TREE-STRUCTURED AND DIRECT PARAMETRIC REGRESSION MODELS FOR THE SUBDISTRIBUTION OF COMPETING RISKS

ABDUL KUDUS

DOCTOR OF PHILOSOPHY
UNIVERSITI PUTRA MALAYSIA

2008

TREE-STRUCTURED AND DIRECT PARAMETRIC REGRESSION MODELS FOR THE SUBDISTRIBUTION OF COMPETING RISKS

By

ABDUL KUDUS

Thesis Submitted to the School of Graduate Studies,
Universiti Putra Malaysia, in Fulfilment of the Requirements
for the Degree of Doctor of Philosophy

August 2008

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirements for the degree of Doctor of Philosophy

TREE-STRUCTURED AND DIRECT PARAMETRIC REGRESSION MODELS FOR THE SUBDISTRIBUTION OF COMPETING RISKS

By

ABDUL KUDUS

August 2008

Chair: Associate Professor Noor Akma Ibrahim, PhD

Faculty: Institute for Mathematical Research

Traditionally, the regression analysis for competing risks survival time is based on the cause-specific hazard that treat failures from causes other than the cause of interest as censored observations. That includes technique such as the Cox proportional hazard model. The modelling of hazard rate may or may not match the objective of investigator. It is often more desirable to investigate the subdistribution function, because cause-specific hazard doesn't obviously give the information about proportion of individuals experiencing a cause of interest. Furthermore, the subdistribution and cause-specific hazard function are not interchangeable. Thus, if we intended to draw inference from

subdistribution function, then we must model on subdistribution function directly or indirectly.

Sometimes, we do not only intend to investigate the relationship between response and covariates through regression analysis, but also we want to identify the presence of subgroup of individuals in our data. We could then utilize tree-structured regression for this purpose.

In this thesis, we developed statistical methods for competing risks data analysis through direct, indirect and parametric subdistribution modelling. Indirect model is employed via hazard of subdistribution. Evaluation of the performance of proposed methods is conducted through series of simulation studies as well as real data application.

We developed four methods: 1) a method to categorize continuous covariate by considering the competing risks survival time outcome variables, called outcome-oriented categorization method, 2) a tree-structured competing risks regression to extract meaningful sub-groups of subjects determined by the value of covariates, 3) a hybrid model which boost the available subdistribution hazards regression by

augmenting it with tree-structured regression resulted from the previous step, 4) two kinds of parametric direct subdistribution model. These models are constructed based on non-mixture cure model. The first model is developed by taking into account the fraction of individuals who did not experience the event of interest in the long term. The second model is developed by reparameterizing the first model in order to mimic Gompertz distribution which allows no immune fraction.

Research finding is as follows: 1) Method of outcome-oriented categorization based on deviance statistic is the best. The application of the method to contraceptive discontinuation data showed good result. 2) Regression tree for competing risks data can uncover the structure of data and yield the sub-group of individuals with a clear description based on their covariates. The application of the method to contraceptive discontinuation data showed good result. Extensive Monte Carlo simulation suggests the method has good performance in identifying the structure of data. 3) Application of the hybrid model to the contraceptive discontinuation data showed that the hybrid model is better than the available subdistribution regression in terms of AIC. 4) By using some well known kernel distribution, the parametric direct subdistribution models are developed. The

maximum likelihood estimations are carried out simultaneously for all causes of event. In Bone Marrow Transplantation (BMT) data analysis, the first proposed model gave noticeably good fit to the nonparametric counterpart. The second proposed model is fitted to contraceptive discontinuation data and showed that Gompertz-like subdistribution with Gompertz kernel is the best fit.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

MODEL REGRESI BERSTRUKTUR POKOK DAN BERPARAMETER LANGSUNG BAGI SUBTABURAN RISIKO BERSAING

Oleh

ABDUL KUDUS

Ogos 2008

Pengerusi: Profesor Madya Noor Akma Ibrahim, PhD

Fakulti: Institut Penyelidikan Matematik

Analisis regresi bagi masa mandirian dengan risiko bersaing biasanya berdasarkan kepada bahaya sebab-spesifik yang memperlakukan kegagalan kerana peristiwa bersaing sebagai tertapis. Ianya termasuk teknik seperti model bahaya berkadaran Cox. Pemodelan kadar bahaya mungkin atau tidak mungkin sesuai dengan tujuan penyelidik. Sering kali diinginkan untuk menyelidiki fungsi subtaburan, disebabkan fungsi bahaya sebab-spesifik tidak memberikan maklumat yang jelas tentang kadar dari individu yang mengalami punca yang diperhatikan. Tambahan pula, fungsi subtaburan dan fungsi bahaya sebab-spesifik tidak boleh ditukarganti. Dengan demikian, jika kita bermaksud

melakukan pentakbiran fungsi subtaburan, maka kita perlu melakukan pemodelan bagi fungsi subtaburan.

Kadang-kala, kita tidak hanya bermaksud untuk menyelidiki hubungan antara pembolehubah bersandar dengan pembolehubah peramal, tetapi juga kita ingin mengenali kehadiran subkumpulan individu dalam data. Kita dapat menggunakan regresi berstruktur-pepohon untuk maksud ini.

Dalam tesis ini, kami membina kaedah berstatistik bagi analisis data risiko bersaing melalui pemodelan subtaburan langsung, tidak langsung dan berparameter. Model tidak langsung dibina melalui bahaya subtaburan. Penilaian prestasi dari kaedah yang dicadangkan dilakukan melalui simulasi dan penerapan data nyata.

Kami membina empat kaedah: 1) kaedah untuk mengkategori pembolehubah selanjar dengan mempertimbangkan pembolehubah bersandar masa mandirian risiko bersaing yang disebut dengan kaedah pengkategorian berorientasi kesudahan, 2) regresi risiko bersaing berstruktur pepohon untuk memaparkan subkumpulan individu yang bermakna yang ditentukan oleh nilai-nilai pembolehubah peramal, 3) model

hibrid yang memperkembangkan regresi bahaya subtaburan yang sedia ada dengan penambahan regresi berstruktur pepohon yang dihasilkan pada langkah terdahulu, 4) dua jenis model subtaburan langsung berparameter. Model-model berkenaan dibina berdasarkan model pulih tak campur. Model pertama dibina dengan mempertimbangkan pecahan individu yang tidak mengalami peristiwa yang diperhatikan dalam penggal yang panjang (pecahan imun). Model kedua adalah pemparameteran dari model pertama agar menyerupai taburan Gompertz yang membenarkan tiada pecahan imun.

Penemuan-penemuan dari disertasi ini adalah: 1) Kaedah pengkategorian berorientasi kesudahan berdasarkan statistik devians adalah yang terbaik. Penerapan dari kaedah ini untuk data pemutusan alat pencegah kehamilan menunjukkan hasil yang baik. 2) Regresi pepohon bagi data risiko bersaing dapat menggali struktur data dan menghasilkan subkumpulan individu yang mempunyai gambaran yang jelas berdasarkan nilai-nilai pembolehubah peramalnya. Penerapan kaedah ini pada data pemutusan alat pencegah kehamilan menunjukkan hasil yang baik. Simulasi Monte Carlo memperlihatkan bahwa kaedah ini mempunyai prestasi yang baik dalam mengenali struktur data. 3) Penerapan model hibrid pada data pemutusan alat pencegah

kehamilan menunjukkan bahwa ianya lebih baik daripada regresi subtaburan sedia ada. 4) Pembinaan model subtaburan langsung berparameter dilakukan dengan menggunakan beberapa taburan kernel yang sudah dikenali. Penganggaran kebolehjadian maksimum dilakukan secara serentak bagi semua punca peristiwa. Dalam analisis data pencedungan sumsum tulang, model pertama yang dicadangkan memberikan lengkung suaian yang baik terhadap tandingan tak berparameternya. Model kedua disuaikan pada data pemutusan alat pencegah kehamilan dan menunjukkan bahwa subtaburan seakan Gompertz dengan kernel Gompertz adalah yang terbaik.

ACKNOWLEDGEMENTS

In The Name of ALLAH, The Most Merciful and Most Beneficent

All praises do to Allah, Lord of the universe. Only by His grace and mercy this thesis can be completed.

This work was carried out with a hope to contribute towards the expansion of our currently limited knowledge on survival data analysis. The completion of this thesis would have been impossible if not for the assistance and direct involvement of so many kindhearted individuals. Thus, I am very much indebted to my previous mentors and I have no way of repaying such a debt except to express my sincerest gratitude.

First and foremost, I am very grateful to my adviser Assoc. Prof. Dr. Hj. Noor Akma Ibrahim, for her strong support, guidance, and patience for the very enriching and thought provoking discussions and lectures which helped to shape the thesis. She was always there to provide everything I needed in the laboratory. I would also like to thank her for providing financial support during the period of my study through IRPA research fund and Fundamental Research Grant Scheme.

I am also grateful to Assoc. Prof. Dr. Mohd. Rizam Abu Bakar and Assoc. Prof. Dr. Isa Daud in their capacities as members of Supervisory Committee. Thank you for the comments and suggestions, which contributed a lot towards the improvement of the final manuscript. I am also indebted to the staff of the Institute for Mathematical Research, Universiti Putra Malaysia for their help and cooperation.

Special thanks are extended to Dean of Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA) and Rector of Universitas Islam Bandung who allowed me to study at the PhD level. Special thanks are also extended to other INSPEM's postgraduate room members who helped me in every way possible. Acknowledgement is also extended to Indonesian Student Association (PPI-UPM) that joined us in sweet friendship and made life easier during my stay in Malaysia.

I wish to express my deepest gratitude to my parents, brothers and sisters for their prayers, continuous moral support and unending encouragement. Last but not least, I wish especially to acknowledge my beloved wife, Rela Umul Hasanah, and my dearest daughter Haifa Qathrunnada for their love, support, patience and understanding.

I certify that an Examination Committee met on 25th August 2008 to conduct the final examination of Abdul Kudus on his Doctor of Philosophy thesis entitled "Tree-structured and Direct Parametric Regression Models for the Subdistribution of Competing Risks" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Examination Committee were as follows:

Malik bin Hj. Abu Hassan, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Habshah Midi, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Kassim Haron, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

M. Ataharul Islam, PhD
Professor
Department of Statistics
University of Dhaka
Bangladesh
(External Examiner)

HASANAH MOHD. GAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 29 January 2009

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Noor Akma Ibrahim, PhD
Associate Professor
Institute for Mathematical Research
Universiti Putra Malaysia
(Chairman)

Isa Daud, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

Mohd. Rizam Abu Bakar, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

_____
HASANAH MOHD. GAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 12 February 2009

DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Universiti Putra Malaysia or at any other institution.

ABDUL KUDUS

Date: 24 November 2008

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criteria
AID	Automatic Interaction Detection
ALL	Acute Lymphoblastic Leukemia
AML-high	Acute Myeloctic Leukemia high-risk second remission or untreated first relapse
AML-low	Acute Myeloctic Leukemia low-risk first remission
BMT	Bone Marrow Transplant
CART	Classification and Regression Trees
c.d.f	Cumulative distribution function
CGVHD	Chronic Graft versus Host Disease
IDHS	Indonesian Demography and Health Survey
IUD	Intra Uterine Device
LAD	Least Absolute Deviation
LS	Least Square
MLE	Maximum likelihood estimation
MSPE	Mean Squared Prediction Error
RMSE	Root Mean Square Error
SS	Sum of Squared residuals

CHAPTER 1

INTRODUCTION

1.1 Background

Survival analysis is the phrase used to describe the analysis of data that correspond to the time from a well-defined time origin until the occurrence of some particular events or end-points. It is important to state what the event is and when the period of observation starts and finish. In medical research, the time origin will often correspond to the recruitment of an individual into an experimental study, and the end-point is the death of the patient or the occurrence of some adverse events. Survival data are rarely Normally distributed, but are skewed and comprise typically of many early events and relatively few late ones. It is these features of the data that necessitate the special method *survival analysis*.

The specific difficulties relating to survival analysis arise largely from the fact that only some individuals have experienced the event and, consequently, survival times will be unknown for a subset of the study group. This phenomenon is called censoring and it may arise in the following ways: (a) a patient has not (yet) experienced the relevant outcome, such as relapse or death, by the time the

study has to end; (b) a patient is lost to follow-up during the study period; (c) a patient experiences a different event that makes further follow-up impossible. For example, a typical animal study or clinical trial starts with a fixed number of animals or patients to which a treatment is applied. Because of time or cost considerations, the investigator may terminate the study or report the result before all subjects realize their events. In this instance, if there are no accidental losses or subject withdrawals, all censored observations have times equal to the length of the study period. Generally, censoring times may vary from individual to individual. Such censored survival time underestimated the true (but unknown) time to event. Visualising the survival process of an individual as a time-line, their event (assuming it were to occur) is beyond the end of the follow-up period. This situation is often called *right censoring*. Most survival data contain right censored observation.

In general, the presence of censoring warrants special methods of analysis whereby standard graphical methods of data exploration and presentation, notably scatter diagram, cannot be used.

The statistical issues become more complicated in studies that have multiple end-points. Here, a unit is exposed to several risks at the same time, but eventual failure of the unit is due to only one

of these risks, we call this the competing risks setting. Consider the example of Hoel (1972), based on a laboratory experiment in which mice were given a dose of radiation at 6 weeks of age. The causes of death were recorded as Thymic Lymphoma, Reticulum Cell Sarcoma, or other. Another example is from a study of breast cancer patients (Boag, 1949), where the cause of death was recorded as “cancer” or “other”.

More examples can be obtained from various field of research. In contraceptive discontinuation studies, common competing risks are failure and abandonment. With respect to failure, contraceptive abandonment is competing risk event. Similarly, with respect to abandonment, failure is competing risk event. Each of the two endpoints, failure and abandonment are of interest. In engineering application, competing risks arise in the analysis of series systems of components. Here, failure of any of the components causes the system to fail. One observes the time at which the system fails and which component caused the system to fail. Based on these data, inference regarding the lifetime of a particular component is made.

The objective of competing risks data analysis is to isolate the effect of a given risk, or a subset of risks, acting on a population. According to Seal (1977), the use of competing risks dates back to

1760 and evolved out of a controversy over smallpox inoculation. Smallpox inoculation in the 1700s was administered by applying leeches to the body, a practice that could lead to acute illness and death. Physicians argued whether the benefits of inoculation outweighed the initial risk of death. Daniel Bernoulli, in a 1760 memoir entitled "Essai d'une nouvelle analyse de la mortalité causée par le petite vérole; et des advantages de l'inoculation pour le prévenir", tried to estimate the expected increase in lifespan, if smallpox were eliminated. This calculation could then be used to weigh the pros and cons of smallpox inoculation.

Similarly, in the modern treatment of competing risks we are interested in isolating the effect of individual risks, for example, when we wish to assess a new treatment for one kind of disease. In a long-term study of this treatment on a sample of individuals, some will die of causes other than this disease. The appropriate analysis of this problem must account for the competing effects of death from other causes.

1.2 Mathematical Formulation of Competing Risks

Let T_i, $i = 1,\ldots,n$ be n independent positive random variables with common continuous distribution F. Independent of T_i's, let U_i, $i = 1,\ldots,n$ be also independent positive random variable with possibly

non-continuous common distribution G, and d_i, $i = 1,\ldots,n$ be the failure type associated with T_i, where $d_i = 1,\ldots,J$. A typical competing risks problem is to make statistical inference on F based on censored observation (Y_i, D_i), defined by

$$Y_i = \min(T_i, U_i), \quad D_i = I(T_i \le U_i)$$

where $I(\bullet)$ is an indicator random variable of the specified event. The following points should be emphasized:

- The pair (T_i, d_i) from different subjects in the sample are assumed to be *iid*.
- The different failure types within each subject are not assumed to be independent.
- Each subject can experience at most one failure type.

The overall hazard function is defined by

$$l(t) = \lim_{Dt \to 0} \frac{P(t \le T_i \le t + Dt \mid T_i > t)}{Dt} \tag{1.1}$$

and the overall survival function is given by

$$S(t) = P(T_i > t) = \exp\left(- \int_0^t l(u) du \right) \tag{1.2}$$

The cause-specific hazard function is defined by

$$l_j(t) = \lim_{Dt \to 0} \frac{P(t \le T_i \le t + Dt, d_i = j \mid T_i > t)}{Dt} \tag{1.3}$$

This quantity indicates the rate at which subjects who have yet to experience any of the competing risks are experiencing the j^{th} competing cause of failure.

The failure time density function for failure type j is defined by

$$f_j(t) = l_j(t)S(t) \tag{1.4}$$

The subdistribution (cumulative incidence) for failure type j is defined by

$$\begin{aligned} F_j(t) = P(T_i \le t, d_i = j) &= \int_0^t l_j(u)S(u)du \\ &= \int_0^t l_j(u)\exp(-L(u))du \end{aligned} \tag{1.5}$$

where $L(u) = \int_0^u l(v)dv = \int_0^u \sum_{j=1}^{J} l_j(v)dv$

This is the probability of experiencing the j^{th} event in the setting where competing risks are acknowledged to exist. Note that the value of $F_j(t)$ depends not only on the rate at which the specific cause of interest is occurring, but also on the rates at which all the competing risks occur. $F_j(t)$ is not a true distribution function due to its properties: it is non-decreasing with $F_j(0) = 0$ and $F_j(\infty) = P(d = j) < 1$. These curves have a straightforward interpretation and observe that

$$S(t) = 1 - \sum_{j=1}^{J} F_j(t) \tag{1.6}$$

Sometimes investigators may be interested in identifying risk factor for a particular outcome or in comparing groups of patients after adjusting for important prognostic factors. Such kind of statistical procedure is called regression. The most commonly used regression model for analyzing survival data is the Cox proportional hazards model (Cox, 1972). The Cox model is a regression model for the hazard rate, or instantaneous risk, of a given outcome. It is often used in the presence of competing risks to model the cause-specific hazard rate (1.3). When the outcome is single endpoint, there is 1 to 1 correspondence between the hazard rate and the survival probability as estimated by the Kaplan Meier estimator. For competing risks data, this relationship does not hold, and estimates the probability that a patient has experienced the event of interest, the subdistribution function (1.5), depend on the hazard rates for all the competing risks. That is why, it is worth to model the subdistribution function directly. Hence, we model the subdistribution for failure from cause j conditional on the covariate vector Z, $F_j(t;Z) = P(T \leq t_i, d_i = j \mid Z)$.

1.3 Scope

The thesis focus is on the problem of regression methods for subdistribution of competing risks. The regression methods which model the relationship between predictor variable(s) and competing

risks survival time outcome include nonparametric exploratory regression, hybrid regression and parametric regression. In this thesis, we used tree-structured regression as a nonparametric exploratory statistical method. Tree-structured regression is combined with semiparametric regression for subdistribution of competing risks proposed by Fine and Gray (1999) to form the hybrid regression. The parametric version of regression for subdistribution of competing risks is developed based on non-mixture cure model which regards individuals who are not yet experiencing the event of interest as a cure fraction.

1.4 Problem Statements

The common approach to summarize the various endpoints in a competing risks study is to generate a series of Kaplan-Meier curve, one for each endpoint. Kaplan-Meier estimator is not an appropriate statistics when there are competing risks because it estimates the probability of the event occurring in imaginary patient who cannot experience other events. For example, a Kaplan-Meier estimator of relapse is an estimate of the probability of relapsing in a patient who can never die.

Kaplan-Meier estimator of survival function which treats competing risk events as censoring events implied that one assumes a

hypothetical latent event time for the endpoint of interest. The use of Kaplan-Meier curve in competing risks analysis relies on the strong assumption of independence between different competing risks events. This assumption is clearly violated when one uses censoring to account for nonindependent competing event, where various competing risks are often dependent. For example, those at high risk of death with chronic graft versus host disease (CGVHD) are thought to be at lowest risk of relapse (Weiden *et al.*, 1981).

Fine and Gray (1999) and Klein and Moeschberger (2003) have advocated the use of subdistribution function which takes consideration the presence of other events, regardless of independence, within a competing risks framework. Unlike the Kaplan-Meier method, the subdistribution method provides a breakdown of the expected distribution of patients into the possible endpoints, or states, at each point in time, such that the sum of individual event rates (including the "no event rate") will always be 100%. This contrast with Kaplan-Meier method, where the sum will exceed 100% (Southern *et al.*, 2006).

In most applications the effects of covariates on the competing risks probabilities are modeled through the cause-specific hazard rates (1.3), the most typical being Cox (1972) regression models

(Prentice *et al.* 1978). This model can be fitted by treating occurrences of the competing risks as censored observation. Cause-specific hazard functions are invaluable in quantifying the instantaneous risk for alive individuals. However, they may not be appropriate if one desires summary probabilities for the different causes. It is important to note that modelling the hazard rate may or may not match the objective of investigator. It is often more desirable to investigate the subdistribution function (1.5) directly, because cause-specific hazard doesn't obviously give the information about proportion of patient experiencing a cause of interest. Information on that proportion may be more relevant to the clinical management of disease.

In light of the fact that the subdistribution function (1.5) and cause-specific hazard rate (1.3) for a given risk are not interchangeable, it is of interest to investigate the effect of covariates on the subdistribution function directly. This no interchangeability means that the properties of cause-specific hazard do not translate directly into properties of subdistribution since these curves are functions of all the competing risks' cause-specific hazard rates. Besides that, the estimates of cause-specific hazard regression often do not agree with impression drawn from plots of subdistribution for each level of covariates (Klein and Andersen, 2005).

Allison (1995) and Lunn and McNeil (1995) had employed previous work on analyzing the effect of certain factors on competing risks which concentrated on examining their effect on the corresponding cause-specific hazards. As noted previously, the effect of a factor on the cause-specific hazard for a particular type of failure can be quite different than its effect on the subdistribution of that type of failure.

To our knowledge, only a few works has been done on direct modelling for the subdistribution function. Fine and Gray (1999) proposed a model for the subdistribution hazard of the subdistribution function, building on earlier work by Gray (1988) and Pepe (1991). This approach will directly assess the importance of covariates on the subdistribution curve. The modelling assumes directly that the complementary log-log of the subdistribution function is on the proportional hazard form. Fine (2001) proposed a semi-parametric transformation model for the subdistribution of a competing risk. Andersen *et al.* (2003), Klein and Andersen (2005) and Klein (2006) proposed pseudo observation approach based on multistate model representation for the competing risks. This approach is computationally extensive. Another recent approach is direct parametric regression analysis (Jeong and Fine, 2006). They

utilized a simple form of Gompertz distribution for the improper baseline subdistribution of the event of interest.

Many medical studies focus on the relationship between the time to some event, such as death or relapse, and covariates measured at the time at which therapy is initiated. When these covariates are discrete or categorical an interpretation of the effects of the covariates on outcome is relatively simple. Using a proportional hazard model the effect of a binary covariate on outcome is interpreted in terms of the relative risk of a patient with the characteristic as compared to a patient without the characteristic. When the covariate is continuous the interpretation of the effect of the covariate on outcome is more difficult. Here one typically reports the relative risk of a patient with a one-unit increase in the covariate. Most clinical investigators would rather have the continuous covariate converted into a binary covariate reflecting high and low risk values of the covariates. While this model may not be optimal for a continuous covariate, it is the model that is most often reported in the medical literature. There are a number of graphical techniques (see Klein and Moeschberger, 2003), such as martingale residual plots, which can be used to check if a threshold model is correct, but quite often the decision to use such a model is made by the clinical investigator on the grounds that it is more understandable than a model which treats the covariate as

continuous. Once a decision is made to use a threshold model, the problem is to determine the cutpoint between high risk and low risk patients. In some cases, the cutpoint can be determined from the literature. Often cutpoints need to be determined from the data. Selection of the cutpoint can be made either by a data-oriented or outcome-oriented approach (Schulgen *et al.*, 1994). In the data-oriented approach, cutpoints are based on the distribution of the covariate in the study population. For example, the median could be used. The outcome-oriented approach picks a cutpoint for which the discretized covariate has the largest effect on outcome.

Some authors have proposed the method to determine cutpoints for survival data outcome (Jespersen 1986, Contal and O'Quigley 1999, Lausen and Schumacher 1992, 1996, Mandrekar *et al.* 2003, Tableman and Kim 2004) and for continuous longitudinal data (Abdolell *et al.*, 2002). By considering multiple causes of failure in competing risks data, it is naïve to use all the above methods for addressing competing risks data. The direct use of those methods may mislead the result. Hence, it is important to develop method for dealing with competing risks data. In this thesis, we propose cutpoint determination method for competing risks data by using direct modelling of subdistribution function.

Often in practice, one primary goal of competing risks survival analysis is to extract meaningful sub-groups of patients determined by the prognostic factor such as patient characteristics that are related to the cause of disease. Although, the existing competing risks regression model is powerful in studying the association between covariates and competing risks survival times, usually they are problematic in prognostic classification.

Recently a large amount of tree-structured methods have been developed for the analysis of univariate and multivariate survival data (Gordon and Olshen 1985, Segal 1988, Davis and Andersen 1989, LeBlanc and Crowley 1992 and 1993, Huang *et al.* 1998, Segal 1992, Zhang 1998, Su and Fan 2004, Gao *et al.* 2004), but there are no prognostic classification methods for competing risks survival data. Analysis of competing risks survival data is complex due to the presence of more than one cause of failure.

The following is the study that motivates our research for prognostic classification in competing risks survival data.

Example: Contraceptive discontinuation study

Contraceptive method is one kind of mode in family planning program. This program aims to decrease the rate of reproduction by means of controlled birth scheduling. While using the

contraceptive method, women are expected not to be pregnant. Thus it is important to identify characteristics that may relate to the discontinuation of contraceptive method, which may cause the pregnancy. Some statistical modelling of the contraceptive discontinuation data are proposed by Islam (1994), Karia *et al.* (1998), Ali and Cleland (1999), Steele (2003), and Steele *et al.* (2004).

As proposed by Steele (2004), the interest of this kind of research is to focus on the last episode of the time of contraceptive use until it discontinues. The outcome variables are measured in time scale from the use to the discontinuation. We focus on three types of discontinuation in a competing risks framework. The outcomes we consider are failure, contraceptive abandonment while in need of family planning, and switching to another contraceptive method. A discontinuation is defined as a contraceptive failure if the woman reported that she became pregnant while using the method. Thus, this definition includes both failures of the method itself and failure owing to incorrect or inconsistent use of the method. Adoption of different method within one month of discontinuation is classified as a method switch, whereas continuation of nonuse for one month or more is classified as contraceptive abandonment.

Clearly, contraceptive failure is of interest because it leads directly to an unintended pregnancy. Contraceptive abandonment is also an important outcome to study because it leads to immediate risk of unintended pregnancy. Method switching also may lead to an increased risk of unintended pregnancy if use of a modern method is discontinued in favor of a less effective, traditional method. Contraceptive failure is somewhat different from the other two outcomes in that it presumably is an unintentional event, whereas contraceptive abandonment and switching suggest some decision-making and choice on the part of the woman.

We consider some covariates which are supposed to be able to explain the rate of discontinuation. The important one is the contraceptive method. For this analysis, contraceptive methods were grouped into three categories: pills and injectables, IUDs and implants, and other modern methods (mainly condoms). Traditional methods and sterilization were excluded from this study. Pills and injectables were grouped together because they are both short-term hormonal methods. IUDs and implants are longer-term reversible methods that require a health worker to remove them. As such, they are fundamentally different from other reversible methods in that they require the user to be proactive to discontinue use and to have contact with the health system at the time of discontinuation. The other covariates are woman's

education (primary or lower, secondary, university), household economic status (1 – 7 scores), area of residence (urban, rural), age of the women at the start of the episode of use (years), and religion (Moslem, non-Moslem).

Unlike the ordinary competing risks models which only concern the estimation of covariate effects on the time to discontinue, the question to this study is “Which stratification group is more likely to experience each type of discontinuation?”, and the interest is in identifying small groups of woman that have similar characteristics of contraceptive discontinuation. For this purpose, we will develop tree-structured regression for competing risks outcome.

Direct modelling of subdistribution function utilized some link functions to model the linear relationship between the subdistribution function with the covariates. This link function is capable of modelling their linear effects, when it is well-known that tree-based methods are not efficient to represent linearity. The tree method is excellent at handling categorical predictors while linear regression defines dummy variables and may result in messy model form, especially when the number of categories is large. Linear regression may fail to model nonlinearity while tree methods, via step functions, often provide satisfactory approximations. Detecting interaction among covariates could be a

daunting task in linear regression while a tree model does automatic interaction detection. Perhaps, if such kind of linear regression and tree models are well combined, the resulting model is able to improve model fit without a loss of interpretability which is often a challenging work in modelling. It is of interest to obtain a hybrid model for competing risk developed from these two models.

The most widely used analyses of competing risks data in practical applications are nonparametric and semiparametric. A major advantage of this approach is that there is no need to assume an underlying distribution form for subdistribution function, which is difficult in the competing risks setting, owing to the impropriety of this function. Of course, such flexibility arises at the cost of efficiency loss relative to parametric models, especially with small sample size. On the other hand, the parametric models permit extrapolation of long-term event probabilities, which are of inherent interest and which cannot generally be identified from nonparametric and semiparametric models. Moreover, parametric regression models are amenable to formal maximum likelihood inferences. Jeong and Fine (2006) proposed a direct parametric subdistribution regression by using Gompertz distribution for baseline distribution. Gompertz distribution is used because it can exhibit an improper subdistribution which is needed for modelling

subdistribution. In this thesis we propose parametric cure model for baseline subdistribution function. Various well known kernel distribution functions can be used in cure model. Therefore, we can develop parametric subdistribution function systematically.

1.5 Research Objectives

The special feature of competing risks data which extend the single type of failure time data analysis to multiple types of failure time and the importance of relaxing the assumption of independency among types of failures made subdistribution function approach for addressing competing risks data analysis important. In addition, the development of regression model based on subdistribution function is important for accounting for predictor variable(s) in the analysis. To do so, we focus on the development regression model for the subdistribution of competing risks through four subtopics, namely (1) cutpoint determination method, (2) regression tree method, (3) hybrid method and (4) parametric model. These regression methods are alternative modelling of competing risks, beside regression for cause-specific hazards. Details on competing risks regression model are presented in Chapter 2.

In view of the problems and the importance of the studies stated in the previous section, there are five primary objectives that address the research problem.

i) To develop cutpoint determination method for competing risks data by using modelling of subdistribution function. Several methods are proposed and their performances are evaluated through simulation to select one of them as the best method for being used in the development of tree-structured method.

ii) To develop tree-structured regression for competing risks outcome based on subdistribution function. The simulation is conducted to asses its performance in identifying the subgroup of subjects contained in data.

iii) To develop a hybrid model for competing risk which is constructed by semiparametric competing risks regression based on subdistribution and its tree-structured regression counterpart.

iv) To develop a parametric direct model of subdistribution function based on parametric non-mixture cure model with plateau.

v) To develop a parametric direct model of subdistribution function based on Gompertz-like distribution.

1.6 Outline of the Thesis

Chapter 2 gives the literature review of this research work. To facilitate readers, a review of competing risks data analysis along with its subdistribution modelling which comprises test for making comparison between subdistribution function and subdistribution regression model are presented in section 2.1 and 2.2. The discussion continues with the presentation of outcome-oriented cutpoint determination method in section 2.3. We considered two methods, the cutpoint determination based on two-sample statistic and based on subdistribution regression. The key idea of the landmark is the Classification and Regression Trees (CART) (Breiman *et al.* 1984) described in 2.4. A review of the tree-structured method for survival data is presented in 2.5. The hybrid model that combined linear regression with tree-structured modelling is discussed in 2.6. Finally, the review of the parametric regression for competing risks data analysis is in 2.7.

In Chapter 3, several different ways to determine outcome oriented optimal cutpoints for competing risks are compared and discussed. The methods can be classified into two-sample statistic and subdistribution regression method. We carried out some simulation works to select the optimal method. The optimal method is then applied to contraceptive discontinuation data.

Bootstrap validation and permutation test are also performed to evaluate the resulted cutpoint.

In Chapter 4, a computationally convenient method is proposed to extend CART algorithm to competing risks survival data. Trees are developed in a conventional way. At each split only the subset of data contained in a given node is considered and a local best split is searched by evaluating all allowable candidates in the current node. We assume a conditional proportional hazard for subdistribution of competing risk structure between the two derived daughter nodes, with the best split to be chosen such that the separation between the two daughter nodes are maximized. We assume a conditional proportional subdistribution hazard structure between the two daughter nodes derived from the same parent node but the subdistribution hazard structures between nodes from different parent may not be proportional. Therefore, we call these survival trees as node specific baseline subdistribution hazard tree because the terminal nodes do not share the common baseline subdistribution hazard. The proposed method is exemplified by contraceptive discontinuation data and its performance is also evaluated by data generated from proportional hazard for subdistribution of competing risks models.

In Chapter 5, we extend the hybrid model to accommodate competing risks regression. Firstly, the best semiparametric subdistribution regression model is selected by means of AIC statistic. Then, we augment it with tree-structured regression model. We illustrate the proposed procedure with application to contraceptive discontinuation data.

In Chapter 6, two new parametric subdistribution regression models are proposed. First, we consider subdistribution of competing risks modelling using parametric cure model. Parametric cure model is adopted for subdistribution function because it has cure fraction parameter which is similar with the proportion of individuals who do not experience the event of interest in the competing risks framework. Moreover, cure model can be developed based on well known distribution function. So, its associated improper subdistribution function can be constructed systematically. Second, a Gompertz-like subdistribution is developed to relax the presence of cure fraction parameter. Sometimes the events occur at fairly steady rate over the entire time period of observation. One would expect that this subdistribution curve would plateau at later times. Subdistribution of such event is better described by a proper distribution. Here, we propose a reparameterization procedure for parametric cure model to take the advantage of Gompertz

distribution which can exhibit proper and improper distribution. Parametric inference is conducted by means of maximum likelihood function. A simulation study is used to evaluate the efficiency of the parameter estimates. We illustrate the first model using Bone Marrow Transplant Data, and contraceptive discontinuation data for the second model.

Finally, in chapter 7 a summary of the research work is given and several considerations for further researches are also listed.

CHAPTER 2

LITERATURE REVIEW

In the previous chapter, the competing risks problem has been defined and a brief introduction has been given. In this chapter, the literature related to regression for subdistribution of competing risks will be reviewed. The discussion will begin with the introduction of some basic statistical quantities in competing risks setting and their estimations based on a censored sample in Section 2.1 and followed by competing risks modeling based on subdistribution function in Section 2.2. Then, in Sections 2.3 and 2.4, the outcome-oriented cutpoint determination method and Classification and Regression Tree (CART) will be respectively discussed. The CART extension to survival time data, known as survival tree, will be presented in Section 2.5 followed by its extension for handling multivariate survival time data in Section 2.6. The combination of Cox proportional hazard model and regression tree, called tree-augmented regression tree, will be discussed in Section 2.7 and finally, parametric modeling of regression for competing risks will be presented in Section 2.8.

2.1 Competing Risks

Often in life-testing situation, failure of an individual can be identified as one or more of J ($J \geq 2$) mutually exclusive, but possibly dependent cause of failure. In other words, each individual is subject to J distinct risks referred to as competing risks threatening its life. Occurrence of one event precludes observation of the other events on the same individual (it is assumed that patient can fail only from one cause). Associated with cause j, there is nonnegative absolutely continuous random variable X_j representing the lifetime of individual when no other potential risks are present. Suppose the termination time of an individual is defined as the time to the first failure. Thus, lifetime of an individual is given by $T = \min\{X_1,...,X_J\}$. The available information is usually given by the pair (T,d), where d indicates the cause(s) of failure, i.e. $d = j$ if $T = X_j$. The competing risks concept can appropriately be applied to many areas of study, such as industrial reliability analysis, market transaction analysis, and clinical trial on paired organs.

A fundamental parameter in competing risks data analysis is the *cause-specific hazard rate,* defined by (1.3). The j^{th} cause-specific hazard is the rate of failure at time t from cause j

among individuals who are still alive at time t. This quantity is often called the *crude hazard rate*.

A related quantity is the *cumulative cause-specific hazard* for cause j, defined as follows:

$$L_j(t) = \int_0^t l_j(u)du, \quad j = 1,\ldots,J \tag{2.1}$$

$L_j(t)$ is also known as the *crude cumulative hazard rate*. It should be emphasized that the exponential of the negative cumulative crude hazard rate does not have a clear probabilistic interpretation and is not related to any proper survival function. Cause-specific hazard rates affect the overall hazard rate of the time to failure, l, the latter being the sum of all J cause-specific hazard rates:

$$l(t) = \sum_{j=1}^{J} l_j(t) \tag{2.2}$$

The cause-specific hazard can be derived from the joint survival function of the J competing risks, $S(t_1,\ldots,t_J) = P(X_1>t_1,\ldots,X_J>t_J)$:

$$l_j(t) = -\frac{\partial}{\partial t_j}\log\{S(t_1,\ldots,t_J)\}\Big|_{t_1=\ldots=t_J=t} \tag{2.3}$$

This relationship was derived by Gail (1975) and Tsiatis (1975). Note that the survival function of the time to failure, $T = \min(X_1, \ldots, X_J)$, is $S_T(t) = S(t,\ldots,t)$.

Marginal hazard rate for the *j*th cause of failure can be found by differentiating –log $S_j(t)$, where $S_j(t) = S(0,\ldots,0, t_j, 0, \ldots, 0)$ is the marginal survival function of the random variable X_j. When the potential failure times are independent then the marginal and crude hazard rates are identical. This need not be the case when the risks are dependent. It is also not possible to identify from competing risks data whether the failure times are independent or not because for every dependent system of $X_1, \ldots, X_J$ there is a set of independent random variables that will have the same cause-specific hazard rates. However, the independent system of risks will have different marginal distributions from the original dependent set of variables (see, for example, Klein and Moeschberger, 1987, Basu and Klein, 1982).

In competing risks problems, one is often interested in a probability which summarizes the likelihood of the occurrence of a particular competing risk. An excellent overview of the methods for summarizing competing risks failure time data is provided by Pepe and Mori (1993), Gooley *et al*. (1999), and Klein and Moeschberger (2003). Usually, three probabilities may be computed, each of them having its own interpretation. These are the crude, net, and conditional probabilities.

The probability which best describes survival experience in the presence of competing risks is the *crude probability*. The crude probability is the probability of failure from a particular cause when there are other risks acting in the population. Crude probabilities are typically expressed by the *subdistribution function*, defined by (1.5). It is a way of describing the probability distribution for a specific cause of failure in the presence of all causes. Crude probability refers to quantities derived from the probability distribution of the observable random variable, T and d, where T is time to failure, and $d = 1, \ldots, J$ is cause of failure. The subdistribution function denotes the proportion of all individuals who are observed to fail from cause j at or before time t in the presence of all cause of failure. For example, if $d = 1$ represent death from breast cancer, then the chance that a woman dies from breast cancer between ages 40 and 60 would be equal to $[F_1(60) - F_1(40)]$. Note that $F_1(\infty)$ is the proportion of individuals who will be observed to die from breast cancer, and $\sum_{j=1}^{J} F_j(t) = F(t)$ defines the distribution function for death from any cause, $F(t) = P(T \leq t)$. We denote the overall survival distribution as $S(t) = 1 - F(t)$ as defined by (1.6).

Another probability which is being often reported is the *net probability*. It is the probability of failing from a particular

cause in a hypothetical world where all other causes of failure were removed. An example of a net probability is the chance that an individual will die from heart disease in the counterfactual world where one can only die from heart disease. In the latent failure time model, this is a quantity 1 - $S_j(t)$ which is interpreted as the probability of failure from cause j by time t if it is impossible to experience another failure. However, these are rarely the probabilities of clinical interest and generally should not be considered in the competing risks problems.

The third probability used to summarize competing risks data is the *conditional probability* function for the competing risks. For a particular risk, j, let F_j and F_{-j} be the subdistribution functions for risk j and for all other risks lumped together, respectively. Then the conditional probability function is defined by

$$CP_j(t) = \frac{F_j(t)}{1 - F_{-j}(t)} \tag{2.4}$$

This quantity represents conditional probability of failure from the jth cause occurring by time t given the patient does not fail from other causes prior to t.

An alternative formulation of the competing risks problem is in terms of a multistate model. It was originally proposed by

Prentice *et al.* (1978) and recently discussed by Andersen *et al.* (2002). This approach does not require the construction of potential failure times for each cause of failure and removes some of the confusion between the crude and net probabilities of occurrence of a competing risk. In the multistate model which is illustrated in Figure 2.1, there are $J + 1$ states a subject may be in at any given point in time. The initial state 0 is transient and is the state that the subject is alive. The other J states are absorbing states corresponding to failure from cause j, $j = 1, \ldots, J$.

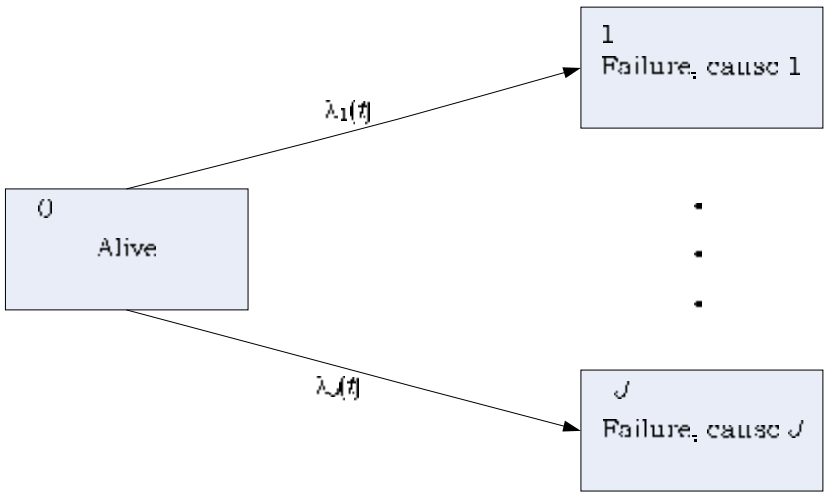

Figure 2.1. The competing risks model

Here, the transition intensities l_j, $j = 1,\ldots, J$, are the cause-specific hazard rates defined by (1.3) and conditional probabilities defined by

$$P_{0j}(s,t) = P(\text{state } j \text{ at time } t \mid \text{state 0 at time } s),\ s < t,$$

are transition probabilities. Probability $P_{00}(0,t)$ is the probability of a subject being alive at time t and $P_{0j}(0,t)$ represents the probability of failure of cause j before time t. Note that $P_{0j}(0,t)$ is the subdistribution probability of failure of type j by time t and is quantified by the subdistribution function for the cause j evaluated at time t.

The latent failure time approach and multistate model formulation form the basis for the analysis of competing risks data and enables us to look at the problem from different perspectives.

In the following passage we will consider estimation of the basic competing risks quantities based on a censored sample of competing risks data where each subject may fail due to one of J ($J \geq 2$) causes. For each of n subjects, we observe a pair of random variables (T_i, d_i), where T_i is an on-study time, and d_i is an indicator of the cause of removal from the study defined as follows:

$$d_i = \begin{cases} 0, & \text{if observation was censored} \\ j, & \text{if individual } i \text{ failed from caused } j, \text{where } j = 1,\ldots,J \end{cases} \quad (2.5)$$

For further developments, it is convenient to introduce the counting process notation. A formal and rigorous survey of counting processes and their applications can be found in books

by Andersen *et al.* (1993) and Fleming and Harrington (1991). Here, we will introduce the notation and approaches to be used in the sequel.

The counting process notation replaces the time and censoring indicator (T_i,d_i) with two functions of time: $N^i(t)$, which counts the number of times the unit has been observed to "fail" by time t, and $Y^i(t)$, which is 1 when the unit is under observation and 0 otherwise.

For ordinary survival data this means $N^i(t) = 0$ and $Y^i(t) = 1$ for $t < T_i$, $N^i(t) = \Delta_i$ and $Y^i(t) = 1$ for $t = T_i$, and $N^i(t) = \Delta_i$ and $Y^i(t) = 0$ for $t > T_i$. The notation $dN^i(t)$ means the jump in N^i at time t . This is zero except at the time of a failure, when it is 1.

As a final completion, integral notation is used to indicate sums over a time point. For example, the notation $\int_0 Z_i \, dN^i(t)$ means the sum of $Z_i \times dN^i(t)$ over all time points. $dN^i(t)$ is defined as

$$dN^i(t) = \begin{cases} 1, & \text{at the time of failure } T_i \\ 0, & \text{otherwise} \end{cases}$$

Thus

$$Z_i \times dN^i(t) = \begin{cases} Z_i, & \text{if the unit fails at time } T_i \\ 0, & \text{if the unit is censored} \end{cases}$$

To express it in another way, let the process $Y^i(t) = I(T_i \geq t)$ be an indicator of unit i being at risk just before time t. Then the total number of units at risk at time t is $Y(t) = \sum_{i=1}^{n} Y^i(t)$.

For competing risk problem, consider a counting process:

$$N_j^i(t) = I(T_i \leq t, d_i = j), \text{ for } i=1,2,\ldots,n$$

Note that $N_j^i(t)$ is a step function, which is zero until unit i fails from cause j and then jumps to one. The process $N_j(t) = \sum_{i=1}^{n} N_j^i(t)$ is also a counting process which simply counts the number of failures of type j in the sample at or prior to time t. Throughout, a subscript replaced by "•" will denote summation over that index. After adopting this notation, the total number of failures by time t is $N_\bullet(t) = \sum_{j=1}^{J} N_j(t)$.

In the counting process notation, the data (T_i, d_i), $i = 1,\ldots,n$, are represented by $\{Y_i(\bullet), N_j^i(\bullet)\}$, $i = 1,\ldots, n$, $j = 1,\ldots, J$.

The crude cumulative hazard rate can be estimated by the Nelson-Aalen estimator (Nelson, 1972, Aalen, 1978) defined as follows:

$$\hat{L}_j(t) = \int_0^t \frac{dN_j(u)}{Y(u)} \qquad (2.6)$$

This is the usual Nelson-Aalen estimator one obtains if failures from any cause other than the cause of interest are treated as censored observations. The estimated variance of this estimator (Aalen, 1978) is given by

$$\hat{Var}\left[\hat{L}_j(t)\right] = \int_0^t \frac{dN_j(u)}{Y^2(u)} \tag{2.7}$$

Estimates of the crude hazard rate itself can be found by a smoothing technique such as the kernel smoothing proposed by Ramlau-Hansen (1983) (see Klein and Moeschberger 2003, for details).

While the estimates of the crude hazard rates are helpful in understanding the failure mechanism, they do not directly lead to estimators of the competing risks probabilities. Subdistribution function is the main quantity used to draw inference about competing risks data. The estimator of the subdistribution function for cause *j* at time *t* is defined by

$$\hat{F}_j(t) = \int_0^t \hat{S}(u^-) d\hat{L}_j(u), \quad j = 1,\ldots,J \tag{2.8}$$

where $\hat{S}$ is the Kaplan-Meier estimator for the overall survival function *S* obtained by treating any one of the competing risks as an event:

$$\hat{S}(t) = \prod_{s \le t}\left[1 - \frac{dN_\bullet(s)}{Y(s)}\right] \tag{2.9}$$

and $\hat{L}_j$ is the Nelson-Aalen estimator for the cumulative cause-specific hazard L_j obtained by formula (2.6) where failure from the *j*th cause is considered as an event.

$\hat{F}_j(t)$ provides an estimate of the probability of the failure of type j prior to time *t* where a subject is at risk for experiencing any of the J competing risks. This estimator was first proposed by Kalbfleish and Prentice (1980) and recently discussed in Satagopan *et al.* (2004). The nonparametric estimation of $\hat{F}_j(t)$ is is carried out in a two-step process as illustrated in Appendix A.

Andersen *et al.* (1993) provides an estimator of the variance of $\hat{F}_j(t)$:

$$\begin{aligned} \hat{Var}\left[\hat{F}_j(t)\right] = &\int_0^t \hat{S}^2(u^-)\left[\hat{F}_j(t) - \hat{F}_j(u)\right]^2 \frac{dN_\bullet(u)}{Y^2(u)} \\ &+ \int_0^t \hat{S}^2(u^-)\left[1 - 2\left\{\hat{F}_j(t) - \hat{F}_j(u)\right\}\right]\frac{dN_j(u)}{Y^2(u)} \end{aligned} \tag{2.10}$$

In reporting the results of a study, investigators often try to describe all failure types separately. The method most frequently employed is the complement of the usual Kaplan-Meier estimator. Using this approach the estimated probability of failure from cause *j* before time *t* is $1-\hat{S}_j(t)$, obtained by treating deaths from the cause *j* as events, and occurrences of

other failures as censored observations. This is an estimate of the net probability which has interpretation usually irrelevant in dealing with competing risks. Various authors have criticized the use of the Kaplan-Meier estimator in the context of competing risks (Pepe and Mori, 1993, Gooley *et al.*, 1999). The Kaplan-Meier estimator only depends on the rates of the cause of interest and does not depend on the rates of the occurrence of other competing risks. Under these circumstances, the Kaplan-Meier estimator is biased, especially if competing events are not independent (see Klein and Moeschberger, 1984). This estimator also has no meaning in the multistate model formulation for competing risks. Despite all the criticism, inappropriate use of the complement of the Kaplan-Meier estimator to represent the probability of occurrence of one out of several endpoints is quite common. The misuse of these methods for estimation purposes stems from a lack of thorough understanding among researchers of the assumptions required to obtain interpretable Kaplan-Meier estimates. A lack of knowledge concerning the mechanics of calculating these estimates may likely to contribute to the misuse. This may lead to incorrect results while making further conclusions, for example, evaluating effect of some factors on survival time, comparing time to specific event in several groups of patients, etc.

2.2 Modelling Based on Subdistribution

The subdistribution function is the primary measure summarizing the likelihood of a specific event in the competing risks setting. Differences in the subdistribution curves would reflect differences in the probabilities of a specific event being observed in distinct populations in the presence of other competing risks. Standard inference has been based on the cause-specific hazards with the main focus being on the cause of interest while occurrences of other events are treated as censored observations. Statistical methods for comparing cause-specific hazards are easy to apply and are routinely used in practice. However, differences in crude hazard rates for a particular risk do not translate directly into differences between subdistribution curves since these curves are functions of all the competing risks' cause-specific hazard rates. As an example of this, suppose there are two types of failure, j=1,2, two groups of subject, k=1,2, and suppose all cause-specific hazards are constant, with the cause-specific hazard for both types of failure being $l_{11} = l_{21} = 3$ in group 1, while in group 2, $l_{12} = 2$ for type 1 failure and $l_{22} = 1$ for type 2 failure. Then the subdistribution functions for type 1 failure are $F_{11} = (1 - e^{-6t})/2$ in group 1 and $F_{12} = 2(1 - e^{-3t})/3$ in group 2, so $F_{11}(t) < F_{12}(t)$ for $t > (\log 3)/3$

even though $l_{11} > l_{12}$ (see Figure 2.2). As a consequence, the hypothesis of equality of the subdistribution functions for failures of a specific type is not equivalent to the hypothesis of equality of the cause-specific hazard functions for failures of that particular type. In this section, we will present two techniques for comparing subdistribution functions, namely 2-sample statistic and regression modeling.

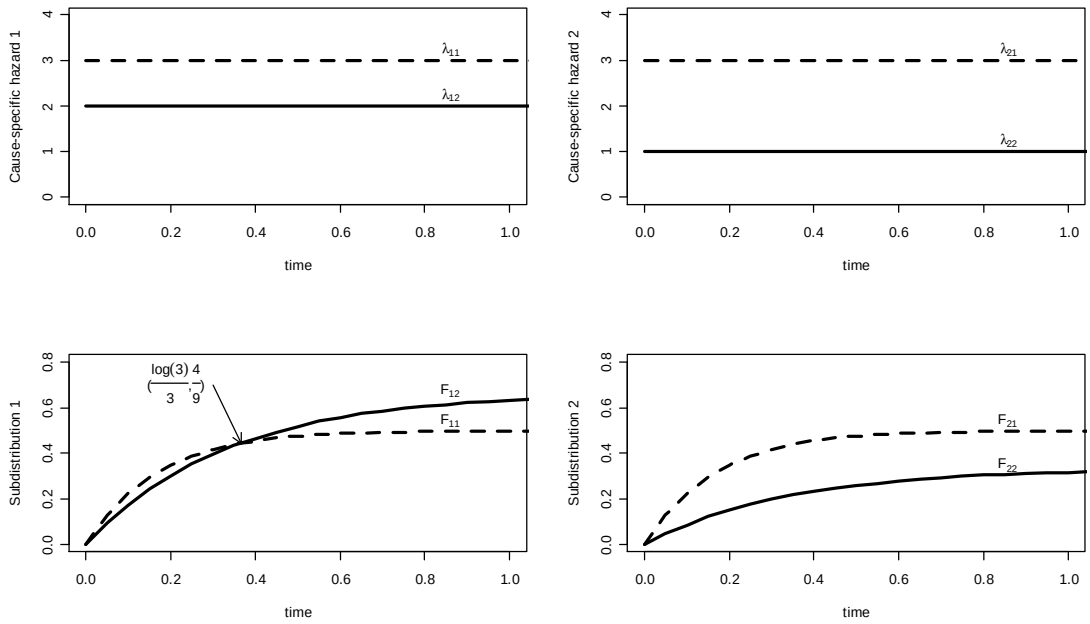

Figure 2.2. The unequivalency between cause-specific hazard and subdistribution.

2.2.1 Hypothesis Testing for Comparison across Populations

In this subsection, we will use the same notation as before. It will be assumed that there are only two types of failure ($J = 2$). The failure type of special interest is taken to be type 1. Consider a problem of comparing the subdistribution functions for the

cause of interest among K ($K \geq 2$) populations. Comparison of subdistributions across populations could be made either directly (Gray 1988, Pepe *et al.* 1993) or using regression model for the subdistribution hazard, usually through a proportional hazard model (Fine and Gray, 1999). Both procedures are presented in Appendix B and C, respectively.

In the `cmprsk` package in R 2.7.1 statistical software, we are able to apply the `cuminc` and `crr` function to compare the subdistribution function (Gray, 1988) and to fit the proportional subdistribution regression model described in Fine and Gray (1999), respectively.

2.2.2 Hypothesis Testing Based on Regression Models

After fitting proportional subdistribution hazard regression models, focus usually shifts on testing a hypothesis about the parameter vector b. In this section, we will focus on the composite hypothesis where a subset of the b's is an object of interest. The hypothesis then is H_0: $b_1 = b_{10}$, where $b = (b_1, b_2)$. Here, b_1 is a $p \times 1$ vector of the coefficients of interest and b_2 is the vector containing the remaining q components of the parameter vector b. We also partition the Information matrix I into

$$\mathrm{I} = \begin{pmatrix} \mathrm{I}_{11} & \mathrm{I}_{12} \\ \mathrm{I}_{21} & \mathrm{I}_{22} \end{pmatrix},$$

where I_{11} is of dimension $p \times p$, I_{22} is of dimension $q \times q$, I_{21} is $p \times q$, $\mathrm{I}_{21}^{/} = \mathrm{I}_{21}$. Notice that partitioned information matrix has an inverse which is also a partitioned matrix with

$$\mathrm{I}^{-1} = \begin{pmatrix} \mathrm{I}^{11} & \mathrm{I}^{12} \\ \mathrm{I}^{21} & \mathrm{I}^{22} \end{pmatrix} \tag{2.11}$$

The Wald test of H_0: $b_1 = b_{10}$ is based on the estimators of parameter vector b, $\hat{b}$. The following quadratic form defines the statistic:

$$X_W^2 = \left(\hat{b}_1 - b_{10}\right)' \left[\mathrm{I}^{11}\left(\hat{b}\right)\right]^{-1} \left(\hat{b}_1 - b_{10}\right) \tag{2.12}$$

where I^{11} is $p \times p$ principal submatrix of I^{-1} as defined by formula (2. 11).

2.3 Outcome-oriented Cutpoint Determination

The other topic that should be presented in the literature review is on the determination of cutpoint in analyzing survival time data. The available cutpoint determination methods for various types of outcome data are reviewed in this section. There are two different ways to decide optimal cutpoints based on the relationships between independent variable and outcome variables. First, some

researchers prefer to use two-sample statistic for comparing two groups of data. The other methods are based on statistic from regression analysis. Both methods can be applied in the analysis of competing risks survival data.

2.3.1 Cutpoint Determination Based on Two-sample Statistic

Cutpoint is searched along covariate Z which give us the largest difference between individuals in the two data-defined groups. That is, the outcome of the groups with $Z < g$ is as different from the group with $Z \geq g$ as possible based on some statistics.

Since the cutpoint is unknown, it is of interest to estimate and test a hypothesis about the cutpoint. A straightforward approach is to use a properly standardized maximum selected rank statistic as the test statistic. The null hypothesis H_0 for analyzing the significance of a cutpoint is that the event $Z < g$ has no influence on the distribution of T for all g:

$$H_0: P(T \leq t \mid Z < g) = P(T \leq t \mid Z \geq g) \text{ for all } t, g \in R$$

Under this null hypothesis, the standardized form of the log rank statistic allows the estimation of g using a maximization technique. To provide a reasonable amount of data in both categories and to allow the asymptotic argument cited by Lausen and Schumacher

(1992), the hypothetical cutpoint g will be restricted to an interval and the sample quartiles for the bounds of the interval will be used; i.e. $g \in \left[F_n^{-1}(e_1), F_n^{-1}(e_2)\right]$, where $0 < e_1 < e_2 < 1$. Therefore the maximally selected rank statistic is defined as the maximization of the log rank statistics in the range between $F_n^{-1}(e_1)$ and $F_n^{-1}(e_2)$. This estimate may not be unique. The minimum of this maximally selected hypothetical cutpoints can be used for the cutpoint estimate as suggested by Lausen and Schumacher (1992).

Mandrekar *et al.* (2003) proposed outcome-oriented approaches for cutpoint determination methods in survival analysis based on log-rank. They developed SAS code for the implementation and found that both methods gave the same cutpoint. However, they warned that sometimes the estimated cutpoint is close to the boundary which may be real or may be due to the presence of outlier. That situation is called end-cut preference which can be avoided by using trimming (LeBlanc and Crowley, 1993).

Abdolell *et al.* (2002) proposed cutpoint determination for continuous longitudinal outcome data. The split corresponding to maximum deviance difference is selected as cutpoint.

O'Brien (2004) proposed a new approach for choosing the number of categories and the location of category cutpoints when a continuous exposure variable needs to be categorized to obtain tabular summaries of the exposure effect. The optimum categorization is defined as the partition that minimizes a measure of distance between the true expected value of the outcome for each subject and the estimated average outcome among subjects in the same exposure category. To estimate the optimum partition, an efficient nonparametric estimate of the unknown regression function is substituted into a formula for the asymptotically optimum categorization. Although categorization is a generally inefficient method of smoothing data, he showed that information loss could be substantially reduced by choosing the cutpoints adaptively.

2.3.2 Cutpoint Determination Based on Regression Model

Mandrekar (2003) develop cutpoint determination method based on log-likelihood of Cox regression. The result showed that optimal cutpoint based on this method was similar with one found by using log-rank method. Tableman and Kim (2004) also developed cutpoint determination method for survival time response based on Cox regression model. Profile likelihood proposed by van der Vaart (1998) is used for identifying cutpoint. Furthermore, bootstrap

procedure is carried out to validate the obtained cutpoint. They found that bootstrap density histogram of cutpoints showed a very similar shape to the profile likelihood.

2.4 CART

In this section, some previous works on regression tree is discussed. The review is started by reviewing CART work of Breiman *et al.* (1984) which developed for addressing continuous response variable. Breiman *et al.* (1984) focus on the least squares regression trees. In addition, the next subsection gives an overview of different approaches in an effort to extend tree methods to handle independent failure times, namely, univariate survival trees, which have been widely discussed in the literature.

The essential idea of CART is growing large tree and then prune it to obtain the best-sized tree by evaluating the subtrees of a large tree. They used a pruning algorithm to identify a sequence of nested subtrees, and then validated the performance of each subtree with test sample or resampling depending on the sample size available.

A large initial tree, G_{max}, is grown so that no significant structure would be missed. This large tree can also be used to explore the data structure.

At this stage, CART claims a node terminal if one of the following situations occurs:

- The node contains less than $n_{min,}$ say, 10 observations. This threshold sample size, $n_{min,}$ could be made on a case by case basis. However, usually it is set small enough to construct a sufficiently large initial tree.
- The node is pure. A node is *pure* if all the responses, y_i, in that node are identical or the node contains only identical covariate values.

Once we have a large initial tree, a tree of optimal size needs to be chosen from the subtrees of this large. One possibility is to consider all its possible subtrees. But this would be computationally overwhelming since the total number of subtrees increases much more rapidly as the size of the initial tree grows. The idea of CART is to obtain a small sequence of subtrees via a computationally efficient pruning method. This method is termed the *minimal cost-complexity pruning* algorithm.

From a small sequence of subtrees we need to select one or several appropriately sized trees from the nested sequence. A reasonable way is to base the selection on the mean squared prediction error (MSPE) for each subtree. However, since the MSPE depends on the scale in which the response was measured, CART uses the *relative mean squared error* to guide the tree selection.

2.5 Survival Trees

CART is not directly applicable to survival data because many observations are censored. Additionally, the major focus in survival analysis is on the survival or hazard function rather than the mean function. In extending tree based techniques to cope with univariate or independent failure times, some approaches allow the direct use of the CART procedure by defining appropriate prediction error terms, while the others have made modifications to CART in an effort to overcome the difficulties naturally associated with censored failure times.

In general, there are two approaches for the development of regression tree for survival time response. The first approach was based on minimizing within node error, and the second approach was based on maximizing between node differences. For the first approach, the appropriate measures of within-node prediction

errors must be defined first. This allows for the direct adoption of the CART algorithm. A brief survey of related literature in this approach includes Gordon and Olshen (1985), Davis and Anderson (1989), Therneau, Grambsch, and Fleming (1988), and LeBlanc and Crowley (1992).

The prediction error defined by Gordon and Olshen (1985) is based on the distance between the Kaplan-Meier estimate of the survival function, $\hat{S}$, and a step function, $\hat{d}$, which is chosen such that it has mass at most one finite point, and it minimizes the distance between any step function restricted to at most one single jump and $\hat{S}$.

The "distance" can be measured by a class of L^p Wasserstein metrics or ordinary L^p metrics between two distribution functions, F_1 and F_2. However, LeBlanc (1989) showed the L^p or L^p Wasserstein distances are quite sensitive to censoring from the simulation study.

Davis and Anderson (1989) proposed *exponential survival trees* by modelling the survival distribution with its simplest form, the exponential distribution, which assumes a constant hazard function $l_i(t) = l$.

The proposed algorithm partitions each non-terminal node on the basis of an exponential log-likelihood loss. The split selected is the partition that minimizes the loss among all possible binary splits defined by the covariates.

In a paper on exploring the properties of martingale-based residuals, Therneau *et al.* (1990) suggested the martingale residuals could be used as the response input for the standard CART as the martingale residuals themselves are rather informative. They also comment that this allows direct use of the commercially available CART program as the software for univariate survival trees is not yet widely available.

LeBlanc and Crowley (1992) developed a procedure called *relative risk trees* for obtaining tree-structured relative risk estimates for censored survival data.

Rather than minimizing within-node variability, Ciampi (1986), Segal (1988), and LeBlanc and Crowley (1993) proposed the second approach that are based on maximizing between-node difference since defining error terms is relatively difficult for survival data and usually requires considerable computation. To measure the between node difference, familiar two-sample rank statistics

designed for censored data, such as the logrank statistic, is commonly suggested. Such tree procedure is termed as *trees by goodness of split*.

Ciampi *et al.* (1986) proposed using the logrank statistic to split data. For every permissible split, they compute the log-rank statistic, which tests the difference in survival between two groups induced by the split. The split with the largest logrank statistic is selected and the data are partitioned accordingly. The procedure is repeated and the data continue to branch off until no further split produces a significant c^2 statistic. Recall that under the null hypothesis of no difference the logrank statistic follows a c^2 distribution. However, this might lead to poor results according to Breiman *et al.* (1984).

Segal (1988) first formalized the idea of growing trees via maximizing between-node difference. He proposed the Tarone-Ware (1977) or Harrington-Fleming (1982) classes of two-sample tests as splitting statistics. However, since only internal nodes have an associated splitting statistic for any tree grown by goodness of split, the cost-complexity pruning algorithm of CART cannot be directly adapted. Alternatively, Segal proposed a bottom-up pruning algorithm.

LeBlanc and Crowley (1993) also suggested splitting by the logrank statistic and they proposed a pruning algorithm analogous to CART.

2.6 Multivariate Survival Tree

Multivariate survival data arises when each subject may experience multiple failures or individuals under study are naturally clustered. The former case is termed as *multiple event times* and the latter is called *clustered failure times*. There are very few published tree methods regarding multivariate survival data. Su and Fan (2001, 2004) proposed two approaches to handle correlated failure time by using tree based methods. In the first approach, the splitting criterion is based on maximizing between-node difference in survival, where the difference is measured by a robust log rank statistic derived from the marginal approach to multivariate survival data by Wei, Lin and Weissfeld (1989). In the second approach, the data is split based on a likelihood ratio test. They introduced a gamma distributed frailty to account for the dependence among survival times.

Gao *et al.* (2004) extended CART algorithm to multivariate survival data in a similar manner as Su and Fan (2004). They proposed method intended to provide an exploratory data analysis for

multivariate survival data, and it was complimentary rather than competitive to those parametric or semi-parametric methods. The splitting rule is defined as Wald statistic evaluating covariate effect. The current node *h* is separated into two daughter nodes such that the ratio of the hazard between the two daughter nodes is maximized. Thus, they created trees by maximizing between-node separation. Segal's pruning method is utilized to avoid computational burden (Segal, 1988). This method also has the advantage to provide a sequence of candidate trees facilitating the investigators to select a proper tree with additional knowledge relevant to the scientific question.

Gao *et al.* (2006) mentioned the drawback of the previous methods. Though survival trees developed above were computationally convenient, they suffered a drawback that the overall model structure was unclear. In trees constructed by this method, all the resultant groups are completely unrelated. Instead of using the available method, they assumed a proportional hazards structure within the whole tree and thus present a clear model structure. As a consequence, a global optimal split is obtained at each partitioned because the best split is searched over the whole tree.

2.7 Tree-augmented Regression Trees

Su and Tsai (2005) proposed a hybrid model that combined the Cox proportional hazards regression with tree modelling. The proposed model is called tree-augmented Cox proportional hazards models. The motivation is the Cox proportional hazards regression and tree-structured modellings complement each other. Therefore, combining them can yield a hybrid model that provides a more efficient way to model survival data and improves fit without loss of interpretability. The resulting model provides a natural adequacy checking for the functional form specification in the best Cox proportional model.

2.8 Parametric regression for Competing Risks

Larson and Dinse (1985) postulates a mixture model that express the distribution of survival time and its corresponding cause of failure. The mixture model was composed of marginal distribution of failure type and the conditional distribution of time to failure, given type of failure. Failure type is modeled with a multinomial distribution and failure times conditioned on failure type with piece-wise exponential distribution.

Maller and Zhou (2002) proposed a similar modelling based on parametric mixture model of the joint distribution of (T,d). The subdistribution function of cause j is represented by product of conditional c.d.f of cause j and probability of failure type j.

$$F_j(t) = P(T \le t \mid d = j) \times P(d = j)$$

Conditional c.d.f $P(T \le t \mid d = j)$ is modelled by using a well-known parametric distribution and $P(d = j)$ with a multinomial distribution.

Jeong and Fine (2006) proposed two ways of full parameterization of the subdistribution without covariate, i.e. by parameterizing the survival and cause specific hazards or parameterizing the subdistribution directly. The first way is called cause-specific hazard approach, whereas the second is direct subdistribution approach.

For the cause-specific hazard approach, $l_j(t)$, j=1,...,J is modelled by using well-known parametric model. This can yield full parametric representation of $F_j(t)$. But, sometimes $F_j(t)$ cannot be expressed in explicit form due to integral equation which cannot be obtained analytically.

To parameterize the subdistribution function directly they used Gompertz distribution, which the asymptote of the subdistribution might be less than 1. They claimed that Larson and Dinse (1985) model was contained in their formulation.

Parametric regression analysis of subdistribution function is proposed by Jeong and Fine (2007). They extended the subdistribution modelling by incorporating covariate. Gompertz distribution is used for the baseline subdistribution of the event of interest.

2.9 Literature Review Summary

This chapter has reviewed the relevant published results related to the thesis. Clearly, there are cutpoint determination methods exist for continuous and single endpoint survival data, but not for competing risks problem. By utilizing two-sample test and regression for subdistribution of competing risks we will develop such kind of method for cutpoint determination in competing risks framework. Also, there are numerous regression tree methods available for continuous and single endpoint survival data with a few extension for handling multivariate and recurrence survival data, but not for competing risks problem. In addition, the effort to

combine linear regression and regression tree which yield a hybrid model that provides a more efficient way to model data and improves its fit is only available for single endpoint survival data, and we are interested in extending the method to accommodate competing risks. Lastly, the development of parametric regression for subdistribution of competing risks has not been fully developed systematically. The usefulness of cure model in the analysis of competing risks data has not been thoroughly investigated.

CHAPTER 3

OUTCOME-ORIENTED CUTPOINT DETERMINATION METHODS FOR COMPETING RISKS

In medical research, continuous variables are often converted into categorical variables by grouping values into two or more categories. The usual approach in clinical and psychological research is to dichotomize such continuous variables, whereas in epidemiological studies it is customary to create several categories allowing investigation of a possible dose-response relation. It is done to make the analysis and interpretation of results simple. Furthermore, clinical decision making often requires two classes, such as normal/abnormal, cancerous/benign, treat/do not treat, and so on.

It is quite often the decision to categorize continuous variable is made by the clinical investigator on the grounds that it is more understandable than a model which treats the covariate as continuous. Once a decision is made to use a threshold model, the problem is to determine the cutpoint between high risk and low risk patients. In some cases, the cutpoint can be determined from the literature. Often cutpoints need to be determined from the data. Selection of the cutpoint can be made either by a data-oriented or outcome-oriented approach (Schulgen *et al.*, 1994). In

the data-oriented approach, cutpoints are based on the distribution of the covariate in the study population. For example, the median could be used. The outcome-oriented approach picks a cutpoint for which the discretized covariate has the largest effect on outcome.

In the outcome-oriented approach, there are two different ways to decide optimal cutpoints based on the relationships between independent variable and outcome variables. First, some researchers prefer to use two-sample statistic for comparing two groups of data. The other methods are based on statistic from regression analysis. Both methods had been extended for addressing single type of failure time data (Jespersen 1986, Contal and O'Quigley 1999, Lausen and Schumacher (1992, 1996), Mandrekar *et al.* 2003 and Tableman and Kim 2004).

Let $(Z_1,T_1),\ldots, (Z_n,T_n)$ be numbers of n bivariate observations, and assume the marginal distribution of Z is continuous. The effect of Z on the dependent variable T is of interest, but the functional relationship between Z and T is unknown. A simple way to categorize Z is to define two groups of individuals whose Z values are either less than or equal, or greater than a certain cutpoint, g. Here we seek a cutpoint which give us the largest difference between individuals in the two data-defined groups. That is, the

outcome of the groups with $Z < g$ is as different from the group with $Z \geq g$ as possible based on some statistics.

For the case of single type of failure time response data, *T*. The procedure search all possible cutpoints; and for each cutpoint, g_k, we compute a particular statistic such as log rank statistic based on the groups defined by *Z* being less than or equal the cutpoint or greater than the cutpoint. That is, at each event time, t_i, we find the total number of deaths, d_i, and the total number at risk, r_i. We also find the total number of deaths with $Z < g_k$, d_i^+ and the total number at risk with $Z \geq g_k$, r_i^+. We then compute the log rank statistic

$$S_k = \sum_{i=1}^{D} \left[d_i^+ - d_i \frac{r_i^+}{r_i} \right] \tag{3.1}$$

where *D* is the total number of distinct death times. The estimated cutpoint $\hat{g}$ is the value of g_k which yields the maximum S_k.

For addressing continuous longitudinal outcome data, Abdolell *et al.* (2002) select the cutpoint corresponding to maximum deviance difference. The method was applied to any set of data *A* which

consists of n cases, where a cutpoint candidate g divided the cases in the parent group A into two groups, i.e. 1st and 2nd group denoted by B and C, respectively. Group B and C consist of n_B and n_C cases, respectively, where $n_B + n_C = n$.

The deviance calculated from data set A can be expressed as

$$D_A(\hat{\mu}, t) = \sum_{i=1}^{n} D(\hat{\mu}, t_i) \tag{3.2}$$

The deviance of the partition B and C can be expressed as the sum of the deviances of B and C, so that

$$\begin{aligned} D_{B,C}(\hat{\mu}_B, \hat{\mu}_C, t) &= D_B(\hat{\mu}_B, t) + D_C(\hat{\mu}_C, t) \\ &= \sum_{i=1}^{n_B} D(\hat{\mu}_B, t_i) + \sum_{i=1}^{n_C} D(\hat{\mu}_C, t_i) \end{aligned} \tag{3.3}$$

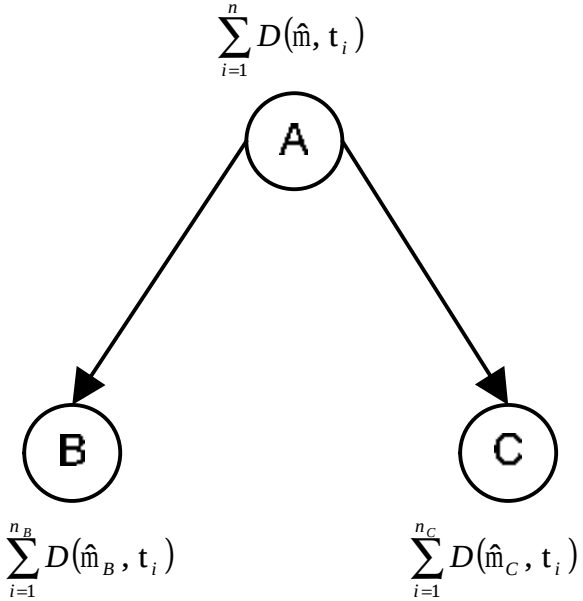

Figure 3.1. The cutpoint determination based on deviance.

For all possible cutpoint g_k which split A into B and C, the change in deviance between A and the partition B and C provides a measure of goodness-of-fit (Clark and Pregibon, 1992), and defined as

$$\begin{aligned} DD(g_k, A) &= D_A(\hat{\mu}, t) - D_{B,C}(\hat{\mu}_B, \hat{\mu}_C, t) \\ &= \sum_{i=1}^{n} D(\hat{\mu}, t_i) - \sum_{i=1}^{n_B} D(\hat{\mu}_B, t_i) - \sum_{i=1}^{n_C} D(\hat{\mu}_C, t_i) \end{aligned} \tag{3.4}$$

Take $\hat{g}$ to be the split such that

$$DD(\hat{g}, A) = \max DD(g_k, A) \tag{3.5}$$

The best split is that split on the predictor variable Z which best separates the high response values from the low ones, in another words it is that split which maximizes $DD(g_k, A)$.

The method of categorization which consider the number and the location of categories is proposed by O'Brien (2004). The purpose of the study was to obtain tabular summaries of the exposure effect, in such away that minimizes a measure of distance between the true expected value of the outcome for each subject and the estimated average outcome among subjects in the same exposure category.

Cutpoint determination based on regression model is carried by assuming model Cox proportional hazards model of the form

$$l_i(t \mid Z,g) = l_0(t)\exp\{b_g \times I(Z_i<g)\} \tag{3.6}$$

Estimate of g is based on finding the parameter at cutpoint g which maximizes a test statistics for testing H_0: $b_g = 0$. Possible tests are the score, Wald or likelihood ratio tests. For any of these tests one computes the value of test statistics for all possible value of g in the range of data. One can show that the values of the statistics change only at value of $g = Z_i$ where Z_i is an observed covariate value so that the statistics needs to be computed at only a finite number of potential values for g.

The score, Wald and likelihood ratio tests are computed based on the basic proportional hazard partial log likelihood given by

$$l(b_g, g) = \sum_{i=1}^{D} b_g I(Z_i < g) - \sum_{i=1}^{D} \log\left[\sum_{i' \in R_i} \exp\{b_g I(Z_{i'} < g)\}\right] \tag{3.7}$$

The score statistics for a fixed g is given by

$$U(b_g, g) = \frac{\partial l(b_g, g)}{\partial b_g} = \sum_{i=1}^{D} I(Z_i < g) - \sum_{i=1}^{D} \frac{\sum_{i' \in R_i} I(Z_{i'} < g)\exp\{b_g I(Z_{i'} < g)\}}{\sum_{i' \in R_i} \exp\{b_g I(Z_{i'} < g)\}} \tag{3.8}$$

and the Fisher information is given by

$$I(b_g, g) = -\frac{\partial U(b_g, g)}{\partial b_g} = \sum_{i=1}^{D} \frac{\sum_{i' \in R_i} [I(Z_{i'} < g)]^2 \exp\{b_g I(Z_{i'} < g)\}}{\sum_{i' \in R_i} \exp\{b_g I(Z_{i'} < g)\}} - \sum_{i=1}^{D} \frac{\left[\sum_{i' \in R_i} I(Z_{i'} < g) \exp\{b_g I(Z_{i'} < g)\}\right]^2}{\left[\sum_{i' \in R_i} \exp\{b_g I(Z_{i'} < g)\}\right]^2} \quad (3.9)$$

The value of b_g, $\hat{b}_g$, that maximizes (3.7) is the profile maximum likelihood estimator for the value of g. Estimate of g are found by finding the Z_i that maximizes the likelihood ratio test statistics for testing $b_g = 0$ defined by

$$LR(g) = 2\{\log[l(\hat{b}_g, g)] - \log[l(0, g)]\} \quad (3.10)$$

or by maximizing the Wald test of $b_g = 0$ given by

$$Z(g) = \hat{b}_g^2 I(\hat{b}_g, g) \quad (3.11)$$

or by maximizing the sore test given by

$$S_C(g) = \frac{U(0, g)}{I(0, g)} \quad (3.12)$$

Note that maximizing the likelihood ratio statistics is equivalent to maximizing the profile likelihood since $\log[l(0,g)]$ is the same for all g.

All of the above methods are for addressing single type of failure time response. To my knowledge there is no cutpoint determination method for competing risks study. Based on a view of extension of single failure time data, it is of interest to develop

cutpoint determination method for competing risks by extending the existing method for addressing single failure time data.

We propose five methods of dichotomization, one method is created based on two-sample statistic for comparing competing risks data, and the remaining four methods are created based on regression for subdistribution of competing risks. Monte Carlo simulation is conducted to asses the performance of the five proposed methods based on some statistical indicators. The procedure to generate competing risks survival time data as well as its censored data is discussed. The final part deals with the application of the method to contraceptive discontinuation data. Permutation test is used to asses the level of significance associated with the optimal split and bootstrap confidence interval is obtained for the optimal cutpoint.

3.1 Cutpoint Determination Method via Two-sample Statistic

In this section we focus on the method based on two-sample Gray's statistic for subdistribution comparison. Let Z be the risk factor of interest measured as a continuous variable and (T,d) be the outcome variable and its indicator of failure type. In some cases of competing risk survival analysis, the outcome of interest T can be censored in which the $d = 0$. The population is divided into two groups based on the cutpoint g : subjects with the value

of Z less than value of the cutpoint g and subjects with the value of Z greater or equal than cutpoint g (see figure 3.1). Let W be the set of K distinct value of the continuous covariate Z. Then, based on one hypothetical cutpoint $g \in W$, calculate Gray's statistic for making comparison of jth subdistribution between group 1 ($Z < g$) and group 2 ($Z \geq g$). The optimal cutpoint is that the value of W, g, that maximizes the value of c_G^2, where

$$c_G^2 = \int_0^t W(t)\left[\frac{d\hat{F}_{j1}(t)}{1-\hat{F}_{j1}(t^-)} - \frac{d\hat{F}_{j2}(t)}{1-\hat{F}_{j2}(t^-)}\right] \tag{3.13}$$

g therefore gives the value of the continuous covariate that gives the maximum difference of subdistribution between the subject in the two groups defined by the cutpoint.

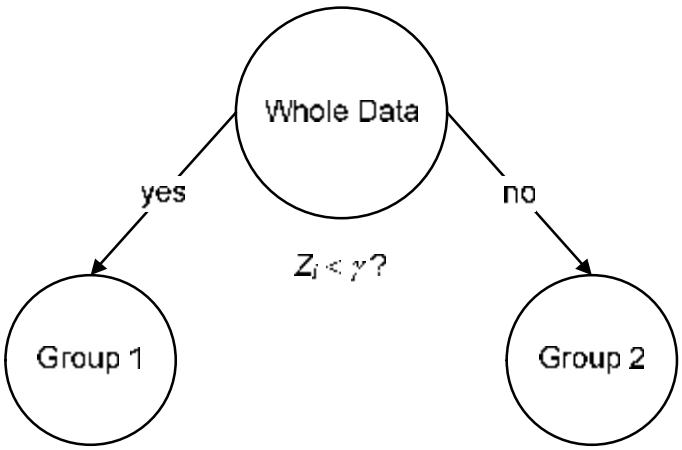

Figure 3.2. Data partition based on cutpoint g

3.2 Cutpoint Determination Method via Regression Analysis

Consider that observations are n replicates of a time to event T, a cause of interest d, and a covariate Z, (T_i,d_i,Z_i), $i=1,2,\ldots,n$. Let $d=1$ denote the failure cause of interest. The cutpoint determination based on regression approach is developed by adopting the similar method for addressing single type of failure time response as formulated in (3.6). The proportional subdistribution hazards model with a cutpoint, g, is defined by

$$\tilde{l}_j(t;Z_i) = \tilde{l}_{j0}(t)\exp\left[b_g I(Z_i < g)\right] \tag{3.14}$$

where b_g and g are unknown parameters to be estimated, and $\tilde{l}_{j0}(\bullet)$ is unspecified function. By considering $\tilde{l}_j(t;Z_i) = -\frac{d}{dt}\left\{\log\left[1-F_j(t;Z_i)\right]\right\}$, then model (3.14) is equivalent to

$$F_j(t;Z_i) = 1-\exp\left\{-\exp\left[b_g I(Z_i < g)\right]\times \tilde{l}^{*}_{j0}(t)\right\} \tag{3.15}$$

where $\tilde{l}^{*}_{j0}(t) = \log\int_0^t \tilde{l}_{j0}(u)du$.

Complementary log-log transformation on (3.15) yielded

$$g\left[F_j(t;Z)\right] = \tilde{l}^{*}_{j0}(t) + b_g I(Z < g) \tag{3.16}$$

where, $g(u) = \log(-\log(1-u))$.

In this model, g is cutpoint, b_g is the effect of having a Z-value less than or equal to g. The statistical inference for b_g and g is based on the partial log-likelihood function

$$l(b_g, g) = \sum_{i=1}^{n} I(d_i = j)\left(b_g I(Z_i < g) - \log \sum_{i' \in R_i} \exp\left[b_g I(Z_{i'} < g) \right] \right) \quad (3.17)$$

where risk set R_i consists of all individuals who have not yet failed of the cause of interest or who will never experience this event type: $R_i = \{i' : (T_{i'} \geq t_i) \cup (T_{i'} < t_i \cap d_{i'} \neq j)\}$.

For fixed g, the log-likelihood is known to have nice properties. The maximum likelihood estimate $\hat{b}_g$ for b_g is calculated, and the asymptotic distribution of $\hat{b}_g$ is well known. The partially maximized likelihood can also be denoted as

$$l(g) = \sup_{\hat{b}_g} l(b_g, g) = l(\hat{b}_g, g) \quad (3.18)$$

This is a piecewise constant between the unique values of Z which is $Z_{(1)}, Z_{(2)}, \ldots, Z_{(K)}$.

3.2.1 Cutpoint with Maximum Value of Wald Statistic

The Wald statistic for testing the hypothesis H_0: $b_g = 0$, g fixed, is

$$c^2_{W_g} = \frac{\hat{b}_g^2}{\text{var}(\hat{b}_g)} \quad (3.19)$$

and the optimal cutpoint $\hat{g}$ should be one value corresponding to the largest value of $c^2_{W_g}$, i.e.

$$c^2_W = \sup_g \left(c^2_{W_g} \right) = \frac{\hat{b}^2_{\hat{g}}}{\text{var}\left(\hat{b}_{\hat{g}}\right)} \tag{3.20}$$

3.2.2 Cutpoint with Maximum Value of Likelihood Ratio Statistic (Minimum Deviance)

Based on model (3.14) a log likelihood ratio test could also be applied for testing the null hypothesis, H_0: b_g =0, with g varying. The standard log likelihood ratio test is the ratio between the partial likelihood given the null hypothesis and the partial likelihood given the estimated parameters under alternative hypothesis, and takes the logarithm of the ratio, and multiply it by –2, that is

$$c^2_{LR_g} = -2\left[l(0) - l\left(\hat{b}_g, g\right) \right] \ , \tag{3.21}$$

where $l\left(\hat{b}_g, g\right)$ is the usual log likelihood when g is fixed. For varying values of g, taking supremum of $c^2_{LR_g}$ over a certain range of g will be a natural statistic to test the null hypothesis (Davies, 1977).

$$c^2_{LR} = \sup_g c^2_{LR_g} = -2\left[l(0) - l\left(\hat{b}_{\hat{g}}, \hat{g}\right) \right] \tag{3.22}$$

For a given dataset with $l(0)$ fixed, the optimal cutpoint is $\hat{g}$ corresponding to the smallest deviance, $-2l\left(\hat{b}_{\hat{g}},\hat{g}\right)$. Deviance is the summary measure of agreement between model and the data (Collet, 1994).

3.2.3 Cutpoint with Maximum Value of Delta Deviance

By considering deviance as a measure of agreement, we can use it to find a cutpoint $\hat{g}$. The procedure is as follows:

1. For standard set of data $N = \{(T_i,d_i,Z_i),\ i = 1, 2, \ldots, n\}$, suppose there is a candidate as a cutpoint g that divides the cases into $N_1 = \{(T_i,d_i,Z_i),\ Z_i < g,\ i = 1, 2, \ldots, n_1\}$ and $N_2 = \{(T_i,d_i,Z_i),\ Z_i \geq g,\ i = 1, 2, \ldots, n_2\}$, where N_1 and N_2 are defined as the first and second group resulted by the candidate cutpoint g.
2. Fit proportional subdistribution hazard model $\tilde{l}_j(t;Z_i) = \tilde{l}_{j0}(t)\exp(bZ_i)$ to data N and obtained its deviance, $D = -2l\left(\hat{b}\right)$.
3. Fit the similar model to N_1 and N_2
   - fit model $\tilde{l}_j(t;Z_i) = \tilde{l}_{j0g}(t)\exp\left(b_{1g}Z_i\right)$ to data N_1 and the obtained deviance is $D_{1g} = -2l\left(\hat{b}_{1g}\right)$.

- fit model $\tilde{l}_j(t;Z_i)=\tilde{l}_{j0g}(t)\exp(b_{2g}Z_i)$ to data N_2 and the obtained deviance is $D_{2g}=-2l(\hat{b}_{2g})$.

4. The improvement of fitting adequacy corresponding to cutpoint g which partition data into N_1 and N_2 is measured as $DD_g=D-(D_{1g}+D_{2g})$.
5. The optimal cutpoint $\hat{g}$ is one corresponding to the largest value of delta deviance, $DD=\sup_g(DD_g)=D-(D_{1\hat{g}}+D_{2\hat{g}})$.

3.2.4 Cutpoint with Maximum Value of Delta Null Deviance

Instead of using proportional subdistribution hazard model with covariate, we can fit the null model without covariate, $\tilde{l}_j(t;Z_i)=\tilde{l}_{j0}(t)$, and use the similar procedure to find the optimal cutpoint $\hat{g}$, i.e.

$$DD^0=\sup_g(DD_g^0)=D^0-(D_{1\hat{g}}^0+D_{2\hat{g}}^0)$$

3.3 Simulation on Cutpoint Determination

To compare different cutpoint determination methods a Monte Carlo experiment were conducted. One thousand replicates for eighteen combinations resulted from three different sample sizes (*n*

= 20, 200, 2000), two different relative risks ($RR = \exp(b_g) = 2, 5$), and three different percentage of censoring (p_c = 0%, 25%, 50%) were generated for the five proposed cutpoint determination methods. Those eighteen scenarios are summarized in Table 3.1.

Table 3.1. Scenarios for comparing five cutpoint determination methods

	$RR = 2$			$RR = 5$		
	n=20	n=200	n=2000	n=20	n=200	n=2000
p_c = 0%	1	2	3	4	5	6
p_c = 25%	7	8	9	10	11	12
p_c = 50%	13	14	15	16	17	18

Note: number in cell is scenario identity number for corresponding cell

Let Z be the independent variable under consideration, and g be the true cutpoint of Z. The value of independent variable Z is grouped into two subsets, the first group of $n/2$ subjects takes the integer value from 1 to 50, and the second group of $n/2$ takes the integer value from 51 to 100. Hence, the sample size is n and the true cutpoint is g = 51.

3.3.1 Data Generation

Suppose that there are two types of failure, d = 1, 2, where the subdistribution for first type is

$$F_1(t;Z_i) = 1 - \{1 - p(1 - \exp(-t))\}^{\exp[b_g I(Z_i < g)]},\ 0 < p < 1 \qquad (3.23)$$

The value of p related to the probability of first and second failure type and b_g represented the relative risks between 2 groups of

individual (group with $Z_i < g$ and group with $Z_i \geq g$). The subdistribution (3.23) can be expressed in complementary log-log form

$$\log\{-\log[1-F_1(t;Z_i)]\} = \log\{-\log[1-p(1-\exp(-t))]\} + b_g I(Z_i < g)$$

or in term of proportional subdistribution hazard model,

$$\tilde{l}_1(t;Z_i) = \frac{p\exp(-t)}{\{1-p[1-\exp(-t)]\}\log\{1-p[1-\exp(-t)]\}} \times \exp[b_g I(Z_i < g)]$$

Given the above formulation, the probability of occurrence of first type of failure is

$$p_1 = P(d=1;Z_i) = F_1(\infty;Z_i) = 1-(1-p)^{\exp[b_g I(Z_i<g)]}$$

and probability of occurrence of type 2 failure is

$$\begin{aligned} p_2 = P(d=2;Z_i) &= 1- P(d=1;Z_i) \\ &= (1-p)^{\exp[b_g I(Z_i<g)]} \end{aligned}$$

Suppose that conditional probability of failure time given type 2 failure followed an exponential distribution with rate $\exp[b_g I(Z_i < g)]$, that is

$$P(T \leq t \mid d=2, Z_i) = 1-\exp\{-t\exp[b_g I(Z_i < g)]\}$$

Hence, the subdistribution for second type of failure is

$$\begin{aligned} F_2(t;Z_i) &= P(T \leq t \mid d=2, Z_i) \times P(d=2;Z_i) \\ &= \{1-\exp[-t\exp(b_g I(Z_i < g))]\} \times (1-p)^{\exp[b_g I(Z_i<g)]} \end{aligned} \qquad (3.24)$$

How does one generate data once the subdistribution function is known? Note that the subdistribution function for cause j, j = 1, 2, satisfy the following relationship

$$F_j(t;Z_i) = P(T \le t, d = j; Z_i) = P(T \le t \mid d = j, Z_i) \times P(d = j; Z_i), j = 1,2$$

where T is overall survival time and d is the indicator of the cause of failure as well as censoring indicator when d = 0. It follows that the conditional distribution of the survival time, given the cause of failure j can be written as follows

$$P(T \le t \mid d = j, Z_i) = \frac{F_j(t;Z_i)}{P(d = j; Z_i)}, j = 1,2 \qquad (3.25)$$

Data is generated by first selecting the cause of failure with probability $P(d = j;Z_i)$ and then generating from the conditional distribution (3.25) of time to event given the failure cause. The latter step is done by applying the inverse transformation method.

The summary of the procedure to generate data following model (3.23) and (3.24) is as follows:

1. Set the parameter value of p and b_g. Here we use p = 0.66 and $b_g \in \{\ln(2), \ln(5)\}$ where b_g represents the relative risks between two group of observations resulted by cutpoint g.
2. Generate Z using previous scenario with true cutpoint g = 51.
3. Calculate

   $p_1 = P(d = 1; Z_i) = 1 - (1-p)^{\exp[b_g I(Z_i < g)]}$, 0 < p < 1 and

$p_2 = P(d = 2; Z_i) = 1 - p_1$.

4. Generate the cause of failure from set {1,2} with probability $\{p_1, 1-p_1\}$.

5. If cause 1 is selected at step 4, then we generate data from conditional distribution of the survival time, given the cause of failure 1

$$P(T \le t \mid d = 1, Z_i) = \frac{1 - \left[1 - p\{1 - \exp(-t)\}\right]^{\exp[b_g I(Z_i < g)]}}{1 - (1-p)^{\exp[b_g I(Z_i < g)]}}$$

which by means of inverse transform method the corresponding formula for generating data is

$$T = \ln p - \ln\left\{p - 1 + \left[1 - U + U(1-p)^{\exp[b_g I(Z_i < g)]}\right]^{\exp[-b_g I(Z_i < g)]}\right\},$$

where U is uniform (0,1) random variates, otherwise data is generated from

$$P(T \le t \mid d = 2, Z_i) = 1 - \exp\left\{-t \exp\left[b_g I(Z_i < g)\right]\right\}$$

which follows the formula

$$T = -\frac{\ln(1-U)}{\exp\left[b_g I(Z_i < g)\right]},$$

3.3.2 Censored Data Generation

To incorporate the censoring observation, independent uniform censoring over the interval $(0,a)$ was imposed. Parameter a was

chosen to give 25 and 50 percent censoring. In the sequel, p_c represents the censoring proportion.

To find appropriate values of the censoring parameter a, we suppose that each observation has a survival time T and a censoring time C, where T and C are independent of each other. We also assume that T has a distribution with a density $f_T(t)$ and C is uniformly distributed over the interval $(0,a)$. The probability of an observation being censored could be written as

$$\begin{aligned} p_c &= P(C < T \mid T = t) = \int_0^\infty \int_0^t f_C(c) f_T(t) dc dt \\ &= \int_0^\infty f_T(t) \left[\int_0^t f_C(c) dc \right] dt \\ &= \int_0^\infty P(C < t) f_T(t) dt \end{aligned} \tag{3.26}$$

Taking into account that censoring time $C \sim U(0,a)$, we can rewrite equation (3.26) as follows:

$$\begin{aligned} p_c &= \int_0^\infty [1 - S_C(t)] f_T(t) dt \\ &= \frac{1}{a} \int_0^a t f_T(t) dt + S_T(a) \end{aligned} \tag{3.27}$$

where

$$S_C(t) = \begin{cases} 1 - \dfrac{t}{a}, & 0 \le t \le a \\ 0, & t > a \end{cases}$$

$S_C(t)$ and $S_T(t)$ are survival functions of the censoring time and overall survival time, respectively. Complex relationships resulting

from these calculations make equation (3.27) intractable. Therefore, censoring parameters a is obtained by employing numerical method.

Given the subdistribution functions in (3.23) and (3.24), then

$$\begin{aligned} F_T(t;Z_i) &= F_1(t;Z_i) + F_2(t;Z_i) \\ &= 1 - \{1 - p(1-\exp(-t))\}^{\exp[b_g I(Z_i<g)]} \\ &\quad + \{1 - \exp[-t\exp(b_g I(Z_i<g))]\} \times (1-p)^{\exp[b_g I(Z_i<g)]} \end{aligned}$$

and

$$\begin{aligned} f_T(t;Z_i) = \frac{dF_T(t;Z_i)}{dt} &= \frac{\{1-p[1-\exp(-t)]\}^{\exp[b_g I(Z_i<g)]} \times p\exp[-t + b_g I(Z_i<g)]}{1-p[1-\exp(-t)]} \\ &\quad + \exp\{b_g I(Z_i<g) - t\exp[b_g I(Z_i<g)]\} \times (1-p)^{\exp[b_g I(Z_i<g)]} \end{aligned} \tag{3.28}$$

and

$$\begin{aligned} S_T(t;Z_i) &= 1 - F_T(t;Z_i) \\ &= \{1 - p(1-\exp(-t))\}^{\exp[b_g I(Z_i<g)]} \\ &\quad - \{1 - \exp[-t\exp(b_g I(Z_i<g))]\} \times (1-p)^{\exp[b_g I(Z_i<g)]} \end{aligned} \tag{3.29}$$

For a given value of p_c, proportion of censoring, the corresponding censoring parameter a is searched by solving the equation (3.27) by considering (3.28) and (3.29) for $f_T(t;Z_i)$ and $S_T(t;Z_i)$, respectively.

For group of observations with $Z_i \geq 0$, equation (3.27) becomes

$$p_c = \frac{1-\exp(-a)}{a} \tag{3.30}$$

hence for p_c = 0.25 the root is $a_{0.25}$ = 3.9207, and for p_c = 0.50 the root is $a_{0.50}$ = 1.5936.

For the second group of observations with $Z_i < 0$, $p = 0.66$ and $b_g = \ln(2)$, which correspond to $p_1 = 0.88$, equation (3.27) becomes

$$p_c = \frac{0.7244 - 0.4488\exp(-a) - 0.2756\exp(-2a)}{a} \qquad (3.31)$$

hence for $p_c = 0.25$ the root is $a_{0.25} = 2.7823$, and for $p_c = 0.50$ the root is $a_{0.50} = 1.0807$.

The other scenario for group of observations with $Z_i < 0$, $p = 0.66$ and $b_g = \ln(5)$, equation (3.18) becomes

$$p_c = \frac{0.3471 - 0.0441\exp(-a) - 0.0856\exp(-2a) - 0.1108\exp(-3a)}{a} + \frac{-0.0806\exp(-4a) - 0.0259\exp(-5a)}{a}$$

for $p_c = 0.25$ the root is $a_{0.25} = 1.3045$, and for $p_c = 0.50$ the root is $a_{0.50} = 0.5038$.

The summary is given in Table 3.2.

Table 3.2 Parameter for simulating censored observation for comparison of cutpoint determination with $p = 0.66$

Scenario		p_c	a
$Z_i \geq 0$		0.25 0.50	3.9207 1.5936
$Z_i < 0$	$b_g = \ln(2)$	0.25 0.50	2.7823 1.0807
	$b_g = \ln(5)$	0.25 0.50	1.3045 0.5038

3.3.3 Statistical Indicators for Assessing the Performance of Cutpoint Determination Methods

The following passages introduce the definition of each statistical indicator which will be applied in comparing the different cutpoint methods and in concluding the optimal cutpoints of each variable. All of these criteria are very important in measuring the validity of an estimate, and have been used in various statistical considerations.

(1) Mean

$\bar{\hat{g}} = \frac{\sum_{i=1}^{m} \hat{g}_i}{m}$, where m is the number of repetitions, and $\hat{g}_i$ is an estimate of the cutpoint parameter g from the ith replicate. The mean is an arithmetic average of a set of m estimated cutpoints, $\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_m$. This can be used to estimate the true cutpoint.

(2) Bias

Bias is the expected deviation of an estimate from the true cutpoint g. In the simulation approach, bias could be estimated by the mean difference, $Estimated\ \ bias = \bar{\hat{g}} - g$.

(3) Absolute relative estimated bias (%)

$$ARE\ \ Bias = \left(\frac{|Estimated\ \ Bias|}{g} \right) \times 100\%$$

(4) Estimated standard errors

$$SE = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{g}_i - \bar{\hat{g}}\right)^2}$$

This is a measure of the cutpoint estimator's variability around its mean.

(5) Estimated root mean square errors

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\hat{g}_i - g\right)^2}$$

This criteria is used to asses the variability and the square of the bias of an estimator. A slightly bias estimator which highly concentrated about the true cutpoint may be preferable to an unbiased estimator that is less concentrated. Thus, this general criterion allows for both bias and unbiased estimators to be compared, and it agrees with the variance criterion if attention is restricted to unbiased estimators.

Table 3.3. The comparison of mean of the estimated cutpoints determined by five cutpoint determination methods based on 1000 simulations and true cutpoint equal to 51 under selected relative risks ($\exp(b_g)$ = 2, 5), sample sizes (n = 20, 200, 2000) and censoring percentage (p_c = 0%, 25%, 50%).

	$\exp(b_g)$ = 2			$\exp(b_g)$ = 5		
	n = 20	n = 200	n = 2000	n = 20	n = 200	n = 2000
p_c = 0%						
c_G^2	54.103	55.447	56.355	52.624	54.217	47.390
c_W^2	46.110	55.392	45.929	51.762	44.201	49.269
D	46.207	45.683	55.029	48.916	56.757	51.103
DD	44.234	52.624	44.743	43.191	51.862	43.981
DD^0	44.048	57.719	52.768	47.353	47.816	44.740
p_c = 25%						
c_G^2	50.963	49.002	46.160	53.025	54.334	52.195
c_W^2	53.047	47.306	53.312	58.014	50.646	56.183
D	49.470	53.066	47.412	50.691	49.140	43.486
DD	42.953	51.100	43.219	47.656	50.338	57.458
DD^0	42.290	54.789	57.338	50.140	52.387	49.311
p_c = 50%						
c_G^2	52.054	52.869	47.215	54.240	44.138	45.750
c_W^2	54.759	54.689	45.226	59.339	50.060	48.566
D	50.408	45.449	54.794	52.530	56.523	50.869
DD	33.202	49.577	56.696	40.643	48.815	55.934
DD^0	32.990	44.359	54.408	48.347	51.332	53.881

Table 3.4. The comparison of bias of the estimated cutpoints determined by five cutpoint determination methods based on 1000 simulations and true cutpoint equal to 51 under selected relative risks ($\exp(b_g)$ = 2, 5), sample sizes (n = 20, 200, 2000) and censoring percentage (p_c = 0%, 25%, 50%).

	$\exp(b_g)$ = 2			$\exp(b_g)$ = 5		
	n = 20	n = 200	n = 2000	n = 20	n = 200	n = 2000
			p_c = 0%			
c_G^2	3.103(1)	4.447(3)	5.355(4)	1.624(2)	3.217(3)	-3.610(3)
c_W^2	-4.890(3)	4.392(2)	-5.071(3)	0.762(1)	-6.799(5)	-1.731(2)
D	-4.793(2)	-5.317(4)	4.029(2)	-2.084(3)	5.757(4)	0.103(1)
DD	-6.766(4)	1.624(1)	-6.257(5)	-7.809(5)	0.862(1)	-7.019(5)
DD^0	-6.952(5)	6.719(5)	1.768(1)	-3.647(4)	-3.184(2)	-6.260(4)
			p_c = 25%			
c_G^2	-0.037(1)	-1.998(2)	-4.840(3)	2.025(3)	3.334(5)	1.195(1)
c_W^2	2.047(3)	-3.694(4)	2.312(1)	7.014(5)	-0.354(1)	5.183(3)
D	-1.530(2)	2.066(3)	-3.588(2)	-0.309(1)	-1.860(4)	-7.514(5)
DD	-8.047(4)	0.100(1)	-7.781(5)	-3.344(4)	-0.662(2)	6.458(4)
DD^0	-8.710(5)	3.789(5)	6.338(4)	-0.860(2)	1.387(3)	-1.689(2)
			p_c = 50%			
c_G^2	1.054(2)	1.869(2)	-3.785(2)	3.240(3)	-6.862(5)	-5.250(5)
c_W^2	3.759(3)	3.689(3)	-5.774(5)	8.339(4)	-0.940(2)	-2.434(2)
D	-0.592(1)	-5.551(4)	3.794(3)	1.530(1)	5.523(4)	-0.131(1)
DD	-17.798(4)	-1.423(1)	5.696(4)	-10.357(5)	-2.185(3)	4.934(4)
DD^0	-18.010(5)	-6.641(5)	3.408(1)	-2.653(2)	0.332(1)	2.881(3)

*)Number in parentheses is rank for corresponding scenario. The resulted rank sums are c_G^2 = 50, c_W^2 = 52, D = 47, ΔD = 62 and ΔD^0 = 59.

Table 3.5. The comparison of absolute relative estimated bias of the estimated cutpoints determined by five cutpoint determination methods based on 1000 simulations and true cutpoint equal to 51 under selected relative risks ($\exp(b_g)$=2, 5), sample sizes (n = 20, 200, 2000) and censoring percentage (p_c = 0%, 25%, 50%).

	$\exp(b_g)$ = 2			$\exp(b_g)$ = 5		
	n = 20	n = 200	n = 2000	n = 20	n = 200	n = 2000
			p_c = 0%			
c_G^2	6.084(1)	8.720(3)	10.500(4)	3.184(2)	6.307(3)	7.078(3)
c_W^2	9.588(3)	8.612(2)	9.943(3)	1.495(1)	13.332(5)	3.394(2)
D	9.398(2)	10.425(4)	7.899(2)	4.087(3)	11.289(4)	0.202(1)
DD	13.266(4)	3.183(1)	12.269(5)	15.312(5)	1.690(1)	13.763(5)
DD^0	13.631(5)	13.174(5)	3.466(1)	7.150(4)	6.242(2)	12.274(4)
			p_c = 25%			
c_G^2	0.073(1)	3.918(2)	9.490(3)	3.971(3)	6.537(5)	2.343(1)
c_W^2	4.014(3)	7.243(4)	4.533(1)	13.752(5)	0.694(1)	10.163(3)
D	3.000(2)	4.051(3)	7.036(2)	0.607(1)	3.647(4)	14.734(5)
DD	15.778(4)	0.196(1)	15.256(5)	6.557(4)	1.297(2)	12.662(4)
DD^0	17.078(5)	7.429(5)	12.427(4)	1.686(2)	2.719(3)	3.313(2)
			p_c = 50%			
c_G^2	2.067(2)	3.664(2)	7.422(2)	6.352(3)	13.454(5)	10.295(5)
c_W^2	7.370(3)	7.233(3)	11.322(5)	16.351(4)	1.843(2)	4.773(2)
D	1.160(1)	10.885(4)	7.440(3)	3.001(1)	10.829(4)	0.258(1)
DD	34.898(4)	2.791(1)	11.168(4)	20.309(5)	4.284(3)	9.675(4)
DD^0	35.314(5)	13.021(5)	6.683(1)	5.202(2)	0.651(1)	5.649(3)

*)Number in parentheses is rank for corresponding scenario. The resulted rank sums are c_G^2 = 50, c_W^2 = 52, D = 47, ΔD = 62 and ΔD^0 = 59.

Table 3.6. The comparison of standard errors of the estimated cutpoints determined by five cutpoint determination methods based on 1000 simulations and true cutpoint equal to 51 under selected relative risks ($\exp(b_g)$ = 2, 5), sample sizes (n = 20, 200, 2000) and censoring percentage (p_c = 0%, 25%, 50%).

	$\exp(b_g)$ = 2			$\exp(b_g)$ = 5		
	n = 20	n = 200	n = 2000	n = 20	n = 200	n = 2000
			p_c = 0%			
c_G^2	11.362(2)	5.649(5)	4.535(5)	8.251(3)	2.475(4)	2.407(5)
c_W^2	14.346(5)	4.679(4)	3.699(4)	9.309(5)	3.186(5)	2.102(3)
D	13.509(4)	4.082(2)	2.234(1)	7.679(2)	1.982(2)	0.234(1)
DD	9.884(1)	2.844(1)	3.369(3)	8.951(4)	2.456(3)	2.307(4)
DD^0	11.556(3)	4.338(3)	3.099(2)	4.518(1)	0.927(1)	0.988(2)
			p_c = 25%			
c_G^2	25.142(4)	9.574(4)	7.962(4)	12.862(2)	3.909(2)	2.909(1)
c_W^2	26.648(5)	10.670(5)	8.546(5)	17.257(4)	7.062(4)	6.959(5)
D	22.843(3)	9.447(2)	5.515(3)	12.483(1)	4.260(3)	3.029(2)
DD	19.092(1)	8.034(1)	4.436(1)	18.330(5)	7.708(5)	4.516(3)
DD^0	20.208(2)	9.449(3)	5.188(2)	13.782(3)	3.786(1)	4.841(4)
			p_c = 50%			
c_G^2	23.451(3)	9.854(3)	3.544(1)	18.955(2)	8.989(4)	2.285(1)
c_W^2	26.765(5)	12.423(5)	11.191(5)	19.293(3)	5.461(3)	5.988(3)
D	24.264(4)	8.909(2)	9.247(3)	15.663(1)	4.387(2)	7.393(4)
DD	22.330(1)	10.827(4)	8.909(2)	23.605(5)	11.839(5)	8.894(5)
DD^0	22.709(2)	7.628(1)	9.956(4)	20.516(4)	2.966(1)	2.453(2)

*)Number in parentheses is rank for corresponding scenario. The resulted rank sums are c_G^2 = 55, c_W^2 = 78, D=42, ΔD = 54 and ΔD^0 = 41.

Table 3.7. The comparison of root mean square errors of the estimated cutpoints determined by five cutpoint determination methods based on 1000 simulations and true cutpoint equal to 51 under selected relative risks ($\exp(b_g)$=2, 5), sample sizes (n = 20, 200, 2000) and censoring percentage (p_c = 0%, 25%, 50%).

	$\exp(b_g)$ = 2			$\exp(b_g)$ = 5		
	n = 20	n = 200	n = 2000	n = 20	n = 200	n = 2000
			p_c = 0%			
c_G^2	11.773(1)	7.187(4)	7.016(4)	8.405(3)	4.058(3)	4.338(3)
c_W^2	15.150(5)	6.416(2)	6.275(3)	9.336(4)	7.508(5)	2.722(2)
D	14.328(4)	6.702(3)	4.606(2)	7.954(2)	6.089(4)	0.256(1)
DD	11.974(2)	3.273(1)	7.106(5)	11.875(5)	2.602(1)	7.388(5)
DD^0	13.481(3)	7.996(5)	3.566(1)	5.805(1)	3.316(2)	6.337(4)
			p_c = 25%			
c_G^2	25.129(4)	9.775(3)	9.314(5)	13.014(2)	5.136(3)	3.143(1)
c_W^2	26.713(5)	11.286(5)	8.849(3)	18.620(4)	7.067(4)	8.674(5)
D	22.883(3)	9.665(2)	6.578(1)	12.480(1)	4.646(2)	8.101(4)
DD	20.710(1)	8.031(1)	8.955(4)	18.624(5)	7.732(5)	7.879(3)
DD^0	21.996(2)	10.176(4)	8.189(2)	13.802(3)	4.031(1)	5.125(2)
			p_c = 50%			
c_G^2	23.463(1)	10.025(1)	5.185(1)	19.221(2)	11.305(4)	5.726(2)
c_W^2	27.014(3)	12.953(5)	12.588(5)	21.009(4)	5.538(2)	6.461(3)
D	24.259(2)	10.493(3)	9.991(2)	15.730(1)	7.052(3)	7.390(4)
DD	28.547(4)	10.915(4)	10.571(4)	25.767(5)	12.033(5)	10.167(5)
DD^0	28.975(5)	10.111(2)	10.519(3)	20.677(3)	2.983(1)	3.783(1)

*)Number in parentheses is rank for corresponding scenario. The resulted rank sums are c_G^2 = 47, c_W^2 = 69, D=44, ΔD = 65 and ΔD^0 = 45.

3.3.4 Simulation Results

Table 3.3 showed the results of the mean based on 1000 simulations from first type of failure time, a predictor variable with true cutpoint equal to 51 and eighteen scenarios corresponding to Table 3.1. The pattern of the resulted means obtained from all of five methods are not quite clear. We can not distinguish which method had closer mean to the true cutpoint 51 compared to the

other. A slightly clear pattern is obtained by using rank. For each scenario we rank the methods. Rank 1 is given to the method with smallest bias, and rank 5 for the largest bias as well. Since we have 18 scenarios, so the smallest rank sum will be 18 and the largest is 90. The rank and rank sum of estimated bias are shown in the parentheses and notes of Table 3.4. Deviance (*D*) method has slightly smaller rank sum of estimated bias compared to another which is 47 and its pattern is displayed in Figure 3.3.

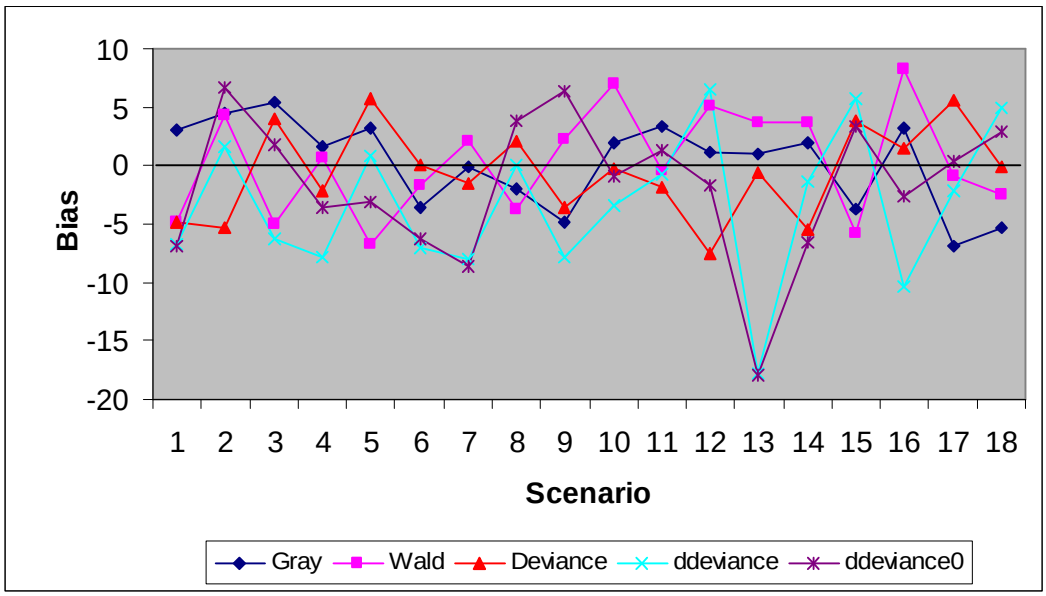

Figure 3.3. Simulation result in term of bias for eighteen scenarios

The absolute relative estimated bias had similar pattern with the bias (see Table 3.5). It is not surprise, since this quantity had direct linear relationship with bias. Therefore, in terms of absolute relative bias deviance method also had slightly smaller absolute relative bias and smallest rank sum.

The result is slightly different for the performance based on standard error (see Table 3.6 and Figure 3.4). The delta null deviance (DD^0) has the smallest rank sum and deviance (D) the second smallest. As percentage of censoring increased, the standard error also increased. For each percentage of censoring, the standard error decreased as the sample size and relative risk increased.

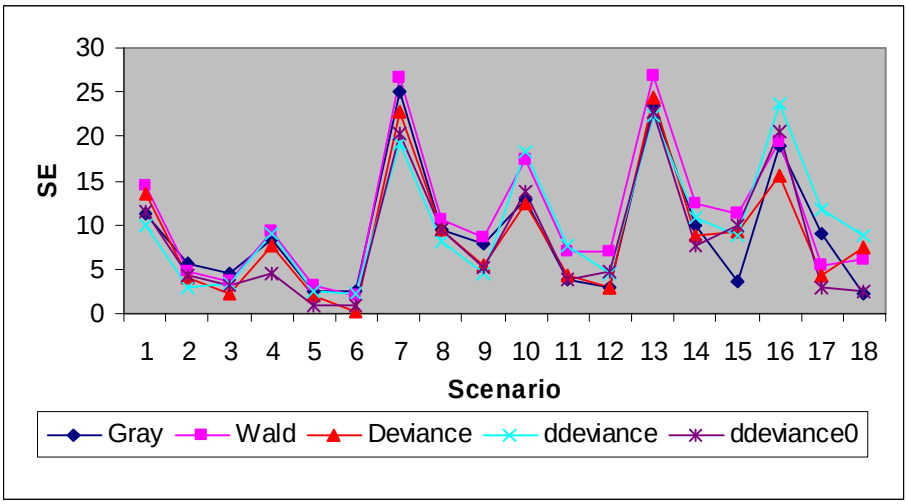

Figure 3.4. Simulation result in term of standard error (SE) for eighteen scenarios

The performance based on root mean square error showed the advantage of deviance (D) method (see Table 3.7 and Figure 3.5). Deviance method gets the best result with the smallest rank sum, and delta null deviance for the second smallest rank sum, although the pattern is almost similar with the performance based on SE.

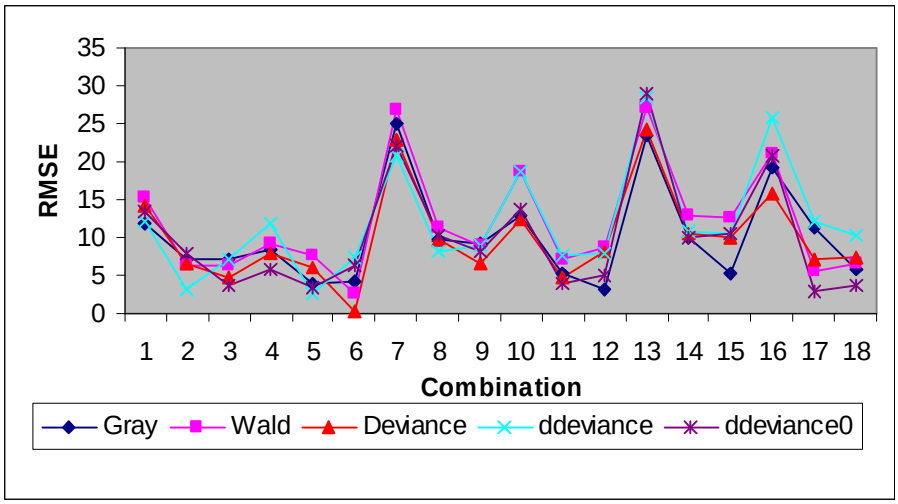

Figure 3.5. Simulation result in term of root mean square error (RMSE) for eighteen scenarios

Overall, comparing the results from all of the indicators, the deviance method has the smallest rank sum among the studied methods in deciding an optimal cutpoint (see Table 3.8). We can get more insight on the performance of five proposed methods through the overall rank sum, because it contains information of all four indicators.

Table 3.8. Overall rank sum for five cutpoint determination methods.

	Rank Sum of the Criteria				
	Bias	ARE Bias	SE	RMSE	Overall
C_G^2	50	50	55	47	202
C_W^2	52	52	78	69	251
D	47	47	42	44	180
DD	62	62	54	65	243
DD^0	59	59	41	45	204

3.4 Application: Contraceptive discontinuation data

To illustrate our approach we consider a sample of 2631 women drawn from the database of the Indonesian Demography and Health Survey (IDHS) 2002. This is the national retrospective database consisting of data on time to contraceptive discontinuation. All subjects were investigated on the history of last episode of contraceptive discontinuation. We observed the length of time of the last contraception use, and we focused on three types of discontinuation in a competing risks framework. The outcomes we considered were failure, contraceptive abandonment while in need of family planning, and switching to another contraceptive method. A discontinuation is defined as a contraceptive failure if the woman reported that she became pregnant while using the method. Thus, this definition includes both failures of the method itself and failure owing to incorrect or inconsistent use of the method. Adoption of different method within one month of discontinuation is classified as a method switch, whereas continuation of nonuse for one month or more is classified as contraceptive abandonment. Clearly, contraceptive failure is of interest because it leads directly to an unintended pregnancy. Contraceptive abandonment is also important outcome to study because it leads to immediate risk of unintended pregnancy. Method switching also may lead to an increased risk of

unintended pregnancy if use of a modern method is discontinued in favor of a less effective, traditional method. Contraceptive failure is somewhat different from the other two outcomes in that it presumably is an unintentional event, whereas contraceptive abandonment and switching suggest some decision-making and choice on the part of the woman. For dichotomization procedure we considered only one covariate which supposed to be able to explain the rate of discontinuation which was age of the women at the start of the episode of use (years). Here we used deviance method, since it was slightly better than the rest based on simulation result in section 3.3.

3.4.1 Optimal Cutpoint

A cutpoint determination method based on deviance statistic was implemented in order to systematically partition the subjects into two groups as determined by the age at start of contraceptive use. The optimal cutpoint selected was one associated with the observed minimum deviance achieved by splitting the subjects into two age groups. For the time to the occurrence of contraceptive failure as dependent variable, an age at start of contraceptive use threshold of 34.167 years was associated with a minimum deviance of 1061.321 (see figure 3.6). The optimal cutpoint on age at start of contraceptive use for the second and third type of

discontinuation, which were abandonment and switching, was 38 years with their corresponding minimum deviance 13796.55 and 14061.17, respectively (see Figures 3.7 and 3.8).

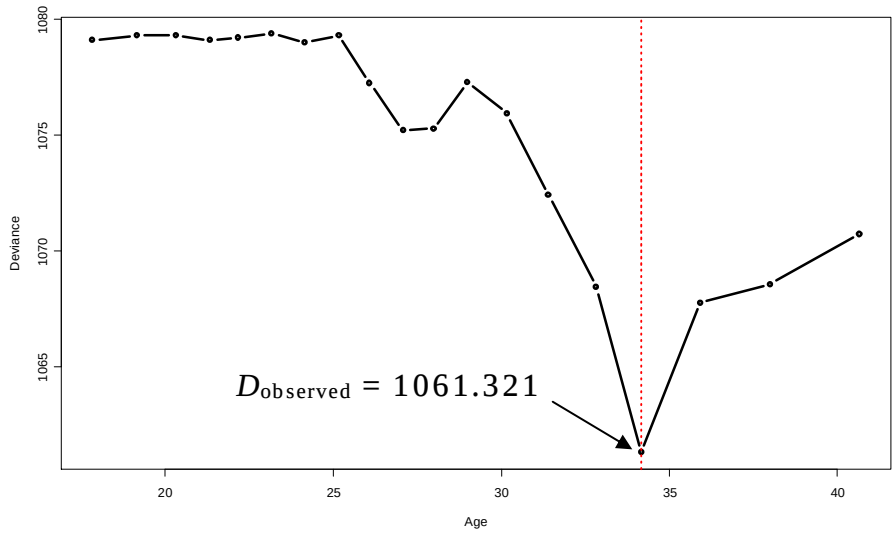

Figure 3.6. The plot of cutpoint criterion *D* for dependent variable time to occurrence of failure against cutpoint on age. *D* bottoms at age 34.167 years.

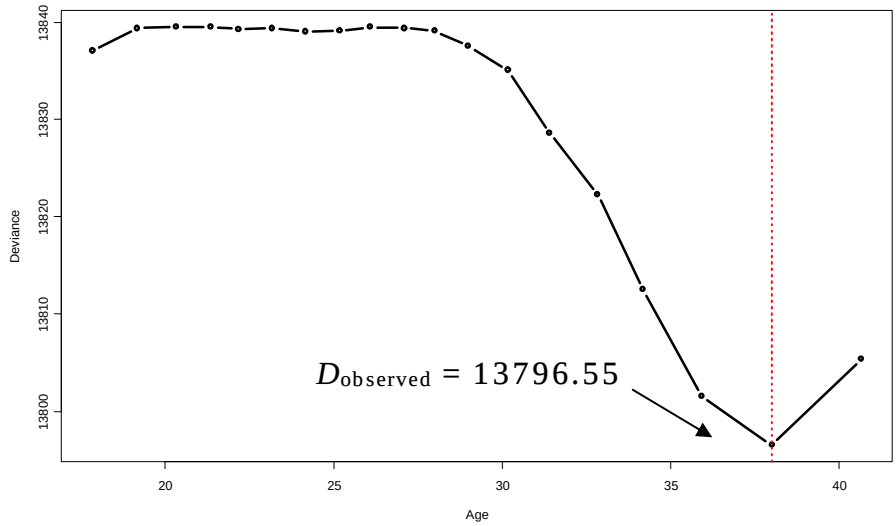

Figure 3.7. The plot of cutpoint criterion *D* for dependent variable time to occurrence of abandonment against cutpoint on age. *D* bottoms at age 38 years.

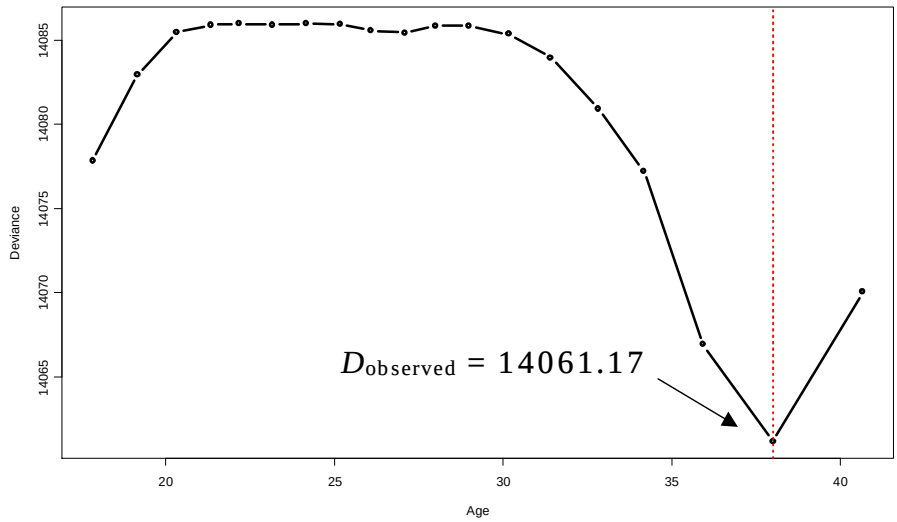

Figure 3.8. The plot of cutpoint criterion *D* for dependent variable time to occurrence of switching against cutpoint on age. *D* bottoms at age 38 years.

3.4.2 Permutation Test

It is useful to assess the strength of evidence of an association between the response and predictor variable which is categorized into a dummy variable using a cutpoint obtained from the results of cutpoint determination method. A distribution-free permutation test is used to asses the strength of evidence of an association between predictor variable which has been dichotomized into 'high' and 'low' levels via the cutpoint determination method proposed in this chapter and the response variable (Venables and Ripley 2002, Mielke and Berry 2007). We implemented the permutation test in the following steps:

1. Calculate D_{observed}, the optimal deviance obtained from the observed data.
2. Permute the observed data to obtain B permutation data sets N_b, where $b = 1, \ldots, B$. This is simply achieved by randomly permuting the predictor values while holding the response variable fixed.
3. For each of these data sets, compute the optimal deviance statistic D_b, where $b = 1, \ldots, B$.
4. The p-value is obtained from the permutational distribution of the optimal deviance statistic. It is calculated as

$$p = \frac{\#D < D_{\text{observed}}}{B}$$

For sample of size n, we can make up to $n!$ permutation sample. It is typically computationally impossible to obtained all permutation sample of large size of the observed data and to compute an optimal deviance statistic for each of those data sets. Therefore, as an approximation, we can choose B sufficiently large in order to achieve as many significant digits as desired for the p-value.

Here we performed a permutation test to provide an assessment of the strength of evidence of association between:

1. age at start of contraceptive use (dichotomized at 34.167) and probability of discontinuation due to failure.

2. age at start of contraceptive use (dichotomized at 38) and probability of discontinuation due to abandonment.
3. age at start of contraceptive use (dichotomized at 38) and probability of discontinuation due to switching.

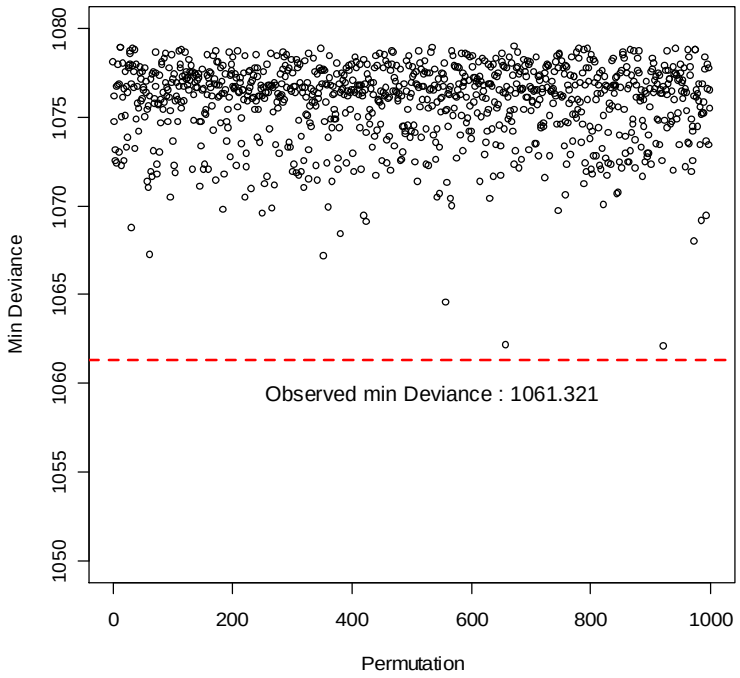

Figure 3.9. Permutation plot of the sequence of D_b for time to occurrence of failure as dependent variable, b = 1, ..., 1000.

A total of 1000 permutations of the data were obtained, and for each permutation a sequence of deviance corresponding to the sequence of possible binary splits was computed. From each of these 1000 sequences, the optimal split was selected, resulting in a sequence of 1000 optimal splits; each optimal split was associated with an optimal deviance. The sequences of optimal deviance for the 1000 permutations of the three types of discontinuation were

plotted in Figure 3.9-3.11. From 1000 results there is no permutation produced an optimal deviance less than the observed value. In the other words, the *p*-value of the test is less than 0.001. It means that there is strong evidence on the association of age with all the three types of discontinuation time.

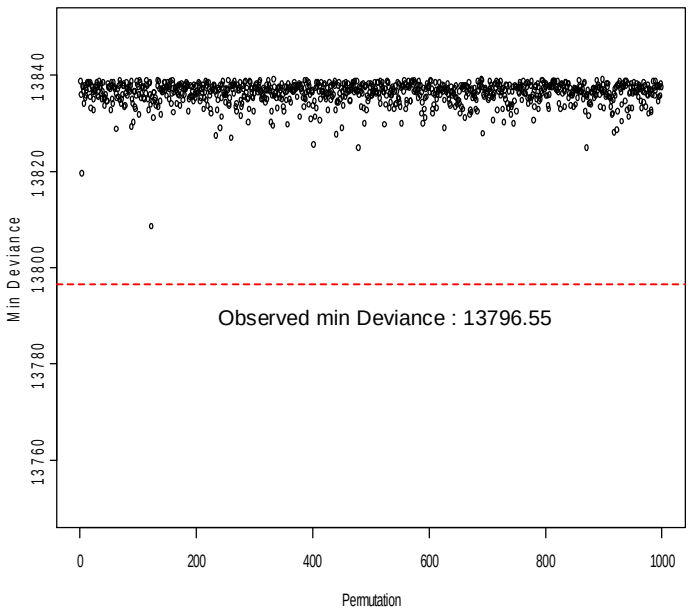

Figure 3.10. Permutation plot of the sequence of D_b for time to occurrence of abandonment as dependent variable, b = 1, ..., 1000.

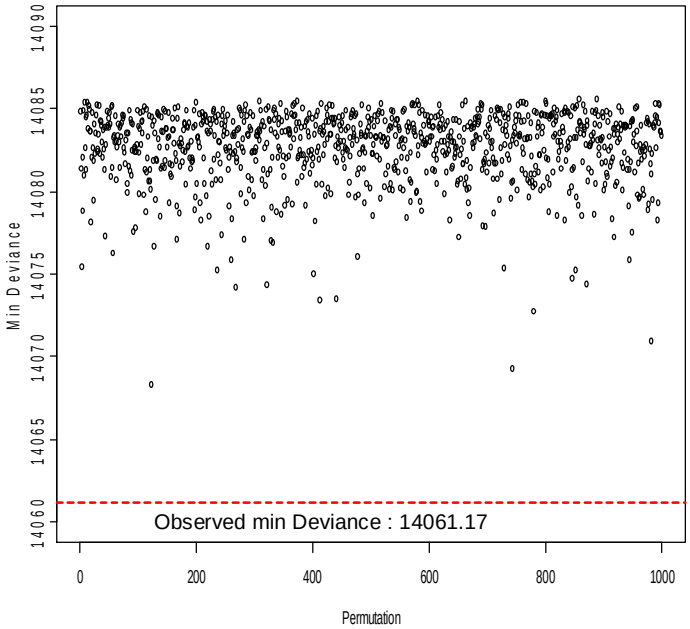

Figure 3.11. Permutation plot of the sequence of D_b for time to occurrence of switching as dependent variable, b = 1, ..., 1000.

3.4.3 Bootstrap Confidence Interval

Here we introduce the steps for the construction of bootstrap confidence intervals for true cutpoint g. The advantage of the bootstrap is that there is no assumption about the distribution of optimal cutpoint.

We considered p-Bootstrap method based on the percentiles of the bootstrap distribution suggested by Efron and Tibshirani (1993) to construct the confidence intervals for true cutpoint g. Other existing alternatives for the p-Bootstrap, not considered in this thesis, also could be used to construct confidence intervals. For a

complete review of available approaches to bootstrap confidence intervals see Efron and Tibshirani (1993) and Davison and Hinkley (1997).

Let $N = (t,d,z)$ be the observed data where $t = (t_1, \ldots, t_n)$ is the vector of lifetime data, $d = (d_1, \ldots, d_n)$ is the vector of indicators of censored observations and $z = (z_1, \ldots, z_n)$ is the vector of predictor variable. The p-Bootstrap procedure is as follows:

1. Random select, with replacement from N, a bootstrap sample $(t_1^*, d_1^*, z_1^*), \ldots, (t_n^*, d_n^*, z_n^*)$.
2. From the bootstrap sample in 1, find the optimal cutpoint $\hat{g}^*$.
3. Repeat steps 1 and 2, B times.
4. From $\hat{g}^* = \left(\hat{g}^*_{(1)} \le \hat{g}^*_{(2)} \le \ldots \le \hat{g}^*_{(B)}\right)$ find a $100\times(1 - a)\%$ bootstrap confidence interval given by $\left(\hat{g}^*_{(q_1)}, \hat{g}^*_{(q_2)}\right)$ where $q_1 = [(a/2)B]$ and $q_2 = B - q_1$.

A bootstrap 90% percentile interval was computed to provide an assessment of the accuracy of the observed optimal cutpoint on age at start of contraceptive use of:

1. 34.167 years for discontinuation due to failure.
2. 38 years for discontinuation due to abandonment.
3. 38 years for discontinuation due to switching.

A total of 1000 bootstrap data sets were generated from original data set. For each bootstrap data set, we employed the deviance cutpoint determination method to obtain the optimal deviance and its associated age cutpoint. Histograms of the 1000 replicates for three types of contraceptive discontinuation are presented in Figure 3.12 – 3.14. For first type of discontinuation, which is failure, the 90% bootstrap percentile interval for the observed optimal cutpoint on age at start of contraceptive use is (31.58,38.08). This confidence interval is slightly moderate than those two others, which are (35.92,40.75) and (17.83,40.75) for the occurrence of abandonment and switching, respectively. Figure 3.12 showed that most of cutpoint is located between the range of age 30 and 40. The shorter range is for abandonment which is between age 37 and 39, even though there are several cutpoints located below and above those bounds (Figure 3.13). Different result is obtained for the cutpoint for switching which reveals small portion of cutpoint located at young age about 18, and the rests are around age 38 (Figure 3.14).

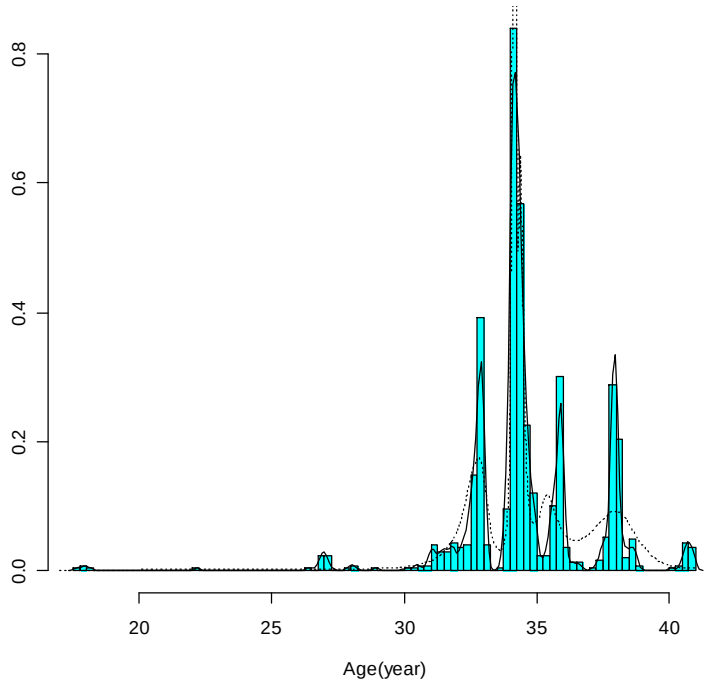

Figure 3.12. Histogram of 1000 bootstrap replications of the optimal cutpoint $\hat{g}$ on age at start of contraceptive use for the discontinuation due to failure.

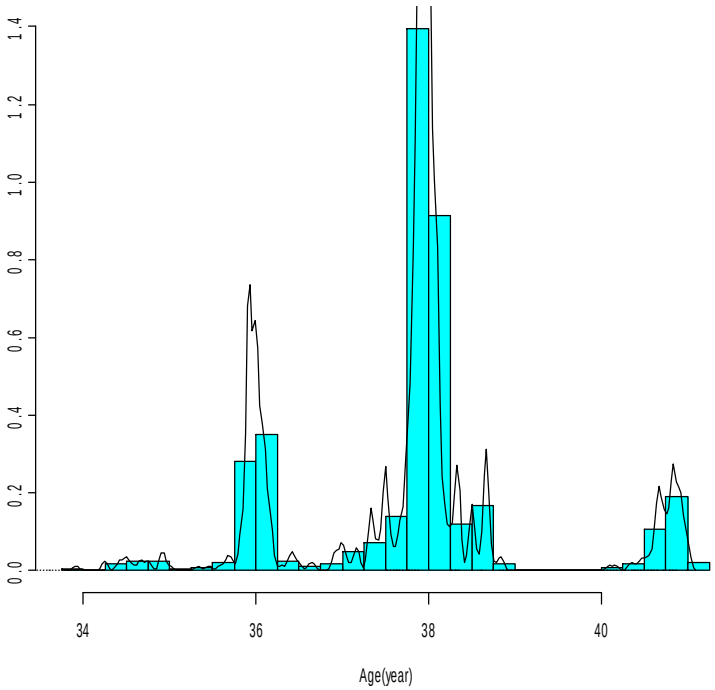

Figure 3.13. Histogram of 1000 bootstrap replications of the optimal cutpoint $\hat{g}$ on age at start of contraceptive use for the discontinuation due to abandonment.

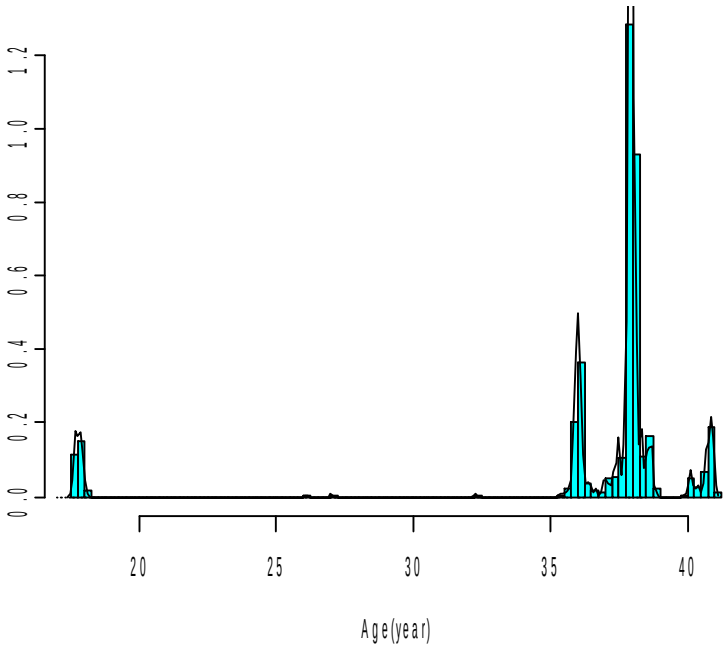

Figure 3.14. Histogram of 1000 bootstrap replications of the optimal cutpoint $\hat{g}$ on age at start of contraceptive use for the discontinuation due to switching.

3.5 Summary

This chapter develops new cutpoint determination method for handling competing risks survival data. A comprehensive simulation study on cutpoint determination methods indicates that the Deviance procedure has better statistical indicators across a variety of methods. The application of Deviance method on contraceptive discontinuation data showed its advantages to search age cutpoint corresponding to failure, abandonment and switching event.

Since there is no statistical distribution for minimum of Deviance, we assessed its significance by using permutation test. With 1000 repetitions, we cannot find smaller minimum Deviance compared to the obtained one, so the *p* value for the obtained minimum Deviance is less than 0.001.

Bootstrap procedure was employed for constructing the confidence interval of obtained cutpoint. The effect of end-cut preference was shown emerged in the result of bootstrap histogram. This effect indicated there is a small portion of optimal cutpoint located near to the boundaries at youngest and oldest age of women. Hence, the obtained confidence interval may be wider than it should be.

CHAPTER 4

TREE-STRUCTURED REGRESSION FOR SUBDISTRIBUTION OF COMPETING RISKS

In this chapter, survival trees are generalized to the competing risks case. Trees for competing risks outcome are grown by minimizing deviance. Therefore, only internal nodes have associated deviance statistics. The tree structure is different from CART because, for original trees, each node, either terminal or internal, has an associated impurity measure. This is why the CART pruning procedure is not directly applicable to such type of trees. However, Segal's pruning algorithm (Segal, 1988), which exerts little computational burden, has resulted in trees that have become well-developed tools.

Our modified tree technique not only provides a convenient way of handling competing risks survival data, but also extend the applied scope of tree-structured methods in a more general sense, especially for those situations where defining prediction error terms is relatively difficult. Growing trees by a deviance statistic, together with the Segal's pruning, offer a feasible way of performing tree analysis. Moreover trees by deviance statistic derived from likelihood-based statistic have an easy extension to handle nonstandard data structures, including binary, categorical,

longitudinal data, etc. For instance, Su *et al.* (2004) studied tree methods by maximizing log-likelihood for continuous response data.

In our proposed competing risks survival trees, the between-node difference is measured by a deviance statistic, which is derived from a likelihood ratio test statistic in a subdistribution hazards regression approach to competing risks survival data developed by Fine and Gray (1999). Reasons accounting for this choice is:

1. The subdistribution approach obviously gives the information about proportion of patient experiencing a cause of interest;
2. By considering this splitting procedure as an important aspect of tree regression, we should then utilize the splitting procedure which has been proven to have good performance. From chapter 3, the cutpoint determination based on deviance is better than the four other statistics.

In Section 4.1, the deviance statistic for the significance of a typical split is first derived. Subsequently, how to use the deviance statistic to grow a large tree is illustrated. Section 4.2 describes a pruning procedure adapted from Segal's (1988) approach. In Section 4.3 we apply the method to the contraceptive discontinuation data. In Section 4.4 the performance of the method is assessed by simulation studies.

4.1 Growing a Large Tree

The initial tree is grown by considering the used of splitting statistic and stopping rule. Splitting statistic is needed to make sure that the resulted data partitioning is the best one. The stopping rule is to maintain the effectiveness of the growing procedure.

4.1.1 The Splitting Statistic

We consider a typical setting for competing risks survival data. Suppose that there are n individuals and each subjected to J ($J \geq 2$) cause of failures. Let T_i^* be the time when ith unit experiences one of the jth type of failures, and let C_i be the corresponding censoring time, where $j = 1, 2, \ldots, J$; $i = 1, 2, \ldots, n$. The sample consists of the set of vectors $\{(T_i, d_i, Z_i) : i = 1, 2, \ldots, n\}$. Here, $T_i = \min(T_i^*, C_i)$ is the observed failure times; $d_i = I(T_i^* < C_i)$, where $I(\bullet)$ is the indicator function; $Z_i \in \Re^p$ denotes the covariate vector for the ith unit. Since recursive partitioning handles covariates one by one, we assume $p = 1$ for the ease of illustration. In order to ensure identifiability, we also assume that the failure time T_i^* is independent of the censoring time C_i conditional on the covariate Z_i, for any $i = 1, \ldots, n$.

In the subdistribution approach by Fine and Gray (1999), the subdistribution hazard for each type of failure is formulated with the proportional hazards model (Cox, 1972). Since we only consider splitting on a single covariate, the subdistribution hazard function $\tilde{l}(t)$ is assumed to take the following form:

$$\tilde{l}_j(t;Z_i) = \tilde{l}_{j0}(t)\exp\left[b_g^j I(Z_i < g)\right], j = 1,\ldots,J \qquad (4.1)$$

where $\tilde{l}_{j0}(t)$ is an unspecified baseline subdistribution hazard function and b_g^j is an unknown regression parameter corresponds to cutpoint g and cause of failure j. We assume that there is a change point effect of Z_i on the subdistribution hazard function with cutpoint g.

When there is no censoring, b_g^j can be estimated in exactly the same way as in the Cox model for right-censored data using a modified risk set. Here the risk set, $R(t)$, at time t is all individuals yet to experience any event plus all those individuals who has experienced event other than the jth event at a time prior to t. The risk set leads to a partial likelihood:

$$L\left(b_g^j\right) = \prod_{i=1}^{n}\left(\frac{\exp\left(b_g^j I(Z_i < g)\right)}{\sum_{k\in R(t_i)}\exp\left(b_g^j I(Z_k < g)\right)}\right)^{I(d_i=j)} \qquad (4.2)$$

The log partial likelihood is

$$l\left(b_g^j\right) = \sum_{i=1}^{n} I(d_i = j)\left(b_g^j I(Z_i < g) - \log\sum_{k\in R(t_i)}\exp\left(b_g^j I(Z_i < g)\right)\right) \qquad (4.3)$$

Based on counting process $N_i(t) = I(T_i \le t,\ d_i = j)$ and $Y_i(t) = 1 - N_i(t-)$ the score function when there is no censoring is

$$U\left(b_g^j\right) = \sum_{i=1}^{n} \int_0^{\infty} \left[I(Z_i < g) - \frac{\sum_{k \in R(s)} Y_k(s) I(Z_k < g) \exp\left\{b_g^j I(Z_k < g)\right\}}{\sum_{k \in R(s)} Y_k(s) \exp\left\{b_g^j I(Z_k < g)\right\}} \right] dN_i(s) \quad (4.4)$$

which is of the form of the usual Cox score function. Value of b_g^j that solves the score equation (4.4) is the desired estimators. In this case, the usual information calculations can be used to find the estimated standard errors of the estimated b_g^j.

When there is right censoring an inverse probability of censoring weighting technique is used. Here we let $C(t)$ be the probability of not being censored at time t. This $C(t)$ is estimated consistently by the usual Kaplan-Meier estimator that treats occurrences of competing risk as censored observations and occurrences of censoring as an event. We define a time dependent weight function, $w_i(t)$ for each observation by

$$w_i(t) = \begin{cases} \dfrac{\hat{C}(t)}{\hat{C}(\min(t, T_i))}, & \text{if } N_i(t) \text{ is observable} \\ 0, & \text{otherwise} \end{cases} \quad (4.5)$$

Note that $w_i(t)$ is nonzero for censored observations up to the time of censoring. Using this weight, an estimating equation for b_g^j is given by

$$U\left(b_g^j\right)=\sum_{i=1}^{n}\int_0^{\infty}\left[I\left(Z_i<g\right)-\frac{\sum_{k\in R(s)}w_k(s)Y_k(s)I\left(Z_k<g\right)\exp\left\{b_g^j I\left(Z_k<g\right)\right\}}{\sum_{k\in R(s)}w_k(s)Y_k(s)\exp\left\{b_g^j I\left(Z_k<g\right)\right\}}\right]w_i(s)dN_i(s) \quad (4.6)$$

Given the estimated b_g^j, the deviance is defined as $-2l\left(\hat{b}_g^j\right)$. This is the summary measure of agreement between model and the data, where the smaller value corresponds to better goodness of fit (Collet, 1994).

The estimated variance of $\hat{b}_g^j$ is given in Fine and Gray (1999). They suggested using a "sandwich" estimator.

The splitting function is defined as $R(g,h)=-2l\left(\hat{b}_g^j\right)$, which is called deviance. This statistic can be derived from likelihood ratio for testing the significance of b_g^j which $\hat{b}_g^j$ is its maximum likelihood estimator. In summary, when a tree is constructed, a proportional subdistribution hazard structure is assumed within each node. The splitting function $R(g,h)$ is evaluated at each allowable split, and the best cutpoint g^* is chosen such that $R(g^*,h)=\min_{g\in h}R(g,h)$. This process is applied recursively until all the nodes cannot be further split.

4.1.2 Algorithm to Grow Tree

To grow a tree, the deviance statistic is evaluated for every possible binary split of the predictor space Z. The split, s, could be of several forms: splits on a single covariate, split on linear combinations of predictors, and boolean combination of splits. The simplest form, in which each split relates to only one covariate, can be described as follows:

1. if Z_k is ordered, then the data will be split into two groups specified by $\{ Z_k < g\}$ and $\{ Z_k \geq g\}$ respectively;
2. if Z_k is nominal, then any subset of the possible nominal values could induce a split.

The "best split" is defined to be the one corresponding to the minimum deviance statistic. Subsequently the data are divided into two groups according to the best split.

Apply this splitting scheme recursively to the sample until the predictor space is partitioned into many regions. There will be no further partition to a node when any of the following occurs:

1. The node contains less than, say 10 or 20, observations, if the overall sample size is large enough to permit this. We suggest using a larger minimum node size than used in CART where the default value is 5. As shown by LeBlanc, M. and Crowley, J.

(1993), even for large sample sizes, "end-cut preference" (see Section 11.8, CART) can be a problem. He showed that it is important to use a larger minimum node size in order to avoid this unattractive effect;

2. All the observed times in the subset are censored, which results in unavailability of the deviance statistic for any split;
3. All the observations have identical covariate vectors or the node has only complete observations with identical survival times. In these situations, the node is considered as 'pure'.

The whole procedure results in a large tree G_0, which could be used for the purpose of data structure exploration.

4.2 Algorithm to Prune Tree

The idea of pruning is to iteratively cut off branches of the initial tree, G_0, in order to locate a limited number of candidate subtrees from which an optimally sized tree is selected. Besides the cost-complexity pruning of CART (Breiman *et al.* 1984), many pruning methods have been proposed in the literature. See, for example, Nibett and Bratko (1986), Mingers (1987) and Quinlan (1993). Those methods frequently used cross-validation and bootstrap resampling techniques to determine an appropriate tree size. However, the extensive computation of our proposed method

forbids the application of those methods. Therefore, to our proposed method, we adopt Segal's pruning algorithm (Segal, 1988) which exerts little computational burden. The step for adopting this algorithm is as follows:

- Initially grow a large tree.
- To each of the internal nodes in the original tree, assign the maximal splitting statistics contained in the corresponding branch. This statistic reflects strength of linking for the branch to the tree.
- Among all these internal nodes, finds the one with the smallest statistic. That is, find the branch that has the weakest link and then prune off this branch from the tree.
- The second pruned tree can be obtained in a similar manner by applying the above two steps to the first pruned tree.
- Repeating this process until the pruned tree contains only the root node, a sequence of nested trees is finally obtained.

The desired tree can be obtained by plotting the size of these trees against their weakest linking statistics. Usually the tree corresponding to the "kink" point in the curve is chosen as the best one.

4.3 Data Analysis

As an illustration of competing risks trees, we revisit the contraceptive discontinuation data drawn from the database of the Indonesian Demographic and Health Survey (IDHS) 2002. Beside age of start of contraceptive use, we also consider some additional covariates which suppose to be able to explain the rate of discontinuation. The important one is the contraceptive method. For this analysis, contraceptive methods were grouped into three categories: pills and injectables, IUDs and implants, and other modern methods (mainly condoms). The other covariates were woman's education (primary or lower, secondary, university), household social and economic status (1 – 7 scores), area of residence (urban, rural), and religion (Moslem, non-Moslem).

4.3.1 Subdistribution Hazard Regression

First, result of subdistribution hazard regression model by Fine and Gray (1999) is presented for comparison (Table 4.1). For the first type of risk (i.e., failure), the result in Table 4.1.(a) shows that Age and IUDs/Implants are statistically significant with *p*-value less than 5%. The older women tend to have lower discontinuation rate due to failure, and the IUDs/Implants user have lower

discontinuation rate due to failure than the user of the other methods.

Table 4.1.(b) showed the effect of covariate on subdistribution hazard of discontinuation due to abandoning. Age, Education and Method of contraception affected the abandoning rate. The sign of coefficient shows that the older women tend to have higher abandoning rate than younger ones. Whereas, the women with primary or lower education have high rate of abandoning, because both signs of coefficients for education covariates are negative. Again, the women with IUDs/implants have lower rate of abandoning.

For the third risk, switching, factors of education and contraceptive methods are two covariates which are statistically significant (see Table 4.1.(c)) .

Table 4.1. Subdistribution hazard regression for contraceptive discontinuation data

(a) Type 1 risk: failure

Variable	Coefficients	SE of Coefficients	P-value
Social Economic Status	-0.06316	0.08329	0.450
Age	-0.03931	0.01601	0.014
Residence	0.27060	0.28730	0.350
Religion	0.49220	0.59200	0.410
Secondary	-0.15390	0.36120	0.670
University	0.32110	0.39970	0.420
IUDs/Implants	-1.09200	0.51880	0.035
Other Methods	-0.06132	0.72440	0.930

(b) Type 2 risk: abandoning

Variable	Coefficients	SE of Coefficients	P-value
Social Economic Status	0.01883	0.023120	0.42000
Age	0.01544	0.005143	0.00270
Residence	0.06970	0.070450	0.32000
Religion	-0.01111	0.191700	0.95000
Secondary	-0.20910	0.083210	0.01200
University	-0.39600	0.111400	0.00038
IUDs/Implants	-0.18280	0.084400	0.03000
Other Methods	-0.32410	0.264000	0.22000

(c) Type 3 risk: switching

Variable	Coefficients	SE of Coefficients	P-value
Social Economic Status	-0.039020	0.02210	0.077000
Age	-0.004583	0.00475	0.330000
Residence	0.028400	0.07453	0.700000
Religion	-0.110000	0.19570	0.570000
Secondary	0.207500	0.09556	0.030000
University	0.509800	0.11660	0.000012
IUDs/Implants	-0.094370	0.08305	0.260000
Other Methods	0.663900	0.20270	0.001100

4.3.2 Regression Trees for Subdistribution Hazard

Type 1 risk: failure

The large initial tree G_0 was grown with minimum node size 100, or event number of type j (j = 1, 2, 3) at least 10. To avoid the end-cut preference, we set the minimum number of observations to be split restricted to be at least 20 (LeBlanc and Crowley, 1993). The initial tree for discontinuation due to failure has 12 terminal nodes as displayed in Figure 4.1. By comparing Table 4.1(a) and Figure 4.1, it is not surprising to see the first cut on age, since age is the most significant covariate in subdistribution hazard regression for discontinuation due to failure. The cutpoint age is 34.17 years with minimum deviance statistic 1061.32. Then the group of younger women is split according to IUDs/implants status, the other significant covariate in subdistribution hazard regression, with deviance statistic 990.32. Again age emerged as the splitter for the non-IUDs/implants user with the cutpoint 29.55 years (deviance statistic 911.09). The rest of the splitting is according to the other covariates.

The Segal's pruning approach would involve in the building up of a sequence of nested subtrees in which one of them will be picked up as the best tree. The plot of the size of subtree and the linking

statistic is shown in Figure 4.2. The “kink” point located at node 6 corresponds to subtree of size 4. Hence the final tree is of size 4 as shown in Figure 4.3. This tree leads to 4 groups of women, namely old women (age ≥ 34.17 years, node 3), young women (age < 34.17 years) with IUD/implant method (node 5), medium age women (29.55 ≤ age < 34.17) with non-IUD/implant method (node 7), and younger women (age < 29.55) with non-IUD/implant method (node 6). The subdistribution functions for those 4 groups are shown in Figure 4.4. The group of old women has the least failure incidence during the study period, and the medium age women with non-IUD/implant method experienced the most failure incidences. Approximately there are 7 percent of medium age women with non-IUD/implant method experiencing the failure of contraceptive method in the end of study period.

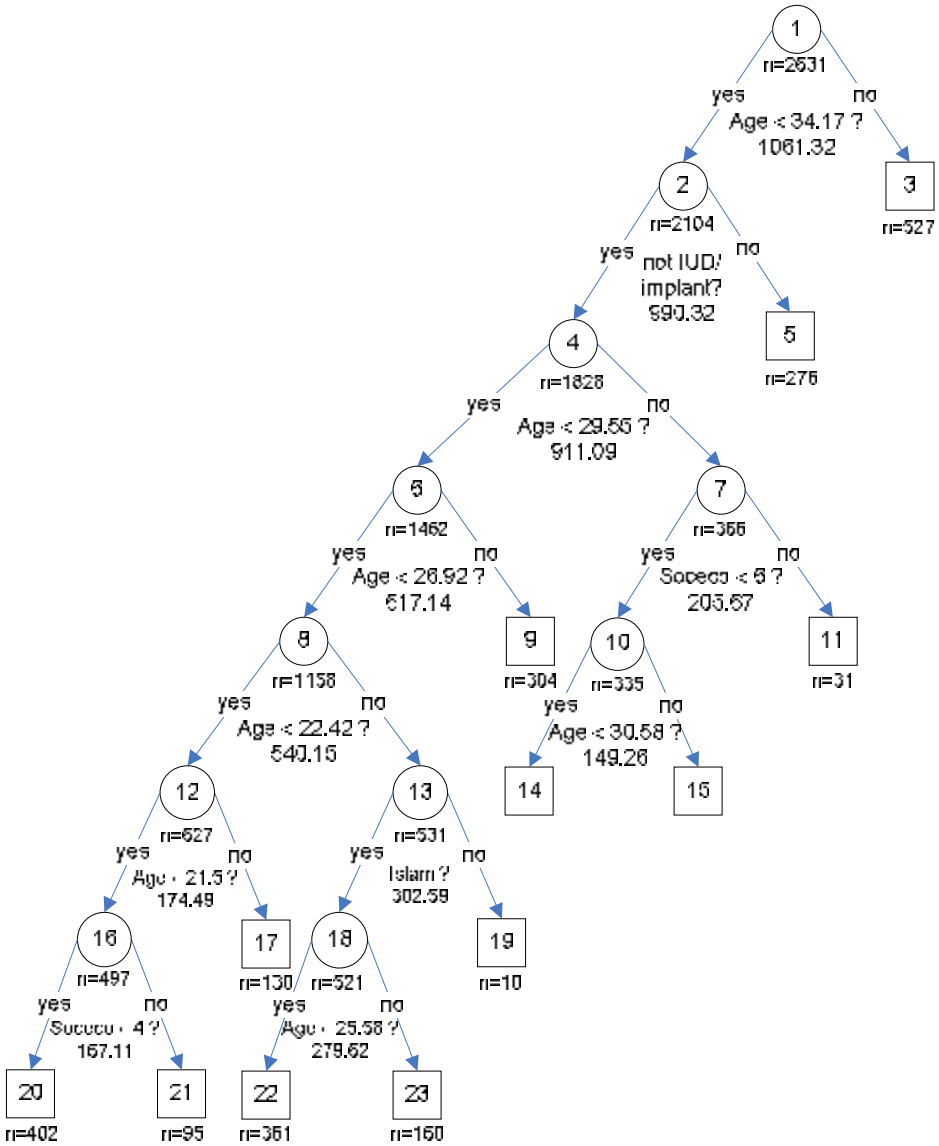

Figure 4.1. Initial tree for discontinuation due to failure (node size, split and corresponding deviance statistic)

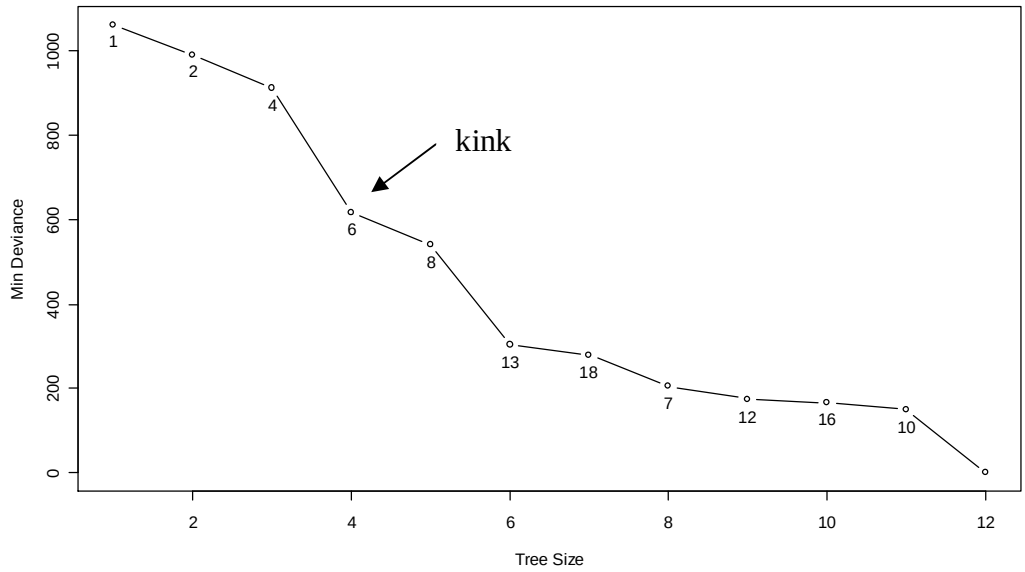

Figure 4.2. Nested subtrees of Segal's pruning for discontinuation due to failure (point label is internal node number)

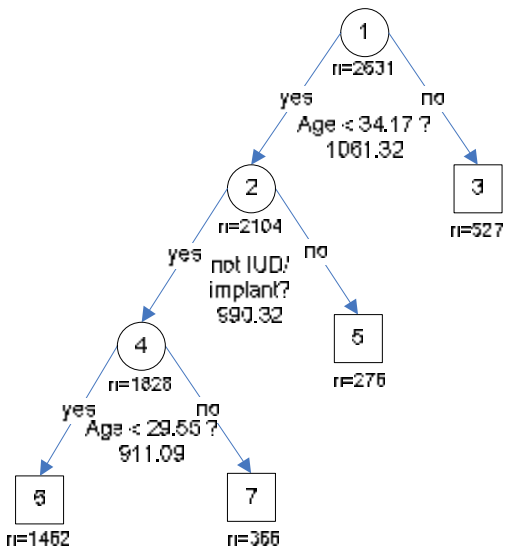

Figure 4.3. Final tree for discontinuation due to failure

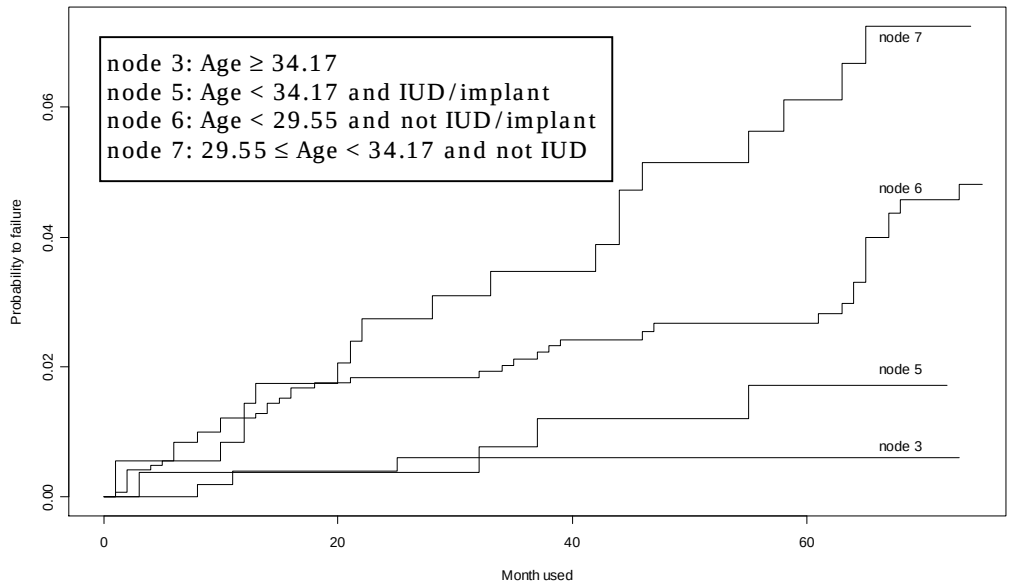

Figure 4.4. Failure subdistribution curve for 4 groups of women

Type 2 risk: abandoning

The initial tree for risk of type 2 contains 12 terminal nodes (Figure 4.5). The first split was on age at cutpoint 38 years. The group of younger women was further split by age at cutpoint 17.67 years. The medium age (17.67 ≤ age < 38) was further split by status of primary education. Contraception methods, residence and religion were not present as splitter in the initial tree. This result is slightly inconsistent with subdistribution hazard regression (Table 4.1(b)) which reveals that IUD's/implant's status is one of significant factors of probability to abandonment.

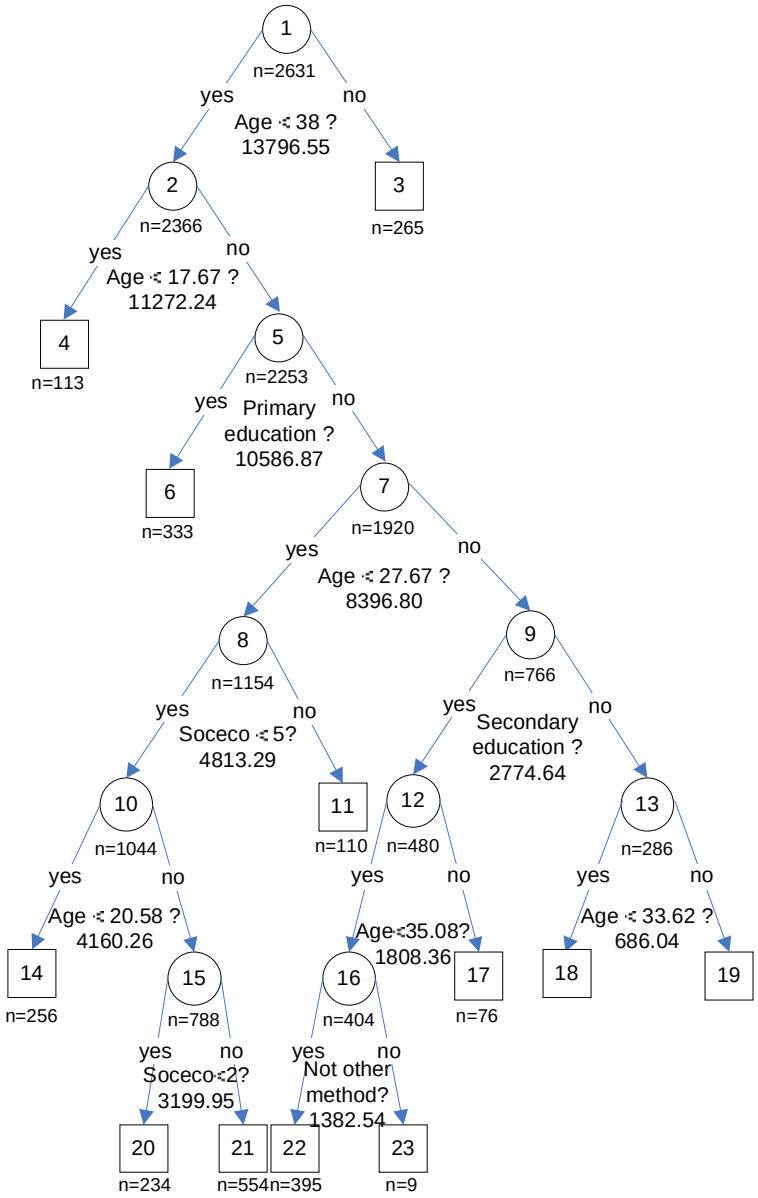

Figure 4.5. Initial tree for discontinuation due to abandoning (node size, split and corresponding deviance statistic)

The best tree for the probability of abandoning has 2 terminal nodes after pruning at node 2. Figure 4.6 shows that the "kink" was located at node 2, which means node 2 was the weakest branch, so it can be pruned off. The two final groups were old women (age ≥ 38, node 3) and younger ones (age < 38 years, node

2), presented in Figure 4.7. In terms of probability to abandonment the result of grouping was consistent with subdistribution hazard regression (Table 4.1(b)), because this group of older women has greater probability to abandonment compared to the younger one (see Figure 4.8). However, when node 2 was further split into node 4 and node 5, we find that node 4 (very young women with age < 17.67 years) has the most risk to abandonment up to about 4 years (dashed line in Figure 4.9). This is rational since the younger women have a strong willingness to be pregnant, so they most probably abandon the contraceptive methods.

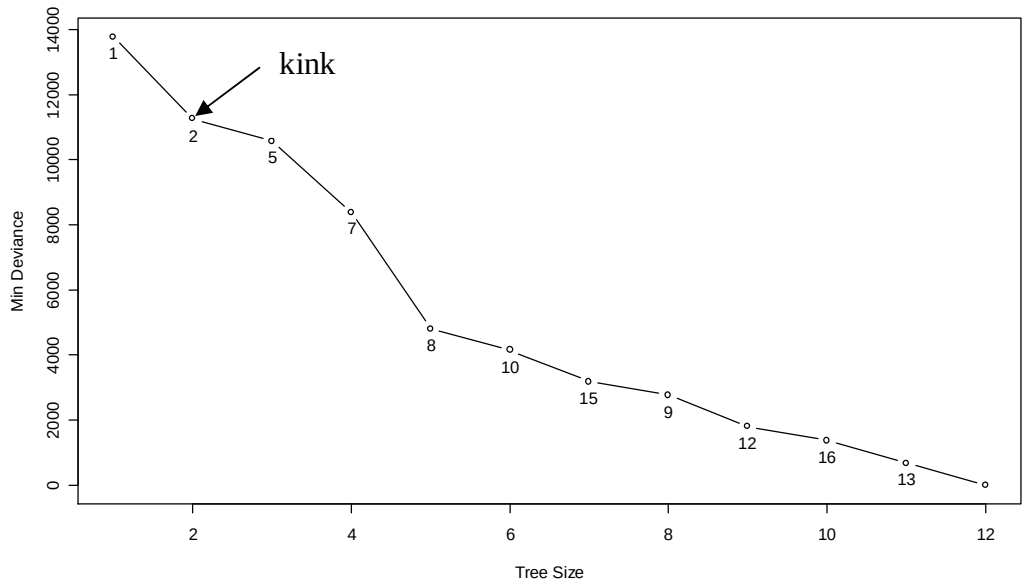

Figure 4.6. Nested subtrees for abandoning risk (point label is internal node number)

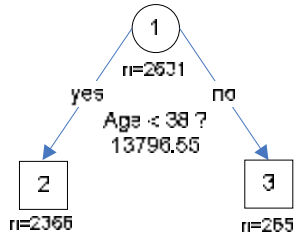

Figure 4.7. Final tree for discontinuation due to abandoning

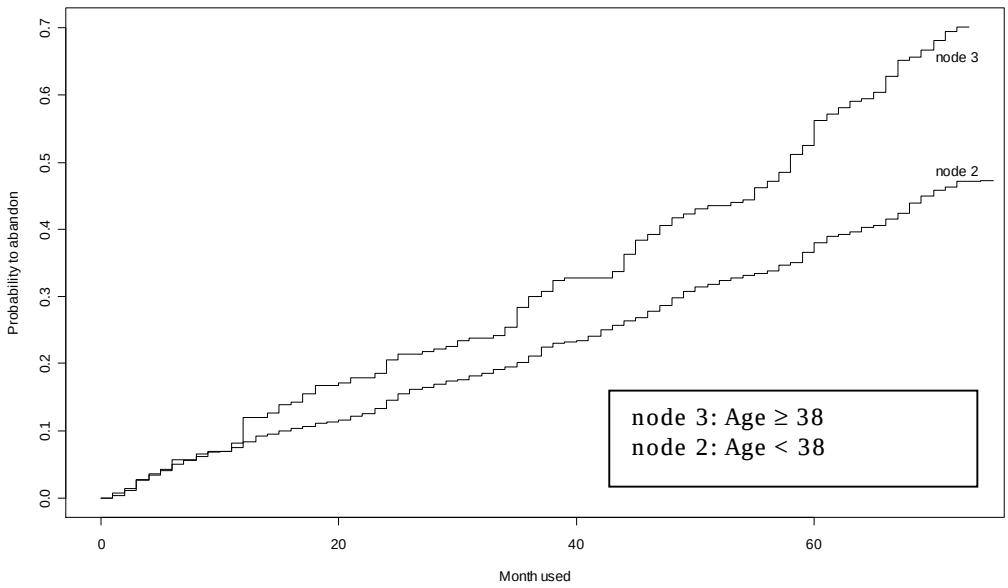

Figure 4.8. Subdistribution function of abandoning for 2 groups of women

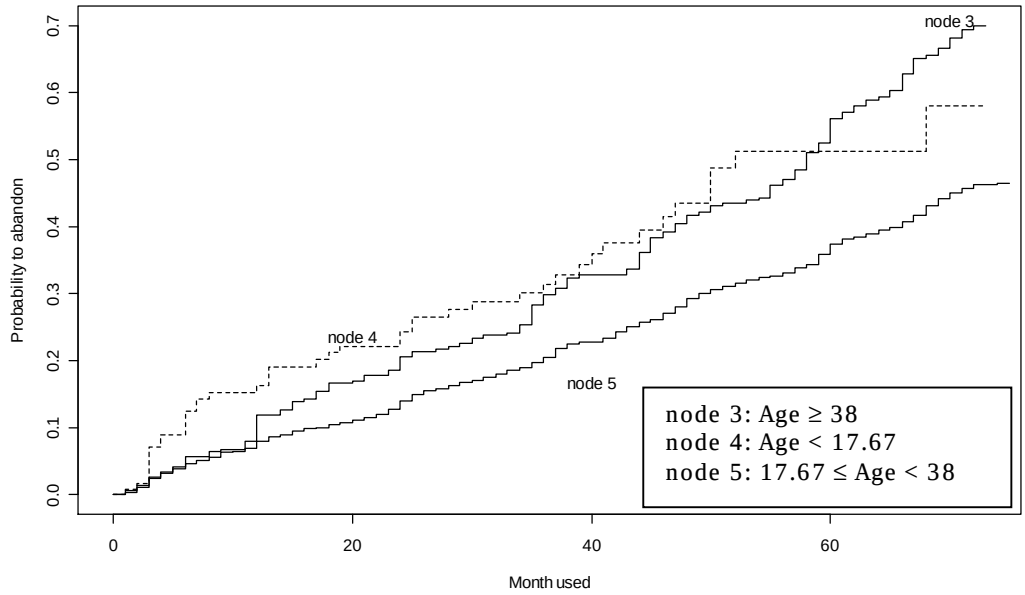

Figure 4.9. Subdistribution function of abandoning for 3 groups of women after breaking down node 2 into node 4 and node 5

Type 3 risk: switching

Regression tree analysis on time to discontinuation due to switching gave inconsistent result compared to its result based on subdistribution hazard regression (Figure 4.10 and Table 4.1(c)). The result showed that age appeared as the first splitter, whereas it was not a significant covariate in subdistribution hazard regression. The initial tree with eleven terminal nodes presented age as first splitter, breaking the subjects into old and younger women. Further split for the younger group of women was based on education which split the subjects into university-educated

women and secondary or lower-educated women. The splitting on younger university-educated women was according to contraceptive method (IUDs/Implants versus other methods). For the younger secondary or lower-educated women, node 4 was further split by age at cutpoint 17.25 years, and so on (Figure 4.10).

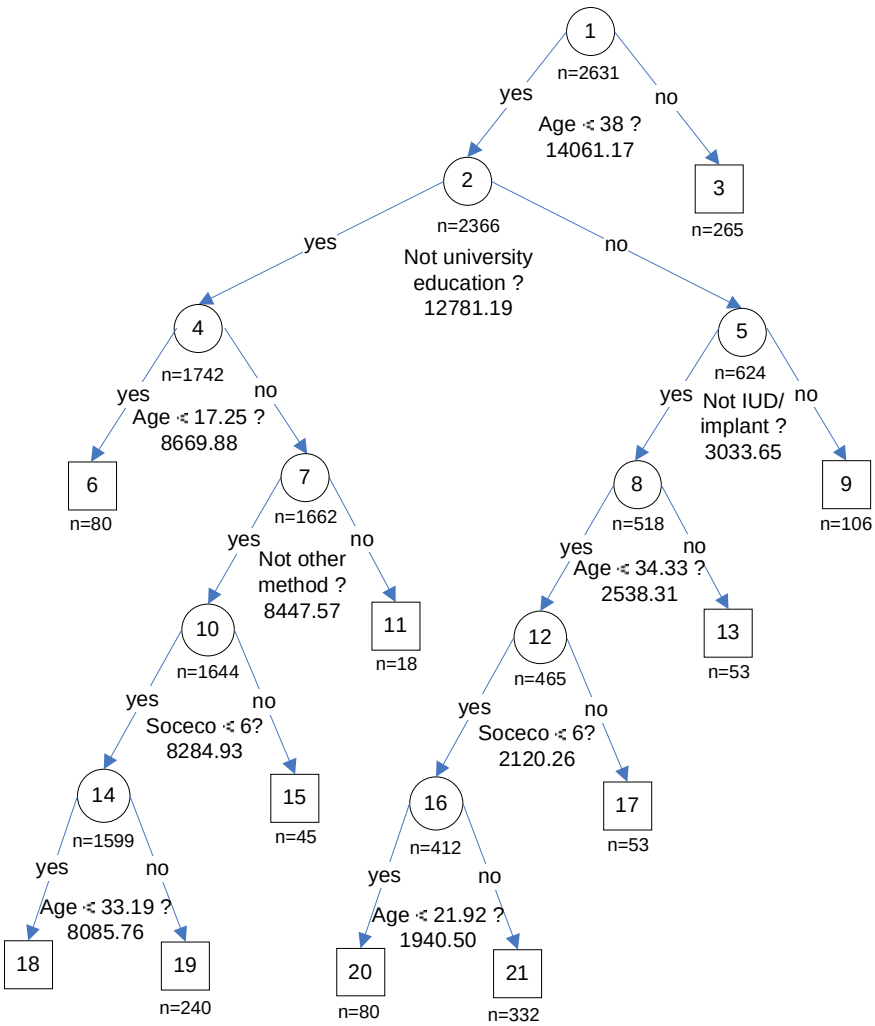

Figure 4.10. Initial tree for discontinuation due to switching (node size, split and corresponding deviance statistic)

The pruning strategy offers a sequence of subtree as presented in Figure 4.11. The "kink" is located at node 4 which gives a three-node final tree (Figure 4.12). The first group is old women (age ≥ 38 years) at node 3. The second is younger women with non-university education level (node 4). The third group is the younger women with university education level (node 5). From Figure 4.13 we obtain some important points. The younger university-educated women (node 5) were the most probable to switch, whereas the older women on node 3 were the least probable to switch during the study period. It might be related to their knowledge on the availability of other contraceptive methods. Hence, for the university-educated women, instead of using only a particular method they can easily switch to another method because of this knowledge.

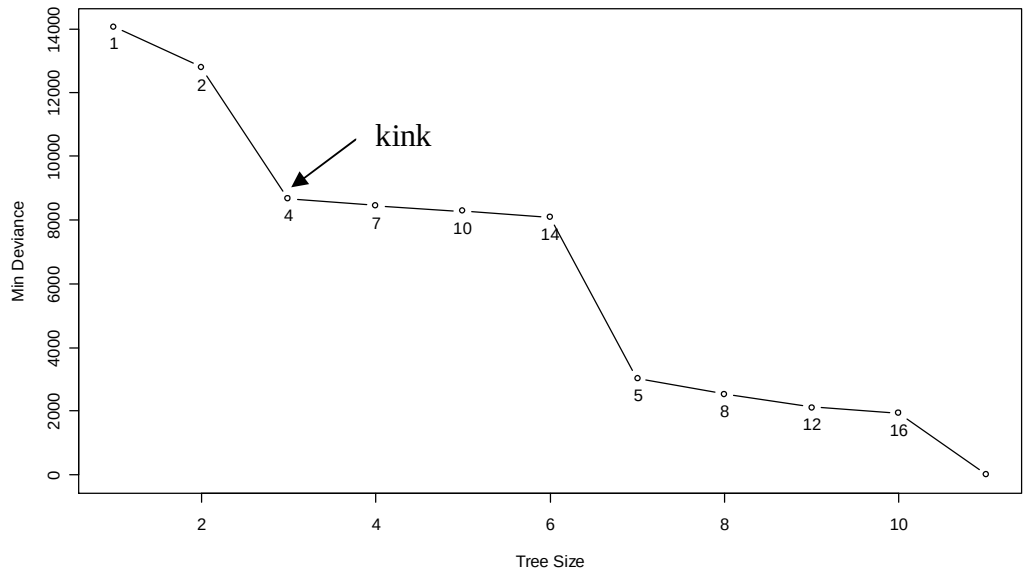

Figure 4.11. Nested subtrees for switching risk (point label is internal node number)

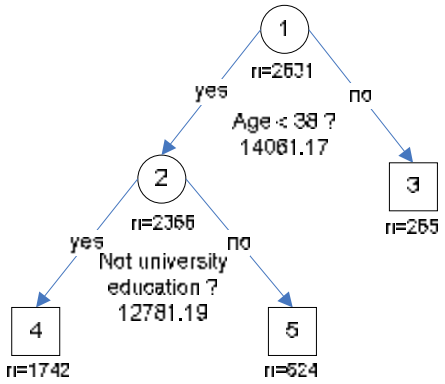

Figure 4.12. Final tree for discontinuation due to switching

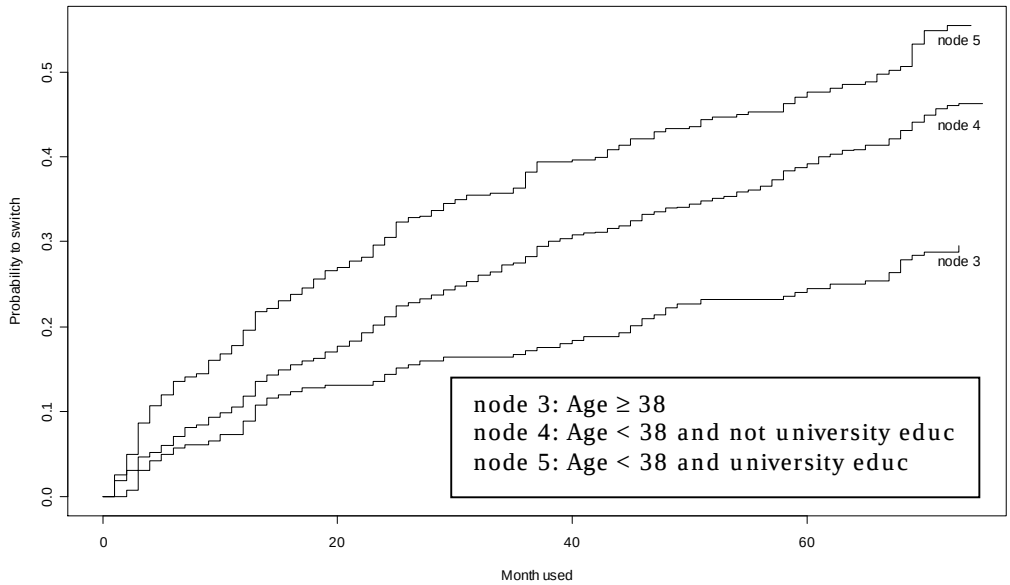

Figure 4.13. Subdistribution of switching curve for 3 groups of women

4.4 Simulation Studies

This section contains simulated experiment design to investigate the performance of the tree procedure in detecting data structure under a variety of scenarios.

Data sets are generated from the following subdistribution models using the methods described in Section 3.3:

$$F_1(t;Z_{1i},Z_{2i}) = 1-\{1-p(1-\exp(-t))\}^{\exp[b_{11}I(Z_{1i}>2)+b_{12}I(Z_{2i}>0)]},\ 0<p<1 \quad (4.7)$$

$$F_2(t;Z_i) = \{1-\exp[-t\exp(b_{21}I(Z_{1i}>2)+b_{22}I(Z_{2i}>0))]\} \times (1-p)^{\exp[b_{11}I(Z_{1i}>2)+b_{12}I(Z_{2i}>0)]} \quad (4.8)$$

This setting fulfill proportional subdistribution hazard model for first cause of failure only. To have a better evaluation of the ability to correctly identify these groupings, two extraneous variables (Z_3 and Z_4), which are not related to outcome, are also included in the data. The variables Z_1 and Z_4 have discrete uniform distribution taking values from 1 to 5, and Z_2 and Z_3 are binary variable taking value 0 and 1 with the same probability. This scenario leads to true tree G_1 or G_2 below.

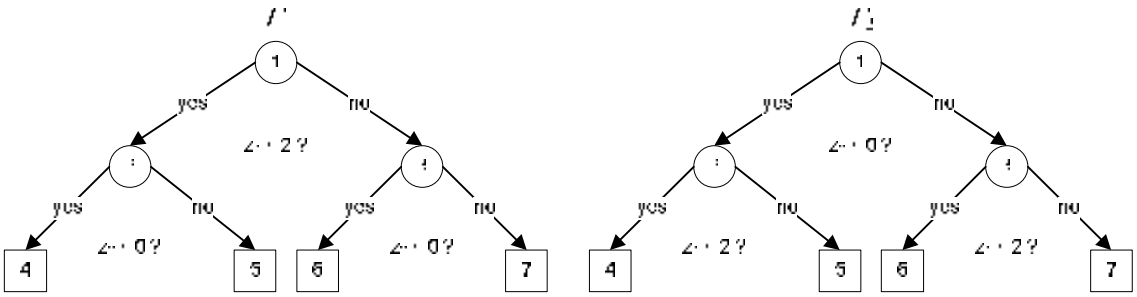

Figure 4.14. True tree for simulation

For (p,b_{11},b_{12},b_{21},b_{22}) = (0.6, 1, –1, 1, 1) resulting in 60% type 1 failure and 40% type 2 failure. In addition to the complete data, censoring times are also introduced independently from a uniform distribution (see Section 3.3.2 and Fine and Gray(1999)) to obtain

approximately 23%, 47% and 71% censoring. Each scenario has 1000 replications and each replication consist of 400 competing risks survival time data.

We examined the capability of the method to identify the data structure of true tree (figure 4.14). The resulted trees were categorized according to their ability to identify correct data structures. From the simulation we classified three categories of the capability:

1. “Correct”, if the resulted tree contains the true tree.
2. “Partially recognized”, if the optimal tree contains the part of true tree (Figure 4.15), which are G_1 or G_2 or G_3 or G_4.
3. Otherwise is a “Failed” category.

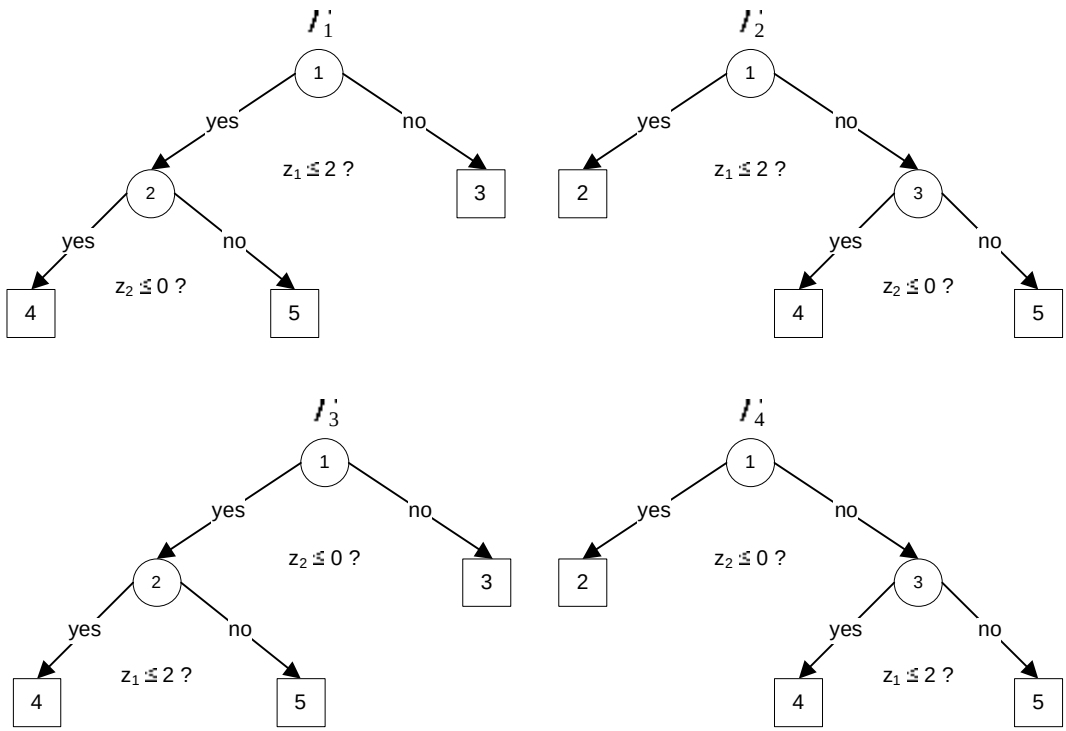

Figure 4.15. Part of true tree

Table 4.2. Simulation result on investigating the capability in identifying data structures of 1000 repetitions

Censoring(%)	Capability		
	Correct	Part	Fail
0	918	76	6
23	807	176	17
47	588	353	59
71	6	618	376

Table 4.2 summaries the simulations which show that the proposed method performs well in identifying true data structures. For no censored competing risks data the method could correctly identify more than 90% of the true structure of data, less than 10% of the true data structure could partially identified and less than 1% of them which failed to be identified. The performance is slightly decreased when the percentage of censoring is increased up to 23% and it is about 50% of data structure which could be correctly identified when the percentage of censoring is 47%. The performance gets worst for high censoring percentage; even though it is more than 50% of the true data structure could be partially identified.

4.5 Summary

This chapter proposes tree-structured regression for subdistribution of competing risks. A splitting rule that select the best partition is based on Deviance statistic which showed good

performance in selecting the best cutpoint presented in previous chapter. To apply this method we revisit the contraceptive discontinuation data and simple comparison is made with the result from Fine and Gray's subdistribution model. Since the trees method is aimed for stratifying individuals into groups, then the resulted groups is characterized by covariates which is significant in Fine and Gray's model. Hence, there is consistency on the result from both methods.

Extensive Monte Carlo simulation suggests the method has good performance in identifying the structure of data. The best performance is obtained for no censoring competing risks survival time data, even though the performance decreased as the increasing in the percentage of censoring. The result is common encountered in the simulation study for survival data.

CHAPTER 5

HYBRID MODEL FOR SUBDISTRIBUTION OF COMPETING RISKS

In this chapter, we studied a hybrid model that combines subdistribution hazards regression (Fine and Gray, 1999) with tree-structured models for subdistribution of competing risks (see Chapter 4). This hybridization will result in a new model having the merits inherited from both components.

One primary motivation for this research stems from the interesting observation that subdistribution hazards regression and tree-based models tend to complement each other in many aspects: The subdistribution hazards regression model is meant to model the linear relationship between the complementary log-log transformation of subdistribution of competing risk survival time response and the predictors (see eq. (2.25)), while it is well-known that tree-based methods are not efficient to represent linearity; the tree method is excellent at handling categorical predictors while subdistribution hazards regression defines dummy variables and may result in messy model forms, especially when the number of categories is large; subdistribution hazards regression may fail to model nonlinearity while tree methods, via step functions, often provide satisfactory approximations; detecting interaction among

covariates could be a daunting task in subdistribution hazards regression while a tree model does automatic interaction detection. On the other hand, both subdistribution hazards regression and tree methods give meaningful interpretations and are able to handle large high-dimensional data. It is rational, if subdistribution hazards regression and tree models are well combined, then the resulting model is able to improve model fit without a loss of interpretability.

Our motivation was stimulated further by the question of how to combine subdistribution hazards regression with tree-structured models. In this chapter, we propose a hybrid model which augments the subdistribution hazards model with its corresponding tree-structured regression. The main idea is to first fit the 'best' subdistribution hazards regression model and then use a tree structure as an augmentative tool to explain the remainder that has been left out by the first fit. The rest of this chapter is organized as follows; we propose a method fitting the hybrid model in section 5.1. Section 5.2 explains the use of hybrid model for competing risks by using the contraceptive discontinuation data as an illustration.

5.1 Hybrid Competing Risks Regression Model

5.1.1 Model Structure

Assume that there are J modes of failure, with potential failure time (X_{1i}, X_{2i},..., X_{Ji}) for individual i. Failure of the ith individual occurs at time $T_i^* = \min(X_{1i}, X_{2i},..., X_{Ji})$, and is due to cause $d_i^* = j$ if $T_i^* = X_{ji}, j = 1,..., J$. We suppose that the follow-up of individual i is censored at time C_i, so that what is actually observed is the time to failure or last follow-up $T_i = \min(T_i^*, C_i)$, and the failure indicator d_i, with $d_i = d_i^*$ if $T_i^* \le C_i$, and $d_i = 0$ otherwise. We assume independence across individuals and that C_i is independent of T_i^*. Besides that we also have covariates $Z_i = (Z_{1i}, Z_{2i},..., Z_{pi},) \in \Re^p$, may be a mixture of continuous and discrete variables. Thus, observed data typically consists of n independent (T_i, d_i, Z_i). Let the true model be

$$\log\left[-\log\left(1 - F_j(t_i; Z_i)\right)\right] = \tilde{l}_{j0}^*(t_i) + g(Z_i) \tag{5.1}$$

and the best subdistribution hazard model be

$$\log\left[-\log\left(1 - F_j(t_i; Z_i)\right)\right] = \tilde{l}_{j0}^*(t_i) + b_j^T Z_i^0 \tag{5.2}$$

where $Z_i^0 \in \Re^q$, $q < p$, is a q-dimensional vector selected from Z_i. Suppose that regression parameters are estimated by means of maximum likelihood method in which the log-likelihood function is

$$l(b_j) = \sum_{i=1}^{n} I(d_i = j)\left(b_j^T Z_i^0 - \log \sum_{i' \in R(t_{i'})} \exp(b_j^T Z_{i'}^0) \right) \qquad (5.3)$$

The 'best' model is searched from some candidates with subset covariates $Z_i^0 \subseteq Z_i$ based on Akaike information criterion (*AIC*) (Sakamoto *et al.*, 1986):

$$AIC = -2l(\hat{b}_j) + 2q \qquad (5.4)$$

Model with the smallest *AIC* is the 'best' model.

However, model (5.2) may not give an adequate and satisfactory fit for a given data set. To possibly make improvements, we consider the following hybrid model

$$\log[-\log(1 - F_j(t_i; Z_i))] = \tilde{l}_{j0}^*(t_i) + b_j^T Z_i^0 + a_j^T Z_i^{(G)} \qquad (5.5)$$

where the vector $Z_i^{(\Gamma)}$ comes from a tree-structured regression *G*. The tree structure *G* gives a piecewise constant approximation of $g(Z_i) - b_j^T Z_i^0$, which is the difference between the true model and the fitted subdistribution hazards model. The advantages of this hybrid approach are:

1. The subdistribution hazards model (5.2) captures global patterns, and the tree structure detects local properties left over out by model (5.2), such as nonlinear patterns and complex interactions.

2. The tree structure G not only provides useful diagnostic information about the subdistribution hazards model (5.2), but also reveals clues about how to make amendments.
3. Model (5.5) enhances the subdistribution hazards model's predictive accuracy.

5.1.2 Algorithm of Hybridization

To obtain the augmentation tree structure G, we start with the 'best' subdistribution hazards model (5.2) and then adopt the backward fitting idea of regression trees, which consists of three steps: (i) growing a large initial tree G_0 by incorporating the linear model effect, (ii) pruning it back to a nested sequence of subtrees, and (iii) selecting the optimal tree size. The detailed steps are given below.

Growing a Large Tree

To split the data, consider the following model:

$$\log\left[-\log\left(1-F_j\left(t_i;Z_i\right)\right)\right]=\tilde{l}^{*}_{j0}\left(t_i\right)+b_j^{T}Z_i^{0}+aI\left(Z_{ki}<g\right) \qquad (5.6)$$

The indicator function $I(Z_{ki} < g)$ corresponds to a binary split of the data according to a continuous predictor Z_k. If the predictor

is discrete with values in $D = \{d_1, \ldots, d_r\}$, then any form of $I(Z_{ki} \in A)$ with $A \subset D$ is considered.

The best split s^* is the one associated with the least deviance from fitting model (5.6) with the available predictors (see Subsection 3.2.2). We then iteratively estimate the most significant change-point covariate effect with the smallest deviance to split the data, which yields a large initial tree G_0.

Pruning

Given the large initial tree G_0, a sequence of nested subtree is searched and the best subtree will be selected adopting Segal's pruning algorithm (Segal, 1988). The algorithm is as follows:

- Initially grow a large tree G_0.
- To each of the internal nodes in the original tree, assign the maximal splitting statistics contained in the corresponding branch. This statistic reflects strength of linking for the branch to the tree.
- Among all these internal nodes, find the one with the smallest statistic. That is, find the branch that has the weakest link and then prune off this branch from the tree.
- The second pruned tree can be obtained in a similar manner by applying the above two steps to the first pruned tree.

- Repeating this process until the pruned tree contains only the root node, a sequence of nested trees is finally obtained.

The best tree G can be obtained by plotting the size of these trees against their weakest linking statistics. Usually the tree corresponding to the "kink" point in the curve is chosen as the best one.

For a given tree structure G in model (5.5), let $\tilde{G}$ denotes the set of all terminal nodes in G and $|\cdot|$ represent cardinality. Then, we define an $n \times |\tilde{G}|$ matrix $Z^{(G)}$ such that

$$Z_{hi}^{(G)} = \begin{cases} 1, & \text{if } i\text{th observation} \in h\text{th terminal node of } G \\ 0, & \text{otherwise} \end{cases}$$

The 'best' model (5.2) is augmented by the 'best' tree G to form the hybrid model (5.5).

5.2 Example: Contraceptive Discontinuation Data

To illustrate, we revisit the contraceptive discontinuation data drawn from the database of the Indonesian Demography and Health Survey (IDHS) 2002. The response variable is the time to discontinuation of using a particular contraceptive method, and the 6 independent variables are listed in Table 5.1.

Table 5.1. Variable descriptions for contraceptive discontinuation data

Var	Name	Description
1.	soceco	household social and economic status (score 1-7)
2.	age	age of start of contraceptive (years)
3.	resid	area of residence (0=rural,1=urban)
4.	relig	religion (0=Moslem, 1=non-Moslem)
5.	educ	woman's education (0= primary or lower, 1=secondary, 2=university)
6.	method	contraceptive methods (1=pills and injectables, 2=IUDs and implants, 0= other modern methods (mainly condoms))

This contraceptive discontinuation data have been explored in Chapter 3 and 4. Next, we apply the hybrid approach to get a better insight of these data. Categorical variables with more than two categories should be converted to dummy variables first. Women's education levels (educ) is converted to two dummy variables namely educ1 and educ2. Variable contraceptive method (method) is also converted to two dummy variables, method1 and method2. The dummy variables that we used are presented in Table 5.2.

Table 5.2. Dummy variables construction

educ	educ1	educ2
0	0	0
1	1	0
2	0	1

method	method1	method2
0	0	0
1	1	0
2	0	1

Type 1 risk: failure

By means of AIC criterion, the best subdistribution hazards model for first risk of discontinuation (failure) contained age, educ2 and method2 predictor variables as presented in Table 5.3. This model has the minimum AIC which was 1070.987.

Table 5.3. The best subdistribution hazards regression for discontinuation due to failure

Variable	Coefficients	SE of Coefficients	P-value
age	-0.03658	0.01511	0.016
educ2	0.44630	0.25360	0.079
method2	-1.10400	0.51890	0.033
AIC = 1070.987			

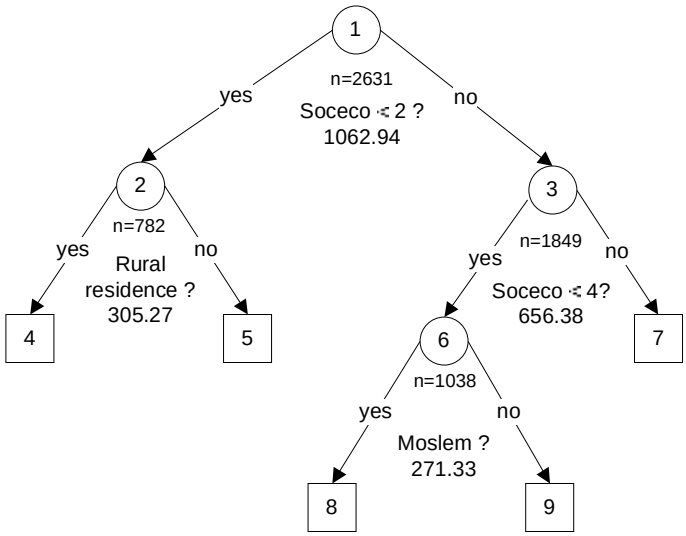

Figure 5.1. The large initial augmentation tree for discontinuation due to failure

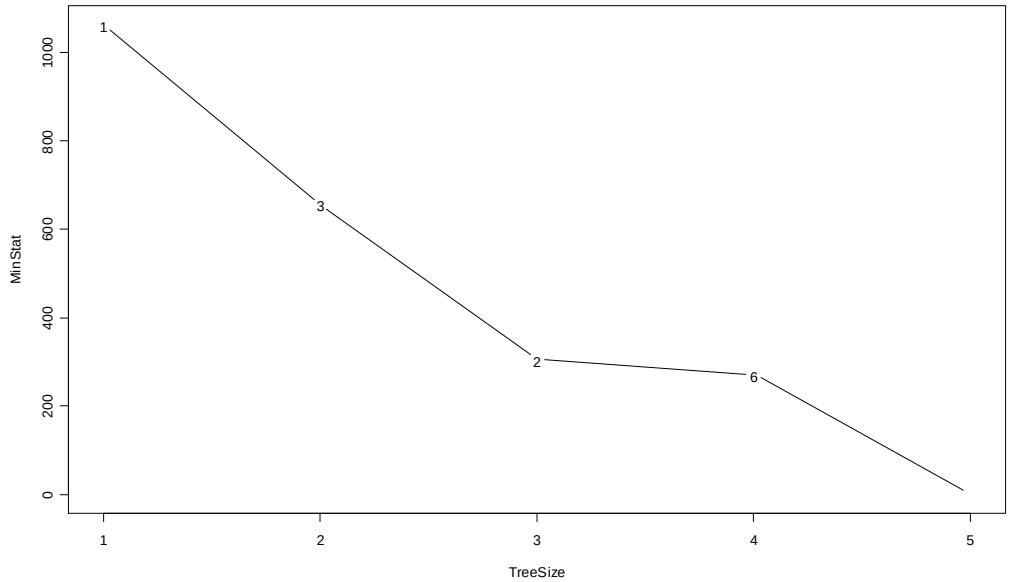

Figure 5.2. Nested subtrees of Segal's pruning for the augmentation trees (first risk, discontinuation due to failure)

Given that the best AIC model contained three predictor variables (age, educ2 and method2), it is interesting to use the hybrid method to boost it. To proceed, we constructed the augmentation tree. The initial tree and sequence of nested subtrees for pruning are presented in Figure 5.1 and Figure 5.2, respectively. The final augmentation tree as shown in Figure 5.3 has three terminal nodes. The first split was according to soceco < 2. For those women with soceco ≥ 2, their discontinuation time further differed by soceco < 4. This indicates that the effect of this predictor on discontinuation time has not been represented by the best AIC model (Table 5.3). Here, we augmented the best AIC subdistribution hazards regression model with a tree structure resulted in a hybrid model, which provided a feasible way of exploiting the merits of

both methods. The fitted hybrid model is displayed in Table 5.4. To compare with the best AIC model, we performed likelihood ratio test. The resulted *p*-value was 0.1718812 which indicated that hybrid model did not constitute a substantial improvement over best AIC model.

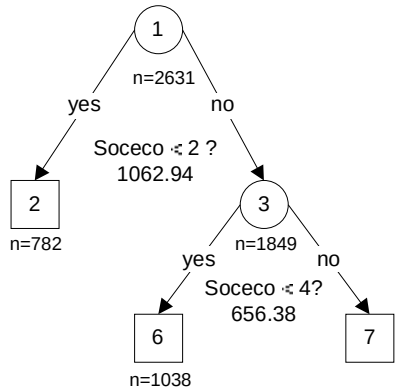

Figure 5.3. The final augmentation tree for discontinuation due to failure

Table 5.4 The hybrid regression for discontinuation due to failure

Variable	Coefficients	SE of Coefficients	P-value
age	-0.03543	0.01498	0.018
educ2	0.49770	0.25360	0.050
method2	-1.14400	0.51690	0.027
node6	-0.54210	0.30360	0.074
node7	-0.17020	0.28280	0.550
AIC = 1071.465			

Type 2 risk: abandonment

For second risk (abandonment), the minimum AIC, 13817.70, was attained by subdistribution hazards model which contained covariates age, educ1, educ2 and method1 (see Table 5.5). All of the covariates were significant at 5% level. The initial augmentation tree had five terminal nodes (see Figure 5.4). The

Segal's pruning algorithm revealed that the weakest branch was node 2 (kink location), which resulted in final augmentation trees of size 2 (see Figure 5.5). The splitter was soceco at cutpoint 3 (see Figure 5.6). Hence the hybrid model contained five covariates. The additional covariate for the final model was node3, dummy variable for soceco ≥ 3. Even though this additional dummy variable and likelihood ratio test were not significant at a = 5%, but its AIC statistic was slightly smaller, 13816.76, compared to the initial model (see Table 5.6).

Table 5.5. The best subdistribution hazards regression for discontinuation due to abandonment

Variable	Coefficients	SE of Coefficients	p-value
age	0.01709	0.004863	0.00044
educ1	-0.17970	0.080320	0.02500
educ2	-0.32750	0.098430	0.00088
method1	0.20060	0.082440	0.01500
AIC = 13817.70			

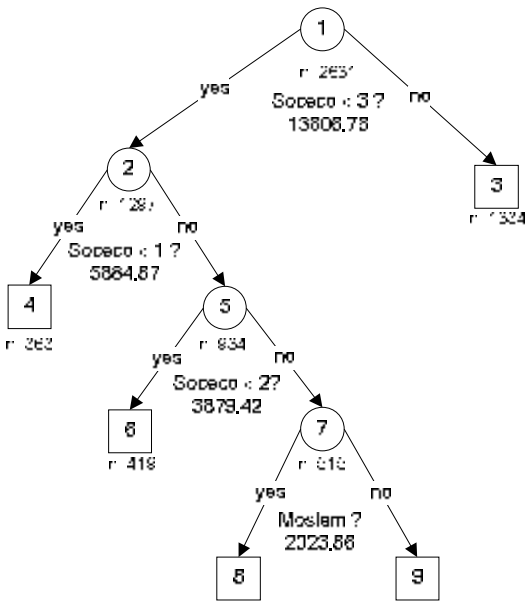

Figure 5.4. The large initial augmentation tree for discontinuation due to abandonment

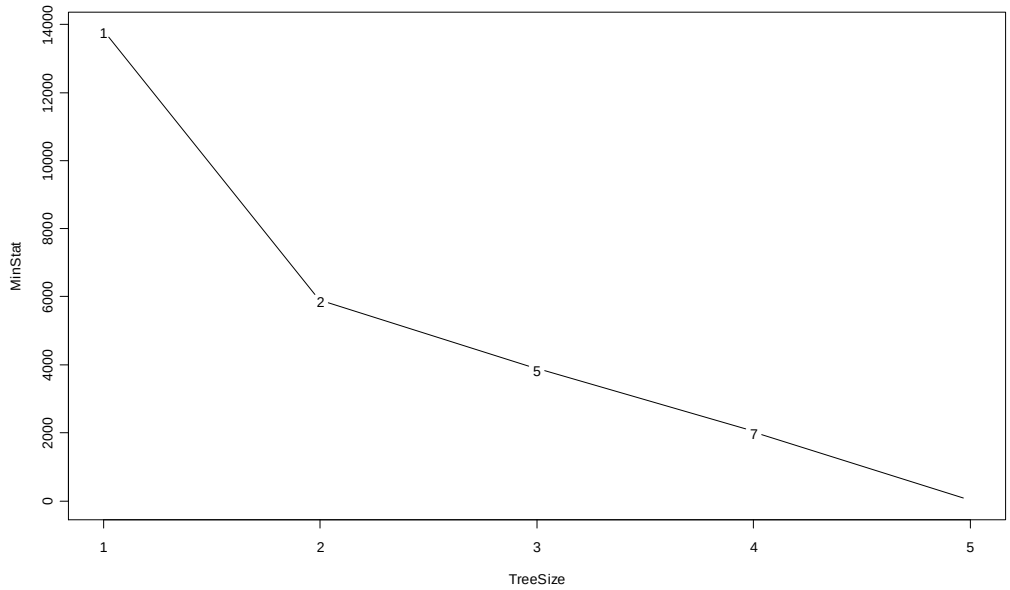

Figure 5.5. Nested subtrees of Segal's pruning for the augmentation trees (second risk, discontinuation due to abandonment)

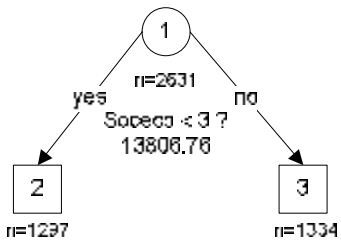

Figure 5.6. The final augmentation tree for discontinuation due to abandonment

Table 5.6. The hybrid regression for discontinuation due to abandonment

Variable	Coefficients	SE of Coefficients	*p*-value
age	0.01529	0.005027	0.00240
educ1	-0.21150	0.082560	0.01000
educ2	-0.40070	0.105900	0.00015
method1	0.20020	0.082530	0.01500
node3	0.12330	0.069560	0.07600
AIC = 13816.76			

Type 3 risk: switching

Subdistribution hazards model for risk to switching with covariates soceco, educ1, educ2, method1 and method2 had minimum AIC, 14056.20. All covariates were significant at a = 5% with the least p-value for covariate educ2, university-educated level (see Table 5.7).

Table 5.7. The best subdistribution hazards regression for discontinuation due to switching

Variable	Coefficients	SE of Coefficients	P-value
soceco	-0.04172	0.02022	0.03900
educ1	0.23090	0.09241	0.01200
educ2	0.53350	0.11320	0.00000
method1	-0.63930	0.20060	0.00140
method2	-0.75120	0.20710	0.00029
AIC = 14056.20			

The initial large augmentation tree had eleven terminal nodes (see Figure 5.7). Almost all of the splitting was based on covariate age. It seemed that age was an important covariate associated with time to switching, although it was not present in the best AIC subdistribution model.

Segal's pruning algorithm yielded a final augmentation tree of size 2. Kink location at node 2 leads to additional dummy covariate node 3 (age ≥ 38). These are presented in Figure 5.8 and 5.9.

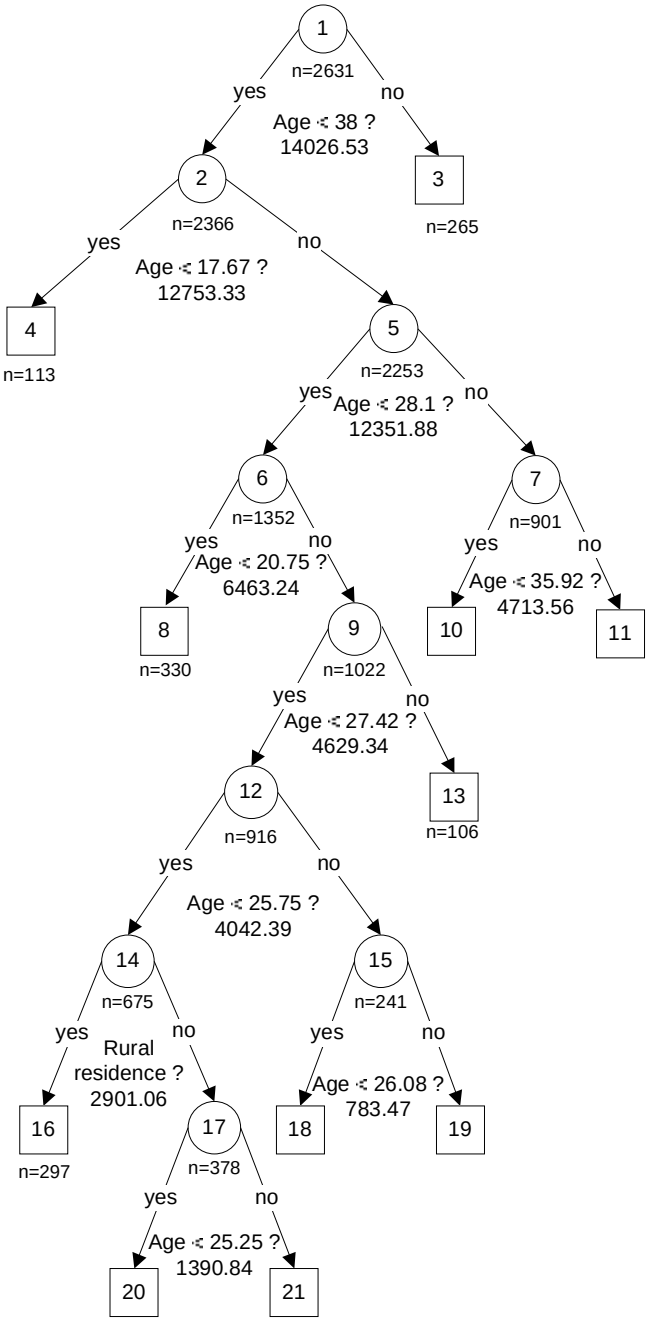

Figure 5.7. The large initial augmentation tree for discontinuation due to switching

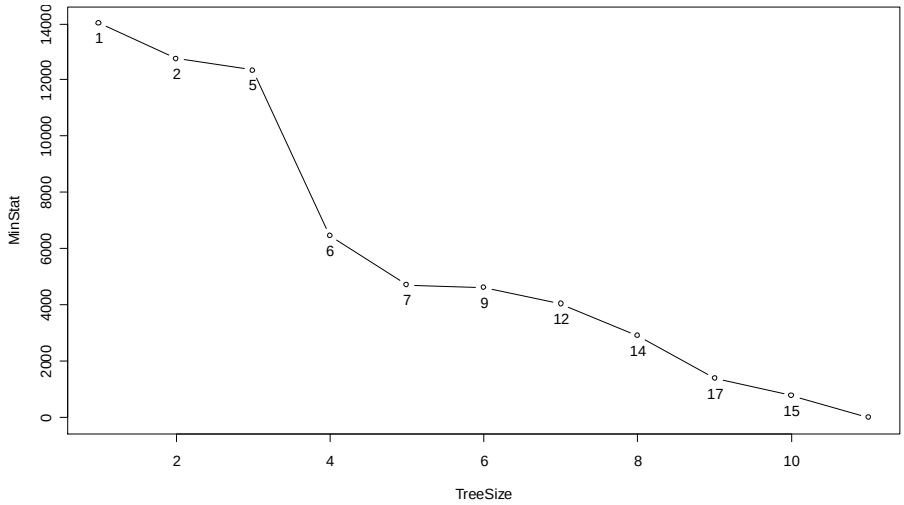

Figure 5.8. Nested subtrees of Segal's pruning for the augmentation trees (third risk = switching)

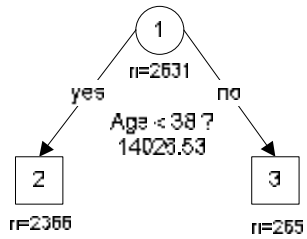

Figure 5.9. The final augmentation tree for discontinuation due to switching

The final hybrid subdistribution hazard model had smaller AIC than the initial one. In addition, its *p*-value of likelihood ratio test was also significant (*p*-value = 9.202903×10^{-6}), even though covariate soceco was not significant (see Table 5.8).

Table 5.8. The hybrid regression for discontinuation due to switching

Variable	Coefficients	SE of Coefficients	*p*-value
soceco	-0.03142	0.02023	0.12000
educ1	0.17560	0.09286	0.05900
educ2	0.45420	0.11400	0.00007
method1	-0.67340	0.20410	0.00097
method2	-0.76150	0.21030	0.00029
node3	-0.51770	0.12390	0.00003
AIC = 14038.53			

5.3 Summary

In this chapter we proposed a new methodology to analyze competing risks survival data, which in particular combined the merits of Fine and Gray's proportional hazards model for subdistribution of competing risks and our developed regression trees for competing risks. Although semiparametric model of Fine and Gray's has been extensively studied, to our knowledge no extension which combined it with regression trees methodology. We first searched the best AIC model of Fine and Gray's model and then augment it with tree. The augmentation was aimed to boost the best AIC model by adding the terms which represented the left over parts of the best AIC model.

Application of the proposed method to the contraceptive discontinuation data showed that the hybrid models had the lower AIC compared to best AIC model of Fine and Gray's. It showed that the proposed method is better than the current one.

CHAPTER 6

PARAMETRIC REGRESSION FOR SUBDISTRIBUTION OF COMPETING RISKS BASED ON NON-MIXTURE CURE MODEL

The competing risks data consist of the failure time (T) and the cause of failure (d). Modelling latent survival time and joint distribution of (T, d) are two approaches which generally used for addressing competing risks data. The first approach assumes independency among latent survival time. However, statistical analysis for independent competing risks data under various parametric models has been considered (David and Moeschberger, 1978). The techniques used for analysing right censored data can be implemented if the approach (i) is used and standard parametric, semiparametric or nonparametric models can be used without any complications. Even when latent failures are not independent multivariate parametric forms can be assumed and statistical analysis can be done (Moeschberger, 1974). By using second approach, cause specific hazard rate is used to model the competing risks data (Kalbfleisch and Prentice, 1980). Nonparametric techniques for estimating and testing cause-specific hazards have been developed. There has been some discussion about the use of Cox's proportional hazards model, which is a semiparametric model, for the cause-specific hazards (Crowder,

2001). However, there has been very little done in specifying parametric models for the cause-specific hazards or for the subdistribution functions.

To our knowledge only Jeong and Fine (2006) which model subdistribution parametrically. Recently parametric subdistribution model had been extended to account for covariate in regression setting (Jeong and Fine 2007). A difficulty in specifying a parametric subdistribution survival distribution is because it is an improper distribution function. One way to overcome this difficulty is to specify a parametric model for the subdistribution function which could in practice take any form of the improper distribution function. To do so, Jeong and Fine (2006, 2007) utilize Gompertz distribution, which can take the form of improper distribution, for modeling parametric subdistribution function. Unfortunately, this approach could not be used for developing subdistribution function systematically. It means that we can not use various other well known distribution functions which commonly used in the survival analysis, e.g. exponential, Weibull, gamma and generalized gamma. Standard methods can be employed for fitting parametric subdistribution functions. Analytical methods like maximum likelihood estimation can be used to estimate the unknown parameters.

To address this issue, we model the subdistribution function for each cause of failure directly. By considering the cure fraction parameter in the non-mixture cure model of Yakovlev and Tsodikov (1996) as the proportion of individuals who don't experience the event of interest in the competing risks setting, we could develop subdistribution model based on non-mixture cure model. Beside that, non-mixture cure model which contained kernel distribution make it is possible to develop subdistribution model based on various well known kernel distributions.

Our proposed parametric model for subdistribution of competing risks might take the form of improper subdistribution function. Improper means the asymptotic value of the function cannot reach value 1 following the usual property of distribution function.

The non-mixture cure model and its relationship with the proposed direct modelling of subdistribution are presented in Section 6.1. In Section 6.2, we developed maximum likelihood estimation for the proposed model. In Section 6.3, simulation study was performed to evaluate the efficiency of the parameter estimates. The data analysis example is in Section 6.4. We also proposed new subdistribution model which had the property like Gompertz distribution and called it Gompertz-like subdistribution in Section 6.5. Finally, Section 6.6 is the conclusion of the chapter.

6.1. Parametric Subdistribution

6.1.1 Univariate Model

The development of subdistribution model is carried out by considering non-mixture cure model. Non-mixture cure model had been proposed by Yakovlev and Tsodikov (1996), Tsodikov (1998) and Chen *et al.* (1999) where it assumed bounded cumulative hazard $H(t)$ as $t \to \infty$,

$$H(t) \le q^*, \ \lim_{t\to\infty} H(t) = q^*$$

One way to enforce the above property is to write $H(t) = q^*F^*(t)$, where $F^*(t)$ is the distribution function of a nonnegative random variable, called kernel distribution. Then the survival distribution is $S(t) = e^{-q^*F^*(t)}$ and the distribution function can be written as

$$F(t) = 1 - \left[\exp(-q^*)\right]^{F^*(t)}, \ q^* > 0 \qquad (6.1)$$

In this model, the distribution function converges to $1-p$, where $p = \exp(-q^*)$, overtime; hence, this is not a proper distribution. The parameter p is interpreted to be the cure fraction. In the competing risks framework, we utilized cure fraction to model the proportion of individuals who don't experience the event of interest.

For this purpose, the subdistribution function for cause *j* can be formulated by

$$F_j(t) = 1 - \left[\exp\left(-q_j^*\right)\right]^{F_j^*(t)} \tag{6.2}$$

where $P\left(d^* = j\right) = \lim_{t\to\infty} F_j(t) = 1 - \exp\left(-q_j^*\right)$ which is the probability of failure due to cause *j*. In the other hand, $P\left(d^* \neq j\right) = \exp\left(-q_j^*\right)$ which is the proportion of individuals who don't experience the j^{th} cause and $\sum_{j=1}^{J} F_j(t) = F(t)$. Chen *et al.* (1999) gave a Bayesian discussion on model (3) and Sposto (2002) applied it to pediatric cancer data.

Exponential kernel

Let us assumed $F^*(t)$ be an exponential distribution with parameter *k*. Then the resulted subdistribution is

$$F_j\left(t; q_j^*, k_j\right) = 1 - \left[\exp\left(-q_j^*\right)\right]^{1-\exp(-k_j t)} \tag{6.3}$$

It is clear that subdistribution (6.3) is improper when $k_j > 0$ and $0 < q_j^* < \infty$. This model encompasses Gompertz subdistribution which was previously proposed by Jeong and Fine (2006).

Weibull kernel

Suppose that Weibull kernel distribution is used in the construction of subdistribution of competing risks. The corresponding subdistribution model is formulated by

$$F_j\left(t;q_j^*,k_j,a_j\right)=1-\left[\exp\left(-q_j^*\right)\right]^{1-\exp(-k_j t^{a_j})} \tag{6.4}$$

where $0<q_j^*<\infty$, $k_j>0$ and $a_j>0$. Note that this distribution includes mixture model which was proposed by Larson and Dinse (1985) and Maller and Zhou (1996, 2002).

We can also use another kernel distribution such as Gompertz, gamma and generalized gamma. Table 6.1. showed the resulted subdistribution based on those three kernel distributions.

Table 6.1. Kernel distribution and the resulted subdistribution for three distributions

Distribution	Kernel distribution	Subdistribution
Gompertz	$1-\exp\left\{t_j\left[1-\exp\left(r_j t\right)\right]/r_j\right\}$	$1-\left[\exp\left(-q_j^*\right)\right]^{1-\exp\left\{t_j\left[1-\exp\left(r_j t\right)\right]/r_j\right\}}$
Gamma	$I(k_j t,g_j)^*$	$1-\left[\exp\left(-q_j^*\right)\right]^{I\left(k_j t,g_j\right)}$
Generalized gamma	$I\left(k_j t^{a_j},g_j\right)$	$1-\left[\exp\left(-q_j^*\right)\right]^{I\left(k_j t^{a_j},g_j\right)}$

* $I\left(t,g\right)=\frac{1}{G\left(g\right)}\int_0^t u^{g-1}e^{-u}du$ and $G\left(g\right)=\int_0^\infty u^{g-1}e^{-u}du$

6.1.2 Regression Model

Frequently, however, the model can be improved by including relevant explanatory variables $\boldsymbol{Z} = (Z_1,...,Z_K)^T$ also known as covariates. When covariates are included in the model, the primary question of interest concerns the relationship between the time at which the subject fails from any cause T and the explanatory variable Z. For instance, this is the case when treatments need to be compared or when risk factors are identified for a particular disease. This covariate can be incorporated into cure fraction parameter as follows

$$F_j(t) = 1 - \left\{\exp\left[-q_j \exp(\boldsymbol{z}' b_j)\right]\right\}^{F_j^*(t)} \tag{6.5}$$

where b_j is $K\times1$ parameter vector and z is a time independent $K\times1$ covariate vector without constant.

For direct regression modelling of the subdistribution function, we could also use some well known distribution as kernel, such as exponential, Weibull, Gompertz, gamma and generalized gamma. For example, if kernel distribution is Weibull then (6.5) became

$$F_j(t; q_j, k_j, a_j, b_j, \boldsymbol{z}) = 1 - \left\{\exp\left[-q_j \exp(\boldsymbol{z}' b_j)\right]\right\}^{1-\exp(-k_j t^{a_j})} \tag{6.6}$$

Reparameterized version of (6.6) is parametric regression of proportional hazard of subdistribution proposed by Jeong and Fine (2007).

The complementary log-log transformation of (6.5) is

$$\log[-\log\{1-F_j(t)\}] = \log[q_j F_j^*(t)] + \mathbf{z}' b_j \tag{6.7}$$

This is the parametric version of proportional subdistribution hazard regression proposed by Fine and Gray (1999) where $F_j^*(t)$ is fully specified by kernel distribution.

6.2. Maximum Likelihood Estimation

Suppose that competing risks data consist of observations $t_1,\ldots,t_n$ on the lifetimes of n individuals and $d_1,\ldots,d_n$ as censor indicators, where

$$d_i = \begin{cases} 1, & \text{if individual } i \text{ is uncensored} \\ 0, & \text{if individual } i \text{ is censored,} \end{cases} \tag{6.8}$$

and, if d_i=1, i.e. individual i dies, we also observe the cause of death which takes values in $\{1,\ldots,J\}$. Thus we can define, and observe, the indicators

$$d_{ji} = \begin{cases} 1, & \text{if individual } i \text{ dies of cause } j \\ 0, & \text{otherwise} \end{cases} \tag{6.9}$$

for $1 \le i \le n$, $1 \le j \le J$.

Inference for the subdistribution function is through the likelihood function

$$L(y) = \prod_{i=1}^{n} \left[\prod_{j=1}^{J} \{ f_j(t_i) \}^{d_{ji}} \right] [S(t_i)]^{1-d_i} , \qquad (6.10)$$

where $S(t)$ is overall survival function, $S(t) = 1 - \sum_{j=1}^{J} F_j(t)$, and $f_j(t) = dF_j(t)/dt$ is an improper probability subdensity function for the jth cause-specific event. Formulating the likelihood in terms of the direct subdistribution function differs from the traditional formulation via the cause-specific hazard function in Prentice *et al.* (1978). Since the likelihood function (6.10) cannot be factored into a product of J cause-specific functions, then all parameters should be estimated simultaneously.

From (6.10), the log-likelihood function is given by

$$l(y) = \sum_{i=1}^{n} \left\{ \left[\sum_{j=1}^{J} d_{ji} \log\{ f_j(t_i) \} \right] + (1 - d_i) \log\{ 1 - \sum_{j=1}^{J} F_j(t_i) \} \right\} \qquad (6.11)$$

We presented the likelihood function for kernel distribution of exponential, Weibull, Gompertz, gamma and generalized gamma in appendix D.

We assume a parameterization for subdistribution F_j of the form

$$F_j(t) = F(t, y_j) \tag{6.12}$$

Here y_j is a (1+q+K)-vector of parameters, $y_j = (q_j, f_{j1},\ldots,f_{jq}, b_{j1},\ldots,b_{jK})$ where q_j represents parameter for cure fraction,($f_{j1},\ldots, f_{jq}$) for kernel distribution and ($b_{j1},\ldots, b_{jK}$) for covariates, for each cause of failure j=1,...,J. Set all parameters into a $p\times1$ vector: $y = (y_1,\ldots, y_J)$, where $p = (1+q+K)J$. Let u(y) be the $p\times1$ score vector which contains the first derivative of $l(y)$ with respect to y-parameters. Also, let the $p\times p$ matrix I(y) be the observed information matrix which contains negative second derivatives of $l(y)$, so that

$$\mathbf{I}(y) = -\frac{\partial^2 l(y)}{\partial y \partial y'} \tag{6.13}$$

Setting the resulting score function equal to 0, the maximum likelihood estimator (MLE) $\hat{y}$ can be obtained by means of an iterative procedure such as Newton-Raphson. When the iterative procedure has converged, the variance covariance matrix of the parameter estimates can be approximated by the inverse of information matrix, evaluated at $\hat{y}$, that is, $\mathbf{I}^{-1}(\hat{y})$.

The asymptotic normality of the MLE is used for testing the hypothesis $y = y_0$. For large samples $\hat{y}$ has a p-variate normal distribution with mean y and variance-covariance estimated consistently by $\mathbf{I}^{-1}(\hat{y})$. The test statistic is

$$c_W^2 = \left(\hat{y} - y_0\right)' \mathrm{I}\left(\hat{y}\right)\left(\hat{y} - y_0\right) \tag{6.14}$$

which has a limiting chi-squared distribution with p degrees of freedom when $y = y_0$.

Wald test can also be used to test hypothesis about subset of y. Let $\psi = (\psi_1', \psi_2')'$, where y_1 is a $p_1 \times 1$ vector of the y's of interest and y_2 is the vector of the remaining $p_2 = p - p_1$ y's. We wish to test the hypothesis that $y_1 = y_{10}$. To construct the Wald test we partition the information matrix as

$$\mathbf{I} = \begin{pmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{pmatrix} \tag{6.15}$$

where $\mathrm{I}_{11}(\mathrm{I}_{22})$ is the $p_1 \times p_1 (p_2 \times p_2)$ sub matrix of negative second partial derivatives of the log-likelihood with respect to $y_1(y_2)$ and I_{12} and I_{21} the matrices of mixed second partial derivatives. The Wald test is

$$c_W^2 = (\hat{y}_1 - y_{10})' \left(\mathbf{I}^{11}(\hat{y})\right)^{-1} (\hat{y}_1 - y_{10}) \tag{6.16}$$

where $\mathbf{I}^{11}(\hat{y})$ is the upper $p_1 \times p_1$ submatrix of $\mathbf{I}^{-1}(\hat{y})$.

For one-parameter case, the more commonly used form is $\hat{y}_{jk} \sim N\left(y_{jk}, \mathbf{I}^{-1}(y_{jk})\right)$. The $100(1-a)\%$ approximate confidence interval for the unknown y_{jk}-parameters is the interval with limits

$\hat{y}_{jk} \pm z_{a/2} SE(\hat{y}_{jk})$, where $z_{a/2}$ is the upper $a/2$ point of the standard normal distribution.

6.3 Simulation

Monte Carlo simulations were carried out to investigate the performance of the maximum likelihood estimates of model parameters. The following subdistribution models were used to generate data

$$F_j(t; y_j) = 1 - \{\exp[-q_j \exp(\boldsymbol{z}' b_j)]\}^{F_j^*(t; f_j)}, j = 1, \ldots, J$$

Let $P(d_j = 1 \mid \boldsymbol{z}) = \lim_{t \to \infty} F_j(t; y_j) = 1 - \exp[-q_j \exp(\boldsymbol{z}' b_j)]$ is the probability to failure due to cause j and $P(d^* = 0 \mid z)$ is the immune fraction. Because we have constraint $\sum_{j=1}^{J} P(d_j = 1 \mid \boldsymbol{z}) + P(d^* = 0 \mid \boldsymbol{z}) = 1$, then the subdistribution of J-th cause is derived based on the first J–1 subdistributions, where $P(d_J = 1) = 1 - P(d^* = 0 \mid \boldsymbol{z}) - \sum_{j=1}^{J-1} P(d_j = 1)$. Hence the J^{th} subdistribution can be expressed as follows

$$F_J(t; q_1, \boldsymbol{\beta}_1, \ldots, q_{J-1}, \boldsymbol{\beta}_{J-1}, P(d^* = 0 \mid \boldsymbol{z}), f_J) = 1 - \{(J-1) + P(d^* = 0 \mid \boldsymbol{z}) - \sum_{j=1}^{J-1} \exp[-q_j \exp(\boldsymbol{z}' \boldsymbol{\beta}_j)]\}^{F_J^*(t; f_J)}$$

For example, if we use one covariate, two causes of failure, J=2, no immune fraction, $P(d^* = 0 \mid z)$=0 and Weibull kernel, then the first subdistribution is

$$F_1(t;q_1,b_1,k_1,a_1,\boldsymbol{z}) = 1 - \left\{\exp\left[-q_1 \exp(b_1 z)\right]\right\}^{1-\exp(-k_1 t^{a_1})}$$

and the second is

$$F_2(t;q_1,b_1,k_2,a_2,\boldsymbol{z}) = 1 - \left\{1 - \exp\left[-q_1 \exp(b_1 z)\right]\right\}^{1-\exp(-k_2 t^{a_2})}$$

This formulation involves six parameters $y = (q_1, a_1, k_1, b_1, a_2, k_2)$.

Sample of size 200 is generated by the following steps. First, select the cause of failure from set {1,2} randomly and then generate failure time from its conditional distribution given the failure cause. Independent uniform censoring over interval $(0,b_0)$ and $(0,b_1)$, for z=0 and z=1 respectively, was used with the endpoints of these intervals chosen to give censoring percentage (CP) 25% and 50%. Let the true values of parameter q_1=a_1=k_1=b_1=1, and k_2=a_2=2. For dichotomous covariate z, the true subdistribution is given in Figure 6.1.

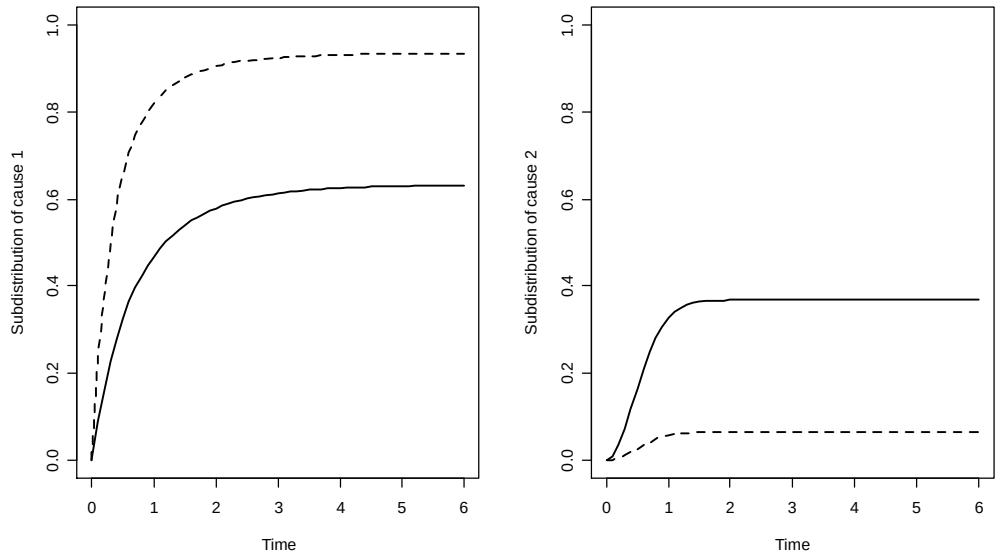

Figure 6.1. The true subdistribution function for 1st cause (left) and 2nd cause (right), z=0 (dashed) and z=1 (solid).

One thousand samples were generated, and the following computations were carried out from the simulation study to asses the estimates.

i. Mean, $\overline{\hat{y}} = \frac{1}{1000}\sum_{i=1}^{1000}\hat{y}_i$

ii. Estimated bias = $\overline{\hat{y}} - y$

iii. Absolute relative estimated bias (%) = $\left(\frac{|\text{Est. bias}|}{y}\right)\times 100\%$

iv. Estimated standard errors = $\sqrt{\frac{1}{1000-1}\sum_{i=1}^{1000}(\hat{y}_i - \overline{\hat{y}})^2}$

v. Estimated root mean square errors (RMSE) = $\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\hat{y}_i - y)^2}$

The results of maximum likelihood estimates for $\hat{y}$ are shown in Table 6.2.

Table 6.2. Simulation result on the efficiency of the parameter estimates

CP	Parameter	True Parameter	Mean	Est. Bias	ARE Bias(%)	Est. SE	Est. RMSE
0%	q_1	1	1.010878	0.010878	1.087814	0.0225	0.1503686
	a_1	1	1.016035	0.016035	1.603524	0.0057	0.0769869
	k_1	1	1.011416	0.011416	1.141581	0.0193	0.1392948
	b_1	1	1.018148	0.018148	1.814797	0.0462	0.2155957
	a_2	2	2.096106	0.096106	4.805285	0.0898	0.3146344
	k_2	2	2.097115	0.097115	4.855728	0.2149	0.4733885
25%	q_1	1	1.024535	0.024535	2.453483	0.0296	0.1738442
	a_1	1	1.018544	0.018544	1.854417	0.0082	0.0921788
	k_1	1	1.013122	0.013122	1.312164	0.0321	0.1795273
	b_1	1	1.012004	0.012004	1.200418	0.0533	0.2311902
	a_2	2	2.105945	0.105945	5.297244	0.1329	0.3794285
	k_2	2	2.220640	0.220640	11.032021	0.5263	0.7579164
50%	q_1	1	1.000385	0.000385	0.038541	0.0851	0.2915301
	a_1	1	1.032530	0.032530	3.252967	0.0166	0.1329781
	k_1	1	1.220879	0.220879	22.087940	0.4326	0.6935007
	b_1	1	1.021363	0.021363	2.136338	0.0695	0.2644442
	a_2	2	2.214819	0.214819	10.740980	0.3411	0.6219948
	k_2	2	3.169036	1.169036	58.451785	18.4589	4.4505180

It appears from Table 6.2 that the estimates obtained by the maximum likelihood method are quite close to the true parameter values, in the sense that they have rather small biases. For instance, for no censoring (*CP*=0%) the absolute relative estimated bias for all estimated parameters was less than 5%. All biases were positive and less than 10%. The estimated standard errors were also small, ranging from 0.0057 to 0.2149. The estimated root mean square (*RMSE*) was also found to be small indicating that the maximum likelihood method has a good performance. However, all were increased as CP increased. Overall, the *MLE* performed very well for the parameter.

6.4 Application to Bone Marrow Transplant (BMT) Data

BMT is a standard treatment for acute leukemia. Recovery following bone marrow transplantation is a complex process. Transplantation can be considered a failure when patient's leukemia returns (relapse) or when he or she dies while in remission (treatment related death) (Klein and Moeschberger, 2003). There are three types of leukemia patients, namely acute lymphoblastic leukemia (ALL), acute myeloctic leukemia low-risk first remission (AML-low), and AML high-risk second remission or untreated first relapse (AML-high). We examined the probabilities for relapse and for death in remission by using univariate model for each type of leukemia patients (ALL, AML-low and AML-high) and regression model with two indicator variables, z_1 and z_2, where z_1=1 for AML-low group, z_1=0 otherwise and z_2=1 for AML-high group, z_2=0 otherwise.

6.4.1 Univariate Models for Leukemia Patients

We considered subdistribution model (6.2) with five kernel distributions, $F_j^*(t)$, as previously discussed in subsection 6.1.1, namely exponential, Weibull, Gompertz, gamma and generalized gamma.

Exponential

The subdistribution model with Exponential kernel is given in (6.3), where $j = 1,2$ whereby

$$j = \begin{cases} 1, & \text{for failure due to relapse} \\ 2, & \text{for failure due to death} \end{cases} \tag{6.17}$$

The model is fitted to the three groups of patient in the BMT data. The resulted estimated model for patient group ALL, AML-low and AML-high gave the log-likelihood = –197.307, –229.571 and –261.082 or AIC = 402.614, 467.142 and 530.164, respectively. All estimated subdistribution curves along with their nonparametric estimates are presented in Figure 6.2. It's clear that AML-high patients has the highest incidence of relapse, followed by ALL and AML-low patients. However, we cannot distinguish the incidence of death for the three groups of patient.

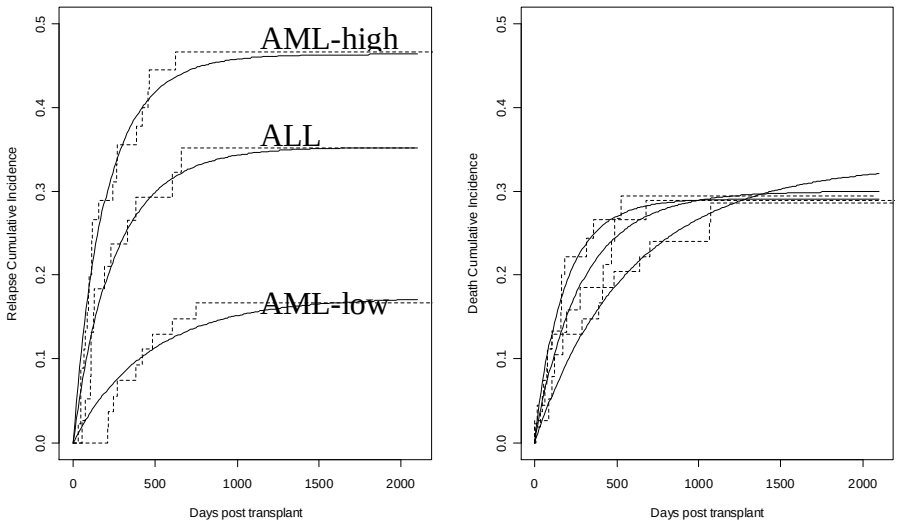

Figure 6.2 The estimated subdistribution curve with Exponential kernel for relapse (left) and death (right).

Weibull

The fitting of subdistribution model with Weibull kernel (6.4) to the three groups of patient (ALL, AML-low and AML-high) resulted in estimated models with log-likelihood = –196.175, –224.568 and –259.886 or AIC = 404.35, 461.136 and 531.772, respectively. Even though all the log-likelihood values were greater than log-likelihood of exponential kernel, but not all of their AICs were less than AIC of Exponential kernel. Only AML-low group has smaller AIC. The other two groups have greater AIC, which means that Weibull kernel are less suitable for ALL and AML-high groups. The estimated subdistribution curves are presented in Figure 6.3. Again we cannot distinguish the incidence of death for the three groups of patient.

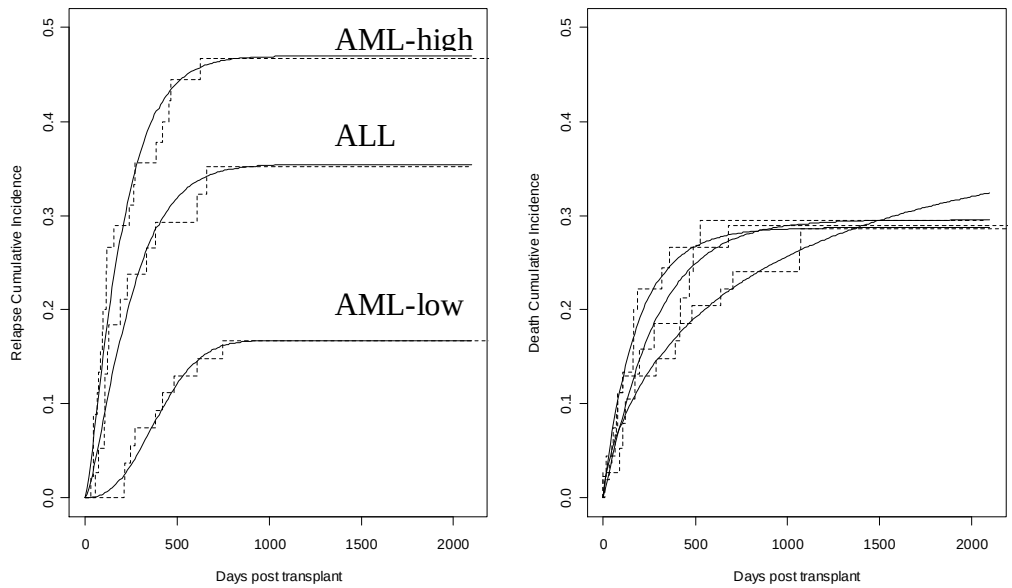

Figure 6.3. The estimated subdistribution curve with Weibull kernel for relapse (left) and death (right).

Gompertz

Figure 6.4 shows the estimated subdistribution curves resulted from fitting Gompertz kernel to the three groups of patient in BMT data. The log-likelihood values were –195.444, –226.031 and –260.294 or AIC = 402.888, 464.062 and 532.588 for ALL, AML-low and AML-high patients, respectively. In terms of AIC, all the three groups are less suitable modelled by Gompertz kernel compared to Exponential (ALL and AML-high patient groups) and Weibull kernel (AML-low patient group).

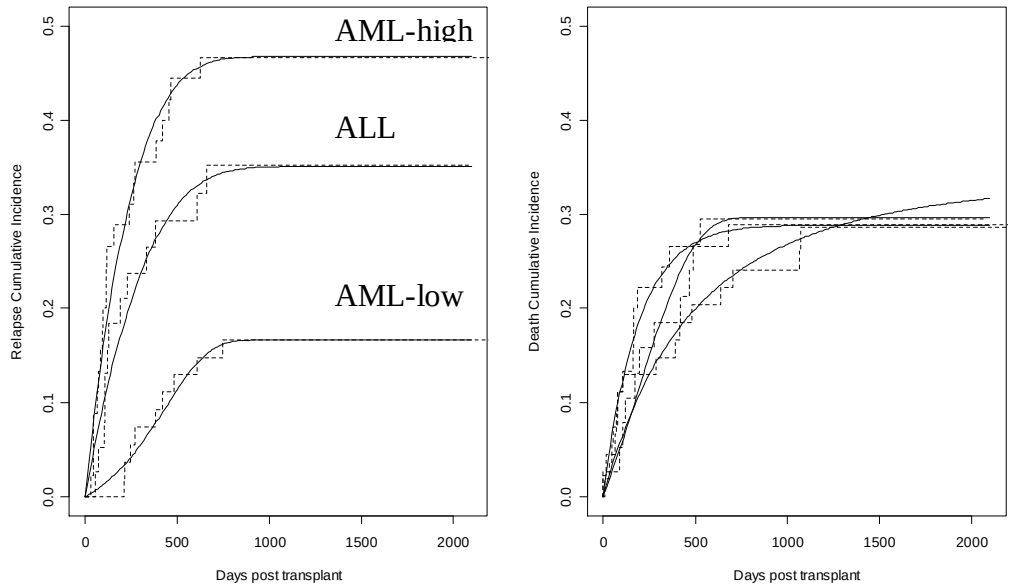

Figure 6.4. The estimated subdistribution curve with Gompertz kernel for relapse (left) and death (right).

Fitting of the rest two kernel distributions, Gamma and Generalized Gamma, are not well suited particularly for AML-low group of patients. For this group of patients, the AIC statistic was larger than the others. We summarized all the log-likelihoods and

AIC statistics in Table 6.3 as well as their estimated curves in Figure 6.5 (Gamma kernel) and Figure 6.6 (Generalized Gamma kernel).

Table 6.3. Summary of the fitting results

	Log-lik			AIC		
	ALL	AML-l	AML-h	ALL	AML-l	AML-h
Exponen	–197.307	–229.571	–261.082	402.614	467.142	530.164
Weibull	–196.175	–224.568	–259.886	404.350	461.136	531.772
Gompertz	–195.444	–226.031	–260.294	402.888	464.062	532.588
Gamma	–196.919	–1194.543	–298.905	405.838	2401.086	609.810
GGamma	–196.014	–817.031	–259.582	408.028	1650.062	535.164

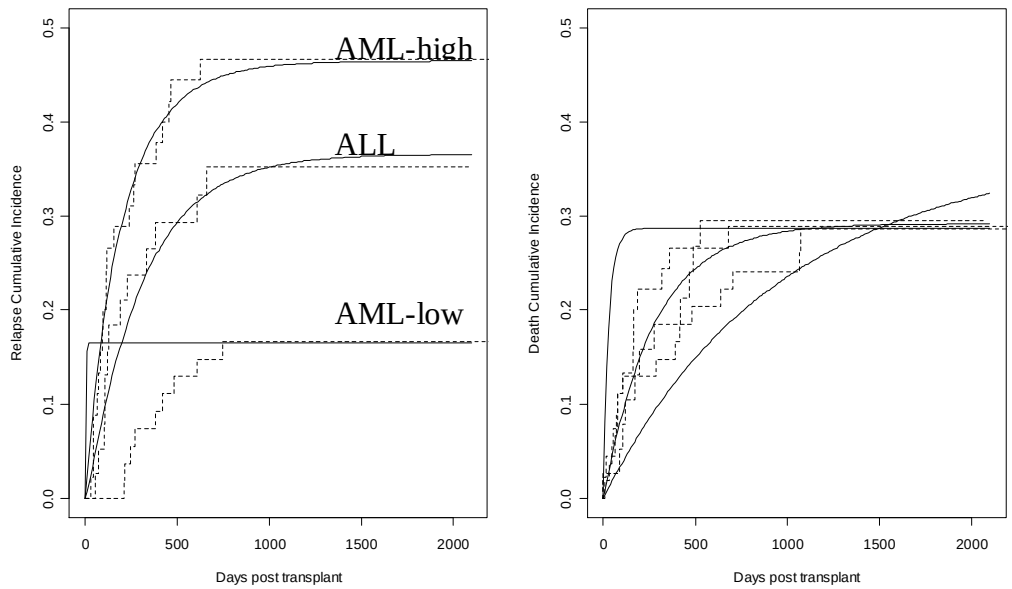

Figure 6.5. The estimated subdistribution curve with Gamma kernel for relapse (left) and death (right).

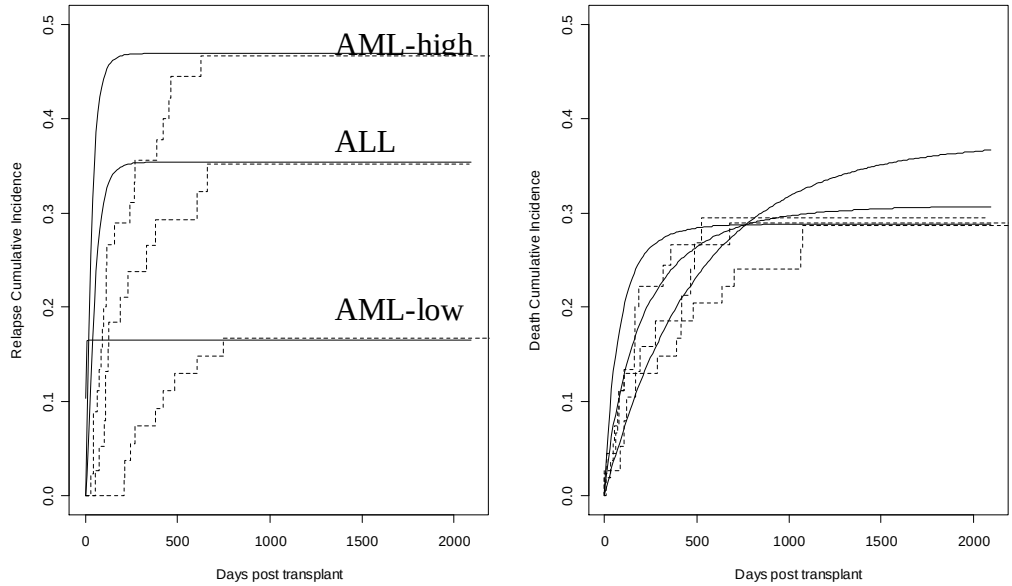

Figure 6.6 The estimated subdistribution curve with Generalized Gamma kernel for relapse (left) and death (right).

6.4.2 Regression Models for Leukemia Patients

We considered the model

$$F_j(t;q_j,b_{j1},b_{j2},f_j)=1-\left\{\exp\left[-q_j\exp(b_{j1}z_1+b_{j2}z_2)\right]\right\}^{F_j^*(t;f_j)},\ j=1,2$$

with the kernel $F_j^*(t;f_j)$ to be exponential, Weibull, Gompertz, gamma and generalized gamma.

The proposed models are fitted to BMT data. The SAS® NLP procedure with trust region optimization method was used as nonlinear optimization subroutine to address the maximum likelihood estimation (SAS, 2004). The result is presented in Table

Table 6.4. Parameter estimates (standard errors) for the BMT data

Cause	Kernel	Kernel parameter estimates			$\hat{q}$	$\hat{b}_1$	$\hat{b}_2$	AIC
c.o.f 1	Exponential	$\hat{k}$ =0.003(0.001)			0.397(0.115)	-0.798(0.440)	0.463(0.362)	1401.293
Relapse	Weibull	$\hat{k}$ =0.0003(0.0003)	$\hat{a}$ =1.393(0.167)		0.402(0.116)	-0.820(0.440)	0.478(0.362)	1396.797*
	Gompertz	$\hat{r}$ =0.002(0.001)	$\hat{t}$ =0.002(0.000)		0.401(0.116)	-0.816(0.440)	0.473(0.362)	1398.482
	Gamma	$\hat{k}$ =0.001(0.0000003)	$\hat{g}$ =0.255(0.038)		0.209(0.032)	-0.797(0.001)	0.522(0.002)	1573.079
	Gen. Gamma	$\hat{k}$ =0.084(0.005)	$\hat{g}$ =0.203(0.031)	$\hat{a}$ =0.251(.)	0.185(0.034)	-0.804(0.596)	0.511(0.381)	1744.596
c.o.f 2	Exponential	$\hat{k}$ =0.002(0.000)			0.409(0.119)	-0.126(0.381)	-0.122(0.400)	
Death	Weibull	$\hat{k}$ =0.007(0.005)	$\hat{a}$ =0.819(0.120)		0.416(0.124)	-0.117(0.382)	-0.147(0.402)	
	Gompertz	$\hat{r}$ =-0.001(0.001)	$\hat{t}$ =0.003(0.001)		0.435(0.148)	-0.114(0.382)	-0.136(0.402)	
	Gamma	$\hat{k}$ =0.004(0.00004)	$\hat{g}$ =0.253(0.040)		0.209(0.032)	-0.105(0.001)	-0.082(0.001)	
	Gen. Gamma	$\hat{k}$ =0.016(0.000)	$\hat{g}$ =0.196(0.030)	$\hat{a}$ =0.402(.)	0.186(0.033)	-0.105(0.437)	-0.083(0.549)	

c.o.f: cause of failure
*the minimum AIC

6.4. Standard error of the maximum likelihood estimates was obtained by Normal asymptotic approach, whereas the variance-covariance matrix was an inverse of Hessian matrix. However, the parameters estimate for generalized gamma kernel distribution are unrealistic. Two of the parameter estimates of generalized gamma were not available for the standard error, and it also gives a warning: *Optimization cannot be completed*. The same problem was reported in Koti (2004).

Weibull kernel shows the minimum AIC statistic, 1396.796. Figure 6.7 shows comparative plot of the nonparametric subdistribution using Satagopan (2004) method and parametric subdistribution using Weibull kernel estimates for three groups of leukemia patients. Parametric curves agree reasonably well with nonparametric, although there is some evidence of lack of fit in the first few days for relapse particularly for AML-low group. *P*-value for effect of dummy variable AML-low to probability of time to relapse was 0.064. The negative sign of $\hat{b}_{11}$ indicated that the relapse subdistribution for AML-low group was lower than the others. There was no difference in the subdistribution death of the three groups.

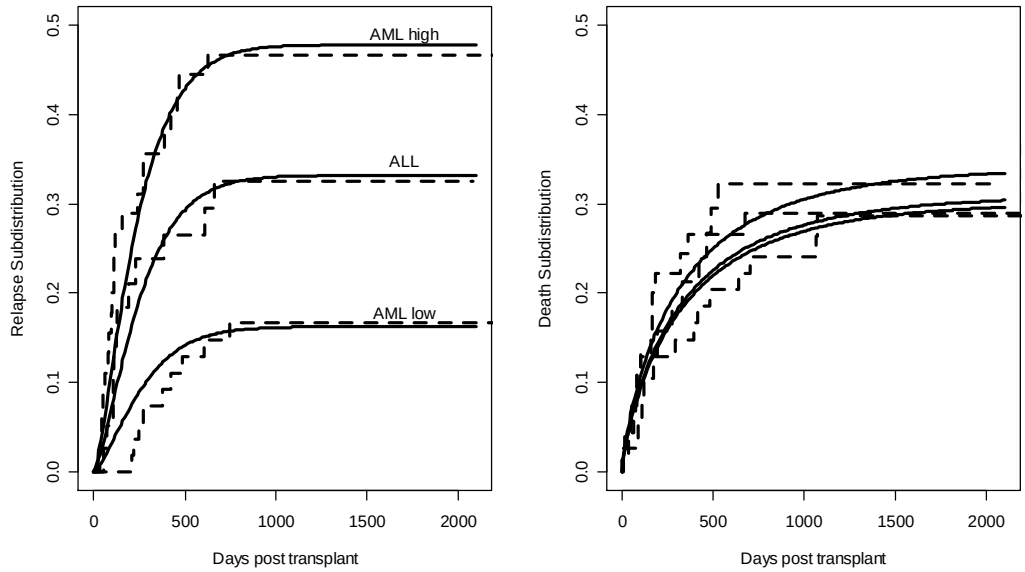

Figure 6.7. Estimated subdistribution functions for relapse (left) and death (right) using nonparametric (dashed) and parametric with Weibull kernel (solid).

6.5. Parametric Gompertz-like Subdistribution

Sometimes the event of interest occurred at a fairly steady rate over the entire time period of study. This makes the subdistribution function seems to behave like proper distribution with asymptote at 1. Thus, the improper subdistribution might not be suitable for modelling such kind of phenomena. Figure 6.8 shows the illustration of the cause of interest which occurred at a fairly steady rate over the first 50 days. Suppose that the period of study is less than 30 days, then most of the competing cause occurred during the first 10 days and then plateau off even though a few events were still occurring towards the end of the follow-up period. However, the cause of interest has not plateau off up to 30 days of

follow-up period. We can use proper subdistribution model for cause of interest and improper subdistribution for competing cause for modelling these phenomena.

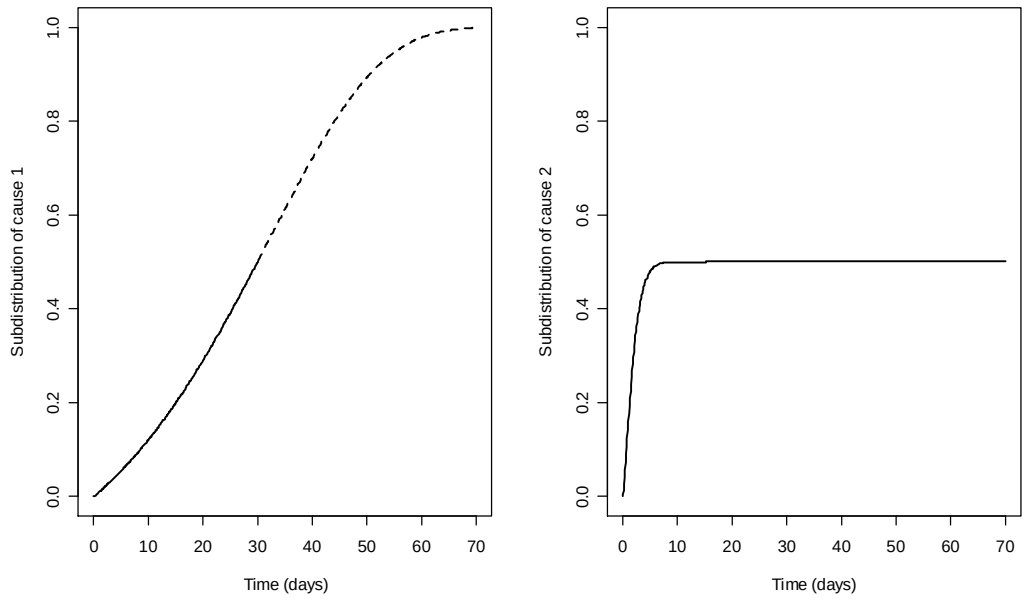

Figure 6.8. Illustration of proper subdistribution for cause of interest (left) and improper subdistribution for competing cause (right).

6.5.1. Univariate Gompertz-like Subdistribution Model

Let us revisit Gompertz distribution,

$$F(t) = 1 - \exp\{t[1 - \exp(rt)]/r\} \tag{6.18}$$

One of the nice properties of Gompertz distribution is the asymptote of the cumulative distribution function which may be less than 1. If $r \geq 0$, then asymptote for large t of equation (6.18) is 1 which shows that it is a proper distribution. Whereas, an improper case of equation (6.18) occurs when $r < 0$. Next, we will

develop some subdistribution functions which may be proper or improper depend on the sign of parameter. Such kind of subdistribution will be called as Gompertz-like subdistribution.

Gompertz-like subdistribution with exponential kernel

Given a subdistribution for j^{th} cause with exponential kernel as expressed in (6.3), and after reparameterization of

$$q_j^* = -\frac{t_j}{r_j} \quad \text{and} \quad k_j = -r_j,$$

the resulted subdistribution is

$$F_j(t; t_j, r_j) = 1 - \exp\{t_j[1 - \exp(r_j t)]/r_j\} \tag{6.19}$$

It is clear that (6.19) is the usual Gompertz subdistribution which will be proper when $r \geq 0$ and improper when $r < 0$ with the plateau equals to $1 - \exp(t_j / r_j)$.

Gompertz-like subdistribution with Weibull kernel

Reparameterization for subdistribution with Weibull kernel (6.4) is carried out in the same manner with

$$q_j^* = -\frac{t_j}{r_j} \quad \text{and} \quad k_j = -r_j$$

Thus, the resulted Gompertz-like subdistribution is

$$F_j\left(t;t_j,r_j,a_j\right)=1-\exp\left\{t_j\left[1-\exp\left(r_j t^{a_j}\right)\right]/\,r_j\right\} \tag{6.20}$$

Again, its property depends on the sign of r_j. For improper case, its plateau is also equals to $1-\exp\left(t_j/r_j\right)$.

Gompertz-like subdistribution with Gompertz kernel

Let us consider the other kind of Gompertz distribution,

$$F(t)=1-\exp\left[-q\left(1-e^{-kt}\right)\right] \tag{6.21}$$

by developing the subdistribution function based on non-mixture model with Gompertz kernel (6.21), we have

$$F_j\left(t;q_j^*,q_j,k_j\right)=1-\left[\exp\left(-q_j^*\right)\right]^{1-\exp\left[-q_j\left(1-e^{-kjt}\right)\right]} \tag{6.22}$$

The reparameterization of (6.22) with

$$q_j^*=-\frac{t_j}{r_j} \quad \text{and} \quad k_j=-r_j$$

gives

$$F_j\left(t;t_j,r_j,q_j\right)=1-\exp\left\{t_j\left[1-\exp\left(-q_j\left(1-e^{r_j t}\right)\right)\right]/\,r_j\right\} \tag{6.23}$$

Thus, (6.23) has the form of Gompertz-like subdistribution which may be proper or improper depends on the sign of r. For $r < 0$, the asymptote of (6.23) will be equal to $1-\exp\left[t_j\left(1-e^{-q_j}\right)/\,r_j\right]$.

6.5.2. Parametric Regression with Gompertz-like Subdistribution Model

We can incorporate covariates into cure parameter of Gompert-like subdistribution model. The utilization of exponential kernel resulted in parametric regression,

$$F_j(t;t_j, r_j, b_j, \boldsymbol{z}) = 1 - \exp\{\exp(\boldsymbol{z}' b_j)t_j[1 - \exp(r_j t)]/ r_j\} \quad (6.24)$$

and for Weibull kernel it is

$$F_j(t;t_j, r_j, a_j, b_j, \boldsymbol{z}) = 1 - \exp\{\exp(\boldsymbol{z}' b_j)t_j[1 - \exp(r_j t^{a_j})]/ r_j\} \quad (6.25)$$

and also by using Gompertz kernel, we have

$$F_j(t;t_j, r_j, q_j, b_j, \boldsymbol{z}) = 1 - \exp\{\exp(\boldsymbol{z}' b_j)t_j[1 - \exp(-q_j(1 - e^{r_j t}))]/ r_j\} \quad (6.26)$$

Given competing risks data, we can fit univariate Gompertz-like subdistributions (6.19), (6.20) and (6.23) as well as their regression model counterparts (6.24) – (6.26) to the data. Maximum likelihood estimation can be used for these purposes. To do so, we have derived the likelihood function as shown in Appendix E.

6.6 Application to Contraceptive Discontinuation Data

We revisited the contraceptive discontinuation data and we fitted Gompertz-like subdistribution to data. As previously discussed,

the data was length of time to contraceptive discontinuation which was categorized into three causes of discontinuation namely failure, abandonment and switching. First part was on fitting Gompertz-like distribution to data without covariates. The second part, we incorporated covariates to the model and regression analysis was carried out.

Univariate analysis

Given competing risks sample data (t_i,d_i), $i = 1,\ldots,n$, where $d_i \in \{0,1,2,3\}$, or $(t_i,d_{1i},d_{2i},d_{3i},)$, where $d_{ji} \in \{0,1\}$, $j = 1,2,3$, we fitted Gompertz-like subdistribution with kernel of exponential, Weibull and Gompertz as expressed in (6.19), (6.20) and (6.23).

Exponential kernel

Gompertz-like subdistribution with exponential kernel (6.19) was fitted to contraceptive discontinuation data and the result is shown in Table 6.5. We have to estimate all parameters simultaneously, because the likelihood function (6.10) cannot be factored into a product of 3 cause-specific functions. Therefore, we only obtained

one log-likelihood value ($l(y)$ = -10407.68753). Negative sign of r_3 and its small *P*-value showed that the subdistribution of time to discontinuation due to switching (3rd cause) has the improper form. Figure 6.9 displayed the estimated curve along with its nonparametric counterpart and the fitting agrees reasonably well with each other.

Table 6.5. Result of fitting Gompertz-like subdistribution with exponential kernel to contraceptive discontinuation data

Cause of failure	Parameter	Estimate	Std. error	*P*-value
1	t_1	0.000622	0.000130	1.75×10^{-6}
	r_1	–0.00660	0.005867	2.60×10^{-1}
2	t_2	0.005691	0.000349	6.75×10^{-57}
	r_2	0.012364	0.001476	8.54×10^{-17}
3	t_3	0.011585	0.000616	3.68×10^{-74}
	r_3	–0.00996	0.001652	1.88×10^{-9}
	$l(y)$	–10407.68753		
	AIC	20827.38		

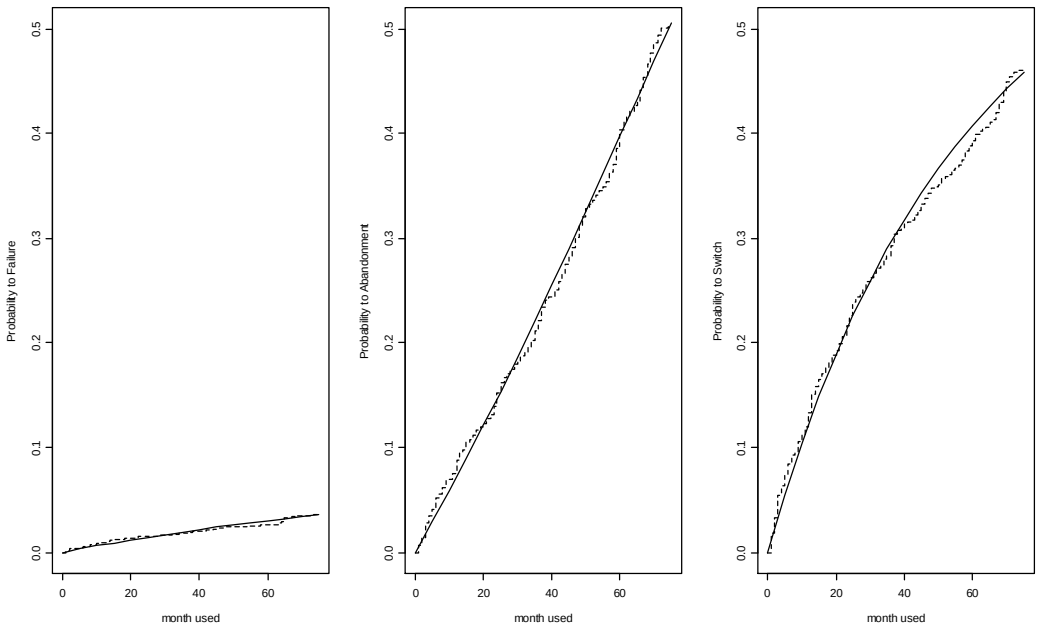

Figure 6.9. Curve fitting of Gompertz-like subdistribution with exponential kernel to contraceptive discontinuation data.

Weibull kernel

Fitting model (6.20) to data gave the likelihood function –10404.72578 (see Table 6.6). This value is less than exponential kernel likelihood. However, its AIC is slightly greater which means the exponential kernel fits better. The result is also similar for negative sign of r_3 which showed improper subdistribution function of the 3rd cause. Plot of estimated curve of subdistribution using nonparametric and parametric with Weibull kernel is presented in Figure 6.10. The fitting of both curves are quite well.

Table 6.6. Result of fitting Gompertz-like subdistribution with Weibull kernel to contraceptive discontinuation data

Cause of failure	Parameter	Estimate	Std. error	*P*-value
1	t_1	0.001105	0.000474	1.98×10^{-2}
	r_1	0.008007	0.02986	7.89×10^{-1}
	a_1	0.784187	0.157194	6.48×10^{-7}
2	t_2	0.00667	0.001099	1.48×10^{-9}
	r_2	0.019747	0.009538	3.85×10^{-2}
	a_2	0.936311	0.064797	1.33×10^{-45}
3	t_3	0.013998	0.001688	1.74×10^{-16}
	r_3	–0.00875	0.002589	7.40×10^{-4}
	a_3	0.928885	0.04199	1.37×10^{-99}
	$l(y)$	–10404.72578		
	AIC	20827.45		

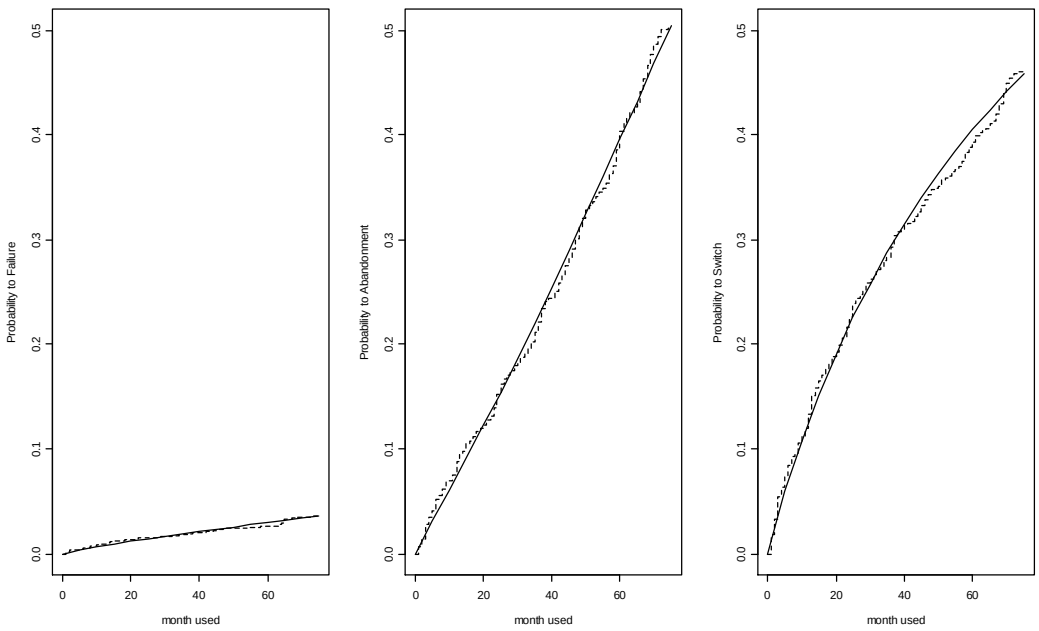

Figure 6.10. Curve fitting of Gompertz-like subdistribution with Weibull kernel to contraceptive discontinuation data.

Gompertz kernel

We obtained the likelihood function –10407.62246 when we fitted parametric Gompertz-like subdistribution with Gompertz kernel (6.23) to contraceptive data (Table 6.7). This value corresponds to 20833.24 of AIC value. This value is bigger than the previous two AIC values, which means Gompertz kernel did not fit the data well. Again, r_3 has negative sign which showed improper subdistribution for the 3rd cause and this agrees with estimated curve shown in Figure 6.11.

Table 6.7. Result of fitting Gompertz-like subdistribution with Gompertz kernel to contraceptive discontinuation data

Cause of failure	Parameter	Estimate	Std. error	P-value
1	t_1	0.034756	0.040212	3.87×10^{-1}
	r_1	–0.00653	0.005806	2.61×10^{-1}
	q_1	0.017901	0.020728	3.88×10^{-1}
2	t_2	0.664669	0.252604	8.56×10^{-3}
	r_2	0.012209	0.001451	6.50×10^{-17}
	q_2	0.008567	0.003257	8.57×10^{-3}
3	t_3	0.57312	0.223136	1.03×10^{-2}
	r_3	–0.00983	0.001635	2.04×10^{-9}
	q_3	0.020234	0.007884	1.03×10^{-2}
	$l(y)$	–10407.62246		
	AIC	20833.24		

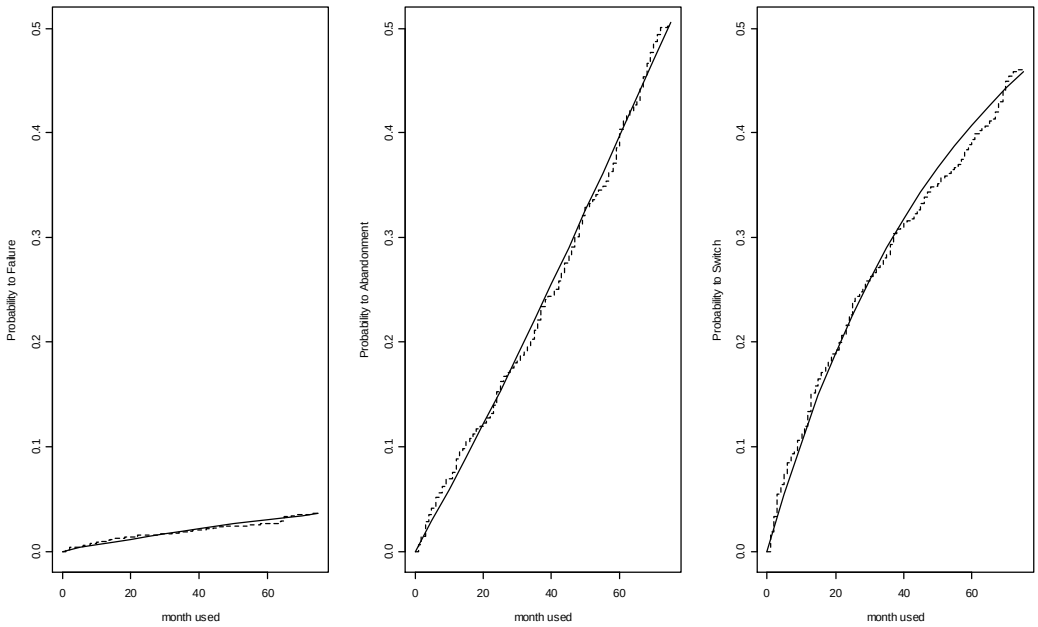

Figure 6.11. Curve fitting of Gompertz-like subdistribution with Gompertz kernel to contraceptive discontinuation data.

Regression Analysis

We modified univariate models to take into account the covariates effect to probability of discontinuation. The resulted regression models are expressed in (6.24)-(6.26) and we fitted those three models to contraceptive discontinuation data. The covariates had been defined in section 4.3.

Exponential kernel

The result of model fitting with exponential kernel is shown in Table 6.8. This fitting gives likelihood function as –10416.57532

which is equals to AIC value 20893.15. By using this model, we found that probability of discontinuation due to failure is affected by Religion covariate. This result is different with semiparametric modelling of Fine and Gray which was presented in Table 4.1(a). For the discontinuation due to abandonment, Age and University education level are two covariates which are significant. This result agreed with semiparametric regression of Fine and Gray (Table 4.1(b)), even though secondary education level and IUD/Implant method are not significant. Whereas, for the discontinuation due to switching beside education and method covariates there are two others covariates which are significant namely social economic status and age.

Weibull kernel

Table 6.9 gives the result of fitting model (6.25) to the data. Probability to failure is affected by age. Social economic status, age and education factor are the significant factors for the probability to abandonment. The significant factors for probability to switching are education factor and contraception method.

Table 6.8. Regression of contraceptive discontinuation using Gompertz-like subdistribution with exponential kernel

Cause of failure	Parameter	Estimate	Std. error	P-value
1	t_1	0.00168	0.001672	3.25×10^{-1}
	r_1	0.005914	0.006681	3.76×10^{-1}
	Social Economic Status	-0.06044	0.082202	4.62×10^{-1}
	Age	-0.04112	0.021564	5.67×10^{-2}
	Residence	0.180885	0.271054	5.05×10^{-1}
	Religion	1.475557	0.657311	2.49×10^{-2}
	Secondary	-0.29854	0.377395	4.29×10^{-1}
	University	0.524297	0.438745	2.32×10^{-1}
	Pills/Injection	0.04691	0.741995	9.50×10^{-1}
	IUDs/Implants	-1.63706	0.913911	7.34×10^{-2}
2	t_2	0.002379	0.000708	7.92×10^{-4}
	r_2	0.023856	0.001732	9.79×10^{-42}
	Social Economic Status	0.039698	0.021039	5.93×10^{-2}
	Age	0.020537	0.004715	1.38×10^{-5}
	Residence	0.014608	0.067415	8.28×10^{-1}
	Religion	0.062844	0.177222	7.23×10^{-1}
	Secondary	-0.15573	0.087524	7.53×10^{-2}
	University	-0.26154	0.114962	2.30×10^{-2}
	Pills/Injection	0.094764	0.238678	6.91×10^{-1}
	IUDs/Implants	-0.0178	0.247231	9.43×10^{-1}
3	t_3	0.012674	0.003193	7.40×10^{-5}
	r_3	-0.00141	0.001817	4.37×10^{-1}
	Social Economic Status	-0.05281	0.021714	1.51×10^{-2}
	Age	0.011296	0.005114	2.73×10^{-2}
	Residence	0.091762	0.068808	1.82×10^{-1}
	Religion	0.055785	0.180474	7.57×10^{-1}
	Secondary	0.302452	0.09684	1.81×10^{-3}
	University	0.670814	0.11812	1.50×10^{-8}
	Pills/Injection	-0.78003	0.17355	7.27×10^{-6}
	IUDs/Implants	-0.9433	0.185076	3.70×10^{-7}
	$l(y)$	-10416.57532		
	AIC	20893.15		

Table 6.9. Regression of contraceptive discontinuation using Gompertz-like subdistribution with Weibull kernel

Cause of failure	Parameter	Estimate	Std. error	P-value
1	t_1	0.002548	0.003383	4.51×10^{-1}
	r_1	0.044685	0.062208	4.73×10^{-1}
	a_1	0.734406	0.163079	6.98×10^{-6}
	Social Economic Status	-0.07029	0.088488	4.27×10^{-1}
	Age	-0.06364	0.024144	8.44×10^{-3}
	Residence	0.370969	0.264681	1.61×10^{-1}
	Religion	1.036631	0.873039	2.35×10^{-1}
	Secondary	-0.50087	0.3464	1.48×10^{-1}
	University	0.283585	0.407988	4.87×10^{-1}
	Pills/Injection	1.05101	1.039881	3.12×10^{-1}
	IUDs/Implants	-0.17939	1.072255	8.67×10^{-1}
2	t_2	0.00291	0.001025	4.54×10^{-3}
	r_2	0.045624	0.017866	1.07×10^{-2}
	a_2	0.869848	0.063216	1.19×10^{-41}
	Social Economic Status	0.049111	0.022415	2.85×10^{-2}
	Age	0.023565	0.004973	2.27×10^{-6}
	Residence	0.008399	0.066083	8.99×10^{-1}
	Religion	0.088932	0.207938	6.69×10^{-1}
	Secondary	-0.18946	0.086948	2.94×10^{-2}
	University	-0.44184	0.11507	1.26×10^{-4}
	Pills/Injection	0.20711	0.271428	4.46×10^{-1}
	IUDs/Implants	0.145378	0.274732	5.97×10^{-1}
3	t_3	0.020734	0.006064	6.38×10^{-4}
	r_3	0.00446	0.005769	4.40×10^{-1}
	a_3	0.87279	0.043937	6.20×10^{-82}
	Social Economic Status	-0.04349	0.022316	5.14×10^{-2}
	Age	0.007864	0.005566	1.58×10^{-1}
	Residence	0.093903	0.068188	1.69×10^{-1}
	Religion	-0.34902	0.192872	7.05×10^{-2}
	Secondary	0.255511	0.097517	8.84×10^{-3}
	University	0.504893	0.117324	1.74×10^{-5}
	Pills/Injection	-0.78513	0.194889	5.77×10^{-5}
	IUDs/Implants	-0.81459	0.20298	6.16×10^{-5}
	$l(y)$	-10372.40251		
	AIC	20810.81		

Table 6.10. Regression of contraceptive discontinuation using Gompertz-like subdistribution with Gompertz kernel

Cause of failure	Parameter	Estimate	Std. error	P-value
1	t_1	0.001791	0.034835	9.59×10^{-1}
	r_1	0.000886	0.009151	9.23×10^{-1}
	q_1	0.968525	18.80398	9.59×10^{-1}
	Social Economic Status	0.002955	0.084077	9.72×10^{-1}
	Age	-0.06343	0.024072	8.46×10^{-3}
	Residence	0.247677	0.258665	3.38×10^{-1}
	Religion	1.007604	0.790109	2.02×10^{-1}
	Secondary	-0.59107	0.345069	8.68×10^{-2}
	University	0.066024	0.378178	8.61×10^{-1}
	Pills/Injection	0.76728	0.89264	3.90×10^{-1}
	IUDs/Implants	-0.5395	0.926109	5.60×10^{-1}
2	t_1	0.004146	0.003242	2.01×10^{-1}
	r_1	0.010952	0.003366	1.15×10^{-3}
	q_1	0.657719	0.475613	1.67×10^{-1}
	Social Economic Status	0.016557	0.020575	4.21×10^{-1}
	Age	0.019896	0.005148	1.14×10^{-4}
	Residence	0.049314	0.064131	4.42×10^{-1}
	Religion	-0.0927	0.188302	6.23×10^{-1}
	Secondary	-0.13897	0.0865	1.08×10^{-1}
	University	-0.3303	0.10773	2.19×10^{-3}
	Pills/Injection	0.083262	0.249314	7.38×10^{-1}
	IUDs/Implants	0.051049	0.252617	8.40×10^{-1}
3	t_1	0.014218	0.038428	7.11×10^{-1}
	r_1	-0.0024	0.003495	4.92×10^{-1}
	q_1	1.092092	2.93768	7.10×10^{-1}
	Social Economic Status	-0.03737	0.021181	7.78×10^{-2}
	Age	0.00717	0.005715	2.10×10^{-1}
	Residence	0.084217	0.06715	2.10×10^{-1}
	Religion	-0.38757	0.182289	3.36×10^{-2}
	Secondary	0.259237	0.098914	8.82×10^{-3}
	University	0.519373	0.113352	4.82×10^{-6}
	Pills/Injection	-0.80098	0.180891	9.9×10^{-6}
	IUDs/Implants	-0.83497	0.190142	1.17×10^{-5}
	$l(y)$	-10361.05516		
	AIC	20788.11		

Gompertz kernel

Among the three Gompertz-like subdistribution functions, Gompertz kernel has the best fit. It has the largest log-likelihood

and also the minimum AIC (Table 6.10). Only age which is significant to the probability of failure. Again, age and university education level are two covariates which are significant to the probability to abandonment. For the discontinuation due to switching, we observed that education, contraception method and religion are the significant factors.

Since Gompertz kernel has the best fit, then simple conclusion can be drawn based on this result as follows: 1) the older women tend to have less chance to failure, but more to abandoning their current contraception method, 2) university educated women have less probability to abandon their contraceptive methods, but more to switching their methods and this happens for secondary educated women too, 3) Moslem women have more probability to switching, where religion = 0 for Moslem and religion = 1 otherwise, 4) other modern contraception method users tend more to switching compared to pill/injection and IUD/implant user.

6.7. Summary

We have demonstrated the use of standard parametric survival models in case of competing risks by modeling subdistribution functions through non-mixture cure model. Any standard parametric survival distributions like exponential, Weibull,

gamma and generalized gamma can be employed for kernel of the non-mixture cure model which can then be used to specify the improper or proper subdistribution functions.

The study also reveals that subdistribution function which constructed by using non-mixture cure model with Weibull kernel distribution is a generalization of Jeong and Fine (2007), the recent parametric modelling of the subdistribution function. Inferential procedure by using maximum likelihood estimation indicates that likelihood function cannot be factored into a product of cause-specific functions. So, we fitted all subdistribution jointly. In BMT data analysis, the proposed model gave noticeably good fit to the nonparametric counterpart, except for the first few days of AML-low group of patients.

Another issue is regarding the event which occurs at a fairly steady rate over the entire time period. This kind of subdistribution is better described by a proper distribution. We have proposed Gompertz-like subdistribution to solve this problem. However, we cannot extrapolate the subdistribution model to estimate the long term probabilities, because its subdistribution value will reach one as time goes to infinity.

We can incorporate covariates through parameter other than cure fraction. It may increase the fitting, but we may lose the linear form of the complementary log-log transformation.

CHAPTER 7

SUMMARY, GENERAL CONCLUSION AND RECOMMENDATION FOR FUTURE RESEARCH

7.1 Summary

The problem of summarizing and making inference about competing risks quantities is important in many areas of application. In studies that have multiple endpoints, each subject may fail due to one of several possible causes. Failure rates from every cause are often related and their complex action affects overall survival time of an individual. In this work, we have presented the main results in modelling competing risks data.

Subdistribution function is the important quantity in summarizing and describing competing risks data. Subdistribution curve estimates the chance of ultimately experiencing a particular cause.

The focus of this thesis was on regression methods for subdistribution of competing risks. Traditionally, this has been done by standard statistical methods based on the cause specific hazard rate that treat failures from causes other than the cause of interest as censored observations. This includes such technique as the Cox proportional hazard model.

As was demonstrated earlier, differences in cause specific hazard rates for a particular risk do not translate directly into differences between subdistribution curves since these curves depend on all the competing risks cause specific hazard rates. In this thesis, we developed various regression methods for subdistribution of competing risks.

Chapter 3 developed the optimal cutpoint determination method for continuous predictor variable in competing risk survival data analysis through direct modelling of subdistribution function. Five methods were considered namely the two-sample, Wald, likelihood ratio (deviance), delta deviance, and delta null deviance. To compare those five methods and conclude the optimal cutpoints, a Monte Carlo experiment was conducted and five statistical indicators were used, which were mean, bias, absolute relative bias, standard error and root mean square error. All of these criteria are very important in measuring the validity of an estimate, and have been used in a variety of statistical considerations. The deviance method was recommended, since the estimated cutpoints from this method had the smallest overall rank sums which represent the small quantities of those four statistical indicators (bias, absolute relative bias, standard error and root mean square error). We illustrated the method by using contraceptive discontinuation data. The optimal cutpoint for age were 34.167,

38 and 38 years for the time to the occurrence of contraceptive failure, abandonment and switching, respectively.

In general, this cutpoint methodology study is very important in the health science field. Implications for health policy attention are obvious. Simple, but yet accurate, guidelines for people will allow for easy implementation.

Tree methods provide an excellent way of exploring data. Brieman *et al.* (1984) have developed the powerful CART algorithm to obtain trees of optimal sizes, though what really constitutes an optimal tree remains unsolved. In the spirit of CART, we developed a method for competing risks survival trees.

The development of a tree-structured methodology for the analysis of competing risk data based on deviance statistic is the goal of Chapter 4. Our modelling is based on the subdistribution hazard, since it has closed relation with subdistribution function which is the important properties of competing risk data. For the our method based on deviance statistic derived from likelihood ratio test to examine the effect of one threshold covariate in subdistribution proportional hazard setting, we have adopted the pruning strategy proposed by Segal (1988).

The method works well for the contraceptive discontinuation data. We developed the stratification of group of women in term of their risks to each discontinuation type (failure, abandoning and switching). The method also offers the tool for exploring the competing risk data thoroughly. Plots of subdistribution estimates are useful description of the subjects within nodes of the tree.

The proposed method intends to provide an exploratory data analysis for competing risks data, and it is complimentary rather than competitive to those parametric or semi-parametric methods.

We evaluated the method by several simulation studies. Simulation results showed that the proposed method performed well for group identifications, but the capability of identification decreased as percentage of censoring increased.

In Chapter 5, we augment the best AIC subdistribution hazard regression model with a tree-structured regression counterpart. This method is to boost the model. Tree-structured regression amends deficiencies in the subdistribution hazard, and extends the application of trees. This hybrid model gives a better insight of the data.

The application of the proposed method to contraceptive discontinuation data showed that for the risk to discontinuation due to failure the hybrid model did not constitute a substantial improvement over best AIC model. However, the other two hybrid models (abandonment and switching risk) have lower AIC compared to the corresponding best subdistribution hazard regression models. For the two risks, the augmented tree can boost the model.

Sometimes parametric model has more advantages compared to nonparametric and semiparametric model. It gave the efficient estimation when the model fits data well. For this purpose, subdistribution modelling through parametric cure model with covariates has been developed and presented in Chapter 6. Some of the well known kernel distributions were utilized. Weibull kernel distribution is a generalization of the recent parametric modelling of the subdistribution function. Inferential procedure by using maximum likelihood estimation indicated that likelihood function cannot be factored into a product of cause-specific functions. Therefore, we fitted all subdistributions jointly.

Simulation results showed that maximum likelihood procedure performed well for parameter estimation. In BMT data analysis, the proposed model gave noticeably good fit to the nonparametric

counterpart, except for the first few days of AML-low group of patients.

The Gompertz-like subdistribution model is useful when the form of subdistribution might be proper. The model is used for the case of occurrence at a fairly steady rate over the entire time period. The application of this model to contraceptive discontinuation data gave the reasonable result.

In conclusion, the proposed models are useful in studying the relationship between covariates and competing risks survival time data through modelling subdistribution function. In particular, regression trees, hybrid and parametric model might be considered as another way of modelling competing risks data.

7.2 Direction for Further Research

There are some directions in which further research can proceed. Firstly, we can develop model which grows trees based on homogeneity measure of residual similar with survival trees proposed by Therneau *et al.* (1990). Unfortunately, for subdistribution hazards regression the residual is only available for subjects with uncensored data (Schoenfeld residual). To proceed on this idea, we have to construct a residual which is defined for

all subjects (censored and uncensored data) first which is similar to Martingale residual.

Secondly, we can study the cutpoint determination, tree-structured and hybrid methods based on parametric regression on subdistribution function (Chapter 6). Of course, we have to select the suitable underlying kernel distribution first. The kernel distributions can also be extended to a more broad range of well known distributions, such as lognormal and loglogistic.

For the parametric regression of subdistribution function, we can incorporate covariates through parameter other than cure fraction. It may increase the fitting, but we may lose the linear form of the complementary log-log transformation.

REFERENCES

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics* 6:701-726.

Abdolell, M., LeBlanc, M., Stephens, D. and Harrison, R.V. (2002). Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Statistics in Medicine* 21:3395-3409.

Ali, M. and Cleland, J. (1999). Determinants of contraceptive discontinuation in six developing countries. *J. biosoc. Sci.* 31, 343-360.

Allison, P. D. (1995). Survival Analysis Using the SAS System. SAS Institute, Cary, NC.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, New York: Springer Verlag.

Andersen, P. K., Abildstrom, S. Z., and Rosthoj, S. (2002). Competing risks as a multi-state model. *Statistical Methods in Medical Research* 11, 203-215.

Andersen, P. K., Klein, J. P., and Rosthoj, S. (2003). Generalized linear models for correlated pseudo-observations, with application to multi-state models. *Biometrics* 90:15-27.

Basu, A. P. and Klein, J. P. (1982). Some recent results in competing risks theory. In *Survival Analysis*, Vol. 2, J. Crowley and R. A. Johnson, Eds., IMS Monograph Series, Hayward, CA.

Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. Roy. Statist. Soc. Ser. B* 11:15–44

Breiman, L, Friedman, J, Olshen, R., and Stone, C. (1984). *Classification and regression trees*. Chapman and Hall, New York.

Breslow, N. (1972). Contribution to the discussion of paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B* 34:216-217.

Chen M.-H., Ibrahim J.G., Sinha D. (1999). A New Bayesian Model for Survival Data with a Surviving Fraction. *Journal of the American Statistical Association*, 94(447):909-919.

Ciampi, A., Thiffault, J., Nakache, J-P. and Asselain, B. (1986). Stratification by stepwise regression, correspondence analysis and recursive partition. *Computational Statistics and Data Analysis* 4:185-204.

Clark, L. and Pregibon, D. (1992). Tree-Based Models. In Chambers, J.M., Hastie T.J. *Statistical Model in S*. chapter. London: Chapman & Hall.

Collet, D. (1994). *Modelling Survival Data in Medical Research*. Chapman & Hall. London.

Contal, C., O'Quigley, J. (1999). An application of changepoint methods in studying the effect of age on survival in breast cancer. *Comput. Statist. Data Anal.* 30:253–270.

Cox, D. R. (1972). Regression models and life-tables. *J. Royal Statist. Soc.* B 34, 187-220.

Crowder, M. J. (2001). *Classical Competing Risks*. Chapman and Hall/CRC, London.

David, H.A. and Moeschberger, M.L. (1978). *The Theory of Competing Risks*. Griffin, London.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64:247-254.

Davis, R. and Anderson, J. (1989). Exponential survival trees. *Statistics in Medicine* 8, 947-962.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge Univ. Press.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Fine, J. P. (2001). Regression modeling of competing crude failure probabilities. *Biostatistics* 2, 85-97.

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, 496-509.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*, Wiley, New York.

Gail, M. (1975). A review and critique of some models used in competing risk analysis. *Biometrics* 31:209-222.

Gao, F., Manatunga, A. K. and Chen, S. (2004). Identification of prognostic factors with multivariate survival data. *Computational Statistics and Data Analysis* 45, 813-824.

Gao, F., Manatunga, A. K. and Chen, S. (2006). Developing multivariate survival trees with a proportional hazards structure. *Journal of Data Science* 4:343-356

Gooley, T. A., Leisenring, W., Crowley, J., and Storer, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine* 18:695-706.

Gordon, L. and Olshen, R. (1985). Tree-structured survival analysis. *Cancer Treatment Reports* 69, 1065-1069.

Gray, R. J. (1988). A class of *k*-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics*, 16:1141-1154.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* 69:553-566.

Hawkins, D. M. and Kass, G. V. (1982). Automatic Interaction Detection. In Hawkins, D. M. *Topics in Applied Multivariate Analysis* (pp.269-302) Cambridge University Press. London.

Hoel, D.G. (1972). A representation of mortality data by competing risks. *Biometrics* 28:475–488.

Holt, J. D. (1978). Competing risk analysis with special reference to matched pair experiments. *Biometrika* 65:159-166.

Huang, X., Chen, S. and Soong, S. (1998). Piecewise exponential survival trees with time-dependent covariates. *Biometrics* 54, 1420-1433.

Islam, M. A. (1994). Multistate survival models for transitions and reverse transitions: an application to contraceptive use data. *Journal of the Royal Statistical Society A* 157, 441-455.

Jeong, J.-H. and Fine, J. (2006). Direct parametric inference for the cumulative incidence function. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 55(2):187-200.

Jeong, J.-H. and Fine, J. (2007). Parametric regression on cumulative incidence function. *Biostatistics*. 8(2):184-196.

Jespersen, N.C.B. (1986). Dichotomizing a continuous covariate in the Cox regression model. Statist. Res. Unit Univ. Copenhagen Res. Report 86 (2).

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Karia, S. R. (1998). Analysis of contraceptive failure data in Intrauterine Devices studies. *Contraception* 58, 361-374.

Klein, J. P. (2006). Modelling competing risks in cancer studies. *Statistics in Medicine* 25:1015-1034.

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

Klein, J. P. and Moeschberger, M. L. (1984). Asymptotic bias of the product limit estimator under dependent competing risks. *Indian Journal of Productivity, Reliability and Quality Control* 9:1-7.

Klein, J. P. and Moeschberger, M. L. (1987). Independent or dependent competing risks: does it make a difference? *Communication in Statistics – Simulation* 16:507-533.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: techniques for censored and truncated data*. 2nd Edition. New York:Springer.

Koti, K. M, (2004). On estimating the gamma accelerated failure-time models. *Handbook of Statistics, Vol. 23*. 479-493.

Larson, M. G. and Dinse, G. E. (1985). A mixture model for the regression analysis of competing risks data. *Appl. Stat.* 34:201-211

Lausen, B., Schumacher, M. (1992). Maximally selected rank statistics. *Biometrics* 48:73–85.

Lausen, B., Schumacher, M. (1996). Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. *Comput. Statist. Data Anal.* 21:307–326.

LeBlanc, M. (1989). *Regression trees for censored survival data*. unpublished PhD dissertations. University of Washington, Department of Biostatistics.

LeBlanc, M. and Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics* 48, 411-425.

LeBlanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association* 88, 457-467.

Lunn, M. and D. McNeil (1995). Applying Cox regression to competing risks. *Biometrics* 51, 524-532.

Maller, R. A. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. Wiley, Chichester.

Maller, R. A. and Zhou, X. (2002). Analysis of parametric models for competing risks. *Statist. Sin*. 12:725-750.

Mandrekar, J.N., Mandrekar, S.J. and Cha, S.S. (2003). Cutpoint determination methods in survival analysis using SAS®. SUGI 28 Proceedings. http://www2.sas.com/proceedings/sugi28/261-28.pdf

Mielke, P. W. and Berry, K. J. (2007). Permutation Methods: A Distance Function Approach. New York: Springer-Verlag.

Mingers, J. (1987). Expert Systems - Rule Induction with Statistical Data. *Journal of the Operational Research Society*, 38:39-47.

Moeschberger, M.L. (1974). Life tests under dependent competing causes of failure. *Technometrics* 16, 39-47.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association* 58, 415-435.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics* 14:945-965.

Nelson, W. (1969). On estimating the distribution of random vector when only the coordinate is observable. *Technometrics* 12:923-924.

Niblett, T. and Bratko, I. (1986). Learning decision rules in noisy domains, Expert Systems 86, Cambridge University Press (*Proceedings of Expert Systems 86 Conf.*), Brighton.

O'Brien, S. M. (2004). Cutpoint selection for categorizing a continuous predictor. *Biometrics* 60:504-509.

Pepe, M. S. (1991). Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association* 86:770-778.

Pepe, M. S. and Mori, M. (1993). Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine* 12:737-751.

Prentice, R. L., J. D. Kalbfleisch, A. V. Peterson, Jr., N. Fluornoy, V. T. Farewell and N. E. Breslow. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* 34, 541-554.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publ. San Mateo, California.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Annals of Statistics* 11: 453-466.

Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986). *Akaike Information Statistics*. D. Reidel Publishing Company. Tokyo.

SAS Institute Inc. (2004). *SAS/OR 9.1 User's Guide: Mathematical Programming*. Cary, NC: SAS Institute Inc.

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

Schulgen G., Lausen, B. Olsen, J. H. and Schumacher, M. (1994). Outcome-oriented cutpoint in analysis of quantitative exposure. *American Journal of Epidemiology* 140:172-184.

Seal, H. (1977). Studies in the history of probability and statistics. xxxv multiple decrement or competing risks. *Biometrika*. 64:429-439.

Segal, M. R. (1988). Regression trees for censored data. *Biometrics* 44, 35-47.

Segal, M. R. (1992). Tree-structured method for longitudinal data. *Journal of the American Statistical Association* 87, 407-418.

Southern,D. A., Faris, P. D., Brant, R., Galbraith, P. D., Norris, C. M., Knudtson, M. L., Ghali, W. A. and for the APPROACH Investigators. (2006). Kaplan-Meier methods yielded misleading results in competing risk scenarios. *Journal of Clinical Epidemiology*, 59:1110-1114.

Sposto R. (2002). Cure model analysis in cancer: An application to data from the Children's Cancer Group. *Statistics in Medicine*, 21(2):293-312.

Steele, F. (2003). Selection effects of source of contraceptive supply in an analysis of discontinuation of contraception: multilevel modeling when random effects are correlated with an explanatory variable. *Journal of the Royal Statistical Society A* 166, 407-423.

Steele, F., Goldstein, H., and Browne, W. (2004). A general multilevel multistate competing risks model for event history data, with an application to a study of contraceptive use dynamics. *Statistical Modelling* 4, 145-159.

Su, X. G. and Fan, J. J. (2001). *Multivariate Survival Trees by Goodness of split.* Technical Report 367. Department of Statistics, University of California, Davis.

Su, X. G. and Fan, J. J. (2004). Multivariate survival trees: a maximum likelihood approach based on frailty models. *Biometrics* 60:93-99.

Su, X. G. and Tsai, C. L. (2005). Tree-augmented Cox proportional hazards models. *Biostatistics* 6:486-499.

Tableman, M. and Kim, J. S. (2004). *Survival Analysis Using S: Analysis of Time-to-Event Data*. Boca Raton: Chapman and Hall/CRC.

Tarone, R. E. and Ware, J. H. (1977). On distribution-free tests for equality for survival distributions. *Biometrika* 64:156-160.

Therneau, T.M., Grambsch, P. M., and Fleming, T.R. (1990). Martingale based residuals for survival models. *Biometrika* 77:147-160.

Torgo, L. (1999). *Inductive Learning of Tree-based Regression Models*. Ph.D. Thesis. University of Porto. http://www.liacc.up.pt/~ltorgo/PhD/th.pdf

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

Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics*, 54(4):1508-1516.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. London: Cambridge University Press.

Venables, W. N. and Ripley, B. D. (2002). Modern Applied Statistics with S 4th edition. New York: Springer-Verlag.

Weiden, P.L., Sullivan, K. M., Fluornoy, N., Storb, R., and Thomas, E. D. (1981). Antileukemic Effect of Chronic Graft-Host Disease. *New England Journal of Medicine* 304:1529-1553.

Yakovlev, A. and Tsodikov, A. (1996). *Stochastic Models of Tumor Latency and Their Biometrical Applications*. World Scientific. Singapore.

Zhang, H. P. (1998). Classification tree for multiple binary responses. *Journal of the American Statistical Association* 93, 180-193.

APPENDIX

APPENDIX A

The procedure of nonparametric estimation of subdistribution function with illustration

The Procedure

In the first step, we calculate the Kaplan-Meier estimate of the overall survival from any event. Both the event of interest as well as the competing risk event are considered 'events'. In the second step, the conditional probabilities of experiencing the event of interest are calculated. The subdistributions are estimated using these probabilities. The step-by-step calculations are detailed below.

Step 1:

Calculate the overall survival probability of being 'event-free'.

1. An 'event' is any event – the onset of the event of interest or the competing risk event. Anyone not experiencing the 'event' (i.e. event free) is considered censored.
2. The Kaplan-Meier survival probabilities corresponding to the 'event' are calculated using the usual procedure.

Step 2:

Calculate the cumulative probability of experiencing the event of interest.

1. Consider the interval between event-of-interest times t_{r-1} and t_r. (Note that a competing risk event may occur in this interval.)
2. The probability of failure for the event of interest, jth cause, is defined as one minus the probability of survival given by $l_j(t_r) = 1-(n_r-d_r)/n_r$, where n_r is the number of units at risk before time t_r and d_r is the number of event of interest at time t_r.
3. Now, consider the overall survival probability of surviving any 'event' (both the event of interest and the competing risk event) up to, but not including, time t_r. This can be obtained from the calculation step 1, and is denoted by $S(t_{r-1})$.
4. Accounting for competing risk, the incidence of the event of interest for this interval is estimated as the product $S(t_{r-1})\times l_j(t_r)$. This can be interpreted as the joint probability of experiencing the event of interest in this time interval given that the units survived both the event of interest and the competing risk event in all prior intervals.
5. The subdistribution to the end of this time interval is defined as the sum of the incidence in this interval and all previous time intervals.

The subdistribution of every distinct type of failure can be calculated as described above. The subdistribution of any event by a given time will be the sum of the incidence of all distinct failures by that time.

Illustration

Suppose that we consider the Bone Marrow Transplant (BMT) data. BMT is a standard treatment for acute leukemia. Recovery following bone marrow transplantation is a complex process. Transplantation can be considered a failure when patient's leukemia return (relapse) or when he or she dies while in remission (treatment related death) (Klein and Moeschberger, 2003). We apply the method to estimate subdistribution function of relapse for acute lymphoblastic leukemia (ALL) patients.

Table A.1. Illustration of the subdistribution of relapse for BMT data set using the competing approach

(a) BMT data for ALL patients[a]

Patient number	Follow-up time	Status	Patient number	Follow-up time	Status
1	1	2	20	418	2
2	55	1	21	466	2
3	74	1	22	487	2
4	86	2	23	526	2
5	104	1	24	530	0
6	107	2	25	609	1
7	109	1	26	662	1
8	110	1	27	996	0
9	122	1	28	1111	0
10	122	2	29	1167	0
11	129	1	30	1182	0
12	172	2	31	1199	0
13	192	1	32	1330	0
14	194	2	33	1377	0
15	226	0	34	1433	0
16	230	1	35	1462	0
17	276	2	36	1496	0
18	332	2	37	1602	0
19	383	1	38	2081	0

[a]The follow-up (in days), event status (1=relapse, 2=treatment related death, or 0=alive).

(b) An illustration of estimating subdistribution accounting for competing risk for BMT data, type ALL patients, listed above

Step 1

Time interval (t_r)	# at risk (n_r)	# of events (d_r)	Survival probability ($(n_r-d_r)/n_r$)	Overall survival ($S(t_r)$)
[0,1)	38	0	1	1
[1,55)	38	1	37/38	37/38
[55,74)	37	1	36/37	36/38
[74,86)	36	1	35/36	35/38
[86,104)	35	1	34/35	34/38
[104,107)	34	1	33/34	33/38
[107,109)	33	1	32/33	32/38
[109,110)	32	1	31/32	31/38
[110,122)	31	1	30/31	30/38
[122,129)	30	2	28/30	28/38
[129,172)	28	1	27/28	27/38
[172,192)	27	1	26/27	26/38
[192,194)	26	1	25/26	25/38
[194,230)	25	1	24/25	24/38
[230,276)	23	1	22/23	(24/38)×(22/23)
[276,332)	22	1	21/22	(24/38)×(21/23)
[332,383)	21	1	20/21	(24/38)×(20/23)
[383,418)	20	1	19/20	(24/38)×(19/23)
[418,466)	19	1	18/19	(24/38)×(18/23)
[466,487)	18	1	17/18	(24/38)×(17/23)
[487,526)	17	1	16/17	(24/38)×(16/23)
[526,609)	16	1	15/16	(24/38)×(15/23)
[609,662)	14	1	13/14	(24/38)×(15/23)×(13/14)
[662,∞)	13	1	12/13	(24/38)×(15/23)×(12/14)

Step 2

Time interval (t_r)	# at risk (n_r)	# of events (d_{r1})	Failure probability ($l_1(t_r)$)	Survival up to time t_j ($S(t_{r-1})$)	Incidence $S(t_{r-1})\times l_1(t_r)$	Subdistribution ($F_1(t_r)$)(%)
[0,55)	38	0	0	1	0	0
[55,74)	37	1	1/37	37/38	1/38	1/38
[74,104)	36	1	1/36	36/38	1/38	2/38
[104,109)	34	1	1/34	34/38	1/38	3/38
[109,110)	32	1	1/32	32/38	1/38	4/38
[110,122)	31	1	1/31	31/38	1/38	5/38
[122,129)	30	1	1/30	30/38	1/38	6/38
[129,192)	28	1	1/28	28/38	1/38	7/38
[192,230)	26	1	1/26	26/38	1/38	8/38
[230,383)	23	1	1/23	24/38	12/437	104/437
[383,609)	20	1	1/20	(24/38)×(20/23)	12/437	116/437
[609,662)	14	1	1/14	(24/38)×(15/23)	90/3059	902/3059
[662,∞)	13	1	1/13	(24/38)×(15/23)×(13/14)	90/3059	992/3059 =0.3243

The estimation of subdistribution of relapse is outlined in Table A.1.b. The overall survival from any event given under Step 1 of Table A.1 is calculated using the Kaplan-Meier approach. In step 2 of Table A.1.b, we calculate the probability of event of interest, that is relapse, from 55 days up to 74 days is 1/37. From step 1 we know that probability of surviving (i.e. neither relapse nor dead) up to but not including 55 days is 37/38. Therefore, the incidence of relapse from 55 days up to 74 days is $1/37 \times 37/38 = 1/38$.

Figure A.1 illustrates the subdistribution corresponding to relapse. It is to be noted that the subdistribution of any event is the sum of the subdistribution of the event of interest and the subdistribution of the competing risk events. Therefore the cumulative event among the leukemia patients is the sum of the cumulative relapse and the cumulative death.

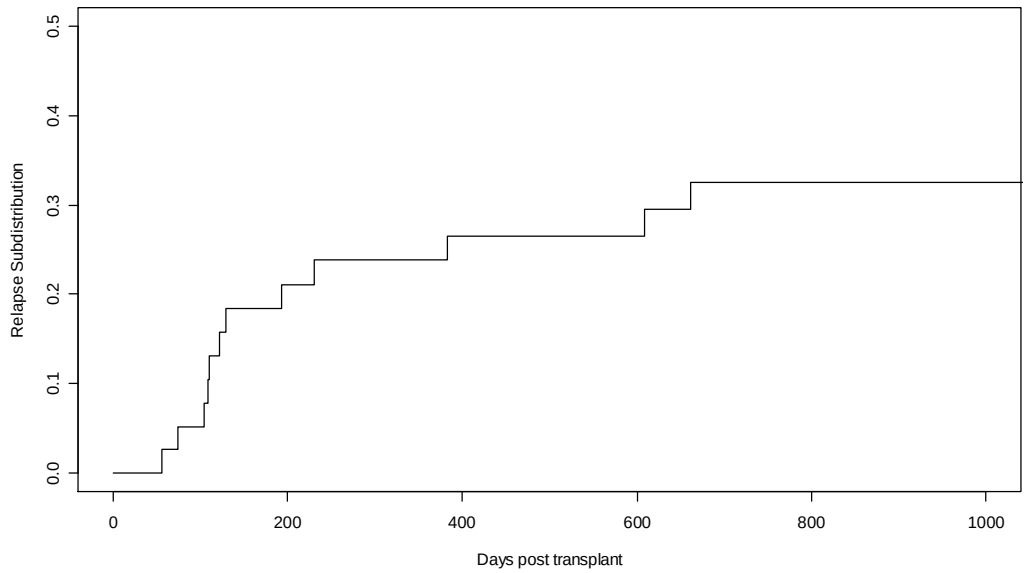

Figure A.1. Subdistribution of Relapse

APPENDIX B

Two sample test for comparing subdistribution function (Gray, 1988)

Data consist of paired observation (T_{ik}, d_{ik}) where T_{ik} is the time on study and d_{ik} is an indicator of the cause of removal from the study, where i is index for individual, $i = 1,\ldots, n_k$; and k is index for group $k = 1,\ldots, K$. We will emphasize on the case $K = 2$.

Let $F_{jk}(t)$ be the subdistribution function for cause j in group k at time t. The hypothesis of interest is the equality of subdistribution failure type 1 across populations, i.e.:

$$H_0 : F_{11}(t) = \ldots = F_{1K}(t) = F_1^0(t), \text{ for all } t \le t \quad \text{versus}$$

H_1 : at least one of the $F_{1k}(t)$'s is different for some $t \le t$.

where $F_1^0(\bullet)$ is an unspecified subdistribution function. Inference is to the subdistribution functions for all time points less than t, which is usually taken to be the largest time on study. The $F_{jk}(t)'s$ are assumed to be continuous with subdensities $f_{jk}(t)$.

The test statistic is based on the (improper) random variable, X_{ik}^*, $i = 1,\ldots, n_k$; $k = 1,\ldots, K$. This random variable is defined by

$$X_{ik}^{*} = \begin{cases} T_{ik}, & \text{if } \delta_{ik} = 1 \\ \infty, & \text{if } \delta_{ik} > 1 \end{cases} \tag{B.1}$$

Then $P\left(X_{ik}^{*} \le t\right) = P\left(T_{ik} \le t, d_{ik} = 1\right) = F_{1k}(t)$ and the hazard rate for X_{ik}^{*} is $\tilde{l}_{ik}(t)$ given by

$$\tilde{l}_{ik}(t) = \frac{dF_{ik}(t)/dt}{1 - F_{ik}(t)} = \frac{f_{ik}(t)}{1 - F_{ik}(t)} \tag{B.2}$$

Let $\hat{F}_{1k}$ be the estimated subdistribution function for cause 1 and sample k and $\hat{F}_{1}^{0}(t)$ be a similar estimator based on the pooled sample. Let $\hat{S}_{k}\left(t^{-}\right)$ be the left-hand limit of the Kaplan-Meier estimate of the overall survival function in sample k obtained by considering failure from any cause as an event. $\hat{S}_{k}\left(t^{-}\right)$ is defined to be 0 when $Y_k(t) = 0$. The K sample statistic will be defined by assigning a score to each group which compares subdistribution hazard $\tilde{l}_{jk}$ for each group to a combined estimate of this hazard under the null hypothesis. Gray (1988) define

$$X_{k}(t) = I(t_{k} \ge t) Y_{k}(t) \left[1 - \hat{F}_{1k}\left(t^{-}\right)\right] / \hat{S}_{k}\left(t^{-}\right) \tag{B.3}$$

The quantities t_k represent the largest time on study in group k. An estimate of the cumulative subdistribution hazard function for the cause of interest in sample k, $\tilde{L}_{1k}(t) = \int_{0}^{t} \tilde{l}_{1k}(u)du$, is given by

$$\hat{\tilde{L}}_{1k}(t) = \int_0^t \frac{d\hat{F}_{1k}(u)}{1-\hat{F}_{1k}(u^-)} = \int_0^t \frac{dN_{1k}(u)}{R_k(u)}, \text{ for } t \le t_k \tag{B.4}$$

The expression for $\hat{\tilde{L}}_{1k}$ suggests taking

$$\hat{\tilde{L}}_1^0(t) = \int_0^t \frac{dN_{1\bullet}(u)}{R_\bullet(u)} \tag{B.5}$$

as an estimator for $\tilde{L}_1^0$, the null value of $\tilde{L}_{1k}$. This estimator is consistent under the null hypothesis.

K sample tests are based on scores of the form

$$Z_k = \int_0^{t_k} W_k(t)\left\{d\hat{\tilde{L}}_{1k} - d\hat{\tilde{L}}_1^0\right\}(t) \tag{B.6}$$

where $W_k(\bullet)$ is suitably chosen weight functions. When the null hypothesis is true, $Z = (Z_1, \ldots, Z_k)'$ has an asymptotic *K*-variate normal distribution with zero mean and covariance matrix S which can be consistently estimated by $\hat{\boldsymbol{\Sigma}}$ with components given by

$$\begin{aligned} \hat{s}_{jj'}^{\,2} = &\sum_{k=1}^{K} \int_0^{t_j \wedge t_{j'}} a_{jk}(t) a_{j'k} h_k^{-1} d\hat{F}_1^{\,0}(t) \\ &+ \sum_{k=1}^{K} \int_0^{t_j \wedge t_{j'}} b_{2jk}(t) b_{2j'k} h_k^{-1} d\hat{F}_{2k}(t) \end{aligned} \tag{B.7}$$

where

$$a_{jk}(t) = d_{jk}(t) + b_{1jk}(t)$$

$$b_{ljk}(t) = \left[I(l=1) - \left(1-\hat{F}_1^{\,0}(t)\right) / \hat{S}_k^{\,0}(t^-)\right]\left[c_{jk}(t_j) - c_{jk}(t)\right]$$

$$c_{jk}(t) = \int_0^t d_{jk}(u)\, d\hat{\hat{L}}_1^0(u)$$

$$d_{jk}(t) = n^{-1} W_j(t)\left[I(j = k) - \hat{h}_k(t) / \hat{h}_\bullet(t)\right] / \left[1 - \hat{F}_1^0(t)\right]$$

Here, $\hat{h}_k(t) = n^{-1} I(t \le t_k) Y_k(t) / \hat{S}_k(t^-)$ and an estimate of $F_1^0(t)$ is given by

$$\hat{F}_1^0(t) = n^{-1} \int_0^t \frac{dN_{1\bullet}(u)}{h_\bullet(u)} \qquad \text{(B.8)}$$

In practice the weight functions $W_k(t)$ are generally of the form $L(t)R_k(t)$, for some function $L(t)$. In this case, $\sum_{k=1}^{K} Z_k = 0$, so only K - 1 of the scores are linearly independent. An appropriate K-sample test statistic can then be formed by using a quadratic form consisting of K - 1 components of Z and their estimated variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_0$:

$$X^2 = (Z_1(t), \ldots, Z_{K-1}(t))\, \boldsymbol{\Sigma}_0^{-1}\, (Z_1(t), \ldots, Z_{K-1}(t))' \qquad \text{(B.9)}$$

When the null hypothesis is true, this statistic has an asymptotic chi-squared distribution with K – 1 degrees of freedom.

The form of the test statistic (B.9) is clear when only two groups are being compared. For this case it is proposed that tests be based on a score of the form

$$\int_0^t W(t)\left[\frac{d\hat{F}_{11}(t)}{1-\hat{F}_{11}(t^-)}-\frac{d\hat{F}_{12}(t)}{1-\hat{F}_{12}(t^-)}\right] \tag{B.10}$$

where $W(\cdot)$ is a weight function. This statistic compares weighted averages of the subdistribution hazards $f_{1k}/(1 - F_{1k})$ in two groups. With the $W_k(t)$ in (B.6) being of the form $L(t)R_k(t)$, and setting $W(t) = L(t)R_1(t)R_2(t)/[R_1(t)+R_2(t)]$ in (B.10), it can be verified that (B.6) has the desirable property of reducing to (B.10) when only two groups are being compared.

APPENDIX C

The proportional subdistribution hazard regression model (Fine and Gray, 1999)

We assume that there are only two competing risks and the risk indexed by 1 is of the main interest. Let T be the failure time, let d be an indicator of the cause of the removal from the study, and let Z be a $p\times1$ covariate vector. Our interest is in modeling the subdistribution function for failure from cause j conditional on the covariates, $F_j(t;Z) = P\{T \le t,\ d = j \mid Z\}$. Inference will be based on a sample of size n consisting of the triplets (T_i, d_i, Z_i), where T_i is the time on study for the ith patient, d_i is the event type indicator, and Z_i is the vector of covariates for the ith individual.

The complement of subdistribution function is equal to

$$1 - F_j(t;Z) = P\left(T > t \cup \left(T \le t \cap d \ne j\right) \mid Z\right) \qquad \text{(C. 1)}$$

and the subdistribution hazard is given by

$$-\frac{d}{dt}\ln\left[1-F_j(t)\right]=\frac{\frac{d}{dt}F_j(t)}{1-F_j(t)},\quad \text{for } j=1,2,\ldots,J$$

$$=\frac{\lim_{\Delta\to 0}\frac{F_j(t+\Delta)-F_j(t)}{\Delta}}{1-F_j(t)}$$

$$=\frac{\lim_{\Delta\to 0}\frac{P\left(t\le T<t+\Delta,d=j\right)}{\Delta}}{1-F_j(t)}$$

$$=\frac{\lim_{\Delta\to 0}\frac{P\left(t\le T<t+\Delta,d=j\right)}{\Delta}}{P\left[T>t\cup\left(T\le t,d\ne j\right)\right]}$$

$$=\lim_{\Delta\to 0}\frac{P\left[t\le T<t+\Delta,d=j\mid T>t\cup\left(T\le t,d\ne j\right)\right]}{\Delta}$$

$$=\tilde{\lambda}_j(t) \qquad \text{(C.2)}$$

$\tilde{\lambda}_j$ can be seen as the hazard rate function of the improper variable X^* as previously defined by

$$X_{ik}^{*}=\begin{cases}T_{ik}, & \text{if } \delta_{ik}=1\\ \infty, & \text{if } \delta_{ik}>1\end{cases}$$

The difference between cause-specific hazard rate and subdistribution hazard is in the risk set. For the cause-specific hazard, the risk set decreases at each time point at which there is a transition to another event. For $\tilde{\lambda}_j$, individuals who fail due to risk other than *j* remain in the risk set. If there is no censoring, they remain in the risk set forever. If there is censoring, they remain in the risk set until their potential censoring time.

The model assumes a proportional hazard form for $\tilde{l}_j$. In this model, given a covariate *Z*, the conditional subdistribution hazard rate $\tilde{l}_j(t;Z)$ is a product of an arbitrary baseline hazard rate $\tilde{l}_{j0}(t)$ and the exponential of the covariate. That is,

$$\tilde{l}_j(t;Z) = \tilde{l}_{j0}(t)\exp(b'Z), \tag{C.3}$$

from (C.2) $\tilde{l}_j(t) = -\frac{d}{dt}\log[1-F_j(t)]$, so this model corresponds to a semiparametric transformation model

$$F_j(t;Z) = 1 - \exp\left\{-\exp(b'Z)\times \tilde{l}^*_{j0}(t)\right\} \tag{C. 4}$$

where $\tilde{l}^*_{j0}(t) = \int_0^t \tilde{l}_{j0}(u)du$, and it can be expressed in linear model form as

$$g\{F_j(t;Z)\} = \tilde{l}^*_{j0}(t) + b'Z \tag{C. 5}$$

with a link function *g* being complementary log-log function, $g(u)$ = $\log(-\log(1-u))$ and $\tilde{l}^*_{j0}(\cdot)$ being some unspecified function.

First assume that there are no censored observations: everyone is seen to progress to an end point. The risk set consists of all individuals who have not yet failed of the cause of interest or who

will never experience this event type: $R_i = \{i' : (T_{i'} \geq t_i) \cup (T_{i'} \leq t_i \cap d_{i'} \neq j\}$. The risk set leads to a partial likelihood for the improper distribution $F_j(t;Z)$:

$$L(b) = \prod_{i=1}^{n} \left(\frac{\exp(b' Z_i)}{\sum_{i' \in R_i} \exp(b' Z_{i'})} \right)^{I(d_i = j)} \tag{C.6}$$

The log partial likelihood is

$$l(b) = \sum_{i=1}^{n} I(d_i = j) \left(b' Z_i - \log \sum_{i' \in R_i} \exp(b' Z_{i'}) \right) \tag{C. 7}$$

The score obtained by differentiating the log partial likelihood with respect to b:

$$\mathrm{U}(b) = \sum_{i=1}^{n} I(d_i = j) \left(Z_i - \frac{\sum_{i' \in R_i} Z_{i'} \exp(b' Z_{i'})}{\sum_{i' \in R_i} \exp(b' Z_{i'})} \right) \tag{C.8}$$

Defining the counting process $N_i(t) = I\{T_i \leq t, d_i = j\}$ and $Y_i(t) = 1 - N_i(t^-)$, we obtain as score function

$$\mathrm{U}(b) = \sum_{i=1}^{n} \int_0^{\infty} \left[Z_i - \frac{\sum_{i'} Y_{i'}(s) Z_{i'} \exp(b' Z_{i'})}{\sum_{i'} Y_{i'}(s) \exp(b' Z_{i'})} \right] dN_i(s) \tag{C.9}$$

Next, suppose some individuals may be censored, but that for each individual the potential censoring time C_i is known. This occurs if censoring is due to cutoff date of analysis. Then the risk set consists of all individuals who did not yet pass their censoring time and who have not (yet) failed of the cause of interest $R_i = \{i' : (C_{i'} \wedge T_{i'} \geq t_i) \cup (T_{i'} \leq t_i \cap d_{i'} \neq j \cap C_{i'} \geq t_i\}$. The subdistribution hazard

incorporating this type of censoring is equal to the subdistribution hazard with complete data:

$$\begin{aligned}\tilde{l}_j^*(t;Z) &= \lim_{Dt\to 0}\tfrac{1}{Dt}P\left(t \le T < t+Dt, d=j \mid C \wedge T \ge t \cup \left(T \le t \cap d \ne j \cap C \ge T\right) \mid Z\right) \\ &= \lim_{Dt\to 0}\frac{\tfrac{1}{Dt}P\left(t \le T < t+Dt, d=j, C \ge t \mid Z\right)}{P\left(\left[T \ge t \cup \left(T \le t \cap d \ne j\right)\right] \cap C \ge t \mid Z\right)} \\ &= \lim_{Dt\to 0}\frac{\tfrac{1}{Dt}P\left(t \le T < t+Dt, d=j \mid Z\right)P\left(C \ge t \mid Z\right)}{P\left(\left(T \le t \cap d \ne j\right) \cup T \ge t \mid Z\right)P\left(C \ge t \mid Z\right)} \\ &= \tilde{l}_j(t;Z) \end{aligned} \qquad \text{(C.10)}$$

The estimator of the parameter vector b is obtained by solving the equation U(b) = 0. Using standard counting process techniques, the usual results are obtained: if b_0 is the true value of b, then $n^{1/2}\left(\hat{b} - b_0\right)$ is asymptotically normal with limiting covariance matrix I^{-1} with I consistently estimated by

$$\hat{\mathrm{I}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathrm{S}_1^{(2)}\left(\hat{b},t\right)}{\mathrm{S}_1^{(0)}\left(\hat{b},t\right)} - \mathrm{E}\left(\hat{b},t\right)^{\otimes 2}\right]I\left(d_i = j\right),$$

where

$$\mathrm{E}\left(b,t\right) = \frac{\mathrm{S}_1^{(1)}\left(b,t\right)}{\mathrm{S}_1^{(0)}\left(b,t\right)},$$

$$\mathrm{S}_1^{(h)}\left(b,t\right) = \frac{1}{n}\sum_{i=1}^{n}Y_i(t)Z_i^{\otimes h}\exp\left\{b'Z_i\right\},\ h = 0,1,2$$

and, for a vector v, $\mathrm{v}^{\otimes 0} = 1, \mathrm{v}^{\otimes 1} = \mathrm{v}, \mathrm{v}^{\otimes 2} = \mathrm{vv}'$.

Under general right censoring, inverse probability of censoring weighting (IPCW) techniques is used. Now we have

$$U(b) = \sum_{i=1}^{n} \int_0^{\infty} \left[Z_i - \frac{\sum_{i'} w_{i'}(s) Y_{i'}(s) Z_{i'} \exp(b Z_{i'})}{\sum_{i'} w_{i'}(s) Y_{i'}(s) \exp(b Z_{i'})} \right] w_i(s) dN_i(s) \qquad \text{(C.11)}$$

with $w_i(t) = I(C_i \geq T_i \wedge t) \frac{\hat{G}(t)}{\hat{G}(X_i \wedge t)}$. Here X = min(T,C) and $\hat{G}$ is the Kaplan-Meier estimate of the survival function of the censoring random variable $P(C \geq t)$, obtained from data $\{X_i, 1-d_i, i = 1, \ldots, n\}$.

Note that $\frac{G(t)}{G(X_i \wedge t)} = P(C_i \geq t \mid C_i \geq T_i \wedge t)$ and

$$w_i(t) = \begin{cases} I(t \leq C_i) & \text{if } C_i \leq T_i \\ 1 & \text{if } t \leq T_i \leq C_i \\ \frac{P(C_i \geq t)}{P(C_i \geq T_i)} & \text{if } T_i \leq t \leq C_i \\ \frac{P(C_i \geq t)}{P(C_i \geq T_i)} & \text{if } T_i \leq C_i \leq t \end{cases} \qquad \text{(C.12)}$$

Note that $w_i(t)$ is non-zero for censored individuals up to time of censoring.

The consistent estimator of the parameter vector b is obtained by solving equation U(b) = 0 (equation 2.31). Taking a Taylor series expansion of $U(\hat{b})$ around b_0, the true value of b, a first order approximation holds:

$$n^{1/2}(\hat{b} - b_0) \approx I^{-1}\{n^{-1/2} U(b_0)\}$$

where I^{-1} is the limit of the negative of the inverse of the partial derivative matrix of the score function evaluated at b_0. With the right-censored data, a consistent estimate for I is given by

$$\hat{\mathrm{I}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathrm{S}_2^{(2)}\left(\hat{b}, T_i\right)}{\mathrm{S}_2^{(0)}\left(\hat{b}, T_i\right)} - \mathrm{E}\left(\hat{b}, T_i\right)^{\otimes 2}\right] I\left(d_i = j\right),$$

where

$$\mathrm{E}\left(b, u\right) = \frac{\mathrm{S}_2^{(1)}\left(b, u\right)}{\mathrm{S}_2^{(0)}\left(b, u\right)},$$

and

$$\mathrm{S}_2^{(h)}\left(b, u\right) = \frac{1}{n}\sum_{i=1}^{n} w_i\left(u\right) Y_i\left(u\right) Z_i^{\otimes h} \exp\left\{b' Z_i\right\}, \quad h = 0,1,2$$

and, for a vector v, $\mathrm{v}^{\otimes 0} = 1, \mathrm{v}^{\otimes 1} = \mathrm{v}, \mathrm{v}^{\otimes 2} = \mathrm{vv}'$.

It can be shown that $n^{-1/2}\mathrm{U}\left(b_0\right)$ has a normal limiting distribution with zero mean and covariance matrix W which can be consistently estimated with the empirical covariance matrix

$$\hat{\mathrm{W}} = n^{-1}\sum_{i=1}^{n}\left(\hat{\eta}_i - \hat{y}_i\right)^{\otimes 2},$$

where

$$\hat{\eta}_i = \int_0^{\infty}\left\{Z_i - \frac{\mathrm{S}_2^{(1)}\left(\hat{b}, u\right)}{\mathrm{S}_2^{(0)}\left(\hat{b}, u\right)}\right\} w_i\left(u\right) d\hat{M}_i^{1}\left(u\right),$$

$$\hat{y}_i = \int_0^{\infty}\frac{\hat{\mathrm{q}}\left(u\right)}{\sum_{i=1}^{n} I\left(T_i \geq u\right)} d\hat{M}_i^{2}\left(u\right),$$

$$\hat{\mathrm{q}}\left(u\right) = -\sum_{i=1}^{n}\int_0^{\infty}\left\{Z_i - \frac{\mathrm{S}_2^{(1)}\left(\hat{b}, s\right)}{\mathrm{S}_2^{(0)}\left(\hat{b}, s\right)}\right\} w_i\left(u\right) d\hat{M}_i^{1}\left(u\right) I\left(s \geq u > T_i\right),$$

$$\hat{M}_i^{1}\left(t\right) = I\left(T_i \leq t, d_i = j\right) - \int_0^{t}\left\{1 - I\left(T_i < s, d_i = j\right)\right\}\exp\left\{\hat{b}' Z_i\right\} d\hat{L}^{1}\left(s\right),$$

$$\hat{M}_i^2(t) = I(T_i \le t, d_i = 0) - \int_0^t I(T_i \ge s) d\hat{L}^2(s),$$

$$\hat{L}^1(t) = \frac{1}{n} \sum_{i=1}^{n} \int_0^t \frac{1}{S_2^{(0)}(\hat{b}, u)} w_i(u) dN_i(u),$$

$$\hat{L}^2(t) = \int_0^t \frac{\sum_i d\{I(T_i \le u, d_i = 0)\}}{\sum_{i=1}^{n} I(T_i \ge u)}.$$

Given the true value of b, b_0, the distribution of $n^{1/2}(\hat{b} - b_0)$ can be approximated by a normal distribution with mean zero and covariance matrix $\hat{S} = \hat{\mathbf{I}}^{-1}\hat{W}\hat{\mathbf{I}}^{-1}$. Inference about covariate effects on the subdistribution function can be based on this asymptotic result. Hypothesis testing problem based on this model will be addressed in the next subsection.

Appendix D

Likelihood function for kernel distribution of exponential, Weibull, Gompertz, gamma and generalized gamma in modelling subdistribution based non-mixture cure model

Since the derived-subdistribution function which utilized non-mixture cure model is expressed by $F_j(t) = 1 - \{\exp[-q_j \exp(\boldsymbol{z}' b_j)]\}^{F_j^*(t)}$, then for:

1. Exponential kernel with parameter k

subdistribution: $F_j(t) = 1 - \{\exp[-q_j \exp(\boldsymbol{z}' b_j)]\}^{1-e^{-k_j t}}$

subdensity: $f_j(t) = q_j k_j \exp\left(\boldsymbol{z}' b_j - k_j t - q_j e^{\boldsymbol{z}' b_j} + q_j e^{\boldsymbol{z}' b_j - k_j t}\right)$

likelihood:

$$l(y) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} d_{ji} \left[\log(q_j k_j) + \boldsymbol{z}' b_j - k_j t - q_j e^{\boldsymbol{z}' b_j} + q_j e^{\boldsymbol{z}' b_j - k_j t} \right] + \left(1 - \sum_{j=1}^{J} d_{ji}\right) \log\left[\sum_{j=1}^{J} \exp\left(-q_j e^{\boldsymbol{z}' b_j} + q_j e^{\boldsymbol{z}' b_j - k_j t}\right) - (J-1) \right] \right\}$$

where $y = (y_1, \ldots, y_J)$ with $y_j = (q_j, k_j, b_{j1}, \ldots, b_{jK})$ for $j=1,\ldots,J$.

2. Weibull kernel with parameter (k,a)

subdistribution: $F_j(t; q_j, k_j, a_j, b_j, \boldsymbol{z}) = 1 - \{\exp[-q_j \exp(\boldsymbol{z}' b_j)]\}^{1-\exp(-k_j t^{a_j})}$

subdensity: $f_j(t) = q_j k_j a_j t^{a_j - 1} \exp\left(\boldsymbol{z}' b_j - k_j t^{a_j} - q_j e^{\boldsymbol{z}' b_j} + q_j e^{\boldsymbol{z}' b_j - k_j t^{a_j}}\right)$

likelihood:

$$l(y) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} d_{ji} \left[\log(q_j k_j a_j) + (a_j - 1)\log t + \boldsymbol{z}' b_j - k_j t^{a_j} - q_j e^{\boldsymbol{z}' b_j} + q_j e^{\boldsymbol{z}' b_j - k_j t^{a_j}} \right] + \left(1 - \sum_{j=1}^{J} d_{ji}\right) \log\left[\sum_{j=1}^{J} \exp\left(-q_j e^{\boldsymbol{z}' b_j} + q_j e^{\boldsymbol{z}' b_j - k_j t^{a_j}}\right) - (J-1) \right] \right\}$$

3. Gompertz kernel with parameter (r,t)

subdistribution:

$$F_j(t;q_j, r_j, t_j, b_j, \boldsymbol{z}) = 1 - \left\{\exp\left[-q_j \exp(\boldsymbol{z}' b_j)\right]\right\}^{1-\exp\left\{\frac{t_j}{r_j}\left(1-e^{r_j t}\right)\right\}}$$

subdensity:

$$f_j(t) = q_j t_j \exp\left(\boldsymbol{z}' b_j + r_j t + \frac{t_j}{r_j}\left(1 - e^{r_j t}\right) - q_j e^{\boldsymbol{z}' b_j} + q_j e^{\boldsymbol{z}' b_j + \frac{t_j}{r_j}\left(1-e^{r_j t}\right)} \right)$$

likelihood:

$$l(y) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} d_{ji} \left[\log\left(q_j t_j\right) + \boldsymbol{z}' b_j + r_j t + \frac{t_j}{r_j}\left(1 - e^{r_j t}\right) - q_j e^{\boldsymbol{z}' b_j} + q_j e^{\boldsymbol{z}' b_j + \frac{t_j}{r_j}\left(1-e^{r_j t}\right)} \right] \right.$$
$$\left. + \left(1 - \sum_{j=1}^{J} d_{ji}\right) \log\left[\sum_{j=1}^{J} \exp\left(-q_j e^{\boldsymbol{z}' b_j} + q_j e^{\boldsymbol{z}' b_j + \frac{t_j}{r_j}\left(1-e^{r_j t}\right)} \right) - (J-1) \right] \right\}$$

4. Gamma kernel with parameter (k,g)

subdistribution: $F_j(t) = 1 - \left\{\exp\left[-q_j \exp(\boldsymbol{z}' b_j)\right]\right\}^{I\left(k_j t, g_j\right)}$

subdensity: $f_j(t) = \frac{1}{G\left(g_j\right)} \left\{ q_j k_j^{g_j} t^{g_j - 1} \exp\left[\boldsymbol{z}' b_j - k_j t - q_j e^{\boldsymbol{z}' b_j} I\left(k_j t, g_j\right)\right] \right\}$

likelihood:

$$l(y) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} d_{ji} \left[\log q_j + g_j \log k_j + \left(g_j - 1\right) \log t + \boldsymbol{z}' b_j - k_j t - q_j e^{\boldsymbol{z}' b_j} I\left(k_j t, g_j\right) - \log G\left(g\right) \right] \right.$$
$$\left. + \left(1 - \sum_{j=1}^{J} d_{ji}\right) \log\left\{ \sum_{j=1}^{J} \exp\left[-q_j e^{\boldsymbol{z}' b_j} I\left(k_j t, g_j\right) \right] - (J-1) \right\} \right\}$$

where *I*(.) is incomplete gamma function

$$I(t,g) = \frac{1}{G(g)} \int_0^t u^{g-1} e^{-u} du$$

5. Generalized gamma kernel with parameter (k,g,a)

subdistribution: $F_j(t) = 1 - \left\{\exp\left[-q_j \exp(\boldsymbol{z}' b_j)\right]\right\}^{I\left(k_j t^{a_j}, g_j\right)}$

subdensity:

$$f_j(t) = \frac{1}{G\left(g_j\right)} \left\{ q_j a_j k_j^{g_j} t^{a_j g_j - 1} \exp\left[\boldsymbol{z}' b_j - k_j t^{a_j} - q_j e^{\boldsymbol{z}' b_j} I\left(k_j t^{a_j}, g_j\right)\right] \right\}$$

likelihood:

$$l(y) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} d_{ji} \left[\log(q_j a_j) + g_j \log k_j + (a_j g_j - 1) \log t + \mathbf{z}' b_j - k_j t^{a_j} - q_j e^{\mathbf{z}' b_j} I\left(k_j t^{a_j}, g_j\right) - \log G(g) \right] \right.$$
$$\left. + \left(1 - \sum_{j=1}^{J} d_{ji}\right) \log \left\{ \sum_{j=1}^{J} \exp\left[- q_j e^{\mathbf{z}' b_j} I\left(k_j t^{a_j}, g_j\right) \right] - (J-1) \right\} \right\}$$

APPENDIX E

Likelihood function for parametric regression with Gompertz-like subdistribution model

1. Exponential kernel

subdistribution: (6.24)

subdensity: $f_j(t) = t_j \exp\left[\boldsymbol{z}' b_j + r_j t + \frac{t_j}{r_j} e^{\boldsymbol{z}' b_j}\left(1 - e^{r_j t}\right)\right]$

likelihood:

$$l(y) = \sum_{i=1}^{n}\left\{\sum_{j=1}^{J} d_{ji}\left[\log\left(t_j\right) + \boldsymbol{z}' b_j + r_j t + \frac{t_j}{r_j} e^{\boldsymbol{z}' b_j}\left(1 - e^{r_j t}\right)\right] + \left(1 - \sum_{j=1}^{J} d_{ji}\right)\log\left[\sum_{j=1}^{J}\exp\left(\frac{t_j}{r_j} e^{\boldsymbol{z}' b_j}\left(1 - e^{r_j t}\right)\right) - (J-1)\right]\right\}$$

where $y = (y_1, \ldots, y_J)$ with $y_j = (t_j, r_j, b_{j1}, \ldots, b_{jK})$ for $j=1,\ldots,J$.

2. Weibull kernel

subdistribution: (6.25)

subdensity: $f_j(t) = t_j a_j t^{a_j - 1} \exp\left[\boldsymbol{z}' b_j + r_j t^{a_j} + \frac{t_j}{r_j} e^{\boldsymbol{z}' b_j}\left(1 - e^{r_j t^{a_j}}\right)\right]$

likelihood:

$$l(y) = \sum_{i=1}^{n}\left\{\sum_{j=1}^{J} d_{ji}\left[\log\left(t_j a_j\right) + (a_j - 1)\log t + \boldsymbol{z}' b_j + r_j t^{a_j} + \frac{t_j}{r_j} e^{\boldsymbol{z}' b_j}\left(1 - e^{r_j t^{a_j}}\right)\right] + \left(1 - \sum_{j=1}^{J} d_{ji}\right)\log\left[\sum_{j=1}^{J}\exp\left(\frac{t_j}{r_j} e^{\boldsymbol{z}' b_j}\left(1 - e^{r_j t^{a_j}}\right)\right) - (J-1)\right]\right\}$$

3. Gompertz kernel

subdistribution: (6.26)

subdensity:

$$f_j(t) = t_j q_j \exp\left[\boldsymbol{z}'\boldsymbol{b}_j + r_j t - q_j\left(1 - e^{r_j t}\right) + \frac{t_j}{r_j} e^{\boldsymbol{z}'\boldsymbol{b}_j}\left(1 - e^{-q_j\left(1-e^{r_j t}\right)}\right)\right]$$

likelihood:

$$l(y) = \sum_{i=1}^{n}\left\{\sum_{j=1}^{J} d_{ji}\left[\log\left(t_j q_j\right) + \boldsymbol{z}'\boldsymbol{b}_j + r_j t - q_j\left(1 - e^{r_j t}\right) + \frac{t_j}{r_j} e^{\boldsymbol{z}'\boldsymbol{b}_j}\left(1 - e^{-q_j\left(1-e^{r_j t}\right)}\right)\right]\right.$$
$$\left. + \left(1 - \sum_{j=1}^{J} d_{ji}\right)\log\left[\sum_{j=1}^{J}\exp\left(\frac{t_j}{r_j} e^{\boldsymbol{z}'\boldsymbol{b}_j}\left(1 - e^{-q_j\left(1-e^{r_j t}\right)}\right)\right) - (J-1)\right]\right\}$$

BIODATA OF THE AUTHOR

Abdul Kudus was born in Subang, West Java, Indonesia on the 21th March 1969. He had his early education at Sekolah Dasar Negeri Bendungan 1 Subang, West Java. After finishing his lower secondary education at Sekolah Menengah Pertama Negeri Binong - Subang, he continued his studies at Sekolah Menengah Atas Negeri 1 Subang. He then enrolled at Department of Statistics, Institut Pertanian Bogor (IPB) and obtained his bachelor degree in 1994. After graduation, he worked as research consultant at Central for Economics and Social Studies (CESS) Jakarta for one year. In the middle of 1995 he got offer to be a lecturer in Statistics at Bandung Islamic University. He continued his studies at IPB for the master degree in Statistics and graduated in 1999. In May 2003 he registered as a post-graduate student at the Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia to pursue his Ph.D in the field of Survival Analysis.

LIST OF PUBLICATIONS

Thesis related Journal Publications, Seminars and Exhibitions

Journal Publications:

1. Kudus, A. and Ibrahim, N. A. Weighted two-sample test for comparing subdistribution function of a competing risk. *Statistika* 5(1):11-18 (2005).

2. Ibrahim, N. A., Kudus, A., Daud, I. and Abu Bakar, M. R. Outcome-oriented cutpoint determination methods for competing risks. *Journal of Quality Measurement and Analysis*. (2007). Accepted.

3. Ibrahim, N. A., Kudus, A., Daud, I. and Abu Bakar, M. R. Decision Tree for Competing Risks Survival Probability in Breast Cancer Study. *Int. J. of Biomed. Sci.* Vol. 3 No. 1. pp:25-29 (2008).

4. Kudus, A., Ibrahim, N. A., Daud, I. and Abu Bakar, M. R. Tree-structured regression for subdistribution of competing risks. *Journal of Mathematics and Statistics*. Submitted.

5. Kudus, A., Ibrahim, N. A., Daud, I. and Abu Bakar, M. R. Parametric model for subdistribution of competing risks based on non-mixture cure model with covariates. *Int. J. of Data Analysis Techniques and Strategies*. Submitted.

6. Kudus, A., Ibrahim, N. A., Daud, I. and Abu Bakar, M. R. On Gompertz-like subdistribution of competing risks with application. *American Journal of Applied Sciences*. Submitted.

Proceeding:

1. Kudus, A. and Ibrahim, N. A. Bootstrap confidence interval for the median failure time of three-parameter Weibull distribution. *Proceedings of the 2007 International Conference of Applied and Engineering Mathematics*. Imperial College London, London, U.K., 2-4 July, 2007. pp: 836-839.

2. Kudus, A., Ibrahim, N. A., Daud, I. and Abu Bakar, M. R. Hybrid model for subdistribution of competing risks. *Proceedings of 2nd International Conference on Mathematical Sciences (ICoMS 2007).* Ibnu Sina Institute, Unversiti Teknologi Malaysia, May 28-29, 2007. Accepted.

3. Kudus, A. and Ibrahim, N. A. Modeling cumulative incidence using parametric cure model. *Prosiding Seminar Kebangsaan Sains Kuantitatif 2006*, 19-21 Disember, 2006. Langkawi

4. Kudus, A. and Ibrahim, N. A. Median survival time of Weibull distribution. *Prosiding Konferensi Kebangsaan Pemodelan Matematik dan Statistik*. Jabatan Matematik, Fakulti Sains, Universiti Putra Malaysia, 5-6 September 2006.

5. Ibrahim, N. A., Kudus, A., Abu Bakar, M. R. and Daud, I. Competing risk data generation based on subdistribution function. *Proceedings of the first Int. Conf. on Math. and Stat. (ICoMS-1),* June 19-21, 2006, Bandung - Indonesia. pp:313-320.

6. Kudus, A. and Ibrahim, N. A. SAS macros for generating dependent competing risks data with exponentially distributed marginal cause of failure. *Proceedings of the International Conference on Statistics and Mathematics and Its Applications in the Development of Science and Technology*, Bandung October 4-6, 2004. pp:103-110. ISBN: 979-99168-0-1

7. Kudus, A., Ibrahim, N. A., Abu Bakar, M. R. and Daud, I. 2005. Regression trees for competing risks survival data. *Proceedings of International Conference on Applied Mathematics*. Bandung, August 22-26, 2005.

Paper Presented:

1. Kudus, A., Ibrahim, N. A., Daud, I. Simulation on the identification of multiple high leverage point (HLP) in censored survival regression. Paper submitted to Simposium Kebangsaan Sains Matematik ke-16. Hotel Renaissance, Kota Bharu, 3 – 5 June 2008.

2. Kudus, A., Ibrahim, N. A., Daud, I. and Abu Bakar, M. R. On simulation of competing risks bivariate Weibull distribution. Paper submitted to Simposium Kebangsaan Sains Matematik ke-16. Hotel Renaissance, Kota Bharu, 3 – June 2008.

3. Kudus, A., Ibrahim, N. A., Daud, I. Group deleted generalized potentials for diagnostics of censored survival regression. Paper presented at the 9th Islamic Countries Conference on Statistical Sciences 2007 (ICCS-IX). Concorde Hotel, Shah Alam, 12-14 December 2007.

4. Kudus, A. and Ibrahim, N. A. Decision tree for competing risks survival probability in breast cancer. Paper presented at Intelligent Systems and Information Technology Symposium 2007, Institute of Advance Technology – UPM. 30 – 31 October 2007.

5. Kudus, A., Ibrahim, N. A., Daud, I. and Abu Bakar, M. R. Decision tree for competing risks survival probability in breast cancer study. Poster presented at Exhibition of Invention, Research & Innovation (PRPI) 2007. UPM.

6. Kudus, A., Ibrahim, N. A., Daud, I. and Abu Bakar, M. R. CIPred: Cumulative Incidence Prediction. Poster presented at Exhibition of Invention, Research & Innovation (PRPI) 2006. UPM.

7. Kudus, A., Ibrahim, N. A., Daud, I. and Abu Bakar, M. R. Cutpoint determination method in competing risks data analysis. Poster presented at Exhibition of Invention, Research & Innovation (PRPI) 2005. UPM.

8. Kudus, A., Ibrahim, N. A., Daud, I. and Abu Bakar, M. R. Regression trees for the subdistribution of a competing risk. Paper presented at International Statistics Conference. Kuala Lumpur, December 27-30, 2005.

9. Kudus, A., Ibrahim, N. A., Daud, I. and Abu Bakar, M. R. Competing risks Cox proportional hazard modeling using SAS. SAS User Malaysia Forum 2005. Kuala Lumpur. 21 September 2005.

10. Kudus, A., Ibrahim, N. A. Tree-structured survival analysis using piecewise multiple linear models. Paper presented at Seminar on Mathematical Research. Universiti Malaysia Sabah. August, 2003.

Award:

1. Ibrahim, N. A., Kudus, A., Daud, I., Abu Bakar, M. R and Suliadi. CRRT – Software for Breast Cancer Survivor Prediction. Exhibition of Invention, Research & Innovation (PRPI) 2008 UPM. Silver medal.

2. Kudus, A., Ibrahim, N. A., Daud, I. and Abu Bakar, M. R. Decision tree for competing risks survival probability in breast cancer study. Exhibition of Invention, Research & Innovation (PRPI) 2007 UPM. Bronze medal.