University of Alberta

Cumulative Total Incidence for Estimating the Burden of Recurrent Events and Risk vs. Rate Concepts and Regression Models in Epidemiology

by

Huiru Dong

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Epidemiology

School of Public Health

Fall 2013
Edmonton, Alberta

Dedication

I would like to dedicate this thesis to my greatest parents, Weichang Dong and Aimin Yan.

Abstract

In the first part of this thesis, a straightforward intuitive method for descriptive survival analysis, termed “cumulative total incidence”, is proposed to measure the total burden of recurrent events in a population by a given time. Using data from the Childhood Cancer Survivor Study, demonstrate the utility of this method contrasting this method to cumulative incidence. In the second part of this thesis, the concepts of risk and rate, and their relationship are discussed in the framework of survival analysis. Regression approaches for estimating the association between factors on event risk and event rate are discussed. Using data from the Childhood Cancer Survivor Study on two competing outcomes, we further demonstrate how competing-risk event affects the estimated association of covariates of interest with event risk and rate.

Acknowledgements

This thesis is under support by Alberta Ingenuity Centre for Machine Learning.

Foremost, I would like to express my sincere gratitude to my greatest supervisor Dr. Yutaka Yasui for his continuous support of my study and research, for his patience, motivation, enthusiasm, and immense knowledge. During the two years studying with him, I learned not only the academic knowledge, but also the wisdom of life. I am extremely thankful to have you as my supervisor, Yutaka!

Besides my supervisor, I would like to thank the rest of my thesis committee: Dr. Yan Yuan and Prof. Gian Jhangri, for their encouragement and insightful comments.

I would also like to acknowledge my friends in Dr. Yasui's biostatistics research team. Thank you Qi, Yan, Leah, Noha, Qiaozhi, Stacey, He, Kaviul, Kai, Jian, Jesus and Shomoita for all your support and encouragement; especially, I want to thank Shahab, Conrado, Xuan, and Linwei for their friendship, I would not have had such wonderful life in Canada without you.

Last but not the least, I would like to thank my family: my parents Weichang Dong and Aimin Yan, for all the support and understanding throughout my life.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CCSS	Childhood Cancer Survivor Study
CI	Cumulative Incidence
CRAN	Comprehensive R Archive Network
CTI	Cumulative Total Incidence
KM	Kaplan-Meier
1-KM	Complement of Kaplan-Meier
RT	Radiation Therapy
SMN	Second Malignant Neoplasm
SN	Subsequent Neoplasm

Chapter 1: Introduction

1.1 Outline

In survival analysis, more than one type of event can be of interest or concern. For example, if cancer recurrence after therapy for childhood cancer is the event of interest, death from any cause may preclude the onset of cancer recurrence (i.e., individuals who experience death are no longer at risk for cancer recurrence) and therefore an event that needs to be taken into account in the analysis. In this situation, death is regarded as *competing risk event* for the event of interest, because it precludes or fundamentally alters the probability of the occurrence of the event of interest [1]. Occurrences of competing risk events are different from and cannot be treated as censoring.

To estimate the risk or probability of developing an event of interest by a given time, the complement of a Kaplan-Meier product limit estimate (1-KM) has been used [2]. The 1-KM method is not appropriate with competing risk setting because it does not distinguish competing risk from censoring. The cumulative incidence (CI) is the alternative method, which accounts for competing risks by properly removing individuals who had a competing-risk event from the risk set[1,3].

In many studies, however, the event of interest is a recurrent event: each individual may experience the event of interest multiple times over the study period [4] . To measure the total burden of such events in a population, the CI method is not appropriate because it only considers the first occurrence of the

event of interest for each individual in the analysis, and subsequent occurrences are not included. Therefore, an approach that can reflect a summarization of all events in the population by a given time is needed.

Chapter 2 of this thesis will propose a straightforward, intuitive method for this purpose, termed "cumulative total incidence" (CTI), which summarizes all events that occur in the population by a given time, not just the first event of each subject. The mathematical relationship between CTI with CI will be given. Detailed calculation of CTI is described using a simple hypothetical example initially, followed by a real example from the Childhood Cancer Survivor Study (CCSS). In the CCSS example, we will contrast CTI and CI for the outcome of subsequent neoplasms to demonstrate differences in these two approaches and the utility of CTI.

One may be interested in factors related to the risk of experiencing an event in a given follow-up period, or those related to the rate of experiencing that event. To introduce covariates in competing-risk settings for the need of assessing the association of factors with event occurrences, considering event risk or event rate, many multivariable statistical regression models have been developed [5-7]. Amongst the different regression approaches, two most widely used approaches are Cox proportional hazards model, and Fine and Gray regression on "subdistribution" hazards. In Chapter 3, we will review these two common approaches, and explain how their result interpretations are connected to the key concepts of risk and rate in epidemiology. Using data from CCSS, we illustrate the differences between these two regression approaches in an analysis

of several hypothesized risk factors with time to two competing outcomes: second malignant neoplasm and death from any cause. The purpose of this chapter is to clarify the differences of the two time-to-event regression methods in conjunction with the corresponding epidemiological concepts of risk and rate in the presence of competing risk. With the illustration using the CCSS data, we hope that this will be helpful in choosing the most appropriate method based on research questions.

In Chapter 4, conclusions will be summarized and areas of interest for future research will be discussed.

1.2 Data

Chapter 2 and Chapter 3 use data from CCSS, a 26-institution retrospective cohort study investigating long-term effects of cancer and its therapy, among 5-year survivors of childhood cancer. The CCSS is funded by the US National Cancer Institute. The CCSS cohort is composed of individuals with a confirmed diagnosis of leukemia, Hodgkin lymphoma, non-Hodgkin lymphoma, neuroblastoma, soft tissue sarcoma, bone cancer, central nervous system malignancy, or kidney cancer before the age of 21 years between January 1, 1970 and December 31, 1986, who survived at least 5 years after diagnosis. A detailed description of the CCSS study design has been published previously [8].

Numerous reports show that childhood cancer survivors are at increased risk for developing neoplasm following the primary childhood cancer. These subsequent neoplasms include subsequent malignant neoplasms, non-malignant

meningioma, and non-melanoma skin cancers [8, 9] . Subsequent neoplasm is a recurrent event (i.e., a survivor could experience it more than once) and its occurrence affects the quality of life in cancer survivors greatly, and also their healthcare service utilization. Chapter 2 uses subsequent neoplasms as outcome and estimates the total burden for this recurrent event using the novel statistical method proposed in the chapter. Chapter 3 uses second malignant neoplasms as the outcome and death from any cause as a competing-risk event, to fit both Cox cause-specific proportional hazards model and Fine and Gray subdistribution proportional hazards model, in order to assess associations of various factors if interest with the occurrence of the first second malignant neoplasm, and illustrate the differences between these two regression approaches.

1.3 Software

All the analyses were conducted using R, which is open source software for statistical computing and graphics, and can be downloaded freely from the Comprehensive R Archive Network (CRAN) site (http://cran.r-project.org/). We developed R code for the CTI method proposed in Chapter 2. In Chapter 3, Cox cause-specific proportional hazards model was fit using the R package *survival*, and Fine and Gray subdistribution proportional hazards model was fit using the R package *cmprsk*. All the R codes used for data manipulation, CTI fitting/drawing and model fitting can be found in Appendices of the thesis.

1.4 References

[1] Gooley, T. A., Leisenring, W., Crowley, J., 1999, "Estimation of Failure Probabilities in the Presence of Competing Risks: New Representations of Old Estimators," Statistics in Medicine, **18**(6) pp. 695-706.

[2] Kaplan, E.L., &Meier, P., 1958, "Nonparametric estimation from incomplete observations." Journal of the American Statistical Association, **53**(282), 457-481.

[3] Satagopan, J. M., Ben-Porat, L., Berwick, M., Robson, M., Kutler, D., & Auerbach, A. D., 2004, "A note on competing risks in survival data analysis." British Journal of Cancer, **91**(7), 1229-1235.

[4] Rothman, K.J., 2012, "Epidemiology: an introduction," Oxford University Press, .

[5] Varadhan, R., Weiss, C. O., Segal, J. B., 2010, "Evaluating Health Outcomes in the Presence of Competing Risks: A Review of Statistical Methods and Clinical Applications," Medical Care, **48**(6) pp. S96-S105.

[6] Gerds, T. A., Scheike, T. H., and Andersen, P. K., 2012, "Absolute Risk Regression for Competing Risks: Interpretation, Link Functions, and Prediction," Statistics in Medicine, **31**(29) pp. 3921-3930.

[7] Klein, J. P., and Andersen, P. K., 2005, "Regression Modeling of Competing Risks Data Based on Pseudovalues of the Cumulative Incidence Function," Biometrics, **61**(1) pp. 223-229.

[8] Robison, L. L., Mertens, A. C., Boice, J. D., 2002, "Study Design and Cohort Characteristics of the Childhood Cancer Survivor Study: A Multi-institutional Collaborative Project," Medical and Pediatric Oncology, **38**(4) pp. 229-239.

[9] Friedman, D. L., Whitton, J., Leisenring, W., 2010, "Subsequent Neoplasms in 5-Year Survivors of Childhood Cancer: The Childhood Cancer Survivor Study," Journal of the National Cancer Institute, **102**(14) pp. 1083-1095.

Chapter 2: Estimating the burden of recurrent events in the presence of competing risks: The method of cumulative total incidence

2.1 Introduction

In many clinical studies, it is of interest to estimate the cumulative probability of developing an event by a given time. The complement of a Kaplan-Meier (KM) product limit estimate (1-KM) has been widely used to estimate this cumulative probability. However, when there are competing-risk events, which are events whose occurrence either precludes the occurrence of the event of interest or fundamentally alters its probability of occurrence [1], the method of cumulative incidence (CI) should be used. The KM method is not appropriate when competing-risk events are present because it does not distinguish competing-risk events from censoring, which can result in inflated cumulative probability estimates. The CI method properly removes individuals who had a competing-risk event from the risk set.

The CI approach estimates the cumulative probability of the *first* event of interest over time: subsequent occurrences of the event of interest are not included. When examining the probability of event occurrence within subpopulations of subjects defined by different treatment or other risk factors (i.e., etiological inference), it may be sensible for the analysis to only consider the first occurrence of the event in each subject, particularly if the occurrence of

first event changes the underlying risk and/or biology of the subsequent event (e.g., by treatment).

In many studies, however, the outcome variable of interest is a recurrent event: each individual in the study may experience the event of interest multiple times over the study period [2]. Examples of such outcomes include hospitalization, injuries, repeated heart attacks, and fractures in osteoporosis studies. When it is of interest to measure the total burden of such recurrent events in a population, we would like a methodology that allows a meaningful summarization of all events that occur in the population, not just the first event of each subject [3-5].

To fully describe the disease burden for recurrent events in the presence of competing risks, we propose a straightforward and intuitive method, hereafter referred to as the "cumulative total incidence" (CTI), for estimating the total number of events of interest that would be the average number of events to occur in a population member by a given time. The organization of this paper is as follows: we will propose the CTI; explore the relationship between the CTI estimate and the CI estimate for first event; describe the calculation method of CTI with a hypothetical study; illustrate the use of CTI with data from Childhood Cancer Survivor Study; and close with a discussion regarding some important points that need to be considered when using CTI.

2.2 Estimation of CTI

Our notation is consistent with that of Gooley *et al.* [1] who provided an intuitive form and a clear demonstration of the mechanics of the CI.

In contrast to CI, which is defined as the proportion of a closed population at risk that develops the first occurrence of an event of interest within a given period of time [6], the CTI proposed in this paper refers to the average number of events of interest (first-ever or recurrent) per individual in a population within a given period of time. CI includes only the first occurrence of the event of interest for each individual and describes the average risk of experiencing at least one event in a population, whereas, CTI is a summarization of all the events that occur in the population at risk and reflects the burden of the event of interest in a population.

To estimate CTI, we assume there are n_0 individuals initially at risk in the study. Each individual could experience three distinct kinds of events at time t_j during follow up: (1) occurrence of the event of interest; (2) occurrence of a competing-risk event; and (3) censoring.

The times at which any of the three events occurs can be ordered as $t_1 \le t_2 \le \ldots \le t_n$. We further define the following:

e_j: The number of events of interest occurring at time t_j (including first-ever or recurrent);

r_j: The number of individuals who experience a competing-risk event at time t_j;

c_j : The number of individuals who are censored at time t_j ;

n_j : The number of individuals who are at risk and under observation of the study *beyond* time t_j .

In contrast to the usual CI setting, when measuring the total number of events is of interest, regardless of whether first or later occurrences, individuals can experience the event of interest several times and still remain "at risk" in the study. Thus, individuals can only experience a competing-risk event or censoring outcome once and are removed from the risk set, while those experiencing an event remain in the risk set, which means:

$$n_j = n_0 - \sum_{k=1}^{j} (r_k + c_k) \qquad \text{(Eq. 1)}$$

and the overall KM estimator of survival probability is expressed as

$$KM(t) = \prod_{j=1}^{s} (1 - \frac{r_j}{n_{j-1}}) \qquad \text{(Eq. 2)}$$

where s is the largest j such that $t_j < t$. This survival probability at a given time is the conditional probability that an individual remains at risk for the event of interest at that time. Since individuals who experienced the event of interest are still at risk of experiencing the event of interest again, they will not affect this survival probability. Given the Equation (1) of $KM(t)$, the CTI by time t is estimated by:

$$CTI(t) = \sum_{j=1}^{s} \frac{e_j}{n_{j-1}} KM(t_j) \ . \qquad \text{(Eq. 3)}$$

The $CTI(t)$ can be interpreted as the estimate of expected number of events per person by a given time who have not experienced a competing-risk event by time t. Therefore, the product $CTI(t)\times n_0$ is the estimate of total expected number of events of interest by time t, which can be a more relevant and clearly interpretable measure of overall disease burden in a population than considering only the first event that occurs for each subject.

For calculating the the cumulative probability of the first event of interest at time *t*, $CI(t)$, Gooley *et al*. [1] gave the following formula:

$$CI(t) = \sum_{j=1}^{s} \frac{e_j}{n^*_{j-1}} \prod_{k=1}^{j-1} (1 - \frac{e_k}{n^*_{k-1}}) \prod_{k=1}^{j-1} (1 - \frac{r_k}{n^*_{k-1}}) \qquad \text{(Eq. 4)}$$

where

$$n^*_j = n_0 - \sum_{k=1}^{j} (e_k + r_k + c_k) \qquad \text{(Eq. 5)}$$

Note that the notation defined for Equations (1)-(3) also apply to Equation (4) and (5), but the follow-up time stops after the individual experiences the first occurrence of the event of interest. Thus, $\frac{e_k}{n_{k-1}}$ and $\frac{r_k}{n_{k-1}}$ are the estimate of the hazard of failure from the event of interest and the competing-risk event, respectively, at time t_k. Individuals are removed from the risk set after the first occurrence of the event of interest, occurrence of a competing-risk event, or censoring.

Comparison of Equations (3) and (4) illustrates a critical difference between CTI and CI. For CI, the cumulative probability of the first event of

interest depends on survival free of both the event of interest and the competing-risk event. After experiencing the first occurrence of the event of interest, the individual should not remain in the risk set, because that person cannot provide any additional information about the first occurrence of the event from continued observation. For CTI, however, the survival probability only depends on survival free of a competing-risk event. Because the individual can remain in the risk set after experiencing the event of interest, the number of event occurrences is no longer the same as the number of individuals who experience the event. Therefore, CTI estimates the "average number of events" per person in the population rather than the proportion of individuals who experience the event of interest.

2.3 Relationship between CTI and CI

If we assume individuals can experience at most m recurrences of the event of interest during the study, it is possible to calculate the CI for the first event occurrence ($CI_1(t)$) and also for the second event occurrence ($CI_2(t)$), and so on, until the m th event occurrence ($CI_m(t)$). Thus, the total expected number of event occurrences by time t can be estimated as

$$CI_1(t)\times n_0 + CI_2(t)\times n_0 + \mathrm{K} + CI_m(t)\times n_0 \qquad \text{(Eq. 6)}$$

Therefore, the cumulative total event estimate is equivalent to the sum of CIs for each incremental number of events, i.e.

$$CTI(t)\times n_0 = \sum_{p=1}^{m} CI_p(t)\times n_0 \qquad \text{(Eq. 7)}$$

Here $CI_p(t)$ represents the CI for the $p^{th}(p=1,2,\text{K},m)$ occurrence of the event of interest by time t. We could simplify the Equation (7) as

$$CTI(t)=\sum_{p=1}^{m} CI_p(t) \qquad \text{(Eq. 8)}$$

For calculating $CI_p(t)$, we only treat the p^{th} occurrence of event as the event of interest. The population at risk for the p^{th} occurrence of event would consist of those individuals who have had the $(p-1)^{th}$ or less occurrence of the event of interest, and who are not censored or experience competing-risk event before p^{th} occurrence of the event of interest. After having the p^{th} event, individuals would leave the population at risk for the p^{th} occurrence of event.

For calculating marginal $CI_p(t)$, we further define the following for p^{th} occurrence of event of interest:

e_{pj}: The number of individuals who experience the event of interest at time t_j;

r_{pj}: The number of individuals who experience a competing-risk event at time t_j;

c_{pj}: The number of individuals who are censored at time t_j;

n_{pj}: The number of individuals who are under study *beyond* time t_j.

Since there are n_0 individuals initially at risk, we have $n_{p0}=n_0$ and

$$n_{pj}=n_0-\sum_{k=1}^{j}(e_{pk}+r_{pk}+c_{pk}) \qquad \text{(Eq. 9)}$$

Note that e_j is the number of events of interest by time t_j, regardless of whether it was the first occurrence or not; therefore, we have $e_j = \sum_{p=1}^{m} e_{pj}$. The mathematical proof of Equation (8) can be found in the appendix.

2.4 Illustrative example

In the following example of a recurrent event outcome, we show step-by-step calculations of the CTI and further illustrate the relationship between CTI and CI. We assume five participants were enrolled at the beginning of a study (Figure 2-1). Subject 1 was alive at the end of study and was considered censored at t_8. Subject 2 was lost to follow-up at t_1 and treated as censored. Subject 3 died from a competing-risk event at t_5. Subject 4 experienced the event of interest three times (at t_2, t_6, and t_7), and was alive at the end of study. Subject 5 experienced the event of interest once at t_3 and died at $t_4 (t_3 = t_4)$.

First, we calculate the overall KM survival probability, taking competing-risk events into account (Table 2-1).

Note that when the event of interest occurs, it does not change the number of individuals at risk for the next time interval because these individuals are still at risk for another occurrence of the event of interest.

Next, we calculate the cumulative average number of events per person by time t_j based on Equation (3) (Table 2-2).

From Table 2-2, we see that $CTI(t_8) = 1$, which means that, by time t_8, the cumulative average number of the event of interest per person is estimated to be 1. In other words, since we have 5 individuals initially at risk, if we have no censoring, we could expect to see the event of interest occurring 5 times during follow-up, regardless of whether it is the first occurrence or not.

Since the maximum number of the event of interest experienced was three in this example, we need to calculate only $CI_1(t_j)$, $CI_2(t_j)$, and $CI_3(t_j)$ and we illustrate that $CI_1(t_j) + CI_2(t_j) + CI_3(t_j)$ is equivalent to $CTI(t_j)$ (Table 2-3).

2.5 An example of use in practice

To illustrate the use of CTI and contrast it with the use of CI, we use the Childhood Cancer Survivor Study (CCSS), a large cohort study designed to investigate long-term effects of cancer and therapy, among 5-year survivors of childhood cancer. The CCSS cohort consists of 5-year survivors of childhood cancer diagnosed before the age of 21 years between 1970 and 1986 in one of 26 collaborating pediatric oncology centers. A detailed description of the CCSS study design has been published previously [7, 8].

Numerous reports show that childhood cancer survivors are at increased risk for developing neoplasm following the childhood cancer. These subsequent neoplasms (SNs) include subsequent malignant neoplasms, non-malignant meningioma, and non-melanoma skin cancers [7, 9]. The occurrence of an SN affects the quality of life in cancer survivors greatly [9], and also their healthcare

service utilization/needs. Radiation therapy (RT) has been consistently reported to increase the risk of SN [9-11]. To understand the total burden of SNs among childhood cancer survivors by RT exposure, CTI proposed in this paper is useful.

Specifically, we report CTI and CI estimates of SNs for a cohort of 12,588 survivors, starting from the CCSS cohort entry (5 years after the original childhood cancer diagnosis). The survivors are stratified by whether they received RT treatment (RT group) or not (No RT group) in the 5-year period following the childhood cancer diagnosis prior to the CCSS cohort entry. Death from any cause was treated as a competing-risk event for occurrences of SN, and survivors were censored at the date of last contact.

Among 8,469 survivors who received RT treatment, 1,229 had at least one SN and a total of 2,112 occurrences of SN were reported after the CCSS cohort entry: of the 1,229 survivors with SN, 840 (68.3%) had only one occurrence of SN and 389 (31.7%) had multiple SNs. Among 4,119 survivors who did not receive RT treatment, a total of 221 occurrences of SN were reported among 178 individuals. Of the 178 survivors with an SN, 147 (82.6%) had only one occurrence and 31 (17.4%) had multiple SNs.

Figure 2-2a shows the estimated CTI curves and 95% confidence intervals calculated by bootstrapping individual survivors [12]. The CTI analysis with all SN occurrences reveals that at 39 years since diagnosis, there would be 56.0 SNs occurring per 100 survivors (CTI=0.56) in the RT group, compared to 16.1 SNs per 100 survivors in the No RT group. In other words, we could expect

to observe more than one SN per every two survivors in the RT group, whereas, less than one SN per every six survivors in the No RT group. This result suggests an approximately 3.5 times higher number of SNs for the survivors who received RT treatment compared with those who did not receive RT treatment at 39 years since diagnosis. Figure 2-2b shows the estimated CI curves and 95% confidence intervals that include only the first SN occurrence for each survivor. It reveals that the probabilities of developing at least one SN at 39 years since diagnosis are 0.26 and 0.10 in the RT and No RT groups, respectively.

2.6 Discussion

To capture the burden of recurrent events in a population by a given time in the presence of competing risk, we proposed in this paper the CTI approach. We mathematically proved and empirically showed the equivalence between the CTI and the sum of CIs for incremental numbers of events in the population.

When analyzing data with recurrent events, based on the scientific questions of the study, one should first clearly establish whether the percentage of people who experience occurrence of at least one event is of main interest, or if a summary of the total number of events occurring in a population is of primary interest. For the former, CI can be estimated. For the latter, the proposed CTI would be useful.

An important characteristic of the CTI is that it is not a probability. Its possible range is not from 0 to 1 (this is the range of CI which is a probability); it can be any positive number. This is because we are measuring the average

number of events per member of a certain population, rather than the proportion of the population that develops the event of interest. Also, it is not interpretable without specification of the time period to which it applies: this is also true for CI. Even when applied to the same population size, an average of two events per person can reflect a dramatically different burden of disease for a 50-year time period compared to a 1-year time period. From the illustrative example, we can see that a meaningful interpretation can be given as an average of one event per every X individuals initially at risk, or an average of Y events per 100 individuals initially at risk. For statistical inference with CTI including significance test and confidence intervals, analytic methods need to be developed; as a valid alternative inference method, however, we applied bootstrapping individuals [12].

Differences between CTI and CI become larger when the repeated occurrence of events is frequent. As shown in Figure 2-2 of the CCSS example, CTI in the No RT group led to a similar estimate as CI, while they differed appreciably in the RT group. This reflects the more frequent repeated occurrences of SNs in the RT group relative to the No RT group. In addition, the discrepancy of CTI from CI becomes greater over time. Thus, the CI analysis that incorporates only the first occurrence of SN would underestimate the total burden of SNs more severely with longer follow-up time.

For quantifying incidence of recurrence events, a traditional measure in epidemiology is “rate” which is defined as the total number of events devided by the total person-time at risk for the event [2]. The denominator takes into

account the number of individuals in a cohort, as well as the length of time contributed by each individual. Because the selection of the time unit can be arbitrary, rate does not necessarily have implication for the period of time over which it is actually measured [2]. Rate reflects the fundemental force of all events of interest occurring in a population and it can rise or fall with time. This is different with the CTI we proposed in this paper, as the denominator of CTI only depends on the number of individuals at risk. It has a direct implication for the length of time over which the CTI applies. CTI reflects the burden of the event of interest in a population, and it is a cumulative measure that cannot decrease with the length of risk period.

Finally, the CTI provides an additional approach for describing the occurrence of an outcome that can occur more than once during the period of observation. We do not propose that the CTI should replace other metrics such as CI, standardized incidence ratio, or absolute excess risk in describing the occurrence of an outcome. Rather the CTI provides a new dimension, which reflects a total burden of the event of interest within a population.

Table 2-1 Calculation of overall survival probability, taking competing-risk events into account

Time interval	# at risk n_{j-1}	# censored c_j	# of event of interest e_j	# of competing-risk events r_j	Survival probability $1 - r_j / n_{j-1}$	Overall survival probability $KM(t_j)$
[Time 1, Time 2)	5	1	0	0	1	1
[Time 2, Time 3)	4	0	1	0	1	1
[Time 3, Time 4)	4	0	1	0	1	1
[Time 4, Time 5)	4	0	0	1	3/4	3/4
[Time 5, Time 6)	3	0	0	1	2/3	1/2
[Time 6, Time 7)	2	0	1	0	1	1/2
[Time 7, Time 8)	2	0	1	0	1	1/2
[Time 8	2	2	0	0	1	1/2

Table 2-2 Calculation of CTI

Time interval	# at risk	# of event of interest	Probability of event	Survival up to t_j	Average # of event	Cumulative total incidence
[Time 1, Time 2)	5	0	0	1	0	0
[Time 2, Time 3)	4	1	1/4	1	1/4	1/4
[Time 3, Time 4)	4	1	1/4	1	1/4	1/2
[Time 4, Time 5)	4	0	0	1	0	1/2
[Time 5, Time 6)	3	0	0	3/4	0	1/2
[Time 6, Time 7)	2	1	1/2	1/2	1/4	3/4
[Time 7, Time 8)	2	1	1/2	1/2	1/4	1
[Time 8	2	0	0	1/2	0	1

Table 2-3 Equivalence of the sum of CIs and CTI

Time interval	$CI_1(t_j)$	$CI_2(t_j)$	$CI_3(t_j)$	$CI_1(t_j)+CI_2(t_j)+CI_3(t_j)$	$CTI(t_j)$
[Time 1, Time 2)	0	0	0	0	0
[Time 2, Time 3)	1/4	0	0	1/4	1/4
[Time 3, Time 4)	1/2	0	0	1/2	1/2
[Time 4, Time 5)	1/2	0	0	1/2	1/2
[Time 5, Time 6)	1/2	0	0	1/2	1/2
[Time 6, Time 7)	1/2	1/4	0	3/4	3/4
[Time 7, Time 8)	1/2	1/4	1/4	1	1
[Time 8	1/2	1/4	1/4	1	1

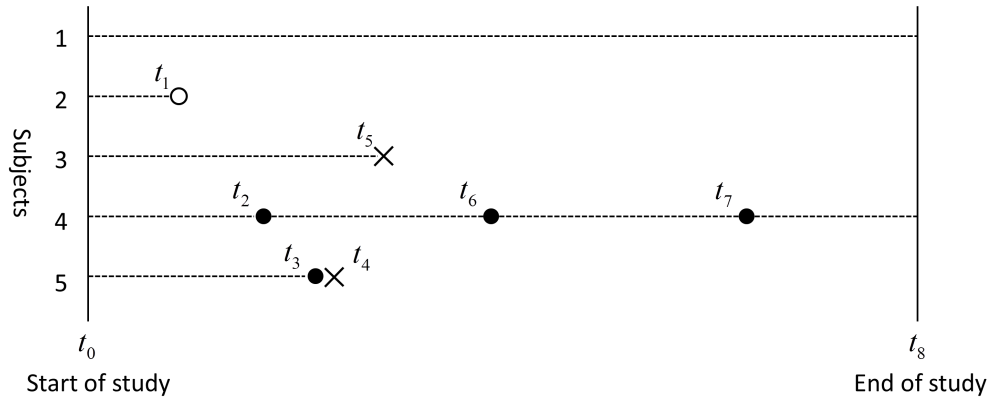

Figure 2-1 A visual representation of a hypothetical study which has a recurrent event outcome. A dashed line represents the follow-up period of each individual. A solid dot represents the occurrence of the event of interest, an open dot represents censoring, and a cross represents the occurrence of the competing-risk event.

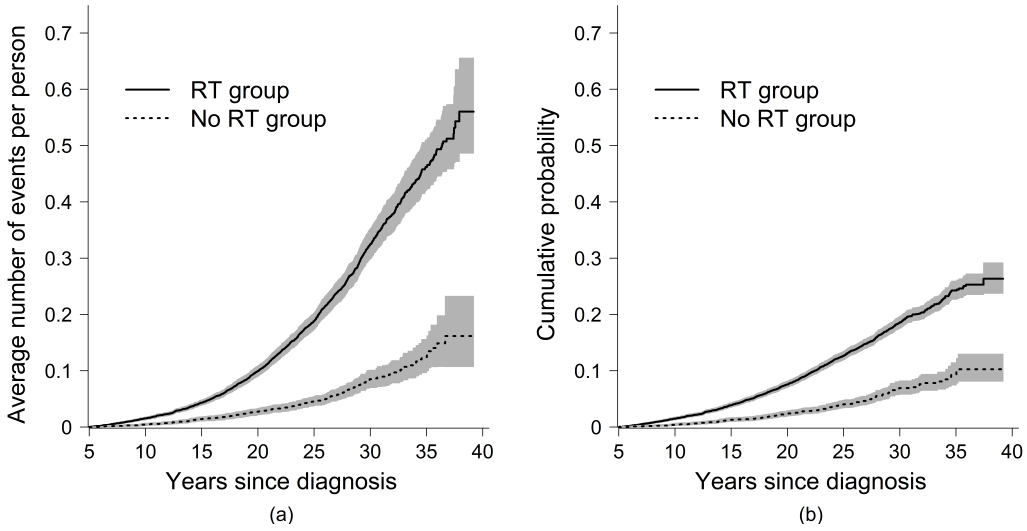

Figure 2-2 CTI curves and 95% confidence intervals calculated by Bootstrapping method (panel a). CI curves and 95% confidence intervals (panel b).

2.7 References

[1] Gooley, T. A., Leisenring, W., Crowley, J., 1999, "Estimation of Failure Probabilities in the Presence of Competing Risks: New Representations of Old Estimators," Statistics in Medicine, **18**(6) pp. 695-706.

[2] Rothman, K.J., 2012, "Epidemiology: an introduction," OUP USA, .

[3] Glynn, R. J., and Buring, J. E., 1996, "Ways of Measuring Rates of Recurrent Events," BMJ, **312**(7027) pp. 364-367.

[4] Cumming, R. G., Kelsey, J. L., and Nevitt, M. C., 1990, "Methodologic Issues in the Study of Frequent and Recurrent Health Problems. Falls in the Elderly." Annals of Epidemiology, **1**(1) pp. 49-56.

[5] Broderick, J., Brott, T., Kothari, R., 1998, "The Greater Cincinnati/Northern Kentucky Stroke Study Preliminary First-Ever and Total Incidence Rates of Stroke among Blacks," Stroke, **29**(2) pp. 415-421.

[6] Rothman, K.J., Greenland, S., and Lash, T.L., 2008, "Modern epidemiology," Lippincott Williams & Wilkins, .

[7] Robison, L. L., Mertens, A. C., Boice, J. D., 2002, "Study Design and Cohort Characteristics of the Childhood Cancer Survivor Study: A Multi-institutional Collaborative Project," Medical and Pediatric Oncology, **38**(4) pp. 229-239.

[8] Robison, L. L., Armstrong, G. T., Boice, J. D., 2009, "The Childhood Cancer Survivor Study: A National Cancer Institute–supported Resource for Outcome and Intervention Research," Journal of Clinical Oncology, **27**(14) pp. 2308-2318.

[9] Friedman, D. L., Whitton, J., Leisenring, W., 2010, "Subsequent Neoplasms in 5-Year Survivors of Childhood Cancer: The Childhood Cancer Survivor Study," Journal of the National Cancer Institute, **102**(14) pp. 1083-1095.

[10] Sklar, C. A., Mertens, A. C., Mitby, P., 2002, "Risk of Disease Recurrence and Second Neoplasms in Survivors of Childhood Cancer Treated with Growth Hormone: A Report from the Childhood Cancer Survivor Study," Journal of Clinical Endocrinology & Metabolism, **87**(7) pp. 3136-3141.

[11] Meadows, A. T., Friedman, D. L., Neglia, J. P., 2009, "Second Neoplasms in Survivors of Childhood Cancer: Findings from the Childhood Cancer Survivor Study Cohort," Journal of Clinical Oncology, **27**(14) pp. 2356-2362.

[12] Good, P.I., 2001, "Resampling methods: A practical guide to data analysis," Springer, .

Chapter 3: Risk and rate regression models for survival analysis in the presence of competing risks

3.1 Introduction

Risk and rate are two important concepts that have been frequently used in epidemiologic studies. William Farr [1] first gave a clear, accurate description of two different ways for measuring mortality, named the "probability of dying" and the "rate of mortality", in his paper "On prognosis" published in 1838. These two terms refer to the concept of "risk" and "rate", respectively. Morgenstern *et al.* [2] described in detail how these two different measures of event occurrences are used in epidemiology, and discussed issues that need to be considered in deciding which quantity to be estimated in a particular study.

Descriptive methods have been developed and commonly used for measuring risk and rate in epidemiologic studies. For example, the complement of a Kaplan-Meier (1-KM) product limit estimate has been used to describe probability of an event occurrence by a given time. This is not appropriate, however, in the presence of competing-risk events because the 1-KM method does not distinguish competing-risk events from censoring [3, 4]. A competing-risk event differs from censoring because it will either preclude or fundamentally alter the probability of the event occurrence [3]: censoring is an event with which the study becomes unable to observe the subject's event occurrences but the nature of the event occurrence process in the subject is unchanged. In the presence of competing risk, cumulative incidence is the alternative method,

which accounts for competing risk events by properly removing individuals who had a competing-risk event from the risk set.

Multiple covariates can be considered in making association inference in regression analysis. Varadhan *et al.* [5] gave a review of statistical methods that can be used to evaluate the effect of a healthcare intervention in the presence of competing risks. Gerds *et al.* [6] compared the statistical properties, such as size of prediction error and fit of the model, across different link functions for regression of the cumulative incidence of an event. Klein and Andersen [7] proposed an approach to regression modeling based on "pseudovalues" (known from jackknife techniques and used in a generalized estimating equation to obtain estimates of model parameters) of the cumulative incidence function. Amongst many different regression approaches, the two widely used are Cox proportional hazards model, and Fine and Gray regression on subdistribution hazards [8-11]. In this paper, we will focus on these two common approaches, and explain how they are connected to the epidemiologic concepts of risk and rate. We feel this is needed because the method of Fine and Gray is often depicted as the method to be used for time-to-event regression in the presence of competing risk which is not accurate: the key difference between these methods corresponds to the difference between risk and rate in epidemiological concepts. This clarification can be helpful for researchers in clinical and public health sciences in choosing the appropriate methods for answering their scientific questions.

The organization of the paper is as follows: (1) review of the concept of risk and rate with a simple hypothetical example; (2) explanation of the relationship between risk and rate in survival analysis with the use of simple statistical concepts; (3) description of two standard regression approaches for the analysis of epidemiologic data with respect to risk and rate; (4) illustration of the use of these methods in an analysis that estimates the association of several factors with the occurrence of second malignant neoplasm (SMN) with data from Childhood Cancer Survivor Study (CCSS); (5) a discussion regarding important points to be considered when using a risk-related approach and a rate-related approach.

3.2 Concept of risk and rate

Consider the following example of epidemiological studies. A cohort of patients diagnosed with a particular cancer is followed up for the primary objective of evaluating recurrence of the cancer, comparing Group A and Group B of the patients (e.g., treatment A vs. treatment B of the cancer). Some patients may be censored at the end of follow up without having developed a recurrence, the event of interest, or died without a recurrence, a competing-risk event. The following two objectives must be distinguished: (1) comparing the chance (*risk*) of developing a recurrence within, say, 5 years, between Group A and Group B; and (2) comparing the *rate* of developing a recurrence per 1,000 person-years of at-risk, between Group A and Group B. Hospitals and healthcare providers may wish to know (1), the risk, in order to prepare for proper resource allocation for

recurrence treatment, while researchers may be interested in (2), a more mechanistic question on Treatment A vs. Treatment B as to which is better for preventing recurrence of this cancer.

The *risk* is defined as the probability of a patient developing a recurrence over a specified period of time [2]. This quantity can be estimated in a sample by the proportion of the patients who have developed a recurrence over the specified period of time. If a patient dies before developing a recurrence, the patient cannot develop a recurrence subsequently. Therefore, death is a *competing risk* against the event of interest (cancer recurrence) and affects the risk. For rate, death (before recurrence) would terminate the person-time at risk for recurrence (i.e., the denominator of the rate), but otherwise the rate is not directly influenced by mortality. The rate quantifies the recurrence number we expect per unit of person-time at risk for recurrence [12]. This quantity can be estimated simply by the ratio of the number of patients who have developed a recurrence over the specified period of time, to the total person-time at risk for recurrence observed in the sample.

Risk and rate highlight different aspects of disease occurrence. Risk is the probability of an event of interest, whereas rate reflects the force, or rapidity of an event of interest. Risk and rate are two fundamentally distinct concepts, and the choice of their use should be driven by the scientific questions.

3.3 The relationship between risk and rate in survival analysis

In survival analysis, *hazard rate* is often used as the quantity to characterize the event process. Hazard rate is a mathematical concept referring to the "instantaneous rate" of an event *at* a certain time point among those who survived to that time. Mathematically, it refers to the theoretical limit approached by an incidence rate as its time interval is narrowed toward zero [12].

The concept that corresponds to risk is *cumulative incidence*, which is a measure of the average risk of developing an event of interest in a population in a specified time interval [12]. Cumulative incidence depends on both the hazard rate of dying (mortality rate) and the hazard rate of the event of interest.

When there is no competing risk, "surviving" to and being at risk for the event of interest at a certain time is defined by not experiencing the event of interest before that time. Therefore, the hazard rate is the only determinant of the cumulative incidence of the event of interest. This means if a factor is associated with a higher hazard rate, then it is also associated with higher cumulative incidence.

When competing risks exist, each individual has the possibility of experiencing an event of interest, a competing-risk event or censoring. Survival to a given time not only depends on the hazard of the event of interest, but also on the hazard of competing risk events. These hazard rates in the context of competing risks are referred to as *cause-specific hazard rates*. The cause-specific hazard rate represents the hazard rate of occurrence of a specific type of event

indexed by its type ("cause"), i.e., event of interest, or competing-risk event, at a given time among "survivors" to that time, i.e., subjects who have not experienced the event of interest and have not experienced competing-risk events. Those who have experienced either event can no longer develop the event of interest; therefore they will be removed from the risk sets for the event of interest following their experience of either event.

In competing risk settings, since cumulative incidence depends on the cause-specific hazard rate of competing-risk events as well as the cause-specific hazard rate of the event of interest, it is *no longer* defined sorely by the (cause-specific) hazard for the event of interest.

Many researchers have suggested that in competing-risk settings, the cause-specific hazard rate of the event of interest is the fundamental mechanistic measure of association which provides a better understanding of the underlying event process of interest, whereas cumulative incidence is more descriptive [5] and focuses on what actually happens in the presence of competing risks.

Using the example in the previous section as illustration, suppose treatment A and treatment B both have no effect on the hazard of breast cancer recurrence, which means the cause-specific hazard rates of breast cancer recurrence are similar in both groups. However, if treatment A can increase the cause-specific hazard rate of death (competing-risk event), relative to treatment B, then we will have a lower chance to observe breast cancer recurrence in Group A. This is because that more Group-A subjects are expected to die due to

the higher cause-specific hazard rate of death in Group A, which reduces their chance of developing breast cancer recurrence, relatively to Group-B subjects.

3.4 Risk and Rate regression models

This section will discuss two commonly-used regression approaches that can be used to evaluate the impact of factors on risk and rate of event of interest.

3.4.1 Cox proportional hazards model

The Cox proportional hazard model has become a widely used procedure for modeling the relationship of covariates to event occurrence in terms of the event's hazard rate. In the presence of competing risks, Cox proportional hazard model can be constructed for the cause-specific hazard rate of event of interest. The standard form of Cox proportional hazards model to cause-specific hazards is

$$h_j(t, X) = h_{j0}(t)\exp\{X'\beta_j\} \quad \text{(Eq. 1)}$$

X is a vector of covariates, $h_{j0}(t)$ is the baseline cause-specific hazard rate for event type j, and $\exp\{X'\beta_j\}$ is the relative change in the cause-specific hazard rate corresponding to a one-unit change in a corresponding covariate. Note that the baseline cause-specific hazard may vary with time, but the regression coefficient of a covariate is the log hazard ratio for event type j associated with a unit change of the covariate, adjusting for the other covariates, regardless of the time at which it is computed.

Since the cause-specific hazard is measured, individuals at risk for event j are those who have not experienced event j and have not experienced competing-risk events. With respect to the cause-specific hazard rate modeling, competing risk events as well as censoring end the at-risk status for observing an event of interest: thus, competing-risk events are treated the same as censoring [13]. Note that this handling of competing-risk events is proper in making statistical inference on the cause-specific hazard rate of the event of interest. It should not be regarded as the inability of Cox proportional hazards models to incorporate competing risks: this is a common misconception.

3.4.2 Fine and Gray subdistribution proportional hazards model

When competing risks exist, the cause-specific hazard rate of the event of interest is not the only determinant of the cumulative incidence of that event. Fine and Gray [8] introduced a regression approach that models the hazard-rate-like quantity for the cumulative incidence of the event of interest. That is, Fine and Gray modeled *risk* (cumulative incidence), not *rate* (cause-specific rate), with explanatory variables. They defined "subdistribution hazard" by:

$$h_j^*(t) = \lim_{\Delta t \to 0} \Pr\{T \in [t, t+\Delta t] \; and \; J = j | T \geq t \; or \; (T < t \; and \; J \neq j)\} / \Delta t \quad \text{(Eq. 2)}$$

where T is time to the first failure which measured from time zero, J is the indicator for event type. Mathematically, however, this is not a proper hazard function because its corresponding cumulative distribution function can never

reach 1 even when t increases. Thus, interpreting this "subdistribution hazard" as *rate* is not appropriate: it is a mathematical quantity useful for modeling cumulative incidence of the event of interest in competing-risk settings.

The standard form for Fine and Gray subdistribution proportional hazards model is

$$h_j^*(t, X) = h_{j0}^*(t)\exp\{X'\beta_j\} \quad \text{(Eq. 3)}$$

It is in the same proportional hazards form as Cox's model but it is for the subdistribution hazard associated with an event of interest. $h_{j0}^*(t)$ is the baseline subdistribution hazard for event j and $\exp\{X'\beta_j\}$ is the relative subdistribution hazard associated with covariates.

Note that the risk set for the subdistribution hazard contains both the individuals who have survived without any type of event to time t and those who have had a competing risk event. For cause-specific hazard, the risk set decreases at each time point at which there is a competing-risk event, whereas, the risk set for subdistribution hazard maintains the individuals who fail from competing risk events, even though they will never develop the event of interest. Fine and Gray noted that the risk set associated with the subdistribution hazard is "unnatural" [8], however, one can think of these individuals as an observed "placeholder" for the proportion of the population that cannot have the event of interest [14].

3.5 An example in practice

We use CCSS, a large retrospective cohort study designed to investigate long-term effects of cancer and therapy, among 5-year survivors of childhood cancer.

The CCSS cohort consists of 5-year survivors of childhood cancer diagnosed before the age of 21 years during the period from January 1, 1970 through December 31, 1986 in one of 26 collaborating pediatric oncology centers. A detailed description of the CCSS study design and cohort characteristics has been published previously [15].

Many reports show that childhood cancer survivors are at increased risk for developing second and subsequent malignant neoplasms [16, 17]. Radiation therapy has been shown to increase the risk of SMNs [18-20]. Other factors, such as type of childhood cancer diagnosis and age at diagnosis, have been reported to be associated with the occurrence of SMNs [16]. In the following analysis, we focus on SMN: the third and subsequent malignant neoplasms are not included in our analysis. Death from any cause was treated as a competing-risk event and survivors were censored at the date of last contact.

Of the total 13,225 childhood cancer survivors in the study cohort, 807 experienced an SMN event, 1,005 died (had a competing-risk event), and 11,413 were censored at the end of follow-up. The characteristics of the cohort are shown in Table 3-1.

We analyzed this cohort by fitting both the Cox cause-specific proportional hazards model and the Fine and Gray subdistribution proportional hazards model. The explanatory variables were selected based on prior knowledge: we did not use any statistical variable selection approach. Table 3-2 displays the estimated regression parameter and its confidence interval

associated with each explanatory variable on the cause-specific and subdistribution hazards of SMN.

By comparing the estimated association between each covariate and SMN incidence, the cause-specific hazard ratio and the subdistribution hazard ratio differ depending on how the covariate is associated with the competing risk (mortality). Our result in Table 3-2 presented three different kinds of scenarios (Table 3-3):

Scenario 1: the cause-specific hazard ratio and the subdistribution hazard ratio are nearly identical. These covariates are not associated with mortality, as the cause-specific hazard ratios for death are close to 1.0 and not statistically significant different from 1.0.

Scenario 2: the cause-specific hazard ratio is larger than the subdistribution hazard ratio. In this setting, the association between the covariate and the cause-specific hazard is stronger than the association between the covariate and the subdistribution hazard. These covariates are associated with an *increase* of the cause-specific hazard for SMN, and at the same time, are associated with an *increase* of the cause-specific hazard rate for death. The increased incidence of competing-risk event prevents the subjects from developing the event of interest. Therefore, we would observe a smaller subdistribution hazard ratio compared to the scenario where the covariates are not associated with competing risk.

Scenario 3: the cause-specific hazard ratio is smaller than the subdistribution hazard ratio. In this setting, the association between the covariate

and the cause-specific hazard is weaker than the association between the covariate and the subdistribution hazard. These covariates are associated with an *increase* of cause-specific hazard for SMN, and at the same time, are associated with a *decrease* of cause-specific hazard rate for death. Therefore, we would observe a greater subdistribution hazard ratio compared to the scenario where the covariates are not associated with competing risk. With the same baseline hazard rate of death, we would expect to observe a larger difference between the estimated cause-specific hazard ratio and subdistribution hazard ratio, if the covariate is associated with a larger decrease of cause-specific hazard rate for death.

3.6 Discussion

Risk and rate are two important concepts that have been frequently used in many epidemiologic studies. In this paper, we reviewed and compared the concepts of risk and rate under a hypothetical cohort study. We further explained the relationship between risk and rate in survival analysis framework. Two commonly-used modeling approaches for estimating the association between factors on risk and rate were discussed. A real data example from CCSS was used to illustrate the utility of the two regression models and the differences in terms of results interpretation.

Both Cox proportional hazards models on cause-specific hazards and Fine and Gray's subdistirbution proportional hazards models can be extended to introduce time-dependent covariates. Software to fit these models has been

written in R. Scrucca *et al.* [21] provided an easy guide for analysts on how to perform a competing risk analysis in R, including cumulative incidence function estimation, testing for equality across subgroups, and computing point-wise confidence intervals.

Cox proportional hazards models give an estimate for hazard ratio, whereas Fine and Gray's proportional subdistribution proportional hazards models give an estimate for subdistribution hazard ratio. As what we showed in the CCSS data analysis, covariates could have different ways of being associated with the event rate and event cumulative incidence. The choice of the model involves a number of theoretical and practical considerations. It has been suggested that the knowledge related to cumulative incidence is central to cost-effectiveness analyses, and it may be useful for policy decisions making [8]. However, it may be more appropriate to use cause-specific hazard to evaluate treatment effect.

Additional attention should be paid when one generalizes the results of the Fine and Gray's subdistribution proportional hazard model. Structures of competing risks influence the cumulative incidence of the event of interest. Therefore, subdistribution hazard ratio estimates cannot be generalized to populations with different competing risks [8]. In contrast, cause-specific hazard ratio estimates can be generalized to populations with similar characteristics regardless of competing risks.

Table 3-1 Characteristics of the Childhood Cancer Survivor Study cohort by second malignant neoplasm (SMN)

Characteristic	Overall cohort (n=13225)	Cases without SMN (n=12418)	Cases with SMN (n=807)
Mean age at diagnosis of childhood cancer (median)	8.3 (6.8)	8.1 (6.5)	11.2 (12.2)
Female N (%)	6226 (47.1)	5733 (46.2)	493 (61.1)
Race N (%)			
Black	517 (3.9)	497 (4.0)	20 (2.5)
White	11620 (87.9)	10883 (87.6)	737 (91.3)
Hispanic/Latin	634 (4.8)	600 (4.8)	34 (4.2)
Other	454 (3.4)	438 (3.5)	16 (2.0)
Childhood cancer disgnosis N (%)			
Acute lymphoblastic leukemia	4036 (30.5)	3890 (31.3)	146 (18.1)
Astrocytomas	1096 (8.3)	1057 (8.5)	39 (4.8)
Hodgkins disease	1750 (13.2)	1483 (11.9)	267 (33.1)
Medulloblastoma, PNET	363 (2.7)	343 (2.8)	20 (2.5)
Non-Hodgkins lymphoma	973 (7.4)	916 (7.4)	57 (7.1)
Other bone tumors	50 (0.4)	47 (0.4)	3 (0.4)
Other leukemia	131 (1.0)	121 (1.0)	10 (1.2)
Acute myeloid leukemia	332 (2.5)	314 (2.5)	18 (2.2)
Ewings sarcoma	380 (2.9)	343 (2.8)	37 (4.6)
Kidney tumors	1141 (8.6)	1110 (8.9)	31 (3.8)
Neuroblastoma	880 (6.7)	843 (6.8)	37 (4.6)
Osteosarcoma	667 (5.0)	625 (5.0)	42 (5.2)
Other CNS tumors	281 (2.1)	267 (2.2)	14 (1.7)
Soft tissue sarcoma	1145 (8.7)	1059 (8.5)	86 (10.7)
Childhood cancer radiation therapy N (%)	8504 (64.3)	7858 (63.3)	646 (80.0)
Alkylating agent score N (%)			
0	5805 (43.9)	5507 (44.3)	298 (36.9)
1	2503 (18.9)	2374 (19.1)	129 (16.0)
2	1655 (12.5)	1528 (12.3)	127 (15.7)
3	1123 (8.5)	1010 (8.1)	113 (14.0)
Treatment era N (%)			
1970-1974	2336 (17.7)	2097 (16.9)	239 (29.6)
1975-1979	3718 (28.1)	3457 (27.8)	261 (32.3)
1980-1986	7171 (54.2)	6864 (55.3)	307 (38.0)
Splenectomy N (%)	692 (0.3)	37 (0.3)	5 (0.6)
Anthracycline exposure, mg/m^2			
None	7583 (57.3)	7102 (57.2)	481 (59.6)
1-100	480 (3.6)	466 (3.8)	14 (1.7)
101-300	1862 (14.1)	1752 (14.1)	110 (13.6)
301+	2100 (15.9)	1963 (15.8)	137 (17.0)
Epipodophyllotoxin exposure, mg/m^2			
None	11570 (87.5)	10838 (87.3)	732 (90.7)
1-1000	280 (2.1)	271 (2.2)	9 (1.1)
1001-4000	245 (1.9)	232 (1.9)	13 (1.6)

4001+	283 (2.1)	267 (2.2)	16 (2.0)
Platinum exposure, mg/m^2			
None	11875 (89.8)	11130 (89.6)	745 (92.3)
1-400	257 (1.9)	249 (2.0)	8 (1.0)
401-750	234 (1.8)	220(1.8)	14 (1.7)
751+	87 (0.7)	80 (0.6)	7 (0.9)

Table 3-2 Estimated regression parameters and confidence intervals associated with explanatory variables on the cause-specific and subdistribution hazards of second malignant neoplasm (SMN)

		Cause-specific hazard ratio (95%CI[+])	Subdistribution hazard ratio (95%CI[+])
Childhood cancer radiation therapy	No	Reference	
	Yes	2.03 (1.60 to 2.57)*	2.00 (1.59 to 2.53)*
Sex	Male	Reference	
	Female	1.83 (1.56 to 2.15)*	1.86 (1.59 to 2.19)*
Age at diagnosis	0-4	Reference	
	5-9	1.19 (0.92 to 1.54)	1.17 (0.90 to 1.52)
	10-14	1.70 (1.32 to 2.21)*	1.67 (1.29 to 2.16)*
	15+	2.09 (1.59 to 2.75)*	2.02 (1.54 to 2.64)*
Treatment era	1970-1974	Reference	
	1975-1979	0.82 (0.66 to 1.01)	0.83 (0.67 to 1.01)
	1980-1986	0.86 (0.69 to 1.07)	0.86 (0.70 to 1.07)
Race	Black	Reference	
	White	1.11 (0.69 to 1.78)	1.10 (0.69 to 1.77)
	Hispanic/Latin	1.20 (0.66 to 2.19)	1.17 (0.64 to 2.12)
	Other	0.84 (0.41 to 1.72)	0.82 (0.40 to 1.69)
Childhood cancer diagnosis	Acute lymphoblastic leukemia	Reference	
	Astrocytomas	1.11 (0.74 to 1.66)	1.05 (0.70 to 1.57)
	Hodgkins disease	2.49 (1.86 to 3.33)*	2.51 (1.88 to 3.36)*
	Medulloblastoma, PNET	1.69 (1.00 to 2.83)*	1.62 (0.97 to 2.72)
	Non-Hodgkins lymphoma	1.00 (0.68 to 1.47)	1.03 (0.70 to 1.52)
	Other bone tumors	1.75 (0.55 to 5.59)	1.80 (0.55 to 5.87)
	Other leukemia	2.87 (1.45 to 5.68)*	2.57 (1.29 to 5.16)*
	Acute myeloid leukemia	0.96 (0.51 to 1.81)	0.95 (0.51 to 1.78)
	Ewings sarcoma	1.61 (1.01 to 2.56)*	1.54 (0.97 to 2.42)
	Kidney tumors	0.84 (0.56 to 1.29)	0.85 (0.55 to 1.30)
	Neuroblastoma	1.41 (0.94 to 2.14)	1.40 (0.92 to 2.15)
	Osteosarcoma	1.44 (0.91 to 2.29)	1.51 (0.95 to 2.39)
	Other CNS tumors	1.29 (0.67 to 2.50)	1.19 (0.61 to 2.32)
	Soft tissue sarcoma	1.62 (1.18 to 2.22)*	1.61 (1.17 to 2.21)*
Splenectomy	No	Reference	
	Yes	0.74 (0.24 to 2.31)	0.74 (0.24 to 2.25)
Alkylating agent score	0	Reference	
	1	1.14 (0.90 to 1.45)	1.14 (0.90 to 1.44)
	2	1.25 (0.97 to 1.60)	1.26 (0.96 to 1.62)
	3	1.22 (0.95 to 1.56)	1.16 (0.91 to 1.49)
Anthracycline exposure, mg/m^2	None	Reference	
	1-100	0.70 (0.38 to 1.27)	0.67 (0.36 to 1.24)
	101-300	1.31 (1.01 to 1.69)*	1.28 (0.99 to 1.66)

	301+	1.41 (1.09 to 1.82)*	1.37 (1.07 to 1.75)*
Epipodophyllotoxin exposure, mg/m^2	None	Reference	
	1-1000	0.89 (0.43 to 1.85)	0.83 (0.40 to 1.73)
	1001-4000	1.15 (0.62 to 2.15)	1.03 (0.55 to 1.96)
	4001+	1.96 (1.15 to 3.32)*	1.87 (1.07 to 3.27)*
Platinum exposure, mg/m^2	None	Reference	
	1-400	0.68 (0.31 to 1.49)	0.63 (0.29 to 1.37)
	401-750	1.55 (0.86 to 2.78)	1.49 (0.82 to 2.70)
	751+	2.36 (1.08 to 5.15)*	1.97 (0.88 to 4.41)

Note: *Statistically significantly different from 1.0 with p-value < 0.05 [+]Confidence interval.

Table 3-3 Comparison of the effects of estimated associations with explanatory variables on the cause-specific and subdistribution hazard ratios, stratified into three scenarios

	Cause-specific hazard ratio for **SMN** (95%CI)	Cause-specific hazard ratio for **Death** (95%CI)	Subdistribution hazard ratio for **SMN** (95%CI)
Scenario 1			
Childhood cancer diagnosis: Hodgkins disease	2.49 (1.86 to 3.33)	0.77 (0.52 to 1.15)	2.51 (1.88 to 3.36)
Childhood cancer diagnosis: Soft tissue sarcoma	1.62 (1.18 to 2.22)	1.11 (0.76 to 1.63)	1.61 (1.17 to 2.21)
Scenario 2			
Platinum exposure, mg/m^2: 751+	2.36 (1.08 to 5.15)	2.74 (1.34 to 5.60)	1.97 (0.88 to 4.41)
Scenario 3			
Sex	1.83 (1.56 to 2.15)	0.69 (0.57 to 0.83)	1.86 (1.59 to 2.19)

3.7 References

[1] Farr, W., 2003, "" On Prognosis" by William Farr (British Medical Almanack 1838; Supplement 199–216)," Sozial-Und Präventivmedizin/Social and Preventive Medicine, **48**(4) pp. 219-224.

[2] MORGENSTERN, H., KLEINBAUM, D. G., and KUPPER, L. L., 1980, "Measures of Disease Incidence used in Epidemiologic Research," International Journal of Epidemiology, **9**(1) pp. 97-104.

[3] Gooley, T. A., Leisenring, W., Crowley, J., 1999, "Estimation of Failure Probabilities in the Presence of Competing Risks: New Representations of Old Estimators," Statistics in Medicine, **18**(6) pp. 695-706.

[4] Satagopan, J., Ben-Porat, L., Berwick, M., 2004, "A Note on Competing Risks in Survival Data Analysis," British Journal of Cancer, **91**(7) pp. 1229-1235.

[5] Varadhan, R., Weiss, C. O., Segal, J. B., 2010, "Evaluating Health Outcomes in the Presence of Competing Risks: A Review of Statistical Methods and Clinical Applications," Medical Care, **48**(6) pp. S96-S105.

[6] Gerds, T. A., Scheike, T. H., and Andersen, P. K., 2012, "Absolute Risk Regression for Competing Risks: Interpretation, Link Functions, and Prediction," Statistics in Medicine, **31**(29) pp. 3921-3930.

[7] Klein, J. P., and Andersen, P. K., 2005, "Regression Modeling of Competing Risks Data Based on Pseudovalues of the Cumulative Incidence Function," Biometrics, **61**(1) pp. 223-229.

[8] Fine, J. P., and Gray, R. J., 1999, "A Proportional Hazards Model for the Subdistribution of a Competing Risk," Journal of the American Statistical Association," **94**(446) pp. 496-509.

[9] Wolbers, M., Koller, M. T., Witteman, J. C., & Steyerberg, E. W., 2009, "Prognostic models with competing risks: methods and application to coronary risk prediction," Epidemiology, **20**(4), 555-561.

[10] Kim, H. T., 2007, "Cumulative incidence in competing risks data and competing risks regression analysis," Clinical Cancer Research, **13**(2), 559-565.

[11] Lim, H., Zhang, X., Dyck, R., & Osgood, N., 2010, "Methods of competing risks analysis of end-stage renal disease and mortality among people with diabetes," BMC medical research methodology, **10**(1), 97.

[12] Rothman, K.J., Greenland, S., and Lash, T.L., 2008, "Modern epidemiology," Wolters Kluwer Health, .

[13] Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., 1978, "The Analysis of Failure Times in the Presence of Competing Risks," Biometrics, pp. 541-554.

[14] Lau, B., Cole, S. R., and Gange, S. J., 2009, "Competing Risk Regression Models for Epidemiologic Data," American Journal of Epidemiology, **170**(2) pp. 244-256.

[15] Robison, L. L., Mertens, A. C., Boice, J. D., 2002, "Study Design and Cohort Characteristics of the Childhood Cancer Survivor Study: A Multi-institutional Collaborative Project," Medical and Pediatric Oncology, **38**(4) pp. 229-239.

[16] Neglia, J. P., Friedman, D. L., Yasui, Y., 2001, "Second Malignant Neoplasms in Five-Year Survivors of Childhood Cancer: Childhood Cancer Survivor Study," Journal of the National Cancer Institute, **93**(8) pp. 618-629.

[17] Meadows, A. T., Baum, E., Fossati-Bellani, F., 1985, "Second Malignant Neoplasms in Children: An Update from the Late Effects Study Group." Journal of Clinical Oncology, **3**(4) pp. 532-538.

[18] DeLaat, C. A., and Lampkin, B., 1992, "Long-term Survivors of Childhood Cancer: Evaluation and Identificationof Sequelae of Treatment," CA: A Cancer Journal for Clinicians, **42**(5) pp. 263-282.

[19] Robison, L. L., 1996, "Second Primary Cancers After Childhood Cancer." BMJ: British Medical Journal, **312**(7035) pp. 861.

[20] Friedman, D. L., Whitton, J., Leisenring, W., 2010, "Subsequent Neoplasms in 5-Year Survivors of Childhood Cancer: The Childhood Cancer Survivor Study," Journal of the National Cancer Institute, **102**(14) pp. 1083-1095.

[21] Scrucca, L., Santucci, A., and Aversa, F., 2007, "Competing Risk Analysis using R: An Easy Guide for Clinicians," Bone Marrow Transplantation, **40**(4) pp. 381-387.

Chapter 4: Conclusion and discussions

In this thesis, I have mainly focused on the following two questions in survival analysis:

1. How to estimate the burden of recurrent events in a population in the presence of competing risks?
2. How to introduce covariates in the context of competing risks to assess the association of risk factors with event occurrence in terms of risk and rate?

In Chapter 2, we proposed a novel statistical approach, termed cumulative total incidence, which can directly capture the burden of recurrent events in a population by a given time in the presence of competing risks. We further mathematically proved and empirically showed the equivalence between the CTI and the sum of CI for incremental numbers of events in the population.

Using the data from the CCSS with subsequent neoplasms as the event of interest, the result of CTI approach suggested an approximately 3.5 times higher number of subsequent neoplasms for the survivors who received radiation therapy treatment compared with those who did not receive radiation therapy treatment at 39 years since diagnosis. This is informative for survivors' follow-up care or guideline/policy making as it quantifies healthcare service needs/utilizations. The cumulative incidence included only the first occurrence of subsequent neoplasm for each survivor and reflected that the probabilities of developing at least one subsequent neoplasm at 39 years since diagnosis was 2.6 times higher for the survivors who received radiation therapy than those who did

not. By comparing the results of the CTI approach and the CI approach stratified by treatment status, one could observe that cumulative incidence analysis that incorporates only the first occurrence of event would underestimate the total burden more severely with longer follow-up time and more frequent recurrence. Therefore, when analyzing data with recurrent events, if a summary of the total number of events occurring in a population is of primary interest, the proposed CTI approach would be useful.

In Chapter 3, we reviewed the concepts of risk and rate and linked the relationship between risk and rate in survival analysis regression framework. By comparing these two fundamentally distinct but mechanically related concepts, we highlighted that risk and rate reflects different aspects of disease occurrence, and the choice of either of them should be driven by the scientific questions.

Cox proportional hazards models on cause-specific hazards and Fine and Gray's regression approach on subdistribution hazards were discussed. A real data example from the CCSS with second malignant neoplasm and death as competing-risk outcomes was used to illustrate the utility of the two regression approaches. The results reveal that the difference between the cause-specific hazard ratio and the subdistribution hazard ratio was dependent on how the covariate is associated with competing-risk event (death). In competing-risk settings, the cumulative incidence for a specific type of event is not determined solely by the cause-specific hazard for that event; it is also influenced by all cause-specific hazards of competing-risk events.

We suggest that future work could include the development of statistical inference methods with CTI approach, including significance test and confidence intervals. It is also of interest to consider regression methods for total cumulative incidence. Statistical software package can be developed to conduct these analyses. The CTI approach could also be examined with other recurrent event.

Appendices

Appendix 1: Mathematical proof of Equation (8) in Chapter 2

The following content shows the mathematical proof of Equation (8).

We assume there are n_0 individuals initially at risk in the study. Each individual could experience three distinct kinds of events at time t_j during follow up: (1) occurrence of the event of interest; (2) occurrence of a competing-risk event; and (3) censoring.

The times at which any of the three events occurs can be ordered as $t_1 \le t_2 \le \ldots \le t_n$.

Individuals could experience the event of interest, (1), multiple times and remain in the study. However, they can only experience outcomes (2) or (3) once. We further define the following:

e_j: The number of events of interest at time t_j (first-ever or recurrent);

r_j: The number of individuals who experience a competing-risk event at time t_j;

c_j: The number of individuals who are censored at time t_j;

n_j: The number of individuals who are at risk and under observation of the study *beyond* time t_j.

If we assume an individual can experience at most m times of the recurrent event in the study, for $p^{th}(p = 1,2,\mathrm{K}\ ,m)$ occurrence of the event of interest:

e_{pj}: The number of individuals who experience the event of interest at time t_j;

r_{pj}: The number of individuals who experience a competing-risk event at time t_j;

c_{pj}: The number of individuals who are censored at time t_j;

n_{pj}: The number of individuals who are under study *beyond* time t_j.

For calculating $CI_p(t)$, we only treat the p^{th} occurrence of event as the event of interest. The population at risk for the p^{th} occurrence of event would consist of those individuals who have had the $(p-1)^{th}$ or less occurrence of the event of interest, and who are not censored or experience competing-risk event before p^{th} occurrence of the event of interest. After having the p^{th} event, individuals would leave the population at risk for the p^{th} occurrence of event.

Since there are n_0 individuals initially at risk, we have $n_{p0} = n_0$ and

$$n_{pj} = n_0 - \sum_{k=1}^{j} (e_{pk} + r_{pk} + c_{pk}) \qquad \text{(Eq. 9)}$$

Note that e_j is the number of events of interest by time t_j, regardless of whether it was the first occurrence or not; therefore, we have $e_j = \sum_{p=1}^{m} e_{pj}$.

Now, we want to mathematically prove that the CTI at time *t* is equivalent to the sum of $CI_p(t)$'s, such that:

$$CTI(t) = \sum_{p=1}^{m} CI_p(t)$$

Here $CI_p(t)$ represents the CI for the $p^{th}(p=1,2,\text{K},m)$ occurrence of the event of interest by time t.

Based on the formula for calculating $CTI(t)$ and $CI(t)$, we have

$$CTI(t)=\begin{cases}0 & if\ s=0\\ \dfrac{e_1}{n_0} & if\ s=1\text{, and}\\ \dfrac{e_1}{n_0}+\sum\limits_{j=2}^{s}\dfrac{e_j}{n_{j\text{-}1}}\prod\limits_{k=1}^{j\text{-}1}(1-\dfrac{r_k}{n_{k\text{-}1}}) & if\ s\geq 2\end{cases}$$

$$\sum_{p=1}^{m}CI_p(t)=\begin{cases}0 & if\ s=0\\ \sum\limits_{p=1}^{m}\dfrac{e_{p1}}{n_{p0}}=\dfrac{e_{11}}{n_{10}} & if\ s=1\\ \sum\limits_{p=1}^{m}\dfrac{e_{p1}}{n_{p0}}+\sum\limits_{p=1}^{m}\sum\limits_{j=2}^{s}\dfrac{e_{pj}}{n_{p(j\text{-}1)}}\prod\limits_{k=1}^{j\text{-}1}(1-\dfrac{r_{pk}}{n_{p(k\text{-}1)}})\prod\limits_{k=1}^{j\text{-}1}(1-\dfrac{e_{pk}}{n_{p(k\text{-}1)}}) & if\ s\geq 2\end{cases}$$

where s is the largest j such that $t_j<t$.

Mathematical Induction is applied for the following proof.

Basis step:

When $s=1$, since $e_{11}=e_1$ and $n_{10}=n_0$, we have $CTI(t)=\sum\limits_{p=1}^{m}CI_p(t)$.

Inductive step:

We assume the equation holds for $s=i$, which means $CTI(t_i)=\sum\limits_{p=1}^{m}CI_p(t_i)$. For $s=i+1$, we have

$$CTI(t_{i+1})=CTI(t_i)+\frac{e_{i+1}}{n_i}\prod_{k=1}^{i}(1-\frac{r_k}{n_{k-1}})$$

$$\sum_{p=1}^{m} CI(t_{i+1}) = \sum_{p=1}^{m} CI(t_i) + \sum_{p=1}^{m} \frac{e_{p(i+1)}}{n_{pi}} \prod_{k=1}^{i} (1 - \frac{r_{pk}}{n_{p(k-1)}}) \prod_{k=1}^{i} (1 - \frac{e_{pk}}{n_{p(k-1)}})$$

By definition,

$$n_{pi} = n_i \times P(\textit{being at risk for } p^{th} \textit{ event at } t = i \,|\, \textit{being at risk for an event at } t = i)$$

The above conditional probability can be calculated by

$$\frac{\prod_{k=1}^{i} (1 - \frac{r_{pk}}{n_{p(k-1)}}) \prod_{k=1}^{i} (1 - \frac{e_{pk}}{n_{p(k-1)}})}{\prod_{k=1}^{i} (1 - \frac{r_k}{n_{k-1}})}$$

Because $e_{i+1} = \sum_{p=1}^{m} e_{p(i+1)}$ and $CTI(t_i) = \sum_{p=1}^{m} CI_p(t_i)$, we can show that

$CTI(t_{i+1}) = \sum_{p=1}^{m} CI_p(t_{i+1})$.

To sum up, we conclude that the CTI is equivalent to the sum of the CIs for incremental numbers of events in the population.

Appendix 2: R code for CCSS data analysis in Chapter 2

```
library(etm)


#####Basic Information from Original Data#####

#read data#
SN=read.csv("Z:/Common/Huiru/Cumulative Total Incidence/SN.csv",h=T)

#the overall number of patients in the study#
#12630 total patients#
length(table(SN$ccssid))

#get only event data, regardless of the group id#
event_data_SN=SN[SN$sn==1,]

#the number of patients who experienced event (first-ever or recurrence)#
#1448 patients#
length(table(event_data_SN$ccssid))

#total number of events#
#2391 total events#
nrow(event_data_SN)

#check the distribution of the number of events#
table(table(event_data_SN$ccssid))



#####Data Manipulation#####

#format study start date#
d_start=as.Date(SN$d_start, "%d%b%Y")

#format study stop date#
d_stop=as.Date(SN$d_stop, "%d%b%Y")

#format patient death date, the competing risks event date#
d_death=as.Date(SN$D_DEATH,"%d%b%Y")

#format SN event date#
sn_date=as.Date(SN$sn_date,"%d%b%Y")

#id#
ccssid=SN$ccssid

#group indicator#
rad=SN$rad_yn

#change the dataset to long format#
#event=0: censoring#
#event=1: event of interest#
#event=2: competing risks event#
id=NULL;entry=NULL;exit=NULL;group=NULL;event=NULL
```

```
for (i in 1:nrow(SN))
{

#No event date or death date#
#Means censoring individuals#
 if (is.na(d_death[i]) & is.na(sn_date[i]))
   { id=c(id,ccssid[i])
    entry=c(entry,as.character(d_start[i]))
    group=c(group,rad[i])
    exit=c(exit,as.character(d_stop[i]))          #date=stop following up date#
    event=c(event, 0)                             #censoring is coded as 0#
    }

#Have event date but no death date#
#Means event of interest individuals#
 if (is.na(d_death[i]) & !is.na(sn_date[i]))
    { id=c(id,ccssid[i])
     entry=c(entry,as.character(d_start[i]))
     group=c(group,rad[i])
     exit=c(exit,as.character(sn_date[i]))        #date=event of interest date#
     event=c(event,1)                             #event of interest is coded as 1#

     if (d_stop[i]>=sn_date[i]) #event individual be censored at following up end date#
       { id=c(id,ccssid[i])
        entry=c(entry,as.character(d_start[i]))
        group=c(group,rad[i])
        exit=c(exit,as.character(d_stop[i]))      #date=stop following up date#
        event=c(event,0)                          #censoring is coded as 0#
       }
    }

#Have death date but no event date#
#Means competing risk event individuals#
 if (!is.na(d_death[i]) & is.na(sn_date[i]))
    { if (d_death[i]>d_stop[i]) #Death happened later, then treat as censoring#
       { id=c(id,ccssid[i])
        entry=c(entry,as.character(d_start[i]))
        group=c(group,rad[i])
        exit=c(exit,as.character(d_stop[i]))      #date=stop following up date#
        event=c(event,0)                          #censoring is coded as 0#
       }

     if (d_death[i]<=d_stop[i]) #Death happened earlier#
       { id=c(id,ccssid[i])
        entry=c(entry,as.character(d_start[i]))
        group=c(group,rad[i])
        exit=c(exit,as.character(d_death[i]))     #date=competing risk event date#
        event=c(event,2)                          #competing risk event is coded as 2#
       }
    }

#Have death date and also event date#
 if (!is.na(d_death[i]) & !is.na(sn_date[i]))
    { id=c(id,ccssid[i])
```

```
      entry=c(entry,as.character(d_start[i]))
      group=c(group,rad[i])
      exit=c(exit,as.character(sn_date[i]))          #recode the event date first#
      event=c(event,1)                              #event of interest is coded as 1#

      if (d_death[i]>d_stop[i]) #Death happened later, then treat as censoring#
        { id=c(id,ccssid[i])
          entry=c(entry,as.character(d_start[i]))
          group=c(group,rad[i])
          exit=c(exit,as.character(d_stop[i]))       #date=stop following up date#
          event=c(event,0)                          #censoring is coded as 0#
        }

      if (d_death[i]<=d_stop[i]) #Death happened earlier#
        { id=c(id,ccssid[i])
          entry=c(entry,as.character(d_start[i]))
          group=c(group,rad[i])
          exit=c(exit,as.character(d_death[i]))      #date=competing risk event date#
          event=c(event,2)                          #competing risk event is coded as 2#
        }
      }
}

data.SN=data.frame(id=id, group=group,entry=entry,exit=exit,event=event)
#remove the duplicated rows#
data_SN=data.SN[!duplicated(data.SN),]
#order the dataset by id and event order#
ii=order(data_SN$id,data_SN$exit)
dataSN=data_SN[ii,]
write.csv(dataSN,"Z:/Common/Huiru/Cumulative Total
 Incidence/dataSN.csv",row.names=FALSE)



#####Prepare Data to Calculate CTI#####

dataSN=read.csv("Z:/Common/Huiru/Cumulative Total Incidence/dataSN.csv",h=T)
#the data is already sorted by id and exit time#
#calculate the follow up time#
dif=as.numeric(difftime(as.Date(dataSN$exit,"%Y-%m-%d"),as.Date(dataSN$entry,"%Y-%m-
 %d"),unit="days"))

#Total number of subjects#
index=table(dataSN$id)
N=length(index)

event.number=NULL
for (i in 1:N)
{
 if (index[i]==1)
   { event.number=c(event.number,0)}
 if (index[i]!=1)
   { event.number=c(event.number,1:index[i])}
}
```

```
#first happened is censoring or competing risk event, will all be coded as 0#
#this is because no matter which time of occurrence is of interest, we need to include them#
#we need to always include event.number==0 subjects for calculating any occurrence of event#

SN.CTI=data.frame(dataSN,dif=dif,event.number=event.number)

#There are total M times of the events that we need to calculate the CI#
M=max(SN.CTI$event.number)

#calculate for each id, what is the max number of event#
id.table=SN.CTI$id[!duplicated(SN.CTI$id)]
max.id=NULL
for (j in 1:N)
{
 max.id[j]=max(SN.CTI$event.number[SN.CTI$id==id.table[j]])
}
freq=match(SN.CTI$id, id.table)
SN.CTI=data.frame(SN.CTI,max_e=max.id[freq])

#generate full dataset for calculating mth event of CI#
data.CTI=NULL
for (i in 1:M)
{
 x=SN.CTI[((SN.CTI$max_e>i)&(SN.CTI$event.number==i))|
 ((SN.CTI$max_e<=i)&(SN.CTI$event.number==SN.CTI$max_e)),]
 data.CTI=rbind(data.CTI,cbind(x,m_event=rep(i,nrow(x))))
}

#This dataset contains the data needed for each time of the occurrence#
#There are some patients have event earlier/on than the start of the study#
aa=c(data.CTI[data.CTI$dif<=0,1],19052013)
#we also need to remove the data for id=19052013 which is wrong data#
data.CTI.final=data.CTI[!(data.CTI$id%in%aa),]
write.csv(data.CTI.final,"Z:/Common/Huiru/Cumulative Total
 Incidence/data_CTI.csv",row.names=FALSE)



#####Recalcuate the Basic Information After Removing the Wrong Records#####

duplicated(aa)
remove.id=aa[!duplicated(aa)]
length(remove.id) #removed 43 patients from the analysis#
#the overall number of patients in the study#
removed.SN=SN[!(SN$ccssid%in%remove.id),]
length(table(removed.SN$ccssid))

#event-data#
event_data_SN=removed.SN[removed.SN$sn==1,]

#total number of events#
nrow(event_data_SN)

#the number of patients who experienced event (first-ever or recurrence)#
length(table(event_data_SN$ccssid))
```

```
table(table(event_data_SN$ccssid))

#separate by group#
length(table(removed.SN[removed.SN$rad_yn==1,]$ccssid))  #number of patients in group1
length(table(removed.SN[removed.SN$rad_yn==2,]$ccssid))  #number of patients in group2

nrow(event_data_SN[event_data_SN$rad_yn==1,])            #number of events in group1
nrow(event_data_SN[event_data_SN$rad_yn==2,])            #number of events in group2

length(table(event_data_SN[event_data_SN$rad_yn==1,]$ccssid))
table(table(event_data_SN[event_data_SN$rad_yn==1,]$ccssid))

length(table(event_data_SN[event_data_SN$rad_yn==2,]$ccssid))
table(table(event_data_SN[event_data_SN$rad_yn==2,]$ccssid))



#####Calculate CIs and CTI#####

SN_CTI=read.csv("Z:/Common/Huiru/Cumulative Total Incidence/data_CTI.csv",h=T)

#There are total M-1 times of the events that we need to calculate the CI#
M=max(SN_CTI$event.number)
M1=max(SN_CTI[SN_CTI$group==1,]$event.number)
M2=max(SN_CTI[SN_CTI$group==2,]$event.number)

time.interval=seq(0,max(SN_CTI$dif),1);length(time.interval)
CTI.base1=rep(0,times=(M-
 1)*length(time.interval));dim(CTI.base1)=c(length(time.interval),(M-1))
CTI.base2=rep(0,times=(M-
 1)*length(time.interval));dim(CTI.base2)=c(length(time.interval),(M-1))

###for group1###
for (j in 1:(M1-1))
{
 CI_group1=data.frame(
 P=summary(etmCIF(Surv(dif, event!=0)~
  group,data=SN_CTI[SN_CTI$m_event==j&SN_CTI$group==1,]
        ,etype=event,failcode=1))[[1]]$'CIF 1'$P,
 Time=summary(etmCIF(Surv(dif, event!=0)~
  group,data=SN_CTI[SN_CTI$m_event==j&SN_CTI$group==1,]
        ,etype=event,failcode=1))[[1]]$'CIF 1'$time)
 CTI.base1[CI_group1$Time+1,j]=CI_group1$P

 for (i in 2:length(time.interval))
  {
   if (CTI.base1[i,j]==0)
     { CTI.base1[i,j]=CTI.base1[i-1,j]}
  }

}

CTI_group1=data.frame(CTI=rowSums(CTI.base1),time=time.interval)

###for group2###
```

```
for (j in 1:(M2-1))
{
 CI_group2=data.frame(
 P=summary(etmCIF(Surv(dif, event!=0)~
  group,data=SN_CTI[SN_CTI$m_event==j&SN_CTI$group==2,],
         etype=event,failcode=1))[[1]]$'CIF 1'$P,
 Time=summary(etmCIF(Surv(dif, event!=0)~
  group,data=SN_CTI[SN_CTI$m_event==j&SN_CTI$group==2,],
         etype=event,failcode=1))[[1]]$'CIF 1'$time)
 CTI.base2[CI_group2$Time+1,j]=CI_group2$P

 for (i in 2:length(time.interval))
  {
   if (CTI.base2[i,j]==0)
     { CTI.base2[i,j]=CTI.base2[i-1,j]}
  }
}

CTI_group2=data.frame(CTI=rowSums(CTI.base2),time=time.interval)

#Save the results#
write.csv(CTI.base1,"Z:/Common/Huiru/Cumulative Total
 Incidence/CTI.base1.csv",row.names=FALSE)
write.csv(CTI.base2,"Z:/Common/Huiru/Cumulative Total
 Incidence/CTI.base2.csv",row.names=FALSE)
write.csv(CTI_group1,"Z:/Common/Huiru/Cumulative Total
 Incidence/CTI_group1.csv",row.names=FALSE)
write.csv(CTI_group2,"Z:/Common/Huiru/Cumulative Total
 Incidence/CTI_group2.csv",row.names=FALSE)



#####Use Bootstrap to Get Confidence Interval for CTI#####

CTI_bootstrap=function(T,groupid)
{
 SN_CTI=read.csv("Z:/Common/Huiru/Cumulative Total Incidence/data_CTI.csv",h=T)
 time.interval=seq(0,max(SN_CTI$dif),1)

 #get the subgroup dataset for groupid#
 SNgroup=SN_CTI[SN_CTI$group==groupid,]

 #the total number of patients in the analysis for groupid: samplesize#
 #each re-sample size#
 samplesize=nrow(SNgroup[SNgroup$m_event==1,])

 #bootstrap table#
 id_table=SNgroup$id[1:samplesize]
 CTI=time.interval

 for (i in 1:T)
  {
   #the following ids has been selected#
   sampleid=sample(id_table,samplesize,replace=T)
```

```
  sample=NULL
  for (m in 1:samplesize)
    { sample=rbind(sample,SNgroup[SNgroup$id==sampleid[m],])}

  #There are total M-1 times of the events that we need to calculate the CI#
  M=max(sample$event.number)

  #calculate CTI#
  CTI.base=rep(0,times=(M-1)*length(time.interval))
  dim(CTI.base)=c(length(time.interval),(M-1))

  for (j in 1:(M-1))
    {
     CI_group=data.frame(
     P=summary(etmCIF(Surv(dif, event!=0)~ group,data=sample[sample$m_event==j,]
      ,etype=event,failcode=1))[[1]]$'CIF 1'$P,
     Time=summary(etmCIF(Surv(dif, event!=0)~ group,data=sample[sample$m_event==j,]
      ,etype=event,failcode=1))[[1]]$'CIF 1'$time)
     CTI.base[CI_group$Time+1,j]=CI_group$P

     for(i in 2:length(time.interval))
       {
        if (CTI.base[i,j]==0)
          {CTI.base[i,j]=CTI.base[i-1,j]}
       }
    }

 CTI=cbind(CTI,bootCTI=rowSums(CTI.base))
 }
write.csv(CTI,paste("Z:/Common/Huiru/Cumulative Total
 Incidence/bootCTI_group",groupid,"_",T,".csv",sep=""),row.names=FALSE)

#the 95% confidence interval#
boot=rep(0,times=length(time.interval)*3);dim(boot)=c(length(time.interval),3)
boot[,1]=time.interval
for (i in 1:length(time.interval))
  {
   boot[i,2]=quantile(CTI[i,2:(T+1)],0.025)
   boot[i,3]=quantile(CTI[i,2:(T+1)],0.975)
  }
write.csv(boot,paste("Z:/Common/Huiru/Cumulative Total
 Incidence/result2_bootCTI_group",groupid,"_",T,".csv",sep=""),row.names=FALSE)
}

CTI_bootstrap(1000,1)
CTI_bootstrap(1000,2)



#####Use Bootstrap to Get Confidence Interval for CI#####

CI_bootstrap=function(T,groupid)
{
 SN_CI=read.csv("Z:/Common/Huiru/Cumulative Total Incidence/data_CTI.csv",h=T)
 time.interval=seq(0,max(SN_CI$dif),1)
```

```
#get the subgroup dataset for groupid#
SNgroup=SN_CI[SN_CI$m_event==1&SN_CI$group==groupid,]

#the total number of patients in the analysis for groupid: samplesize#
#each re-sample size#
samplesize=nrow(SNgroup)

#bootstrap table#
id_table=SNgroup$id[1:samplesize]
CItable=rep(0,times=length(time.interval)*(T+1))
dim(CItable)=c(length(time.interval),(T+1))
CItable[,1]=time.interval

for (i in 1:T)
  {
    #the following ids has been selected#
    sampleid=sample(id_table,samplesize,replace=T)
    sample=NULL
    for (m in 1:samplesize)
      {sample=rbind(sample,SNgroup[SNgroup$id==sampleid[m],])}

    CI_group=data.frame(
    P=summary(etmCIF(Surv(dif, event!=0)~ group,data=sample
        ,etype=event,failcode=1))[[1]]$'CIF 1'$P,
    Time=summary(etmCIF(Surv(dif, event!=0)~ group,data=sample
        ,etype=event,failcode=1))[[1]]$'CIF 1'$time)
    CItable[CI_group$Time+1,(i+1)]=CI_group$P

    for(j in 2:length(time.interval))
      {
       if (CItable[j,(i+1)]==0)
         {CItable[j,(i+1)]=CItable[j-1,(i+1)]}
      }
  }

 write.csv(CItable,paste("Z:/Common/Huiru/Cumulative Total Incidence/
 bootCI_group",groupid,"_",T,".csv",sep=""),row.names=FALSE)

 #the 95% confidence interval#

 boot=rep(0,times=length(time.interval)*3);dim(boot)=c(length(time.interval),3)
 boot[,1]=time.interval
 for (i in 1:length(time.interval))
    {
      boot[i,2]=quantile(CItable[i,2:(T+1)],0.025)
      boot[i,3]=quantile(CItable[i,2:(T+1)],0.975)
    }
 write.csv(boot,paste("Z:/Common/Huiru/Cumulative Total
 Incidence/result_bootCI_group",groupid,"_",T,".csv",sep=""),row.names=FALSE)
}

CI_bootstrap(1000,1)
CI_bootstrap(1000,2)
```

```
#####Compare Bootstrap Results with R Command for CI#####

#get cumulative incidence and confidence interval#
CI=function(groupid)
{
 SN_CTI=read.csv("Z:/Common/Huiru/Cumulative Total Incidence/data_CTI.csv",h=T)
 time.interval=seq(0,max(SN_CTI$dif),1)

 #get the subgroup dataset for groupid for the first occurrence of event#
 SNgroup=SN_CTI[SN_CTI$group==groupid&SN_CTI$m_event==1,]
 CItable=rep(0,times=length(time.interval)*4);dim(CItable)=c(length(time.interval),4)
 CI_group=data.frame(
  Time=summary(etmCIF(Surv(dif, event!=0)~ group,data=SNgroup,
            etype=event,failcode=1))[[1]]$'CIF 1'$time,
  P=summary(etmCIF(Surv(dif, event!=0)~ group,data=SNgroup,
            etype=event,failcode=1))[[1]]$'CIF 1'$P,
  lower=summary(etmCIF(Surv(dif, event!=0)~ group,data=SNgroup,
            etype=event,failcode=1))[[1]]$'CIF 1'$lower,
  upper=summary(etmCIF(Surv(dif, event!=0)~ group,data=SNgroup,
            etype=event,failcode=1))[[1]]$'CIF 1'$upper
             )
  CItable[,1]=time.interval
  CItable[CI_group$Time+1,2]=CI_group$P
  CItable[CI_group$Time+1,3]=CI_group$lower
  CItable[CI_group$Time+1,4]=CI_group$upper

  for (j in 2:4)
    {
     for (i in 2:length(time.interval))
       {
        if (CItable[i,j]==0)
          {CItable[i,j]=CItable[i-1,j]}
       }
    }
 write.csv(CItable,paste("Z:/Common/Huiru/Cumulative Total
 Incidence/result_CI_group",groupid,".csv",sep=""),row.names=FALSE)
}

CI(1)
CI(2)



#####Figure 2 for Paper#####

CTI_group1=read.csv("Z:/Common/Huiru/Cumulative Total Incidence/CTI_group1.csv",h=T)
CTI_group2=read.csv("Z:/Common/Huiru/Cumulative Total Incidence/CTI_group2.csv",h=T)
bootCTI_group1=read.csv("Z:/Common/Huiru/Cumulative Total
 Incidence/result2_bootCTI_group1_1000.csv",h=T)
bootCTI_group2=read.csv("Z:/Common/Huiru/Cumulative Total
 Incidence/result2_bootCTI_group2_1000.csv",h=T)
CI_group1=read.csv("Z:/Common/Huiru/Cumulative Total
 Incidence/result_CI_group1.csv",h=T)
```

```
CI_group2=read.csv("Z:/Common/Huiru/Cumulative Total
 Incidence/result_CI_group2.csv",h=T)

###Figure 2a###
tiff(filename="Z:/Common/Huiru/Cumulative Total Incidence/figure.2a.tiff",width = 1024,
 height = 1024,units = "px")
plot(CTI_group1$time,CTI_group1$CTI,ylim=c(0,0.72),type="n",xlab="",las=1,ylab="",xaxt="
 n",yaxt="n",bty="n")
rect(CTI_group1$time-1,bootCTI_group1[,2],CTI_group1$time,
    bootCTI_group1[,3],border ="gray80", col ="gray80")
lines(CTI_group1$time,CTI_group1$CTI,lwd=4)
rect(CTI_group2$time-1,bootCTI_group2[,2],CTI_group2$time,
    bootCTI_group2[,3],border = "gray80", col = "gray80")
lines(CTI_group2$time,CTI_group2$CTI,lwd=4,lty=3)


axis(side=1, at=seq(0,365.25*40,365.25*5),label=seq(5,45,5),line=-
 2.2,padj=0.5,lwd=2,cex.axis=1.8)
axis(side=2,at=seq(0,0.8,0.1),label=seq(0,0.8,0.1),las=1,line=-2.3,lwd=2,cex.axis=1.8)
mtext(side=1, text="Years since diagnosis", line=2,cex=2)
mtext(side=2, text="Average number of events per person", line=2,cex=2)
legend(500,0.65,c("RT group","No RT group"),bty="n",lty=c(1,3),lwd=4,cex=2)
#legend(750,0.605,c("95%CI (RT group)","95%CI (No RT group)"),pch=15,
#       col=c("pink","light blue"),bty="n",lty=0,lwd=2,cex=1.8)
dev.off()

###figure 2b###
tiff(filename="Z:/Common/Huiru/Cumulative Total Incidence/figure.2b.tiff",width = 1024,
 height = 1024,units = "px")

plot(CI_group1[,1],CI_group1[,2],ylim=c(0,0.72),type="n",xlab="",las=1,ylab="",xaxt="n",yaxt
 ="n",bty="n")
rect(CI_group1[,1]-1,CI_group1[,3],CI_group1[,1],
    CI_group1[,4],border ="gray80", col ="gray80")
lines(CI_group1[,1],CI_group1[,2],lwd=4)
rect(CI_group2[,1]-1,CI_group2[,3],CI_group2[,1],
    CI_group2[,4],border = "gray80", col = "gray80")
lines(CI_group2[,1],CI_group2[,2],lwd=4,lty=3)


axis(side=1, at=seq(0,365.25*40,365.25*5),label=seq(5,45,5),line=-
 2.2,padj=0.5,lwd=2,cex.axis=1.8)
axis(side=2,at=seq(0,0.8,0.1),label=seq(0,0.8,0.1),las=1,line=-2.3,lwd=2,cex.axis=1.8)
mtext(side=1, text="Years since diagnosis", line=2,cex=2)
mtext(side=2, text="Cumulative probability", line=2,cex=2)
legend(500,0.65,c("RT group","No RT group"),bty="n",lty=c(1,3),lwd=4,cex=2)

dev.off()
```

Appendix 3: R code for CCSS data analysis in Chapter 3

```
library(etm)

#read data#
SMN=read.csv("Z:/Common/Huiru/Risk and Rate/smn_huiru_regression.csv",h=T)
#delete the duplicated rows#
SMN=SMN[!duplicated(SMN),]

#####Data Manipulation#####

#format study start date#
d_start=as.Date(SMN$d_start, "%d%b%Y")

#format study stop date#
d_stop=as.Date(SMN$d_stop, "%d%b%Y")

#format patient death date, the competing risks event date#
d_death=as.Date(SMN$D_DEATH,"%d%b%Y")

#format SMN event date#
smn_date=as.Date(SMN$smn_date,"%d%b%Y")

#id#
ccssid=SMN$ccssid

#change the dataset to long format#
#event=0: censoring#
#event=1: event of interest#
#event=2: competing risks event#
id=NULL;entry=NULL;exit=NULL;group=NULL;event=NULL;infor=NULL

for (i in 1:nrow(SMN))
{
#No event date or death date#
#Means censoring individuals#
 if (is.na(d_death[i]) & is.na(smn_date[i]))
    { id=c(id,ccssid[i])
     entry=c(entry,as.character(d_start[i]))
     exit=c(exit,as.character(d_stop[i]))          #date=stop following up date#
     event=c(event, 0)                        #censoring is coded as 0#
     infor=rbind(infor,SMN[i,c(4,8:19)])
     }

#Have event date but no death date#
```

```
#Means event of interest individuals#
 if (is.na(d_death[i]) & !is.na(smn_date[i]))
    { id=c(id,ccssid[i])
      entry=c(entry,as.character(d_start[i]))
      exit=c(exit,as.character(smn_date[i]))          #date=event of interest date#
      event=c(event,1)                              #event of interest is coded as 1#
      infor=rbind(infor,SMN[i,c(4,8:19)])

      if (d_stop[i]>=smn_date[i]) #event individual be censored at following up end date#
        { id=c(id,ccssid[i])
          entry=c(entry,as.character(d_start[i]))
          exit=c(exit,as.character(d_stop[i]))        #date=stop following up date#
          event=c(event,0)                          #censoring is coded as 0#
          infor=rbind(infor,SMN[i,c(4,8:19)])
        }
    }

#Have death date but no event date#
#Means competing risk event individuals#
 if (!is.na(d_death[i]) & is.na(smn_date[i]))
    { if (d_death[i]>d_stop[i]) #Death happened later, then treat as censoring#
        { id=c(id,ccssid[i])
          entry=c(entry,as.character(d_start[i]))
          exit=c(exit,as.character(d_stop[i]))        #date=stop following up date#
          event=c(event,0)                          #censoring is coded as 0#
          infor=rbind(infor,SMN[i,c(4,8:19)])
        }

      if (d_death[i]<=d_stop[i]) #Death happened earlier#
        { id=c(id,ccssid[i])
          entry=c(entry,as.character(d_start[i]))
          exit=c(exit,as.character(d_death[i]))       #date=competing risk event date#
          event=c(event,2)                          #competing risk event is coded as 2#
          infor=rbind(infor,SMN[i,c(4,8:19)])
        }
    }

#Have death date and also event date#
 if (!is.na(d_death[i]) & !is.na(smn_date[i]))
    { id=c(id,ccssid[i])
      entry=c(entry,as.character(d_start[i]))
      exit=c(exit,as.character(smn_date[i]))          #recode the event date first#
      event=c(event,1)                              #event of interest is coded as 1#
      infor=rbind(infor,SMN[i,c(4,8:19)])

      if (d_death[i]>d_stop[i]) #Death happened later, then treat as censoring#
        { id=c(id,ccssid[i])
          entry=c(entry,as.character(d_start[i]))
          exit=c(exit,as.character(d_stop[i]))        #date=stop following up date#
          event=c(event,0)                          #censoring is coded as 0#
          infor=rbind(infor,SMN[i,c(4,8:19)])
        }

      if (d_death[i]<=d_stop[i]) #Death happened earlier#
        { id=c(id,ccssid[i])
```

```
          entry=c(entry,as.character(d_start[i]))
          exit=c(exit,as.character(d_death[i]))       #date=competing risk event date#
          event=c(event,2)                          #competing risk event is coded as 2#
          infor=rbind(infor,SMN[i,c(4,8:19)])
        }
      }
}

data.SMN=data.frame(id=id, entry=entry,exit=exit,event=event, infor)
#remove the duplicated rows#
data_SMN=data.SMN[!duplicated(data.SMN),]
#order the dataset by id and event order#
ii=order(data_SMN$id,data_SMN$exit)
dataSMN=data_SMN[ii,]

#follow up duration#
dif=as.numeric(difftime(as.Date(dataSMN$exit,"%Y-%m-%d"),as.Date(dataSMN$entry,"%Y-
 %m-%d"),unit="days"))

#age at diagnosis of childhood cancer#
age.cancer=as.numeric(difftime(as.Date(dataSMN$D_DX,"%d%b%Y"),as.Date(dataSMN$D_B
 IRTH,"%d%b%Y"),unit="days"))/365.25

#age at stop following-up#
age.stop=as.numeric(difftime(as.Date(dataSMN$exit,"%Y-%m-
 %d"),as.Date(dataSMN$D_BIRTH,"%d%b%Y"),unit="days"))/365.25

#event number indicator#
index=table(dataSMN$id)
N=length(index)

event.number=NULL
for (i in 1:N)
{
 if (index[i]==1)
   { event.number=c(event.number,0)}
 if (index[i]!=1)
   { event.number=c(event.number,1:index[i])}
}

#format other variables#
v.sex=NULL; v.dxgroup=NULL; v.race=NULL; v.era=NULL; v.agedx=NULL; v.sple=NULL
v.anth=NULL; v.epip=NULL; v.plat=NULL

v.sex[dataSMN$SEX=="Male"]=0; v.sex[dataSMN$SEX=="Female"]=1


v.dxgroup[dataSMN$DXGROUP=="Acute lymphoblastic leukemia"]=0
v.dxgroup[dataSMN$DXGROUP=="Astrocytomas"]=1
v.dxgroup[dataSMN$DXGROUP=="Hodgkins disease"]=2
v.dxgroup[dataSMN$DXGROUP=="Medulloblastoma, PNET"]=3
v.dxgroup[dataSMN$DXGROUP=="Non-Hodgkins lymphoma"]=4
v.dxgroup[dataSMN$DXGROUP=="Other bone tumors"]=5
v.dxgroup[dataSMN$DXGROUP=="Other leukemia"]=6
v.dxgroup[dataSMN$DXGROUP=="Acute myeloid leukemia"]=7
```

```
v.dxgroup[dataSMN$DXGROUP=="Ewings sarcoma"]=8
v.dxgroup[dataSMN$DXGROUP=="Kidney tumors"]=9
v.dxgroup[dataSMN$DXGROUP=="Neuroblastoma"]=10
v.dxgroup[dataSMN$DXGROUP=="Osteosarcoma"]=11
v.dxgroup[dataSMN$DXGROUP=="Other CNS tumors"]=12
v.dxgroup[dataSMN$DXGROUP=="Soft tissue sarcoma"]=13

v.alkscore=dataSMN$alkscore5

v.race[dataSMN$race4=="Black,NH"]=0
v.race[dataSMN$race4=="White,NH"]=1
v.race[dataSMN$race4=="Hispanic/Latine"]=2
v.race[dataSMN$race4=="Other"]=3

v.era[dataSMN$y_dx=="1970-1974"]=0
v.era[dataSMN$y_dx=="1975-1979"]=1
v.era[dataSMN$y_dx=="1980-1986"]=2

v.agedx[dataSMN$age_dx=="<5"]=0
v.agedx[dataSMN$age_dx=="5-9.9"]=1
v.agedx[dataSMN$age_dx=="10-14.9"]=2
v.agedx[dataSMN$age_dx==">=15"]=3

v.sple[dataSMN$sple=="no"]=0
v.sple[dataSMN$sple=="yes"]=1

v.anth[dataSMN$anth_5=="None"]=0
v.anth[dataSMN$anth_5=="<=100"]=1
v.anth[dataSMN$anth_5=="101-300"]=2
v.anth[dataSMN$anth_5==">300"]=3

v.epip[dataSMN$epip_5=="None"]=0
v.epip[dataSMN$epip_5=="<=1000"]=1
v.epip[dataSMN$epip_5=="1001-4000"]=2
v.epip[dataSMN$epip_5==">4000"]=3

v.plat[dataSMN$plat_5=="None"]=0
v.plat[dataSMN$plat_5=="<=400"]=1
v.plat[dataSMN$plat_5=="401-750"]=2
v.plat[dataSMN$plat_5==">750"]=3


SMN.all=data.frame(dataSMN,age.cancer=age.cancer,age.stop=age.stop,v.sex=v.sex,v.dxgroup=
  v.dxgroup,v.alkscore=v.alkscore,v.race=v.race,v.era=v.era,v.agedx=v.agedx,v.sple=v.sple,v.ant
  h=v.anth,v.epip=v.epip,v.plat=v.plat,dif=dif,event.number=event.number)

#we only include first event in the regression analysis#
regress.smn=SMN.all[event.number==0 | event.number==1,]

write.csv(SMN.all," Z:/Common/Huiru/Risk and Rate/SMN.all.csv",row.names=FALSE)
write.csv(regress.smn," Z:/Common/Huiru/Risk and Rate/regress.smn.csv",row.names=FALSE)


#####Data Analysis#####
regress.smn=read.csv("Z:/Common/Huiru/Risk and Rate/regress.smn.csv", h=T)
```

```
###basic information###
cases=regress.smn[regress.smn$event==1,]
no.cases=regress.smn[regress.smn$event !=1,]

n.all=nrow(regress.smn)
n.cases=nrow(cases)
n.nocases=nrow(no.cases)

#age at diagnosis of childhood cancer#
mean(regress.smn$age.cancer)
median(regress.smn$age.cancer)

mean(cases$age.cancer)
median(cases$age.cancer)

mean(no.cases$age.cancer)
median(no.cases$age.cancer)

#age at diagnosis of SMN#
mean(cases$age.stop)
median(cases$age.stop)

#gender#
sum(regress.smn$v.sex==1)/n.all
sum(cases$v.sex==1)/n.cases
sum(no.cases$v.sex==1)/n.nocases

#race#
table(regress.smn$race4)/n.all
table(cases$race4)/n.cases
table(no.cases$race4)/n.nocases

#childhood cancer diagnosis#
table(regress.smn$DXGROUP)/n.all
table(cases$DXGROUP)/n.cases
table(no.cases$DXGROUP)/n.nocases

#radiation#
table(regress.smn$rad_yn)/n.all
table(cases$rad_yn)/n.cases
table(no.cases$rad_yn)/n.nocases

#alkylating agent score#
table(regress.smn$alkscore5)/n.all
table(cases$alkscore5)/n.cases
table(no.cases$alkscore5)/n.nocases

#treatment ear#
table(regress.smn$y_dx)/n.all
table(cases$y_dx)/n.cases
table(no.cases$y_dx)/n.nocases

#splenectomy#
table(regress.smn$sple)/n.all
```

```
table(cases$sple)/n.cases
table(no.cases$sple)/n.nocases

#anthracycline exposure#
table(regress.smn$anth_5)/n.all*100
table(cases$anth_5)/n.cases*100
table(no.cases$anth_5)/n.nocases*100

#epipodophyllotoxin exposure#
table(regress.smn$epip_5)/n.all*100
table(cases$epip_5)/n.cases*100
table(no.cases$epip_5)/n.nocases*100

#platinum exposure#
table(regress.smn$plat_5)/n.all*100
table(cases$plat_5)/n.cases*100
table(no.cases$plat_5)/n.nocases*100

###fit Cox Regression###
library(splines)
library(survival)

event.cox=rep(0,times=nrow(regress.smn))
event.cox[regress.smn$event==1]=1

rad=NULL
rad[regress.smn$rad_yn==2]=0
rad[regress.smn$rad_yn==1]=1

regressSMN=data.frame(regress.smn, event.cox=event.cox,rad=rad)

COX=coxph(Surv(dif,event.cox) ~
 factor(rad)+factor(v.sex)+factor(v.dxgroup)+factor(v.alkscore)+factor(v.era)+factor(v.race)+fac
 tor(v.agedx)+factor(v.sple)+factor(v.anth)+factor(v.epip)+factor(v.plat),regressSMN)

summary(COX)

#for competing-risk event#
event.cox.c=rep(0,times=nrow(regress.smn))
event.cox.c[regress.smn$event==2]=1

rad=NULL
rad[regress.smn$rad_yn==2]=0
rad[regress.smn$rad_yn==1]=1

regressSMN=data.frame(regress.smn, event.cox.c=event.cox.c,rad=rad)
COX.c=coxph(Surv(dif,event.cox.c) ~
 factor(rad)+factor(v.sex)+factor(v.dxgroup)+factor(v.alkscore)+factor(v.era)+factor(v.race)+fac
 tor(v.agedx)+factor(v.sple)+factor(v.anth)+factor(v.epip)+factor(v.plat),regressSMN)

summary(COX.c)


###fit Fine & Gray subdistribution hazards model###
library(cmprsk)
```

```
#generate dummy variables#

#radiation#
v.rad=regressSMN$rad

#v.sex#
v.sex=regressSMN$v.sex

#race#
race1=ifelse(regressSMN$v.race==1,1,0)
race2=ifelse(regressSMN$v.race==2,1,0)
race3=ifelse(regressSMN$v.race==3,1,0)


#v.dxgroup#
v.dxgroup1= ifelse(regressSMN$v.dxgroup==1,1,0)
v.dxgroup2= ifelse(regressSMN$v.dxgroup==2,1,0)
v.dxgroup3= ifelse(regressSMN$v.dxgroup==3,1,0)
v.dxgroup4= ifelse(regressSMN$v.dxgroup==4,1,0)
v.dxgroup5= ifelse(regressSMN$v.dxgroup==5,1,0)
v.dxgroup6= ifelse(regressSMN$v.dxgroup==6,1,0)
v.dxgroup7= ifelse(regressSMN$v.dxgroup==7,1,0)
v.dxgroup8= ifelse(regressSMN$v.dxgroup==8,1,0)
v.dxgroup9= ifelse(regressSMN$v.dxgroup==9,1,0)
v.dxgroup10= ifelse(regressSMN$v.dxgroup==10,1,0)
v.dxgroup11= ifelse(regressSMN$v.dxgroup==11,1,0)
v.dxgroup12= ifelse(regressSMN$v.dxgroup==12,1,0)
v.dxgroup13= ifelse(regressSMN$v.dxgroup==13,1,0)

#v.alkscore#
v.alkscore1=ifelse(regressSMN$v.alkscore==1,1,0)
v.alkscore2=ifelse(regressSMN$v.alkscore==2,1,0)
v.alkscore3=ifelse(regressSMN$v.alkscore==3,1,0)

#v.era#
v.era1=ifelse(regressSMN$v.era==1,1,0)
v.era2=ifelse(regressSMN$v.era==2,1,0)

#v.agedx#
v.agedx1=ifelse(regressSMN$v.agedx==1,1,0)
v.agedx2=ifelse(regressSMN$v.agedx==2,1,0)
v.agedx3=ifelse(regressSMN$v.agedx==3,1,0)

#v.sple#
v.sple1=ifelse(regressSMN$v.sple==1,1,0)

#v.anth#
v.anth1=ifelse(regressSMN$v.anth==1,1,0)
v.anth2=ifelse(regressSMN$v.anth==2,1,0)
v.anth3=ifelse(regressSMN$v.anth==3,1,0)

#v.epip#
v.epip1=ifelse(regressSMN$v.epip==1,1,0)
v.epip2=ifelse(regressSMN$v.epip==2,1,0)
```

```
v.epip3=ifelse(regressSMN$v.epip==3,1,0)

#v.plat#
v.plat1=ifelse(regressSMN$v.plat==1,1,0)
v.plat2=ifelse(regressSMN$v.plat==2,1,0)
v.plat3=ifelse(regressSMN$v.plat==3,1,0)

#generate covariates matrix#
cov=cbind(v.rad,v.sex,race1,race2,race3,v.dxgroup1,v.dxgroup2,v.dxgroup3,v.dxgroup4,v.dxgro
  up5,v.dxgroup6,v.dxgroup7,v.dxgroup8,v.dxgroup9,v.dxgroup10,v.dxgroup11,v.dxgroup12,v.d
  xgroup13,v.alkscore1,v.alkscore2,v.alkscore3,v.era1,v.era2,v.agedx1,v.agedx2,v.agedx3,v.sple1
  ,v.anth1,v.anth2,v.anth3,v.epip1,v.epip2,v.epip3,v.plat1,v.plat2,v.plat3)

CR<- crr(regressSMN$dif, regressSMN$event,cov)
summary(CR)
```