Statistical Methods for Competing Risks Model

by

Yao Lu

A thesis submitted to The Johns Hopkins University in conformity with the requirements for the degree of Master of Science.

Baltimore, Maryland

May, 2014

Abstract

"Competing Risks" refers to the study of the time to event where there is more than one type of failure event. The distinct problem can be vital, since not only it can inform the patients what risks they are facing, but also it helps to select appropriate treatment for a particular patient. In Chapter 2 we introduce two methods, cause-specific hazard model and cumulative incidence function, to deal with the competing risks problem. In Chapter 3, we study the prognosis of different patterns of cancer recurrences using data from 209 patients who had surgical resection of pancreatic cancer at the Johns Hopkins Hospital between 1998 and 2007. We analyze different types of tumor recurrences and death as competing risks. We first apply Cox's proportional hazard model to analyze the time from surgery to the composite endpoint of recurrence or death. We then analyze the nonparametric cumulative incidence function under competing risks setting. The conditional cumulative incidence function given each event type will be presented to investigate whether the competing risks have different distribution patterns. Then, the cause-specific hazard model is applied to evaluate the effect of risk factors on the cause-specific hazards, and the results are

ABSTRACT

compared with the conventional survival analysis that ignores the recurrence types. Finally, we use Cox's proportional hazard model with time-dependent covariates to analyze the time from surgery to death. At last, we discuss implications of data analysis and future research.

Primary Reader: Mei-Cheng Wang

Secondary Reader: Chiung-Yu Huang

Acknowledgments

I would like to express my very great appreciation to Dr. Mei-Cheng Wang and Dr. Chiung-Yu Huang. Advice given by Dr. Wang and Dr. Huang has been a great help in my thesis research. Also, I would like to thank the professors who have taught me during my study at Johns Hopkins Bloomberg School of Public Health. My special thanks are extended to my friends who have assisted and encouraged me.

Dedication

This thesis is dedicated to Jufeng Cui and Xinnian Lu.

Contents

CONTENTS

List of Tables

List of Figures

Chapter 1

Introudction and Literature Review

Survival analysis, which is considered a branch of statistics, is widely utilized in epidemiological and medical studies, especially in cancer research. It deals with modelling, estimating and testing for time to event data. In survival analysis, survival time T is time from a defined starting-point to the occurrence of a given event. For example, in a clinical trial, the survival time may be defined as the time from the start of certain treatment to diagnosis of disease. When studying disease with recurrence pattern, the survival time varies with your definition, which may be the time from receiving the treatment to first diagnosis of recurrence, or the time from receiving treatment to death. While the recurrence is considered as an endpoint of survival time in the first scenario, it may be a confounder, effect modifier or mediator

in the second scenario. Also, in the collection of survival data, censoring and other sampling constrains, such as left truncation, often arise. Censoring refers to the scenario that we fail to observe the event. It may due to various reasons, such as the end of study, patients' dropping out of study, or other reasons for loss to follow-up. In analyzing survival data, two functions of time are of particular interest: the survival function and the hazard function. Survival function $S(t)$ is defined as the probability that a person's survival time is larger or equal to time t. The hazard function $h(t)$ is the conditional probability of dying at time t given the subject survived up to that time. Since the survival function $S(t)$ provides us useful summary information, it is often desired and common to estimate the survival function $S(t)$ in exploratory data analysis. The KaplanMeier method (Kaplan and P. [1958]) can be used to estimate the survival function from the observed survival times with the only assumption that the censoring mechanism is independent of survival time. Kaplan-Meier method is based on the idea that the probability of surviving at time t is a product of all survival rates for each period prior to t. Other than estimating the survival function $S(t)$, we can also study the hazard function $h(t)$, based on which the Cox proportional hazard model is formulated (Cox [1972]). Cox proportional hazard model is a semiparametric model, and it includes two components. One is the unspecified baseline hazard, and the other parts is the parametric component, where certain covariates of interest are included, such as age, gender. The model has nice interpretation in terms of hazard and it is semiparametric. Under the independent censoring condition, based

on proportional hazard model, the likelihood function consists of two parts. The first likelihood, which is known as "partial likelihood", only involves the covariates of interest. Thus the computation of the maximum likelihood estimate of those covariates is manageable, and inference can be made. When the survival time is continuous, Breslow (Breslow [1974]) gives a ways to estimate the baseline hazard.

Under the conventional survival analysis settings, where only one type of event can occur during the study, we can use methods described above to estimate survival function or hazard function. However, more complex circumstance arises during the study. There may be the study of any failure type in which there is more than one distinct type of failure but the patient's eventual failure is attributed to precisely one of the cause. This kind of situation is referred to as "**competing risks**". The distinct problem can be vital, since it not only informs the patients what risks they are facing, but also helps to select appropriate treatment for a particular patient. When we want to study one certain type of failure type under the setting of competing risks, the previous two conventional methods cannot be applied here, since there may be dependent censoring occurring. To deal with competing risks problems, the usual formulation of these problems is in terms of latent failure times corresponding to each type of failure. We assume that each person may have a potential failure time for each failure type, which is the latent failure time of certain failure type. Since each person can only die of one failure type, the survival time or failure time we observe is the minimum of these latent failure times and the corresponding failure type is the

type of which the person dies (Cox [1959], Moeschberger [1971]). The latent failure times have two approach of interpretation. Cox [1959] and Moeschberger [1971] defines the failure time for each failure time under competing risks settings to be the time that would be observed if all other types of failure are removed. However, this needs strong assumption that certain failure type will operate exactly the same as under the condition that all other failure types are removed. And usually, the risk of certain failure type will change if other failure types are removed (Cox [1959], Makeham [1874], Cornfield [1957]). The first approach of interpretation, though having a physical meaning, will not be considered. The second approach the latent failure time of failure type j is the observed time of failure if the individual fails of type j, while no physical meaning is attached to unobserved other latent failure times. However, this approach may lead to lack of physical interpretation of the unobserved latent failure times and identifiability problems (Prentice et al. [1978]). Therefore, instead of using latent failure times format, utilizing cause-specific hazard or cumulative incidence function for observed quantities provides us alternative methods to approach competing risks problems.

In the following chapters, we will first introduce cause-specific hazard and cumulative incidence function, including the definitions, inference and its application to study competing risks problems. Then we will apply the methods to study the relationship between survival time and recurrence types among pancreatic cancer patients, and compare it with the Cox proportional hazard model.

Chapter 2

Competing Risks Models

As mentioned before, "competing risks" refers to the study of any failure type in which there is more than one distinct type of failure. In the following sections, we will introduce two popular methods to handle this problem; cause-specific hazard model (Prentice and Breslow [1978]), estimating cumulative incidence function based on subditribution hazard (Fine and Gray [1999]).

A statistical model for competing risks data involves the observed quantities $(T, j, \mathbf{z})$ and the distribution for them, where T is the time to failure or death, and is a positive random variable; $j = 1, \ldots, m$ refers to the type of failure the patient has; $\mathbf{z}$ is the covariate vector, or the covariates we are interested in. The covariate vector $\mathbf{z}$ may be time-depend, and can be written as $\mathbf{z}(t)$. The latent failure times $Y_1, Y_2, \ldots, Y_m$ correspond to each type of failure type $j = 1, \ldots, m$. The time to failure or death is $T = \min(Y_1, \ldots, Y_m)$ and $j = \{p | Y_p \leq Y_k, k = 1, \ldots, m\}$. C_j

is the potential censoring time for failure type $j = 1, \ldots, m$, then potential censoring time $C = \{C_k | T_k \leq T_j, j = 1, \ldots, m\}$. Therefore the observed failure time is $X = \min(T, C)$, and the censoring indicator $\Delta = I(T < C)$. This setting as mentioned before lacks physical interpretation and identifiability problem.

2.1 Cause-Specific Hazard

Assume that the failure time T is continuous. The overall hazard function for is the conditional probability of dying at time t given that a subject survived up to that time. And the hazard for an individual with the covariate vector $\mathbf{z} = \mathbf{z}(\mathbf{t})$ is defined as following,

$$\lambda(t; \mathbf{z}) = \lim_{\Delta t \to 0} P\{t \leq T < t + \Delta t | T \geq t; \mathbf{z}(\mathbf{t})\}/\Delta t.$$

Cause-specific hazard functions (Chiang [1968], Altshuler [1970], Holt [1978], Prentice and Breslow [1978]) are defined by

$$\lambda_j(t; \mathbf{z}) = \lim_{\Delta t \to 0} P\{t \leq T < t + \Delta t, J = j | T \geq t; \mathbf{z}(\mathbf{t})\}/\Delta t,$$

for the failure type $j = 1, 2, \ldots, m$. It simply gives the instantaneous failure rate from cause j at time t, given the regression vector $\mathbf{z}(\mathbf{t})$ (Cox [1972]) and those who survive time t, in the presence of other failure types. By the previous two definitions, we can get that the overall hazard function can be expressed as

$$\lambda(t; \mathbf{z}) = \sum_{1}^{m} \lambda_j(t; \mathbf{z}). \tag{2.1}$$

The overall survival function at time t, which is the probability that a person's survival time is larger or equal to time t, can be written as the function of overall hazard as below,

$$F(t;\mathbf{z}^*) = \exp\left\{-\int_0^t \lambda(u;\mathbf{z})du\right\},$$

and the probability for time to failure and cause of failure

$$f_j(t;\mathbf{z}^*) = \lambda_j(t:\mathbf{z})F(t;\mathbf{z}^*), \tag{2.2}$$

where $\mathbf{z}^* = \mathbf{z}^*(t)$ denotes $\{\mathbf{z}(u); u \leq t\}$, which refers to the history information about covariates up to time t.

Suppose now there are n study subjects, $(t_i, j_i, \delta_i, \mathbf{z}_i)$, where t_i is the observed failure time of subject i, j_i is the cause of failure, δ_i is a censoring indicator, and $\mathbf{z}_i^* = \mathbf{z}_i^*(t)$ is a vector-valued regressor for the ith subject. As usual an independent censoring mechanism will be assumed. The likelihood function under an independent censoring mechanism is

$$\prod_{i=1}^{n}\{[\lambda_{j_i}(t_i;\mathbf{z}_i)]^{\delta_i}F(t_i;\mathbf{z}_i^*)\} = \left(\prod_{i=1}^{n}[\lambda_{j_i}(t_i;\mathbf{z}_i)]^{\delta_i}\prod_{j=1}^{m}\exp\left\{-\int_0^{t_i}\lambda_j[u;\mathbf{z}(u)]du\right\}\right). \tag{2.3}$$

The likelihood function is completely specified by the cause-specific hazard functions λ_j. Rearranging the likelihood factors into a component for each j, the likelihood factor for λ_j is precisely the same as being obtained by regarding all types of failure other than type j as being censored at their time of failure. The likelihood factorization along with standard survival data techniques make it clear that λ_j, the

cause-specific hazard function, has the potential to be directly estimated from the data of the form $(t, j, \delta, \mathbf{z}^*)$.

2.1.1 Inference on the cause-specific regression coefficients

As mentioned before, the jth likelihood factor is precisely the likelihood being obtained by regarding all other failure types as being censored. This implies the usual survival data methods for a single failure type can be used for testing and estimating λ_j. For example, we can use Cox's proportional hazard model (Cox [1972, 1975]) in Holt [1978] and Prentice and Breslow [1978] to model effects of regression covariates in the cause-specific hazard functions as

$$\lambda_j(t; \mathbf{z}) = \lambda_{0j} \exp(\mathbf{z}\beta_j), \qquad j = 1, \ldots, m. \tag{2.4}$$

The partial likelihood (Holt [1978]) for $\beta_j's$ can be written as

$$\prod_{j=1}^{m} \left[\prod_{i=1}^{d_j} \frac{\exp[\mathbf{z}_{j(i)}\beta_j]}{\sum_{l \in R(t_{j(i)})} \exp(\mathbf{z}_l \beta_l)} \right], \tag{2.5}$$

where $t_{j(i)}$, $i = 1, \ldots, d_j$ denotes the d_j times of failure of type j, $R(t_{j(i)})$ is the risk set prior to $t_{j(i)}$. Standard asymptotic likelihood methods can be applied to the partial likelihood for the estimate of the β.

Here the assumption that the jth type of failure is independent of other failure types and censoring. However, no assumption is required concerning the interrelation

among the other causes of failure. Thus the inference of the coefficients can be made without introducing strong model assumptions. We note that the interpretation of the same regression estimate may change under a new set of conditions, for example, certain types of failure have been removed.

With stronger assumption, that different failure types are independent, a stronger interpretation can be made about λ_j. At this time, λ_j is exactly the hazard function for cause j given that no other causes are operative. Note that the specific β_j can be estimated using the jth component of the previous partial likelihood function without restricting other failure types to follow the proportional hazard form.

2.1.2 The study of interrelations among failure types

Failure types j_1 and j_2 will be said to be related if study subjects at high risk for a failure type j_1 are at the same time at high, or low, risk for a failure type j_2. We can import the definition of time-depend risk-indicator for some failure types which can establish a relationship to cause-specific hazard functions for other failure types. For example, j_1 indicates death due to lung cancer and j_2 refers to stroke. We can include time-depend covariate as the indicator of j_2 in the Cox's proportional hazard model for cause-specific hazard for j_1. If there is a positive relationship between j_1 and j_2, then it will indicate individuals at high risk for j_1 is simultaneously at high risk for j_2.

2.1.3 Failure rate estimation following cause removal

Another problem we will address in competing risks is to estimate the failure rates for certain causes given the removal of some or all other causes. This kind of problem is not in general well defined until the mechanism for cause removal is clearly specified, and it is necessary to explore detailed knowledge of the biological mechanism giving rise to failures.

Chiang (Chiang [1968]) asserts that a very strong assumption, that the instantaneous failure rate for cause j under actual conditions, with all m causes operative, is identical to that under new condition where only cause j presents, is needed when you want to base the probability statements for cause j considering it's the only failure type on cause-specific hazard function λ_j.

2.2 Cumulative Incident Function

In the presence of competing risks, the cumulative incidence function is the probability of having jth failure type is $F_j(t) = Pr(T \leq t, J = j)$. It can be written as,

$$F_j(t) = \int_0^t \lambda_j(\mu) S(\mu) d\mu. \tag{2.6}$$

Then a nonparametric estimator can be obtained by first calculating the Kaplan-Meier estimate of the overall survival function $\hat{S}(t)$, and then at the observed time t_i, where $\delta_i = 1$ and $J_i = j$.

Besides the nonparametric estimator, we introduce a semi-parametric estimator for CIF by Fine and Gray [1999]. To simple the procedure, our interest here is to model the CIF for failure type 1,

$$F_1(t; \mathbf{z}) = Pr(T \leq t; J = 1|\mathbf{z}). \tag{2.7}$$

Instead of estimating CIF directly, we consider the class of semiparametric transformation models (Cheng et al. [1995], Cox [1972], Cuzick [1988], Dabrowska and Doksum [1988], Fine et al. [1998], Murphy et al. [1997]). The transformation formula is

$$g\{F_1(t; Z)\} = h_0(t) + Z^T\beta_0 \tag{2.8}$$

where h_0 is a completely unspecified, invertible, and monotone increasing function. On the scale of g, the regression coefficients are a measure of distance from the baseline marginal probability function $g^{-1}\{h_0(t)\}$.

The first step is to try $g = \log\{-\log(1-u)\}$ (Fine and Gray [1999]), since it is corresponding to the popular hazard model. However, we should notice the hazard here is not the usual cause-specific hazard and define it as subdistribution hazard (Gray [1988]);

$$\lambda_1^*(t; \mathbf{z}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr\{t \leq T \leq t + \Delta, J = 1 \mid T \geq t \cup (T \leq t \cap J \neq 1), \mathbf{z}\}, \tag{2.9}$$

then we can get:

$$\lambda_1^*(t;\mathbf{z}) = \{dF_1(t;\mathbf{z})/dt\}/\{1 - F_1(t;\mathbf{z})\} \tag{2.10}$$

$$= -d\,\log\{1 - F_1(t;\mathbf{z})\}dt. \tag{2.11}$$

We can think of λ_1^* as the hazard function for improper random variable $T^* = I(J = 1) \times T + I(J \neq 1) \times \infty$. T^* has the distribution function equal to $F_1(t;\mathbf{z})$. In this scenario, failure from other causes in unobservable, and the estimation of overall survival is equal to the estimation of the subdistribution for individuals who will eventually experience the event of interest. However, in general competing-risks setting, failure from other causes are observable. Therefore, interpretation of g-transformation model for CIF is problematic if viewed in terms of the corresponding hazard function.

Let

$$\lambda_1^*(t;\mathbf{z}) = \lambda_{10}^*(t)\exp\{\mathbf{z}(t)^T\beta_0\}, \tag{2.12}$$

so that

$$F_1(t;\mathbf{z}) = 1 - \exp\left[-\int_0^t \lambda_{10}^*(s)\exp\{\mathbf{z}(s)^T\beta_0\}ds\right]. \tag{2.13}$$

Thus the regression coefficients and baseline hazard from the Cox transformation model for F_1 have a straightforward interpretation that does not depend on the problematic structure of the subdistribution hazard.

In the following sections, we will show that a modified partial likelihood method can be apply on the subdistribution hazards with complete and censoring complete data.

2.2.1 Complete data

This procedure, which can be viewed as a modification of partial likelihood, yields estimates for the regression parameters that are consistent and asymptotically normal (Fine and Gray [1999]). A version of Breslow's estimator (Breslow [1974]) provides a consistent estimate for $\Lambda_{10}^*(t) = \int_0^t \lambda_{10}^*(s)ds$ that is equivalent to a mean 0 Gaussian process.

We define the risk set at the time of failure for ith individual,

$$R_i = \{k : (T_k \geq T_i) \cup (T_k \leq T_i \cap j_k \neq 1)\}.$$

This includes two groups: those who have survived at time t and those who have failed from other causes before time t. It leads to the proper partial likelihood for the improper distribution $F_1(T; \mathbf{Z})$ (Fine and Gray [1999]):

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\{\mathbf{z}_i^T(T_i)\beta\}}{\sum_{j \in R_i} \exp\{\mathbf{z}_j^T(T_i)\beta\}} \right]^{I(j_i=1)} \tag{2.14}$$

Then we can get the maximum likelihood estimate for β and derive the asymptotic normal distribution, which inherited from the ordinary Cox proportional hazard model.

2.2.2 Censoring complete data

In some designed clinical trials, censoring only results from administrative loss-to-follow up, which means the patients have not failed by the time the data are analysed.

Under this condition, the potential censoring time is always observed. We call these data **censoring complete**.

We then redefine the risk set at the time of failure for ith individual (Fine and Gray [1999]):

$$R_i = \{k : (C_k \wedge T_k \geq T_i) \cup (T_k \leq T_i \cap j_k \neq 1 \cap C_k \geq T_i)\}. \tag{2.15}$$

In this setting, an individual with $J \neq 1$ is still "at risk" for fail from cause of interest until censoring time C. If (T, J) and C are conditionally independent given covariates, then the "crude" subdistribution hazard function with censoring-complete data, $\lambda_{1*}\{t; \mathbf{z}\}$, is equivalent to the "net" subdistribution hazard function with complete data, $\lambda_1^*\{t; \mathbf{z}\}$ (Fine and Gray [1999]).

Using the censoring-complete risk set setting, the partial likelihood principle can again be applied to the model for $\lambda_1^*\{t; \mathbf{z}\}$. And the asymptotic results for the censoring-complete data estimation and prediction follow from the complete data derivation.

Chapter 3

A Pancreatic Cancer Study

3.1 Background

Pancreatic cancer is a malignant neoplasm originating from transformed cells arising in tissues forming the pancreas. The most common type of pancreatic cancer, constituting about 95% of these tumors, is adenocarcinoma appearing within the exocrine component of the pancreas. This kind of tumor exhibites glandular architecture on microscopy. The signs and symptoms, which eventually lead to the diagnosis depend on many factors, such as the location, the size, and the tissue type of the tumor. Other information related to physiological abnormality, including abdominal pain, lower back pain, jaundice, which may be caused if the tumor compresses the bile duct, unexplained weight loss, and digestive problems, are also considered.

According to World Health Organization, pancreatic cancer is the fourth most

common cause of cancer-related deaths in the United States and the twelfth most common cause of cancer-related deaths worldwide. Pancreatic cancer has an extremely poor prognosis: for all stages combined, the one-year and five-year relative survival rates are just 25% and 6% respectively; from American Cancer Society, for local disease the five-year survival is approximately 15% while the median survival for locally advanced and for metastatic disease, accounting for over 80% of individuals from National Cancer Institute, is about 10 and 6 months respectively. Individuals vary from each other. However some are diagnosed when they are already in stage IV, therefore only have a few days or weeks to live. Others, who have slower progression, may live a couple of years even if they cannot have the surgery. Men are 30% more likely to get pancreatic cancer than women.

Family history may be considered as a risk factor, since 5-10% of pancreatic cancer patients have a family history of pancreatic cancer (Ghaneh et al. [2007]). The risk of developing pancreatic cancer increases with age. Most cases occur after age 60, while cases before age 40 are rare. Smoking has a risk ratio of 1.74 with respect to pancreatic cancer; a decade of nonsmoking after heavy smoking is associated with a risk ratio of 1.2 (Iodice et al. [2008]). Obesity is also considered as a risk factor for pancreatic cancer (Society [2008]).

However, the prognosis of different patterns of recurrence, particularly in lung, for pancreatic cancer patients who had had surgery has not been well studied. The relationship between survival time and recurrence patterns may help us predict the

prognosis in the future and assign appropriate treatment to the patient given their recurrence pattern.

Here we have three recurrence types, one is having recurrence in lung before death, denoted as **recurrence-in-lung**; one is having recurrence in sites other than lung before death, denoted as **recurrence-in-other-sites**; and the third is not having recurrence before death, denoted as **no-recurrence**.

3.2 Summary of Baseline Covariates

The medical records of 209 patients who had surgical resection of pancreatic cancer and had postoperative follow-up primarily at the Johns Hopkins Hospital between 1998 and 2007 were retrospectively reviewed. Among the 209 patients, 13.4% had recurrence only in lung; about 70% of the patients had recurrence in sites other than lung; 16.74% of the patients did not have metastasis. We perform survival analysis on two types of survival outcomes. One is the time from the surgery to the date of first diagnosis of recurrence. However, the diagnosis of lung recurrence was often delayed. Therefore, the recurrence time we observed here might not be the true recurrence time. The other time period is the time from surgery to death or censoring. Age, gender, cancer staging, margins, lymph node, grading differentiation, vascular invasion, perineural invasion, therapy type are also provided in the data set. The summary information about these covariates are shown below in Table 1. The mean

age of the whole data set is 64.23, with standard deviation 10.94. The three subgroups, recurrence-in-lung, recurrence-in-other-sites and no-recurrence, have similar mean in age. However, the standard deviation of it for the patients with recurrence in lung is smaller, which means age in this subgroup is more concentrated around the mean. Given gender, the difference between the number of female and that of male is quite small in the whole data set, the-recurrence-lung subgroup and recurrence-in-other-sites subgroup. However, the number of male in no-recurrence subgroup is almost twice as large as that of female in this group.

Cancer staging is the process of determining the extent to which a cancer has developed by spreading. The larger the cancer staging is, the poor the prognosis. In Table 3.1, majority, above 90%, of the patients in the study in stage II, which is the moderate prognosis in cancer staging, and only 17 people are either in stage I or III. The three subgroups share similar patterns.

Margin refers to the edge or border of the tissue removed in cancer surgery. The margin is described as negative or clean when the pathologist finds no cancer cells at the edge of the tissue, suggesting that all of the cancer has been removed. The margin is described as positive or involved when the pathologist finds cancer cells at the edge of the tissue, suggesting that all of the cancer has not been removed. Usually positive margin means better prognosis. In Table 3.1, the number of patients with positive margins is close to that of patients with negative margins in the whole data set, which also happens in the recurrence-in-lung and no-recurrence subgroups. However, in the

recurrence-in-other-sites subgroup, 81 patients are with negative margins while only 65 with positive margins. We may want to consider this difference in our later model.

Tumor grade is a way of classifying tumors based on certain features of their cells. The grade of a tumor is directly linked to prognosis. It is to check how much the cancer cells look like normal cells: the more the cancer cells look like normal cells, the lower the tumor grade tends to be. It also consider how many of the cancer cells are in the process of dividing: the fewer cancer cells that are in the process of dividing, the more likely it is that the tumor is slow-growing slowly and the lower the tumor grade tends to be. Well-differentiated means the tumor cells look the most like normal tissue and are slow-growing, moderate-differentiated means the tumor cells fall somewhere in between Grade 1 and Grade 3, and poorly-differentiated means the tumor cells look very abnormal and are fast-growing. In Table 3.1, in the whole data set and also in three subgroups, most patients were poorly or moderate-differentiated. However, in recurrence-in-lung patients, 75% of those are moderate-differentiated, while 50.68% of the recurrence-in-other-sites patients and 37.14% of the no-recurrence patients are moderate-differentiated.

Lymph nodes refers to the indicator of whether the cancer has spread to lymph nodes. The prognosis is poorer if lymph node is positive, since the cancer cell can travel to the rest of the body by the lymph system. In Table 3.1, among all the patients, 86.12% are positive in lymph nodes, 92.86% in recurrence-in-lung, 85.62% in recurrence-in-other-sites, and 82.86% in no-recurrence. Carrying out chi-square

test, we find out the p-value is significant, which indicates the distributions of lymph node in four groups are different. With further examining, the distribution in the recurrence-in-other-sites is different from those of the other three groups.

Vascular invasion is the indicators that we have that cells have a tendency to go into the vascular system and to spread to the rest of the body. Most of the patients without information about vascular invasion were in subgroup of patients with recurrence-in sites other than lung. Perineural invasion, abbreviated PNI, refers to cancer spreading to the space surrounding a nerve. In Table 3.1, among all patients, recurrence-in-other-sites and no-recurrence, more than 90% are "Yes", while 85.71% in recurrence-in-lung are "Yes".

Therapy information provides us which therapy the patient had, radiotherapy, chemotherapy or both. In Table 3.1, in no-recurrence subgroup, about 40% of the patients lose information about the therapy. In other two subgroups, at least 76% of the patients have received the therapy. The p-values of the two therapies are significant, and this may due to the fact that about 40% people in the no-recurrence group do not have information about it, and the percentile is much higher than the other three groups.

3.3 Cox's Proportional Hazard Model Without Recurrence Type Information

3.3.1 Time from surgery to composite endpoint of recurrence and death

The first outcome of interest is time from surgery to composite endpoint, which includes recurrence only in lung, recurrence in sites other than lung and death without having recurrence. Conventional Cox's model with the main effects as covariates is considered here to discover the main effects that may effect the risk. The result is shown in Table 3.2. We can see both the linear and quadratic term of age have significant p-values, 0.03. From the estimate of coefficients, those youngest and oldest patients in the data tend to have a greater risk of the composite event, while the patients with age from 55 to 75 tend to have lower risk. Receiving radiation therapy is found to decrease the risk by 40%. Later, we will compare the coefficients with those we get from cause-specific hazard model in Section 3.4.2.

3.3.2 Time from surgery to death

In this section, we analyze time from surgery to death, but ignore the recurrence types here. The conventional Coxs model with same set of covariates is considered

here. The result is shown in Table 3.3. We can see the linear and quadratic term of age, gender and perneural invasion have significant p-values. Radiation therapy does not show significance here, different from that in Table 3.2. Age seems to follow similar pattern as in Section 3.3.1. The risk of death for those having perneural invasion is about three times as much as that for those not having. In Section 3.5.2, we will compare the results with the coefficients of main effects in the cause-specific hazard model.

3.4 Survival Analysis for Time from Surgery to Recurrence

In this section, we mainly deals with time from surgery to the time when the recurrence was first diagnosed or death without recurrence. We want to see how different baseline characteristics influence the risk of different recurrence types, and how the time from surgery to the recurrence first diagnosed varies across different recurrence types.

3.4.1 Cumulative incidence function

We estimate cumulative incidence function to discover whether recurrence type influences the survival outcome, the time from surgery to recurrence or death if there

was no recurrence. The cumulative incidence curve estimate was from the **cmprsks** cuminc() function (Gray [1988]). We can see from Figure 3.1 that the probability that patients were first diagnosed of recurrence in sites other than lung before time t increases much more rapidly when t goes from 0 to 12, reaches about 0.45 at the end of first year, and then the slope decreases. After 40 months, the line of the cumulative incidence function for recurrence in other sites becomes quite stable and is close to 0.70. The probability that patients were first diagnosed of recurrence only in lung before t increases most slowly among the three risk types when t is in $[0, 40]$. When t is 40, there is a cross-over between the line of recurrence-in-lung subgroup and no-recurrence group. We realize that few people survived very long in no-recurrence subgroup, and it may cause the cross-over here. Finally, the cumulative incident function of recurrence-in-lung and that of no-recurrence both go to about 0.15.

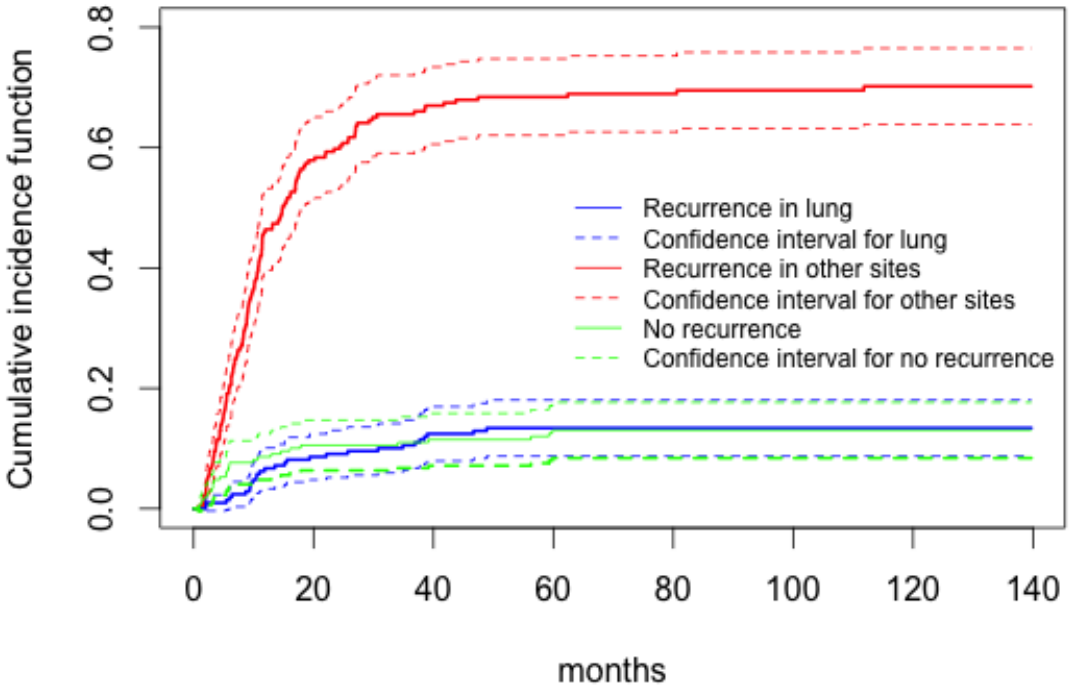

Figure 3.1: The nonparametric estimate of cumulative incidence function and 95% confidence intervals.

Since most patients in this study are in recurrence-in-other-sites subgroup (69.86%), the difference in cumulative incidence functions is larger because of prevalence rates of the failure events. Then we condition on that a failure type occurred during the study period to see if the conditional cumulative incidence functions for the three types are significantly different. The result is in Figure 3.2. The solid lines are the estimate of conditional cumulative incidence functions, and the dashed lines are the 95% confidence intervals of the estimate using the bootstrap resampling method.

From Figure 3.2, the conditional subdistribution of no-recurrence group increases most rapidly within the first 12 months, followed by that of recurrence-in-other-sites subgroup. Then slope of the curve with regards to no-recurrence, decreases as most patients died in this subgroup. The cumulative incidence curve of recurrence-in-lung subgroup increases most slowly within first 40 months. Since the sample sizes vary in different subgroups, the estimate of the cumulative incidence curve of the recurrence-in-other-sites subgroup, with the largest sample size, has the most narrow confidence interval, while the widths of the confidence intervals of other two subgroups are relatively large.

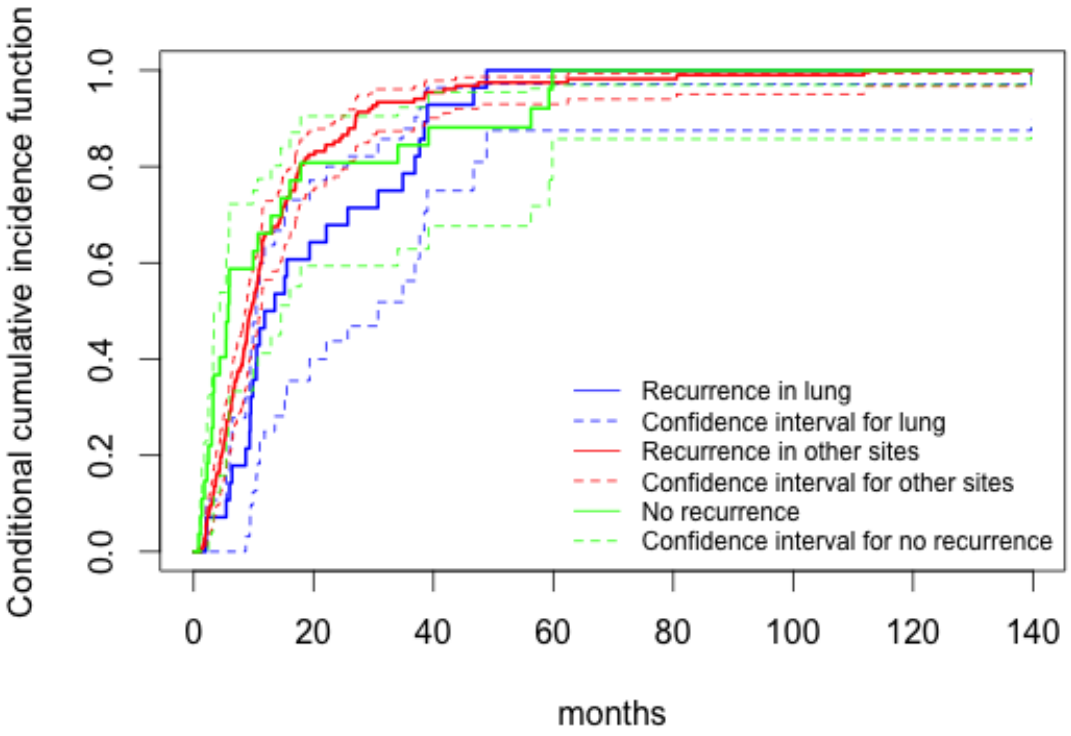

Figure 3.2: Estimated conditional cumulative incidence functions and 95% bootstrap confidence intervals.

3.4.2 Cause-specific hazard

We have reviewed the methods of cause-specific methods. From Section 2.1, we know the jth likelihood factor in the full likelihood function, is precisely the likelihood being obtained by regarding all other failure types as being censored. The usual survival data methods for a single failure type can be used for testing and estimating λ_j. Here we use Cox's proportional hazard model for

$$\lambda_j(t;\mathbf{z}) = \lambda_{0j}\exp(\mathbf{z}\beta_j), \qquad j = 1,\ldots,m. \tag{3.1}$$

3.4.2.1 With only main effects

First we only include main effects in each of the cause-specific hazard model, same covariates with the model in Section 3.3.1. In Table 3.3, none of the main effects in the cause-specific hazard for no-recurrence and recurrence-in-lung subgroups are significant. And only positive margin in the model for recurrence-in-lung subgroup has a p-value of 0.06, close to 0.05. This may due to the small sample size of these two subgroups. We note that the variances of stage, lymph node and chemo therapy in the recurrence-in-lung cause-specific hazard model are large, and so are the ranges of their 95% confidence intervals. This may due to small sample size , and also the unequal distribution of patients in each category of the covariate. For example, only one of the 28 patients in recurrence-in-lung subgroup is in Stage I, while all the other are in Stage II (Table 3.2). Similar thing happens to lymph node and radiation in the no-recurrence cause-specific hazard model too. However, in this model, the perneural invasion and stage has extreme large estimate and variance. When we check the data, we find out that only two people in this subgroup did not have perneural invasion, and they lived up to 133 and 129 months. Regarding the cancer staging information, only one person in no-recurrence subgroup was in Stage I, who lived up 129 months, while two people in the same group was in Stage III, and they lived 3.38 and 1.15

months respectively, both very short. Since these patients are sparse in the sample, it is not reasonable to include perneural invasion and stage in the model. While looking at the cause-specific hazard model for recurrence-in-other-sites, other than the three significant main effects in Table 3.2, we find out the gender is also a significant main effect.

3.4.2.2 With some of the main effects and interaction terms

By exploring different models, we fit proportional hazard model for the cause-specific hazard models for three recurrence types in Table 3.5.

In the recurrence-in-lung subgroup, Table 3.5, the positive margin and the interaction tern of the positive margin and gender have significant p-values, that are 0.006 and 0.03 respectively. Among the females who had recurrence in lung, the cause-specific hazard rate of those with positive margin is about 7 times greater than that of those with negative margin indicator. Among the males who had recurrence in lung, the cause-specific hazard rate of those with positive margins is 1.2 times as large as that of those with negative margins. In the recurrence-in-lung subgroup, the margin indicator influences females much more than males. Age, which is always an important characteristic in cancer study, does not seem to play an important role here, the coefficient associated with it is 1.03 with an insignificant p-value of 0.19. The appearance of cancer cells in lymph nodes increases the cause-specific hazard by 2.82 times, which is consistent with empirical facts. However its 95% confidence

interval is $[0.63, 12.74]$ and standard deviation is large, the accuracy of the estimate can not be assured. The similar circumstance happens to grade differentiation, vascular invasion, perineural invasion, chemo therapy and radiation therapy. The effect of these baseline characteristics are not certain, and more data are needed.

The Table 3.5 shows the results from estimate of the cause-specific hazard for recurrence-in-other-sites. Among all the baseline characteristics, age, gender grade differentiation, radiation therapy and interaction term of age and gender have significant p-values. Among patients having recurrence in sites other than lung, age follows the same pattern as that in Table 3.4. However, the effect of age in females is sightly different from that in males. The youngest and oldest patients are at more risk in this subgroup. The risks of females and males at the same age are also very different: the cause-specific hazard rate of the males is 25 times as large as that of the females. Besides, the hazard of the patients who were poorly-differentiated in recurrence-in-other-sites subgroup at time t is 1.86 times greater than that from the moderate-differentiated patients in the same subgroup. We also notice radiation therapy has a significant p-value while chemo therapy does not. The risk of the patients who had radiation therapy decreases by 87% comparing to those who did not. However, 76.71% of the patients in recurrence-in-other-sites received radiation therapy, and there may be the possibility that those, not receiving radiation therapy, died too early to receive therapy. Further tests need to be carried out to see if radiation therapy really helps to reduce risk. The baseline characteristic vascular invasion, though

its p-value is not significant, has a 95% confidence interval of $[0.98, 2.19]$, the lower end of which is very close to 1, it is reasonable to consider that it does have impact on the cause-specific hazard, and it increases the risk by 0.46 if the vascular invasion appears.

Among patients who had no recurrence, none of the baseline characteristics in our model are significant (see Table 3.5). This may result from the fact that almost 40% of the patients in this 35-patient subgroup had missing values in radiation therapy. The grade differentiation indicator has the smallest p-value of 0.07. When looking at its 95% confidence interval $[0.09, 1.09]$, the upper end of the interval is very close to 1. The differentiated grade of the cancer cell may have effect on the cause-specific hazard for patients not having recurrence. The risk of the patients poorly-differentiated in no-recurrence subgroup is 0.39 times greater than that of the moderate-differentiated patients in the same subgroup. The 95% intervals of lymph node and radiation therapy have large range, which may due to small sample size of this subgroup.

3.5 Survival Analysis for Time from Surgery to Death

In this section, we study time from surgery to death. The three types of recurrence patterns are defined same as in the previous section. Similar procedure is carried out as previous section.

3.5.1 Cumulative incidence function

Instead of using conventional survival analysis, we analyze the problem by regarding it as competing risks problem and tried to estimate cumulative incidence functions of the three subgroups. Based on the same methods as previous section, in Figure 3.3, the probability that patients not having recurrence before time t increases most rapidly when t is less than half a year, and has the value 0.08 at the end of the sixth month. Then its slope decreases. After 6 months, the cumulative incidence curve for recurrence in other sites increases the fastest. The probability of death for patients first diagnosed of recurrence only in lung before time point t increases most slowly among the three risk types. Finally, since recurrence-in-lung and no-recurrence subgroups have similar sample size, the cumulative incidence functions of them both go to about 0.1.

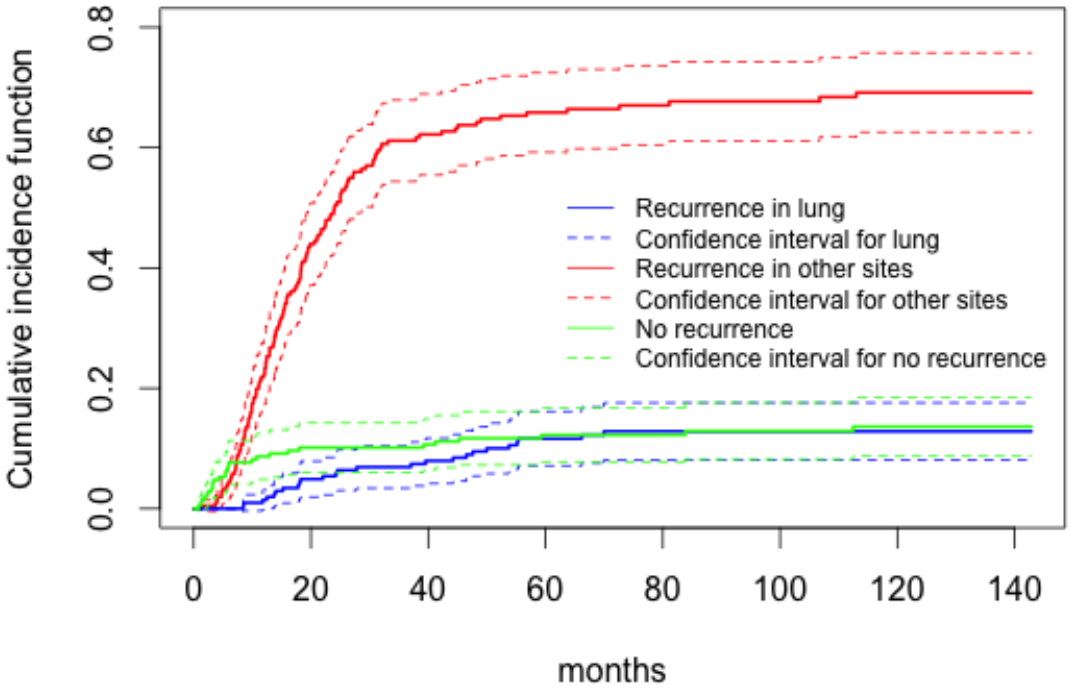

Figure 3.3: The nonparametric estimate of cumulative incidence function and 95% confidence intervals.

Since most patients in this study are in recurrence-in-other-sites subgroup (69.86%), the difference in cumulative incidence functions is larger due to prevalence rates of the failure events. Therefore, we condition on that a recurrence type occurred during the study period to see if the conditional cumulative incidence functions for three recurrence types are significantly different. The result is in Figure 3.4. The solid lines are the estimate of conditional cumulative incidence functions, and the dashed lines are the 95% confidence intervals of the estimate using the bootstrap resampling

method. From Figure 3.4, the conditional subdistribution of no-recurrence group increases most rapidly within the first 20 months, followed by that of recurrence-in-other-sites subgroup. Then, increase rate of the curve, with regards to patients not having recurrence, becomes slower as most patients died in this subgroup. Comparing to Figure 3.2 in Section 3.4.1, the difference between rate of cumulative incidence curve of no-recurrence subgroup and that of recurrence-in-other-sites becomes larger, since those who had recurrence in other sites were alive for a while after recurrence. The cumulative incidence curve of recurrence-in-lung subgroup increases most slowly. And because of difference in sample size in three subgroups, the estimate of the cumulative incidence curve of the recurrence-in-other-sites subgroup, which has the largest sample size, has the most narrow confidence interval, while the widths of the confidence intervals of other two subgroups are relatively large.

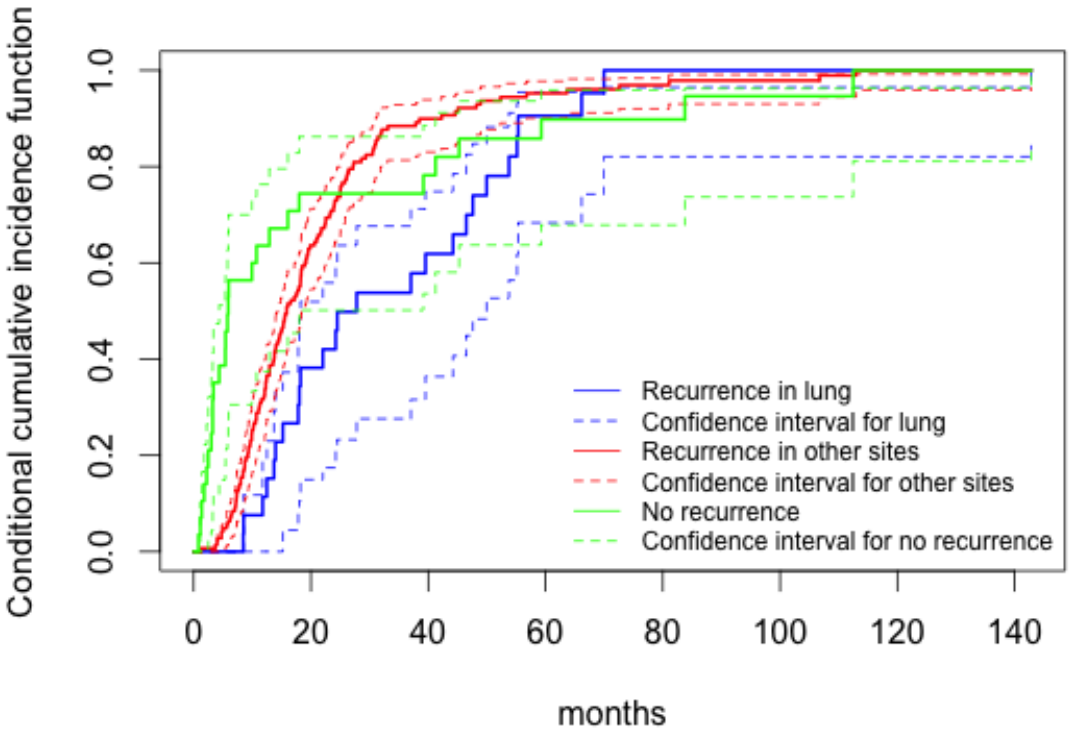

Figure 3.4: Estimated conditional cumulative incidence functions and 95% bootstrap confidence intervals.

3.5.2 Cause-specific hazard

3.5.2.1 With only main effects

First we only include main effects in each of the cause-specific hazard model. None of the main effects in the cause-specific hazard for no-recurrence and recurrence-in-lung subgroups are significant. And only positive margin in the model for recurrence-in-lung subgroup has a p-value of 0.08, close to 0.05. This may due to the small sample size of these two subgroups. We note that the main effects: lymph node,

perneural invasion and chemo therapy, in the recurrence-in-lung cause-specific hazard model have large variance, and the ranges of their 95% confidence intervals are large. Besides the small sample size problem, the unequal distribution of patients in each category of the main effect may also be a reason. For example, only two of the 28 patients in recurrence-in-lung subgroup did not have cancer cell appearing at lymph node, while all the other had. Similar thing happens to lymph node and radiation in the no-recurrence cause-specific hazard model too. However, in this model, estimate of coefficient of stage is really large, which is 17.02, and so is the variance. When we check the data, we find out that only one person in this subgroup was in stage I, and lived up to 142 months, while all other people in this subgroup were in stage II. Because of this sparsity problem, it is not reasonable to include stage in this cause-specific hazard model. While looking at the cause-specific hazard model for recurrence-in-other-sites, perneural invasion does show significance as it does in Table 3.6, but it is 0.07, close to 0.05.

3.5.2.2 With some of the main effects and interaction terms

Then we fit proportional hazard model for the cause-specific hazard models for three recurrence types in Table 3.7.

In the recurrence-in-lung subgroup, Table 3.7, the positive margin indicator and the interaction tern of the margin indicator and gender have significant p-values, that are 0.01 and 0.04 respectively. It agrees with the results in Table 3.5. Among the

females who had recurrence in lung, the cause-specific hazard rate of those whose margins were positive is 5.7 times greater than that of those with negative margin indicator, which is a little smaller than that in Table 3.5. Among the males who had recurrence in lung, the cause-specific hazard rate of those with positive margins is 1.02 times as large as that of those with negative margins, which indicates the margin indicator has less effect on males. Age, which is always an important characteristic in cancer study, does not seem to play a key role here, the coefficient associated with it is 1.01 with an insignificant p-value of 0.57, however, this may due to the small sample size. Though the coefficient of lymph nodes is 5.98, the 95% confidence interval of it is $[0.78, 46.08]$, which means the accuracy of the estimate can not be assured. The similar circumstance happens to grade differentiation, vascular invasion, perineural invasion, chemo therapy and radiation therapy. The effect of these baseline characteristics on cause-specific hazard are not certain, and more data are needed.

The Table 3.7 shows the results from estimate of the cause-specific hazard for recurrence-in-other-sites. Among all the baseline characteristics, age, age^2, gender, vascular invasion and the interaction term of age and gender have significant p-values, which are different from results in Table 3.5, where grade differentiation and radiation therapy are significant. Though the interaction term of age and gender has a significant p-value, the estimate coefficient is 1.05, which indicates the risks of females and males at the same age do vary much from each other. Among patients having recurrence in sites other than lung, the coefficients of age and its quadratic term are

close to those in Table 3.5. The effect of age in this subgroup regarding to different endpoints, recurrence and death, is similar. The risk of the patients who had vascular invasion in recurrence-in-other-sites subgroup at time t is 0.52 times greater than that of patients not having it in the same subgroup, with other covariates held. It agrees with the empirical fact that arise of vascular invasion often means poor prognosis.

Among patients who had no recurrence, Table 3.7, none of the baseline characteristics in our model are significant, which agreed with the results in Table 3.5. This may due to the same reason that almost 40% of the patients in this 35-patient subgroup had missing values in radiation therapy. Age has the smallest p-value of 0.06, and when looking at its 95% confidence interval $[0.36, 1.02]$, the upper end of the interval is very close to 1. Therefore, age may have effect on the risk of patients not having recurrence before death. Similar thing happens to age^2. Perneural invasion, lymph node and radiation have large standard deviations, and this may due to small sample size of this subgroup.

3.5.3 Cox's proportional hazard model with time-dependent covariates

In this subsection, instead of using competing risks models, we set no-recurrence as our reference group, import two time-depend risk indicators for recurrence-in-lung and recurrence-in-other sites, and then apply Cox's proportional hazard model to time from surgery to death or censoring including all the baseline characteristics.

The results are in Table 3.8. The estimate of coefficient related to time-depend risk-indicator of recurrence-in-lung is 0.57 with a p-value of 0.04, which means that the hazard of those who discovered lung recurrence decreases by 43% comparing to those who did not have recurrence in lung, given that all other covariates are the same. The estimate of coefficient related to time-depend risk-indicator of recurrence-in-other-sites is 1.82 with a p-value of 0.002. Then the hazard of those who had recurrence in sites other than lung increases by 82% comparing to that of those who did not have, with all other covariates held. We also consider interaction terms of gender and time-depend risk indicator. The p-values of these interaction terms are not significant, though the estimate values themselves indicates the risk of males having recurrence event increases comparing to those not have recurrence. The quadratic term of age is included in the model, since we discover the youngest patients and oldest patients in the study are likely to die early, while patients from 55 to 75 are most likely to live longer. Among baseline characteristics, margin indicator and

perineural invasion indicator have significant p-values of 0.01 and 0.001 respectively. The hazard of the patients with positive margins increases by 42% comparing to that of patients with negative margins, when holding all other covariates in the model. Also, the patients having perinueral invasion have hazard 2.77 times as large as those without perineural invasion, with the other covariates held.

Furthermore, we study whether the time to have recurrence in lung influence the risk of the patient. We includes two time-depend risk indicators related to recurrence in lung: one indicates whether the patient had diagnosed of recurrence in lung within 6 months after the surgery, and the other indicates whether the diagnosis occurred longer than 6 months after surgery. In the Table: 3.9. recurrence in lung occurring within 6 months after surgery increases the risk by 11%, while recurrence in lung occurring longer than 6 months after surgery decreases the risk by 55%. However, neither of the p-values are significant, it may be the reason that the sample size of patients who only had recurrence in lung is too small. We may want to study more patients who had recurrence only in lung to see if this difference actually exists.

Table 3.1: Summary Table of The Baseline Characteristics

covariates	whoel.data.set	recurrence.in.lung	recurrence.in.other.sites	no.recurrence	p-value
number of patients(%)	209(100)	28(13.40)	146(69.86)	35(16.74)	
Age(SD)	64.23(10.94)	65.25(8.50)	63.38(11.13)	66.97(11.60)	
gender					0.2512
Male(%)	100(47.85)	15(53.57)	71(48.63)	23(65.71)	
Female(%)	109(52.15)	13(46.43)	75(51.37)	12(34.29)	
cancer staging					0.5706
I(%)	11(5.26)	1(3.57)	9(6.16)	1(2.86)	
II(%)	191(91.39)	27(96.43)	133(91.1)	31(88.57)	
III(%)	6(2.87)	0(0)	4(2.74)	2(5.71)	
Unknown(%)	1(0.48)	0(0)	0(0)	1(2.86)	
margins					0.5266
Postive(%)	100(47.85)	15(53.57)	65(44.52)	20(57.14)	
Negative(%)	109(52.15)	13(46.43)	81(55.48)	15(42.86)	
Lymph Nodes					0.001
Yes(%)	180(86.12)	26(92.86)	125(85.62)	29(82.86)	
No(%)	29(13.88)	2(7.14)	21(14.38)	6(17.14)	
Grade Differentiation					0.06209
Poor(%)	108(51.67)	7(25)	68(46.58)	21(60)	
Moderate(%)	94(44.98)	21(75)	74(50.68)	13(37.14)	
Well(%)	3(1.44)	0(0)	4(2.74)	0(0)	
Unknown(%)	4(1.91)	0(0)	0(0)	1(2.86)	
Vascular Invasion					0.5199
Yes(%)	103(49.28)	13(46.43)	71(48.63)	19(54.29)	
No(%)	81(38.76)	13(46.43)	53(36.3)	15(42.86)	
Unknown(%)	25(11.96)	2(7.14)	22(15.07)	1(2.86)	
Perineural Invasion					0.9338
Yes(%)	192(91.87)	24(85.71)	136(93.15)	32(91.43)	
No(%)	12(5.74)	3(10.71)	7(4.79)	2(5.71)	
Unknown(%)	5(2.39)	1(3.57)	3(2.05)	1(2.86)	
Radiation Therapy					0.001286
Yes(%)	156(74.64)	24(85.71)	112(76.71)	20(57.14)	
No(%)	25(11.96)	3(10.71)	20(13.7)	2(5.71)	
Unkown(%)	28(13.4)	1(3.57)	14(9.59)	13(37.14)	
Chemo Therapy					$3.098e^{-06}$
Yes(%)	165(78.95)	26(92.86)	120(82.19)	19(54.29)	
No(%)	20(9.57)	1(3.57)	17(11.64)	2(5.71)	
Unkown(%)	24(11.48)	1(3.57)	9(6.16)	14(40)	

Table 3.2: Proportional Hazard Model of Time from Surgery to Composite Endpoint with Only Main Effects

	coef	exp(coef)	se(coef)	p	lower .95	upper .95
age	-0.20	0.82	0.09	**0.03**	0.68	0.98
age × age	0.002	1.002	0.0007	**0.03**	1.0002	1.0031
gender	-0.33	0.72	0.18	0.06	0.50	1.02
positive margin	0.30	1.34	0.18	0.10	0.94	1.91
lymph node	0.45	1.57	0.30	0.13	0.88	2.81
grade	-0.23	0.79	0.17	0.17	0.57	1.11
vascular invasion	0.22	1.24	0.17	0.21	0.88	1.75
perneural invasion	0.66	1.94	0.37	0.08	0.93	4.03
chemo	-0.29	0.75	0.43	0.50	0.32	1.74
radiation	-0.88	0.41	0.39	**0.02**	0.19	0.88
stage	0.02	1.02	0.48	0.96	0.40	2.61

Table 3.3: Proportional Hazard Model of Time from Surgery to Death with Only Main Effects, Ignoring the Recurrence Types

	coef	exp(coef)	se(coef)	p	lower .95	upper .95
age	-0.25	0.78	0.09	**0.01**	0.65	0.94
age × age	0.0002	1.00	0.0008	**0.01**	1.0005	1.0035
gender	-0.37	0.69	0.19	**0.05**	0.48	1.0006
positive margin	0.29	1.34	0.18	0.10	0.94	1.91
lymph node	0.43	1.53	0.30	0.16	0.85	2.77
grade	-0.05	0.95	0.17	0.75	0.68	1.32
vascular invasion	0.16	1.17	0.18	0.38	0.82	1.66
perneural invasion	1.11	3.04	0.42	**0.01**	1.33	6.95
chemo	-0.14	0.87	0.43	0.73	0.37	2.00
radiation	-0.26	0.77	0.39	0.51	0.36	1.67
stage	-0.03	0.97	0.50	0.95	0.36	2.58

Table 3.4: Cause-Specific Hazard of Time from Surgery to Recurrence With Only Main Effects

	coef	exp(coef)	se(coef)	p	lower .95	upper .95
Recurrence-in-lung						
age	0.08	1.09	0.31	0.79	0.59	2.01
age × age	-0.0004	1.00	0.0002	0.87	0.99	1.0004
gender	-0.25	0.78	0.45	0.58	0.32	1.88
positive margin	0.85	2.34	0.45	**0.06**	0.96	5.71
lymph node	1.01	2.75	1.07	0.35	0.34	22.57
grade	0.58	1.79	0.44	0.19	0.75	4.26
vascular invasion	0.07	1.07	0.41	0.87	0.48	2.37
perneural invasion	0.45	1.57	0.70	0.52	0.40	6.13
chemo	1.26	3.53	1.44	0.38	0.21	59.71
radiation	-1.08	0.34	1.06	0.31	0.04	2.70
stage	-0.15	0.86	1.47	0.92	0.05	15.14
Recurrence-in-other-sites						
age	-0.23	0.79	0.10	**0.02**	0.65	0.97
age × age	0.00	1.00	0.00	**0.03**	1.00	1.00
gender	-0.43	0.65	0.21	**0.04**	0.44	0.98
positive margin	0.18	1.19	0.21	0.39	0.80	1.78
lymph node	0.35	1.42	0.33	0.29	0.74	2.71
grade	-0.34	0.71	0.19	0.08	0.48	1.04
vascular invasion	0.29	1.34	0.20	0.15	0.90	1.99
perneural invasion	0.64	1.89	0.46	0.16	0.78	4.62
chemo	-0.34	0.71	0.47	0.46	0.28	1.77
radiation	-0.93	0.39	0.42	**0.03**	0.17	0.90
stage	-0.01	0.99	0.52	0.98	0.35	2.75
No-recurrence						
age	-0.27	0.77	0.32	0.40	0.41	1.43
age × age	0.00	1.00	0.00	0.34	1.00	1.01
gender	0.33	1.39	0.74	0.66	0.32	5.92
positive margin	0.42	1.52	0.67	0.54	0.41	5.68
lymph node	0.84	2.31	0.99	0.40	0.33	16.06
grade	-0.83	0.44	0.66	0.21	0.12	1.60
vascular invasion	0.07	1.07	0.66	0.91	0.29	3.92
perneural invasion	18.23	$8.28e^7$	$7.85e^3$	1.00	0.00	Inf
chemo	-1.82	0.16	1.97	0.35	0.003	7.60
radiation	-0.50	0.61	1.98	0.80	0.01	29.50
stage	1.19	3.30	3.34	0.72	0.0005	2291.84

Table 3.5: Cause-Specific Hazard of Time from Surgery to Recurrence With Some of the Main Effects and Interaction Terms

	coef	exp(coef)	se(coef)	p	lower .95	upper .95
Recurrence-in-lung						
age	0.03	1.03	0.02	0.19	0.98	1.08
gender	0.71	2.03	0.68	0.30	0.53	7.72
positive margin	2.09	8.09	0.76	**0.006**	1.82	35.98
lymph node	1.04	2.82	0.77	0.18	0.63	12.74
grade	0.67	1.94	0.45	0.14	0.81	4.70
vascular invasion	0.08	1.08	0.40	0.84	0.49	2.39
perneural invasion	0.36	1.43	0.69	0.60	0.37	5.48
chemo	1.56	4.77	1.49	0.30	0.26	89.13
radiation	-0.90	0.41	1.09	0.41	0.05	3.46
positive margin × gender	-1.91	0.15	0.90	**0.03**	0.03	0.86
Recurrence-in-other-sites						
age	-0.28	0.76	0.10	**0.006**	0.62	0.92
age× age	0.002	1.00	0.00	**0.01**	1.00	1.00
gender	-3.14	0.04	1.19	**0.008**	0.004	0.44
positive margin	0.17	1.19	0.21	0.41	0.79	1.79
lymph node	0.37	1.44	0.28	0.19	0.84	2.48
grade	-1.04	0.35	0.52	**0.04**	0.13	0.98
vascular invasion	0.38	1.46	0.21	0.07	0.98	2.19
perneural invasion	0.66	1.93	0.46	0.15	0.79	4.72
chemo	-0.68	0.51	0.49	0.16	0.20	1.32
radiation	-2.04	0.13	0.85	**0.02**	0.02	0.69
age × gender	0.04	1.04	0.02	**0.02**	1.01	1.08
radiation × grade	0.76	2.14	0.55	0.17	0.73	6.31
No-recurrence						
age	0.03	1.04	0.04	0.35	0.96	1.11
gender	0.42	1.53	0.72	0.56	0.37	6.28
positive margin	0.61	1.83	0.69	0.38	0.48	7.06
lymph node	0.72	2.06	0.87	0.40	0.38	11.28
grade	-1.14	0.32	0.63	0.07	0.09	1.09
vascular invasion	0.01	1.01	0.63	0.98	0.30	3.47
chemo	-2.28	0.10	1.92	0.24	0.00	4.42
radiation	-0.39	0.68	1.96	0.84	0.01	31.28

Table 3.6: Cause-Specific Hazard of Time from Surgery to Death with Only Main Effects

	coef	exp(coef)	se(coef)	p	lower .95	upper .95
Recurrence-in-lung						
age	0.06	1.07	0.31	0.84	0.58	1.97
age × age	-0.00	1.00	0.00	0.90	0.99	1.00
gender	-0.53	0.59	0.46	0.25	0.24	1.46
positive margin	0.76	2.14	0.44	0.08	0.91	5.02
lymph node	1.14	3.12	1.07	0.29	0.39	25.13
grade	0.72	2.06	0.45	0.11	0.85	4.97
vascular invasion	-0.14	0.87	0.42	0.75	0.38	2.00
perneural invasion	1.12	3.05	0.83	0.18	0.60	15.48
chemo	0.55	1.74	1.37	0.69	0.12	25.74
radiation	-0.14	0.87	1.03	0.89	0.12	6.58
stage	17.02	$2.468e^{7}$	$5.597e^{3}$	0.9976	0.00	Inf
Recurrence-in-other-sites						
age	-0.22	0.80	0.10	**0.03**	0.65	0.98
age × age	0.00	1.00	0.00	**0.04**	1.00	1.00
gender	-0.47	0.62	0.21	**0.02**	0.42	0.94
positive margin	0.31	1.36	0.20	0.13	0.92	2.01
lymph node	0.24	1.28	0.33	0.46	0.67	2.42
grade	-0.16	0.85	0.19	0.38	0.59	1.22
vascular invasion	0.34	1.41	0.20	0.09	0.95	2.09
perneural invasion	0.83	2.30	0.46	0.07	0.93	5.67
chemo	-0.14	0.87	0.45	0.75	0.36	2.10
radiation	-0.43	0.65	0.43	0.31	0.28	1.50
stage	0.06	1.06	0.54	0.91	0.37	3.05
No-recurrence						
age	-0.48	0.62	0.26	0.06	0.37	1.03
age × age	0.00	1.00	0.00	0.06	1.00	1.01
gender	0.06	1.07	0.66	0.92	0.29	3.90
positive margin	0.29	1.34	0.57	0.61	0.44	4.10
lymph node	1.29	3.64	0.90	0.15	0.63	21.12
grade	-0.68	0.50	0.59	0.25	0.16	1.60
vascular invasion	-0.73	0.48	0.57	0.20	0.16	1.46
perneural invasion	0.94	2.56	0.97	0.33	0.38	17.07
chemo	-1.55	0.21	1.54	0.32	0.01	4.36
radiation	0.27	1.31	1.46	0.85	0.07	22.93
stage	-0.77	0.46	1.45	0.60	0.03	7.93

Table 3.7: Cause-Specific Hazard of Time from Surgery to Death With Some of the Main Effects and Interaction Terms

	coef	exp(coef)	se(coef)	p	lower .95	upper .95
Recurrence-in-lung						
age	0.01	1.01	0.03	0.57	0.97	1.07
gender	0.50	1.65	0.69	0.47	0.42	6.45
positive margin	1.91	6.77	0.74	0.01	1.57	29.11
lymnode	1.79	5.98	1.04	0.09	0.78	46.08
grade	0.54	1.72	0.43	0.21	0.74	4.04
vascular invasion	-0.15	0.86	0.41	0.71	0.38	1.92
perneural invasion	0.90	2.47	0.81	0.26	0.51	11.99
chemo	0.71	2.04	1.40	0.61	0.13	31.86
radiation	-0.06	0.94	1.05	0.95	0.12	7.36
positive margin × gender	-1.90	0.15	0.93	0.04	0.02	0.93
Recurrence-in-other-sites						
age	-0.27	0.76	0.10	**0.01**	0.63	0.93
age× age	0.002	1.00	0.00	**0.02**	1.00	1.00
gender	-3.32	0.04	1.19	**0.01**	0.00	0.37
positive margin	0.36	1.44	0.20	0.07	0.97	2.14
lymph node	0.26	1.29	0.28	0.35	0.75	2.22
grade	-0.17	0.84	0.18	0.34	0.59	1.20
vascular invasion	0.42	1.52	0.20	**0.04**	1.02	2.26
perneural invasion	0.83	2.28	0.46	0.07	0.93	5.62
chemo	-0.20	0.82	0.46	0.66	0.33	2.02
radiation	-0.49	0.61	0.43	0.25	0.26	1.42
age × gender	0.05	1.05	0.02	**0.01**	1.01	1.08
No-recurrence						
age	-0.45	0.64	0.25	0.07	0.39	1.04
age× age	0.0003	1.00	0.0003	0.07	1.00	1.01
gender	0.03	1.03	0.66	0.97	0.28	3.74
positive margin	0.31	1.37	0.57	0.58	0.45	4.17
lymph node	1.05	2.86	0.75	0.16	0.65	12.49
grade	-0.61	0.54	0.58	0.29	0.18	1.67
vascular invasion	-0.66	0.52	0.55	0.23	0.18	1.51
perneural invasion	0.77	2.15	0.89	0.39	0.38	12.18
chemo	-1.57	0.21	1.53	0.31	0.01	4.19
rad	0.28	1.32	1.47	0.85	0.07	23.42

Table 3.8: Cox's Proportional Hazard with Time-Dependent Covariates

	coef	exp(coef)	se(coef)	p	lower .95	upper .95
lung	-0.56	0.57	0.27	**0.04**	0.34	0.97
other	0.60	1.82	0.19	**0.002**	1.25	2.64
age	-0.14	0.87	0.07	**0.03**	0.76	0.99
age × age	0.001	1.00	0.00	**0.04**	1.00	1.00
gender	-0.31	0.74	0.16	0.06	0.54	1.01
stage	-0.08	0.93	0.33	0.82	0.49	1.77
positive margin	0.35	1.42	0.13	**0.01**	1.11	1.82
lymnode	0.29	1.34	0.20	0.15	0.90	2.00
grade	0.04	1.04	0.12	0.75	0.83	1.30
vascular invasion	0.03	1.04	0.12	0.78	0.81	1.32
perneural invasion	1.02	2.77	0.29	**0.001**	1.56	4.92
chemo	0.39	1.48	0.32	0.23	0.79	2.77
radiation	-0.40	0.67	0.29	0.17	0.38	1.19
gender × other	0.28	1.33	0.26	0.27	0.80	2.20
gender × lung	0.26	1.30	0.36	0.47	0.64	2.65

Table 3.9: Cox's Proportional Hazard with Time-Dependent Covariates Regarding Time of Recurrence

	coef	exp(coef)	se(coef)	p	lower .95	upper .95
lung-recurrence $\leq$ 6 months	0.10	1.11	0.26	0.69	0.67	1.83
lung-recurrence $>$ 6 months	-0.81	0.45	0.62	0.19	0.13	1.51
other	0.67	1.95	0.14	$\mathbf{8.79}e^{-7}$	1.49	2.54
age	-0.14	0.87	0.07	**0.03**	0.77	0.99
age $\times$ age	0.001	1.00	0.0005	0.05	1.00	1.00
gender	-0.14	0.87	0.13	0.25	0.67	1.11
stage	0.05	1.05	0.32	0.88	0.56	1.99
positive margin	0.33	1.39	0.13	**0.01**	1.08	1.79
lymnode	0.28	1.32	0.20	0.17	0.89	1.98
grade	0.003	1.00	0.11	0.98	0.80	1.26
vascular invasion	0.02	1.02	0.12	0.88	0.80	1.30
perneural invasion	0.93	2.54	0.29	**0.001**	1.43	4.51
chemo	0.30	1.35	0.31	0.34	0.73	2.48
radiation	-0.33	0.72	0.30	0.26	0.40	1.28

Chapter 4

Discussion

To study time from surgery to composite endpoints, recurrence or death, we analyze the data under competing risks format. The unconditional CIF, though showing large difference, is hard to make inference, since about 70% patients are in the recurrence-in-other-sites subgroup. Instead, we estimate conditional CIF, conditioning on recurrence type. The difference still exists, especially in the first 30 months. However, large standard deviation problem occurres in the recurrence-in-lung and no-recurrence subgroups, which have smaller sample size. To analyze the effect of main effects on the risk, we first use conventional Cox's model with all the main effects as covariates, but ignoring the recurrence type information. Then we include same main effects in the cause-specific hazard models. The cause-specific hazard model does not find any main effects significant when analyzing recurrence-in-lung and no-recurrence subgroup. However, when regarding time from surgery to recurrence for recurrence-

in-other-sites subgroup, other than the significant ones in conventional model, new significant main effect, gender, is found. And when analyzing time from surgery to death for the same subgroup, perneural invasion, significant in conventional model, is not significant here. By making inference about the cause-specific model with main effects and exploring interaction relationships between main effects, we get our final cause-specific hazard models for three recurrence types. The cause-specific hazards find out some covariates that influenced the risk, the significant ones. And different cause-specific hazards have different significant covariates. However, the sample size of two subgroups, recurrence-in-lung and no-recurrence, is very small, resulting in the problem that some covariates have large variances. We cannot make conclusion about how these covariates effect the risk, because of the uncertainty. To overcome the small sample size problem in two subgroups, we, instead, utilize Cox proportional hazard model based on the whole data set and include time-depend risk indicator to test whether certain recurrence type effects patient's risk. The result shows that recurrence-in-lung does decrease patient's risk. Further exploratory shows that diagnosis of recurrence in lung within 6 months after surgery increases the risk, while diagnosis of recurrence in lung after 6 months after surgery decreases risk. However, due to the small number of cases having recurrence in lung, more evidence is needed to test if early recurrence in lung actually increase the risk, while only recurrence in lung occurred after certain time decreases the risk.

The prognosis of different recurrence, particularly in lung, has not been carefully

studied yet. Though it's a quite a new topic, the meaning of it is profound. We want to study the difference in survival or hazard between different recurrence, the pattern of time to different recurrence, what characteristics may influence the pattern of recurrence or survival of various recurrence types. Trying to answer these questions, we use competing risks models. Though "competing risks" refers to the study of any failure type in which there is more than one distinct type of failure, as mentioned before, our settings here are different from conventional competing risks settings. The patients in the study died of cancer, but different in recurrence types. We treat different recurrence events as competing risks, and then apply competing risks models. Our study shows that patients having had recurrence lung survives longer than patients with recurrence in sites other than lung or not having recurrence before death. Moreover, the period from surgery to their diagnosis of recurrence in lung is also longer. Patients not having recurrence before death often die very quickly after surgery. There may be genetic or psychological features in these patients, which effect the recurrence type and also survival. Finding out these features will help doctors make better prediction of the patients, and select appropriate treatment to increase survival.

We study time from surgery to recurrence and time from surgery to death, but do not study the time from recurrence to death. This will be an interesting topic. By careful study of the time from recurrence to death, we may be able to predict the survival time when we discover certain recurrence types in patients.

Appendix A

R Code

```
data <- read.csv("AllDataClean.csv", head=T)

data <- data[,-c(3,4)]

cov1 <- data[,c(2:13)]

levels(cov1$Sex) <- c("F", "F", "M", "M")

levels(cov1$Stage) <- c(" ", "I", "II", "1A", "1B", "2B",

                        "2B", "I", "II", "III", "II", "I", "II", "III")

levels(cov1$Vascular.Invasion) <- c(" ", "N", "N", "Unknown",

                                    "Unknown", "Y", "Y")

levels(cov1$PerineuralInvasion) <- c(" ", "N", "N", "Unknown",

                                     "Unknown", "Y", "Y")

levels(cov1$Adjuvant.RadiationTherapy) <- c("N", "N", "Unknown",

                                            "Unknown", "Y", "Y", "Y")
```

```
levels(cov1$Adjuvent.Chemo) <- c("N", "N", "N", "Unknown", "Unknown",
                                 "Y", "Y", "Y","Y")
cov1 <- data.frame(cov1, as.factor(stage), as.factor(grade))
summary(cov1)


##covariates used in model
age <- as.vector(data$Age)
gender <- as.vector(data$Gender)
stage <- data$Stage123
marg.pos <- data$Margin.pos
lymnode <- data$LymNodes
grade <- data$Grade
vas.inv <- data$Vas.inv
per.inv <- data$Peri.inv
chemo <- data$AdjChemo
rad <- data$AdjRad


covariate <- data.frame(age, gender, stage, marg.pos,
                        lymnode, grade, vas.inv, per.inv, chemo, rad)
cov <- as.matrix(covariate)
```

```
##time to recurrence or time to death

recur <- c()

recur[data$Recur == "LungOnly"] = 1

recur[data$Recur == "LungOther" | data$Recur == "liver" |

        data$Recur == "Local" | data$Recur == "Peritoneal"] = 2

recur[data$Recur == "NoRecur"] = 0

recur = as.numeric(recur)


##time to recurrence and time to death

X <- data$surg.recur

X_censor <- data$censor.DFS

T <- data$time.OS

T_censor <- data$censor.OS


lung_recur <- X[recur==1]

lung_death <- T[recur==1]

other_recur <- X[recur==2]

other_death <- T[recur==2]

no_recur <- X[recur==0]


library("survival")
```

```
##Cox for T ignoring the recurrence type on main effects


cox0 <- coxph(Surv(time=T, event=T_censor) ~ age + I(age^2) + gender +
                marg.pos + lymnode + grade  + vas.inv  + per.inv +
                chemo  + rad + stage, data = covariate)
summary(cox0)


##cause-specific hazard for T
##lung only
t1_censor <- T_censor
for (i in 1:209){
  if(recur[i] != 1){
    t1_censor[i] = 0
  } else
    t1_censor[i] = 1
}
cox1.0 <- coxph(Surv(time=T, event=t1_censor) ~ age + I(age^2) + gender +
                  marg.pos + lymnode + grade  + vas.inv  + per.inv +
                  chemo  + rad + stage, data = covariate)
summary(cox1.0)
```

```
cox1 <- coxph(Surv(time=T, event=t1_censor) ~ age + gender +
                marg.pos + lymnode + grade  + vas.inv + per.inv +
                chemo + rad  + I(marg.pos*gender) , data = covariate)
sumamry(cox1)


##other
t2_censor <- T_censor
for (i in 1:209){
  if(recur[i] != 2){
    t2_censor[i] = 0
  } else
    t2_censor[i] = 1
}
cox2.0 <- coxph(Surv(time=T, event=t2_censor) ~ age + I(age^2) + gender +
                  marg.pos + lymnode + grade  + vas.inv  + per.inv +
                  chemo  + rad + stage, data = covariate)
summary(cox2.0)


cox2 <- coxph(Surv(time=T, event=t2_censor) ~ age + I(age^2) + gender +
                marg.pos + lymnode + grade  + vas.inv + per.inv +
                chemo + rad + I(age*gender), data = covariate)
```

```
summary(cox2)


##no recur
t3_censor <- T_censor
for (i in 1:209){
  if(recur[i] != 0){
    t3_censor[i] = 0
  } else
    t3_censor[i] = 1
}


cox3.0 <- coxph(Surv(time=T, event=t3_censor) ~ age + I(age^2) + gender +
                  marg.pos + lymnode + grade  + vas.inv  + per.inv +
                  chemo  + rad + stage, data = covariate)
summary(cox3)


cox3 <- coxph(Surv(time=T, event=t3_censor) ~ age + I(age^2) + gender +
                marg.pos + lymnode + grade  + vas.inv  + per.inv +
                chemo + rad, data = covariate)
summary(cox3)
```

```
##cause-specific hazard for X

##lung only

x1_censor <- X_censor

for (i in 1:209){

  if (recur[i] != 1)

    x1_censor[i] = 0

}

cox1.2 <- coxph(Surv(time=X, event=x1_censor) ~ age + I(age^2) + gender +

                  marg.pos + lymnode + grade  + vas.inv  + per.inv +

                  chemo  + rad + stage, data = covariate)

summary(cox1.2)


cox1.3 <- coxph(Surv(time=X, event=x1_censor) ~ age+ I(age^2) + gender +

                  marg.pos + lymnode + grade  + vas.inv  + per.inv +

                  chemo + rad + I(marg.pos*gender), data = covariate)

summary(cox1.3)


##other

x2_censor <- X_censor

for (i in 1:209){
```

```
  if (recur[i] != 2)
    x2_censor[i] = 0
}

cox2.2 <- coxph(Surv(time=X, event=x2_censor) ~ age + I(age^2) + gender +
                  marg.pos + lymnode + grade  + vas.inv  + per.inv +
                  chemo  + rad + stage, data = covariate)
summary(cox2.2)

cox2.3 <- coxph(Surv(time=X, event=x2_censor) ~ age + I(age^2) + gender +
                  marg.pos + lymnode + grade  + vas.inv  + per.inv +
                  chemo + rad + I(age*gender) + I(rad*grade), data = covariate)
cox2.3

##no recur
x3_censor <- X_censor
for (i in 1:209){
  if (recur[i] != 0)
    x3_censor[i] = 0
}
```

```
cox3.3 <- coxph(Surv(time=X, event=x3_censor) ~ age + I(age^2) + gender +
                  marg.pos + lymnode + grade  + vas.inv  + per.inv +
                  chemo  + rad + stage, data = covariate)
summary(cox3.3)


cox3.1 <- coxph(Surv(time=X, event=x3_censor) ~ age  + gender +
                  marg.pos + lymnode + grade  + vas.inv  +
                  chemo + rad, data = covariate)
cox3.1


##CIF for T
library("cmprsk")


##failure status (0 = censoring, 1 = no recur, 2 = lung, 3 = others)
censor <- recur + 1
for(i in 1:209){
  if(T_censor[i] == 0)
    censor[i] = 0
}


CIF <- cuminc(T, censor)
```

```
##unconditional cumulative incidence curves

##full time point

tl <- CIF[2][[1]]$time

to <- CIF[3][[1]]$time

t_no <- CIF[1][[1]]$time


estl <- CIF[2][[1]]$est

esto <- CIF[3][[1]]$est

est_no <- CIF[1][[1]]$est


varl <- CIF[2][[1]]$var

varo <- CIF[3][[1]]$var

var_no <- CIF[1][[1]]$var


upper_lung <- estl + 1.96 * sqrt(varl)

lower_lung <- estl - 1.96 * sqrt(varl)


upper_other <- esto + 1.96 * sqrt(varo)

lower_other <- esto - 1.96 * sqrt(varo)
```

```
upper_no <- est_no + 1.96 * sqrt(var_no)

lower_no <- est_no - 1.96 * sqrt(var_no)


plot(to, esto, col="red", type="s", lty=1, xlab="months", ylab="probability of dea

     ylim=c(0,0.8), lwd=2)

lines(to, lower_other, lty=2, col="red", type="s")

lines(to, upper_other, lty=2, col="red", type="s")


lines(tl, estl, lty=1, col="blue", type="s", lwd=2)

lines(tl, lower_lung, lty=2, col="blue", type="s")

lines(tl, upper_lung, lty=2, col="blue", type="s")


lines(t_no, est_no, col="green", type="s", lwd=2)

lines(t_no, lower_no, lty=2, col="green", type="s")

lines(t_no, upper_no, lty=2, col="green", type="s")


legend(60, 0.55, legend=c("Recurrence in lung","Confidence interval for lung",

                          "Recurrence in other sites",

                          "Confidence interval for other sites",

                          "No recurrence",

                          "Confidence interval for no recurrence"),
```

```
       col=c("blue", "blue", "red", "red", "green", "green"), lty=c(1,2, 1, 2, 1,
       bty="n", cex=.75)


##doing bootstrap for recurrence on lung and recurrence in others
##function for estiamte in bootstrap
cif_est <- function(data, tl, to, t_no){
  time <- data[,1]
  censor <- data[,2]
  cif <- cuminc(time, censor)
  time_lung <- cif[2][[1]]$time ## time for lung
  est_lung <- cif[2][[1]]$est ## est at time points


  time_other <- cif[3][[1]]$time
  est_other <- cif[3][[1]]$est


  time_no <- cif[1][[1]]$time
  est_no <- cif[1][[1]]$est


  ##est for lung
  match1 <- match(tl, time_lung)
  for(i in 1:length(tl)){
```

```
    if (!is.na(match1[i])) next
    else {
      match1[i] <- match1[i-1]
    }
  }
  est1 <- est_lung[match1]/max(est_lung)


  ##est for no
  match3 <- match(t_no, time_no)
  for(i in 1:length(t_no)){
    if(!is.na(match3[i])) next
    else{
      match3[i] <- match3[i-1]
    }
  }
  est3 <- est_no[match3]/max(est_no)


  ##est for other
  match2 <- match(to, time_other)
  for(i in 1:length(to)){
    if(!is.na(match2[i])) next
```

```
    else{
      match2[i] <- match2[i-1]
    }
  }
  est2 <- est_other[match2]/max(est_other)



  est <- c(est1, est2, est3)
  return(est)
}


cif_data <- as.matrix(data.frame(T, censor))


boot1 <- matrix(NA, nrow=52+254+54, ncol=1000)
for (i in 1:1000){
  id <- sample(1:209, replace=TRUE)
  sample <- cif_data[id, ]
  est <- cif_est(sample, tl=tl, to=to, t_no=t_no)
  boot1[,i] <- est
}
```

```
##95% CI for lung
boot_lung <- boot1[1:52, ]
estl_c <- estl/max(estl)
lower_lung <- apply(boot_lung, 1, function(x) quantile(x, .025))
upper_lung <- apply(boot_lung, 1, function(x) quantile(x, .975))

##95% CI for other
boot_other <- boot1[53:306, ]
esto_c <- esto/max(esto)
lower_other <- apply(boot_other, 1, function(x) quantile(x, .025))
upper_other <- apply(boot_other, 1, function(x) quantile(x, .975))

##95% CI for no
boot_no<- boot1[307:360, ]
est_no_c <- est_no/max(est_no)
lower_no <- apply(boot_no, 1, function(x) quantile(x, .025))
upper_no <- apply(boot_no, 1, function(x) quantile(x, .975))

##plot
plot(tl, estl_c, type="s", lty=1, col="blue", xlab="months",
     ylab="probability of death", ylim=c(0, 1), lwd=2)
```

```
lines(tl, lower_lung, lty=2, col="blue", type="s")

lines(tl, upper_lung, lty=2, col="blue", type="s")


lines(to, esto_c, col="red", type="s", lwd=2)

lines(to, lower_other, lty=2, col="red", type="s")

lines(to, upper_other, lty=2, col="red", type="s")


lines(t_no, est_no_c, col="green", type="s", lwd=2)

lines(t_no, lower_no, lty=2, col="green", type="s")

lines(t_no, upper_no, lty=2, col="green", type="s")


legend("bottomright", legend=c("Recurrence in lung","Confidence interval for lung"
                               "Recurrence in other sites",
                               "Confidence interval for other sites",
                               "No recurrence",
                               "Confidence interval for no recurrence"),
       col=c("blue", "blue", "red", "red", "green", "green"),
       lty=c(1,2, 1, 2, 1, 2), bty="n", cex=.75)



##CIF for X
```

```
censor2 = recur + 1

for(i in 1:209){

  if(X_censor[i] == 0)

    censor2[i] = 0

}

CIF2 <- cuminc(X, censor2)

print(CIF2)

plot(CIF2, lty=1, col=1:4, xlab="months")


cif_data <- as.matrix(data.frame(X, censor2))


##full time point

tl <- CIF2[2][[1]]$time

to <- CIF2[3][[1]]$time

t_no <- CIF2[1][[1]]$time


estl <- CIF2[2][[1]]$est

esto <- CIF2[3][[1]]$est

est_no <- CIF2[1][[1]]$est


varl <- CIF2[2][[1]]$var
```

```
varo <- CIF2[3][[1]]$var

var_no <- CIF2[1][[1]]$var


upper_lung <- estl + 1.96 * sqrt(varl)

lower_lung <- estl - 1.96 * sqrt(varl)


upper_other <- esto + 1.96 * sqrt(varo)

lower_other <- esto - 1.96 * sqrt(varo)


upper_no <- est_no + 1.96 * sqrt(var_no)

lower_no <- est_no - 1.96 * sqrt(var_no)


plot(to, esto, col="red", type="s", lty=1, xlab="months",

     ylab="probability of event", ylim=c(0,0.8), lwd=2)

lines(to, lower_other, lty=2, col="red", type="s")

lines(to, upper_other, lty=2, col="red", type="s")


lines(tl, estl, lty=1, col="blue", type="s", lwd=2)

lines(tl, lower_lung, lty=2, col="blue", type="s")

lines(tl, upper_lung, lty=2, col="blue", type="s")
```

```
lines(t_no, est_no, col="green", type="s")

lines(t_no, lower_no, lty=2, col="green", type="s", lwd=2)

lines(t_no, upper_no, lty=2, col="green", type="s")


legend(60, 0.55, legend=c("Recurrence in lung","Confidence interval for lung",
                          "Recurrence in other sites",
                          "Confidence interval for other sites",
                          "No recurrence",
                          "Confidence interval for no recurrence"),
       col=c("blue", "blue", "red", "red", "green", "green"),
       lty=c(1,2, 1, 2, 1, 2), bty="n", cex=.75)



##doing bootstrap for recurrence on lung and recurrence in others
##function for estiamte in bootstrap
cif_est <- function(data, tl, to, t_no){
  time <- data[,1]
  censor <- data[,2]
  cif <- cuminc(time, censor)
  time_lung <- cif[2][[1]]$time ## time for lung
  est_lung <- cif[2][[1]]$est ## est at time points
```

```
time_other <- cif[3][[1]]$time
est_other <- cif[3][[1]]$est

time_no <- cif[1][[1]]$time
est_no <- cif[1][[1]]$est

##est for lung
match1 <- match(tl, time_lung)
for(i in 1:length(tl)){
  if (!is.na(match1[i])) next
  else {
    match1[i] <- match1[i-1]
  }
}
est1 <- est_lung[match1]/max(est_lung)

##est for no
match3 <- match(t_no, time_no)
for(i in 1:length(t_no)){
  if(!is.na(match3[i])) next
```

```
    else{
      match3[i] <- match3[i-1]
    }
  }
  est3 <- est_no[match3]/max(est_no)


  ##est for other
  match2 <- match(to, time_other)
  for(i in 1:length(to)){
    if(!is.na(match2[i])) next
    else{
      match2[i] <- match2[i-1]
    }
  }
  est2 <- est_other[match2]/max(est_other)



  est <- c(est1, est2, est3)
  return(est)
}
```

```
boot1 <- matrix(NA, nrow=58+270+54, ncol=1000)

for (i in 1:1000){

  id <- sample(1:209, replace=TRUE)

  sample <- cif_data[id, ]

  est <- cif_est(sample, tl=tl, to=to, t_no=t_no)

  boot1[,i] <- est

}
```

```
##95% CI for lung

boot_lung <- boot1[1:58, ]

estl_c <- estl/max(estl)

lower_lung <- apply(boot_lung, 1, function(x) quantile(x, .025))

upper_lung <- apply(boot_lung, 1, function(x) quantile(x, .975))
```

```
##95% CI for other

boot_other <- boot1[59:328, ]

esto_c <- esto/max(esto)

lower_other <- apply(boot_other, 1, function(x) quantile(x, .025))

upper_other <- apply(boot_other, 1, function(x) quantile(x, .975))


##95% CI for no

boot_no<- boot1[329:382, ]

est_no_c <- est_no/max(est_no)

lower_no <- apply(boot_no, 1, function(x) quantile(x, .025))

upper_no <- apply(boot_no, 1, function(x) quantile(x, .975))


##plot

plot(tl, estl_c, type="s", lty=1, col="blue", xlab="months",

     ylab="probability of event", lwd=2)

lines(tl, lower_lung, lty=2, col="blue", type="s")

lines(tl, upper_lung, lty=2, col="blue", type="s")


lines(to, esto_c, col="red", type="s", lwd=2)

lines(to, lower_other, lty=2, col="red", type="s")

lines(to, upper_other, lty=2, col="red", type="s")
```

```
lines(t_no, est_no_c, col="green", type="s", lwd=2)

lines(t_no, lower_no, lty=2, col="green", type="s")

lines(t_no, upper_no, lty=2, col="green", type="s")


legend("bottomright", legend=c("Recurrence in lung","Confidence interval for lung"
                               "Recurrence in other sites",
                               "Confidence interval for other sites",
                               "No recurrence",
                               "Confidence interval for no recurrence"),
       col=c("blue", "blue", "red", "red", "green", "green"),
       lty=c(1,2, 1, 2, 1, 2), bty="n", cex=.75)


##############################################################################
##Cox's model with time-depend risk indicators


library("survival")
##time to recurrence
X <- data$surg.recur
X_censor <- data$censor.DFS
```

```
##time to death

T <- data$time.OS

T_censor <- data$censor.OS


recur <- c()

recur[data$Recur == "LungOnly"] = 1

recur[data$Recur == "liver" | data$Recur == "Local" |

        data$Recur == "Peritoneal"] = 2

recur[data$Recur == "LungOther"] =3

recur[data$Recur == "NoRecur"] = 0

recur = as.numeric(recur)


lung_recur <- X[recur==1]

lung_death <- T[recur==1]

other_recur <- X[recur==2]

other_death <- T[recur==2]

no_recur <- X[recur==0]

both_recur <- X[recur==3]

both_death <- T[recur==3]
```

```
##create time-dependent risk indicator

##for lung only recurrence

lung_matrix <- matrix(NA, sum(recur==1), 2)

for (i in 1:sum(recur==1)){

  lung_matrix[i, 1] <- 0

  lung_matrix[i, 2] <- 1

}

##for other recurrence

other_matrix <- matrix(NA, sum(recur==2), 2)

for(i in 1:sum(recur==2)){

  other_matrix[i, 1] <- 0

  other_matrix[i, 2] <- 1

}


##for recurrence in both lung and others

lung_matrix2 <- matrix(NA, sum(recur==3), 2)

for (i in 1:sum(recur==3)){

  lung_matrix2[i, 1] <- 0

  lung_matrix2[i, 2] <- 1

}
```

```
other_matrix2 <- matrix(NA, sum(recur==3), 2)

for(i in 1:sum(recur==3)){

  other_matrix2[i, 1] <- 0

  other_matrix2[i, 2] <- 1

}

##for no recurrence


dataframe <- data.frame(age, gender, stage, marg.pos, lymnode,

                        grade, vas.inv, per.inv, chemo, rad, T_censor)

data1 <- dataframe[recur==1, ]

data2 <- dataframe[recur==2, ]

data3 <- dataframe[recur==0, ]

data4 <- dataframe[recur==3, ]


##create dataframe for recurrence-in-lung patients


sum1 <- sum(!is.na(lung_matrix)) ## count of rows


lung_group <- matrix(0, sum1, 15)

colnames(lung_group) <- c("start", "stop", "lung", "other", "censor",

                          "age", "gender", "stage", "marg.pos",
```

```
                          "lymnode", "grade", "vas.inv", "per.inv",
                          "chemo", "rad")


##time to recurrence and death
lung_recur2 <- cbind(0, lung_recur, lung_death)


##create table with both time-dependent covariates and other covariates
row1<-0 #set record counter to 0
for (i in 1:nrow(data1)) { # loop over individuals
  for (j in 1:2) {   # loop over time points
    if (is.na(lung_matrix[i,j])) next #
    else {
      row1 <- row1 + 1 # increment row counter
      start <- lung_recur2[i,j] # start time
      stop <- lung_recur2[i,j+1] # stop time
      lung <- lung_matrix[i,j] ##time-dependent risk indicator
      other <- 0
      censor <-  if (stop == lung_death[i] && data1[i,11] == 0) 0 else 1
      ## censoring indicator
      #construct result
      lung_group[row1,] <- c(start, stop, lung, other, censor,
```

```
                              unlist(data1[i, c(1:10)]))
    }
  }
}


sum2 <- sum(!is.na(other_matrix)) ## count of rows


other_group <- matrix(0, sum2, 15)
colnames(other_group) <- c("start", "stop", "lung", "other", "censor",
                           "age", "gender", "stage", "marg.pos",
                           "lymnode", "grade", "vas.inv", "per.inv",
                           "chemo", "rad")


other_recur2 <- cbind(0, other_recur, other_death)
row2<-0 #set record counter to 0
for (i in 1:nrow(data2)) { # loop over individuals
  for (j in 1:2) {   # loop over time points
    row2 <- row2 + 1 # increment row counter
    start <- other_recur2[i,j] # start time
    stop <- other_recur2[i,j+1] # stop time
    lung <- 0
```

```
    other <- other_matrix[i,j]
    censor <-  if (stop == other_death[i] && data2[i,11] == 0) 0 else 1
    ## censoring indicator
    #construct result
    other_group[row2,] <- c(start, stop, lung, other, censor,
                            unlist(data2[i, c(1:10)]))
  }
}

both_group <- matrix(0, 2*nrow(data4), 15)
colnames(both_group) <- c("start", "stop", "lung", "other", "censor",
                          "age", "gender", "stage", "marg.pos",
                          "lymnode", "grade", "vas.inv", "per.inv",
                          "chemo", "rad")

both_recur2 <- cbind(0, both_recur, both_death)
row4 <- 0
for (i in 1:nrow(data4)) { # loop over individuals
  for (j in 1:2) {   # loop over time points
    row4 <- row4 + 1 # increment row counter
    start <- both_recur2[i,j] # start time
```

```
    stop <- both_recur2[i,j+1] # stop time
    lung <- lung_matrix2[i, j]
    other <- other_matrix2[i, j]
    censor <-  if (stop == both_death[i] && data4[i,11] == 0) 0 else 1
    ## censoring indicator
    #construct result
    both_group[row4,] <- c(start, stop, lung, other, censor,
                           unlist(data4[i, c(1:10)]))
  }
}


no_group <- matrix(0, nrow(data3), 15)
colnames(no_group) <- c("start", "stop", "lung", "other", "censor",
                        "age", "gender", "stage", "marg.pos", "lymnode",
                        "grade", "vas.inv", "per.inv", "chemo", "rad")


row3<-0 #set record counter to 0
no_recur2 <- c(0, no_recur)
for (i in 1:nrow(data3)) { # loop over individuals
  row3 <- row3 + 1 # increment row counter
  start <- no_recur2[1] # start time
```

```
  stop <- no_recur2[2] # stop time
  lung <- 0
  other <- 0
  censor <-  if (stop == no_recur[i] && data3[i,11] == 0) 0 else 1
  ## censoring indicator
  #cinstruct result
  no_group[row3,] <- c(start, stop, lung, other, censor,
                       unlist(data3[i, c(1:10)]))
}
cancer <- as.data.frame(rbind(lung_group, other_group,
                              both_group, no_group))
cox <- coxph(Surv(start, stop, censor) ~ lung + other + age +
               I(age*age) + gender + stage + marg.pos +
               lymnode + grade + vas.inv + per.inv + chemo + rad +
               I(gender*other)+I(gender*lung), data = cancer)
summary(cox)


##with indicator whether the recurrence time is within certain time
##set the cut point for recurrence
cut <- 6
```

```
##create time-dependent risk indicator

##for lung only recurrence

lung_matrix <- matrix(NA, sum(recur==1), 2)

for (i in 1:sum(recur==1)){

  lung_matrix[i, 1] <- 0

  lung_matrix[i, 2] <- 1

}

##for other recurrence

other_matrix <- matrix(NA, sum(recur==2), 2)

for(i in 1:sum(recur==2)){

  other_matrix[i, 1] <- 0

  other_matrix[i, 2] <- 1

}


##for recurrence in both lung and others

lung_matrix2 <- matrix(NA, sum(recur==3), 2)

for (i in 1:sum(recur==3)){

  lung_matrix2[i, 1] <- 0

  lung_matrix2[i, 2] <- 1

}
```

```
other_matrix2 <- matrix(NA, sum(recur==3), 2)

for(i in 1:sum(recur==3)){

  other_matrix2[i, 1] <- 0

  other_matrix2[i, 2] <- 1

}

##for no recurrence


dataframe <- data.frame(age, gender, stage, marg.pos,

                        lymnode, grade, vas.inv, per.inv, chemo, rad, T_censor)

data1 <- dataframe[recur==1, ]

data2 <- dataframe[recur==2, ]

data3 <- dataframe[recur==0, ]

data4 <- dataframe[recur==3, ]


##create dataframe for recurrence-in-lung patients


sum1 <- sum(!is.na(lung_matrix)) ## count of rows


lung_group <- matrix(0, sum1, 17)

colnames(lung_group) <- c("start", "stop", "lung =< 6","lung > 6",

                          "other<=6","other>6", "censor", "age", "gender",
```

```
                          "stage", "marg.pos", "lymnode", "grade",
                          "vas.inv", "per.inv", "chemo", "rad")


##time to recurrence and death
lung_recur2 <- cbind(0, lung_recur, lung_death)


lung_id1 <- (lung_recur > cut)
lung_id2 <- (lung_recur <= cut)
##create table with both time-dependent covariates and other covariates
row1<-0 #set record counter to 0
for (i in 1:nrow(data1)) { # loop over individuals
  for (j in 1:2) {   # loop over time points
    if (is.na(lung_matrix[i,j])) next #
    else {
      row1 <- row1 + 1 # increment row counter
      start <- lung_recur2[i,j] # start time
      stop <- lung_recur2[i,j+1] # stop time
      lung1 <- lung_matrix[i,j] * lung_id1[i]
      ##time-dependent risk indicator
      lung2 <- lung_matrix[i,j] * lung_id2[i]
      other1 <- 0
```

```
      other2 <- 0
      censor <-  if (stop == lung_death[i] && data1[i,11] == 0) 0 else 1
      ## censoring indicator
      #construct result
      lung_group[row1,] <- c(start, stop, lung1, lung2, other1,
                             other2, censor, unlist(data1[i, c(1:10)]))
    }
  }
}

sum2 <- sum(!is.na(other_matrix)) ## count of rows

other_group <- matrix(0, sum2, 17)
colnames(other_group) <- c("start", "stop", "lung =< 6","lung > 6",
                           "other<=6","other>6", "censor", "age", "gender",
                           "stage", "marg.pos", "lymnode", "grade",
                           "vas.inv", "per.inv", "chemo", "rad")

other_recur2 <- cbind(0, other_recur, other_death)
other_id1 <- (other_recur <= cut)
other_id2 <- (other_recur > cut)
```

```
row2<-0 #set record counter to 0
for (i in 1:nrow(data2)) { # loop over individuals
  for (j in 1:2) {   # loop over time points
    row2 <- row2 + 1 # increment row counter
    start <- other_recur2[i,j] # start time
    stop <- other_recur2[i,j+1] # stop time
    lung1 <- 0
    lung2 <- 0
    other1 <- other_matrix[i,j] * other_id1[i]
    other2 <- other_matrix[i,j] * other_id2[i]
    censor <-  if (stop == other_death[i] && data2[i,11] == 0) 0 else 1
    ## censoring indicator
    #construct result
    other_group[row2,] <- c(start, stop, lung1, lung2, other1, other2,
                            censor, unlist(data2[i, c(1:10)]))
  }
}

both_group <- matrix(0, 2*nrow(data4), 17)
colnames(both_group) <- c("start", "stop", "lung =< 6","lung > 6",
                          "other<=6","other>6", "censor", "age", "gender",
```

```
                          "stage", "marg.pos", "lymnode", "grade",
                          "vas.inv", "per.inv", "chemo", "rad")


both_recur2 <- cbind(0, both_recur, both_death)
both_id1 <- (both_recur <= cut)
both_id2 <- (both_recur > cut)
row4 <- 0
for (i in 1:nrow(data4)) { # loop over individuals
  for (j in 1:2) {   # loop over time points
    row4 <- row4 + 1 # increment row counter
    start <- both_recur2[i,j] # start time
    stop <- both_recur2[i,j+1] # stop time
    lung1 <- 0
    lung2 <- 0
    other1 <- other_matrix2[i, j] * both_id1[i]
    other2 <- other_matrix2[i, j] * both_id2[i]
    censor <-  if (stop == both_death[i] && data4[i,11] == 0) 0 else 1
    ## censoring indicator
    #construct result
    both_group[row4,] <- c(start, stop, lung1, lung2, other1, other2,
                           censor, unlist(data4[i, c(1:10)]))
```

```
  }
}


no_group <- matrix(0, nrow(data3), 17)
colnames(no_group) <- c("start", "stop", "lung =< 6","lung > 6",
                        "other<=6","other>6", "censor", "age", "gender",
                        "stage", "marg.pos", "lymnode", "grade",
                        "vas.inv", "per.inv", "chemo", "rad")


row3<-0 #set record counter to 0
no_recur2 <- c(0, no_recur)
for (i in 1:nrow(data3)) { # loop over individuals
  row3 <- row3 + 1 # increment row counter
  start <- no_recur2[1] # start time
  stop <- no_recur2[2] # stop time
  lung1 <- 0
  lung2 <- 0
  other1 <- 0
  other2 <- 0
  censor <-  if (stop == no_recur[i] && data3[i,11] == 0) 0 else 1
  ## censoring indicator
```

```
  #construct result
  no_group[row3,] <- c(start, stop, lung1, lung2, other1, other2,
                       censor, unlist(data3[i, c(1:10)]))
}
cancer <- as.data.frame(rbind(lung_group, other_group,
                              both_group, no_group))
colnames(cancer) <- c("start", "stop", "lung1","lung2", "other1","other2",
                      "censor", "age", "gender",
                      "stage", "marg.pos", "lymnode", "grade",
                      "vas.inv", "per.inv", "chemo", "rad")




cox <- coxph(Surv(start, stop, censor) ~ lung1 + lung2 + other1
             + other2 + age + I(age*age) + gender + stage +
               marg.pos + lymnode + grade + vas.inv + per.inv
             + chemo + rad, data = cancer)
summary(cox)


##include only the recurrence cut for lung recurrence
##set the cut point for recurrence
```

```
cut <- 6


##create time-dependent risk indicator
##for lung only recurrence
lung_matrix <- matrix(NA, sum(recur==1), 2)
for (i in 1:sum(recur==1)){
  lung_matrix[i, 1] <- 0
  lung_matrix[i, 2] <- 1
}
##for other recurrence
other_matrix <- matrix(NA, sum(recur==2), 2)
for(i in 1:sum(recur==2)){
  other_matrix[i, 1] <- 0
  other_matrix[i, 2] <- 1
}


##for recurrence in both lung and others
lung_matrix2 <- matrix(NA, sum(recur==3), 2)
for (i in 1:sum(recur==3)){
  lung_matrix2[i, 1] <- 0
  lung_matrix2[i, 2] <- 1
```

```
}


other_matrix2 <- matrix(NA, sum(recur==3), 2)
for(i in 1:sum(recur==3)){
  other_matrix2[i, 1] <- 0
  other_matrix2[i, 2] <- 1
}
##for no recurrence


dataframe <- data.frame(age, gender, stage, marg.pos, lymnode,
                        grade, vas.inv, per.inv, chemo, rad, T_censor)
data1 <- dataframe[recur==1, ]
data2 <- dataframe[recur==2, ]
data3 <- dataframe[recur==0, ]
data4 <- dataframe[recur==3, ]


##create dataframe for recurrence-in-lung patients


sum1 <- sum(!is.na(lung_matrix)) ## count of rows


lung_group <- matrix(0, sum1, 16)
```

```
colnames(lung_group) <- c("start", "stop", "lung =< 6","lung > 6",
                          "other", "censor", "age", "gender",
                          "stage", "marg.pos", "lymnode", "grade",
                          "vas.inv", "per.inv", "chemo", "rad")


##time to recurrence and death
lung_recur2 <- cbind(0, lung_recur, lung_death)


lung_id1 <- (lung_recur > cut)
lung_id2 <- (lung_recur <= cut)
##create table with both time-dependent covariates and other covariates
row1<-0 #set record counter to 0
for (i in 1:nrow(data1)) { # loop over individuals
  for (j in 1:2) {   # loop over time points
    if (is.na(lung_matrix[i,j])) next #
    else {
      row1 <- row1 + 1 # increment row counter
      start <- lung_recur2[i,j] # start time
      stop <- lung_recur2[i,j+1] # stop time
      lung1 <- lung_matrix[i,j] * lung_id1[i]
      ##time-dependent risk indicator
```

```
      lung2 <- lung_matrix[i,j] * lung_id2[i]
      other <- 0
      censor <-  if (stop == lung_death[i] && data1[i,11] == 0) 0 else 1
      ## censoring indicator
      #construct result
      lung_group[row1,] <- c(start, stop, lung1, lung2, other,
                             censor, unlist(data1[i, c(1:10)]))
    }
  }
}

sum2 <- sum(!is.na(other_matrix)) ## count of rows

other_group <- matrix(0, sum2, 16)
colnames(other_group) <- c("start", "stop", "lung =< 6","lung > 6",
                           "other", "censor", "age", "gender",
                           "stage", "marg.pos", "lymnode", "grade",
                           "vas.inv", "per.inv", "chemo", "rad")

other_recur2 <- cbind(0, other_recur, other_death)
other_id1 <- (other_recur <= cut)
```

```
other_id2 <- (other_recur > cut)

row2<-0 #set record counter to 0

for (i in 1:nrow(data2)) { # loop over individuals

  for (j in 1:2) {   # loop over time points

    row2 <- row2 + 1 # increment row counter

    start <- other_recur2[i,j] # start time

    stop <- other_recur2[i,j+1] # stop time

    lung1 <- 0

    lung2 <- 0

    other <- other_matrix[i,j]

    censor <-  if (stop == other_death[i] && data2[i,11] == 0) 0 else 1

    ## censoring indicator

    #construct result

    other_group[row2,] <- c(start, stop, lung1, lung2, other,

                            censor, unlist(data2[i, c(1:10)]))

  }

}


both_group <- matrix(0, 2*nrow(data4), 16)

colnames(both_group) <- c("start", "stop", "lung =< 6","lung > 6",

                          "other", "censor", "age", "gender",
```

```
                          "stage", "marg.pos", "lymnode", "grade",
                          "vas.inv", "per.inv", "chemo", "rad")


both_recur2 <- cbind(0, both_recur, both_death)


row4 <- 0
for (i in 1:nrow(data4)) { # loop over individuals
  for (j in 1:2) {   # loop over time points
    row4 <- row4 + 1 # increment row counter
    start <- both_recur2[i,j] # start time
    stop <- both_recur2[i,j+1] # stop time
    lung1 <- 0
    lung2 <- 0
    other <- other_matrix2[i, j]
    censor <-  if (stop == both_death[i] && data4[i,11] == 0) 0 else 1
    ## censoring indicator
    #construct result
    both_group[row4,] <- c(start, stop, lung1, lung2, other,
                           censor, unlist(data4[i, c(1:10)]))
  }
}
```

```
no_group <- matrix(0, nrow(data3), 16)
colnames(no_group) <- c("start", "stop", "lung =< 6","lung > 6",
                        "other", "censor", "age", "gender",
                        "stage", "marg.pos", "lymnode", "grade",
                        "vas.inv", "per.inv", "chemo", "rad")

row3<-0 #set record counter to 0
no_recur2 <- c(0, no_recur)
for (i in 1:nrow(data3)) { # loop over individuals
  row3 <- row3 + 1 # increment row counter
  start <- no_recur2[1] # start time
  stop <- no_recur2[2] # stop time
  lung1 <- 0
  lung2 <- 0
  other <- 0
  censor <-  if (stop == no_recur[i] && data3[i,11] == 0) 0 else 1
  ## censoring indicator
  #construct result
  no_group[row3,] <- c(start, stop, lung1, lung2, other,
                       censor, unlist(data3[i, c(1:10)]))
```

```
}

cancer <- as.data.frame(rbind(lung_group, other_group,

                              both_group, no_group))

colnames(cancer) <- c("start", "stop", "lung1","lung2",

                      "other", "censor", "age", "gender",

                      "stage", "marg.pos", "lymnode", "grade",

                      "vas.inv", "per.inv", "chemo", "rad")




cox <- coxph(Surv(start, stop, censor) ~ lung1 + lung2 + other +

               age + I(age*age) + gender + stage +

               marg.pos + lymnode + grade + vas.inv + per.inv +

               chemo + rad, data = cancer)

summary(cox)
```

Bibliography

Bernard Altshuler. Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences*, 6:1–11, 1970.

N. E. Breslow. Covariance Analysis of Censored Survival Data. *Biometrics*, 30(3): 89–99, 1974.

S. C. Cheng, L. J. Wei, and Z. Ying. Analysis of transformation models with censored data. *Biometrika*, 82(4):835–845, 1995.

Chin Long Chiang. Introduction to stochastic processes in biostatistics, 1968.

J. Cornfield. The estimation of the probability of developing a disease in the presence of competing risks. *American Journal of Public Health*, 47:601–607, 1957.

D. R. Cox. The analysis of exponentially distributed lifetimes with two types of failure. *Journal of the Royal Statistical Society. Series B*, 21:411–421, 1959.

D. R. Cox. Regression models and life tables. *Journal of the Royal Statistical Society. Series B*, 34:187–220, 1972.

BIBLIOGRAPHY

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

J. Cuzick. Rank regression. *The Annals of Statistics*, 16(4):1369–1389, 1988.

Dorota M. Dabrowska and Kjell A. Doksum. Estimation and testing in a two-sample generalized odds-rate model. *Journal of the American Statistical Association*, 83 (403):744–749, 1988.

J. P. Fine, Z Ying, and L. G. Wei. On the linear transformation model for censored data. *Biometrika*, 85(4):980–986, 1998.

Jason P. Fine and Robert J. Gray. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94 (446):496–509, 1999.

P. Ghaneh, E. Costello, and J. P. Neoptolemos. Biology and management of pancreatic cancer. *Gut 56*, 8:1134–1152, 2007.

R. J. Gray. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 1988.

J. D. Holt. Competing risk analyses with special reference to matched pair experiments. *Biometrika*, 65(1):159–165, 1978.

S. Iodice, S. Gandini, P. Maisonneuve, and A. B. Lowenfels. Tobacco and the risk of pancreatic cancer: a review and meta-analysis. *Langenbeck's Archives of Surgery*, 393:534–545, 2008.

BIBLIOGRAPHY

E. L. Kaplan and Meier P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(2):457–481, 1958.

W. M. Makeham. On an application of hte theory of the composition of decremental forces. *Journal of the Institute of Actuaries*, 18:317–322, 1874.

H. A. Moeschberger, M. L.and David. Life tests under competing causes of failure and the theory of competing risks. *Biometrics*, 27:909–923, 1971.

S. A. Murphy, A. J. Rossini, and A W van der Vaart. Maximum Likelihood Estimation in the Proportional Odds Model. *Journal of the American Statistical Association*, 1997.

R. L. Prentice and N. E. Breslow. Retrospective studies and failure time models. *Biometrika*, 65(1):153–158, 1978.

R. L. Prentice, J. D. Kalbfleisch, Peterson A. V., N. Flournoy, V. T. Farewell, and N. E. Breslow. The Analysis of Failure Times in the Presence of Competing Risks. *Biometrics*, 34(4):541, 1978.

American Cancer Society. Obesity Linked to Pancreatic Cancer. *Cancer Epidemiology, Biomarkers and Prevention*, 14(2):459–466, 2008.

Vita

YAO LU received the Sc. B. degree in Mathematics from Fudan University in 2012, and enrolled in the Biostatistics ScM program at Johns Hopkins Bloomberg School of Public Health in 2012. She won the First Prize Scholarship in 2009, the Major Scholarship in a series from 2009 to 2011, and Kocherlakota Award in 2013. Her research focuses on survival analysis and competing risks model, and and her thesis have used the methods to study pancreatic cancer recurrence patterns. Besides that, she has explored other fields in Biostatistics, such as genomics, Bayesian method and imaging, and did several projects.

She has rich experience in research as she started working as research assistant from September 2013. Also since September 2007, YAO has been working as teaching assistant for Biostatistics department and helped a lot of students. She hopes to work as a biostatistician or a data scientist in the future.