J., Zhang, X. and Scheike, T. H. (2008) Modeling cumulative incidence function for competing risks data. *Expert review of clinical pharmacology* **1**(3), 391–400.

Zhang, Z. (2017) Survival analysis in the presence of competing risks. *Annals of translational medicine* **5**(3).

Zuccaro, V., Celsa, C., Sambo, M., Battaglia, S., Sacchi, P., Biscarini, S., Valsecchi, P., Pieri, T. C., Gallazzi, I., Colaneri, M. *et al.* (2021) Competing-risk analysis of coronavirus disease 2019 in-hospital mortality in a northern italian centre from smatteo covid19 registry (smacore). *Scientific reports* **11**(1), 1–10.

Mohammad Anamul Haque

CURRICULUM VITAE

Personal Details

Date and Place of Birth: December 13th, 1982 – Moulvibazar (Bangladesh)
Nationality: Bangladeshi

Contact Information

University of Padova, Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.
Tel. +39 3246199269; +44 7888407932
e-mail: mohammadanamul.haque@studenti.unipd.it; enamsust82@gmail.com (personal)

Current Position

Since December 2018 (expected completion: June 2022)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: *Model-based prediction on the cumulative incidence functions and sample size calculation for competing risks survival data.*
Supervisor: Prof. Giuliana Cortese

Research interests

- Clinical trial
- Sample size computations
- Survival analysis
- Statistical softwares

Education

October 2016 – September 2018
Master degree in Bio-statistics.
University of Hasselt, Belgium
Title of dissertation: "User friendly Interface for gsDesign"
Supervisor: Prof. dr. Tomasz Burzykowski
Co-supervisor: dr. Emmanuel Quinaux

Spring 2013 – Spring 2014
MBA (Major in Finance).
North east university Bangladesh
Title of dissertation: "Technical effciency in banking sectors in Bangladesh"
Supervisor: Fathema Rashid Saba

February 2011 – February 2012
Diploma in Banking (Part I, part II).
Institute of Bankers Bangladesh

2004 – 2005 (Exam held in 2009)
Master degree in Statistics.
Shahjalal University of Science and Technology (SUST), Bangladesh

Title of dissertation: “Tea Production in Bangladesh: Trend, Growth & Technical Efficiency”
Supervisor: Prof. Md. Zakir Hossain
Co-supervisor: Prof. Md. Azizul Baten.

2000 – 2004 (Exam held in 2007)
Bachelor degree in Statistics.
Shahjalal University of Science and Technology (SUST), Bangladesh.

Work experience

September 2017 – June, 2018
Research work at International Drug Development Institute, Belgium.
Main responsinility: Create user friendly interface for group sequential design.

September 2010 – Septembe, 2016
Senior offcier, AB Bank Ltd., Bangladesh.
Main responsinility: Analyse and monitor credit management.

June 2010 – August, 2010
Field assistant.
Project work “Food Security Strategies of the People Living in Haor Areas: Status and Prospects.”
Funded by USAID andEuropean Commission
Main responsinility: Questionnaire follow-up.

June 2005 – July, 2008
Field assistant.
Report Card Survey at Transparency International Bangladesh (TIB) .
Main responsinility: Monitor group collaborations.

Teaching experience

February 2009 – Septembe, 2010
Statistics
Higher secendary school
Department of Statistics, Sylhet Science College, Bangladesh.

Awards and Scholarship

2018
Distinction at Masters program in Hasselt university, Belgium.
2014
Distinction at Masters program in NEUB, Bangladesh.
2009
Department First, SUST.

Computer skills

- *Good knowledge of R*
- *Good knowledge of LaTeX*
- *Good knowledge of Microsoft Office*
- *Basic knowledge of SAS*
- *Basic knowledge of HTML*

Language skills

Bengali: native; English: fluent (written/spoken).

Publications

Baten, A., Kamil, A. A., & Haque, M. A. (2010). Productive efficiency of tea industry: A stochastic frontier approach. African journal of Biotechnology, 9(25), 3808-3816.

Baten, M. A., Kamil, A. A., & Haque, M. A. (2009). Modeling technical inefficiencies effects in a stochastic frontier production function for panel data. African Journal of Agricultural Research, 4(12), 1374-1382.

Haque, M. A., Hossain, M. Z., & Alam, M. B. (2010). Concentration and elasticity measurement of selected tea companies in Bangladesh: an econometric analysis. Journal of Socioeconomic Research and Development, 7(3), 873-879.

Articles in conference proceedings

Haque, M. A., & Cortese, G. (2021). Sample Size Computation for Competing Risks Survival Data in Gs design. In Book of Short Paper SIS 2021, (Perna, C., Salvati, N. and Schirripa Spagnolo, F.) pp. 584–589, Pearson, ISBN: 9788891927361.

Working papers

Haque, M. A., & Cortese, G. (2022). Computing sample size under fixed and Gs design in competing risks survival data.

Haque, M. A., & Cortese, G. (2022). A simulation study among regression approaches in competing risks survival data.

Haque, M. A., & Cortese, G. (2022). Real data example from subjects with Covid-19 in competing risks survival data.

Conference presentations

Haque, M. A., & Cortese, G. (2021). Guidelines for Sample Size Calculation under Competing Risks Survival data in Gs-design. Royal Statistical Society (RSS), Manchester, UK, September 06-09 (2021).

Haque, M. A., & Cortese, G. (2021). Sample Size Computation for Competing Risks Survival Data in GS-Design. Preface XIX 1 Plenary Sessions, 584. Italian Statistical Society (SIS), Pisa, Italy, June 21-25 (2021).

References

Prof. dr. Giuliana Cortese
Faculty of Statistical Sciences, University of Padova
Via C. Battisti, 241 - 35121 Padua
+39 049 827 4159
email: gcortese@stat.unipd.it

Prof. dr. Geert MOLENBERGHS
I-BioStat, Director PStat®, American Statistical Association
Universiteit Hasselt, CenStat
Martelarenlaan 42, 3500 Hasselt, Belgium
tel. +32 11 26 8238
email: geert.molenberghs@uhasselt.be