COMPARISON OF COX REGRESSION AND DISCRETE TIME SURVIVAL MODELS

by

HONG YE

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

In partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2016

MAJOR: EDUCATION EVALUATION
AND RESEARCH

Approved By:

____

Advisor Date

____

____

____

DEDICATION

To my beloved family.

ACKNOWLEDGEMENTS

First and foremost, I need to acknowledge my parents for supporting me throughout my educational career. They are always there when I need them.

Academically, I appreciate very much the invaluable support from my dissertation advisor, Prof. Shlomo Sawilowsky. With his encouragement, I started to work on my current project. During my Ph.D. program, he supplied visionary guides to my research, and also provided much support to me. His guidance is always forthcoming. He teaches me to think big and prevents me from lost in details. From him, I learn a lot about how to thinking problems and how to express my ideas to people.

I am also indebted to the other committee members: Prof. Craig Stevens, Prof. Monte Piliawsky, and Prof. Frank Castronova. They served as my committee members and gave valuable suggestions on my prospectus and my dissertation.

I also want to thank Dr. Gail Fahoome who served as a mentor prior to her unfortunate passing. She provided me valuable suggestions and encouragement during the first years I pursued my Ph.D. degree in Wayne State University.

Last but not least, I'd like to thank my husband who provided a lot of support to me. I also want to thank my lovely daughter Annie and my son Kevin who are my spiritual power. I could have never done this without the support of my family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

Survival analysis generally includes a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. The event can be death, relapse of a disease, remission to hospital, or any occurrence of certain events. The question often arises about the occurrence and timing of critical events, and therefore modeling the occurrence of those events becomes important. Examples in health care include how long the patients remain well after treatment, or whether patients with a certain history or characteristics have greater chances for an illness relapse.

Kaplan and Meier (1958) demonstrated how to deal with incomplete observations. Their method is a nonparametric estimator of the survival function, and is used to estimate and graph survival probabilities as a function of time. The Kaplan-Meier curves have become popular in life and medical sciences (Allison, 1982; Barber, Murphy, Axinn, & Maples, 2000; Kaplan & Meier, 1958; Miller, 1981a, 1983).

Cox (1972) proposed the discrete-time survival method for discrete-time data and the proportional hazard modeling for continuous-time data. The Cox proportional hazard regression model is a regression model for the analysis of survival data, and it provides useful information regarding the relationship of the hazard function to predictors. Similarly, Allison (1982) and Judith D. Singer and Willett (1993) proposed discrete-time survival methods.

The Kaplan-Meier curves, life-table ("Life Table,"), and Cox Regression are commonly used methods in medical research (Rich et al., 2010). Some applications are only descriptive, but other applications involve estimating the survival or hazard after adjustment for other predictors. For example, the Cox proportional model is a well-established statistical technique for analyzing survival data. The Cox proportional model is considered as a semi-parametric procedure because the baseline hazard function doesn’t have to be specified. Because the hazard function is not restricted to a specific form, the semi-parametric model has considerable flexibility and is widely used (Han et al., 2003). When the assumption of a particular probability distribution for the data is valid, inferences based on such an assumption are more precise with smaller standard errors and narrower confidence limits.

Although the Cox proportional model is a commonly employed technique in survival analysis, it has some restrictions such as its proportional hazard assumption, meaning the hazard ratio between two sets of covariates is constant over time. The baseline hazard is defined as the hazard function for that individual with zero on all covariates. Because the Cox proportional model is a semi-parametric model, its baseline hazard has no particular form. Thus, the baseline hazard can take parametric form. Under certain circumstances in which parametric assumptions of baseline hazards are met, Cox proportional model will be more powerful (M. Pourhoseingholi et al., 2011).

The discrete-time survival methods have been in use for decades, but they are less visible than continuous survival methods like Cox regression model, especially in the medical and behavioral science area (Altman, De Stavola, Love, & Stepniewska, 1995; Enderlein, 1987). The discrete-time survival method was proposed by Cox in 1972, and it is a type of logistic regression. The discrete-time methods are used more appropriate in the situation with large number of ties. A tie is defined as more than two individuals experience an event at the same time (Allison, 1982). Examples include a person-period dataset by Singer and Willett (1995, 2003), Allison (1982), and Willett and Singer (1993). In 1990, D’Agostino et al. (1990) covered the relationship between a pooled logistic regression and time dependent Cox regression by a variety of samples sizes and proportions of events, and displayed the closeness of this relationship under certain conditions.

The advantages of the discrete-time survival method were summarized by Xie, McHugo, Drake, and Sengupta, (2003) and Sharaf and Tsokos (2014). In most clinical settings, the discrete-time survival methods are useful for longitudinal studies when the data are often collected at discrete-time periods. Discrete-time survival method examines the shape of hazards function, and it is simple to implement using the logistic regression model. In practice, when the time of experiencing event is hard to tell, then using the discrete-time method has more advantages than continuous-time survival method.

Discrete-time vs. Continuous-time Survival Data

Time scales for events can be classified into two categories: continuous or discrete. Survival analysis requires that each individual be observed over some defined interval of time. The time to event or survival time can be measured in days, weeks, years, etc. If the event occurred during that interval, their times are recorded. Most methods of survival analysis require that survival time be measured with respect to some origin time. It is substantively important to choose the origin time because the risk of the event varies as a function of time since the origin. In many cases, the choice of origin is obvious. For instance, if the event is divorce, the origin time is the date of the marriage; if the event is recurrence of cancer, the origin time is the date of last cancer treatment.

An event of interest may occur at any particular instant in time, and time is a continuum and measured as a non-negative real number. If it is known when the events occur for origin time, it is better to treat time as continuous (Allison, 1982; Xie, McHugo, Drake, & Sengupta, 2003). Applications using continuous-time assume that the timing of event is known and is measured in some discrete intervals which are small enough to be treated as a continuous-time scale. Measuring the time in a discrete fashion will place it into bins (e.g. number of months or years). The observations on the transition process are summarized discretely rather than continuously.

Both continuous-time and discrete-time models involve examining the coefficients for each explanatory variable. A positive regression coefficient for an explanatory variable means higher hazard and worse prognosis. Conversely, a negative regression coefficient implies a better prognosis with higher value of that variable. In comparison with continuous survival-time models, such as the Kaplan Meier and Cox proportional hazard methods, the discrete-time survival analysis is relatively unknown and underused in medical research.

Analytic models for survival analysis can be categorized into four general types: parametric models, nonparametric models, semi-parametric models and discrete-time models. Parametric models assume an underlying distribution for the probability function. And parametric statistical procedures are sensitive to the violation of underlying assumptions. Nonparametric procedures include no assumptions regarding the probability density function and use observed data to describe survivor functions and hazard functions. Nonparametric methods are robust with respect to Type I errors for departures from normality, meaning they don't have distribution assumptions. However, they are also sensitive to the violation of other types of assumptions, e.g., independence and homoscedasticity. Similarly, outliers do impact the power properties of nonparametric procedures.

Certain semi-parametric model, such as the Cox proportional model, does not have strong assumptions about the underlying probability function but does include an assumption of proportional hazards among model covariates. Altman, De Stavola, Love, and Stepniewska, (1995) reported that authors of only 5% of studies use Cox models checked the underlying assumption. Outliers also play an important impact on nonparametric procedures. For Cox regression model, a single outlier can lead to violate the assumption of proportionality of hazard. Models such as the logit and complementary log-log are popular choices for discrete-time survival analysis. Key features of this type of analysis needs a properly structured data set with multiple records per respondent. In a parametric model, the maximum likelihood procedure is used to estimate the unknown parameters. In the Cox proportional regression model, the partial maximum likelihood is used for computing average hazard ratios in the presence of non-proportionality of hazards. The maximum likelihood (ML) function is a mathematical expression which describes the joint probability of obtaining the data actually observed on the subjects in the study as a function of the unknown parameters in the model being considered. The likelihood function L is sometimes written notationally as $L(\beta)$where β denotes the collection of unknown parameters.

Event history data are common in many disciplines and its core is focused on time. Time can be regarded as continuous or discrete and this basic distinction affects the analytic approach selected. Singer and Willett (1993) demonstrated the use of discrete-time survival analysis using logistic regression in social sciences. The use of discrete-time survival method has been studied further by Prentice and Gloeckler (1978), as well as many others including Allison (1982), Altman et al. (1995), Barber et al. (2000), Xie et al. (2003), and McCallon (2009).

Purpose of Study

Despite the varied conditions under which discrete-time survival methods have been studied, its statistical properties remain largely unknown. Therefore, the purpose of this study is to research and explicate under what conditions the discrete-time survival method is comparable with Cox regression model respect to hazard estimation.

Definition of Terms

Akaike Information Criterion (AIC): A goodness of fit measure of the relative quality of statistical models for a given set of data.

Assumptions: A statistical test requirement necessary to maintain specified Type I error rates (e.g., p=.05).

Bayesian Information Criterion (BIC): A criterion for model selection among a finite set of models. Lower value indicates a better model. It is closely related to the AIC on the likelihood function

Coefficient: A multiplicative factor in terms of a polynomial, a series or any expression.

Censored: The survival time of an individual is said to be censored when the end-point of interest has not been observed for that individual.

Cox Regression: One type of regression. The dependent variable of Cox regression is the hazard function at a given time.

$$h(t) = h_0(t) \cdot \exp(\beta_i X_i)$$

If taking natural logarithm of both sides:

$$Inh(t) = In(h_0(t) \cdot \exp(\beta_i X_i))$$

Confidence interval: A range of values, calculated from the sample of observations that are believed, with a particular probability, to contain the true parameter value. A 95% confidence interval implies that if the estimation process were repeated again and again, then 95% of the calculated intervals would be expected to contain the true parameter value.

Cumulative distribution function (c.d.f.): is the common method to characterize the distribution of any random variable, which is denoted by:

$cdf: F(t) = P(T \leq t)$, where T is non-negative elapsed time until an event.

Hazard function H(t): The chronological pattern of hazard probabilities over the time. The hazard probability refers to an individual will experience an event within a small time interval given that the individual has survived up to the beginning of the interval.

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < T + \Delta t \mid T \ge t)}{\Delta t}$$

Where the numerator is the probability that an event occurs during a very small interval of time $[t, t + \Delta t)$, given that no event occurred before time t.

Hazard probability: The proportion of the risk set who experience the event in that time periods.

Independent censoring: Censoring is unrelated to event occurrence.

Left-censoring: The event has already occurred before enrollment. This is very rarely encountered.

Logarithms: Logarithms are mainly used in statistics to transform a set of observations to values with a more convenient distribution.

Logrank test: A method for comparing the survival times of two or more groups of subjects. It involves the calculation of observed and expected frequencies of events in separate time intervals.

Monte Carlo Simulation: The use of a computer program to simulate some aspect of reality to make determinations of the nature of reality or change in reality through the repeated sampling via Monte Carlo methods (Sawilowsky & Fahoome, 2002).

Maximum likelihood (ML): an estimate of unknown parameters which uses the joint probability of obtaining the data actually observed on the subjects in the study.

Median lifetime: The time at which half the sample or population had experienced the target event and half have not.

Normality: A state of data distribution which fits the normal or Gaussian curve. It is a parameter assumed for the t and F tests.

P-value: The probability value, or significance level, from a hypothesis test. *p* is the probability of the data arising by chance when the null hypothesis is true.

Regression: The statistical technique used to describe the relationship between the values of two or more variables. When more than one explanatory variable is need to be taken into account, the method is known as multiple regression.

Right-censoring: The event has not occurred by the end of the observation period. Right-censoring is the most common form of censoring.

Risk set: The group of people known to be eligible to experience the event in a particular time period.

SE (se): The standard error of a sample mean or some other estimated statistics. It is the measure of the uncertainty of such an estimate and it is used to derive a confidence interval for the population values.

Standard error: It is defined to be the square root of the estimated variance of the estimate, and is used in the construction of an interval estimate for a quantity of interest.

Survival function S(t): The probability that an individual survives from the time origin to sometime beyond t.

$S(t) = \mathrm{P}(T \geq t) = 1 - F(t)$, where $F(t)$ is probability density function. The distribution function of T is given by:

$$F(t) = \mathrm{P}(T < t) = \int_0^t f(u)\mathrm{d}u$$

The baseline model:

In discrete-time survival analysis, $\beta_0(t)$ is the baseline log hazard profile, and represents the values of the outcome without other predictor variables. The baseline equation can be expanded to account for specific measurements of discrete-time intervals to

$$\text{logit}_{\text{e}} h_j = [\alpha_1 t_1 + \alpha_2 t_2 + ... + \alpha_k t_k]$$

In Cox proportional model, the baseline hazard function is left unspecified but must be positive.

CHAPTER 2 LITERATURE REVIEW

Censorship

Analyzing survival data basically needs censor variable (outcome variable) and time variable (survival time). The survival time is defined as the time to events. Observations are called censored when the information about their survival time is incomplete (Tabachnick & Fidell, 1996). There are several different censorships: right-censoring, left-censoring, and interval-censoring. For example, the survival time of an individual is said to be right censored when the end-point of interest has not been observed for that individual at the end of study, or the individual has lost to follow-up or dropped out from the study. We call this phenomenon right-censoring because the true unobserved event is to the right of our censoring time. Right-censoring is the most common type of censoring assumption we will deal with in survival analysis, and it underestimates the true survival time because the survival time is unknown and the ultimate event time for censored cases is greater than the imputed value which is equal to the length of data collection for the right-censored cases (Clark, Bradburn, Love, & Altman, 2003a; John B. Willett & Singer, 1991). "Most methods of survival analysis do not distinguish among type of right-censoring, but cases that are lost from the study may pose problems because it is assumed that there are no systematic differences between them and the cases that remain". (Tabachnick & Fidell, 1996, p. 537)

Left-censoring refers to the actual time of event of interest occurs less than the observation time. For example, if a patient was examined 6 month after treatment to determine recurrence, then those who had a recurrence would have a survival time that was left censored because their survival time is less than 6 month (Clark et al., 2003a). A problem of interval censoring arises when time to event may be known only up to a time interval. This usually happens at a periodical monitoring. If we consider the previous example and patients are also examined at 6 months, then those who are disease free at 6 months but lost to follow-up between 6 and 12 months are considered interval censored. Some studies even included both right censoring and left censoring observations (Miller, 1981a). Most survival data are right-censored, and methods for interval and left censored data are also available.

Censorship is important and unavoidable in survival analysis since it represents a particular type of missing data. Sometimes, all the subjects in the study experienced the events of interests, and there is no censored case. In most situations, however, not all participants experience the events of interested during the study period. This may occur because participants are no longer able to be tracked. Since right censoring is the most common censorship, this dissertation only involves right censoring. For the right-censored cases, the time to failure is greater than the censoring time, and the censored cases because of loss to follow-up are treated to have same survival prospects as those who continue to be followed. Thus the censoring is uninformative. Informative censoring may occur when patients withdraw from a study because of special condition. Standard methods for survival analysis are valid for uninformative censoring but not for informative censoring in which uninformative censoring carries no prognostic information about subsequent survival experience (Clark et al., 2003a; Clark, Bradburn, Love, & Altman, 2003b).

Willett (1991) wrote:

> Censoring creates an analytic dilemma: What should be done with people who do not experience the target event during the period of data collection? Although the researcher knows something about them – if they ever experience the event, they do so after data collection ends – this knowledge is imprecise. (p. 408)

There are different strategies to deal with censoring and various methods can be used to treat different censored data, including complete data analysis, imputation techniques or analysis based on dichotomized data (Prinja, Gupta, & Verma, 2010; John B. Willett & Singer, 1991). In some investigations, the purpose was to focus exclusively on those subjects with known events times and set aside censored cases (Judith D. Singer & Willett, 1993; John B. Willett & Singer, 1991). It may lead to large bias if the number of censored cases is large (Allison, 1982). Some investigations impute the missing duration data, assigning the duration value to censored cases (John B. Willett & Singer, 1991). Other investigations dichotomize the event histories at a particular time and ask whether the event has occurred by that time. However, "the dichotomization of dependent variable is both arbitrary and wasteful of information" (Allison, 1982, p. 64). It is arbitrary because the cutting point of dichotomization can be set to any number, and usually the cutting line is set to what investigation care about. "It is wasteful of information because it ignores the variation on either side of the cutoff point." (Allison, 1982, p. 64)

No matter which way the studies choose for analyzing the censored cases, summarizing the event history data is the main goal of survival analysis. Survival data are generally described and modeled into two related functions. One way is to use survival function to list estimated survival probabilities chronologically. Survival probabilities represent the proportions of the initial sample that do not experience the event through each of several time intervals. Another different way is looking at the proportion of the risk set who experiences the event in that period rather than the survival proportion. The hazard function involves both non-censored and censored cases, and it is an indirect way to estimate the survival functions.

Survival and Hazard Functions

The Survival Function

Survival analysis aims to analyze longitudinal data on the occurrence of events. Events may include death, injury, onset of illness, etc. The goal of survival analysis is to estimate and compare survival experiences of different groups. Survival data are generally described and modeled in terms of two related probabilities, namely *survival* and *hazard*.

The survival probability is also called the survival function $S(t)$, which is the probability that an individual survives from the time origin to a specified future time t. Survival experience is fundamental to a survival analysis because survival probabilities for different values of t provide crucial summary information from time to event data. The cumulative survival function is described as follows:

$$S(t) = \mathrm{P}(T \geq t) = 1 - F(t)$$

Where $F(t)$ is the c.d.f of $f(t)$. The function $f(t)$ is defined as the probability density function which refers to the probability of the failure time occurring at exactly time t:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t}$$

Life table is one of the oldest survival techniques. The cumulative event-free probabilities for equal distance of time interval are calculated to generate the survival curve. In life table, the censored cases during an interval are assumed to have been followed on average for half the interval. It is also assumed that event occurs uniformly within the interval and withdrawal occurs uniformly within the interval.

All survival functions have similar features - a negative accelerating extinction curve and a monotonically non-increasing function of time. At the beginning of a study, when all the samples are present, the survival probability is 1.00. A common survival analysis technique is the Kaplan-Meier. Kaplan and Meier (1958) and Efron (1967) adapted product limit method to the censored cases tests based on sample cumulative distribution function. When there is no event, the survival curve in a Kaplan-Meier plot will be drawn horizontally over time and only drop (vertically) down at the time of event to the calculated cumulative probability of surviving. Suppose there are k patients have event events in the period of follow-up at distinct times $t_1 < t_2 < t_3 < t_4 < \cdots < t_k$. As events are assumed to occur independently, the probabilities of surviving from one interval to the next may be multiplied together to give the cumulative survival probability. In another way, the probability of being alive at time t_j, $S(t_j)$ is calculated from $S(t_{j-1})$ which is the probability of being alive at t_{j-1}. If the number of patients alive just before t_j is n_j, and d_j is the number of events at t_j, then,

$$S(t_j) = S(t_{j-1})(1 - \frac{d_j}{n_j})$$

The Hazard Function

The hazard is the chronological profile of the probabilities that a portion of the risk set will experience the event during specific time periods, and it is usually denoted by $h(t)$. In another word, it is the probability that an individual who is under observation at a time t has an event at that time. The hazard function represents the instantaneous event rate for an individual who already survived to time t, defined as

$$h(t) = \lim_{dt \to 0} \frac{\Pr\{t \le T \le t + dt \mid T \ge t\}}{dt}$$

The numerator of this expression is the conditional probability that the event will occur in the interval $[t, t + dt)$, given that it has not occurred before, and the denominator is the width of the interval. Dividing one by the other we obtain a rate of event occurrence per unit of time. Taking the limit as the width of the interval goes down to zero, we obtain an instantaneous rate of occurrence. If we already set up the equal time intervals, and the hazard function is straightforward to calculate for the sample population. Under each time interval, identify the risk set and calculate the proportion of group with events during that time interval. Collecting a sequence of hazard probabilities together as a plot over time provides a chronological summary of the risk of event occurring.

The Hazard function has several appealing properties. First, it indicate whether the events occur, and if so, when. The risk of the event occurring during certain time period can be assessed directly. Higher hazard indicate higher risk. Second, both censored and non-censored cases are included in the calculations. Third, in discrete-time survival analysis, the information on variation in the timing of events is not ignored like Cox regression does.

The survival function focuses on not having an event and reflects the cumulative non-occurrence, and the hazard function focuses on the event occurring and relates to incident event rate.

There is a clearly defined relationship between $S(t)$ and $h(t)$, which is given by the calculus formula:

$$h(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\left[\log S(t)\right]$$

The formula above is rarely seen in the survival analysis textbook since most statistical software already incorporates it. Here it is just simply illustrating the relationship. As long as either of $S(t)$ or $h(t)$ is known, the other is automatically determined.

Survival function $S(t)$ is easy to be calculated either from life table or Kaplan-Meier method. Comparing with survival function, there is so simple way to estimate hazard function $h(t)$. The cumulative hazard $H(t)$ is the integral of the hazard which is defined as the area under the hazard function between 0 and t, and it differs from the log-survival curves only by sign. The cumulative hazard $H(t)$ can be treated as the number of events that would be expected for each individual by time t if the events were a repeatable process, and it is used an intermediary measure for estimating $h(t)$. A simple nonparametric method for estimating $H(t)$ is the Nelson-Aalen estimator (Hosmer & Lemeshow, 1999) and kernel smoother to the increments was applied to estimate the hazard (Ramlau-Hansen, 1983).

The importance of hazard function was emphasized by Willett and Singer (1993):

> The hazard function is the cornerstone of survival analysis for several reasons. First, it tells us exactly what we want to know – whether and, if so, when events occur. Its magnitude summarizes the risk of event occurrence in each period… Second, the hazard function involves both noncensored and censored cases…Third, the sample hazard probabilities are computed in every time period that an event occurs – no information is ignored or pooled. Finally, the sample hazard function can be used to estimate the sample survivor functions indirectly in time periods that censoring precludes its direct computation.
>
> Unlike the survivor probabilities, the sample hazard probabilities can be computed in every time period regardless of censoring, censored observations are simply removed from the risk set at the appropriate juncture, reducing the denominator of the hazard quotient. (p. 954)

Hazard function was also emphasized to have many appealing properties, as noted by Singer and Willit (1993):

> The hazard function has many appealing properties which, taken together, explain why it – and not the survivor function – forms the cornerstone of survival analysis. (p. 161)

Collett (2003) mentioned two main reasons for modeling survival data:

> One objective of the modeling process is to determine which combination of potential explanatory variables affects the form of the hazard function. In particular, the effect that the treatment has on the hazard of death can be studied, as can the extent to which other explanatory variables affect the hazard function. Another reason for modeling the hazard function is to obtain an estimate of the hazard function itself for an individual. (p. 56)

Hazard function is the risk of event occurrence instead of survival proportion and its calculation includes both noncensored cases and censored cases, and it doesn't need to throw out the censored cases and no information will be discarded. The sample hazard can be correspondingly computed for each defined time interval to provide a clear picture of pattern of hazard variation over the time. Furthermore, sample survival function can be indirectly calculated from hazard function as long as the data are independent censoring. Independent censoring requires that censoring is unrelated to event occurrence. Under independent censoring, individuals in the risk set don’t differ systematically from censored individuals.

Nonparametric or Parametric Survival Analysis

Survival analysis techniques can be generally classified into nonparametric, parametric, and semi-parametric methods.

The Kaplan Meier method is a nonparametric method. It uses the exact time when the event occurred rather than the intervals of follow-up, and an event rate is calculated at every time point where an event occurs (Kaplan & Meier, 1958). The probability of the event is equal to the number of events at that time divided by the number at risk at that point in time. If there are withdrawals before the time of event, they are subtracted from the number at risk. This is also known as a product-limit method (Kleinbaum & Klein, 2012). Kaplan and Meier (1958) discussed the analysis of right-censored incomplete data and explained the estimation solution via non-parametric maximum likelihood. Maximum likelihood or minimum chi-square can be interpreted as procedures to fit the observations which are selected from an admissible class of distribution. Efron (1967) adapted the product limit method to the censored cases tests based on sample cumulative distribution function. When there is no event, the survival curve in a Kaplan-Meier plot will be drawn horizontally over time and only drop (vertically) down at the time of event to the calculated cumulative probability of surviving. Censoring affects the shapes of survival curve in a situation when a large number of individuals are censored at a single point of time leading to sudden spurious large jumps or large flat sections in survival curves. A low number of individuals at risk especially toward the end of study can also lead to such spurious jump.

There is no specific assumption about the distribution of survival time in Kaplan Meier method. Kaplan (1958) noted:

> "It seems reasonable to call an estimation procedure 'nonparametric' when the class of admissible distribution from which the best-fitting one is to be chosen is the class of all distribution" (p. 459).

In the absence of censorship, several nonparametric tests for survival data, such as log-rank, Wilcoxon, and Kruskal-Wallis tests were developed to compare the survival curves across different time points (Cox & Oakes, 1984; Mantel & Haenszel, 1959; Tarone & Ware, 1977). The log-rank test is equivalent to the Mantel-Haenszel method (Mantel, 1966; Mantel & Haenszel, 1959), and the only difference between log-rank and Mantel-Haenszel is in the way they deal with multiple deaths at exactly the same time point. The Wilcoxon (Breslow, Gehan) test is more sensitive to early survival differences and it gives more weight on earlier cases with events. Contrast with the Wilcoxon test, the log-rank test is more sensitive to later survival difference. If there are more than two groups presented, then a Kruskal-Wallis test is needed. These methods are nonparametric in that they don't make any assumptions about the distribution of survival estimates.

The Kaplan-Meier method has been recognized as an important tool to analyze censored data and is routinely used in many areas, especially in medical research. Miller (1981) introduced various parametric distributions and procedures for survival analysis as well as Kaplan-Meier method, and explained why the Kaplan-Meier method is inefficient, and parametric analysis is recommended especially for the exponential or Weibull distribution.

Miller (1981b) wrote in his introduction:

> The product-limit for Kaplan-Meier estimator is attractive because it is easy to compute and understand. It has an asymptotic normal distribution with an estimated variance that is easily computed by Greenwood's formula. For the underlying probability structure, no assumptions are required other than the basic one of independence between the survival that there is a danger of becoming mentally lazy and not considering parametric modeling. Is there a price to be paid for this easy living? (p. 1077)

Miller also argued that Kaplan-Meier estimator has low efficiencies for high censoring proportions or for surviving fractions that are closer to one or zero. Miller (1983) further examined efficiency of the Kaplan-Meier product-limit estimator to the maximum likelihood estimator of a parametric survival function under a random censoring model. Klein and Moeschberger (1989) compared the efficiencies of Kaplan-Meier method and parametric method, and concluded that parametric estimators outperform the distribution free estimator when a particular parametric model's distribution is assumed under a variety of censoring schemes and underlying failure model. Aranda-Ordaz (1987) examined the comparison of the Kaplan-Meier and the parametric maximum likelihood (MLE) through simulations for several sample sizes, percentages of censorship and proportions of outliers in the sample. The Exponential and Weibull models were used throughout the paper, and it was found that for Weilbull samples the effect can be substantial but for exponential samples it is almost negligible (Aranda-Ordaz, 1987).

Efron (1988) proposed a new modeling approach with the Kaplan-Meier estimator and introduced the techniques of using standard logistic regression to estimate hazard rates and survival curves by providing both estimates and standard errors. From the demonstrated example, it was pointed out that the logistic regression estimation is closely related to Kaplan-Meier curves and the logistic regression approach to the Kaplan-Meier estimate as the number of parameters grows large. That Kaplan-Meier survival estimator is easy to calculate and works well with just a few assumptions had been discussed in many literatures (Meier, Karrison, Chappell, & Xie, 2004; Miller, 1981b; Oakes, 2000).

Meier et al. (2004) considered the discussion for both noncensored data and censored data. For the noncensored data, the Kaplan-Meier estimator was pointed to perform better in estimating the mean when the data are complete, although the parametric estimator may be advantage for point estimation of survival function. And Meier suggested that a parametric estimate of survival curve is necessary in certain extreme situation, such as when the sample size is very small. However, if the functions of survival curve are testing the mean or restricted mean, then the nonparametric approach is preferred over the parametric-based estimate since it is unbiased and entails little or no loss in efficiency.

When comparing two survival distributions, Fisher (1950) argued that

> Even if the original distribution were not exactly normal, that of the mean usually tends to normality, as the size of the sample is increased; the method is therefore applied widely and legitimately to cases in which we have not sufficient evidence to assert that the original distribution was normal. (p. 112)

The logrank test cannot be used to adjust for the effect of explanatory variables. The adjustment for explanatory variables will improve the precision of estimation with the treatment effect.

Breslow (1974) also addressed importance of distribution in multiple regressions for survival data,

> The past few years have witnessed intense activity among statisticians in adapting the powerful methods of multiple regression and covariance analysis for use with censored survival data. Some of these efforts have been directed towards extending traditional least squares methods based on normal distribution theory. However, researchers have found that working with distributions specifically proposed for life testing and survival problems, such as the exponential, Weibull, and Gompertz, often leads to methods which are mathematically more tractable and are conceptually and computationally somewhat simpler than is true for the normal. Regression models proposed for these distributions generally involve the assumption of proportional hazard functions which has long been used in the theory of competing risks. (p. 89)

In contrast to non-parametric distributions, some survival time follows a known distribution is called parametric distribution. The parameters in parametric distributions can be estimated. The classical parametric survival distributions are the exponential, the Weibull, the log-logistic, the lognormal and the generalized gamma. For parametric survival models, time is assumed to follow some distribution whose probability density function $f(t)$ can be expressed in terms of unknown parameters. Once a probability density function is specified, the corresponding survival function and hazard functions can be determined.

Cox Regression

Estimating survival functions, median survival time, and hazard function are descriptive statistics to answer when and whether a sample of subjects has the events of interest. After introduction of the proportional hazards model by Cox (1972), the attention shifted from hypothesis testing to modeling effects of explanatory variables. "Statistical models of hazard express hypothesized population relationships between entire hazard profiles and one or more predictors"(John B. Willett & Singer, 1991). The Cox model is the most commonly used multivariate approach for analyzing survival time data in medical research. The Cox's model permits an analysis in which survival time is treated as continuous variable and explanatory variables can be continuous scale or categorical form. The Cox model is a method for investigating the effect of several variables upon the time a specified event takes to happen. Cox's model includes a simple multiplicative factor of baseline hazard function and the effects of the covariates on the hazard. The baseline hazard is defined as the hazard function for that individual with zero on all covariates. Since the baseline hazard is not assumed to be of a parametric form, Cox's model is referred to as a semi-parametric model for the hazard function. Researchers in medical sciences often tend to prefer semi parametric instead of parametric because of its less assumptions.

Mathematically, the Cox model is written as

$$h(t,\mathrm{X}) = h_0(t) \times e^{\beta'\mathrm{X}}$$

The Cox model formula says that the hazard at time t is the product of two quantities. The first is the baseline hazard function $h_0(t)$ which is left unspecified but must be positive. The $h_0(t)$ involves t but not $X's$. The second quantity is the exponential expression e to the linear sum of $\beta_P X_P$. Suppose for each individual, there are one or more measurements are available, let say variables $x_1,...,x_p$, and their corresponding impact are measured by the size of the respective coefficients $(\beta_1, \beta_2, \cdots, \beta_P)$.

The Cox model can be either one of the following form:

$$h_i(t,x) = h_0(t) e^{\beta_1 x_{i1} + ... + \beta_P x_{iP}}$$

or

$$\log h_i(t,x) = \log h_0(t) + \beta_1 x_{i1} + ... + \beta_P x_{iP}$$

Another important feature of Cox model is that the baseline hazard is a function of t, but not specified function. This property makes the Cox model as a semi-parametric model. In contrast, a parametric model is one in which survival time is assumed to follow a known distribution. The survival or hazard function form is completely specified, except for the values of the unknown parameters.

Breslow (1974) tested three models in the comparison of survival curves for a clinical trial of maintenance therapy of children leukemia. The log linear exponential, linear exponential and nonparametric generalization models were tested for the same data.

Comparing with some small studies with few numbers of factors interests, some studies contain a large number of factors and relatively more information. It is not an easy task and always time-consuming to choose which variables should be included in the regression model

(Hosmer & Lemeshow, 1999). The availability of stepwise methods which contain either backward elimination or forward selection in many software packages makes this procedure easy to use without adding human decision. Adding or removing covariates are fully based on statistical significance at some pre-decided level, however, this automated selection technique has its own disadvantages because it only evaluates a small number of the set of possible models, especially for smaller sample sizes and when few event occur (Clark et al., 2003b), and it is sometimes lack of real meaning since selection of covariates only based on the statistical significance without involve any human's experience.

Adding interaction terms to a regression model can greatly expand understanding of the relationship among the variables in the model and allows more hypotheses to be tested. The interaction effect was also emphasized in many papers (Clark et al., 2003b; Hosmer & Lemeshow, 1999; Thomas & Reyes, 2014). The importance of testing of interaction in regression approach was emphasized in Breslow (1974) and Sawilowsky (1990). With exploring the interaction, the formal analyses may be invalidated. Sawilowsky (1990) reviewed nonparametric techniques for the testing of interaction in experimental design and showed they are robust, powerful, versatile, and easy to compute comparing to parametric methods.

The Cox proportional hazards model is widely used in epidemiological analyses of cohort data (Rothman, Greenland, & Lash, 2007). The hazard function is taken to be a function of the explanatory variables and unknown regression coefficients multiplied by an arbitrary and unknown function of time (Cox, 1972). The coefficients in a Cox model relate to hazard. A positive coefficient indicates a worse prognosis and a negative coefficient indicates a protective effect of the variable with which it is associated. Efron (1977) suggested a simple method for the regression analysis of censored data, and explicated the connection between Cox regression with the Kaplan-Meier estimator of a survival curve. The calculation by Efron showed Cox regression has full asymptotic efficiency under many realistic situations.

Kleinbaum and Klein (2012) wrote,

> A key reason for the popularity of the Cox model is that, even though the baseline hazard is not specified, reasonably good estimates of regression coefficients, hazard ratios of interests, and adjusted survival curves can be obtained for a wide variety of data situations. Another way of saying this is that the Cox PH model is a “robust” model, so that the results from using the Cox model will closely approximate the results for the correct parametric model. (p. 96).

Cox proportional hazard model is one type of event history model. It makes no assumptions about the shape of the hazard function, and it treats the time as continuous. With the growing popularity of the semiparametric Cox proportional hazard model, it is important to find convenient ways to detect if the model is well specified. Kleinbaum and Klein (2012) introduced three approaches in his Chapter III for evaluating the proportional hazard (PH) assumption of the Cox model including a graphical procedure, a goodness-of-fit testing procedure, and a procedure that involves the use of time-dependent variables. The graphical and goodness of fit procedures for proportional hazard model had been discussed in many papers (Arjas, 1988; Moreau, O'Quigley, & Mesbah, 1985; Parzen & Lipsitz, 1999; Wei, 1984).

The likelihood function is a mathematical expression which describes the joint probability of obtaining the data actually observed on the subjects in the study as a function of the unknown parameters in the model being considered. The formula for the Cox model likelihood function is called a partial likelihood function. The phrase partial likelihood considers the probabilities only for those subjects who fail, and does not explicitly consider those subjects who are censored. A detailed description of the mathematics of partial likelihood estimation can be found in Allison (1984), and the general properties are as follows:

> The method relies on the fact that the likelihood function for data arising from the proportional hazards model can be factored into two parts: one factor contains information only about the coefficients β_1 and β_2; the other factor contains information about β_1 and β_2, and the function $\alpha(t)$. Partial likelihood simply discards the second factor and treats the first factor as if it were an ordinary likelihood function. The first factor depends on the order in which events occur, not on the exact times of occurrence. (p. 37)

The effects of covariates are assumed to be constant over time in Cox proportional hazard model. Comparing with Cox proportional hazard model, the discrete-time survival model can allow the effects of covariates varying over the time.

Although Cox proportional hazard model is widely used in the many areas, there are some important limitations. The most significant is the basic assumption that cancels the interaction when the time is not in the equation. Singer and Willet (1991) state that time is crucial for time-varying predictor and time should be included in the model. The other major limitation is that it is lacking a term to represent the observed heterogeneity in the Cox proportional model. The latter one has been found to be especially significant when dealing with repeated events.

Discrete-time Survival Analysis

Cox (1972) introduced the discrete-time hazard model in terms of logit-hazard rather than hazard in his seminal article. The discrete-time survival analysis had been widely used in the educational research and social research (Allison, 1982; Judith D. Singer & Willett, 1993; J. B. Willett & Singer, 1993; John B Willett & Singer, 1995; John B. Willett & Singer, 1991). For example, event history data are usually collected in a retrospective cross-sectional survey, where dates are recorded to the nearest month or year, or event history data are prospectively collected in waves of a panel study.

Allison (1982) discussed about how censoring and time-varying explanatory variables impact standard survival analysis,

> "Although event histories are almost ideal for studying the causes of events, they also typically possess two features – censoring and time-varying explanatory variables – that create major difficulties for standard statistical procedures. In fact, the attempt to apply methods to such data can lead to serious bias or loss of information." (p. 62).

Under the following situations, discrete-time model may be more appropriate: where events can only occur at regular discrete points in time; where the events can occur at any point in time, but available data record only the particular interval of time in which each event occurs.

> "Discrete-time methods have several desirable features. It is easy, for example, to incorporate time-varying explanatory variables into a discrete-time analysis. Moreover, when the explanatory variables are categorical (or can be treated as such), discrete-time models can be estimated by using log-linear methods for analyzing contingency tables. With this approach one can analyze large samples at very low cost. When explanatory variables are not categorical, the estimation procedures can often be well approximated by using ordinary least-squares regression. Finally, discrete-time methods are more readily understood by the methodologically unsophisticated." (p.63).

Willett and Singer (1991) discussed the principles of survival analysis, and showed how they apply into educational research by using two examples: teacher entry into and exit from teaching and student entry into and exit from school. They believed that discrete-time survival analysis is the good choice for educational transitions:

> Of all the survival methods available, we believe that discrete-time survival analysis offers the most promise for exploring educational transitions. The application is natural; educational data are typically collected at regular intervals, not in continuous time. Discrete-time survival analysis does not require dedicated software; it can be implemented using routines available in most standard statistical packages. In addition, it facilitates investigation of the effect of time-varying predictors; it can be used to detect interactions between predictors and time (as when the effects of a predictor fluctuate with the passage of time), and it can be used to study the many competing risks of exit – voluntary and involuntary terminations among teachers and dropping out and graduation among students. (p. 439)

Willett and Singer (1993) explained three obstacles of survival analysis to model educational data:

> First, most readily available software is designed for fitting models that incorporate only time-invariant predictors (those whose values are constant over time). Yet the values of many predictors of educational processes – such as financial aid, the availability of support and remedial programs, and the nature of the peer support network – fluctuate naturally with time. Second, the most popular model in use today (the continuous time proportional hazard model) is predicated on the often unrealistic assumption that the effect of a predictor on event occurrence is constant over time. Yet in many educational applications, the effects of predictors – such as teacher salary or peer pressure – will vary over time. Third, continuous time models (in which researchers assume that they know the precise instant when the event occurs) don't not adapt readily to school contexts, where time is so often measured discretely, in quarters, semesters, or years. (p. 156)

The advantages of discrete-time survival analysis used in education research were also emphasized by Judith D. Singer and Willett (1993) and Allison (1982). First, much history event data are collected in a discrete time manner due to the logistical and financial reasons. Second, the Cox regression model assumes the effect of predictor are constant, however, many effect of predictors will vary with time. These time-varying predictors can be easily included into the discrete-time models. Third, common model violations can be easily tested and remediated for discrete-time model. Finally, discrete-time survival analysis is specified by Cox as a type of logistic regression, and the calculation and estimation don't need additional special statistical software and can be carried out within a standard statistical package(Pierce, Stewart, & Kopecky, 1979; Prentice & Gloeckler, 1978).

Discrete-time survival analysis is a useful analogue to the continuous time proportional hazards model. The smaller the time interval the smaller that difference will be because as the interval width becomes smaller, the logistic model converges to the proportional hazard model (Thompson, 1977). In survival analysis, timing of event occurrence is critically important. Sometime the event occurrence could be thought of in a discrete time framework. For example, in many medical screening programs, the disease status is ascertained only during annual screening or periodic checking up. The particular discrete time interval instead of the exact date of event incidence is the only thing to know.

Sharaf and Tsokos (2014) predicted survival time of localized melanoma patients by using discrete survival time method. The discrete survival time method was able to provide better results when applied on follow-up data sets. Xie et al. (2003) used the discrete time survival method into the mental health research. In the mental health, the interested outcomes are often the onset, relapse, and remission from an illness. When the data are collected at discrete-time periods, then the discrete-time survival analysis model is more suitable than the continuous-time survival model.

Discrete-time survival analysis is not only useful in some medical science but also provide an idea framework whether and when the event happens for educational researcher (Bray , Almirall, Zimmerman, Lynam, & Murphy, 2006; Henry, Thornberry, & Huizinga, 2009; McCallon, 2009; Judith D. Singer & Willett, 1993), occupational and environmental science (Richardson, 2010).

Masyn (2003) modeled single event discrete-time data by using a latent class regression model. The interested events were measured in discrete-time or grouped-time intervals. The data were presented as a set of binary event indicators and observed risk indicators. Time-dependent and time-independent covariates were tested in the models. All the models in this Masyn's dissertation used the domestic violence data with an alcohol treatment intervention. The latent class regression framework was presented in Muthén and Masyn (2005).

A discrete-time survival analysis was conducted for analyzing the departure patterns exhibited by students enrolled in a large, private church-related university over a six-year period (McCallon, 2009). Several risk factors including ethnicity, religious preference, and matriculation status were examined. Discrete-time survival was proven to be an effective procedure in this social study. Henry et al. (2009) initiated the prevention strategies after exploring the relationship between truancy and the onset of marijuana use for the teenagers.

Discrete survival analysis can be treated as one form of logistic regression (Henry et al., 2009; Judith D. Singer & Willett, 1993; Xie et al., 2003). Logistic regression is an efficient and powerful method to analyze the effect of a group of independent variables on a binary outcome by quantifying each independent variable's unique contribution. However, the accuracy of logistic regression model is mainly relying on satisfactions of assumptions as well as the right strategy of building model (Stoltzfus, 2011). Adjusting confounder is also an important issue which is related with logistic regression. Mantel and Haenszel (1959) considered stratification on confounding variables for retrospective studies and suggested the odds ratio should be formed based on the combining estimator from individual strata.

Selection of the Discrete-time Survival Analysis and Cox Regression Survival Analysis

In order to distinguish the difference of Cox regression and logistic regression, it must know that the distinction between rate and proportion. The incidence (hazard) rate refers to the number of new cases of events per population at-risk per unit time. If the event of interested is death, then it is called morality rate. Cumulative incidence refers to the proportion of new cases that develop in a given time period. Cox regression aims to estimate the hazard ratio which is the ratio of incidence rates, while logistic regression aims to estimate the odds ratio which is the ratio of proportions.

Cox regression does not require that you choose some particular probability model to represent survival times, and is therefore more robust than parametric methods (e.g. exponential or Weilbull). Unlike non-parametric Kaplan-Meier method, Cox regression is a semi-parametric and it can accommodate both discrete and continuous measures of event times. Furthermore, both constant and time-dependent covariates can be incorporated into the model over the course of the observation period.

Formalizing a Discrete-Time Survival Analysis Model

For each homogeneous group of individuals, the single event is nonrepeatable. On another way, one individual can experience the event only once. Once the event occurs, it cannot occur again. Repeated events model are not discussed here. To record event occurrence in discrete intervals, divide continuous time into an infinite sequence of contiguous time periods $(0,t_1]$, $(t_1,t_2]$,..., $(t_{j-1},t_j]$,..., and so forth. The letter j represents the period index, and jth period begins right after time t_{j-1} (using the initial parenthesis) and ends at, and includes time t_j (using the including bracket). For example, if the time is measured as years, when an event occurs any time after t_2 (the last day of Year 2) and before t_3 (including the last day of year 3), then the event is accounted as happening the 3rd time interval $(t_2,t_3]$. Adopting common mathematical notation, [brackets] denote inclusions and (parentheses) denote exclusions.

A discrete-time hazard rate h_j can be defined as a conditional probability that a randomly selected individual will experience the target event in time period j, given that he or she did not experience the event prior to j:

$$h_j = \Pr[T = j \mid T \geq j] \tag{1}$$

Where T represent the discrete random variable that indicates the time period j when the event occurs for a randomly selected individual from the population.

The discrete-time hazard rate h_j is a probability whose value lies between 0 and 1. The goal of the discrete-time survival analysis is to estimate these conditional probabilities h_j and investigate their dependence on selected covariates. Thus, the heterogeneities from covariates need to be considered in the hazard model in order to determine whether different types of individuals with their specific covariates have different hazard functions.

Let assume there are P covariates, $Z_p(p=1,2,...,\text{P})$ refers to each specific covariate for the members of population. The vector $z_{ij}=[z_{1ij},z_{2ij},...,z_{Pij}]$ can be used to represent the individual i 's value for each of the P covariates in time period j , in such notation, Z_p can be constant over time or may vary over time. The values of each covariate remain constant within each time periods even they can be different in different time periods. After introducing the individual i and time period j as long as the P covariates, the discrete-time hazard rate h_j can be extended into the following form:

$$h_{ij}=\Pr\left\{T_i=j\,|\,T_i\geq j,Z_{1ij}=z_{1ij,}Z_{2ij}=z_{2ij,}...Z_{Pij}=z_{Pij}\right\} \tag{2}$$

The Equation 2 indicates that the hazard depends on each individual's values on a vector of predictors.

Cox (1972) proposed to re-parameterize the probabilities h_{ij} into a logistic dependence relationship on covaraites and the time periods. The model represents the log-odds of event occurrence as a function of covariates and also has the attributes of the baseline profile. The proposed population discrete-time hazard model is therefore:

$$h_{ij}=\frac{1}{1+e^{-[(\alpha_1D_{1ij}+\alpha_2D_{2ij}+...+\alpha_JD_{Jij})+(\beta_1Z_{1ij}+\beta_2Z_{2ij}+...+\beta_JZ_{Pij})]}} \tag{3}$$

Where $[D_{1ij,}D_{2ij,}...,D_{Jij}]$ is a sequence of dummy variables, with values $[d_{1ij,}d_{2ij,}...,d_{Jij}]$ indexing time periods. J refers to the last time period observed for anyone in the sample, and j_i refers to the last time period when individual i was either observed or experienced the event. The time-period dummy variables are defined consistent to everyone. For example, $d_{1ij}=1$

when $j = 1$, and $d_{1ij} = 2$ when $j = 2$, and so on. The vector $[\alpha_{1,}\alpha_2,...,\alpha_J]$ capture the baseline level of hazard in each time period, and the vector $[\beta_{1,}\beta_2,...,\beta_P]$ represent the effects of predictors on the baseline hazard function.

Taking logistic transformations of both sides of the equation, the equation of (3) changes to the following form:

$$\log(\frac{h_{ij}}{1-h_{ij}}) = (\alpha_1 D_{1ij} + \alpha_2 D_{2ij} + ... + \alpha_J D_{Jij}) + (\beta_1 Z_{1ij} + \beta_2 Z_{2ij} + ... + \beta_J Z_{Pij}) \qquad (4)$$

The Equation 4 above expresses a conditional log-odd which is linear function of a constant term α_j specific to period, and of the values of the predictors period j multiplied by the appropriate slope parameters. With event history data on a random sample of n individuals $[i = 1, 2, ..., n]$, the discrete-time hazard model can be fitted by Equation 3 and corresponding parameters in Equation 3 can be estimated.

The direct connection between logits, odds, and probabilities is shown in the following table 1. From the relationships among them, it is understandable that hazard profiles can also be displayed as odds instead of probabilities. If a hazard probabilities in a time period is 0.4, there is a 40% chance that the event of interest will occur in the period and a 60% chance that it will not , given no prior occurrence. The odds of event occurrence in this period are 0.4 and 0.6, which refers to the odds equal to 4/6 or 0.66.

Table 1

Relationship between logit, odds and probability (Judith D. Singer & Willett, 1993)

Original scale	Desired scale	Transformation
Logit	Odds	$Odds = e^{\text{logit}}$

Odds	Probability	$\text{Probability} = \frac{odds}{1+odds} = \frac{e^{\text{logit}}}{1+e^{\text{logit}}}$
Logit	Probability	$\text{Probability} = \frac{1}{1+e^{\text{logit}}}$

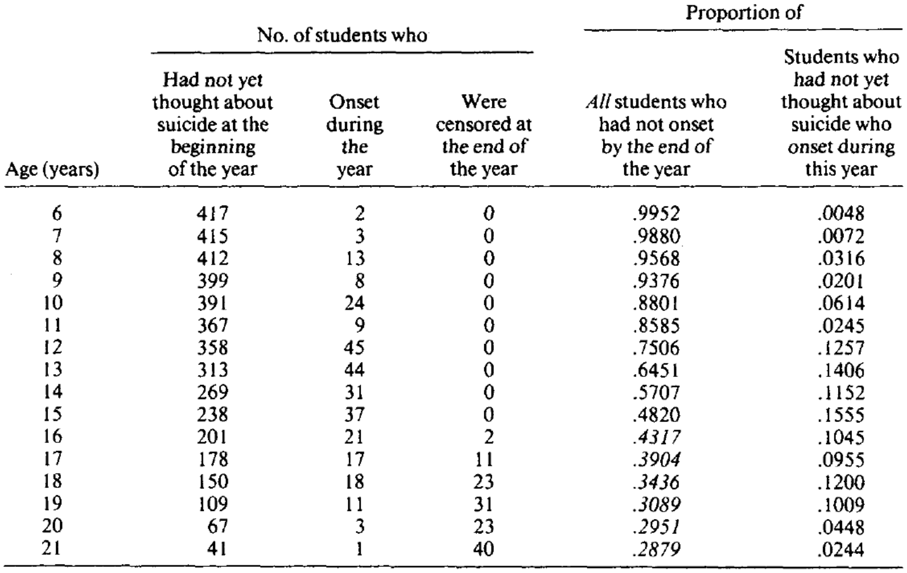
Example of Data Structure

A typical example in the following Table 2, reprinted with permission, displays an example of the traditional method for summarizing event history data by Willett and Singer (1993). The first column lists student age in years. The next three columns tally the number of students who had not yet thought of suicide at the beginning of each age period, the number of students who contemplated suicide during each period, and the censored numbers at the end of the period. The last two columns give one proportion of who had not onset by the end of the period and the proportion of students who had not yet thought about suicide who onset during each period. For example, total 417 students were present at the beginning of the year 6, and 2 students had contemplated suicide during Year 6, then hazard probability in Year 6 is therefore 2/417, or .0048. The corresponding survival probability at the end of Year 6 is .9952 which equals to 1 minus hazard probability 0.048 at Year 6. Then, at the end of Year 6, the remaining risk population was 415.

From Year 16, the hazard function involved both censored and non-censored cases, and the hazard probability was 21/201, or .1045. Two cases were censored during the 16th year, then at the beginning of 17th year, the risk set only considered total 178 students which were taken out 21 events and 2 censored cases from 16th year. The survival probability by the end of 17th year equals to 0.4317*(1 - .0955), or .3904. It is a conditional probability – survival proportion for 17th year based on the population who are event free for their 16th year.

Table 2

What Do Survival Data Look Like? Age at First Suicide Ideation Among 417 College Students



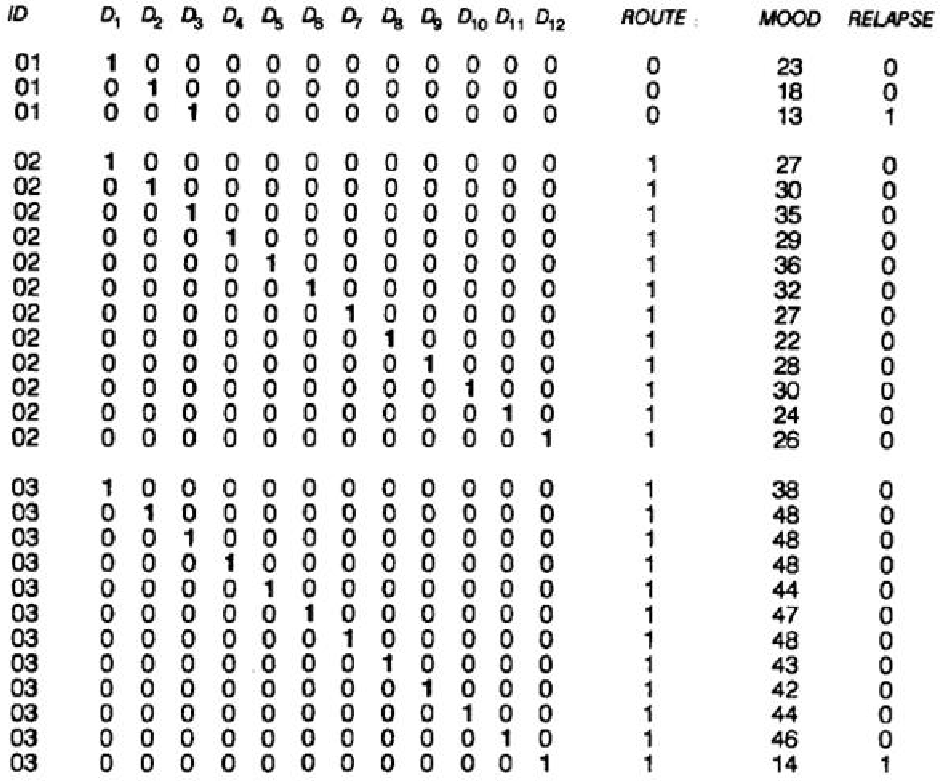
	No. of students who			Proportion of	
Age (years)	Had not yet thought about suicide at the beginning of the year	Onset during the year	Were censored at the end of the year	*All* students who had not onset by the end of the year	Students who had not yet thought about suicide who onset during this year
6	417	2	0	.9952	.0048
7	415	3	0	.9880	.0072
8	412	13	0	.9568	.0316
9	399	8	0	.9376	.0201
10	391	24	0	.8801	.0614
11	367	9	0	.8585	.0245
12	358	45	0	.7506	.1257
13	313	44	0	.6451	.1406
14	269	31	0	.5707	.1152
15	238	37	0	.4820	.1555
16	201	21	2	*.4317*	.1045
17	178	17	11	*.3904*	.0955
18	150	18	23	*.3436*	.1200
19	109	11	31	*.3089*	.1009
20	67	3	23	*.2951*	.0448
21	41	1	40	*.2879*	.0244

Note: Reprinted from "Investigating Onset, Cessation, Relapse, and Recovery: Why You Should, and How You Can, Use Discrete-Time Survival Analysis to Examine Event Occurrence" by John B Willet and Judith D. Singer, 1993, Journal of Consulting and Clinical Psychology Volume 61(6) p. 953. Copyright 1993 by the American Psychological Association Inc.

In many social and education research studies, for instance, the suicide example above in Willet (1991), the survey or the interview was given at 1-year intervals to obtain information. When the main aim of study is to investigate the relationship between some social economic factors and onset of suicide, the social economic factors like family income, parents' marital status, and etc. may change during the follow-up periods. It is not reasonable to keep those factors as constant all the way through the study period, but instead treat each individual measurement as one record and incorporate them in the multiple regressions.

The figure 1 listed the data in person-period format for discrete-time survival model. The first column is the unique identification ID for each subject. The variables D1, D2, …, D12 indicated 12 time intervals from the first time period through the twelve time period. Except the 12 time interval variables, the data also include one categorical variable and one continuous variable. The categorical variable was the primary mode of cocaine ingestion before treatment

(ROUTE, coded 0 = all other routes, 1 = intranasal), and the continuous variable indicated the mood scale of subject (MOOD). The last column (REPLASE) in the right side is the event of interest - suicide occurrence (coded 0 = non-event, 1 = event). The first three records in figure 3 indicate the same person, and the person had the event in the third time interval. The person cocaine ingestion kept same and mood scale changed through the time. The following 12 records were for the person with ID 02, and the person had not experienced the event at the end of 12th time interval. In each time interval, the mood scale was different, and the 2nd person used intranasal cocaine ingestion mode. The third person experienced the event during the 12th time period.



ID	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}	ROUTE	MOOD	RELAPSE
01	1	0	0	0	0	0	0	0	0	0	0	0	0	23	0
01	0	1	0	0	0	0	0	0	0	0	0	0	0	18	0
01	0	0	1	0	0	0	0	0	0	0	0	0	0	13	1
02	1	0	0	0	0	0	0	0	0	0	0	0	1	27	0
02	0	1	0	0	0	0	0	0	0	0	0	0	1	30	0
02	0	0	1	0	0	0	0	0	0	0	0	0	1	35	0
02	0	0	0	1	0	0	0	0	0	0	0	0	1	29	0
02	0	0	0	0	1	0	0	0	0	0	0	0	1	36	0
02	0	0	0	0	0	1	0	0	0	0	0	0	1	32	0
02	0	0	0	0	0	0	1	0	0	0	0	0	1	27	0
02	0	0	0	0	0	0	0	1	0	0	0	0	1	22	0
02	0	0	0	0	0	0	0	0	1	0	0	0	1	28	0
02	0	0	0	0	0	0	0	0	0	1	0	0	1	30	0
02	0	0	0	0	0	0	0	0	0	0	1	0	1	24	0
02	0	0	0	0	0	0	0	0	0	0	0	1	1	26	0
03	1	0	0	0	0	0	0	0	0	0	0	0	1	38	0
03	0	1	0	0	0	0	0	0	0	0	0	0	1	48	0
03	0	0	1	0	0	0	0	0	0	0	0	0	1	48	0
03	0	0	0	1	0	0	0	0	0	0	0	0	1	48	0
03	0	0	0	0	1	0	0	0	0	0	0	0	1	44	0
03	0	0	0	0	0	1	0	0	0	0	0	0	1	47	0
03	0	0	0	0	0	0	1	0	0	0	0	0	1	48	0
03	0	0	0	0	0	0	0	1	0	0	0	0	1	43	0
03	0	0	0	0	0	0	0	0	1	0	0	0	1	42	0
03	0	0	0	0	0	0	0	0	0	1	0	0	1	44	0
03	0	0	0	0	0	0	0	0	0	0	1	0	1	46	0
03	0	0	0	0	0	0	0	0	0	0	0	1	1	14	1

Figure 1. The Person-period Dataset

Note: Reprinted from "Investigating Onset, Cessation, Relapse, and Recovery: Why You Should, and How You Can, Use Discrete-Time Survival Analysis to Examine Event Occurrence" by

John B Willet and Judith D. Singer, 1993, Journal of Consulting and Clinical Psychology Volume 61(6) p. 958. Copyright 1993 by the American Psychological Association Inc.

When Will Cox and Logit Estimates Be Similar?

Cox model had two components: baseline hazard function that is left unspecified but must be positive, and a linear function of a set of k fixed covariates is exponentiated. The Cox model can be written:

$$h_i(t) = h_0(t)e^{\beta_1 x_{i1}+...+\beta_k x_{ik}}$$

or sometimes as:

$$\log h_i(t) = \log h_0(t) + (\beta_1 x_{i1} + ... + \beta_k x_{ik})$$

Where $h_0(t)$ is the baseline hazard function which can take on any form. The x_{ik} is a vector of covariates with coefficients βs.

Cox estimates are effects on log scale, and $\exp(\beta)$ are hazards ratios.

Discrete-time logit model

In order to distinguish the hazard probabilities in Cox regression model, let p_{ti} be the probability that individual i has an event during the interval t in the discrete-time analysis, given that no event has occurred before the start of t.

$$p_{ti} = \Pr(y_{ti} = 1 \mid y_{t-1,i} = 0)$$

p_{ti} is a discrete-time approximation to the continuous-time hazard function $h_i(t)$. The logit model is listed below to expand the data to fit a binary response model.

$$\log(\frac{p_{ti}}{1-p_{ti}}) = \alpha \mathrm{D}_{ti} + \beta \mathrm{x}_{ti}$$

Where D_{ti} is a vector of functions of the cumulative duration by interval t with coefficients α. Changes in p_{ti} with t are captured in the model by $\alpha\mathrm{D}_{ti}$. Here D_{ti} is specified as step function.

$$\alpha D_{ti} = \alpha_1 D_1 + \alpha_2 D_2 + ... + \alpha_q D_q$$

Where D_1, ... , D_q are dummies variables for time interval $t = 1,...,q$ and q is the maximum observed event time.

x_{ti} is a vector of covariates (time-varying or constant over time) with coefficients β.

Logit estimates are effects on log-odds scale, and $\exp(\beta)$ are hazard-odds ratios.

In general, Cox and logit estimates will get closer as the hazard function becomes smaller because:

$$\log(h(t)) \approx \log(\frac{h(t)}{1-h(t)}) \text{ as } h(t) \to 0.$$

The discrete-time hazard will get smaller as the width of the time intervals become smaller. A discrete-time model with a complementary log-log link, $\log(-\log(1-p_t))$, is an approximation to the Cox proportional hazard model, and the coefficients are directly comparable(Steele & Washbrook, 2013).

Model Evaluation

Null Hypothesis

To determine whether the regression coefficient is different from zero, there are several hypothesis tests that can be performed. Let's say the null hypothesis assumes that the predictor variable is 0 for the population. If there is sufficient evidence in the sample to conclude that the regression coefficient is significantly different from 0, then the alternative hypothesis can assume that the predictor variable has some effect on the dependent variable. The z test and the likelihood ratio statistic (alternative Wald statistic) are methods of testing the null hypothesis. The likelihood ratio and the Wald statistic typically give similar results for the same data set, given the sample is large enough (Wright, 1995).

The z test is used for testing the significance of individual parameters. It is calculated by dividing the estimated parameter estimate for that predictor by its standard error. Ratios of 1.96 and 2.58 or larger can be considered significant for an α of 0.05 and 0.01.

The Likelihood Ratio Statistic is similar to the F test in that a large value means the population differs from zero. The probability is associated with the likelihood will determine if it is a significant difference. The likelihood ratio statistic is also used for comparing the fits of full and restricted models. Smaller value indicates a better fit of the model.

The Wald statistic is an alternative method to the likelihood ratio for testing the significance of individual coefficient. It is obtained by comparing the maximum likelihood estimate of the β to an estimate of its standard error. It can be calculated to be asymptotically distributed as a chi-square distribution or it can follow a normal distribution. The SPSS logistical regression and Cox regression procedures use the chi-square distribution. If the β is large, the estimated standard error is inflated, resulting in failure to reject the null hypothesis when the null hypothesis is false.

Goodness of Fit Measures

One of the most common question for any regression method is “How do I know if the model fits the data”. The approaches to answer this question generally can be classified into two categories: measures of predictive power and goodness of fit tests (Allison, 2014).

Maximum likelihood

The estimations of the parameters $\beta's$ in general Cox model are called maximum likelihood (ML). As with logistic regression, the ML estimates of the Cox model parameters are derived by maximizing a likelihood function, usually denoted as L . The likelihood function is a mathematical expression which describes the joint probability of obtaining the data actually observed on the subjects in the study as a function of the unknown parameters (the $\beta's$) in the model being considered.

For continuous-time, the general likelihood equation for censored data is:

$$L = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [1 - F(t_i)]^{1-\delta_i}$$

For discrete-time model, the likelihood is presented as:

$$L = \prod_{i=1}^{n} [\Pr(T_i = t_i)]^{\delta_i} [\Pr(T_i > t_i)]^{1-\delta_i}$$

For both continuous-time and discrete-time models, δ_i is set equal to 1 if i is uncensored; otherwise it is zero.

The partial likelihood (PL) only considers probability for subjects who fail, and it doesn't consider the probabilities for subjects who are censored. The formula for the Cox model likelihood function is actually called a "partial" likelihood function rather than a complete likelihood function since the likelihood for the Cox model does not consider probabilities for all subjects. At the jth failure time, L_j denotes the likelihood of failing at this time, given survival up to this time. Let's say the set of individuals at risk at the jth failure time is called the risk set $R(t_{(j)})$, and this set will get smaller in size as the failure time increases.

$$PL = L_1 \times L_2 \times L_3 \times ... \times L_k = \prod_{j=1}^{k} L_j$$

Although the partial likelihood focuses on subjects who fail, survival time information prior to censorship is used for those subjects who are censored. In another words, a person who is censored after the jth failure time is part of risk set used to compute L_j, even though this person is censored later.

The likelihood function is generally done by maximizing the natural log of L by taking partial derivatives of log of L with respect to each parameter in the model, and then solving a system of equations. Normally computer will do this step by carrying out through iterations.

The log-likelihood function for the Cox proportional hazard model looks like this

$$L(\beta) = \sum_{i=1}^{n} \left\{ c_i \ln\left[h_0(t_i)\right] + c_i x_i \beta + e^{x_i \beta} \ln\left[S_0(t_i)\right] \right\}$$

In logistic regression, the log-likelihood is the criterion for selecting parameters. The likelihood itself is a small number, so the log of the likelihood is multiplied by -2 and approximates a chi-square distribution. Smaller values indicate a better prediction of the dependent variable. In the SAS and SPSS package, the log of the likelihood is commonly abbreviated as -2LL. The likelihood equals to 1 indicates the model perfectly fit, and -2LL equals to 0. The likelihood ratio test is used to test the significance of the coefficients in the model. The -2 Log Likelihood statistics has a chi-square distribution under the null hypothesis that all coefficients in the model are zero. The difference in fit between two nested models is assessed by looking at the change in -2LL, with degrees of freedom equal to the number of β parameters.

R-Squares Statistics

There are many different ways to calculate the R^2 and there is no consensus on which one is best. The R-squares is an analogous to the R^2 in linear regression. It indicates a proportional reduction in chi-square or in the absolute of the log-likelihood (Hosmer & Lemeshow, 1989). In logistic regression, the most common R square is the Cox and Snell R^2:

$$R^2{}_{C\&S} = 1 - (L_M / L_0)^{2/n}$$

Where L_0 is the likelihood function for a model with no predictors, and L_M is the likelihood for the model being estimated.

Other Measures of Goodness of Fit

The Akaike Information Criterion (AIC) and Bayesian information criterion (BIC) can be used to evaluate and compare the goodness of fit of an estimated statistical model (Akaike, 1974; M. A. Pourhoseingholi et al., 2007; M. Pourhoseingholi et al., 2011). A simulated data will be used to compare the properties of two approaches. Two approaches will be fitted into the models on clinical data to demonstrate the closeness of this relationship in different types of situations. The AIC is given by:

$$\mathrm{AIC} = -2 * \mathrm{Log(likelihood)} + k * npar$$

Where $npar$ represents the number of parameters in the fitted model, and $k = 2$ for the usual AIC, or $k = \log(n)$ (n being the number of observations) for BIC. The smaller AIC represents the favor of model with smaller residual error.

Robust Estimation

The robustness estimations in Cox Regression (Bednarski, 1993; Farcomeni & Viviani, 2011; Ten Have, Miller, Reboussin, & James, 2000; van Houwelingen & Putter, 2014) and logistic regression (Bianco & Yohai, 1996; Kordzakhia, 2001; van Houwelingen & Putter, 2014) had been discussed.

The Cox regression model does not make any assumption on the underlying hazard. However it relies on the proportional hazards assumption. The traditional statistical solution of the robustness problem is to extend the proportional hazard model by stratification or the introduction of time-varying effects of the covariates in the Cox model.

CHAPTER 3 METHODOLOGY

Overview of the Research Design

The purpose of this study is to determine whether the discrete-time survival model is comparable to the Cox regression model when both methods are used to investigate the relationship between predictors and outcome variable.

The Cox's proportional hazard model has been proposed for the purpose of exploring the effects of one or multiple variables on survival. Cox's proportional hazard model is analogous to a multiple regression model and enables the difference between survival times of particular groups of patients to be tested while allowing for other factors. In the Cox' proportional model, the dependent variable is the "hazard". The hazard is the probability of experiencing the event given the individuals has survived up to a given point in time.

Comparing with the Cox proportional hazard model, the discrete-time survival model is most likely to be used in educational research when looking at the timing of certain educational events. Regular continuous-time method don't allow for the flexibility inherent in a discrete-time way. Under the discrete-time method, both time-invariant and time-varying predictors can be used, and the interaction of predictors with time can also be tested and implemented into the model.

Data Structures

For a continuous-time Cox hazard model, the data structure is "Person-Level" format. Normally, in a typical person-level data set, each individual in the sample has one record (line). Each record in the dataset indicates i th individual subject with his or her corresponding following information:

Censoring. The variable Y_i indicates whether the individual i experienced the event of interest in the last time period in which he or she was observed. The value of Y_i is 0, if individual i was not censored in time period j_i, and 1, if he or she was.

Duration. The time interval is the length of the individual was observed. The time for subject to developing the event of interest can only be a positive value.

The predictors. The covariates P for individual i are recorded in each time period j up to, and including, time period j_i. The explanatory variables can be continuous format or categorical format.

The discrete-time model needs “Person-Period” instead of “Person-Level” data, thus corresponding data restructure is needed. The notation for the discrete-time model is similar to that for continuous-time survival model. It is assumed that time can take on only positive values $(t = 1, 2, 3, ...)$ and observe a total of n independent individuals $(i = 1, 2, ..., n)$ beginning at some natural starting point $t = 1$. The observation continues until time t_i, at which point either an event occurs or the observation is censored. Censoring here is right-censoring, which means the individual is observed at t_i but not at t_{i+1}. Normally, the time of censoring is independent of the hazard rate for the occurrence of events. A period represented a “year” in both data sets.

Basically, the following items are needed for the discrete-time model:

The time indicators. The set of dummy variables, D_{1ij}, D_{2ij}, ... , D_{Jij} identify the particular time periods to which the record refers. If the individual had the events on the time period j, all of the time indicators take on value 0 except for the j th dummy, D_{jij}, which has value 1.

The predictors. In the jth record, the covariates contain the ith individual's values of the P covariates appropriate for time period j, Z_{1ij}, Z_{2ij}, ... ,Z_{Pij}. Time-invariant predictors have values that are identical in all time periods between 1st period and jth period. Time-varying predictors, on the other hand, have values that may differ from time period to time period.

Censoring. The variable Y records the value y_{ij} that indicates whether the event of interest occurred for individual i in time period j. Its value is 0, if the event of interested did not occur, and 1, if it did.

Sample Data and Hypothesis Testing

Description of medical research data

The medical research data came from in-house prostate cancer database in a large Midwestern county hospital. A total of 1577 intermediate- or high-risk prostate cancer patients with clinical tumor stage T1-T3 N0 M0 who were treated with conventional dose EBRT, high-dose adaptive radiation therapy, EBRT+high-dose-rate brachytherapy boost, or brachytherapy alone between 1984 and 2005 were included. All the patients had minimum 5-year follow-up. Biochemical failure was defined as a rise in the blood level of prostate-specific antigen (PSA) in prostate cancer patients after treatment with surgery or radiation (Roach et al., 2006). After radiation therapy, PSA levels usually fall below 0.3 ng/mL or undetectable levels. In 2005, the American Society for Therapeutic Radiology and Oncology (ASTRO) revised a definition of biochemical failure in Phoenix, Arizona. A rise by 2 ng/mL or more above the nadir PSA is considered the standard definition for biochemical failure after EBRT with or without hormone treatment. Cancer recurrence is a return of the cancer after a period of time in which no cancer could be detected. The odds of a cancer recurring depend on many factors, including type of cancer, its extent within the body at the time of treatment, type of treatment received, and many baseline patient's characteristics.

We will test three hypotheses concerning the timing of biochemical failure, and test how some prognostic variables impact on biochemical failure. The hypotheses are listed below:

H1 Patients with higher risk factor including higher Gleason score, higher pre-RT PSA, and higher clinical tumor stage have a higher hazard of biochemical failure compared to the patients with lower risk factors.

H2 Patients with longer nadir time have a lower hazard of biochemical failure compared to the patients with shorter nadir time.

H3 Patients with lower risk category but have a longer nadir time have a better biochemical control compared to patients with high-risk and with a shorter nadir time.

With H1, the hazards of biochemical failure occurrence between patients with different levels of risk factors were compared. With H2, the hazards of biochemical failure occurrence between patients between different nadir time groups were compared. The patients with shoter time to reach their lowest PSA value were more likely to have biochemical failure occurred earlier. Under H3, the hazards with different levels of NCCN risk group were compared. And the change of hazards of biochemical failure cross-level of risk groups and nadir time groups were also investigated. The patients with lower risk category and longer nadir time will decrease the hazard to develop the biochemical failure. Patients with lower risk factor may be more likely to have good outcome control, and their nadir time may be more likely to be longer than the nadir time for patients with worse risk factors (Vicini et al., 2011).

To construct a model to test these hypotheses, we use the following variables with subscripts denoting the tth calendar year, and the j th patient in the k th risk group.

Y_{tjk} = a dichotomous indicator of whether patient j in risk group k initiates biochemical failure during year t. This is the outcome variable.

p_{tjk} = the hazard of initiating biochemical failure by patient j in risk group k during year t.

$Risk_{tjk}$ = a categorical indicator of the patient j had been classified into one of NCCN risk groups based on the pre-radiation treatment prognostic factors. This is a time-invariant individual level covariate.

$Nadir\ Time_{tjk}$ = a continuous indicator of when the patient j had their lowest PSA value. This is a time-variant individual level covariate.

$Gleason_{tjk}$ = a continuous indicator of the patient j had pre-radiation treatment Gleason score. This is a time-invariant individual level covariate.

$pre-RT\ PSA_{tjk}$ = a continuous indicator of the patient j had pre-radiation treatment PSA value. This is a time-invariant individual level covariate.

$T\ Stage_{tjk}$ = a continuous indicator of the patient j had pre-RT clinical tumor stage. This is a time-invariant individual level covariate.

Analysis Strategies

Survival Function

The analysis begins with an examination of the survival function. The survival function is a plot of the probabilities that an individual will remain in the risk set as a function of time. The risk set contains only cases that are qualified to experience the event in question. The survival function may be expressed as:

$$S(t) = \frac{\text{Numbers of survivors to time } t}{\text{Total number in sample}}$$

However, this study used a modification of this formula known as Kaplan-Meier method for estimating survival probabilities since the Kaplan-Meier method accommodates "censored" individuals. According to Slonim-Nevo and Clark (1989),

> …the Kaplan-Meier approach uses ordered observations rather than grouped data. This approach has the advantage of yielding results that do not depend upon the length of time interval used for grouping, and is especially useful for small sample sizes (p.9).

The Kaplan-Meier estimate is also known as the product limit estimator and can provide a nonparametric estimate of survival outcome of interests. It is calculated as:

$$\mathrm{S_{PL}(t) = \{[r(t_1) - d(t_1)] / r(t_1)\} \times \{[r(t_2) - d(t_2)] / r(t_2)\} \times ... \times \{[r(t_k) - d(t_k)] / r(t_k)\}}$$

Where r is the risk set at time period t and d is the number of individual had events at time t.

Hazard function

The hazard function can help a researcher identifying the high-risk period. Compared to the survival function, the hazard function is more sensitive since it can detect the slope of the survival function. The hazard function describes the probabilities of an event occurring during a particular time interval and provides the subject is at risk of experiencing the event. The higher the hazard is, the higher the risk that the event will occur.

For the prostate cancer data, the hazard refers to the probability that a patient will have biochemical failure during a time interval after he finished the radiation treatment, given that the patient is at risk of having biochemical failure at the beginning of that time interval. Each separate hazard probability is computed only on that time period's risk set. The hazard probability for a patient at time period t is defined as:

$$h(t) = \frac{f(t)}{S(t)}$$

or,

$$h(t) = \frac{f(t)}{1 - F(t)}$$

Where $f(t)$ is the probability density function at time period t, $S(t)$ is the probability of surviving to the time period t without experiencing the events, and $F(t)$ is the cumulative distribution function for T . The hazard function mathematically records changes in the slope of the survival function, thereby allowing researchers to identify high-risk periods.

In discrete-time survival model, the set of the discrete-time hazard probabilities parameters h_j is a function of time period j , which is called as discrete-time hazard function. The function can be plotted whose x-axis is time and y-axis is the population risk of the event occurring in each time period under the condition where the events haven't occurred in any earlier time period.

Statistical Model for Hazards

Cox's Proportional Hazard Model

In 1972, David Cox, a British statistician proposed his model in his published paper entitled "Regression Analysis and Life Tables", and he expressed the hazard rate depends on the predictors and the time period by using a logistic regression model.

$$\log h_i(t) = \log h_0(t) + (\beta_1 x_{i1} + \ldots + \beta_k x_{ik})$$

After we input the predictor variables along with baseline model, total eight models are presented here:

Model A: $\log h(t) = \log h_0(t) + \beta_1 X_1$ (T Stage)

Model B: $\log h(t) = \log h_0(t) + \beta_1 X_1$ (Gleason score)

Model C: $\log h(t) = \log h_0(t) + \beta_1 X_1$ (pre-treatment PSA)

Model D: $\log h(t) = \log h_0(t) + \beta_1 X_1$ (nadir time)

Model E: $\log h(t) = \log h_0(t) + \beta_1 X_1$ (pre-treatment hormone thearpy)

Model F: $\log h(t) = \beta_0(t) + \beta_1 X_1$ (NCCN risk category)

Model G: $\log h(t) = \beta_0(t) + \beta_1 X_1$ (NCCN risk category)+$\beta_2 X_2$ (nadir time)

Model H: $\log h(t) = \beta_0(t) + \beta_1 X_1$ (NCCN risk category)+$\beta_2 X_2$ (nadir time) +$\beta_3 X_3$ (pre-treatment hormone therapy)

Discrete-Time Survival Analysis Model

The proportional hazard model uses duration data and can handle the censoring problem effectively. In hazard modeling, the time of a patient's occurrence an event becomes a part of the dependent variable. Hazard modeling provides decision-makers with additional information which includes 1) characteristics of high-risk patients, 2) high-risk time periods over the course of a patient post-treatment follow-up, 3) the probability of a patient surviving to any given time period (year), 4) the conditional probability of event occurring during any given time period post-treatment. The information above is very important for follow-up visit arrangement and implementation of follow-up care after initial cancer treatment. Both health caregivers and patient's family can work closely to deal with the possibility of cancer recurrence.

The entire hazard function can be modeled as a function of selected predictors. In discrete-time survival analysis, as in linear regression, the initial model (or baseline) contains only the intercept with no predictor variables. The baseline model fits the data with the models with unstructured hazard functions and no covariates. The baseline equation with no predictors (or baseline logit-hazard profile) is:

logit $h(t) = (\alpha_1 t_1 + \alpha_2 t_2 + ... + \alpha_k t_k)$

The alpha parameters (α) are multiple intercepts which have one in each time period. They represent the "baseline logit-hazard function because they capture the time-period by time-period conditional log-odds that individuals whose covariate values are all zero will experience the event in each time period, given that they have not already done so" (Judith D. Singer & Willett, 1993, p. 167).

The baseline equation will be expanded to include predictor variables, as in ordinary least squares regression. The relationship of the logit-transformed hazard profile to a predictor variable, X_1, is

logit $h(t) = (\alpha_1 t_1 + \alpha_2 t_2 + ... + \alpha_k t_k) + \beta_1 \mathrm{X}_1$

Where the β_1 measures the amount of vertical shift in logit-hazard per unit difference in the predictor variable. Using standard statistical packages, the β coefficients and their standard errors can be estimated and inferences can be made with respect to the effects of predictors on survival.

According to Willet and Singer (1991), the baseline logit-hazard model can be written as:

logit $h(t) = \beta_0(t)$

Where $\beta_0(t) = (\alpha_1 t_1 + \alpha_2 t_2 + ... + \alpha_k t_k)$ (p.417)

Using this formulation, eight hazard models, A, B, C, D, E, F, G, and H were constructed to model the hazard associated with prostate cancer patients' recurrence during the first ten years post-treatment. These models are the following:

Model A: logit $h(t) = \beta_0(t) + \beta_1 X_1$(Clinical Tumor Stage)

Model B: logit $h(t) = \beta_0(t) + \beta_1 X_1$(Gleason Score)

Model C: $\text{logit } h(t) = \beta_0(t) + \beta_1 X_1 (\text{pre-treatment PSA})$

Model D: $\text{logit } h(t) = \beta_0(t) + \beta_1 X_1 (\text{nadir time} \geq \text{ 2 year})$

Model E: $\text{logit } h(t) = \beta_0(t) + \beta_1 X_1 (\text{pre-treatment hormone thearpy})$

Model F: $\text{logit } h(t) = \beta_0(t) + \beta_1 X_1 (\text{risk group})$

Model G: $\text{logit } h(t) = \beta_0(t) + \beta_1 X_1 (\text{risk group}) + \beta_2 X_2 (\text{nadir time } \geq \text{ 2 year})$

Model H: $$\begin{aligned} &\text{logit } h(t) = \beta_0(t) + \beta_1 X_1 (\text{risk group}) + \beta_2 X_2 (\text{nadir time } \geq \text{ 2 year}) \\ &+ \beta_3 X_3 (\text{pre-treatment hormone therapy}) \end{aligned}$$

The variable clinical tumor stage (T1, T2, T3), Gleason score (<6, 7, 8-10), pre-tx PSA (<10ng/mL and ≥10 ng/mL), nadir time (< 2 years and ≥ 2 years) NCCN risk category (intermediate- and high-risk) are all categorical variables in the above models. Categories were coded using the digits 0 and 1 for binary variable, and 1, 2, and 3 for variables with multiple values. The disease risk was classified according to National Comprehensive Network (NCCN) criteria, with high risk being T3 or more, initial prostate-specific antigen (iPSA) ≥20 ng/mL, or Gleason score 8 to 10; low risk being T2a or less, iPSA <10 ng/Ml, and Gleason score ≤ 6; and intermediate risk being all the remainders. NCCN risk category definitions can be found in the supplement table (Table 8

).

Model A was used to examine the relationship between hazard to occurring and the time indicator. The model serves as a baseline for determining whether other variables have an effect on the event (cancer recurrence). The model A, B, C, D, and E tested the main effects of tumor stage, Gleason score, pre-tx PSA, and nadir time. Model F consider NCCN risk as a whole instead of investigating each individual tumor stage, Gleason score, and pre-treatment PSA. Model G tested the effect of two variables – NCCN risk group and nadir time group. Model H tested the total effect after considering all three variables together (NCCN risk, nadir time, and if the patient received pre-treatment HT).

Likelihood Ratio Chi-Square Statistics

The likelihood ratio chi-squares were compared by computing a $G^2(\text{df}=1)$ value with the following equation (Wainer, 1990):

$$G^2(1) = G^2{}_2 - G^1{}_2$$

For either the Cox regression model or the discrete-time survival model, the $G^1{}_2$ is the -2(loglikelihood) of model 1, and $G^2{}_2$ is the -2(loglikelihood) of model 2 with additional variables. The statistical difference between two models' chi-squares can also be assessed like likelihood ratio chi-square statistics from the formula above.

Testing Assumptions

Cox's Proportional Hazard Model

Cox's proportional hazard model requires that the hazard ratio is constant over time. For any two individuals at any point in time, the ratio of their hazards is a constant. Two situations of proportional hazards mean here: 1) the hazards of two individuals are constant over time which makes the ratio of the two hazards be constant, 2) the hazards of two individuals varies over time but the rates of the changes in two hazards are the same. Basically, for any time$t$, the ratio of $h_i(t)/h_j(t)=c$, where i and j refer to distinct individuals and c may depend on explanatory variables but not on time (Allison, 1984). Violation of the proportional hazard assumption can occur in many ways. The first violation may involve the inclusion of time-varying variables in the equation, whereby the hazards are no longer proportional, but may become nonproportional. Or if there is an interaction between time and one or more of predictor variables, the proportional hazard assumption is also violated. Care must be taken to check this assumption. Violation of this proportionality assumption can be checked both graphically and statistically. By stratifying the sample according to the categories of a variable, assuming that the influence of other covariates are identical for all categories, and transforming the survivor function, the plotted curves should differ only by a constant factor, β. If there is a change in the distance between two curves, the proportionality assumption may be violated. A statistical test for proportionality would demonstrate that the coefficient β would not be significantly different from zero and the hazard functions of the two categories of the variable should differ only by the constant factor $\exp(\beta)$.

Discrete-time Hazard Model

There are three important assumptions which need to be tested in the discrete-time hazard model (Judith D. Singer & Willett, 1993). The first one is linearity assumption which requires that the vertical displacements in logit hazard are linear per unit of difference in each predictor. Exploratory analysis and statistical reference can be used to check this assumption. The graphical method is easy to implement by visual checking if hazard functions in logit-hazard form of difference stratums have approximately equal vertical displacements. If the displacements are roughly equal, then the linearity assumption is met. Otherwise, the assumption is violated. Generally, the violation of linearity assumption can be resolved by transformation of the predictors or converting the continuous variable into a set of dummy variables.

The second assumption in discrete-time method is no unobserved heterogeneity. The individuals are hypothesized to be different only in their predictors, and all the variations in the hazard profiles across individuals only depend on observed variation in the predictors. Vaupel and Yashin (1985) brought the term of “heterogeneity’s ruses”, and illustrated that the mixing heterogeneous population with different risk profiles can yield a pooled profile that may have a shape entirely different from the component profiles. Selection is the root cause for this. Because of selection without knowing unobserved heterogeneity, the shape of hazard profiles could be hard to explain.

The third assumption is proportional assumption. Both continuous-time and discrete-time survival models involve a proportionality assumption. A simple graphical method can be used for verifying the proportionality assumption. In the preliminary analysis, “if logit-hazard profiles estimated separately within strata are all approximately parallel, then the assumption is met; if they are not, it is violated.”(Judith D. Singer & Willett, 1993, p. 186)

CHAPTER 4 RESULTS

This study was undertaken to examine the effect of certain data characteristics on survival analysis hazard estimation and goodness of fit statistics between Cox regression and discrete-time survival models. The following three conditions were varied to assess the impact on the hazard estimates and goodness of fit statistics: (a) the number of time periods for which the data were coded, (b) the sample size, (c) the number of parameters for which the statistics model was used.

Data Sets

Two levels of time periods and four sample sizes of cases which were generated from the original medical research data, were compared among Model A through Model H (Table 9). The data were coded to reflect the division of time into either ten or five periods. Ten time periods (one year per time period) were chosen to emulate typical time periods found in the medical literature, and follow-up visits were recommended to occur every 6 months for the first 3 years and at least yearly starting the 4th year after patients finished their radiation treatment for prostate cancer. Five time periods (two years per time period) were chosen to compare how it impacts on the hazard estimation comparing to ten time periods.

Singer (1991) described that the statistical power of the discrete-time survival analysis model is affected by sample size. The sample size sets were chosen by using one hundred percent, seventy-five percent, fifty percent, and twenty-five percent of empirical data sets. The four sample size sets were randomly chosen from the original data set. The conditions for each data set are presented in Table 3.

Table 3

Conditions for Simulated Date Sets

Data Sets	Number of Time Periods	Model	Sample Size in Cox Model	Sample Size in Discrete-time Survival Model
Data set 1	10	Model F	1577	9692
Data set 2	10	Model G	1577	9692
Data set 3	10	Model H	1577	9692
Data set 4	5	Model F	1577	6041
Data set 5	5	Model G	1577	6041
Data set 6	5	Model H	1577	6041
Data set 7	10	Model F	1213	7516
Data set 8	10	Model G	1213	7516
Data set 9	10	Model H	1213	7516
Data set 10	5	Model F	1213	4688
Data set 11	5	Model G	1213	4688
Data set 12	5	Model H	1213	4688
Data set 13	10	Model F	809	4939
Data set 14	10	Model G	809	4939
Data set 15	10	Model H	809	4939
Data set 16	5	Model F	809	3097
Data set 17	5	Model G	809	3097
Data set 18	5	Model H	809	3097
Data set 19	10	Model F	422	2490
Data set 20	10	Model G	422	2490
Data set 21	10	Model H	422	2490
Data set 22	5	Model F	422	1576
Data set 23	5	Model G	422	1576
Data set 24	5	Model H	422	1576

The purpose for the empirical example was to examine the biochemical failure hazard from both the Cox regression model and the discrete-time survival model. It was previously established that biochemical failure is related to patient prognostic factors (Vicini et al., 2011). For example, patients with higher pre-treatment PSA, higher Gleason score, higher clinical tumor stage, longer time to reach the nadir PSA, and the lack of hormone therapy before radiation would likely have biochemical failure earlier than patients with lower risk factors, shorter nadir time, and treatment with hormone therapy.

For the Cox regression model or discrete-time survival Model, the Model A through Model F were individually examined on how each single factor impacted the outcome, i.e., biochemical failure (BF). The hazard coefficients estimates were analyzed and compared by using Cox regression model and discrete-time survival model cross the Model G and Model H. Twenty-four data sets were analyzed and compared coefficients for Model F, Model G, and Model H.

In the life tables (Table 10, 11, and 12), the incidence rates for BF were different among the three NCCN risk groups. Table 10 lists the hazard rates for low-risk patients. Table 11 lists the hazard rates and survival proportions of all intermediate-risk patients who had not experienced BF by the end of each year. Patients in the intermediate-risk group had their peak hazard rates between 4th – 5th years after finishing radiation treatment, with almost no biochemical failures after 12 years after radiation treatment. The high-risk group had much earlier BF starting from the 1st year after radiation treatment (Table 12) and with its peak BF of 13% rate occurring around the 5th year after radiation. The intermediate-risk and high-risk groups were more important than low-risk group for the clinicians to investigate how well intermediate- and high-risk patients response the treatment, and if, there is a way to tell the patients and clinicians how much probability they may have the cancer back based on the prognostic factors and follow-up PSA information.

The estimated hazard function and corresponding survival function are displayed in Figures 2 and 3, respectively. The figures provide the same result from life table graphically.

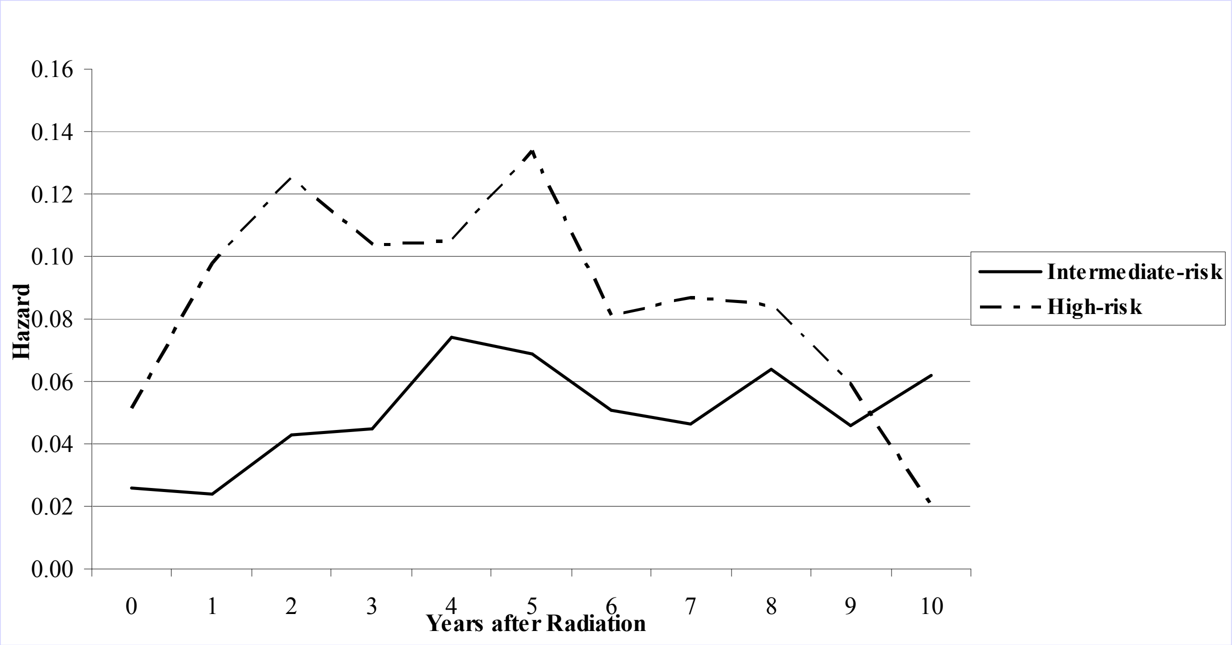


Figure 2. Hazard Functions for Biochemical Failure by NCCN Risk Groups.

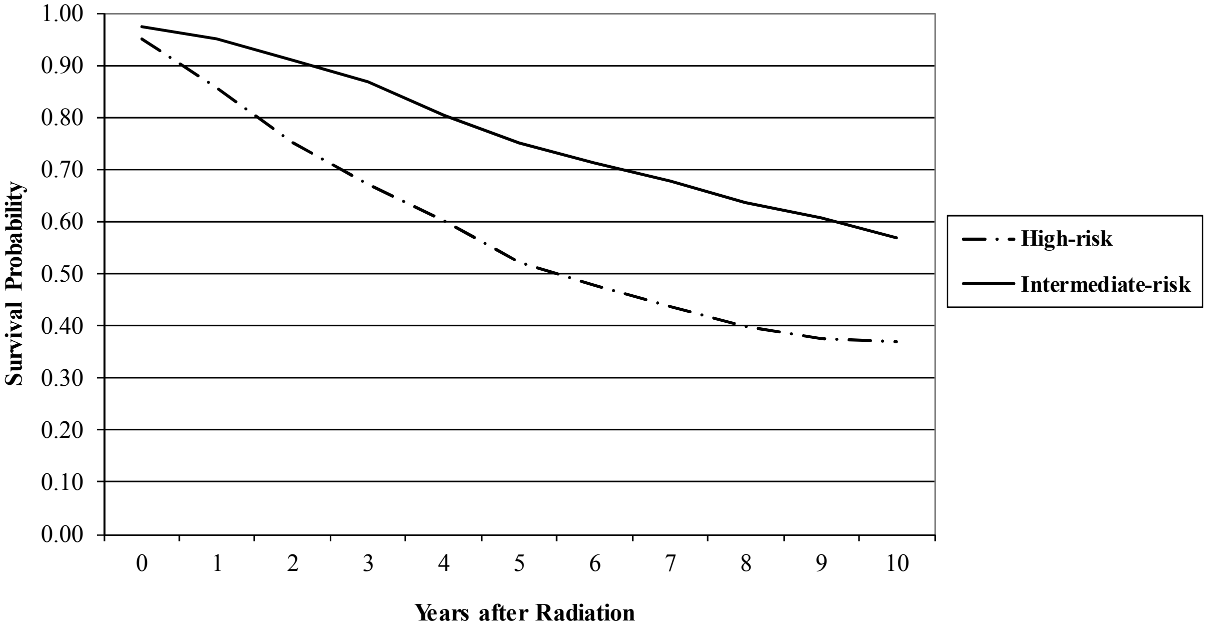


Figure 3. Survival Functions for Biochemical Control by NCCN Risk Groups.

Survival Function and Hazard Function

The survival functions indicated that the high-risk group had higher biochemical failure rate and worse biochemical control (Figure 4 and Figure 5) compared with the intermediate-risk group. The median biochemical failure times for the intermediate-risk group and high-risk group are 14.9 year and 6.4 year. The biochemical control rate was also highly associated with the time when the patients reached their nadir PSA after radiation treatment and if the patients received the hormone therapy before radiation therapy. Patients who took longer time to reach their lowest PSA value were less likely to have BF compare to the patients with a shorter time to PSA nadir (Figure 6 and Figure 7). Patients who received pre-treatment hormone therapy were less likely to have BF comparing to those who did not have hormone therapy before radiation therapy (Figure 8 and Figure 9).

The cumulative hazard rates are shown in the Figure 5, Figure 7, and Figure 9, and survival rates are shown in the Figure 4, Figure 6, and Figure 8.

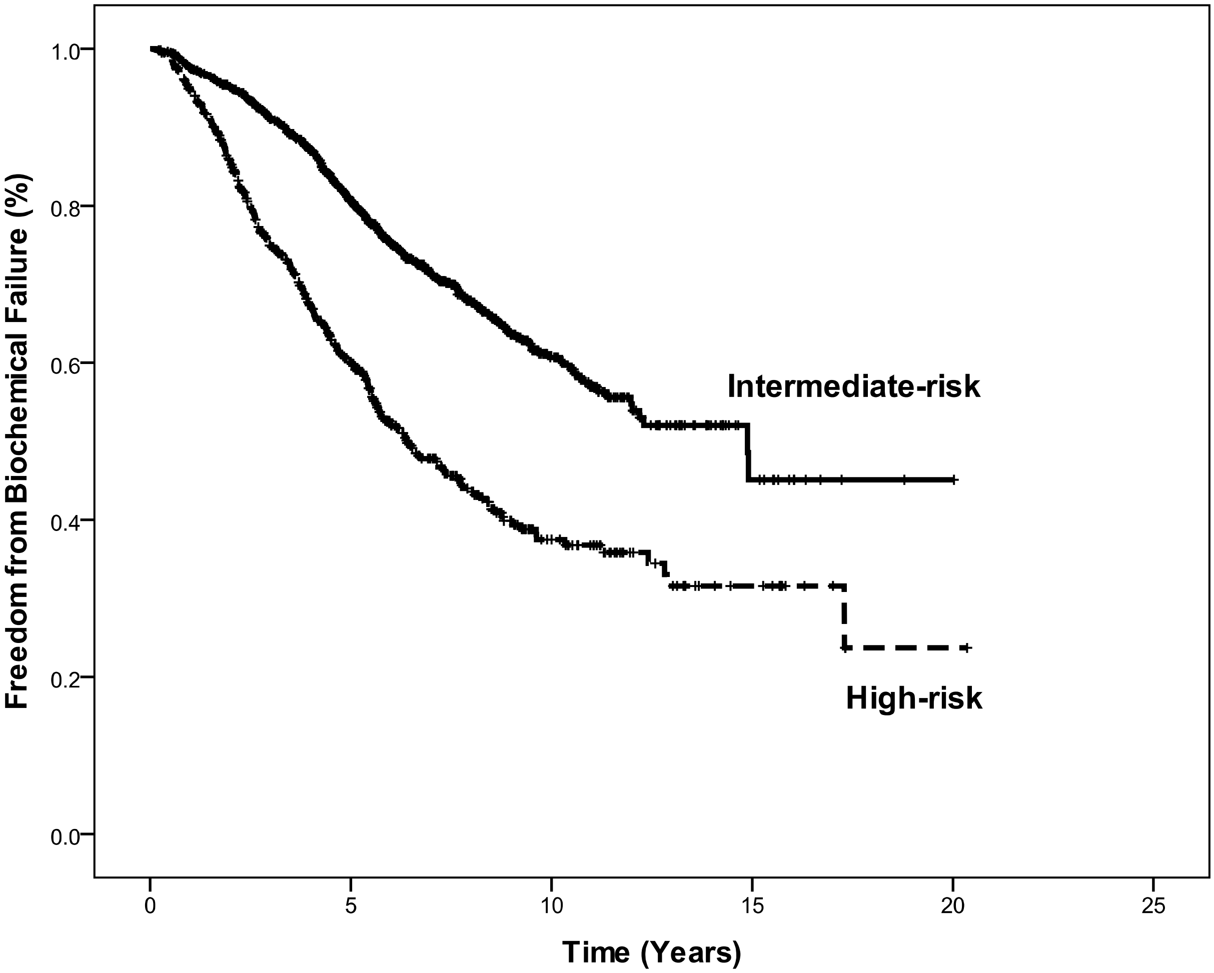


Figure 4. Freedom from Biochemical Failure by Risk Group

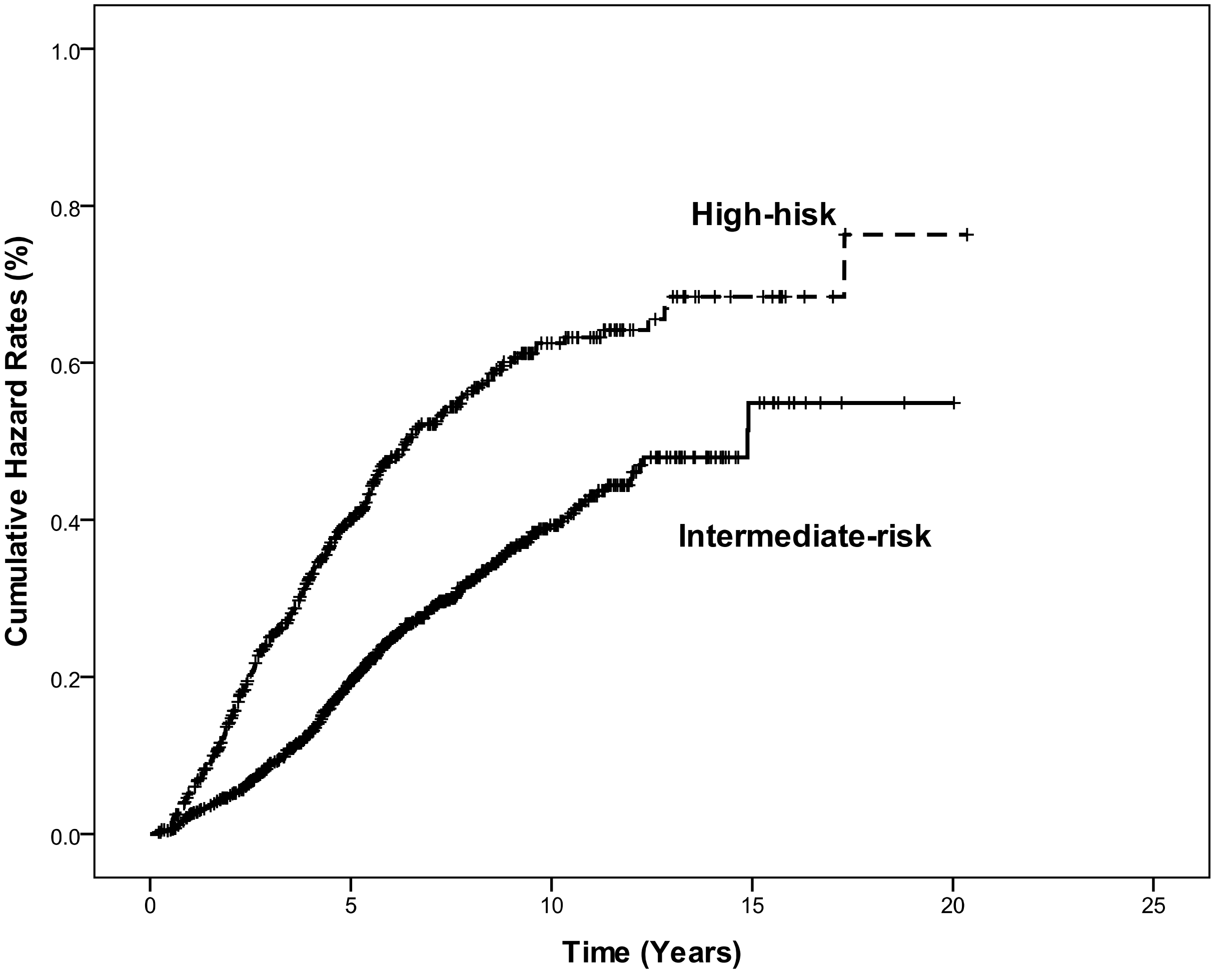


Figure 5. Cumulative Hazard Rates for Biochemical Failure by Risk Groups

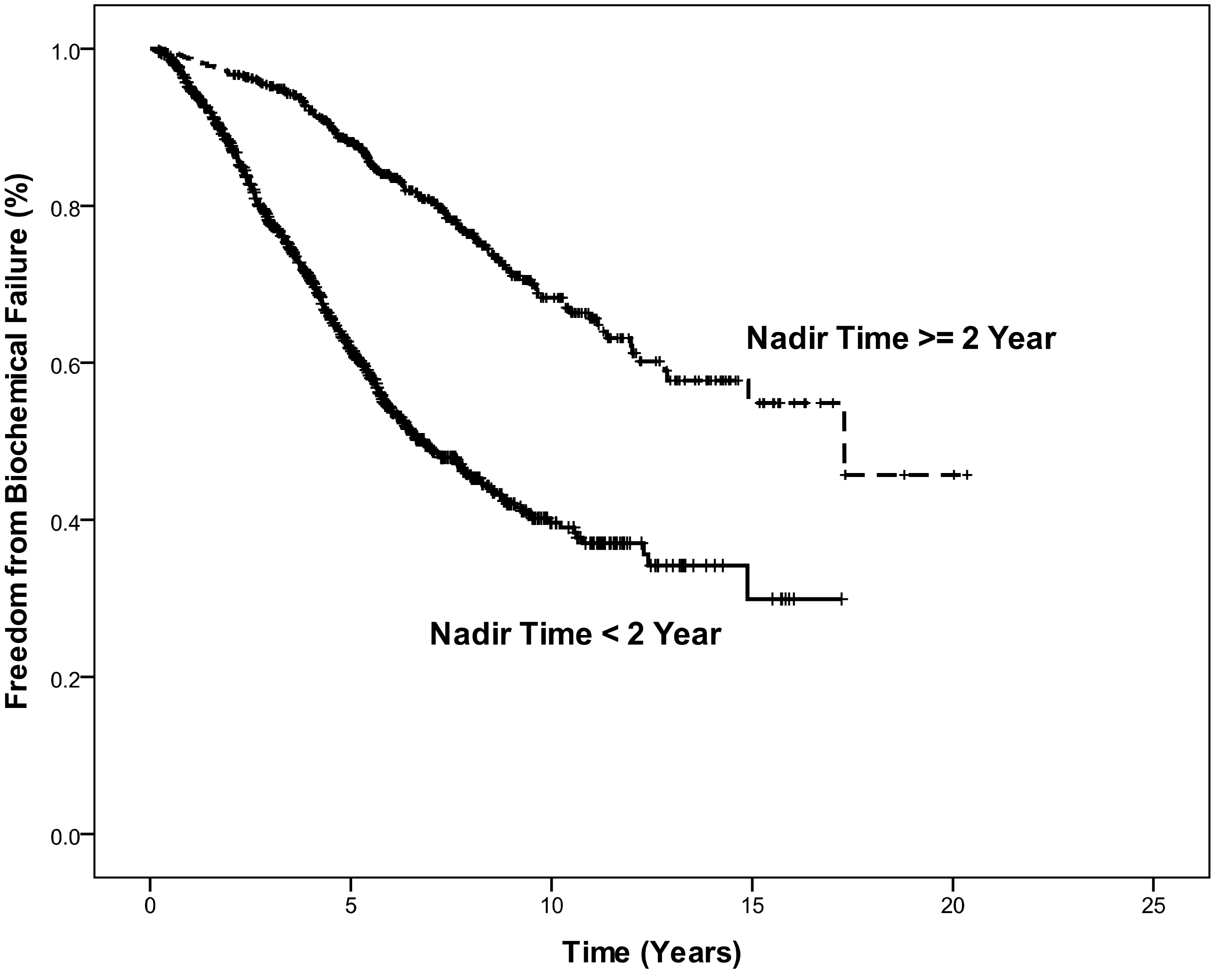


Figure 6. Biochemical Control by Nadir PSA Time Groups

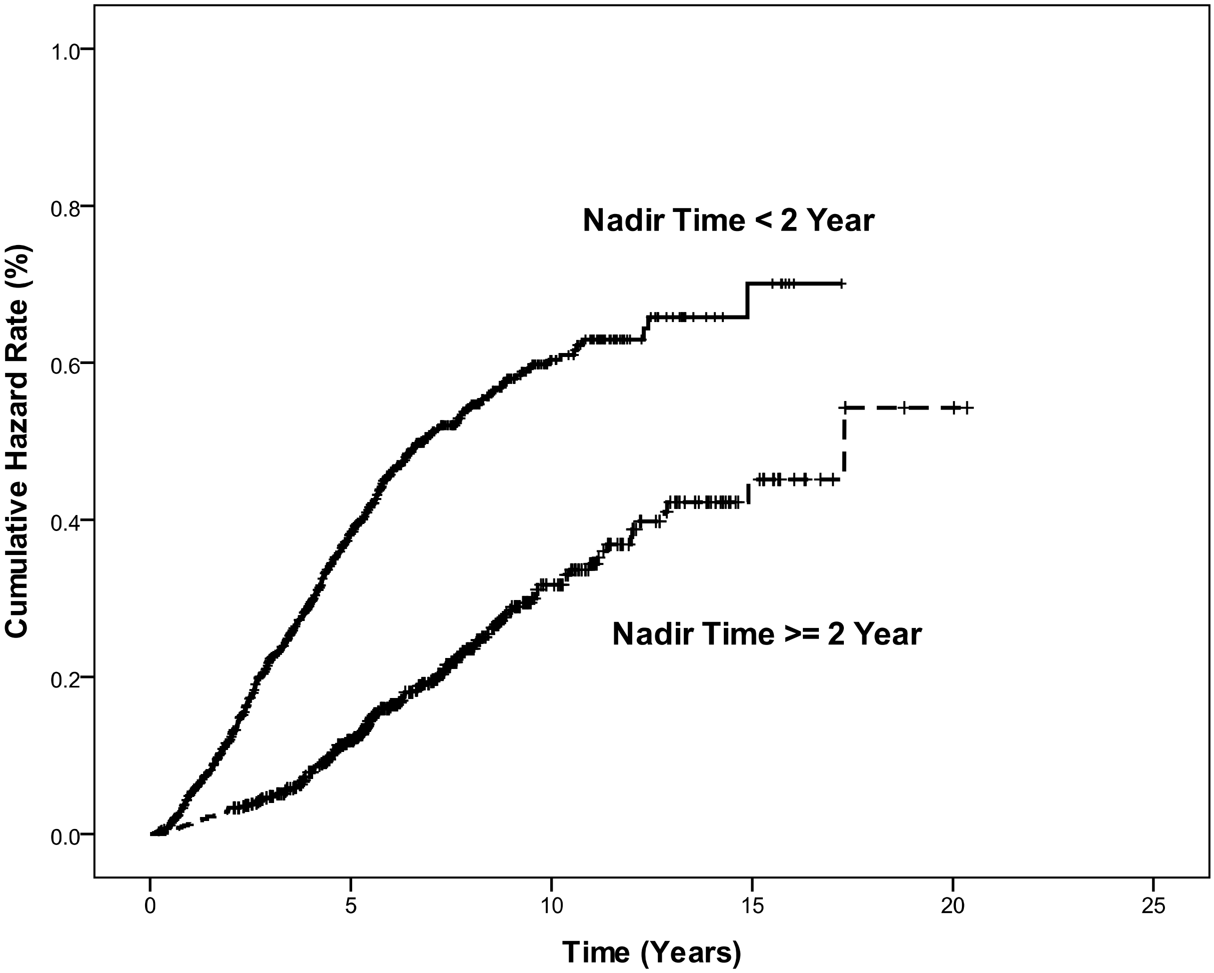


Figure 7. Cumulative Hazard Rate for Biochemical Failure by Nadir PSA Time Groups

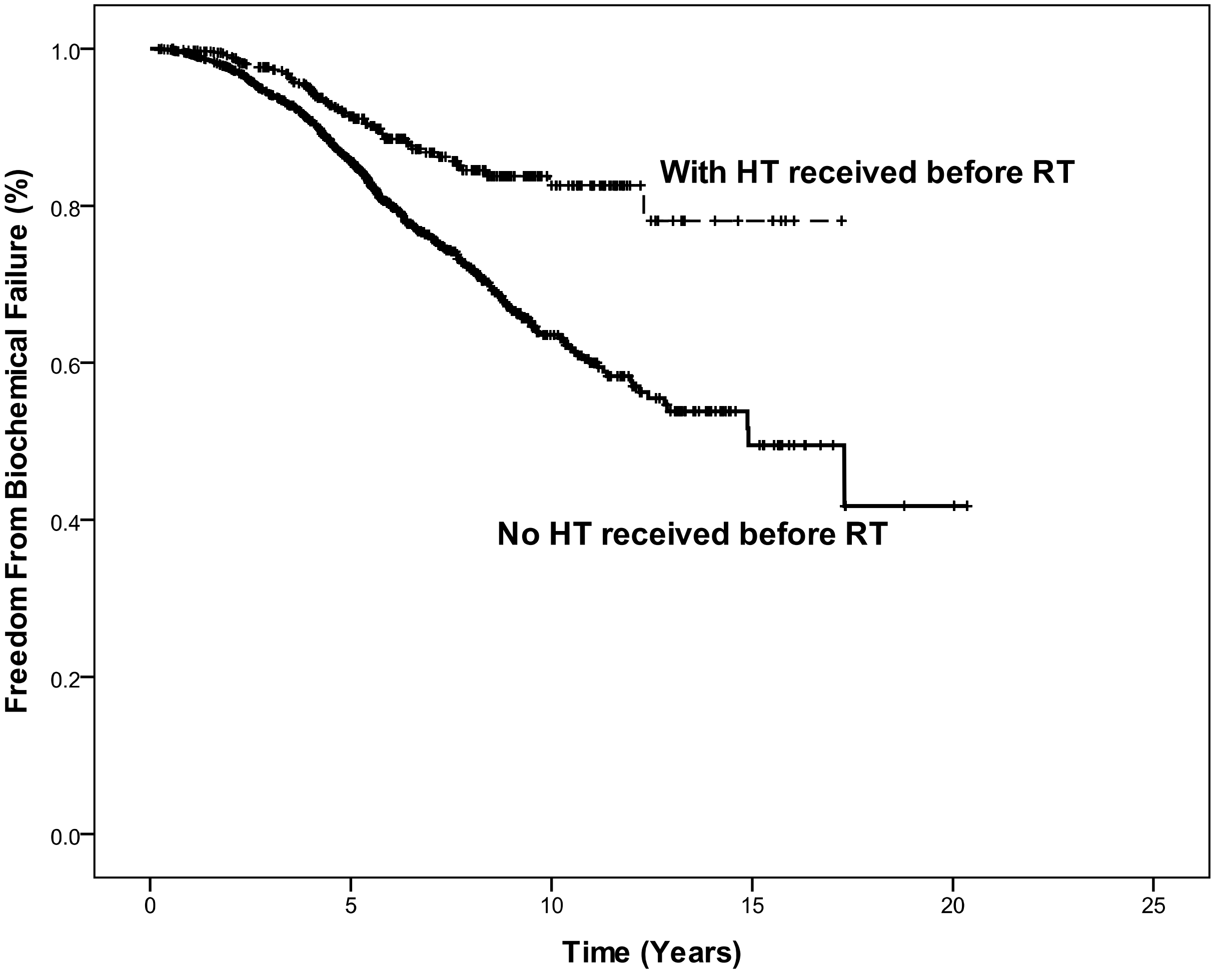


Figure 8. Biochemical Control by Groups Received HT or No HT

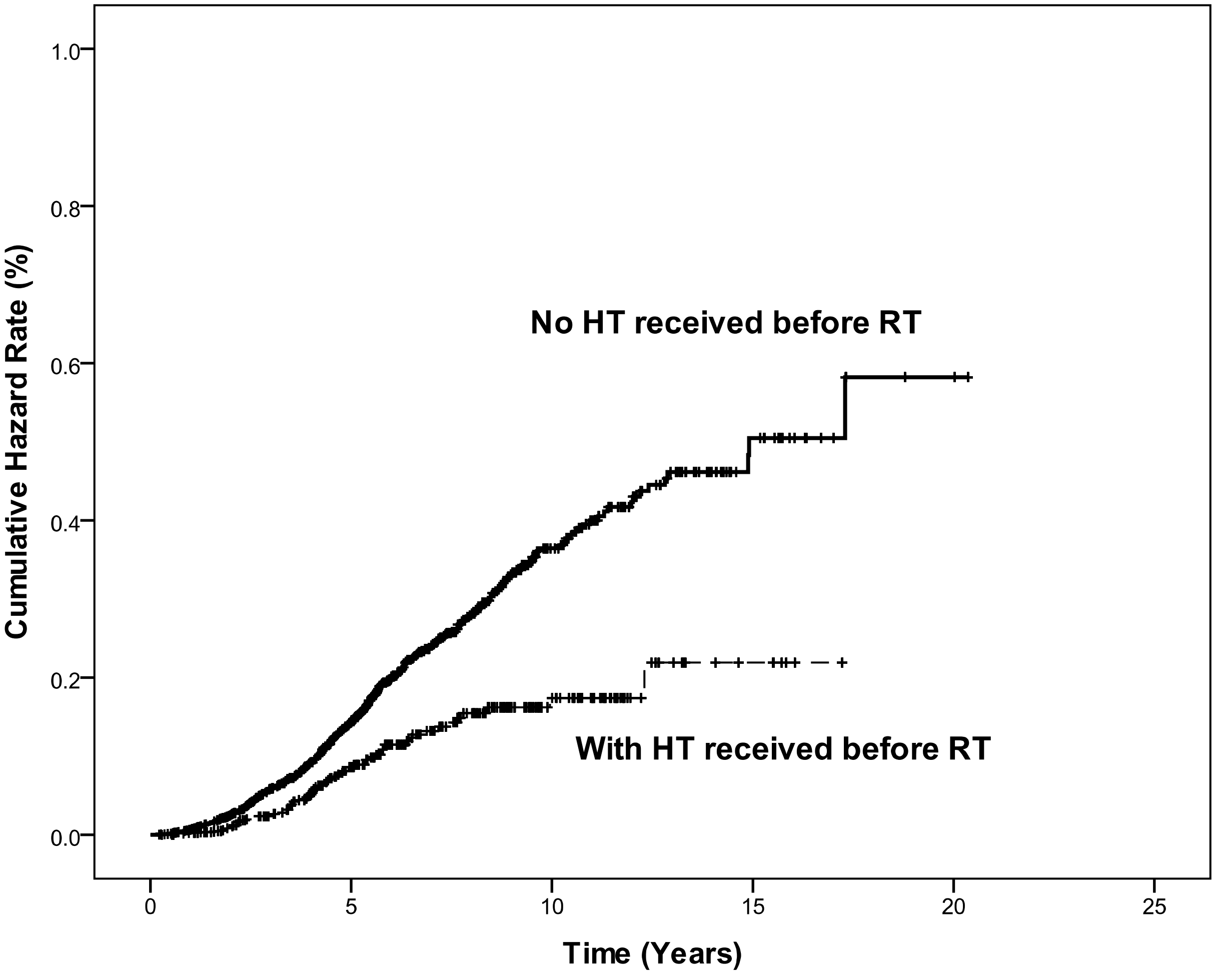


Figure 9. Cumulative Hazard Rate for Biochemical Failure by Groups Received HT or No HT

Cox Model

The univariate analysis was tested for each individual factor for different sample size n=1577, n=1213, n=809, and n=422. Models A through E were tested with one covariate by using the Cox regression model and the discrete-time survival model. The single covariate in Model A to E included clinical tumor Stage, the pre-treatment PSA, tumor's Gleason score, if the patient received the hormone therapy before radiation, and the time to reach the nadir PSA. All single covariates are strong predictors of biochemical failure in the univariate analysis ($p < 0.001$). Model F tested the NCCN risk group as the combination of clinical tumor stage, pre-treatment PSA and tumor Gleason score information. The high risk group had higher risk had more frequent and earlier biochemical failures.

The multivariate analysis was used in Model F, G and H. The likelihood ratio test was used to test the significance of the coefficients in the model. The -2 log likelihood statistics (-2LL) has a chi-square distribution under the null hypothesis that all coefficients in the model are zero. The difference in fit between two nested models is assessed by looking at the change in -2LL, with the degree of freedom equal to the difference between the numbers of parameters in the two models. For example, the group with sample size n=1577, the -2LL for the model H is 7160.8 (Table 13) smaller than in the model G with -2LL=7304.1. The decreased -2LL indicates that the model H was improved relative to model G by taking account the additional variable which is if the patient received the HT before radiation. The summary of Model A through H including -2LL estimates are presented in Table 13. Under other sample size groups, model H had the best representation among model F, G, and H. For the sample size n=1213, n=809, and n=422, the summary of goodness-of-fit for the -2LL estimates are presented under Table 14, 15, and 16.

Figure 10, Figure 11, and Figure 12 are the graphs from the Cox model for cumulative hazard, survival rate and log-hazard against time. Line 1 indicates the baseline hazard for the null model without considering any variable, the line 2 indicates that the hazard increased with the variable of risk factor in the Model F, line 3 is the hazard after considering risk and nadir time variable in the Model G. Line 4 indicates the highest hazard rate after inputting all three factors into the model H.

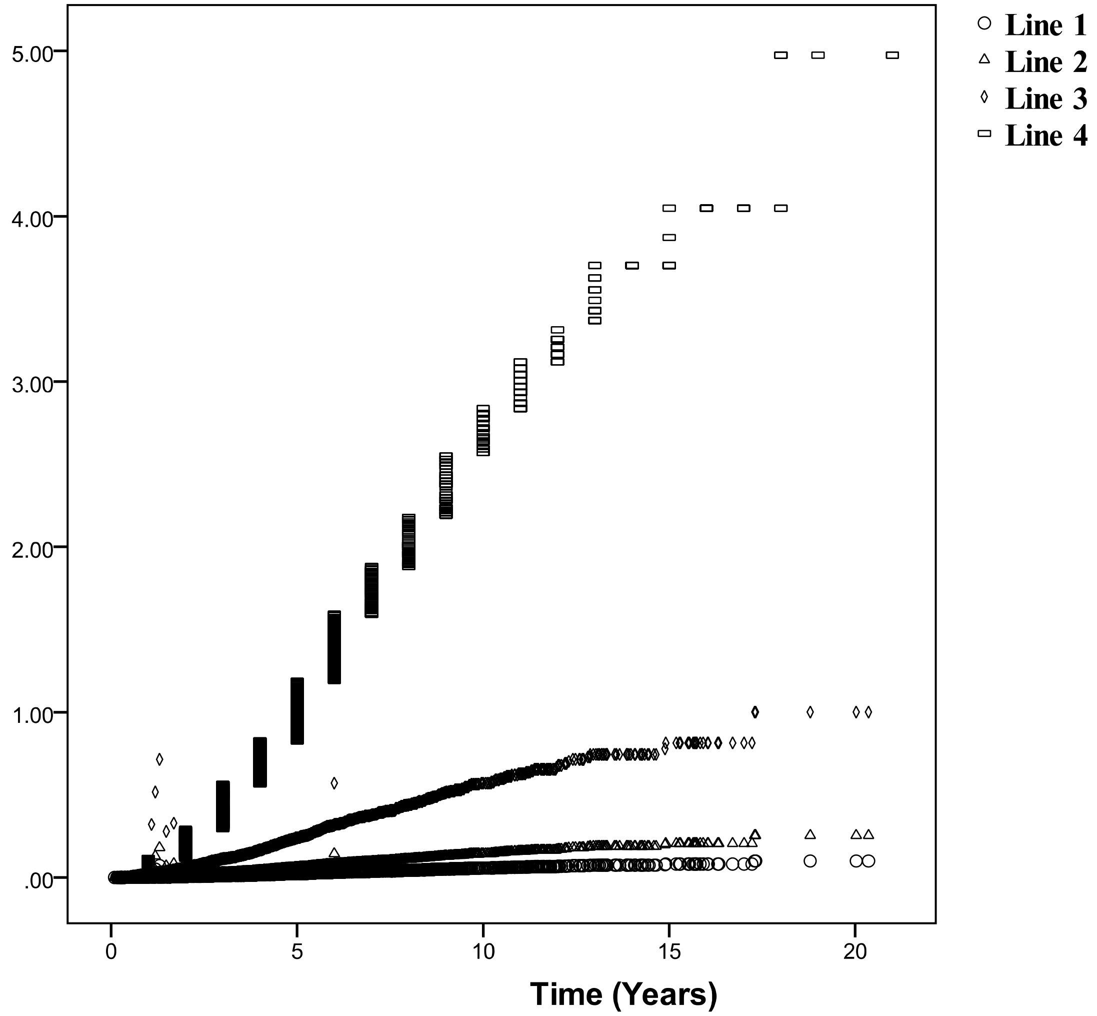


Figure 10. Cumulative Hazard Comparison from Nested Model Null/F/G/H under Sample Size n = 1577

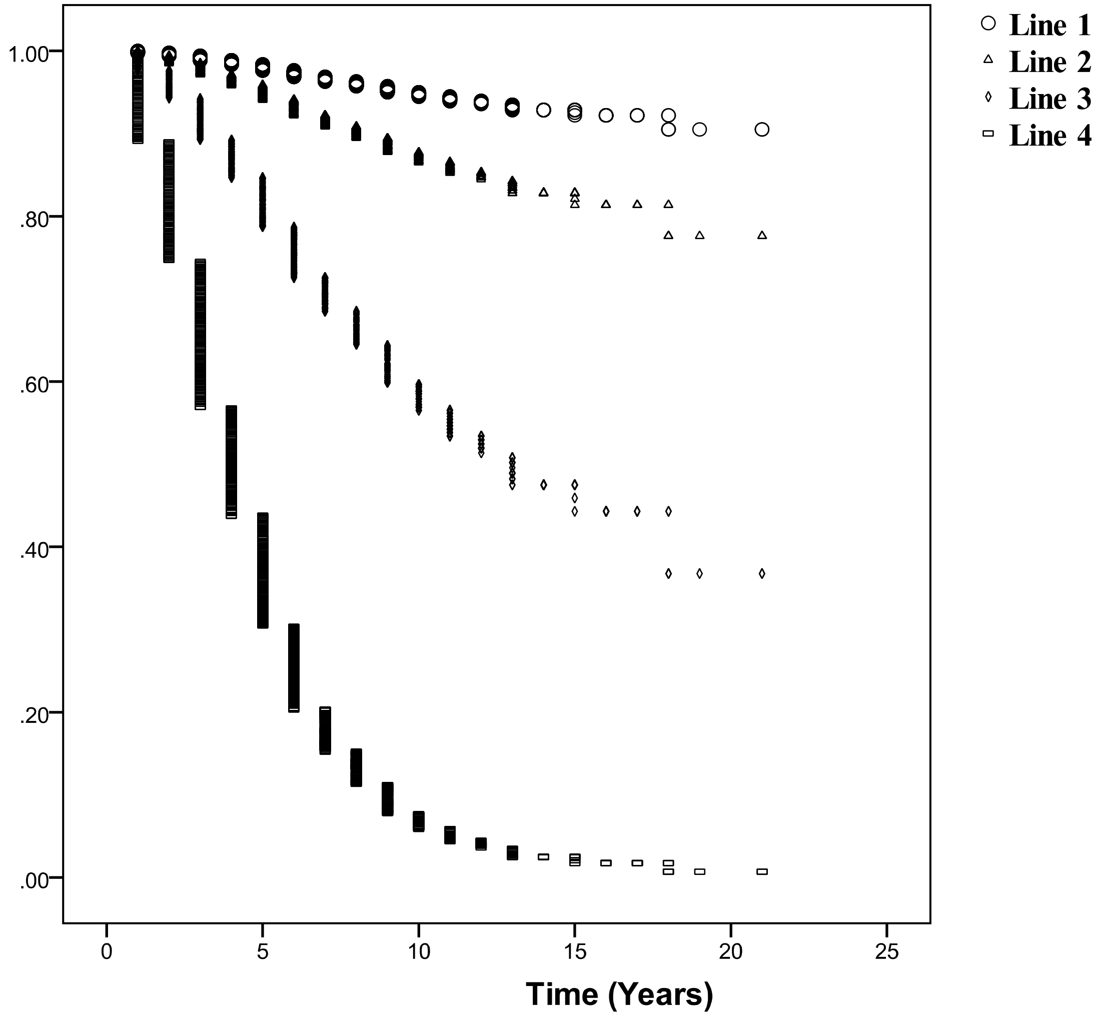


Figure 11. Survival Rate Comparison for Nested Model Null/F/G/H under Sample Size n = 1577

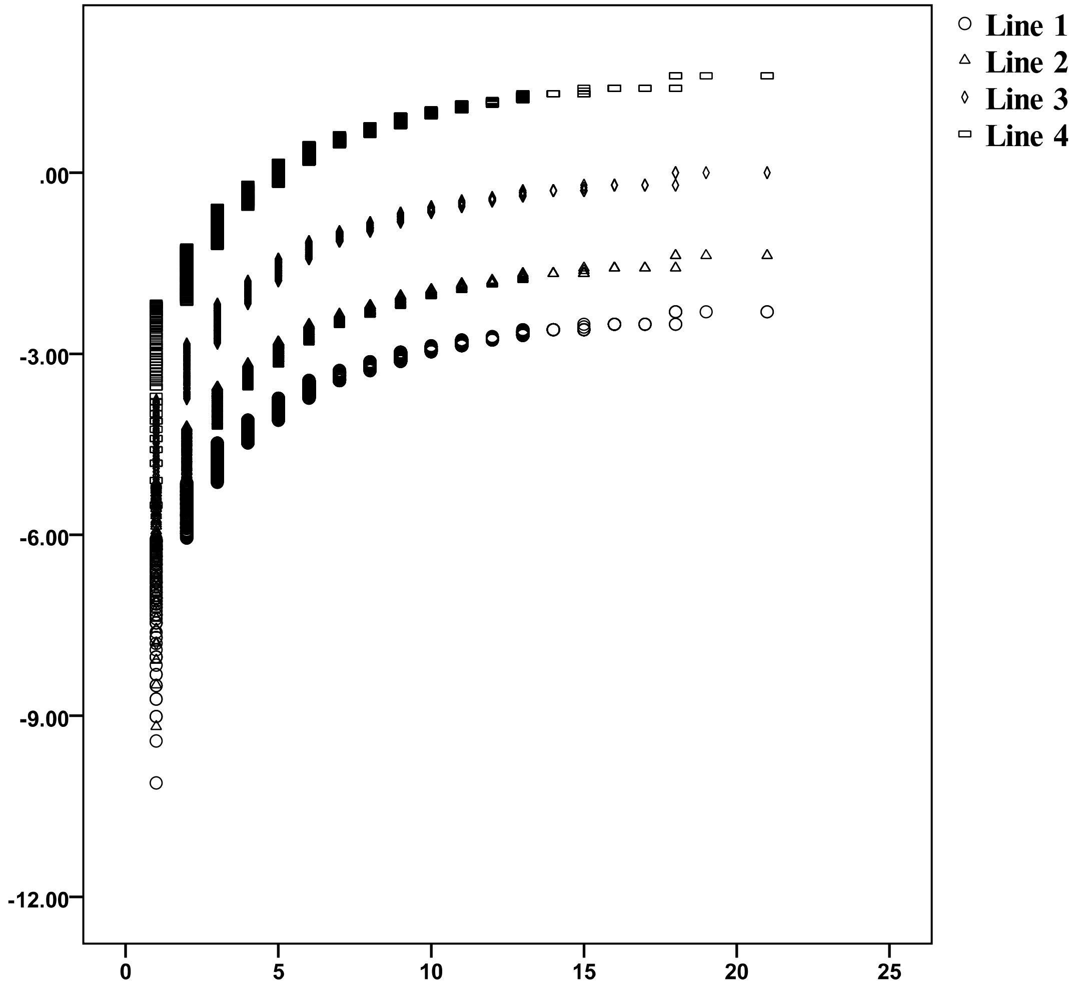


Figure 12. Log Hazard Comparison for Nested Model Null/F/G/H under Sample Size n = 1577

For the subgroup of patients who are under intermediate-/high-risk, nadir time <2 / ≥2 year, and no HT/with HT were investigated their hazard and survival probabilities. Eight groups were investigated to compare their hazards and survival functions: 1) intermediate-risk, nadir time < 2 yrs, and NO HT; 2) intermediate-risk, nadir time < 2 yrs, and with HT; 3) intermediate-risk, nadir time ≥ 2 yrs, and NO HT; 4) intermediate-risk, nadir time ≥ 2 yrs, and with HT; 5) high-risk, nadir time < 2 yrs, and NO HT; 6) high-risk, nadir time < 2 yrs, and with HT; 7) high-risk, nadir time ≥ 2 yrs, and NO HT; 8) high-risk, nadir time ≥ 2 yrs, and with HT. Indicated in the Figure 13 and Figure 14, the solid lines refer to intermediate-risk group (Group 1-4) and the dash lines (Group 6-8) refer to high-risk group.

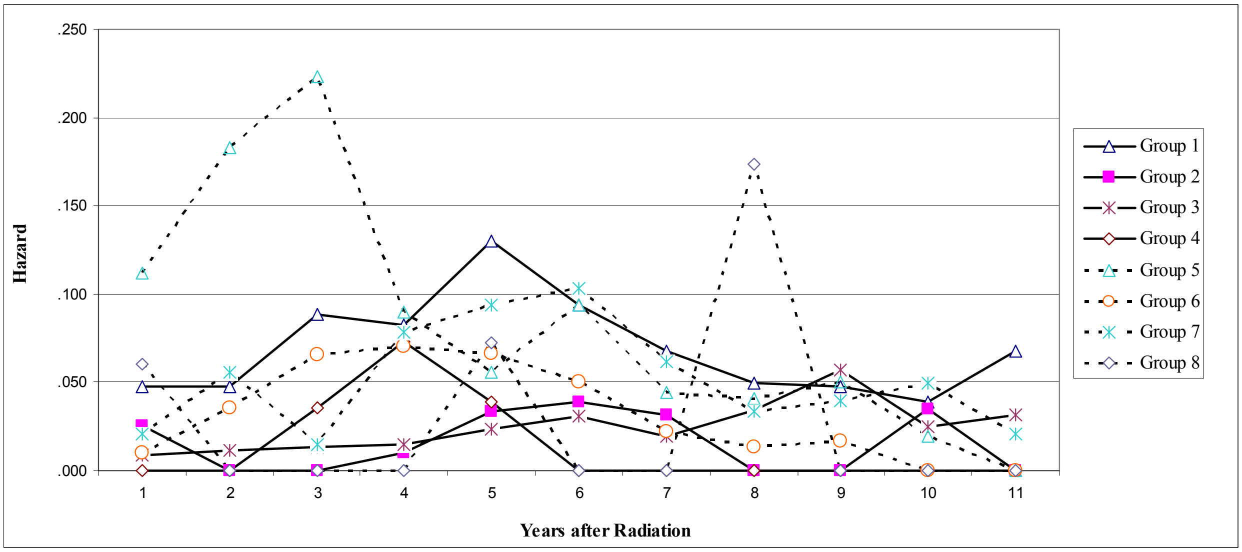


Figure 13. Hazard Functions for Biochemical Failure by Subgroups.

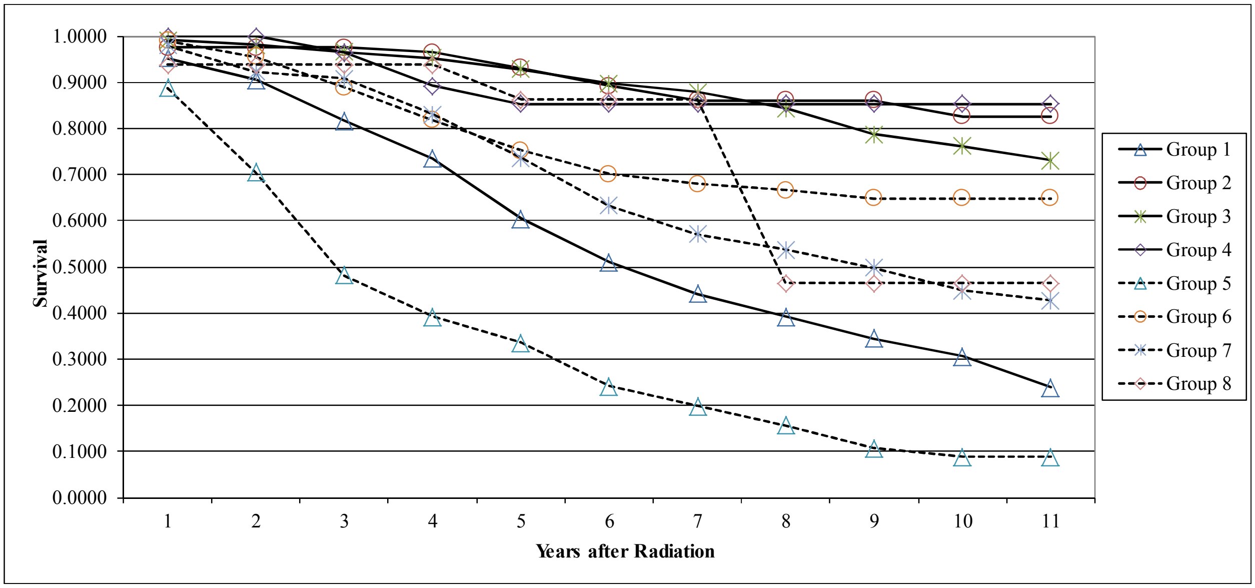


Figure 14. Survival Function for Biochemical Failure by Subgroups

The high-risk group with nadir time less than 2 year and no HT had the highest biochemical failure, and it had worse biochemical control comparing to intermediate-risk with nadir time less than 2 year and no HT. The graphs indicated that group without HT had worse outcome than group with HT. With the same characteristics, the high-risk group had worse outcome than intermediate-risk group. Group with shorter nadir time (< 2 years) had worse outcome than the group with longer nadir time (≥ 2 years). The Figure 30 in Appendix from Cox regression model indicated the same result.

Discrete-time Survival Model

The conversion from person-case format to person-period format was done by using SPSS ver. 22 code (Appendix A2). Under different sample sizes in Cox models, the numbers of records of person-period data in DTSA models were reconstructed correspondingly, depending on the duration of observation of each case and the number of time periods.

The logit model A through H was tested by different sample sizes. The summary statistics for discrete-time survival models under different sample sizes are listed in the Table 17 - 20 for 10 time periods and Table 21 - 24 for 5 time periods.

The univariate analysis was conducted in the discrete-time model, the estimated odds, the estimated hazard, and the estimated logit(hazard) under five time periods and ten time periods were plotted for the Model F, G, and H. The Figure 27, Figure 28, and Figure 29 are the estimated hazard, estimated odds, and the estimated logit(hazard) against time for Model F under ten periods. The plots for Model D and E under ten periods and five periods are present in the Figure 24, Figure 25, and Figure 26.

Analysis of the Effect of Sample Size

Chi-square statistics are affected by sample size. Therefore, the model chi-square and likelihood ratio chi-square are not appropriate statistics to use when comparing data sets for difference due to the sample size. However, a visual analysis of hazard functions of data sets with different sample size by using either the Cox regression model or the discrete-time survival model with same time periods reveal that the smaller sample sizes had higher hazard estimates in general. The hazard estimates for the compared data sets can be ordered in this fashion: n=1577; n=1213; n=809; n=422, and hazard estimates for the compared data sets are plotted in Figure 15. The dash dot line represents the hazard function for the data set with smallest sample size (n=422), and the hazard function is similar with other data set with different sample sizes at beginning. The other lines are for sample size 809, sample size 1213, and sample size 1577.

As shown in Figure 15, the smaller sample size had a larger hazard estimate. When the sample size reaches 500 more, the difference between the hazard estimates among the different sample sizes becomes smaller. After the five year post-treatment, the hazard function in the sample size 422 starts to rise above from other three lines which indicates a higher hazard function estimation. Presented in Figure 16 and Figure 17, the survival function and log hazard function were plotted by different sample sizes.

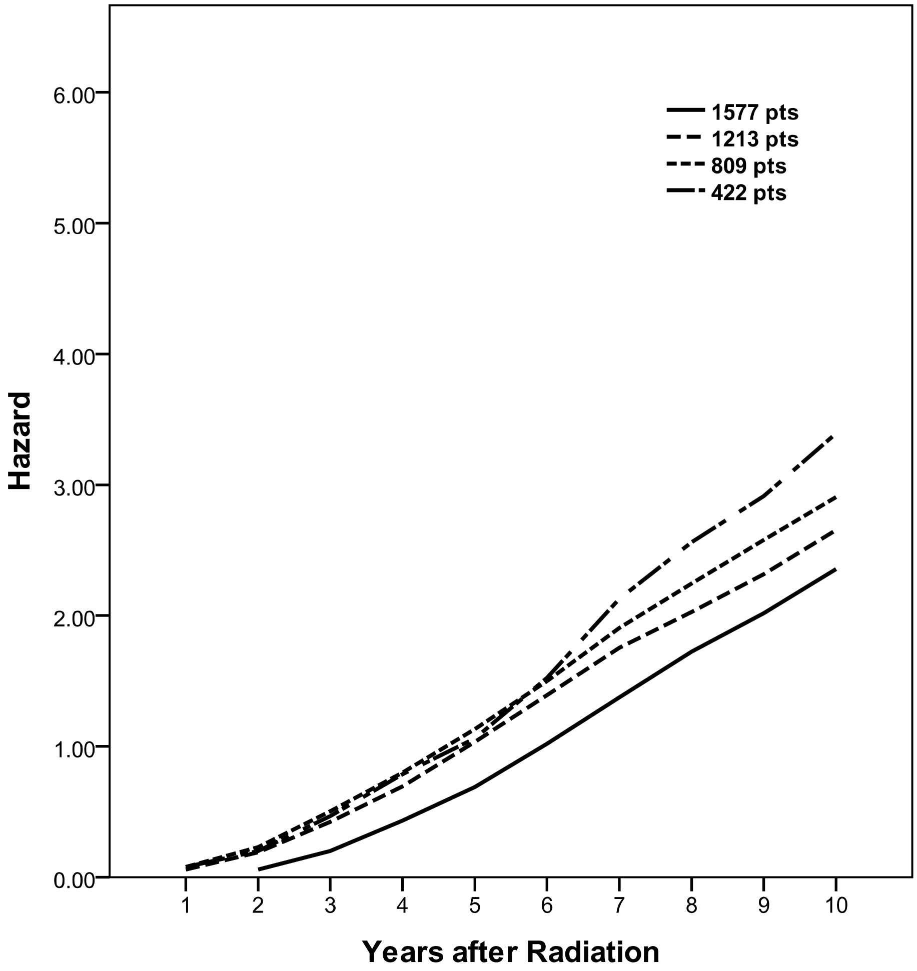


Figure 15. Cumulative Hazard Rate Comparisons in Cox Regression Model H by Different Sample Sizes

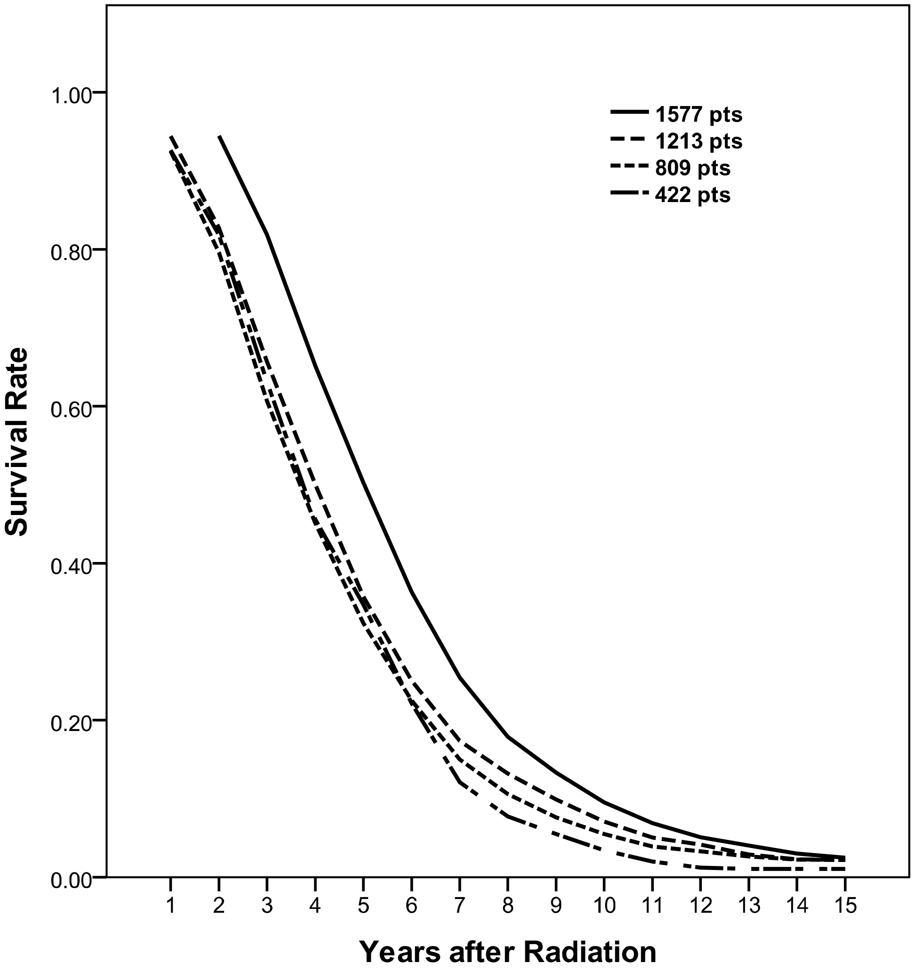


Figure 16. Survival Function Comparisons in Cox Regression Model H by Different Sample Sizes

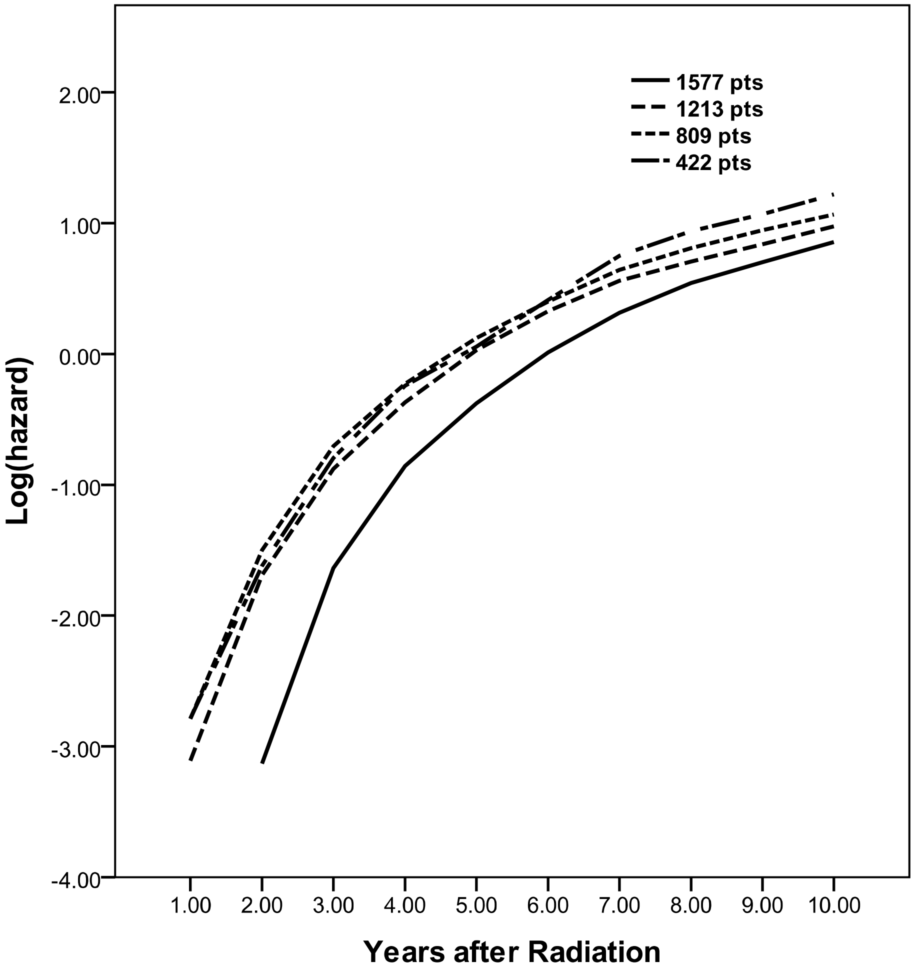


Figure 17. Log Hazard Comparisons in Cox Regression Model H by Different Sample Sizes

For the discrete-time survival model, hazard estimates for the compared data sets can be ordered in the same fashion: n=1577; n=1213; n=809; n=422. Hazard estimates, odds estimates, and logit(hazard) are plotted in Figure 18 through Figure 23 for sample size n=1577 in Appendix for these comparisons. Generally, data sets with smaller sample sizes have higher hazard estimates. The group with sample size n=422 had larger variance due to the small size comparing with the number of variable in the Model F.

The findings from this study indicate that sample size has an impact on survival analysis and hazard estimates. As the sample size decrease, the noise increase. With five time periods, the sample size n = 422 had the highest hazard estimates compare to the other sample size group.

The trends of odds, hazard, and logit(hazard) were similar for the sample size n = 1517, n = 1213, and n = 809.

As shown in the Figure 18, Figure 19 and Figure 20, the odds, logit(hazard) and hazard functions are presented for five-period data sets. The estimations under sample size n = 422 were dramatically different with the other three sample size. Under the ten time periods, the differences of estimations between different sample size groups were decreasing.

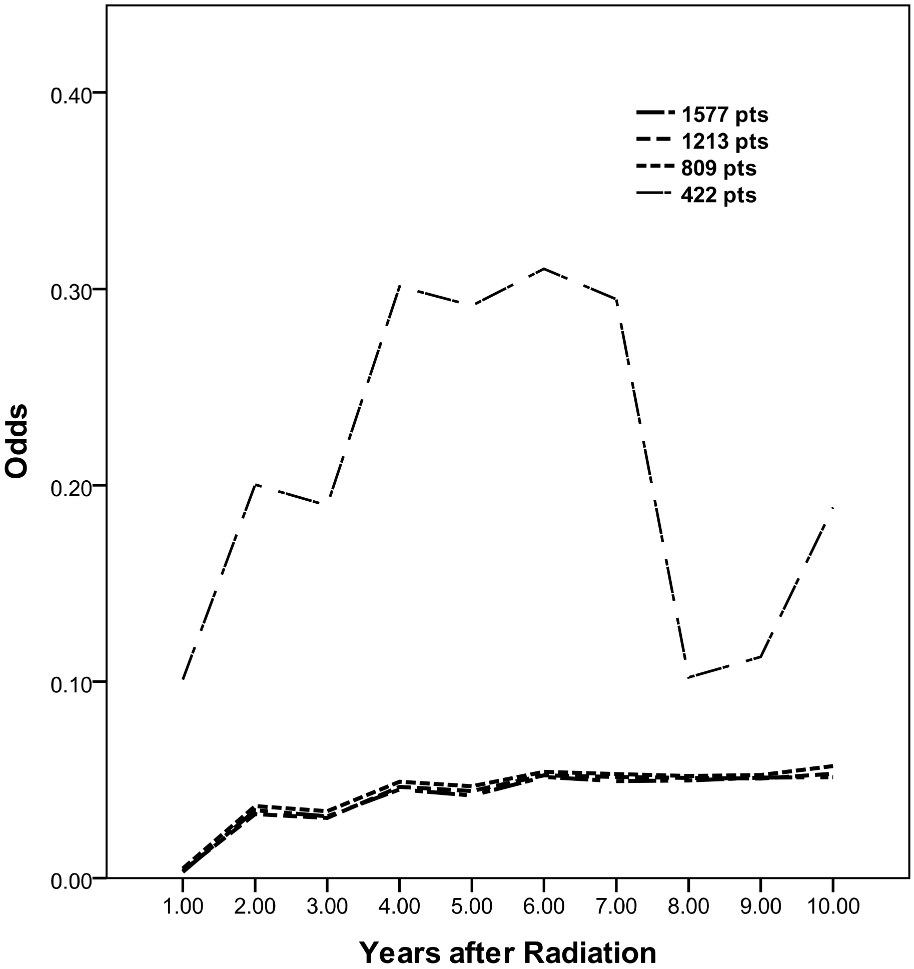


Figure 18. Odds Comparisons in Discrete-time Model F by Different Sample Sizes Under Five Time Periods

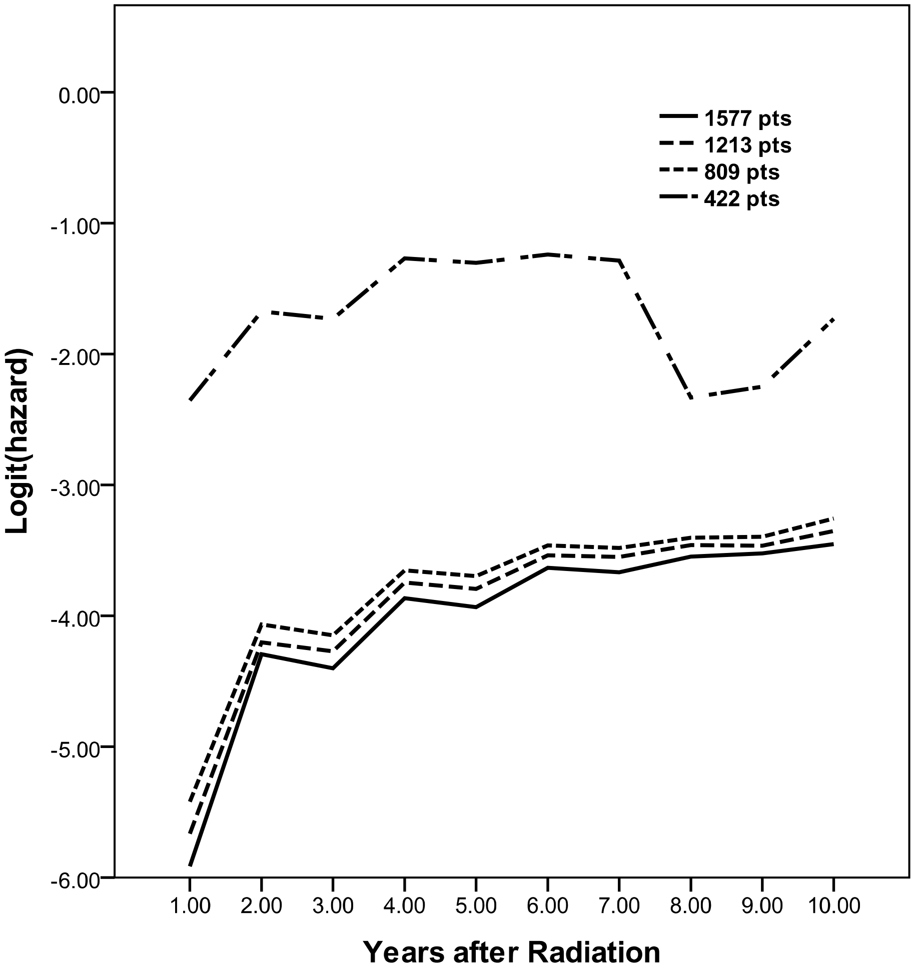


Figure 19. Logit Hazard Comparisons in Discrete-time Model F by Different Sample Sizes Under Five Time Periods

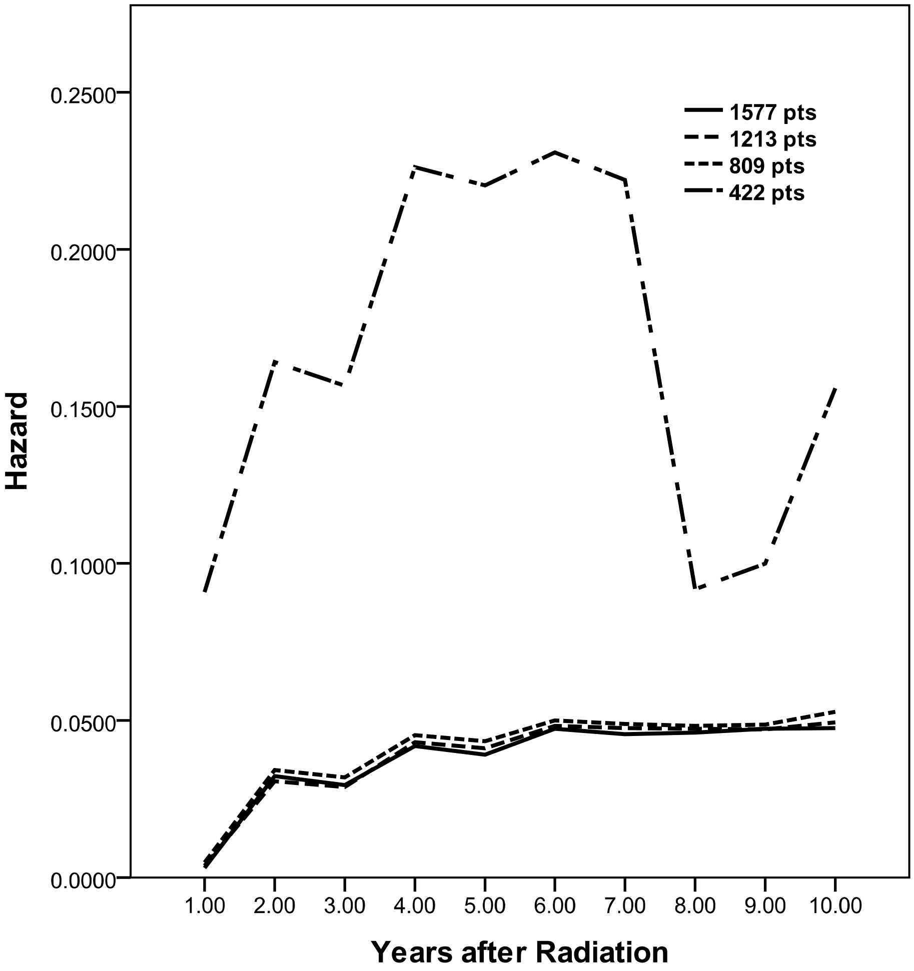


Figure 20. Hazard Comparisons in Discrete-time Model F by Different Sample Sizes Under Five Time Periods

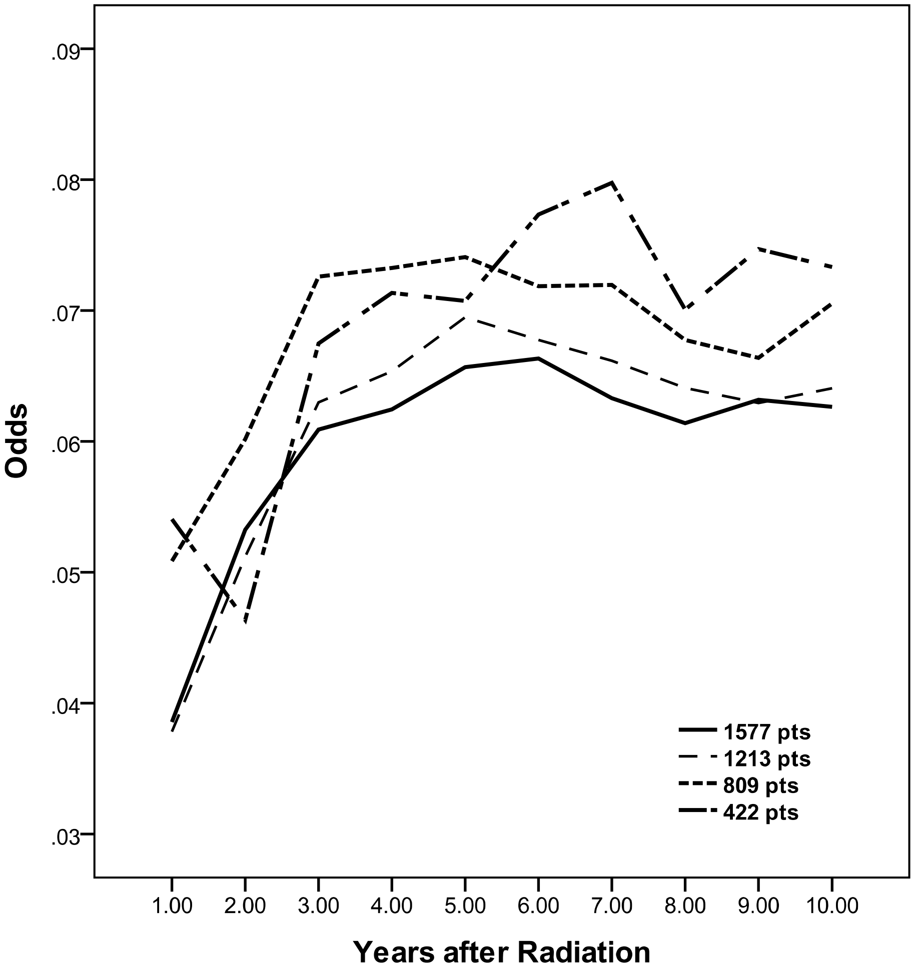


Figure 21. Odds Comparisons in Discrete-time Model F by Different Sample Sizes Under Ten Time Periods

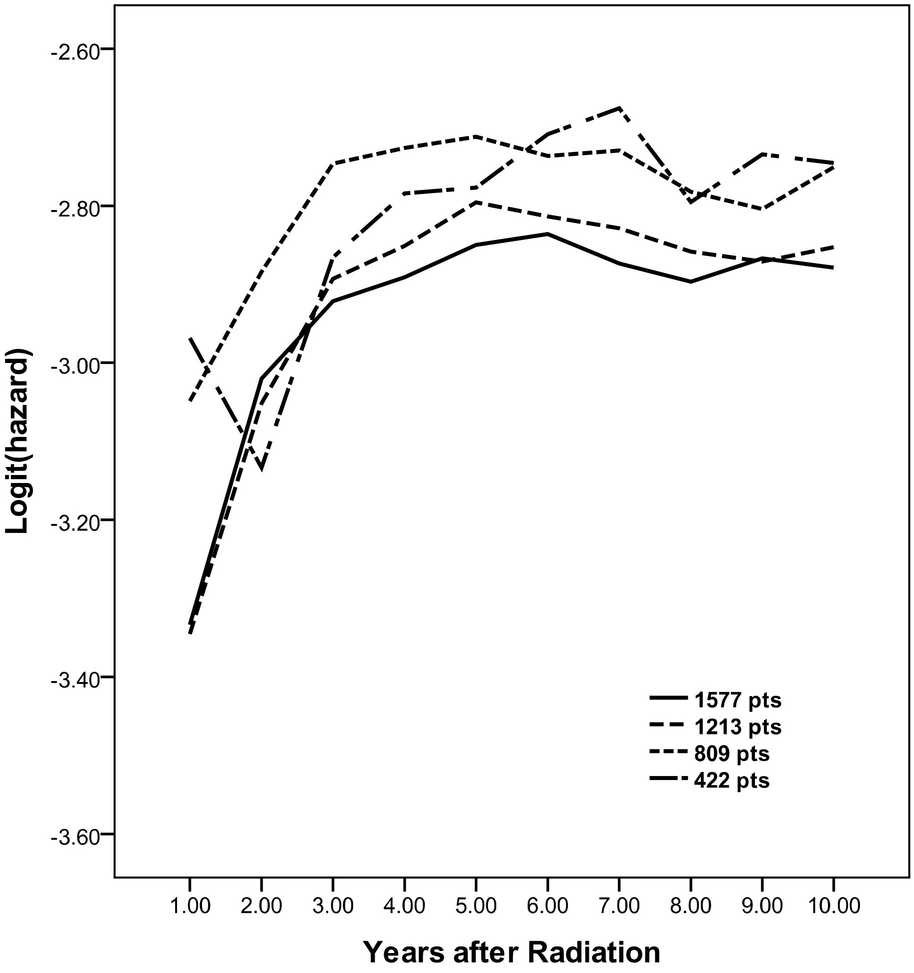


Figure 22. Logit Hazard Comparisons in Discrete-time Model F by Different Sample Sizes Under Ten Time Periods

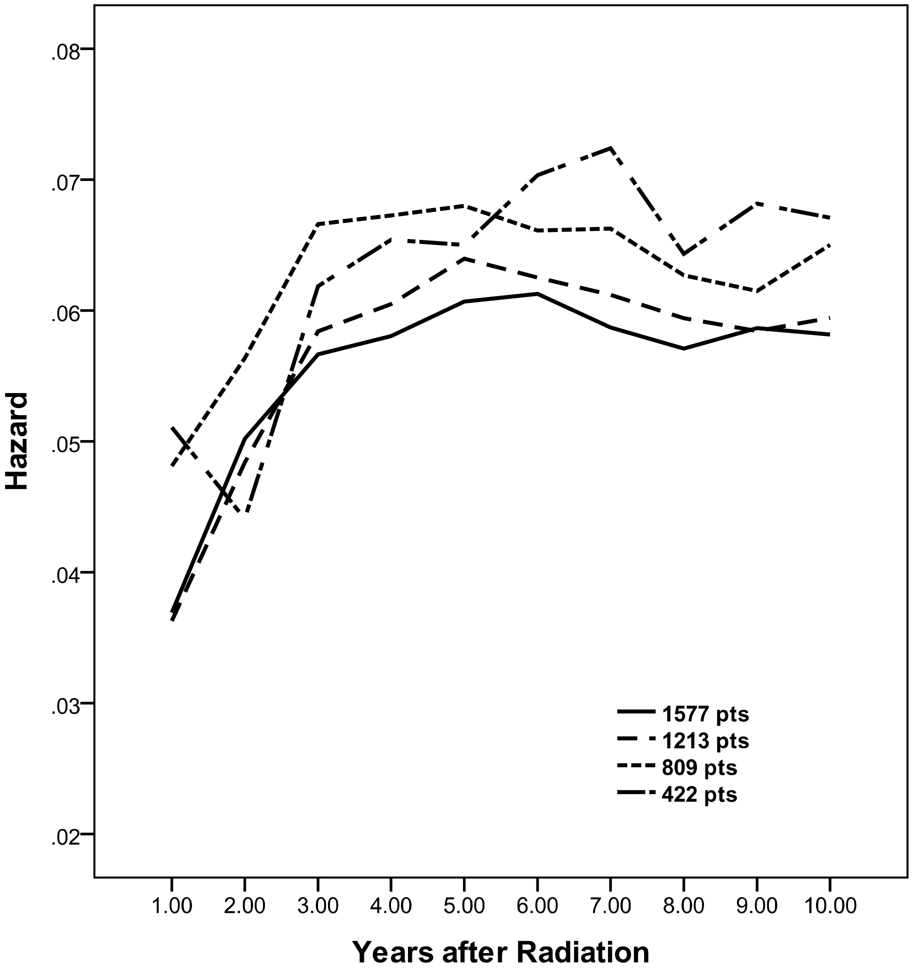


Figure 23. Hazard Comparisons in Discrete-time Model F by Different Sample Sizes Under Ten Time Periods

Analysis of the Effect of Number of Time Periods

Within the same model, the estimated hazard, odds, and logit(hazard) were plotted for the same group of patients after restructuring them into person-period data format in five time periods and ten time periods. Figure 24, Figure 25, and Figure 26 are the hazard, odds, and logit(hazard) against time for Model F, Model G, and Model H under five time periods for the sample size n=1577. For the ten time periods, the odds, logit(hazard) and hazard were plotted in the Figure 27, Figure 28, and Figure 29. The plots of odds, logit(hazard) and hazards for sample size n=1213, n=809, and n=422 are presented in the Figure 35 - 52.

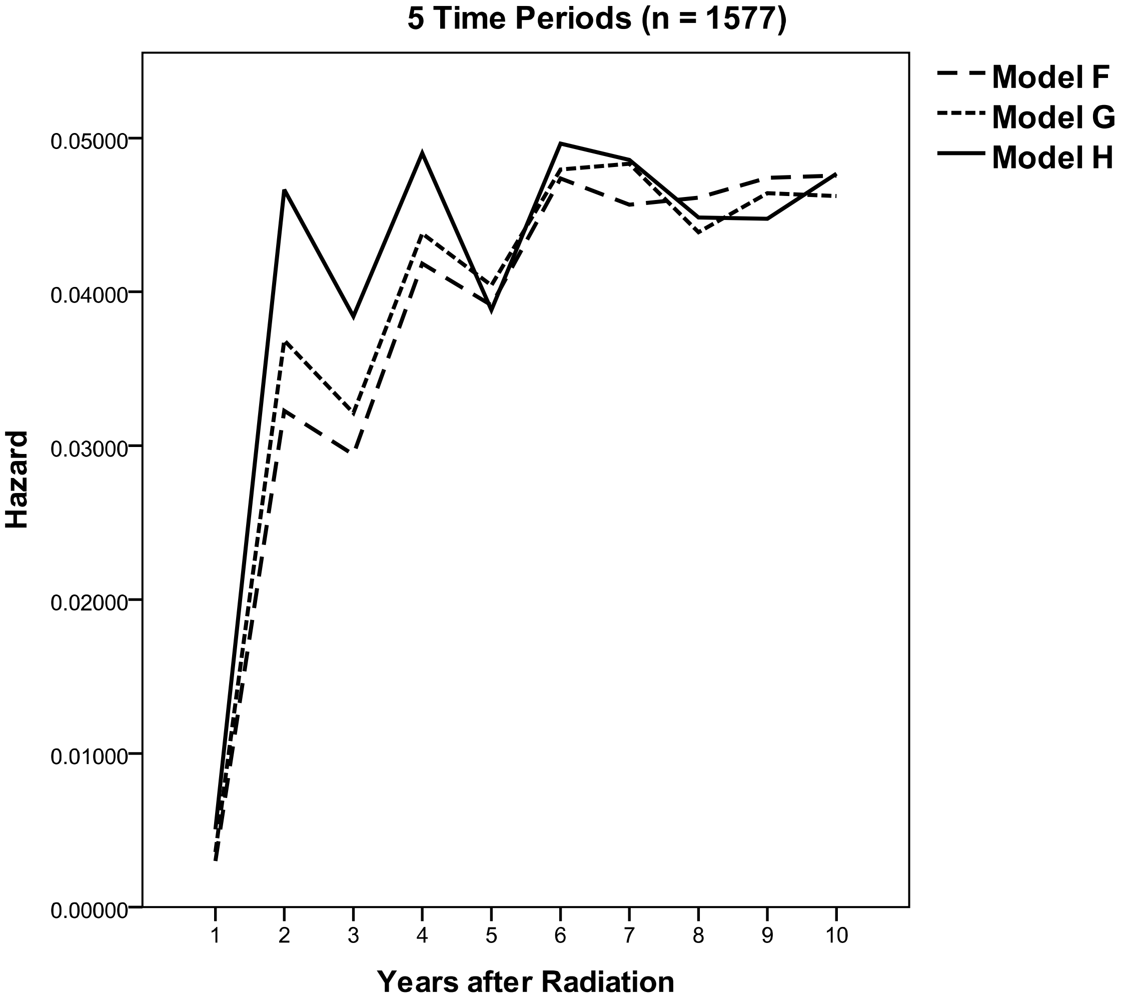


Figure 24. Hazard Comparisons of Model F/G/H by Using Discrete-time Method with Five Time Period Under Sample Size n = 1577

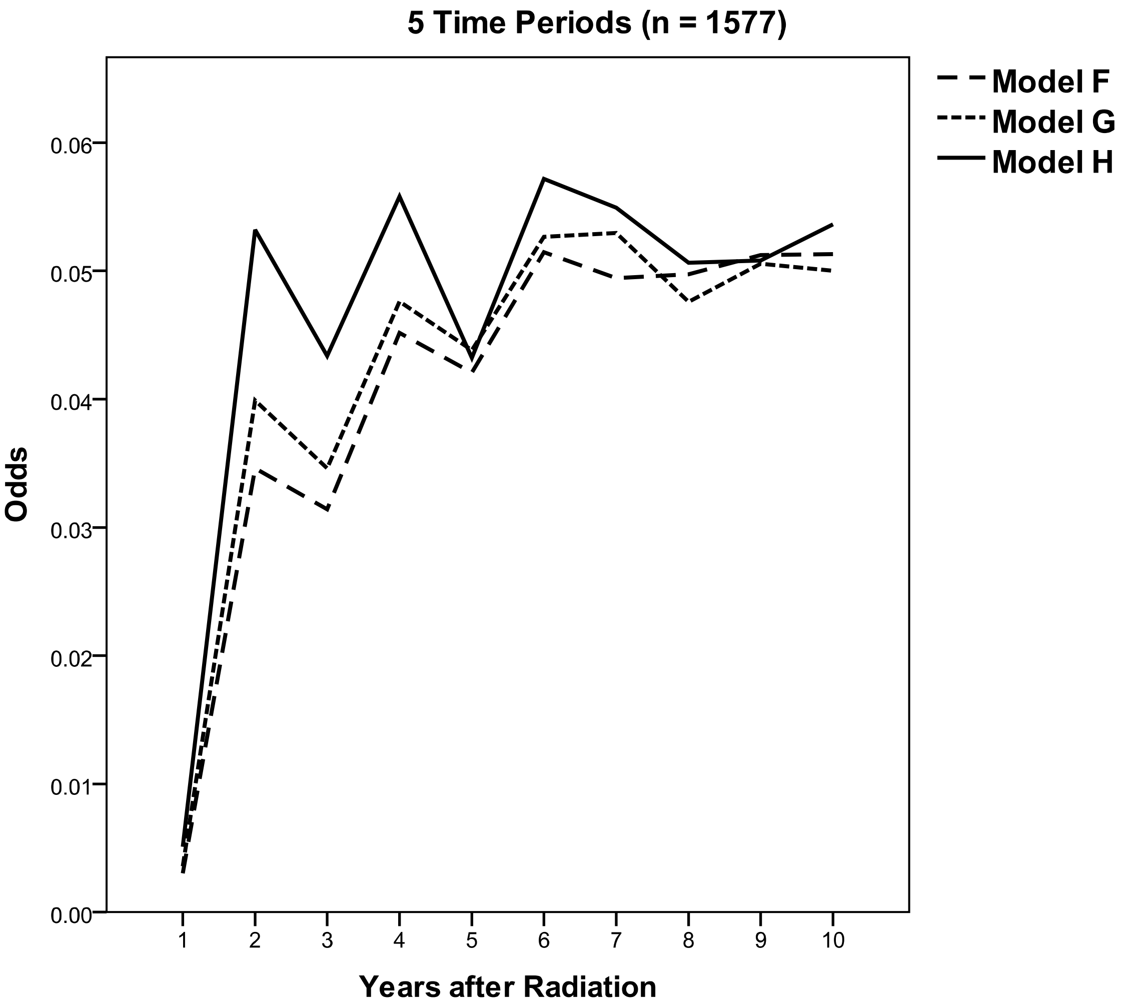


Figure 25. Odds Comparisons of Model F/G/H by Using Discrete-time Method with Five Time Period Under Sample Size n = 1577

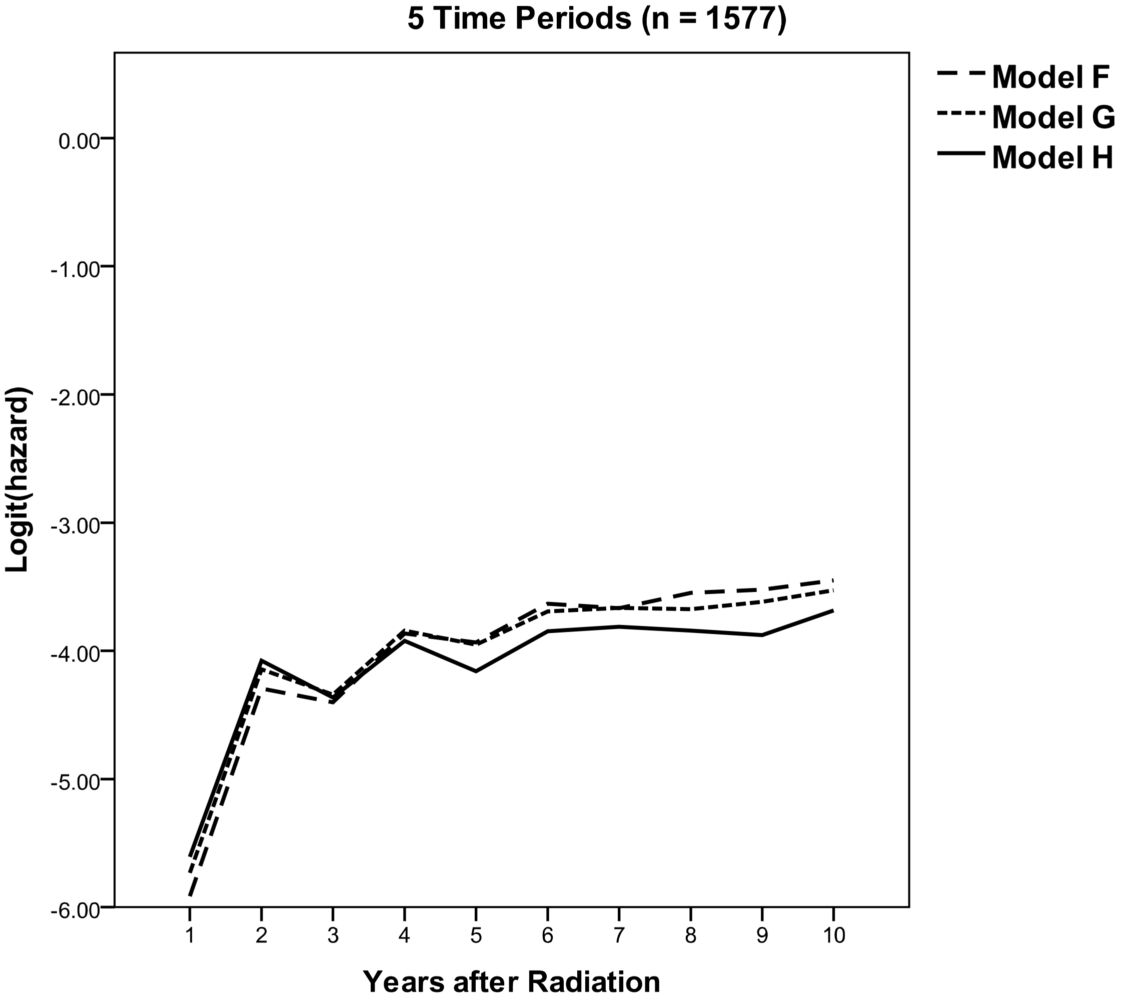


Figure 26. Logit Hazard Comparisons of Model F/G/H by Using Discrete-time Method with Five Time Period Under Sample Size n = 1577

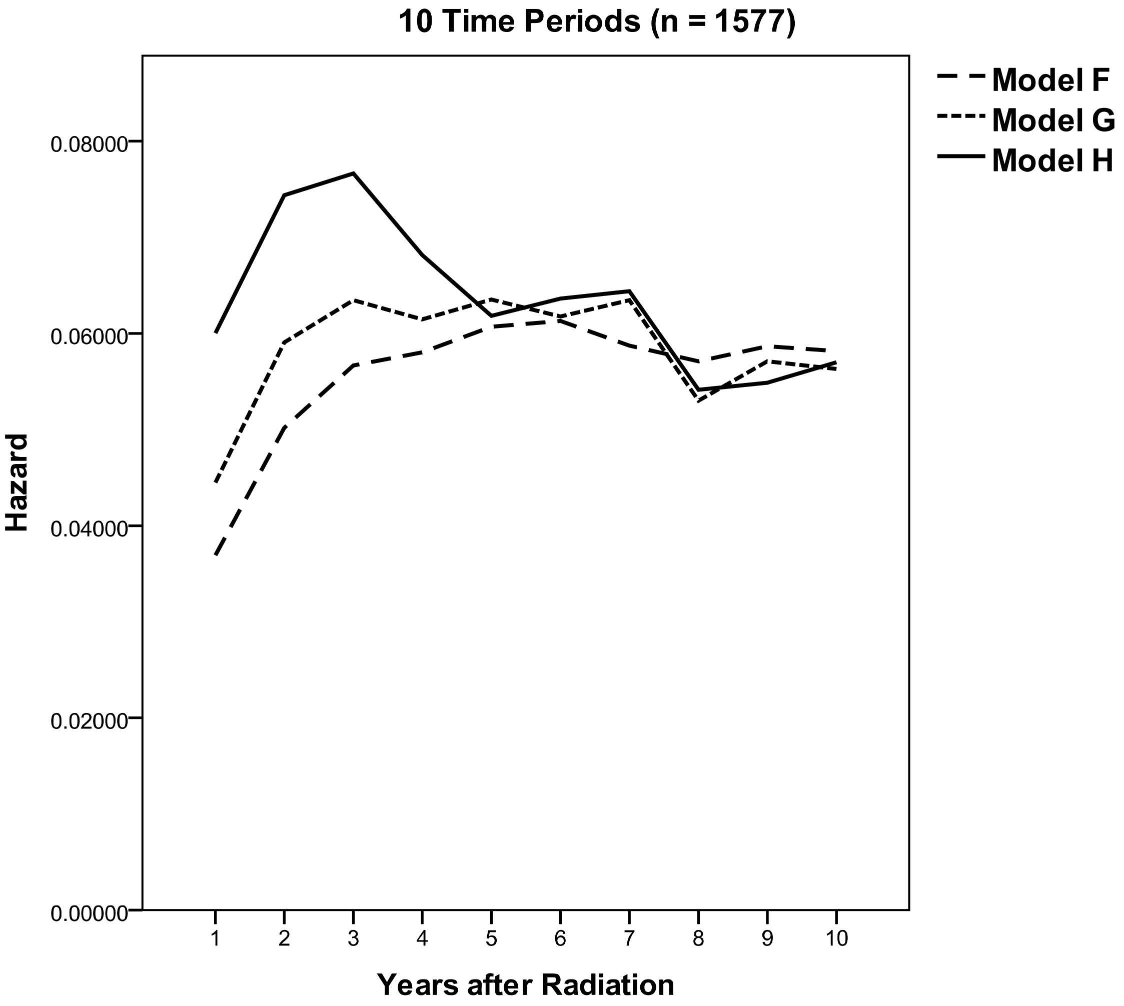


Figure 27. Hazard Comparisons of Model F/G/H by Using Discrete-time Method with Ten Time Period Under Sample Size n = 1577

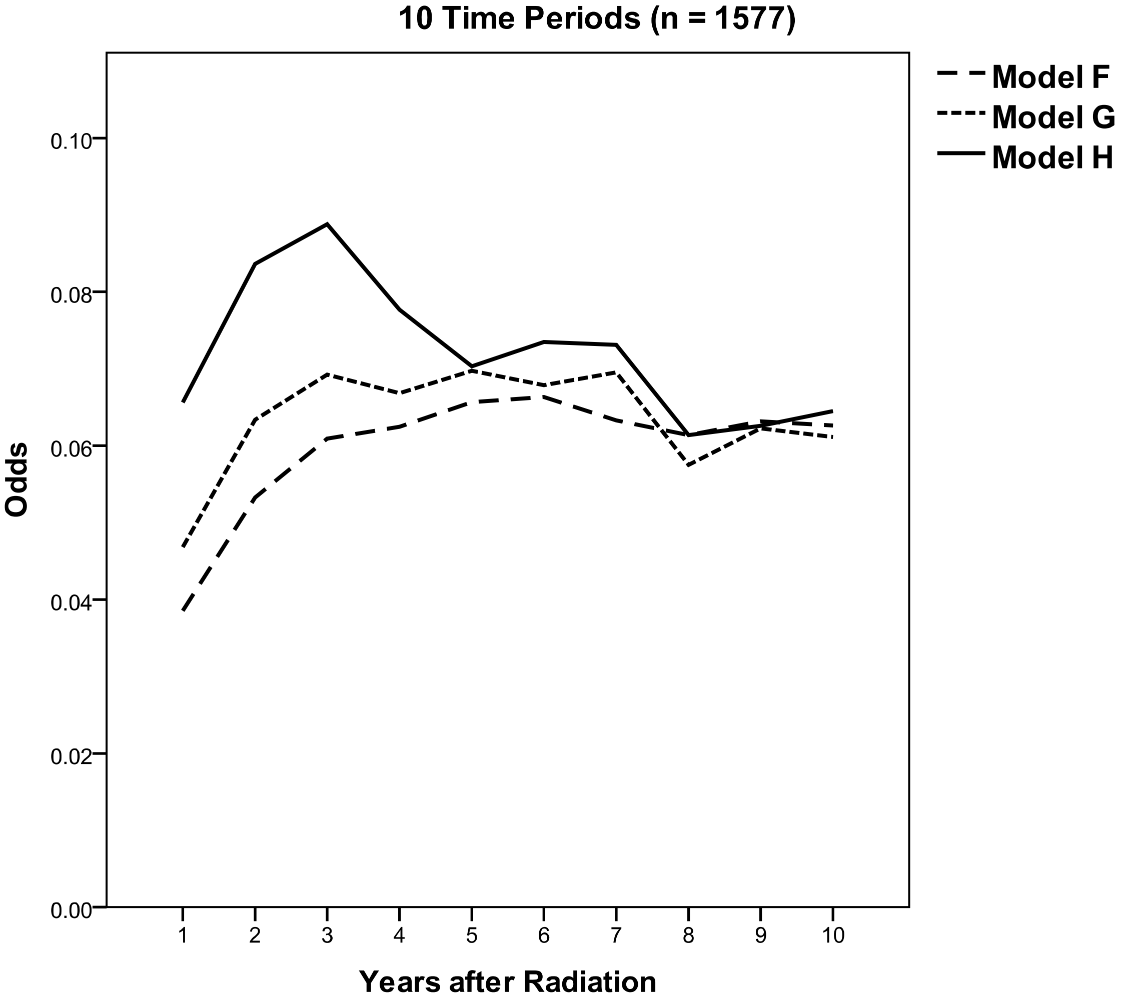


Figure 28. Odds Comparisons of Model F/G/H by Using Discrete-time Method with Ten Time Period Under Sample Size n = 1577

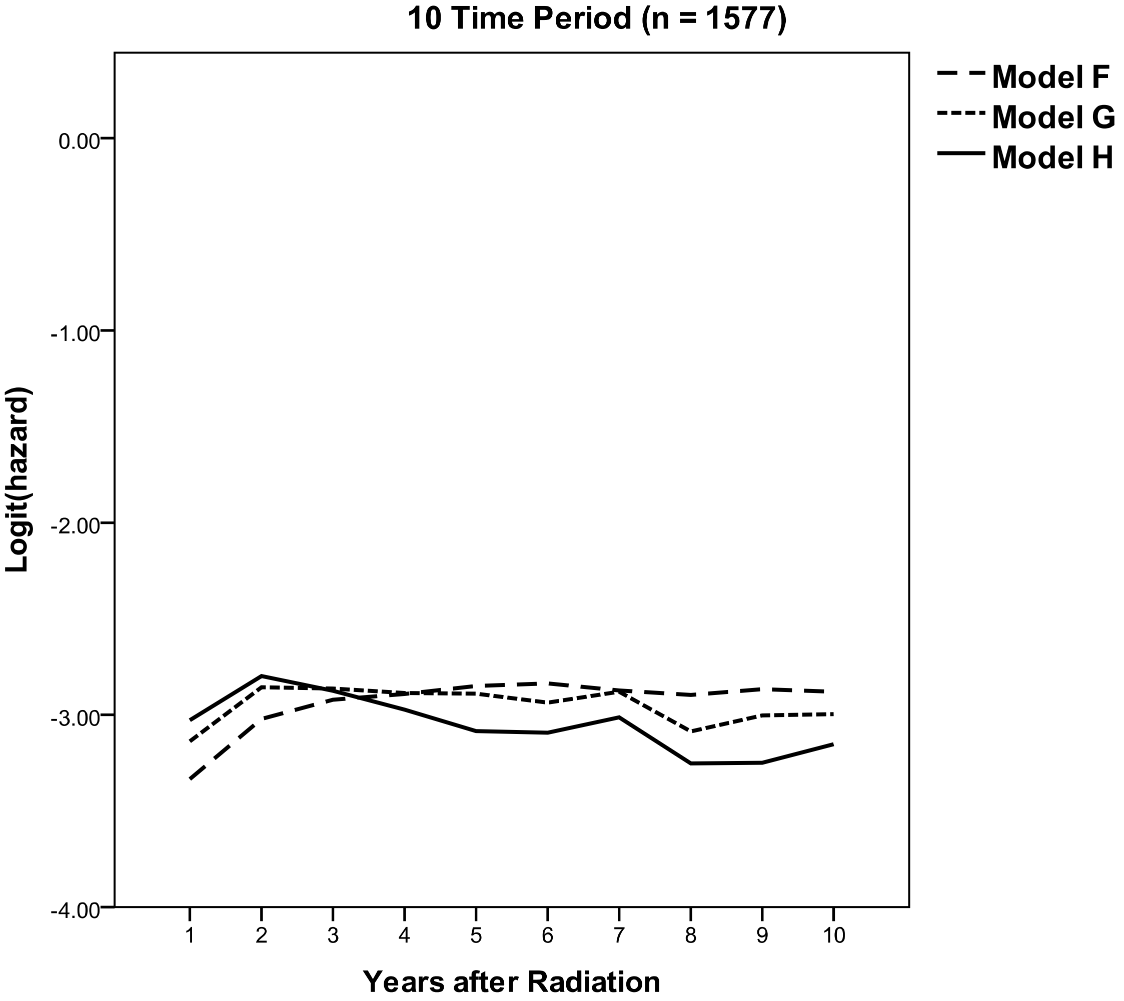


Figure 29. Logit Hazard Comparisons of Model F/G/H by Using Discrete-time Method with Ten Time Period Under Sample Size n = 1577

Analysis of the Effect on Hazard Estimates

The hazard estimates were conducted to compare between Cox regression and discrete-time survival model. For certain group of patients under one sample size, the same population was reconstructed into person-period data format. The hazards were tested in both models. For model F, the comparisons of hazard estimations from Cox model and discrete-time survival model are listed in the Tables 4 - 7 under different sample sizes and time periods. The results indicate that both Cox regression model and discrete-time survival model provided similar hazard estimation.

Presented in Table 4 are the results of hazard estimations from Cox regression and discrete-time survival models with different time periods for the sample size n=1577. In the Cox regression, the patient in high-risk group had 2.541 times more likely to have biochemical failure than the patient in the intermediate-risk group. The patient with nadir time less than 2 years had 3.947 times more likely to have biochemical failure compare to the patients with nadir time greater than 2 years. The patients with no hormone treatment before radiation had 4.974 times more likely to have biochemical failure compare to the patients with hormone treatment before radiation. In the discrete-time survival model with 10 time periods, the patients in high-risk had 2.717 times more likely to have BF compare to intermediate-risk patient. The patients with nadir time less than 2 years and did not receive the hormone treatment before radiation had 4.161 and 5.457 times more likely to have BF compare to group with nadir time longer than 2 years and patients with hormone. Similarly, in the discrete-time model with 5 time periods, the patients in high-risk, with nadir time less than 2 years and no hormone treatment before radiation had 3.242, 3.303 and 5.150 times more likely to have BF compare to patients in intermediate-risk group, patients with nadir time longer than 2 year and patients with hormone treatment before radiation.

Table 4

Model H Comparisons between Cox Regression and Discrete-time Survival Model under Different Time Periods (n = 1577)

	Model H					
	Cox Regression Model		Discrete-time Survival Model 5 Time Periods		Discrete-time Survival Model 10 Time Periods	
	N=1577		N=6041		N=9692	
Variable	*HZ*	*95% CI*	*HZ*	95% *CI*	HZ	95% *CI*
High Risk	2.541***	2.141 - 3.017	3.242***	2.474 - 4.247	2.717***	2.262 - 3.265
Nadir Time < 2 yrs	3.947***	3.242 – 4.805	3.303***	2.463 – 4.429	4.161***	3.389 – 5.109
No HT before RT	4.974***	3.856 – 6.416	5.150***	3.493 – 7.594	5.457***	4.183 – 7.120
Goodness-of-fit						

-2LL	7106.8	1787.9	3825
n parameters	3	8	13
AIC	7166.8	1803.9	3851
BIC	7182.9	1857.6	3944.3

Note. *HZ* = hazard ratio, an *HZ* < 1 indicates a lower risk for the indicator group, *HZ* = 1 no difference between indicator and reference group, *HZ* > 1 indicates a higher risk for the indicator group; CI = confidence interval.

$^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

The hazard estimations for sample size n = 1213, n = 809, and n = 422 are shown in the Table 5, 6, and 7.

Table 5

Model H Comparisons between Cox Regression and Discrete-time Survival Model under Different Time Periods (n = 1213)

	Model H					
	Cox Regression Model		Discrete-time Survival Model 5 Time Periods		Discrete-time Survival Model 10 Time Periods	
Variable	*HZ*	*95% CI*	*HZ*	95% *CI*	HZ	95% *CI*
High Risk	2.481^{***}	1.448 - 2.117	2.908^{***}	2.148 – 3.936	2.649^{***}	2.156 – 3.256
Nadir Time < 2 yrs	3.931^{***}	3.154 – 4.900	3.262^{***}	2.354 – 4.518	4.151^{***}	3.299 – 5.224
No HT before RT	5.145^{***}	3.833 – 6.907	5.853^{***}	3.675 – 9.321	5.693^{***}	4.186 – 7.743
Goodness-of-fit						
-2LL	5443		1421.8		3025.9	
n parameters	3		8		13	
AIC	5449		1437.8		3051.9	
BIC	5464		1489.4		3141.9	

Note. *HZ* = hazard ratio, an *HZ* < 1 indicates a lower risk for the indicator group, *HZ* = 1 no difference between indicator and reference group, *HZ* > 1 indicates a higher risk for the indicator group; CI = confidence interval.

$^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

Table 6

Model H Comparisons between Cox Regression and Discrete-time Survival Model under Different Time Periods (n = 809)

	Model H					
	Cox Regression Model		Discrete-time Survival Model 5 Time Periods		Discrete-time Survival Model 10 Time Periods	
Variable	*HZ*	*95% CI*	*HZ*	95% *CI*	HZ	95% *CI*
High Risk	2.450^{***}	1.946 – 3.083	2.919^{***}	2.030 – 4.197	2.626^{***}	2.053 – 3.360
Nadir Time < 2 yrs	3.730^{***}	2.872 – 4.844	2.918^{***}	1.981 – 4.300	3.897^{***}	2.966 – 5.122
No HT before RT	5.213^{***}	3.651 – 7.444	6.284^{***}	3.513 – 11.24	5.724^{***}	3.947 – 8.302
Goodness-of-fit						
-2LL	3576.7		973.9		2086.3	
n parameters	3		8		13	
AIC	3582.7		989.9		2112.3	

BIC	3596.8	1038.2	2196.9

Note. *HZ* = hazard ratio, an *HZ* < 1 indicates a lower risk for the indicator group, *HZ* = 1 no difference between indicator and reference group, *HZ* > 1 indicates a higher risk for the indicator group; CI = confidence interval.

${}^{*}p < 0.05$. ${}^{**}p < 0.01$. ${}^{***}p < 0.001$.

Table 7

Model H Comparisons between Cox Regression and Discrete-time Survival Model under Different Time Periods (n = 422)

	Model H					
	Cox Regression Model		Discrete-time Survival Model 5 Time Periods		Discrete-time Survival Model 10 Time Periods	
Variable	*HZ*	*95% CI*	*HZ*	*95% CI*	HZ	*95% CI*
High Risk	2.506^{***}	1.826 – 3.439	2.918^{***}	2.171 – 3.922	2.630^{***}	1.870 – 3.700
Nadir Time < 2 yrs	3.966^{***}	2.779 – 5.660	1.493^{*}	1.092 – 2.041	4.173^{***}	2.863 – 6.082
No HT before RT	8.265^{***}	4.712 – 14.496	8.101^{***}	4.724 – 13.891	9.095^{***}	5.076 – 16.295
Goodness-of-fit						
-2LL	1642.1		1244.8		1060.9	
n parameters	3		8		13	
AIC	1648.1		1260.8		1086.9	
BIC	1660.2		1303.7		1162.6	

Note. *HZ* = hazard ratio, an *HZ* < 1 indicates a lower risk for the indicator group, *HZ* = 1 no difference between indicator and reference group, *HZ* > 1 indicates a higher risk for the indicator group; CI = confidence interval.

${}^{*}p < 0.05$. ${}^{**}p < 0.01$. ${}^{***}p < 0.001$.

The model G comparisons under different sample size are listed in the Table 25 - 28.

Effects of Covariates

As in any analysis with covariates, identifying which covariate has significant effects on the response is a major issue of interest. In the case of the tumor recurrence data here, all three factors including if patient received HT before radiation, time to reach nadir PSA, and risk group have a significant effect on the risk of biochemical recurrence. The significance of covariate effects can be assessed by using the confidence intervals of the covariates βs. The approximate $100(1-\alpha)$ per cent confidence intervals $\beta_1, \beta_2, \beta_3$ can be calculated by the formula

$$\text{estimate} \pm z_{\alpha/2} \times \text{STD},$$

where $z_{\alpha/2}$ is the upper $(\alpha/2)$ th-percentile of the standard normal distribution.

CHAPTER 5 CONCLUSIONS

Cox regression models had been utilized in many survival applications in medical data analysis. Compared to the Cox regression model, the discrete-time survival model has been used more frequently in the fields of education and social science in the past decades, but it is still not a familiar method in the medical literature. There is little information on how certain data characteristics impact survival analysis hazard estimates and goodness of fit statistics between Cox regression and discrete-time survival models. This study examined three attributes including sample size, the number of time period, and the number of parameters used in the model, and investigated how these attributes related to the hazard estimation and model fitness. The purpose of this study was to investigate the effects of attributes that could be compared for any survival data set by using the Cox regression model and the discrete-time survival model. Data sets came from experimental medical data and were compared between both models to assess if varying these characteristics caused statistically significant differences among the model chi-squares and likelihood ratios. Hazard estimations were also compared to assess the effects of the varied models and varied characteristics. Based on the results of the study, the sample size does have an effect on hazard estimates. Sample size has been found to have an effect on many statistical procedures. Both the Cox regression and the discrete-time survival model have chi-square distributions, thus it was not appropriate to compare model chi-squares and likelihood ratios across models with different sample sizes. Therefore, only hazard functions were used for comparison.

From the Cox regression model, data sets with smaller sample sizes had higher hazard estimates than the data sets with larger sample size. For the discrete-time survival model, the group with sample size n=422 had relative larger hazard estimates than the groups with sample size n=809, n=1213, and n=1577 with five time periods. For the discrete-time survival model under ten time periods, the hazard estimates crossed over among four sample size groups. However, after removing the group with sample size n=422, the hazard estimates had similar trends. Generally, the group with larger sample size had the smaller hazard estimates. Decreasing the sample size produced larger hazard estimates.

As observed from the results of data sets with different lengths of time periods, the goodness of fit statistics which was measured in more finite units are significantly different with those measured in fewer units. Data sets coded into ten time periods had larger model chi-squares and likelihood ratio values than those coded into five time periods. In both logistic regression and Cox regression, smaller values of chi-square and the likelihood ratio indicate a better fit the model. Data sets with fewer time periods had smaller sample sizes compare with data sets with more time periods. Further work is needed to test if smaller values of the likelihood ratio or chi-square indicate a better fit of the model with different sample size.

Under the five time periods, model H had larger odds, hazard, and logit(hazard) estimate compared to model F and model G because model H involved more variables than the other two models. However, under the ten time periods, the difference between Model H and Model F or Model G became smaller as increasing the time periods from five to ten and decreasing the sample size.

The hazard estimation is the cornerstone when comparing two methods in this study. As we can see from the hazard estimation table (Table 4, 5, 6, and 7), discrete-time survival method provided similar results as Cox regression had. Under the same sample size n=1577, n=1213, or n=809, hazard estimates under ten time periods are closer to the hazard estimates from Cox regression model with narrower confidence interval compared to five time periods. For the data with sample size n=422, the hazard estimates of certain variable in Model G and H did not reach significance under the five time periods, but showed the significance in the Cox regression and discrete-time survival model under ten time periods. It is possible that fewer subjects and events were observed in the shorter period (ten time periods) instead of longer time period (five time periods) due to the random selected sample. In general, the discrete-time survival models provided similar results comparing with ones in Cox regression model, and the strategies for comparing -2LL statistics for the Cox regression model are identical to those for comparing deviance statistics for the discrete-time hazard model.

The aim of this study was trying to identify certain attributes related to the hazard estimation and model fitness in the survival analysis. The strategies of both the Cox regression model and the discrete-time survival model are comparable to provide similar answers for the hazard estimation. In many real life scenarios, especially for cancer care situations, the completion of cancer treatments is not the end point for the patient outcome analysis. After the completion of treatment, cancer patients will experience a series of follow-up care in the long term. Current NCCN guideline suggests several treatment options after radiation therapies which include observation, ADT, clinical trial, and regular laboratory testing. Also, the NCCN guideline provides recommended follow-up care plan for patients. Either additional treatment plan or follow-up plan options could potentially increase the patients' anxiety and doctor's concern. However, if the clinicians have an better understanding regarding the whether the cancer occur, and if so, when the cancer come back based on patients' certain characteristics and follow-up information, then some patients may not need frequent follow-up monitoring after cancer treatment.

RECOMMENDATIONS

Real data were tested under the situations with varied attributes. It would be beneficial if there were more medical data sets with differing characteristics available to be tested. Several recommendations for further research are the following:

1. Twenty four data sets were generated from one prostate oncology data. The comparisons were conducted by changing one attribute, for example, same parameters, same time period, but different sample size. The comparisons of data sets that differ in more than one attribute should be conducted.

2. The $G^2(\mathrm{df}=1)$ value was used to compare the model chi-square and the likelihood ratio, but a statistic to compare hazard estimates needs to be identified and conducted.

3. Sample sizes were chosen randomly from the real data as 100%, 75%, 50%, and 25% of data set. Because the outcome is biochemical failure which is a binomial variable, the choice of sample sizes doesn’t count the balance of proportions of biochemical failure and biochemical control. Future analyses should consider the balanced proportions of event and control cases.

APPENDIX A ADDITIONAL TABLES AND FIGURES

Table 8

Supplement NCCN Risk Guideline for Prostate Cancer Patients

Risk Group	Definition
Low risk	Meeting all three conditions: 1) T1a, T1b, T1c, or T2a 2) Pre-RT PSA < 10 ng/mL 3) Gleason score <=6
Intermediate risk	Meeting at least one from all three conditions: 1) T2b or T2c 2) Pre-RT PSA 10-20 ng/mL 3) Gleason score = 7
High risk	Meeting at least one from all three conditions: 1) T3a, T3b or T4 2) Pre-RT PSA >=20 ng/Ml 3) Gleason score 8 - 10

Table 9

Model Tested

Model	Variable Used	Value
A	Clinical tumor stage	T1a, T1b, T1c, T2a, T2b, T3a, T3b, T4
B	Tumor Gleason score	2-10
C	PSA Value before radiation treatment	>0
D	Time to reach the lowest PSA after radiation treatment	>=0
E	If Patient received hormone therapy before radiation	Yes/No
F	NCCN risk category	Intermediate- and high-risk
G	NCCN risk category + time to reach the lowest PSA after radiation treatment	
H	NCCN risk category + time to reach the lowest PSA after radiation treatment + If Patient received hormone therapy before radiation	

Table 10

Life Table of the Number of Biochemical Failure Cases, Probability of Biochemical Failure, and Cumulative Proportion of Biochemical Control Among 727 Low-risk Patients Over 20 Years

	No. of patients who			Proportion of	
Years After RT	Had not yet experienced BF at the beginning of the year	Were censored at the end of the year	Number of BF at the end of the year	Patient who had BF during this year	All patients who had not experienced BF by the end of year
0-1	727	18	5	0.0070	0.9930
1-2	704	27	7	0.0101	0.9830
2-3	670	42	11	0.0169	0.9663
3-4	617	37	10	0.0167	0.9502
4-5	570	47	17	0.0311	0.9206
5-6	506	88	24	0.0519	0.8728
6-7	394	87	19	0.0542	0.8255
7-8	288	65	10	0.0391	0.7932
8-9	213	50	10	0.0532	0.7510
9-10	153	40	5	0.0376	0.7227
10-11	108	34	6	0.0659	0.6751
11-12	68	16	5	0.0833	0.6188
12-13	47	15	3	0.0759	0.5718
13-14	29	6	0	0.0000	0.5718
14-15	23	5	0	0.0000	0.5718
15-16	18	8	2	0.1429	0.4901
16-17	8	4	0	0.0000	0.4901
17-18	4	2	0	0.0000	0.4901
18-19	2	1	0	0.0000	0.4901

Table 11

Life Table of the Number of Biochemical Failure Cases, Probability of Biochemical Failure, and Cumulative Proportion of Biochemical Control Among 985 Intermediate-risk Patients Over 20 Years

	No. of patients who			Proportion of	
Years After RT	Had not yet experienced BF at the beginning of the year	Were censored at the end of the year	Number of BF at the end of the year	Patient who had BF during this year	All patients who had not experienced BF by the end of year
0-1	985	22	25	0.0257	0.9743
1-2	938	31	22	0.0238	0.9511
2-3	885	47	37	0.0429	0.9102
3-4	801	49	35	0.0451	0.8692
4-5	717	61	51	0.0743	0.8046
5-6	605	134	37	0.0688	0.7493
6-7	434	76	20	0.0505	0.7115
7-8	338	74	14	0.0465	0.6784
8-9	250	63	14	0.0641	0.6349
9-10	173	41	7	0.0459	0.6058
10-11	125	24	7	0.0619	0.5682
11-12	94	27	3	0.0373	0.5471
12-13	64	18	3	0.0545	0.5172
13-14	43	17	0	0.0000	0.5172
14-15	26	11	2	0.0976	0.4668
15-16	13	6	0	0.0000	0.4668
16-17	7	4	0	0.0000	0.4668
17-18	3	1	0	0.0000	0.4668
18-19	2	1	0	0.0000	0.4668
19-20	1	0	0	0.0000	0.4668
20-21	1	1	0	0.0000	0.4668

Table 12

Life Table of the Number of Biochemical Failure Cases, Probability of Biochemical Failure, and Cumulative Proportion of Biochemical Control Among 572 High-risk Patients Over 20 Years

	No. of patients who			Proportion of	
Years After RT	Had not yet experienced BF at the beginning of the year	Were censored at the end of the year	Number of BF at the end of the year	Patient who had BF during this year	All patients who had not experienced BF by the end of year
0-1	572	9	29	0.0511	0.9489
1-2	534	22	51	0.0975	0.8564
2-3	461	25	56	0.1249	0.7494
3-4	380	29	38	0.1040	0.6715
4-5	313	35	31	0.1049	0.6011
5-6	247	45	30	0.1336	0.5208
6-7	172	23	13	0.0810	0.4786
7-8	136	19	11	0.0870	0.4370
8-9	106	22	8	0.0842	0.4002
9-10	76	17	4	0.0593	0.3764
10-11	55	10	1	0.0200	0.3689
11-12	44	16	1	0.0278	0.3587
12-13	27	2	3	0.1154	0.3173
13-14	22	7	0	0.0000	0.3173
14-15	15	2	0	0.0000	0.3173
15-16	13	7	0	0.0000	0.3173
16-17	6	1	0	0.0000	0.3173
17-18	5	3	1	0.2857	0.2266
18-19	1	0	0	0.0000	0.2266
19-20	1	0	0	0.0000	0.2266

Table 13

Goodness-of-fit Summary for Model A – H for Sample Size n = 1577 Using Cox Regression

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Goodness-of-fit								
LL	-3712.5	-3737.9	-3598.6	-3675.65	-3721.1	-3703.3	-3652.05	-3580.4
-2LL	7425.0	7475.8	7197.2	7351.3	7442.2	7406.6	7304.1	7160.8
LR statistics	68.6	11.8	456.1	130.3	39.1	86.1	182.9	390.7
n parameters	1	1	1	1	1	1	2	3
p	<0.001***	0.001**	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***
AIC	7427	7477.8	7199.2	7353.3	7444.2	7408.6	7306.1	7166.8
BIC	7432.4	7483.2	7204.6	7358.7	7449.6	7414.0	7311.5	7182.9

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 14

Goodness-of-fit Summary for Model A – H for Sample Size n = 1213 Using Cox Regression

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Goodness-of-fit								
LL	-2984.3	-2865.4	-2766.15	-2816.35	-2851.75	-2841.4	-2799.85	-2721.5
-2LL	5968.6	5730.8	5532.3	5632.7	5703.5	5682.8	5599.7	5443
LR statistics	46.2	10.9	325.7	104	32.4	62.2	140.8	302.3
n parameters	1	1	1	1	1	1	2	3
p	<0.001***	0.001**	<0.001***	<0.001***	<0.001***	<0.001***	<0.001	<0.001***
AIC	5970.6	5732.8	5534.3	5634.7	5705.5	5684.8	5603.7	5449
BIC	5975.7	5737.9	5539.4	5639.8	5710.6	5689.9	5613.9	5464.3

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 15

Goodness-of-fit Summary for Model A – H for Sample Size n = 809 Using Cox Regression

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Goodness-of-fit								
LL	-1873.9	-1885.35	-1805.2	-1855.45	-1875.4	-1868.75	-1843.25	-1788.35
-2LL	3747.8	3770.7	3610.4	3710.9	3750.8	3737.5	3686.5	3576.7
LR statistics	33.7	8.82	236.2	65.5	24.1	43.9	92.4	207.2
n parameters	1	1	1	1	1	1	2	3
p	$<0.001^{***}$	0.003^{**}	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$
AIC	3749.8	3772.7	3612.4	3712.9	3752.8	3739.5	3690.5	3582.7
BIC	3754.5	3777.4	3617.1	3717.6	3757.5	3744.2	3699.9	3596.8

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 16

Goodness-of-fit Summary for Model A – H for Sample Size n = 422 Using Cox Regression

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Goodness-of-fit								
LL	-875.4	-877.1	-835.1	-867.4	-867.85	-871.8	-861.35	-821.05
-2LL	1750.8	1754.2	1670.2	1734.8	1735.7	1743.6	1722.7	1642.1
LR statistics	11.6	8.3	99.2	25.9	21.5	18.8	38.9	120.8
n parameters	1	1	1	1	1	1	2	3
p	0.001^{**}	0.004^{**}	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$
AIC	1752.8	1756.2	1672.2	1736.8	1737.7	1745.6	1726.7	1648.1
BIC	1756.8	1760.2	1676.2	1740.8	1741.7	1749.6	1734.8	1660.2

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 17

Goodness-of-fit Summary for Model A – H Using Discrete-time Survival Regression with Ten Time Periods (n = 9692 for 1577 patients)

Goodness-of-fit	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
LL	-2068.35	-2096.5	-1975.25	-2035.75	-2078.75	-2062.1		-1912.5
-2LL	4136.7	4193	3950.5	4071.5	4157.5	4124.2	4024.3	3825
LR statistics	109.96	53.7	235.6	175.23	89.21	122.5	232.4	421.64
n parameters	11	11	11	11	11	11	12	13
p	$<0.001^{***}$	0.001^{**}	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$
AIC	4158.7	4215	3972.5	4093.5	4179.5	4146.2	4048.3	3851
BIC	4237.7	4294.0	4051.5	4172.5	4258.5	4225.2	4134.4	3944.3

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 18

Goodness-of-fit Summary for Model A – H Using Discrete-time Survival Regression with Ten Time Periods (n = 7516 for 1213 patients)

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Goodness-of-fit								
LL	-1638.8	-1656.85	-1571.15	-1609.1	-1641.95	-1632.85	-1592.65	-1512.95
-2LL	3277.6	3313.7	3142.3	3218.2	3283.9	3265.7	3185.3	3025.9
LR statistics	84.7	48.6	177.7	141.2	78.4	96.5	177	336.4
n parameters	11	11	11	11	11	11	12	13
P	<0.001***	0.002**	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***
AIC	3299.6	3335.7	3164.3	3240.2	3305.9	3287.7	3185.3	3051.9
BIC	3375.8	3411.9	3240.5	3316.4	3382.1	3363.9	3292.4	3141.9

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 19

Goodness-of-fit Summary for Model A – H Using Discrete-time Survival Regression with Ten Time Periods (n = 4939 for 809 patients)

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Goodness-of-fit								
LL	-1126.8	-1139.5	-1077.45	-1110.8	-1128.85	-1122.75	-1098.4	-1043.15
-2LL	2253.6	2279	2154.9	2221.6	2257.7	2245.5	2196.8	2086.3
LR statistics	57.5	32.05	126.1	89.5	53.4	65.6	114.3	224.7
n parameters	11	11	11	11	11	11	12	13
P	<0.001***	0.005**	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***
AIC	2275.6	2301	2176.9	2243.6	2279.7	2267.5	2220.8	2112.3
BIC	2347.2	2372.6	2248.5	2315.2	2351.3	2339.1	2298.9	2196.9

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 20

Goodness-of-fit Summary for Model A – H Using Discrete-time Survival Regression with Ten Time Periods (n = 2490 for 422 patients)

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Goodness-of-fit								
LL	-582.95	-585.3	-554.15	-575.9	-575.4	-580	-570.1	-530.45
-2LL	1165.9	1170.6	1108.3	1151.8	1150.8	1160	1140.2	1060.9
LR statistics	37.92	33.24	71.4	52.05	53.04	43.9	63.6	142.9
n parameters	11	11	11	11	11	11	12	13
p	$<0.001^{***}$	0.009^{**}	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$
AIC	1187.9	1192.6	1130.3	1173.8	1172.8	1182	1164.2	1086.9
BIC	1251.9	1256.6	1194.3	1237.8	1236.8	1246.0	1234.0	1162.6

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 21

Goodness-of-fit Summary for Model A – H Using Discrete-time Survival Regression with Five Time Periods (n = 6041 for 1577 patients)

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Goodness-of-fit								
LL	-962.45	-975.85	-918.65	-956.05	-967.85	-950.7	-936.9	-893.95
-2LL	1924.9	1951.7	1837.3	1912.1	1935.7	1901.4	1873.8	1787.9
LR statistics	109.96	53.7	236	194.9	171.4	205.6	233.3	319.1
n parameters	6	6	6	6	6	6	7	8
p	<0.001***	0.072	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***
AIC	1936.9	1963.7	1849.3	1924.1	1947.7	1913.4	1887.8	1803.9
BIC	1977.1	2003.9	1889.5	1964.3	1987.9	1953.6	1934.7	1857.6

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 22

Goodness-of-fit Summary for Model A – H Using Discrete-time Survival Regression with Five Time Periods (n = 4688 for 1213 patients)

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Goodness-of-fit								
LL	-767.05	-772.6	-734.75	-757.95	-763.4	-757.95		-710.9
-2LL	1534.1	1545.2	1469.5	1515.9	1526.8	1515.9	1494.6	1421.8
LR statistics	131.4	120.4	175.6	149.6	138.8	149.6	170.9	243.8
n parameters	6	6	6	6	6	6	7	8
p	$<0.001^{***}$	0.15	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$
AIC	1546.1	1557.2	1481.5	1527.9	1538.8	1527.9	1508.6	1437.8
BIC	1584.8	1595.9	1520.2	1566.6	1577.5	1566.6	1553.8	1489.4

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 23

Goodness-of-fit Summary for Model A – H Using Discrete-time Survival Regression with Five Time Periods (n = 3097 for 809 patients)

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Goodness-of-fit								
LL	-525.15	-527.85	-506.05	-520.85	-520.35	-517.85	-512.85	-486.95
-2LL	1050.3	1055.7	1012.1	1041.7	1040.7	1035.7	1025.7	973.9
LR statistics	84.2	78.8	108.7	92.8	93.8	98.8	108.8	160.6
n parameters	6	6	6	6	6	6	7	8
p	0.006^{**}	0.173	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$
AIC	1062.3	1067.7	1024.1	1053.7	1052.7	1047.7	1039.7	989.9
BIC	1098.5	1103.9	1060.3	1089.9	1088.9	1083.9	1082.0	1038.2

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 24

Goodness-of-fit Summary for Model A – H Using Discrete-time Survival Regression with Five Time Periods (n = 1576 for 422 patients)

	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Goodness-of-fit								
LL	-676.25	-672.75	-664.45	-676.75	-651.05	-663.15	-662.45	-622.4
-2LL	1352.5	1345.5	1328.9	1353.5	1302.1	1326.3	1324.9	1244.8
LR statistics	46	53	58	44.9	60.4	72.2	73.6	153.7
n parameters	6	6	6	6	6	6	7	8
p	0.261	0.003^{**}	$<0.001^{***}$	0.654	0.012	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$
AIC	1364.5	1357.5	1340.9	1365.5	1314.1	1338.3	1338.9	1260.8
BIC	1396.7	1389.7	1373.1	1397.7	1346.3	1370.5	1376.4	1303.7

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 25

Model G Comparisons between Cox Regression and Discrete-time Survival Model under Different Time Periods (n = 1577)

	Model F					
	Cox Regression Model		Discrete-time Survival Model 5 Time Periods		Discrete-time Survival Model 10 Time Period	
Variable	*HZ*	*95% CI*	*HZ*	95% *CI*	HZ	95% *CI*
High Risk	1.819***	1.535 - 2.155	2.286***	1.761 - 2.969	1.864***	1.562 - 2.224
Nadir Time < 2 yrs	2.575***	2.122 – 3.125	2.086***	1.572 – 2.768	2.618***	2.147 – 3.192
Goodness-of-fit						
-2LL	7304.1		1873.8		4024.3	
n parameters	2		7		12	
AIC	7308.1		1887.8		4048.3	
BIC	7318.8		1934.7		4134.4	

Note. *HZ* = hazard ratio, an *HZ* < 1 indicates a lower risk for the indicator group, *HZ* = 1 no difference between indicator and reference group, *HZ* > 1 indicates a higher risk for the indicator group; CI = confidence interval.

$^{*}p < 0.05$.

$^{**}p < 0.01$.

$^{***}p < 0.001$.

Table 26

Model G Comparisons between Cox Regression and Discrete-time Survival Model under Different Time Periods (n = 1213)

	Model G					
	Cox Regression Model		Discrete-time Survival Model 5 Time Periods		Discrete-time Survival Model 10 Time Period	
Variable	*HZ*	*95% CI*	*HZ*	95% *CI*	HZ	95% *CI*
High Risk	1.751^{***}	1.448 – 2.117	2.004^{***}	1.495 – 2.685	1.792^{***}	1.469 – 2.184
Nadir Time < 2 yrs	2.597^{***}	2.091 – 3.227	2.059^{***}	1.502 – 2.822	2.635^{***}	2.109 – 3.291
Goodness-of-fit						
-2LL	5599.7		1494.6		3185.3	
n parameters	2		7		12	
AIC	5603.7		1508.6		3209.3	
BIC	5613.9		1553.8		3292.4	

Note. *HZ* = hazard ratio, an *HZ* < 1 indicates a lower risk for the indicator group, *HZ* = 1 no difference between indicator and reference group, *HZ* > 1 indicates a higher risk for the indicator group; CI = confidence interval.

$^{*}p < 0.05$.

$^{**}p < 0.01$.

$^{***}p < 0.001$.

Table 27

Model G Comparisons between Cox Regression and Discrete-time Survival Model under Different Time Periods (n = 809)

	Model F					
	Cox Regression Model		Discrete-time Survival Model 5 Time Periods		Discrete-time Survival Model 10 Time Period	
Variable	*HZ*	*95% CI*	*HZ*	95% *CI*	HZ	95% *CI*
High Risk	1.781^{***}	1.418 – 2.237	2.063^{***}	1.450 – 2.936	1.833^{***}	1.446 – 2.324
Nadir Time < 2 yrs	2.433^{***}	1.883 – 3.143	1.803^{***}	1.242 – 2.617	2.453^{***}	1.886 – 3.191
Goodness-of-fit						
-2LL	3686.5		1025.7		2196.8	
n parameters	2		7		12	
AIC	3690.5		1039.7		2220.8	
BIC	3699.9		1082.0		2298.9	

Note. *HZ* = hazard ratio, an *HZ* < 1 indicates a lower risk for the indicator group, *HZ* = 1 no difference between indicator and reference group, *HZ* > 1 indicates a higher risk for the indicator group; CI = confidence interval.

$^{*}p < 0.05$.

$^{**}p < 0.01$.

$^{***}p < 0.001$.

Table 28

Model G Comparisons between Cox Regression and Discrete-time Survival Model under Different Time Periods (n = 422)

	Model G					
	Cox Regression Model		Discrete-time Survival Model 5 Time Periods		Discrete-time Survival Model 10 Time Period	
Variable	*HZ*	*95% CI*	*HZ*	95% *CI*	HZ	95% *CI*
High Risk	1.744^{***}	1.276 – 2.384	2.161^{***}	1.629 – 2.866	1.764^{**}	1.273 – 2.446
Nadir Time < 2 yrs	2.158^{***}	1.532 – 3.039	0.841	0.629 – 1.124	2.169^{***}	1.524 – 3.087
Goodness-of-fit						
-2LL	1722.7		1324.9		1140.2	
n parameters	2		7		12	
AIC	1726.7		1338.9		1164.2	
BIC	1734.8		1376.4		1234.0	

Note. *HZ* = hazard ratio, an *HZ* < 1 indicates a lower risk for the indicator group, *HZ* = 1 no difference between indicator and reference group, *HZ* > 1 indicates a higher risk for the indicator group; CI = confidence interval.

$^{*}p < 0.05$.

$^{**}p < 0.01$.

$^{***}p < 0.001$.

Table 29

-2loglikelihood Comparisons for Nested Model H (n = 1577)

		Discrete-time Survival Model 5 Time Periods	Discrete-time Survival Model 10 Time Periods
Null	Time Variables	1953.8	4203.1
Model F	Time variables + Risk Group	1901.4	4124.2
Model G	Time variables + Risk Group + Nadir Time Group	1873.8	4024.3
Model H	Time variables + Risk Group + Nadir Time Group + If the patient received HT before RT	1787.9	3825.0

Table 30

-2loglikelihood Comparisons for Nested Model H (n = 1213)

		Discrete-time Survival Model 5 Time Periods	Discrete-time Survival Model 10 Time Periods
Null	Time Variables	1547.2	3322.6
Model F	Time variables + Risk Group	1515.9	3265.7
Model G	Time variables + Risk Group + Nadir Time Group	1494.6	3185.3
Model H	Time variables + Risk Group + Nadir Time Group + If the patient received HT before RT	1421.8	3025.9

Table 31

-2loglikelihood Comparisons for Nested Model H (n = 809)

		Discrete-time Survival Model 5 Time Periods	Discrete-time Survival Model 10 Time Periods
Null	Time Variables	1057.5	2286.3
Model F	Time variables + Risk Group	1035.7	2245.5
Model G	Time variables + Risk Group + Nadir Time Group	1025.7	2196.8
Model H	Time variables + Risk Group + Nadir Time Group + If the patient received HT before RT	973.9	2086.3

Table 32

-2loglikelihood Comparisons for Nested Model H (n = 422)

		Discrete-time Survival Model 5 Time Periods	Discrete-time Survival Model 10 Time Periods
Null	Time Variables	1176.0	1353.7
Model F	Time variables + Risk Group	1160.0	1326.3
Model G	Time variables + Risk Group + Nadir Time Group	1140.2	1324.9
Model H	Time variables + Risk Group + Nadir Time Group + If the patient received HT before RT	1060.9	1244.8

Table 33

Model F Comparisons between Cox Regression and Discrete-time Survival Model under Different Time Periods (n = 1577)

	Model F					
	Cox Regression Model		Discrete-time Survival Model 5 Time Periods		Discrete-time Survival Model 10 Time Period	
Variable	*HZ*	*95% CI*	*HZ*	95% *CI*	HZ	95% *CI*
High Risk	2.163^{***}	1.830 - 2.556	2.208^{***}	1.857 - 2.626	2.614	2.023 - 3.379
Goodness-of-fit						
-2LL	7406.6		4124.2		1901.4	
n parameters	1		11		6	
AIC	7408		4136		1923	
BIC	7413.4		4179.1		1996.8	

Note. *HZ* = hazard ratio, an *HZ* < 1 indicates a lower risk for the indicator group, *HZ* = 1 no difference between indicator and reference group, *HZ* > 1 indicates a higher risk for the indicator group; CI = confidence interval.

$^{*}p < 0.05$.

$^{**}p < 0.01$.

$^{***}p < 0.001$.

Table 34

Model F Comparisons between Cox Regression and Discrete-time Survival Model under Different Time Periods (n = 1213)

	Model F					
	Cox Regression Model		Discrete-time Survival Model 5 Time Periods		Discrete-time Survival Model 10 Time Period	
Variable	*HZ*	*95% CI*	*HZ*	95% *CI*	HZ	95% *CI*
High Risk	2.087^{***}	1.731 - 2.517	2.285^{***}	1.714 - 3.045	2.128^{***}	1.752 - 2.584
Goodness-of-fit						
-2LL	5740.8		1515.9		3265.7	
n parameters	1		6		11	
AIC	5742.8		1527.9		3287.7	
BIC	5747.9		1566.6		3363.9	

Note. *HZ* = hazard ratio, an *HZ* < 1 indicates a lower risk for the indicator group, *HZ* = 1 no difference between indicator and reference group, *HZ* > 1 indicates a higher risk for the indicator group; CI = confidence interval.

$^{*}p < 0.05$.

$^{**}p < 0.01$.

$^{***}p < 0.001$.

Table 35

Model F Comparisons between Cox Regression and Discrete-time Survival Model under Different Time Periods (n = 809)

	Model F					
	Cox Regression Model		Discrete-time Survival Model		Discrete-time Survival Model	
			5 Time Periods		10 Time Period	
Variable	*HZ*	*95% CI*	*HZ*	95% *CI*	HZ	95% *CI*
High Risk	2.097***	1.676 - 2.623	2.285***	1.714 - 3.045	1.668***	1.230 - 2.262
Goodness-of-fit						
-2LL	3737.5		1515.9		1248.1	
n parameters	1		6		11	
AIC						
BIC						

Note. *HZ* = hazard ratio, an *HZ* < 1 indicates a lower risk for the indicator group, *HZ* = 1 no difference between indicator and reference group, *HZ* > 1 indicates a higher risk for the indicator group; CI = confidence interval.

$^{*}p < 0.05$.

$^{**}p < 0.01$.

$^{***}p < 0.001$.

Table 36

Model F Comparisons between Cox Regression and Discrete-time Survival Model under Different Time Periods (n = 422)

	Model F					
	Cox Regression Model		Discrete-time Survival Model 5 Time Periods		Discrete-time Survival Model 10 Time Period	
Variable	*HZ*	*95% CI*	*HZ*	95% *CI*	HZ	95% *CI*
High Risk	1.958***	1.437 - 2.667	2.111***	1.596 – 2.793	1.970***	1.428 – 2.718
Goodness-of-fit						
-2LL	1743.6		1326.3		1159.9	
n parameters	1		6		11	
AIC	1745.6		1338.3		1181.9	
BIC	1749.6		1370.5		1245.9	

Note. *HZ* = hazard ratio, an *HZ* < 1 indicates a lower risk for the indicator group, *HZ* = 1 no difference between indicator and reference group, *HZ* > 1 indicates a higher risk for the indicator group; CI = confidence interval.

$^{*}p < 0.05$.

$^{**}p < 0.01$.

$^{***}p < 0.001$.

Figure 30

Estimated Survival Function in Model G by Using Cox regression

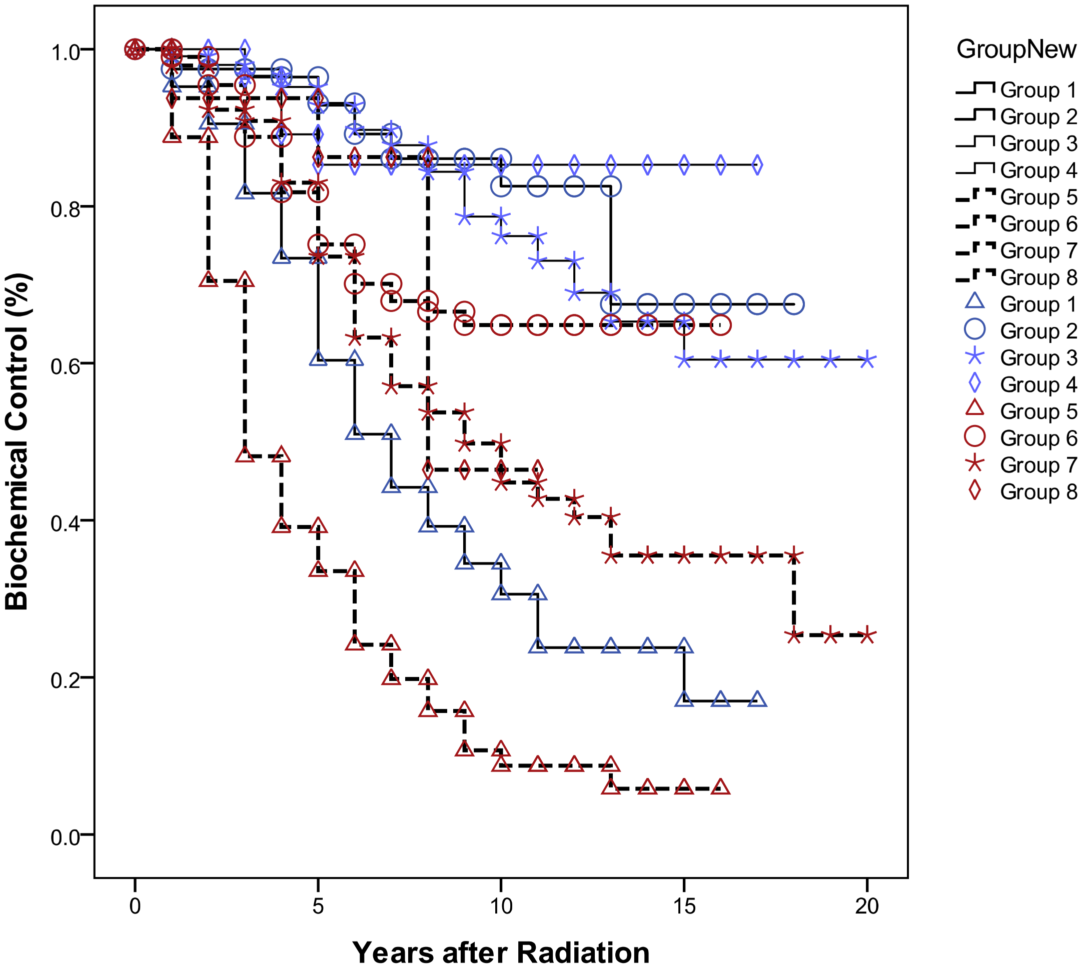


Figure 31

Hazard Comparisons in Discrete-time Model H by Different Sample Sizes

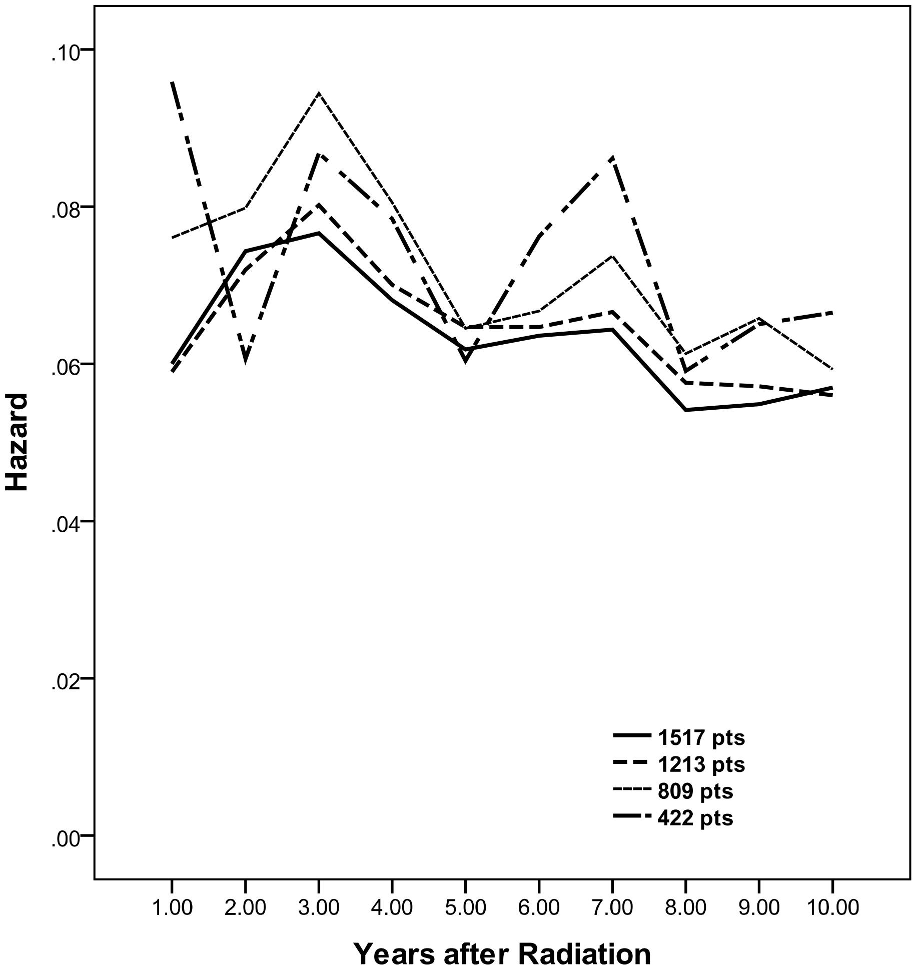


Figure 32

Odds Comparisons in Discrete-time Model H by Different Sample Sizes

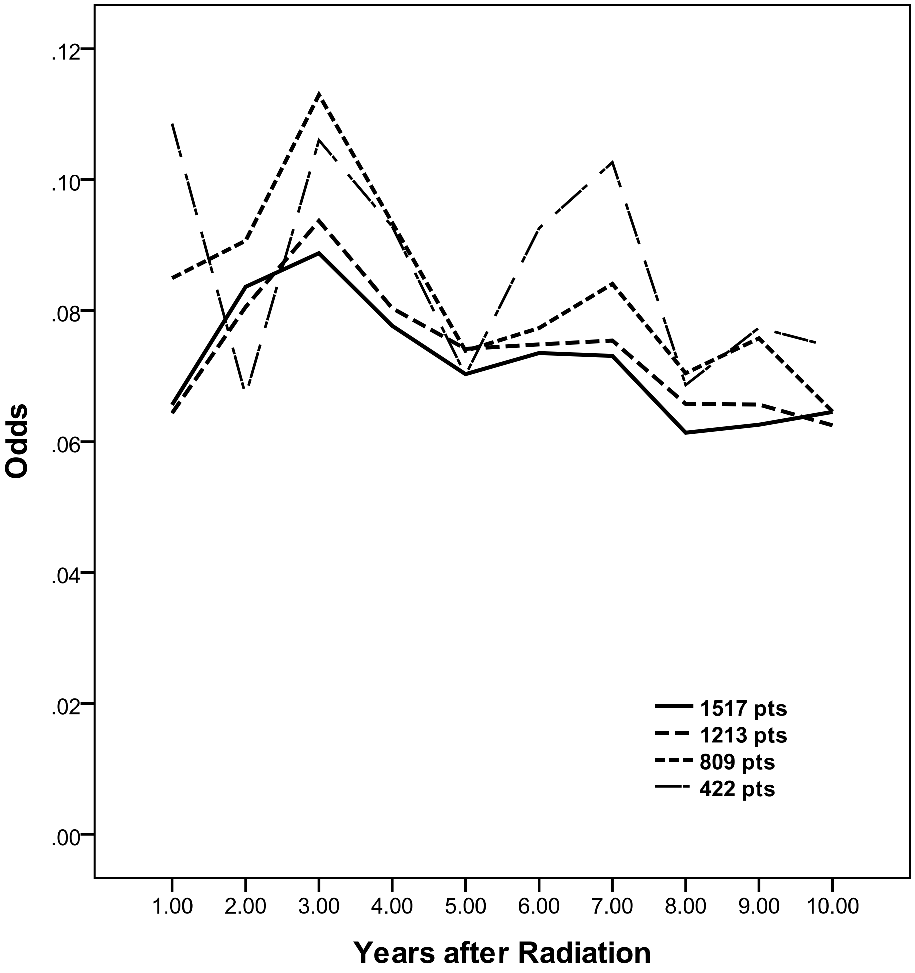


Figure 33

Logit Hazard Comparisons in Discrete-time Model H by Different Sample Sizes

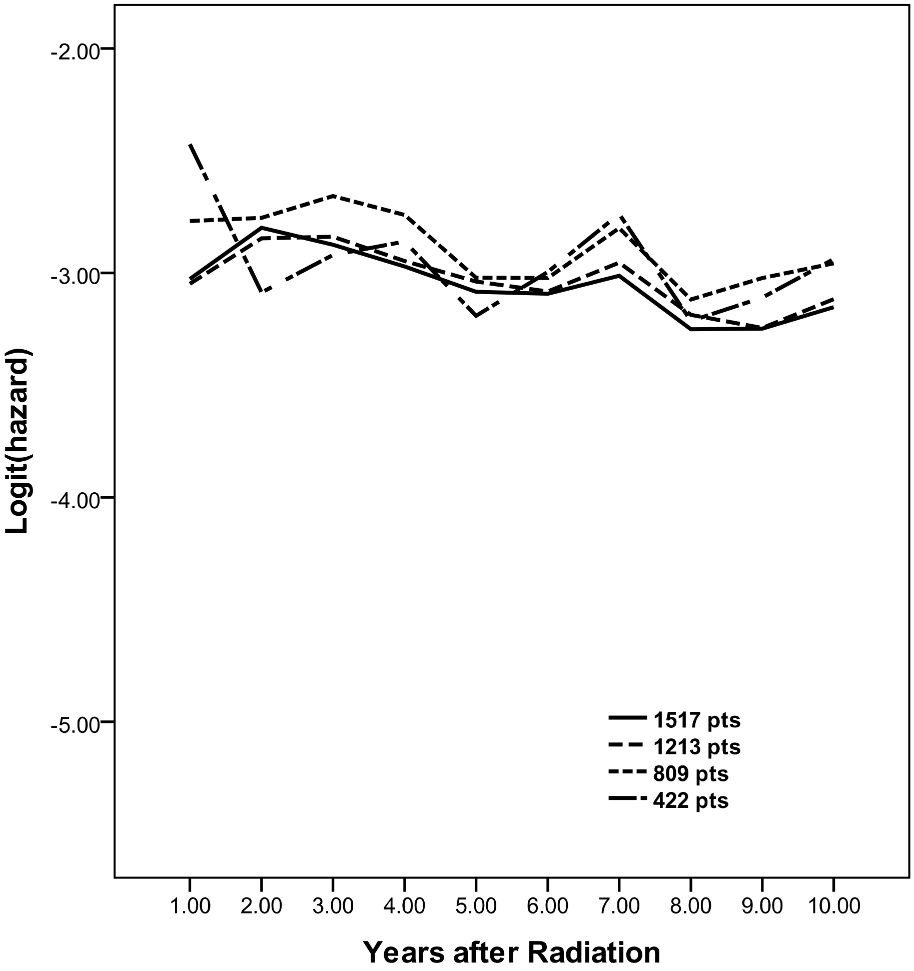


Figure 34

Hazard Comparisons of Model F/G/H by Using Discrete-time Method with Five Time Period under Sample Size 1213

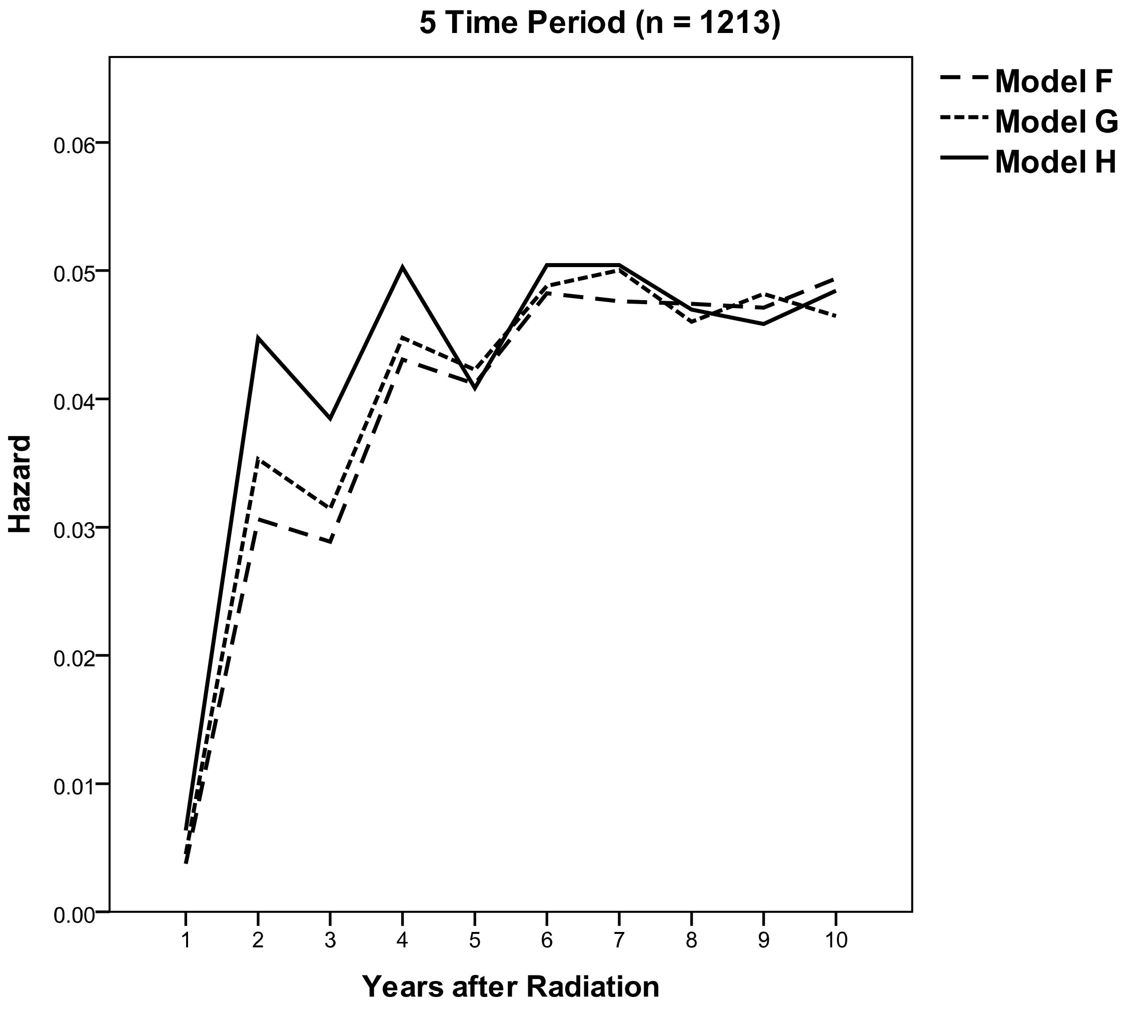


Figure 35

Odds Comparisons of Model F/G/H by Using Discrete-time Method with Five Time Period under Sample Size 1213

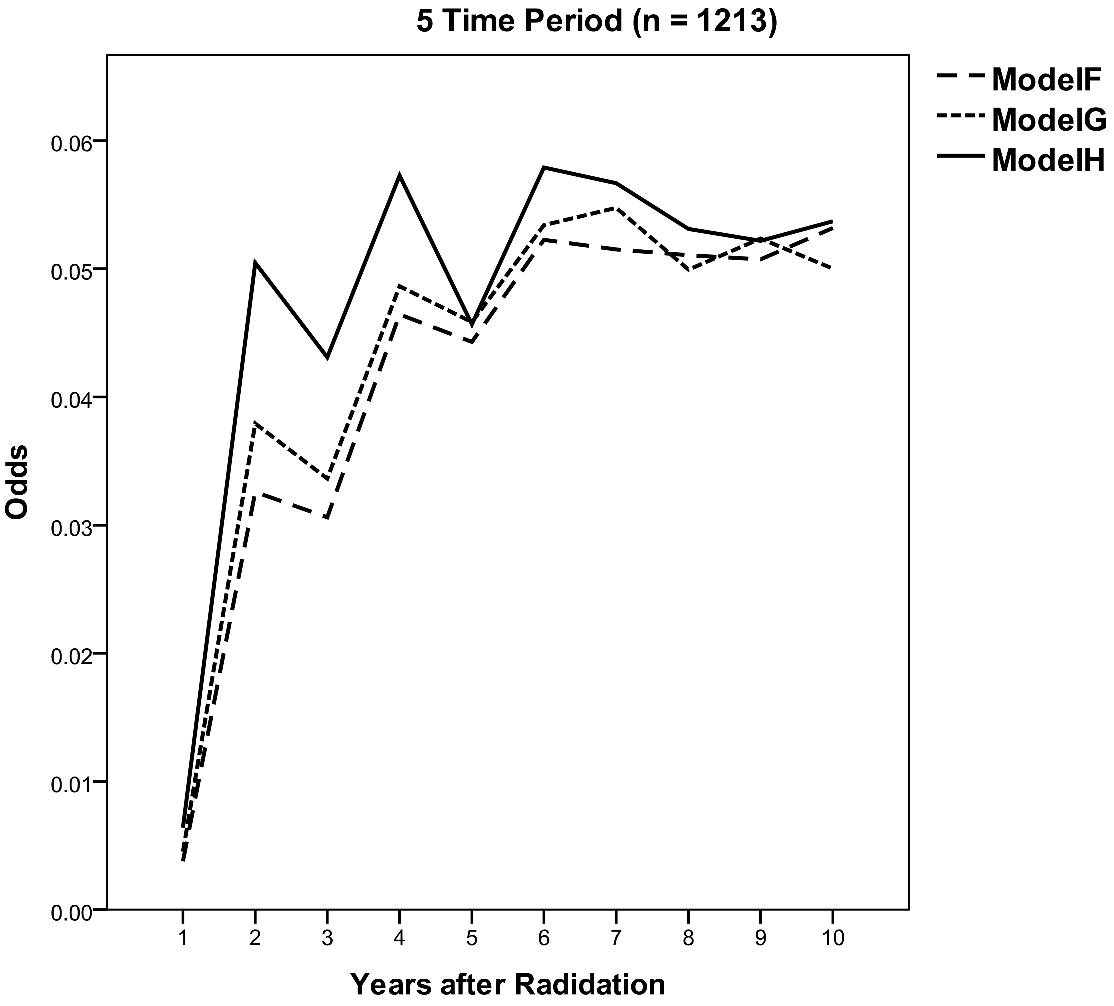


Figure 36

Logit Hazard Comparisons of Model F/G/H by Using Discrete-time Method with Five Time Period under Sample Size 1213

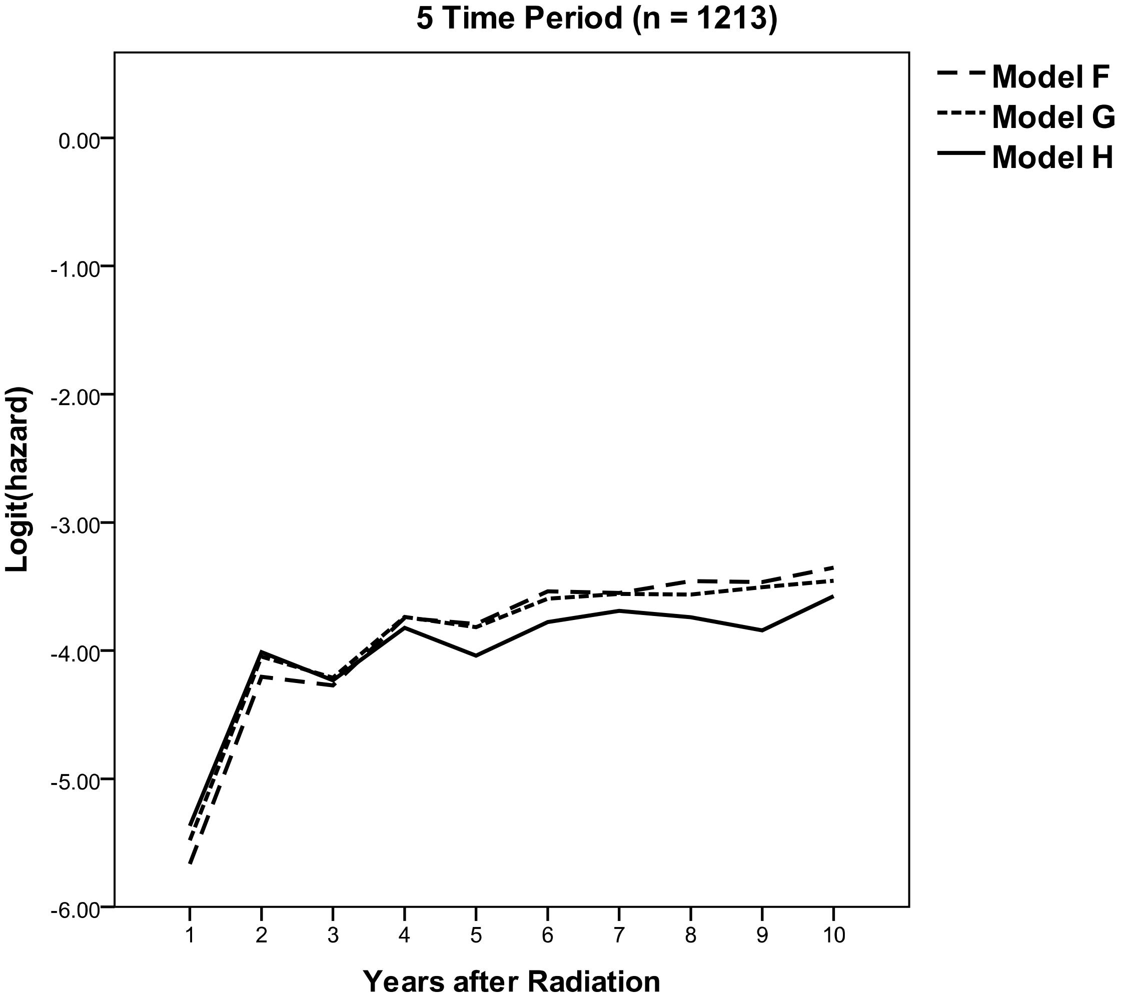


Figure 37

Hazard Comparisons of Model F/G/H by Using Discrete-time Method with Ten Time Period under Sample Size 1213

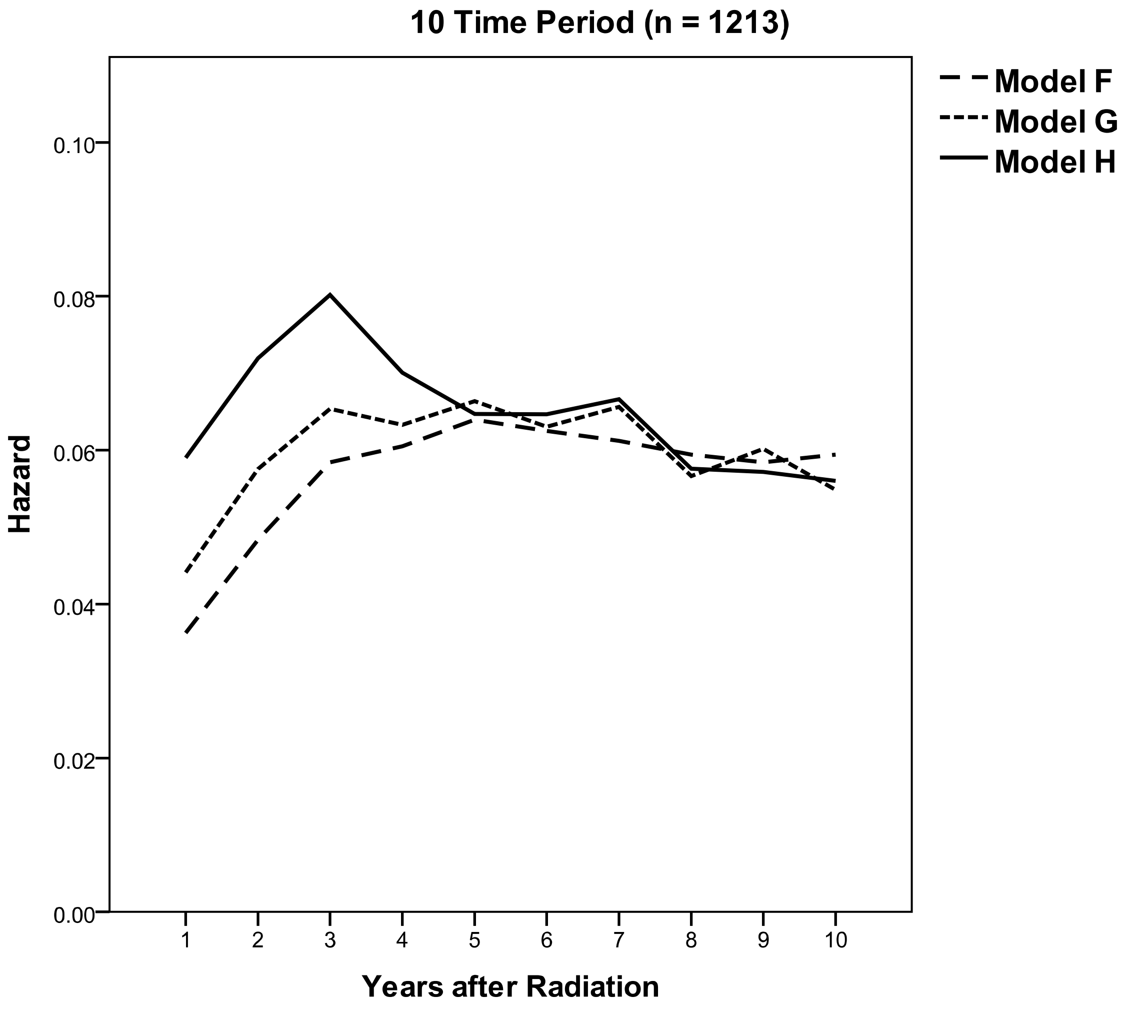


Figure 38

Odds Comparisons of Model F/G/H by Using Discrete-time Method with Ten Time Period under Sample Size 1213

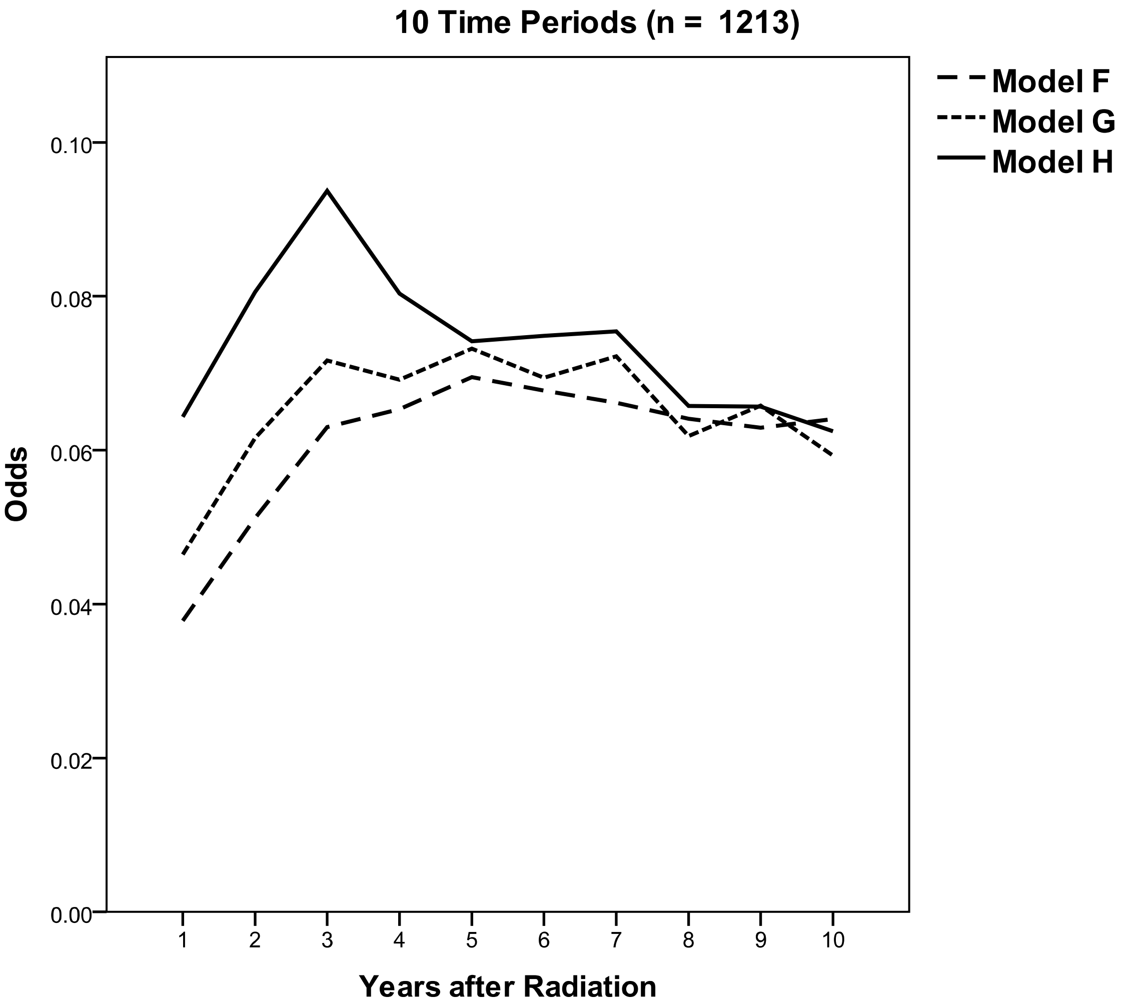


Figure 39

Logit Hazard Comparisons of Model F/G/H by Using Discrete-time Method with Ten Time Period under Sample Size 1213

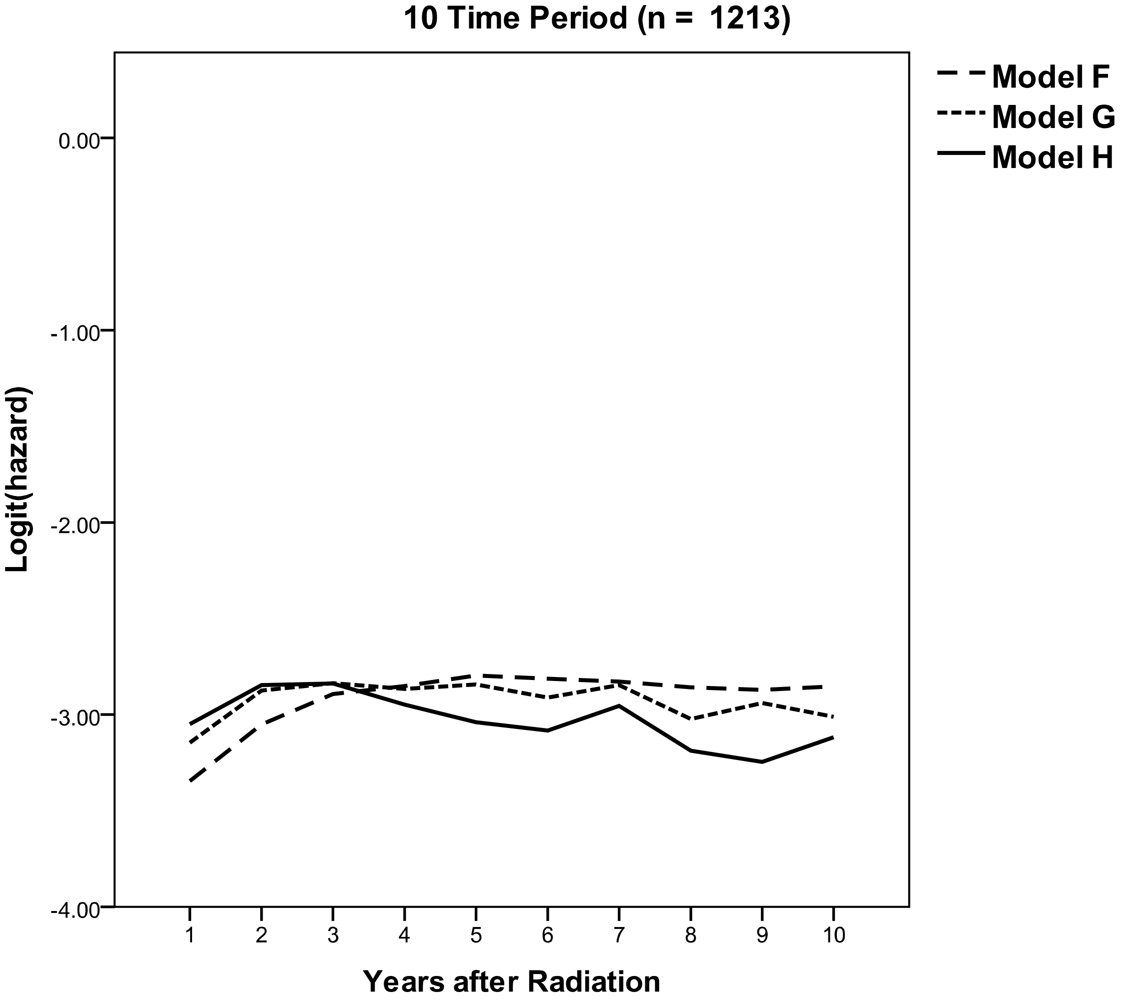


Figure 40

Hazard Comparisons of Model F/G/H by Using Discrete-time Method with Five Time Period under Sample Size 809

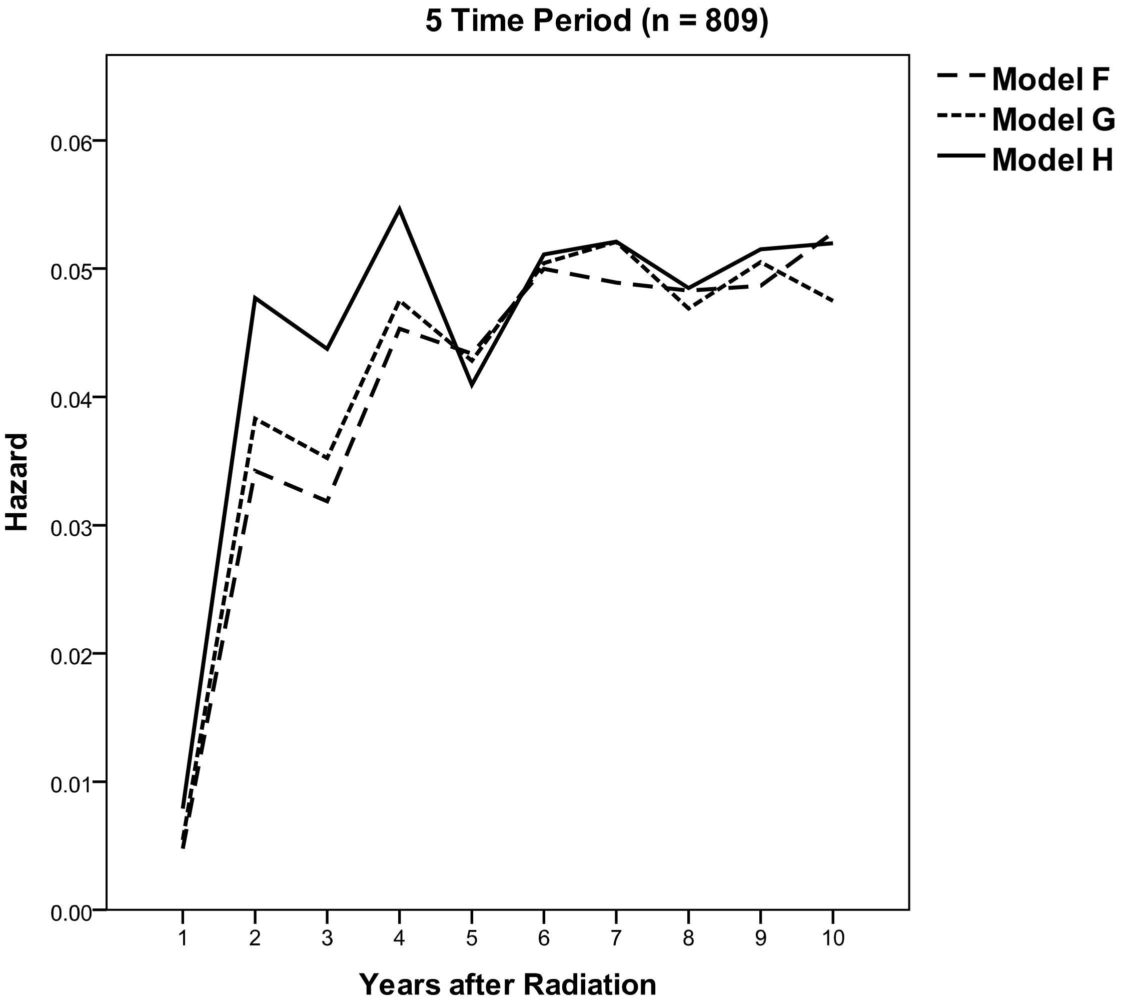


Figure 41

Odds Comparisons of Model F/G/H by Using Discrete-time Method with Five Time Period under Sample Size 809

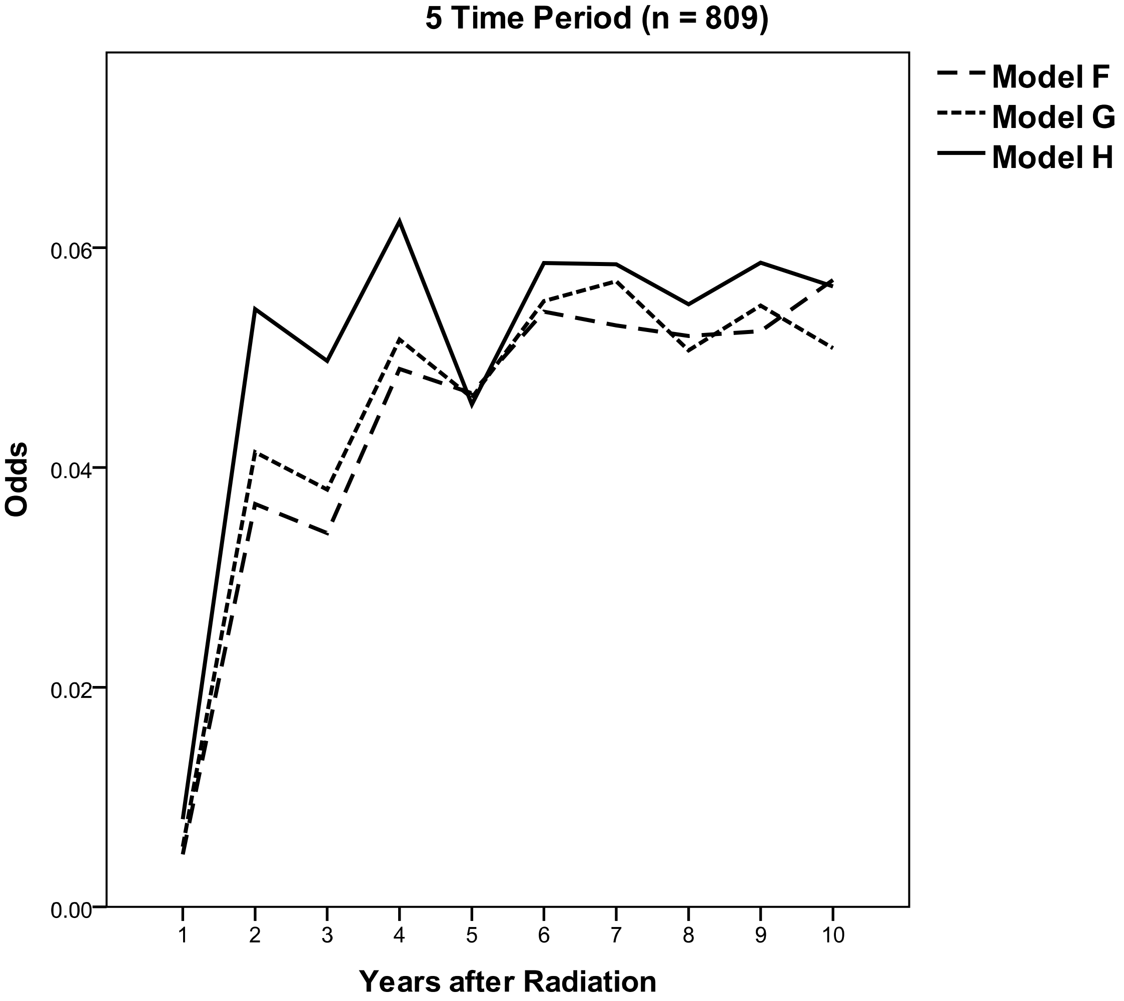


Figure 42

Logit Hazard Comparisons of Model F/G/H by Using Discrete-time Method with Five Time Period under Sample Size 809

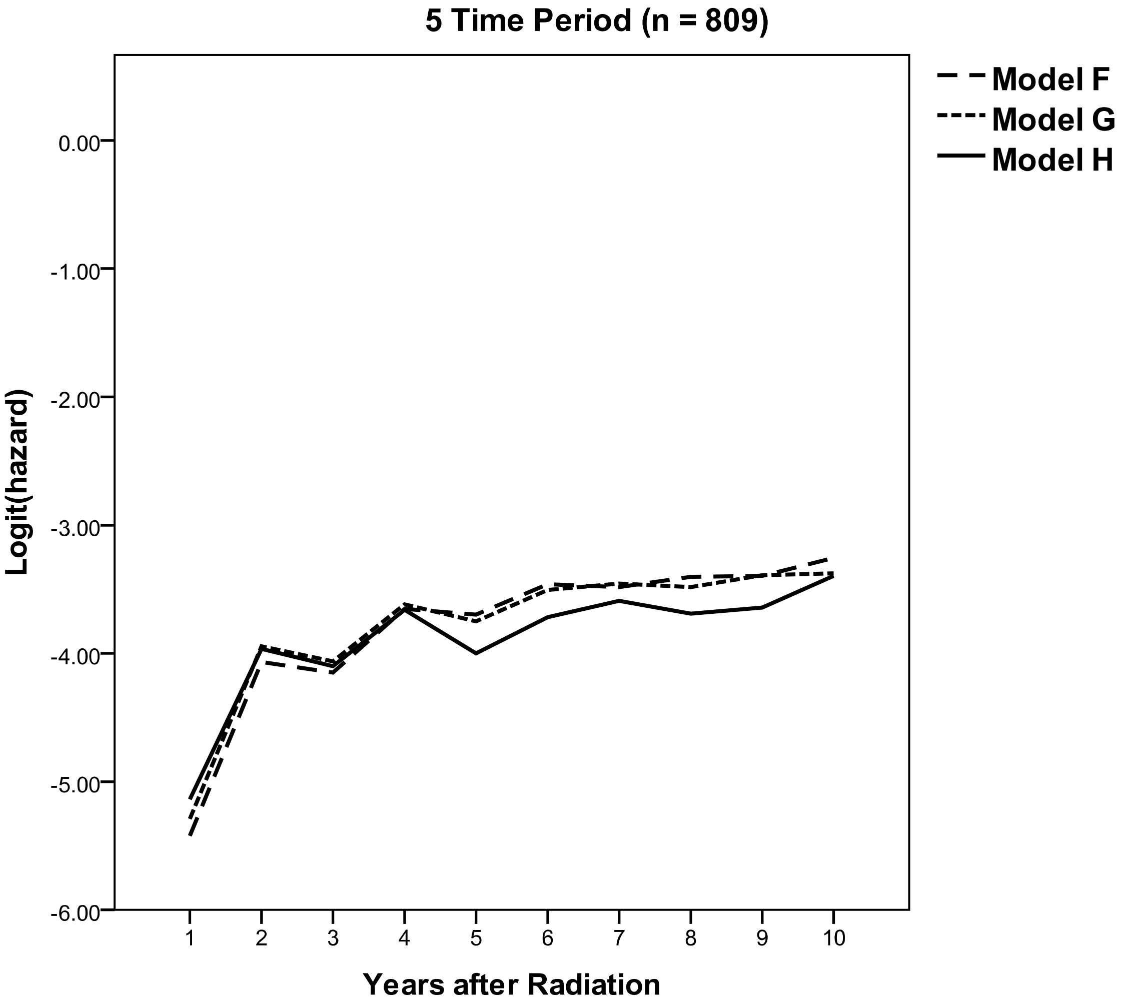


Figure 43

Hazard Comparisons of Model F/G/H by Using Discrete-time Method with Ten Time Period under Sample Size 809

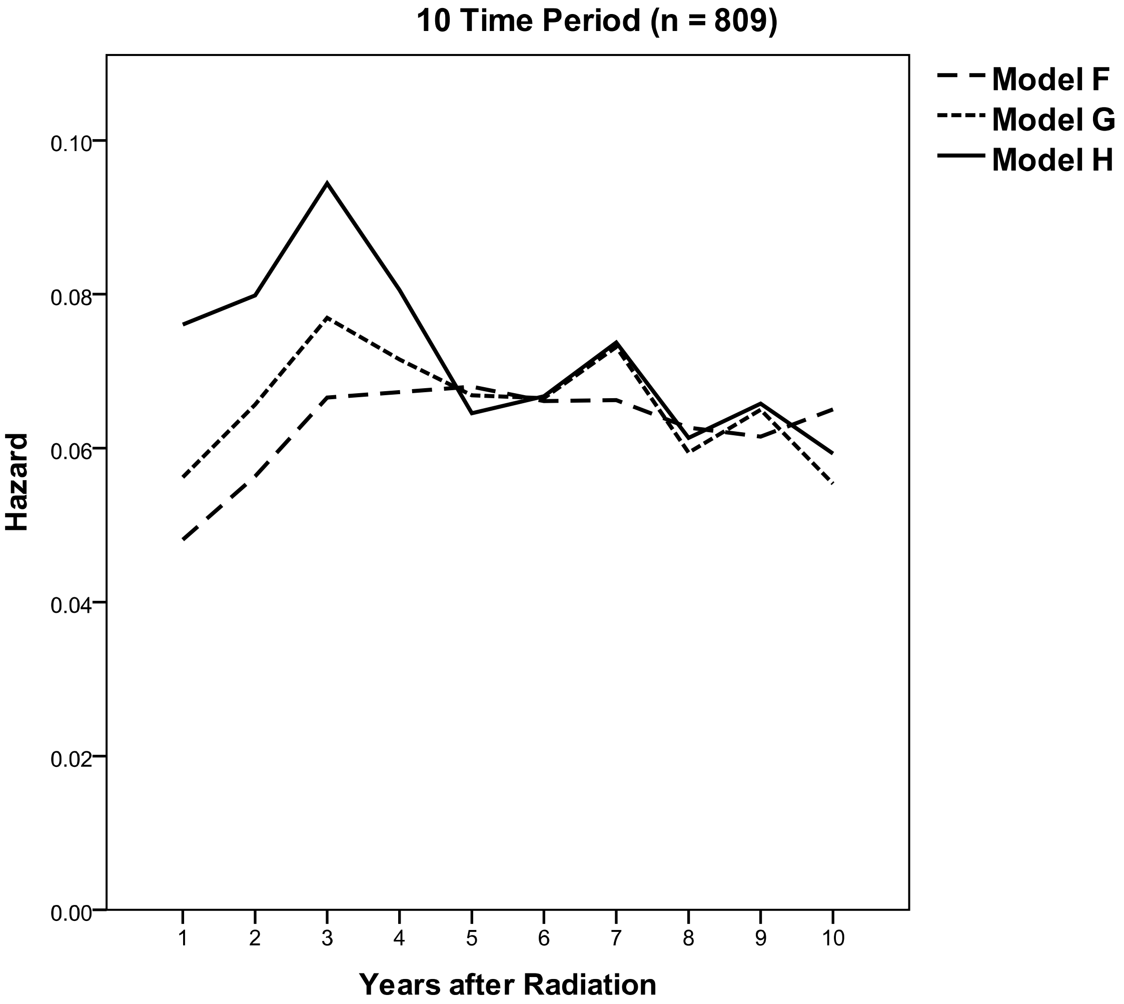


Figure 44

Odds Comparisons of Model F/G/H by Using Discrete-time Method with Ten Time Period under Sample Size 809

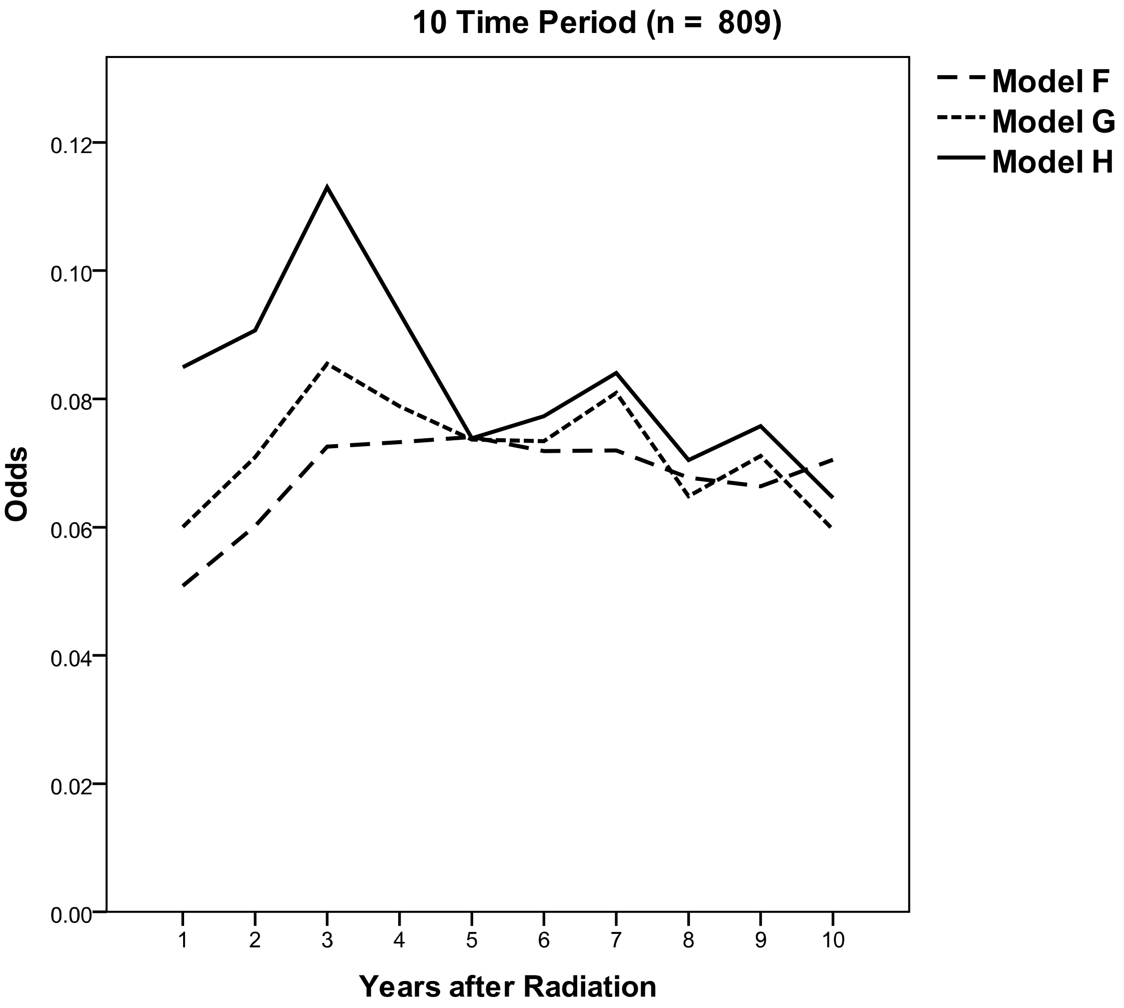


Figure 45

Logit Hazard Comparisons of Model F/G/H by Using Discrete-time Method with Ten Time Period under Sample Size 809

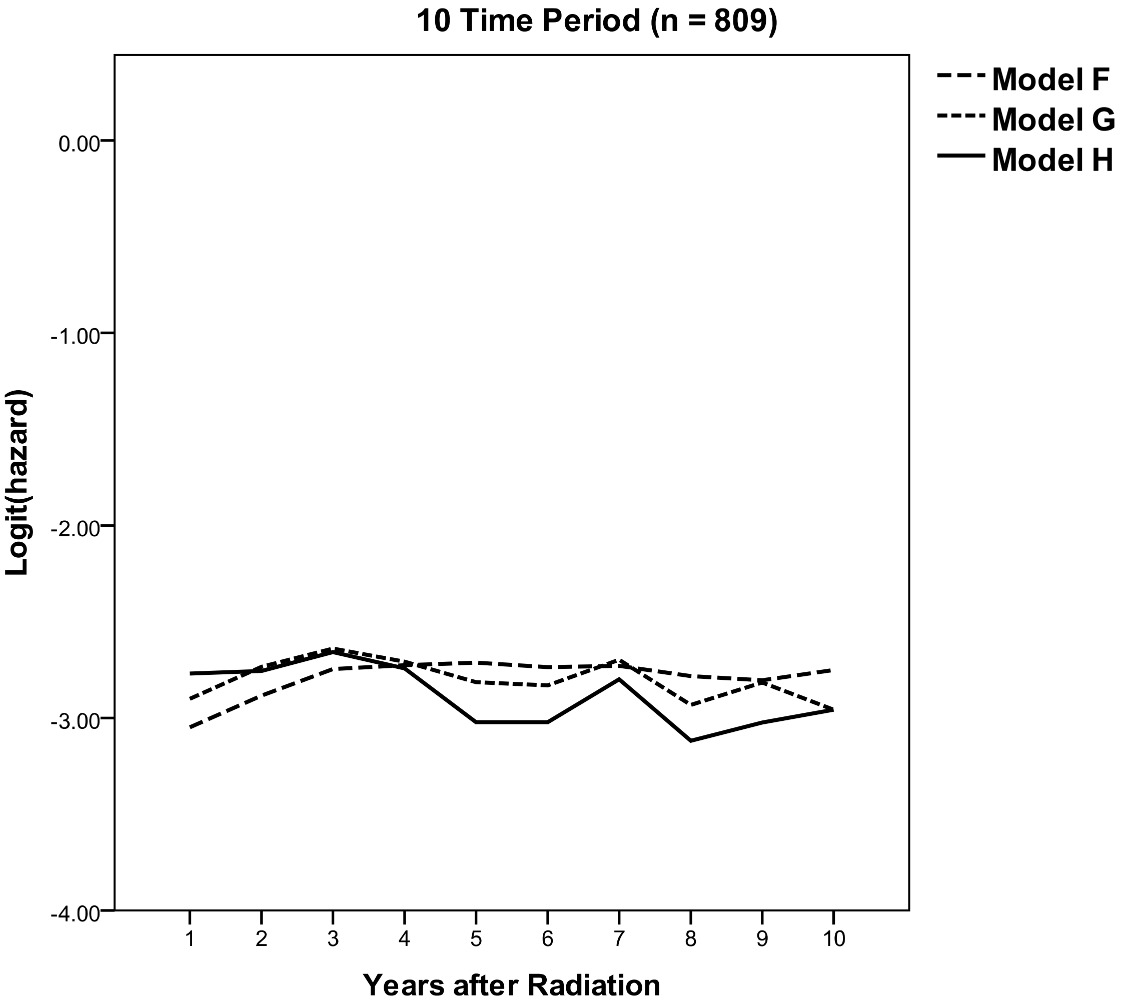


Figure 46

Hazard Comparisons of Model F/G/H by using Discrete-time Method with Five Time Period under Sample Size 422

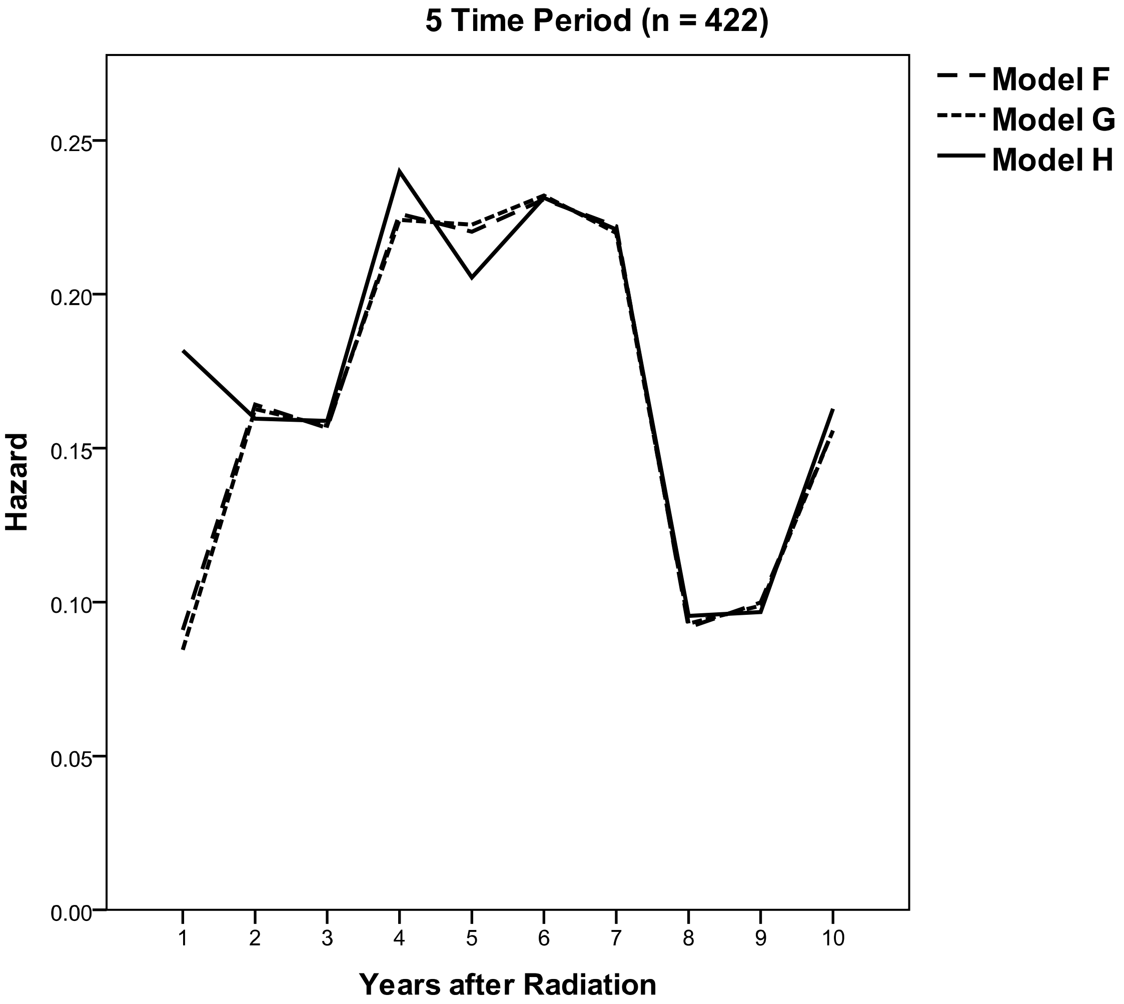


Figure 47

Odds Comparisons of Model F/G/H by using Discrete-time Method with Five Time Period under Sample Size 422

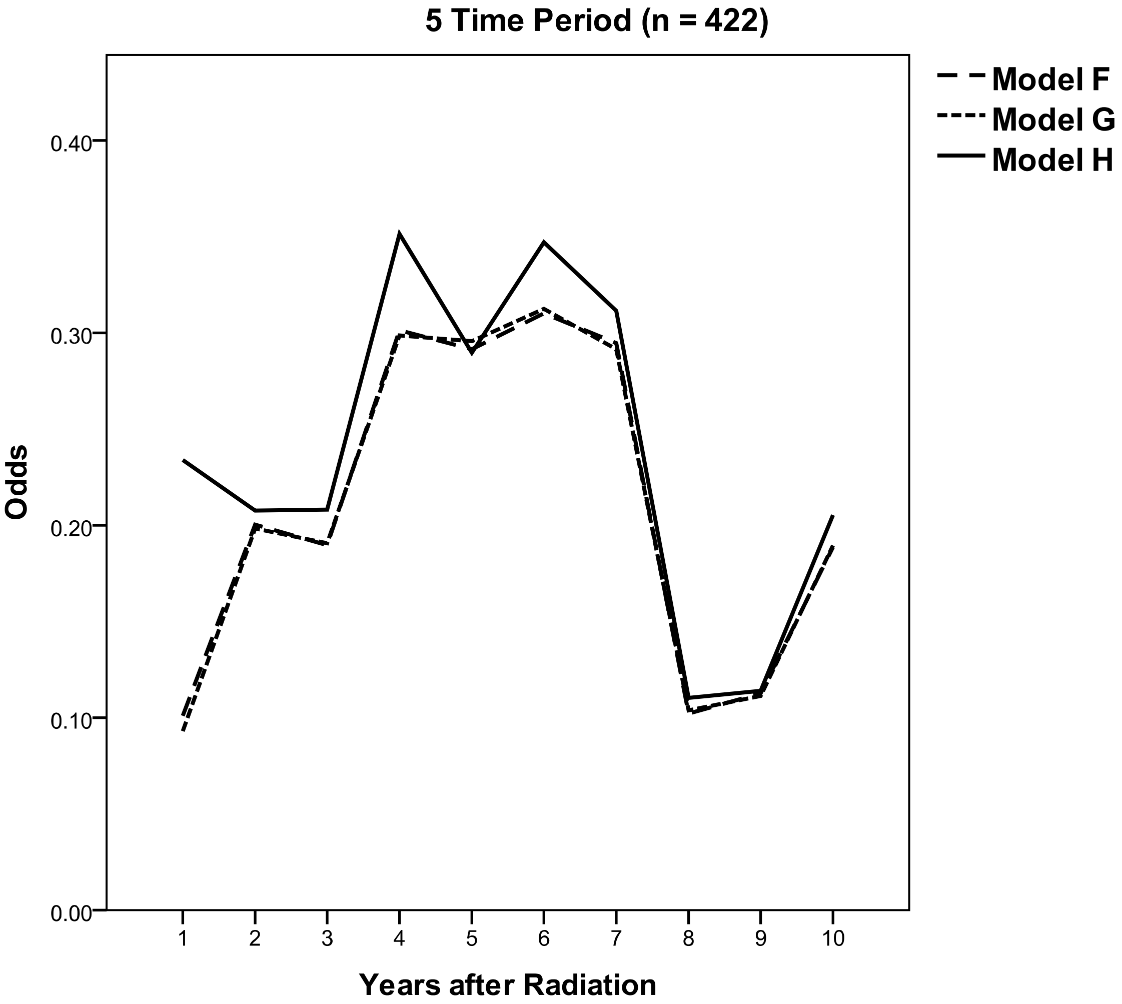


Figure 48

Logit Hazard Comparisons of Model F/G/H by Using Discrete-time Method with Five Time Period under Sample Size 422

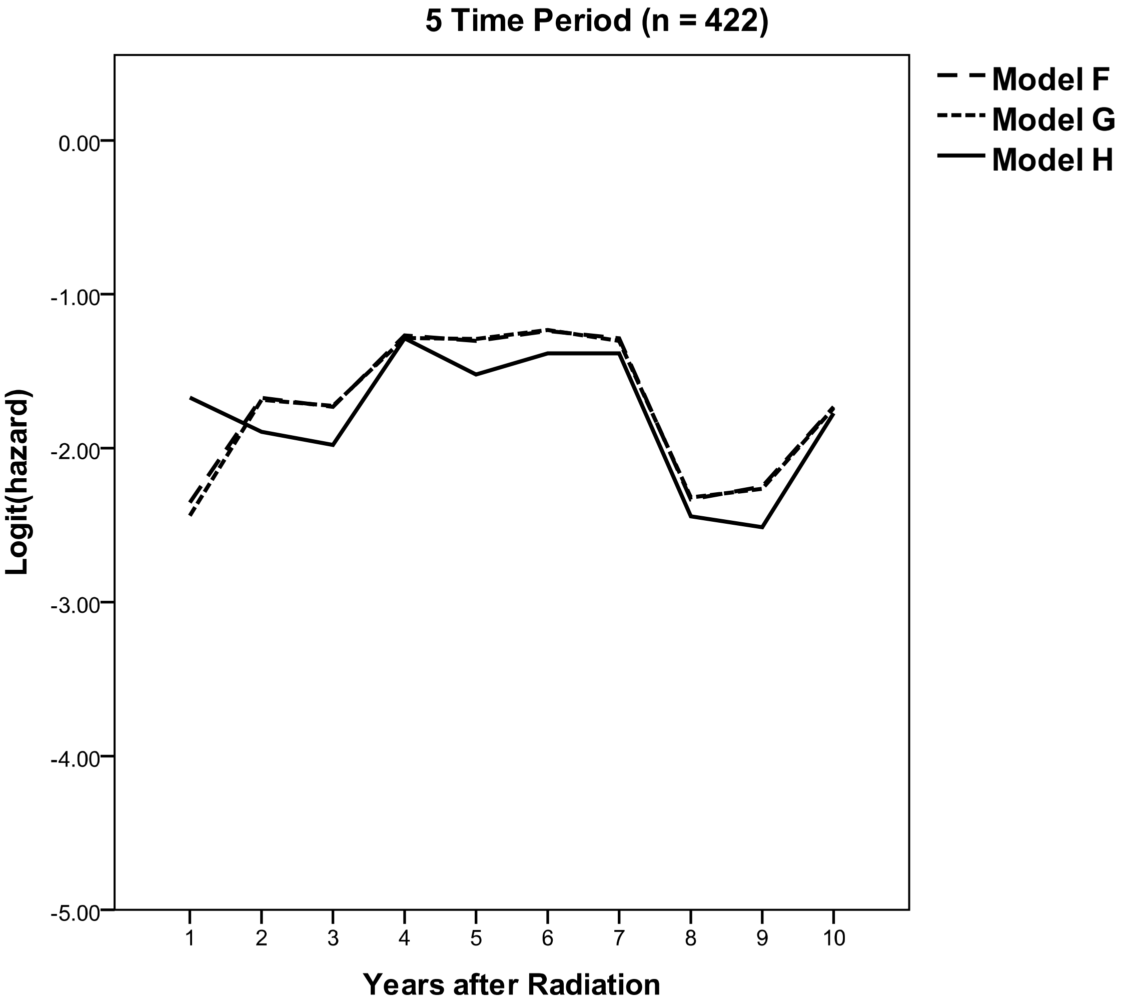


Figure 49

Hazard Comparisons of Model F/G/H by Using Discrete-time Method with Ten Time Period under Sample Size 422

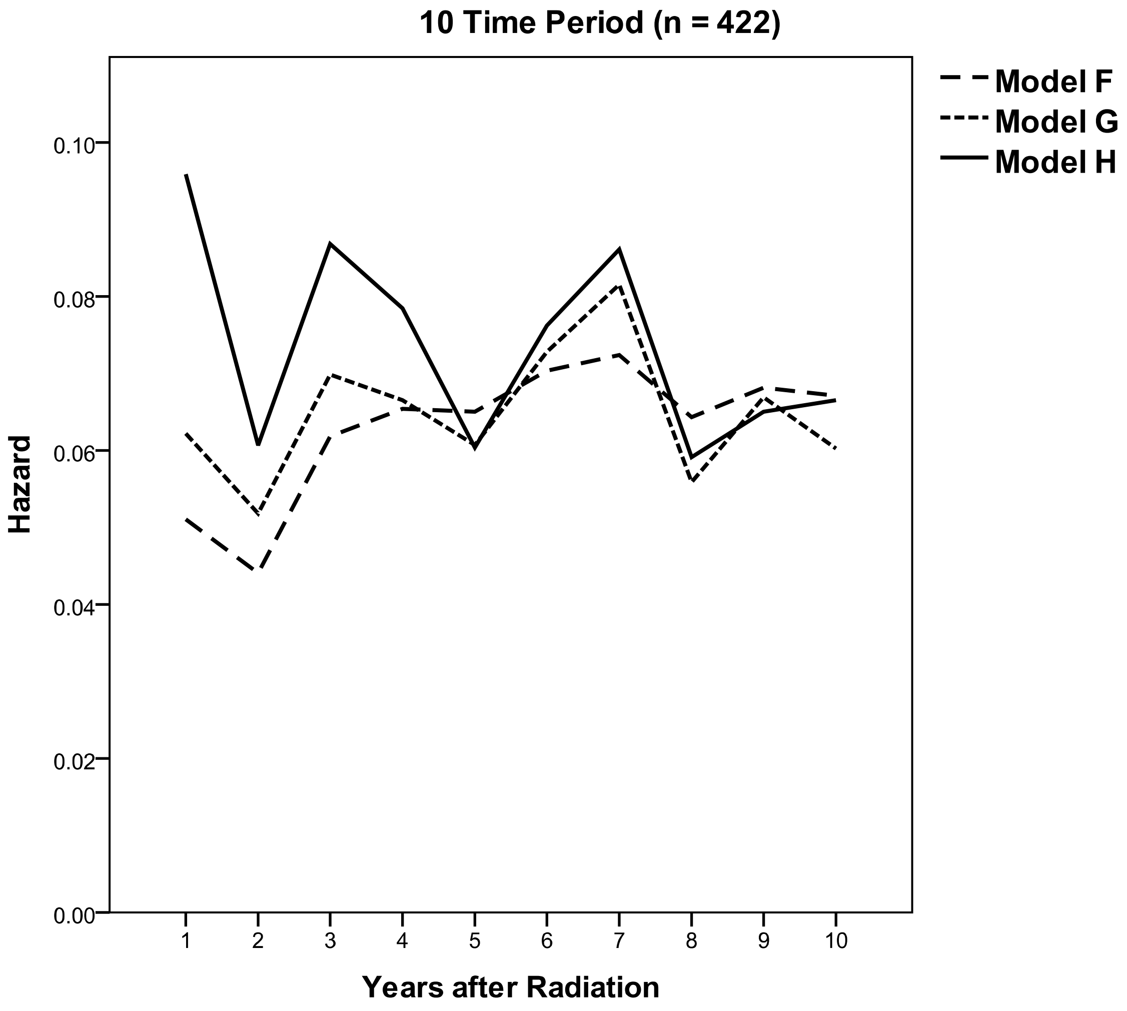


Figure 50

Odds Comparisons of Model F/G/H by Using Discrete-time Method with Ten Time Period under Sample Size 422

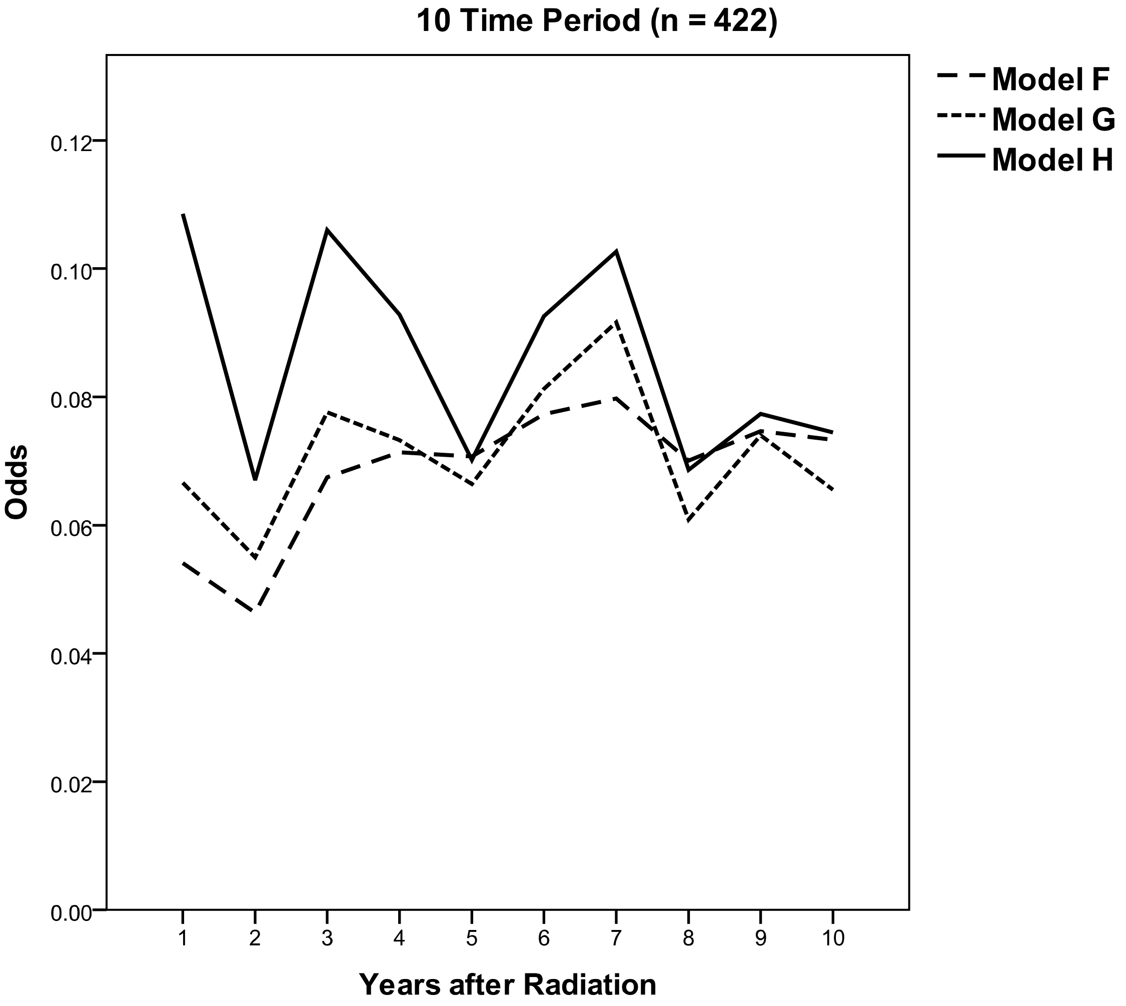


Figure 51

Logit Hazard Comparisons of Model F/G/H by using Discrete-time Method with Ten Time Period under Sample Size 422

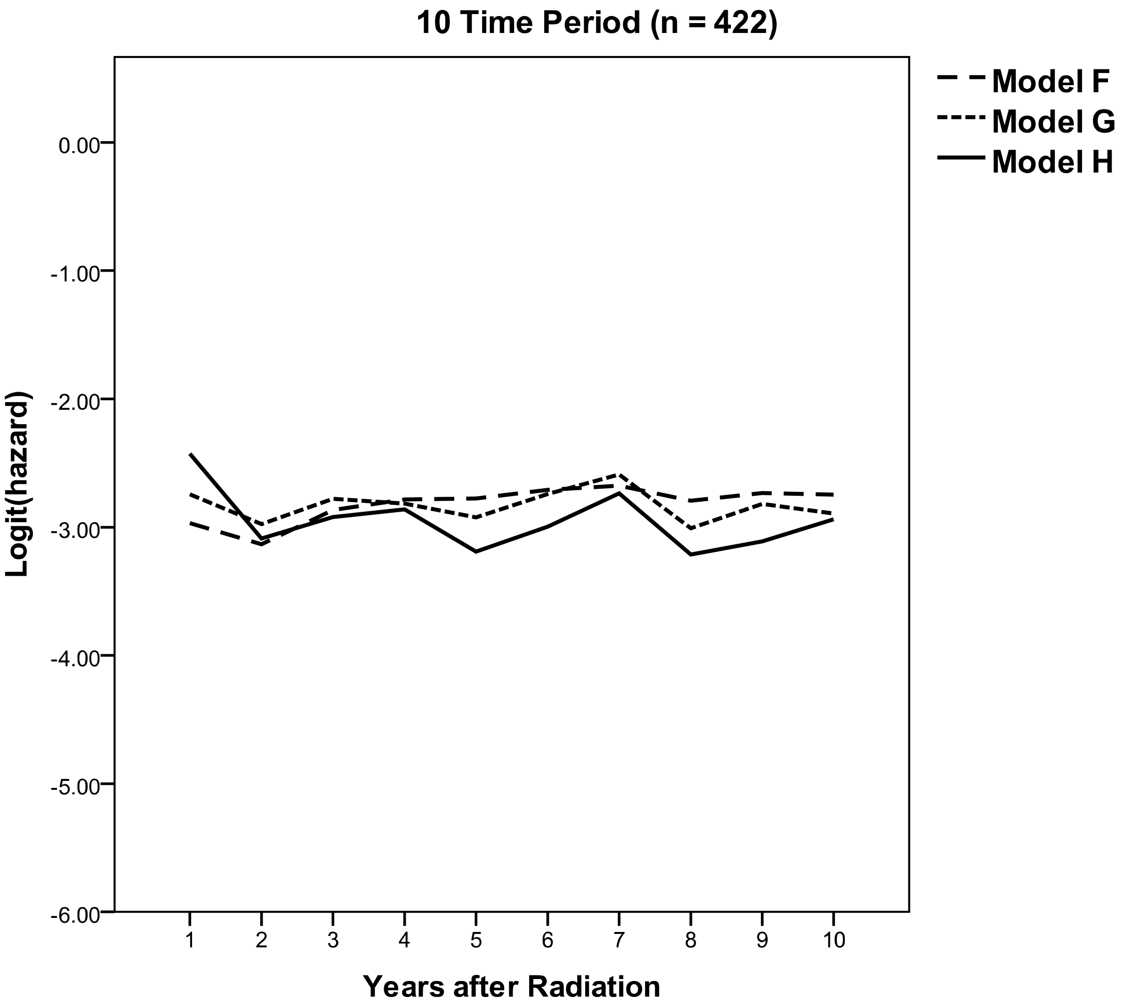


Figure 52

Hazard Comparisons in Discrete-time Model F by Different Time Period Under Sample Size 422

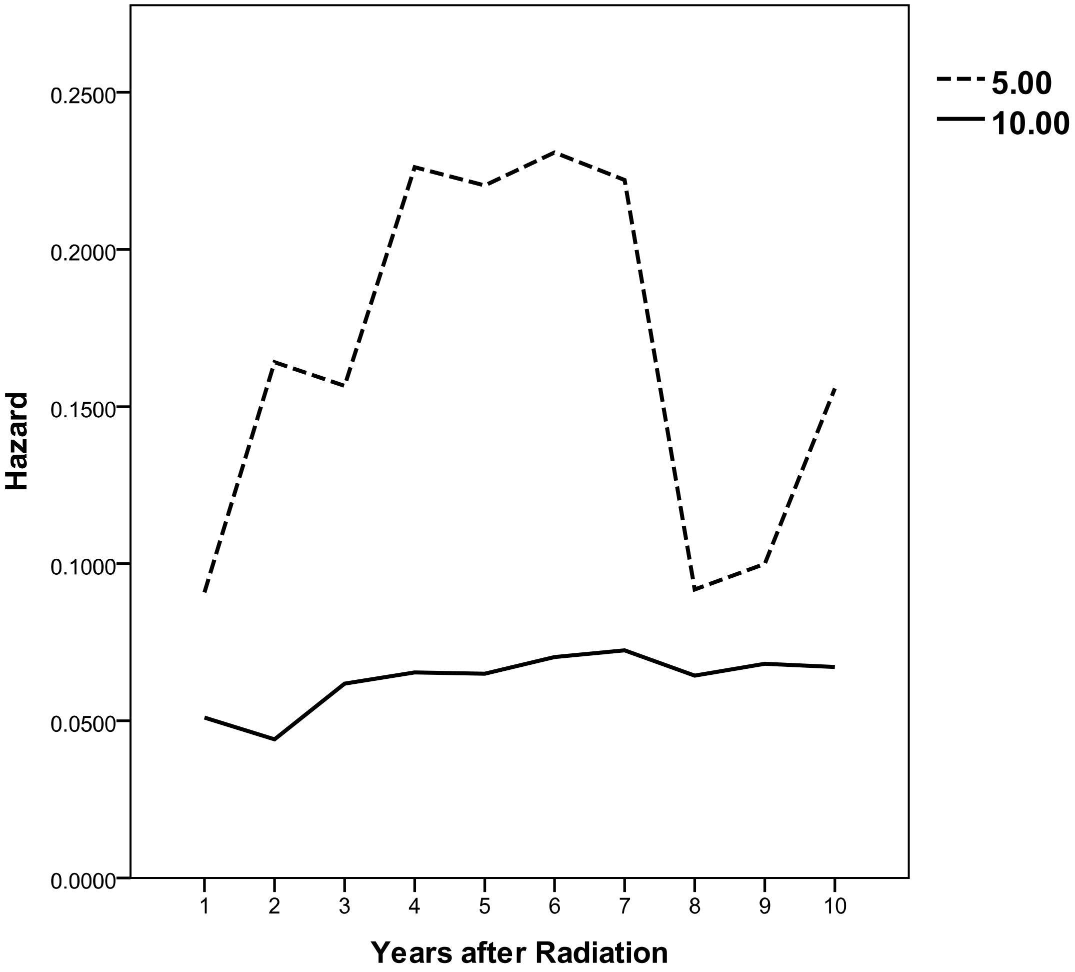


Figure 53

Odds Comparisons in Discrete-time Model F by Different Time Period Under Sample Size 422

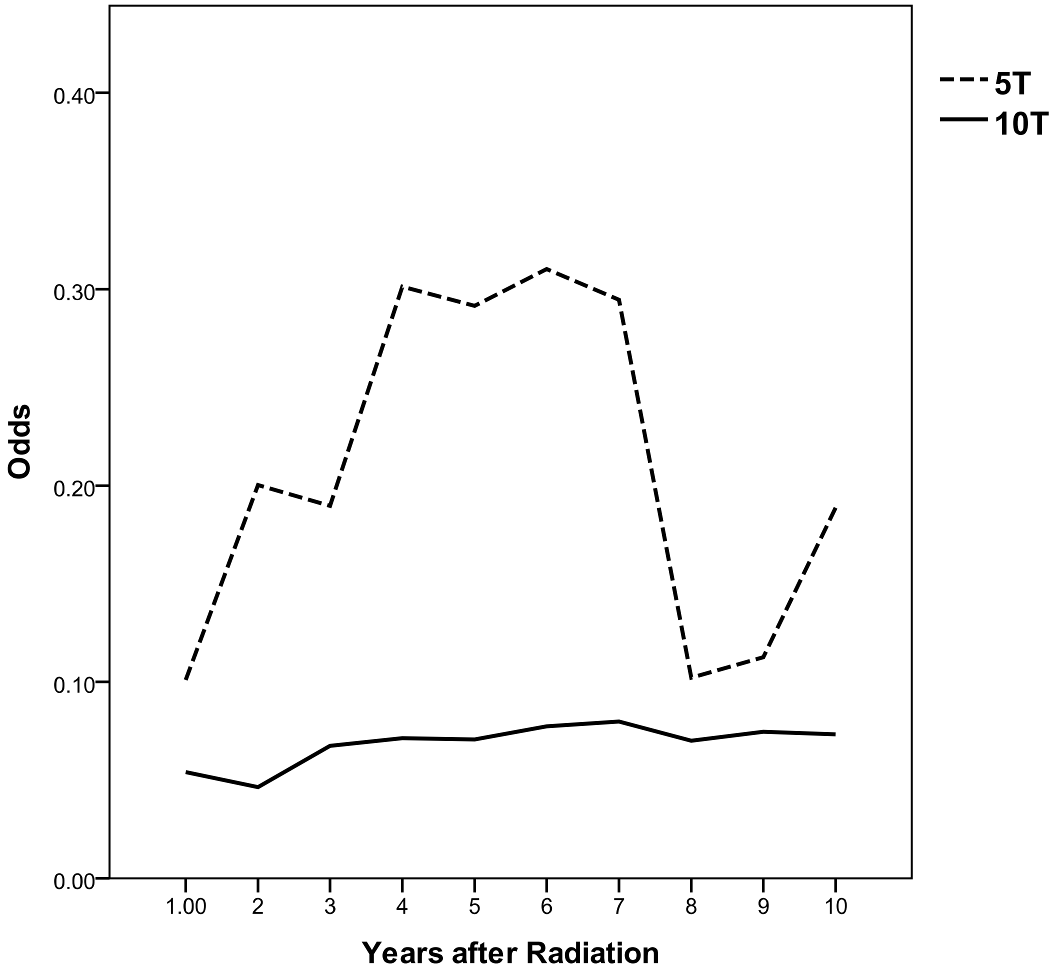


Figure 54

Logit Hazard Comparisons in Discrete-time Model F by Different Time Period Under Sample Size 422

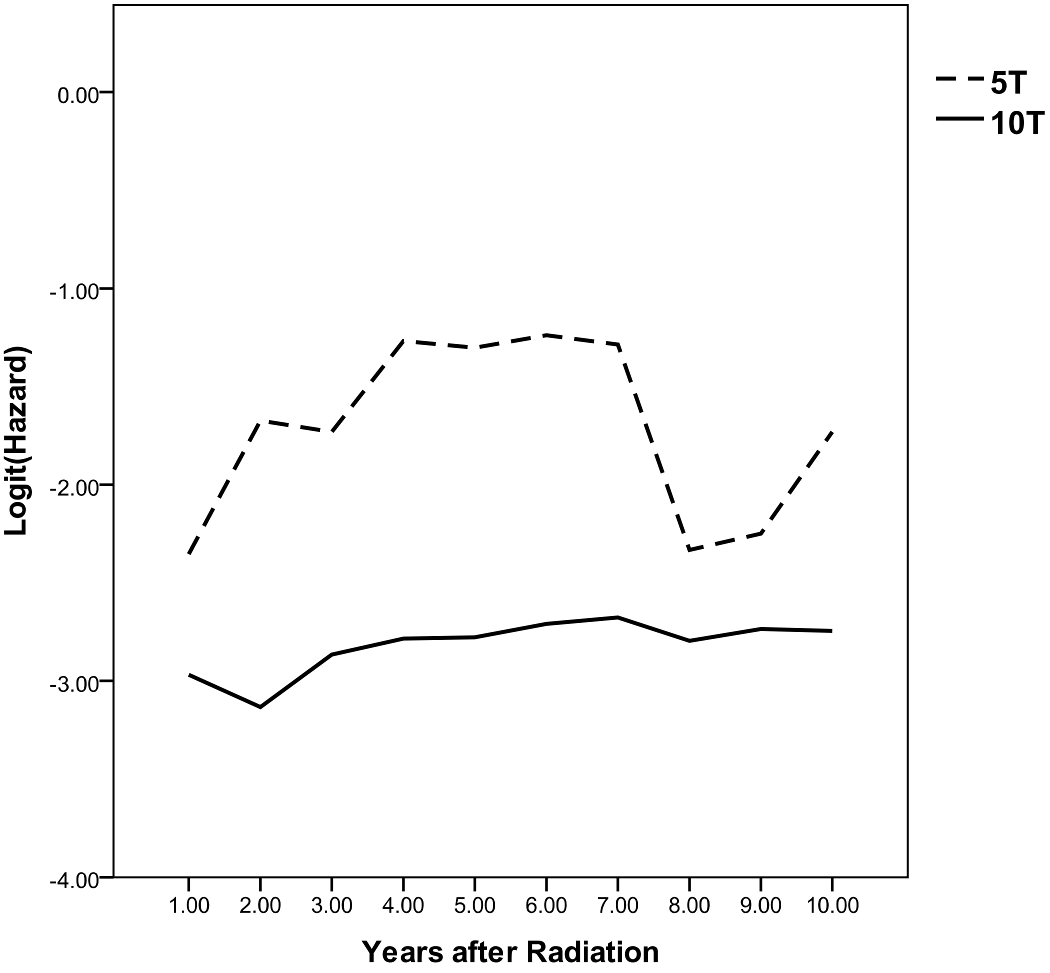


Figure 55

Hazard Comparisons in Discrete-time Model F by Different Time Period Under Sample Size 809

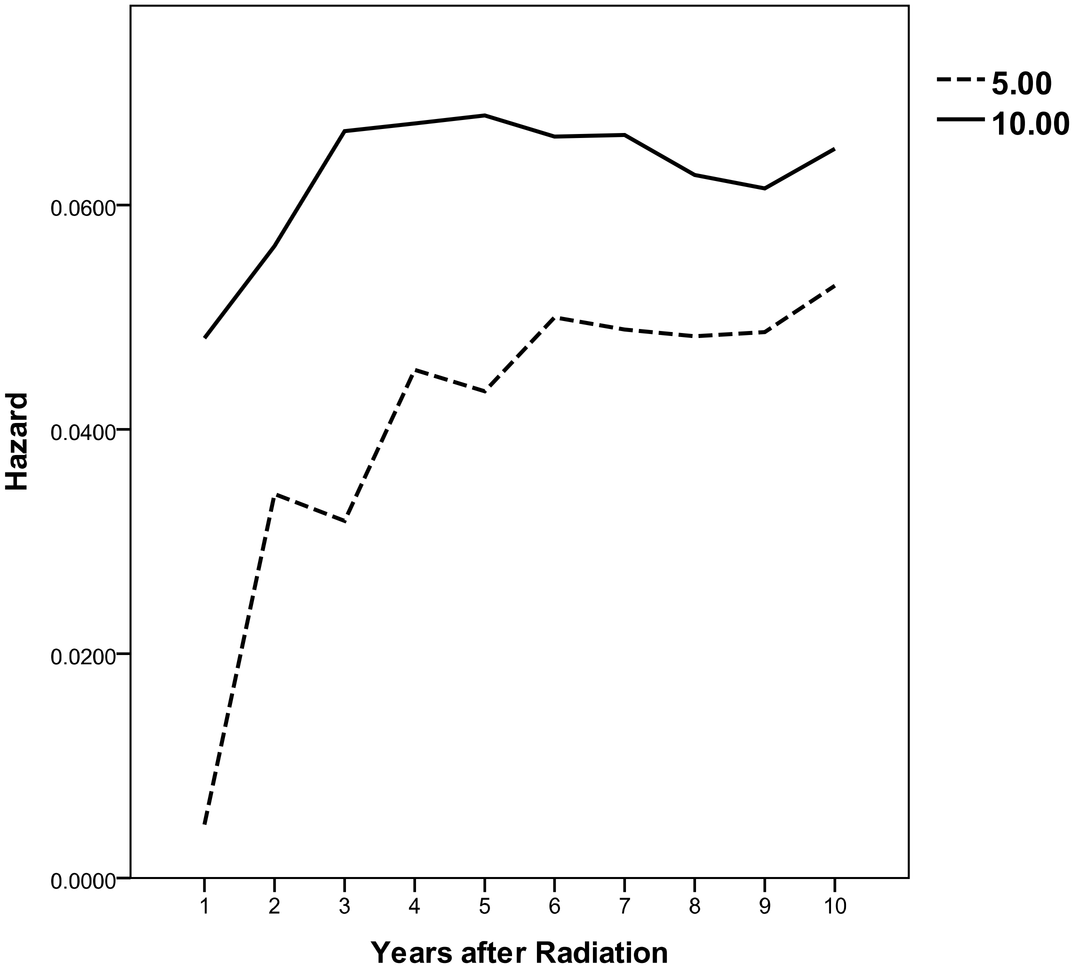


Figure 56

Odds Comparisons in Discrete-time Model F by Different Time Period Under Sample Size 809

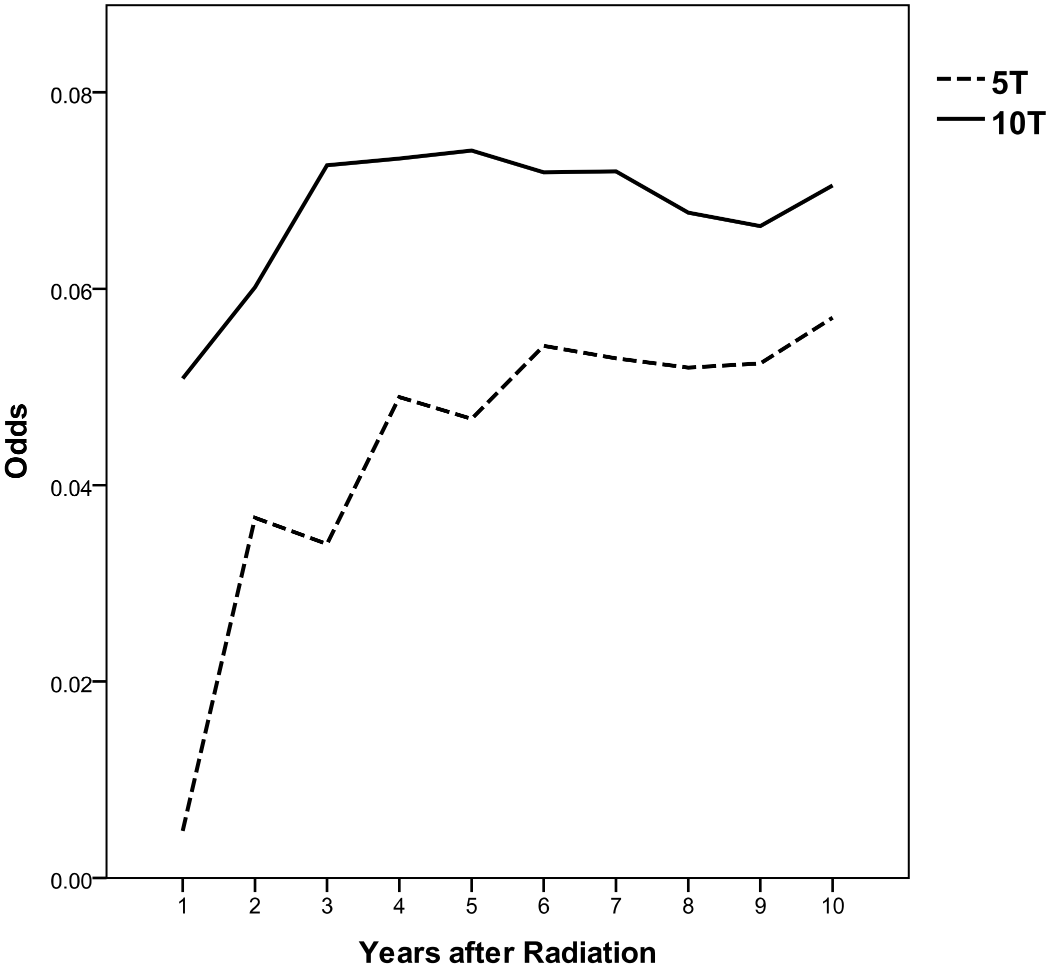


Figure 57

Logit Hazard Comparisons in Discrete-time Model F by Different Time Period Under Sample Size 809

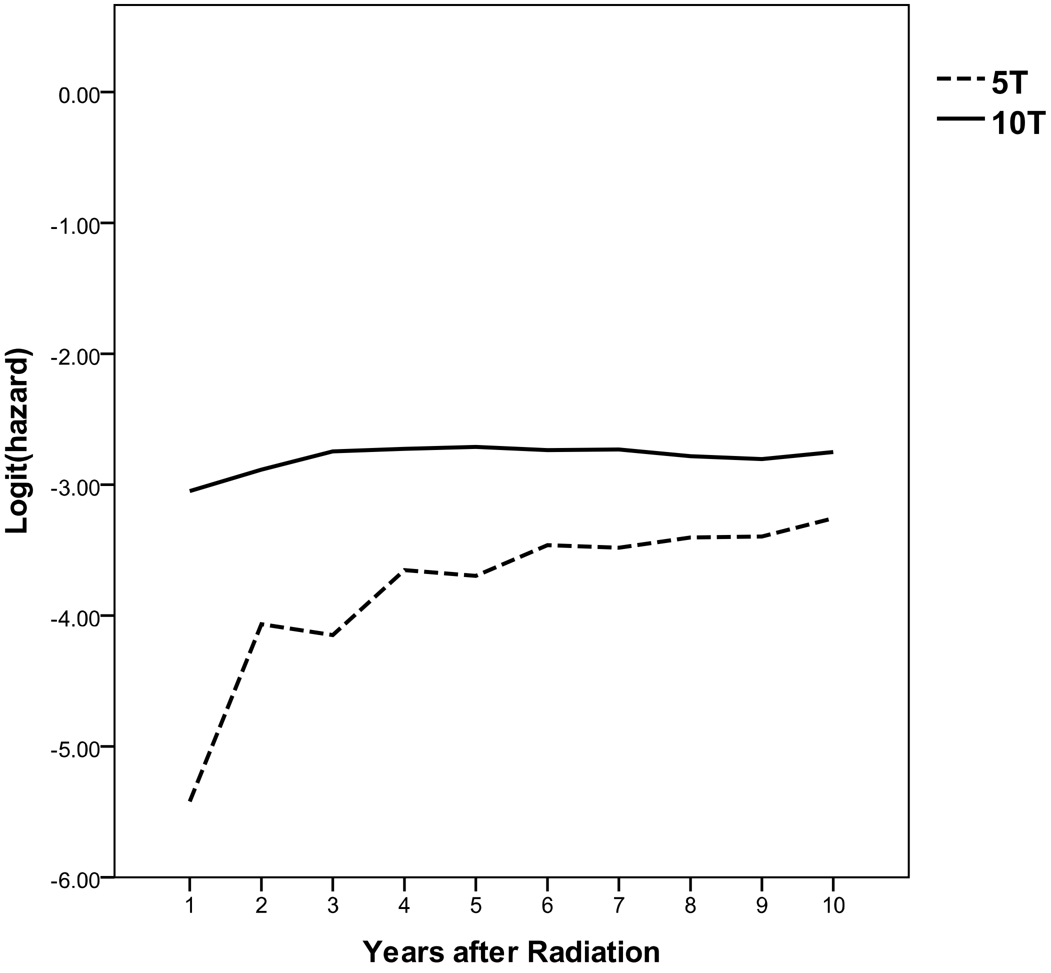


Figure 58

Hazard Comparisons in Discrete-time Model F by Different Time Period Under Sample Size 1213

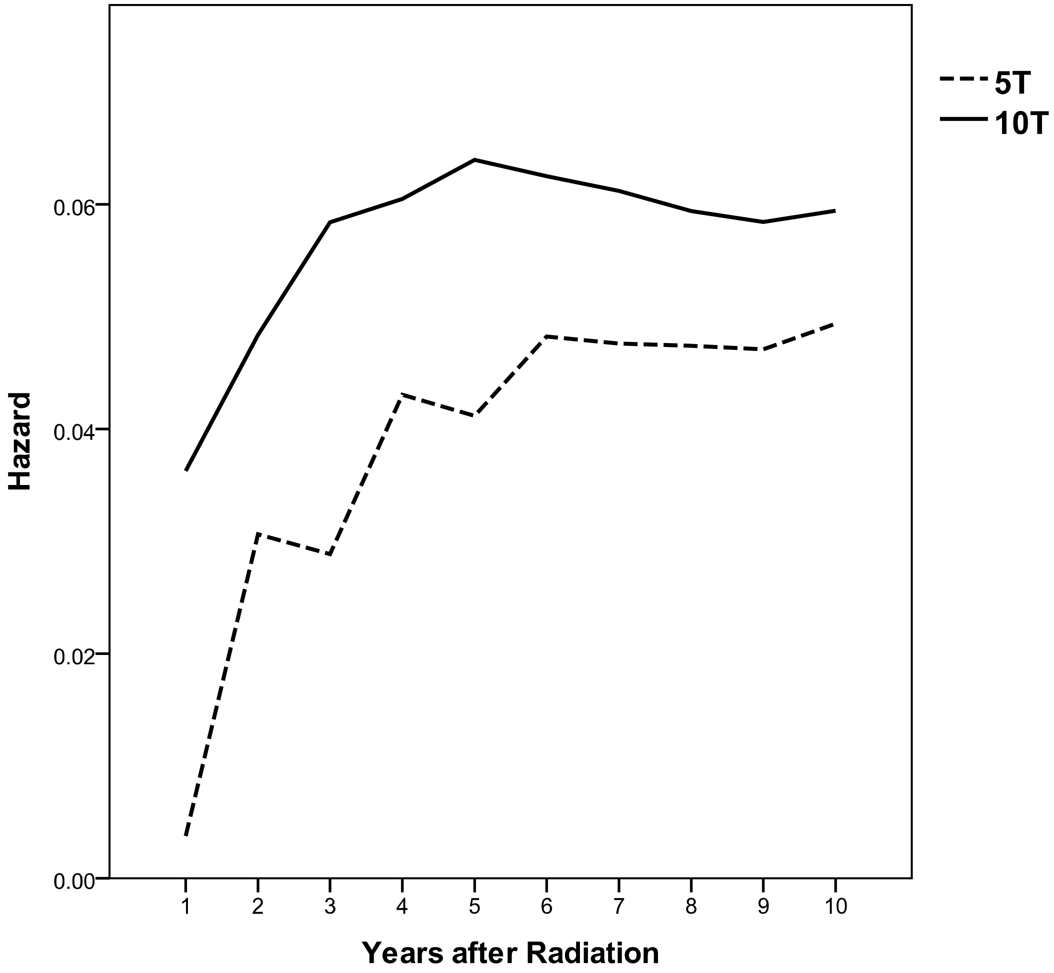


Figure 59

Odds Comparisons in Discrete-time Model F by Different Time Period Under Sample Size 809

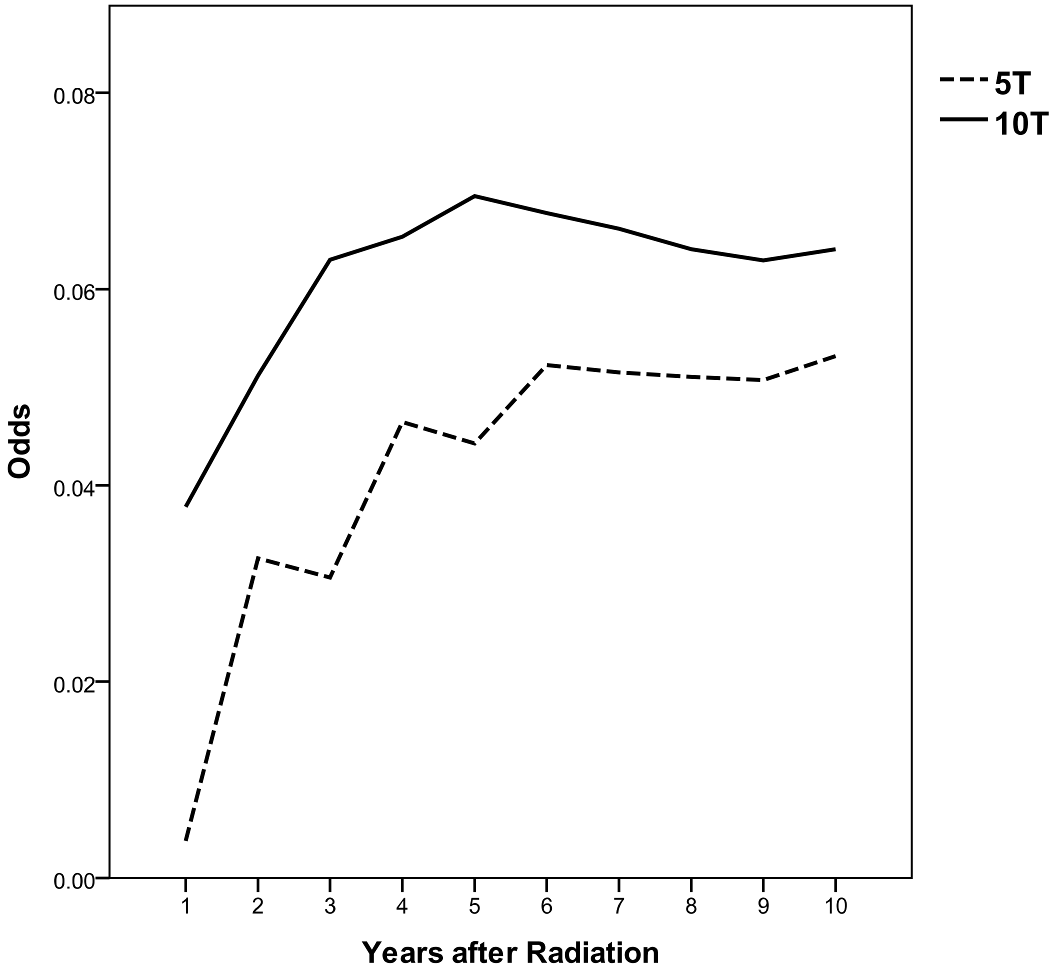


Figure 60

Logit Hazard Comparisons in Discrete-time Model F by Different Time Period Under Sample Size 1213

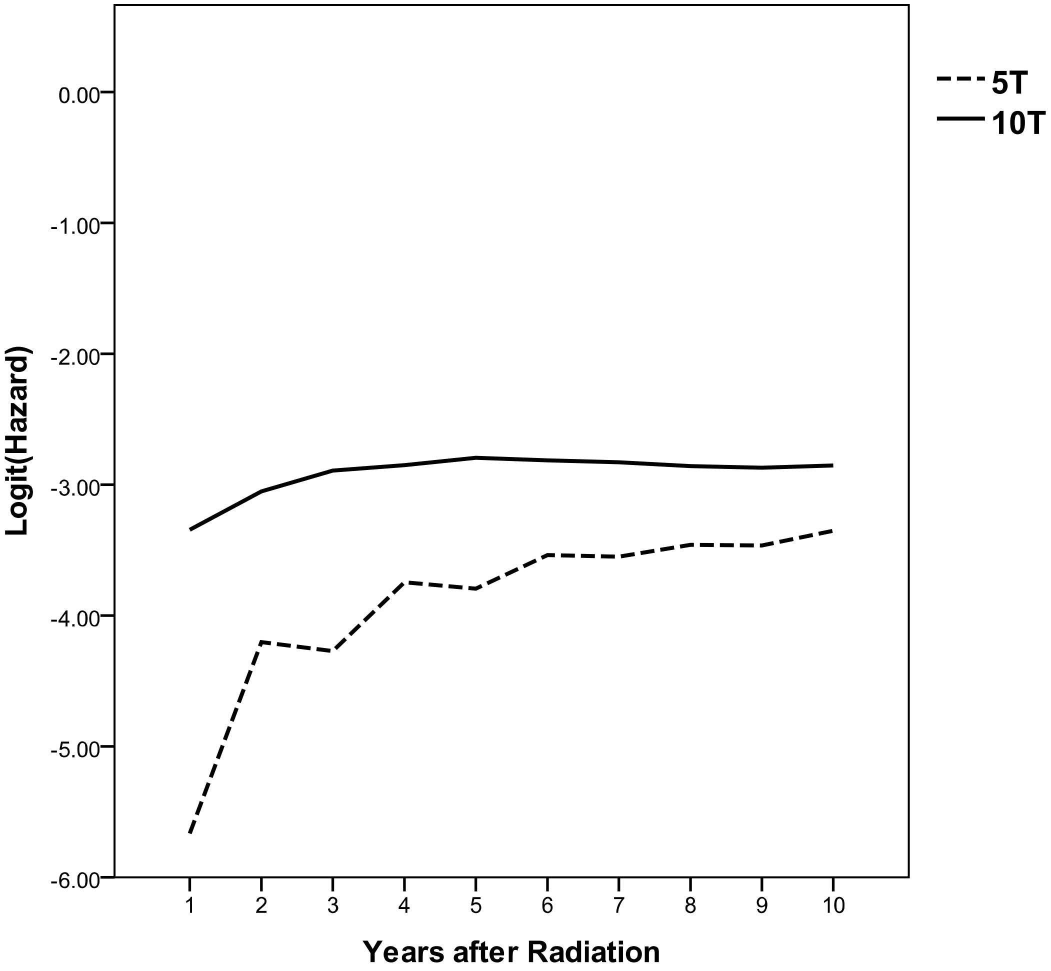


Figure 61

Hazard Comparisons in Discrete-time Model F by Different Time Period Under Sample Size 1517

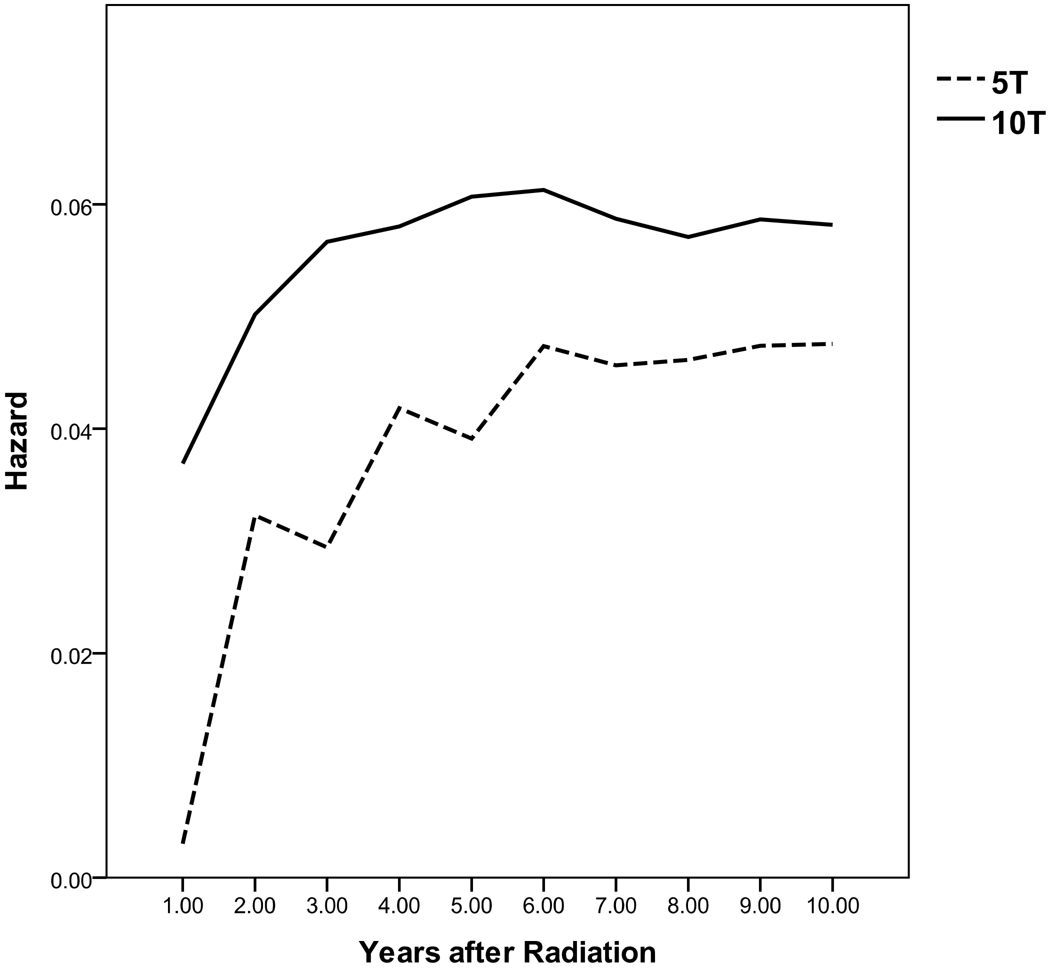


Figure 62

Odds Comparisons in Discrete-time Model F by Different Time Period Under Sample Size 1517

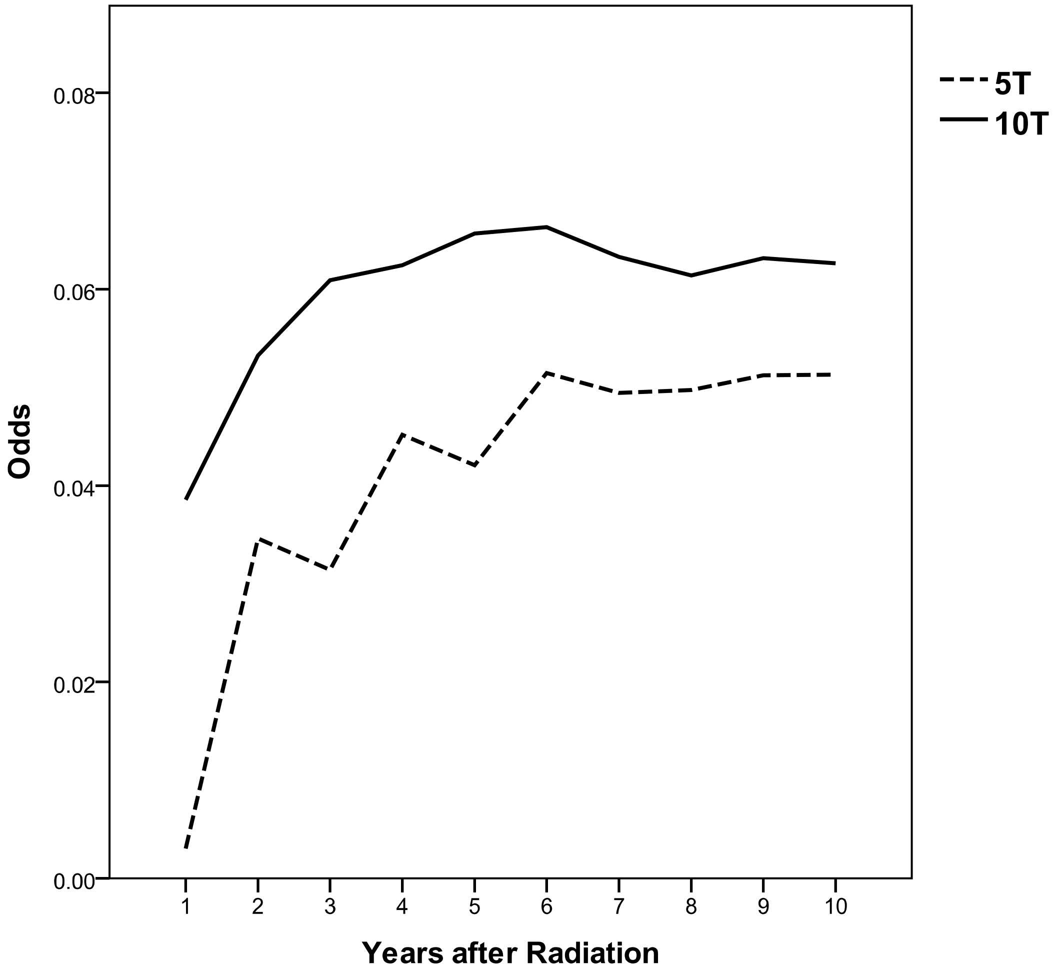
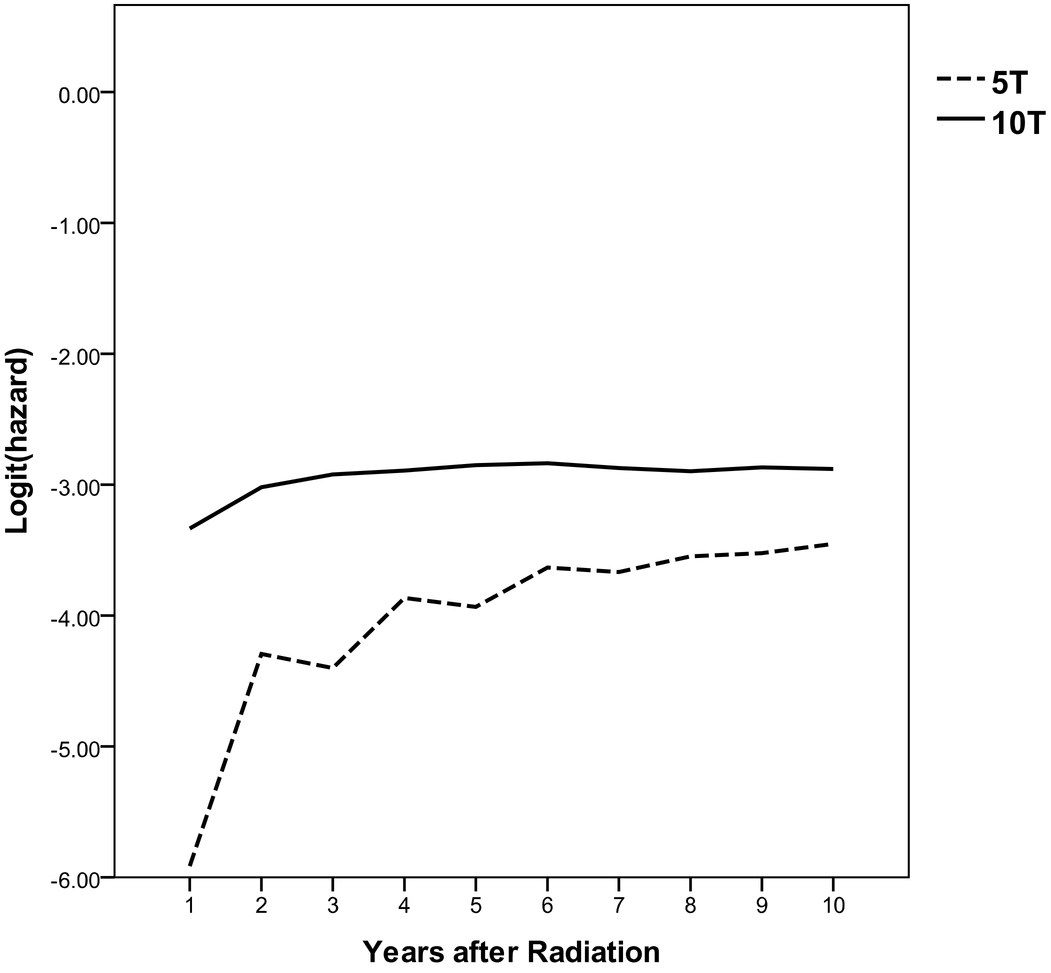


Figure 63

Logit Hazard Comparisons in Discrete-time Model F by Different Time Period Under Sample Size 1517



APPENDIX B

SPSS PROGRAM FOR CREATING THE PERSON-PERIOD DATA SET

```
COMPUTE BFTimeY=TRUNC(BFTimeNew,1) +1.
EXECUTE.
loop #i = 1 to BFTimeMaxY.
compute time_new = #i.
compute event_new = 0.
if #i = BFTimeMaxY and BFN2HT = 1 event_new = 1.
DO IF time_new=1 .
        compute time1=1.
        compute time2=0.
        compute time3=0.
        compute time4=0.
        compute time5=0.
        compute time6=0.
        compute time7=0.
        compute time8=0.
        compute time9=0.
        compute time10=0.
        compute time11=0.
        compute time12=0.
        compute time13=0.
        compute time14=0.
        compute time15=0.
        compute time16=0.
        compute time17=0.
        compute time18=0.
        compute time19=0.
        compute time20=0.
        compute time21=0.
END IF.
DO IF time_new=2 .
        compute time1=0.
        compute time2=1.
        compute time3=0.
        compute time4=0.
        compute time5=0.
        compute time6=0.
        compute time7=0.
        compute time8=0.
        compute time9=0.
        compute time10=0.
        compute time11=0.
        compute time12=0.
        compute time13=0.
        compute time14=0.
        compute time15=0.
        compute time16=0.
        compute time17=0.
        compute time18=0.
        compute time19=0.
        compute time20=0.
        compute time21=0.
END IF.
DO IF time_new=3 .
        compute time1=0.
        compute time2=0.
        compute time3=1.
        compute time4=0.
        compute time5=0.
        compute time6=0.
        compute time7=0.
        compute time8=0.
        compute time9=0.
        compute time10=0.
        compute time11=0.
        compute time12=0.
        compute time13=0.
        compute time14=0.
        compute time15=0.
        compute time16=0.
        compute time17=0.
        compute time18=0.
        compute time19=0.
        compute time20=0.
        compute time21=0.
END IF.
DO IF time_new=4 .
        compute time1=0.
        compute time2=0.
        compute time3=0.
        compute time4=1.
        compute time5=0.
        compute time6=0.
        compute time7=0.
        compute time8=0.
        compute time9=0.
        compute time10=0.
        compute time11=0.
        compute time12=0.
        compute time13=0.
        compute time14=0.
        compute time15=0.
        compute time16=0.
        compute time17=0.
        compute time18=0.
        compute time19=0.
        compute time20=0.
        compute time21=0.
END IF.
DO IF time_new=5 .
        compute time1=0.
        compute time2=0.
        compute time3=0.
        compute time4=0.
        compute time5=1.
        compute time6=0.
        compute time7=0.
        compute time8=0.
        compute time9=0.
        compute time10=0.
        compute time11=0.
        compute time12=0.
        compute time13=0.
        compute time14=0.
        compute time15=0.
        compute time16=0.
        compute time17=0.
        compute time18=0.
        compute time19=0.
        compute time20=0.
        compute time21=0.
END IF.
DO IF time_new=6 .
        compute time1=0.
        compute time2=0.
        compute time3=0.
        compute time4=0.
        compute time5=0.
        compute time6=1.
        compute time7=0.
        compute time8=0.
        compute time9=0.
        compute time10=0.
        compute time11=0.
        compute time12=0.
        compute time13=0.
        compute time14=0.
        compute time15=0.
        compute time16=0.
        compute time17=0.
        compute time18=0.
        compute time19=0.
        compute time20=0.
        compute time21=0.
END IF.
DO IF time_new=7 .
        compute time1=0.
        compute time2=0.
        compute time3=0.
        compute time4=0.
        compute time5=0.
        compute time6=0.
        compute time7=1.
        compute time8=0.
        compute time9=0.
        compute time10=0.
        compute time11=0.
        compute time12=0.
        compute time13=0.
        compute time14=0.
        compute time15=0.
        compute time16=0.
        compute time17=0.
        compute time18=0.
        compute time19=0.
        compute time20=0.
        compute time21=0.
END IF.
DO IF time_new=8 .
        compute time1=0.
        compute time2=0.
        compute time3=0.
        compute time4=0.
        compute time5=0.
        compute time6=0.
        compute time7=0.
        compute time8=1.
        compute time9=0.
        compute time10=0.
        compute time11=0.
        compute time12=0.
        compute time13=0.
        compute time14=0.
        compute time15=0.
        compute time16=0.
        compute time17=0.
        compute time18=0.
        compute time19=0.
        compute time20=0.
        compute time21=0.
END IF.
DO IF time_new=9 .
        compute time1=0.
        compute time2=0.
        compute time3=0.
        compute time4=0.
        compute time5=0.
        compute time6=0.
        compute time7=0.
        compute time8=0.
        compute time9=1.
        compute time10=0.
        compute time11=0.
        compute time12=0.
        compute time13=0.
        compute time14=0.
        compute time15=0.
        compute time16=0.
        compute time17=0.
        compute time18=0.
        compute time19=0.
        compute time20=0.
        compute time21=0.
END IF.
DO IF time_new=10 .
        compute time1=0.
        compute time2=0.
        compute time3=0.
        compute time4=0.
        compute time5=0.
        compute time6=0.
        compute time7=0.
        compute time8=0.
        compute time9=0.
        compute time10=1.
        compute time11=0.
        compute time12=0.
        compute time13=0.
        compute time14=0.
        compute time15=0.
        compute time16=0.
        compute time17=0.
        compute time18=0.
        compute time19=0.
        compute time20=0.
        compute time21=0.
END IF.
end loop.
execute.
```

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on, 19*(6), 716-723. doi: 10.1109/TAC.1974.1100705

Allison, Paul D. (1982). Discrete-Time Methods for the Analysis of Event Histories. *Sociological Methodology, 13*, 61-98. doi: 10.2307/270718

Allison, Paul D. (1984). *Event History Analysis: Regression for Longitudinal Event Data*: SAGE Publications.

Allison, Paul D. (2014). *Measures of Fit for Logistic Regression.* Paper presented at the Proceedings of the SAS Global Forum 2014 Conference, Washington DC.

Altman, D. G., De Stavola, B. L., Love, S. B., & Stepniewska, K. A. (1995). Review of survival analyses published in cancer journals. *Br J Cancer, 72*(2), 511-518.

Aranda-Ordaz, Francisco J. (1987). Relative efficiency of the kaplan-meier estimator under contamination. *Communications in Statistics - Simulation and Computation, 16*(4), 987-997. doi: 10.1080/03610918708812632

Arjas, Elja. (1988). A Graphical Method for Assessing Goodness of Fit in Cox's Proportional Hazards Model. *Journal of the American Statistical Association, 83*(401), 204-212. doi: 10.2307/2288942

Barber, Jennifer S., Murphy, Susan A., Axinn, William G., & Maples, Jerry. (2000). Discrete-Time Multilevel Hazard Analysis. *Sociological Methodology, 30*, 201-235. doi: 10.2307/271134

Bednarski, T. (1993). Robust Estimation in Cox's Regression Model. *Scandinavian Journal of Statistics, 20*(3).

Bianco, AnaM, & Yohai, VíctorJ. (1996). Robust Estimation in the Logistic Regression Model. In H. Rieder (Ed.), *Robust Statistics, Data Analysis, and Computer Intensive Methods* (Vol. 109, pp. 17-34): Springer New York.

Bray , Bethany C., Almirall, Daniel, Zimmerman, Rick S., Lynam, Donald, & Murphy, Susan A. (2006). Assessing the Total Effect of Time-Varying Predictors in Prevention Research. *Prevention Science, 7*(23).

Breslow, N. (1974). Covariance Analysis of Censored Survival Data. *Biometrics, 30*(1), 89-99. doi: 10.2307/2529620

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003a). Survival analysis part I: basic concepts and first analyses. *Br J Cancer, 89*(2), 232-238. doi: 10.1038/sj.bjc.6601118

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003b). Survival analysis part IV: further concepts and methods in survival analysis. *Br J Cancer, 89*(5), 781-786. doi: 10.1038/sj.bjc.6601117

Collett, David. (2003). *Modelling survival data in medical research*: Chapman & Hall.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(2), 187-220. doi: 10.2307/2985181

Cox, D. R., & Oakes, David. (1984). *Analysis of survival data*: CRC Press.

Efron, Bradley. (1967, 1967). *The two sample problem with censored data.* Paper presented at the Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Biology and Problems of Health, Berkeley, Calif.

Efron, Bradley. (1977). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association, 72*(359), 557-565. doi: 10.2307/2286217

Efron, Bradley. (1988). Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve. *Journal of The American Statistical Association, 83*(402), 414-425. doi: 10.1080/01621459.1988.10478612

Enderlein, G. (1987). Cox, D. R.; Oakes, D.: Analysis of Survival Data. Chapman and Hall, London – New York 1984, 201 S., £ 12,–. *Biometrical Journal, 29*(1), 114-114. doi: 10.1002/bimj.4710290119

Farcomeni, A., & Viviani, S. (2011). Robust estimation for the Cox regression model based on trimming. *Biom J, 53*(6), 956-973. doi: 10.1002/bimj.201100008

Fisher, R.A. (1950). *Statistical methods for research workers. Biological monographs and manuals. No. V.*

Han, M., Partin, A. W., Zahurak, M., Piantadosi, S., Epstein, J. I., & Walsh, P. C. (2003). Biochemical (prostate specific antigen) recurrence probability following radical prostatectomy for clinically localized prostate cancer. *J Urol, 169*(2), 517-523. doi: 10.1097/01.ju.0000045749.90353.c7

Henry, Kimberly L., Thornberry, Terence P., & Huizinga, David H. (2009). A Discrete-Time Survival Analysis of the Relationship Between Truancy and the Onset of Marijuana Use. *Journal of Studies on Alcohol and Drugs, 70*(1), 5-15.

Hosmer, D.W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

Hosmer, D.W., & Lemeshow, S. (1999). *Applied Survival Analysis: Time-to-Event*: Wiley.

Kaplan, E.L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association, 53*(282), 457-481. doi: 10.2307/2281868

Klein, John P., & Moeschberger, Melvin L. (1989). The Robustness of Several Estimators of the Survivorship Function with Randomly Censored Data. *Communications in Statistics - Simulation and Computation, 18*(3), 1087-1112. doi: 10.1080/03610918908812808

Kleinbaum, D.G., & Klein, M. (2012). *Survival analysis: A self-learning text (3rd ed.)*. New York: NY:Springer.

Kordzakhia, N. (2001). Robust estimation in the logistic regression model. *Journal of statistical planning and inference, 98*(1-2), 211-223. doi: 10.1016/S0378-3758(00)00312-8

Life Table.). from http://en.wikipedia.org/wiki/Life_table

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep, 50*(3), 163-170.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst, 22*(4), 719-748.

Masyn, K.E. (2003). *Discrete-Time Survival Mixture Analysis for Single and Recurrent Events Using Latent Variables.* (Doctor of Philosophy in Education), University of California, Los Angeles.

McCallon, Mark. (2009). *A Discrete-Time Survival Analysis of Student Departure from College*: VDM Verlag.

Meier, Paul, Karrison, Theodore, Chappell, Rick, & Xie, Hui. (2004). The Price of Kaplan-Meier. *Journal of the American Statistical Association, 99*(467), 890-896. doi: 10.2307/27590457

Miller, Rupert G. (1981a). *Survival analysis* New York: Wiley.

Miller, Rupert G. (1981b). *What Price Kaplan Meier?*

Miller, Rupert G. (1983). What price Kaplan-Meier? *Biometrics, 39*(4), 1077-1081.

Moreau, T., O'Quigley, J., & Mesbah, M. (1985). A Global Goodness-of-Fit Statistic for the Proportional Hazards Model. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 34*(3), 212-218. doi: 10.2307/2347465

Oakes, David. (2000). Survival Analysis. *Journal of the American Statistical Association, 95*(449), 282-285. doi: 10.2307/2669547

Parzen, Michael, & Lipsitz, Stuart R. (1999). A Global Goodness ‐ of ‐ Fit Statistic for Cox Regression Models. *Biometrics, 55*(2), 580-584.

Pierce, Donald A., Stewart, William H., & Kopecky, Kenneth J. (1979). Distribution-Free Regression Analysis of Grouped Survival Data. *Biometrics, 35*(4), 785-793. doi: 10.2307/2530110

Pourhoseingholi, M. A., Hajizadeh, E., Moghimi Dehkordi, B., Safaee, A., Abadi, A., & Zali, M. R. (2007). Comparing Cox regression and parametric models for survival of patients with gastric carcinoma. *Asian Pac J Cancer Prev, 8*(3), 412-416.

Pourhoseingholi, MA., Pourhoseingholi, A., Vahedi, M., Moghimi, Dehkordi B., Safaee, A., Ashtari, S., & MR., Zali. (2011). Alternative for the Cox Regression model: using Parametric Models to Analyze the Survival of Cancer Patients. *Iranian Journal of Cancer Prevention, 4*(1).

Prentice, R. L., & Gloeckler, L. A. (1978). Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data. *Biometrics, 34*(1), 57-67. doi: 10.2307/2529588

Prinja, Shankar, Gupta, Nidhi, & Verma, Ramesh. (2010). Censoring in Clinical Trials: Review of Survival Analysis Techniques. *Indian Journal of Community Medicine : Official Publication of Indian Association of Preventive & Social Medicine, 35*(2), 217-221. doi: 10.4103/0970-0218.66859

Ramlau-Hansen, Henrik. (1983). Smoothing Counting Process Intensities by Means of Kernel Functions. *The Annals of Statistics, 11*(2), 453-466. doi: 10.2307/2240560

Rich, Jason T., Neely, J. Gail, Paniello, Randal C., Voelker, Courtney C. J., Nussenbaum, Brian, & Wang, Eric W. (2010). A PRACTICAL GUIDE TO UNDERSTANDING KAPLAN-MEIER CURVES. *Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery, 143*(3), 331-336. doi: 10.1016/j.otohns.2010.05.007

Richardson, D. B. (2010). Discrete time hazards models for occupational and environmental cohort analyses. *Occup Environ Med, 67*(1), 67-71. doi: 10.1136/oem.2008.044834

Roach, M., 3rd, Hanks, G., Thames, H., Jr., Schellhammer, P., Shipley, W. U., Sokol, G. H., & Sandler, H. (2006). Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: recommendations of the RTOG-ASTRO Phoenix Consensus Conference. *Int J Radiat Oncol Biol Phys, 65*(4), 965-974. doi: 10.1016/j.ijrobp.2006.04.029

Rothman, Kenneth J., Greenland, Sander, & Lash, Timothy L. (2007). *Modern epidemiology 3rd edn*. Philadelphia: Lippincott Williams & Wilkins.

Sawilowsky, S. (1990). Nonparametric Tests of Interaction in Experimental Design. *Review of Educational Research, 60*(1), 91-126. doi: 10.2307/1170226

Sawilowsky, S., & Fahoome, G. (2002). *Statistics Through Monte Carlo Simulation with Fortran*: JMASM.

Sharaf, Taysseer, & Tsokos, Chris P. (2014). Predicting Survival Time of Localized Melanoma Patients Using Discrete Survival Time Method. *Journal of Modern Applied Statistical Methods, 13*(1).

Singer, Judith D, & Willett, John B. (1991). Modeling the days of our lives: using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. *Psychological Bulletin, 110*(2), 268.

Singer, Judith D., & Willett, John B. (1993). It's about Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events. *Journal of Educational Statistics, 18*(2), 155-195. doi: 10.2307/1165085

Slonim-Nevo, Vered, & Clark, Virginia. (1989). An illustration of survival analysis: factors affecting contraceptive discontinuation American teenager. *Social Workers Research and Abstracts*, 7-14.

Steele, Fiona, & Washbrook, Elizabeth. (2013). Discrete-time Event History Analysis: Centre for Multilevel Modelling, University of Bristol.

Stoltzfus, Jill C. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine, 18*(10), 1099-1104. doi: 10.1111/j.1553-2712.2011.01185.x

Tabachnick, B.G., & Fidell, L.S. (1996). *Using Multivariate Statistics*: Pearson Education.

Tarone, Robert E., & Ware, James. (1977). On Distribution-Free Tests for Equality of Survival Distributions. *Biometrika, 64*(1), 156-160. doi: 10.2307/2335790

Ten Have, T. R., Miller, M. E., Reboussin, B. A., & James, M. K. (2000). Mixed effects logistic regression models for longitudinal ordinal functional response data with multiple-cause drop-out from the longitudinal study of aging. *Biometrics, 56*(1), 279-287.

Thomas, Laine, & Reyes, Eric M. (2014). Tutorial: survival estimation for Cox regression models with time-varying coefficients using SAS and R. *Journal of Statistical Software, 61*.

Thompson, W. A., Jr. (1977). On the treatment of grouped observations in life studies. *Biometrics, 33*(3), 463-470.

van Houwelingen, H. C., & Putter, H. (2014). Comparison of stopped Cox regression with direct methods such as pseudo-values and binomial regression. *Lifetime Data Anal*. doi: 10.1007/s10985-014-9299-3

Vaupel, James W., & Yashin, Anatoli I. (1985). Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics. *The American Statistician, 39*(3), 176-185. doi: 10.2307/2683925

Vicini, F. A., Shah, C., Kestin, L., Ghilezan, M., Krauss, D., Ye, H., . . . Martinez, A. A. (2011). Identifying differences between biochemical failure and cure: incidence rates and predictors. *Int J Radiat Oncol Biol Phys, 81*(4), e369-375. doi: 10.1016/j.ijrobp.2011.05.017

Wainer, H. (1990). *Computerized adaptive testing: A primer*. NJ: Erlbaum: Hilladale.

Wei, L. J. (1984). Testing Goodness of Fit for Proportional Hazards Model with Censored Observations. *Journal of the American Statistical Association, 79*(387), 649-652. doi: 10.2307/2288412

Willett, J. B., & Singer, J. D. (1993). Investigating onset, cessation, relapse, and recovery: why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *J Consult Clin Psychol, 61*(6), 952-965.

Willett, John B, & Singer, Judith D. (1995). It's Déjà Vu All over Again: Using Multiple-Spell Discrete-Time Survival Analysis. *Journal of Educational and Behavioral Statistics*, 41-67.

Willett, John B., & Singer, Judith D. (1991). From Whether to When: New Methods for Studying Student Dropout and Teacher Attrition. *Review of Educational Research, 61*(4), 407-450. doi: 10.2307/1170572

Wright, Raymond E. (1995). Logistic regression. In L. G. G. P. R. Yarnold (Ed.), *Reading and understanding multivariate statistics* (pp. 217-244). Washington, DC, US: American Psychological Association.

Xie, H., McHugo, G., Drake, R., & Sengupta, A. (2003). Using discrete-time survival analysis to examine patterns of remission from substance use disorder among persons with severe mental illness. *Ment Health Serv Res, 5*(1), 55-64.

ABSTRACT

COMPARISON OF COX REGRESSION AND DISCRETE TIME SURVIVAL MODELS

by

HONG YE

August 2016

Advisor: Dr. Shlomo Sawilowsky

Major: Education Evaluation and Research

Degree: Doctor of Philosophy

A standard analysis of prostate cancer biochemical failure data is done by conducting two approaches in which risk factors or covariates are measured. Cox regression and discrete-time survival models were compared under different attributes: sample size, time periods, and parameters in the model. The person-period data was reconstructed when examining the same data in discrete-time survival model. Twenty-four numerical examples covering a variety of sample sizes, time periods, and number of parameters displayed the closeness of Cox regression and discrete-time survival methods in situations from a typical cancer study.

AUTOBIOGRAPHICAL STATEMENT

Hong Ye

hongyejenny@gmail.com

Education:

Wayne State University Doctor of Philosophy in Ed. Evaluation & Research	Detroit, MI August 2016
Wayne State University Master of Public Health	Detroit, MI December 2011
Wayne State University Master of Compute Science	Detroit, MI December 2003
North China Electric University Master of Mechanical Engineering	Beijing, China May 2002
Wuhan University Bachelor of Mechanical Engineering	Wuhan, China June 1998

Professional Experience:

06/2008-Present, Analytical Data Manager, Radiation Oncology, Beaumont Health System, Royal Oak, MI

04/2006-06/2008, Research Analyst, Emergency Medicine, Wayne State University School of Medicine, Detroit, MI

02/2004-04/2006, Research Assistant, Internal Medicine, Wayne State University School of Medicine, Detroit, MI

Honors/Awards:

Graduated Professional Scholarship by Wayne State University, 2015-2016
Graduated Professional Scholarship by Wayne State University, 2014-2015
Graduated Professional Scholarship by Wayne State University, 2013-2014
Graduated Professional Scholarship by Wayne State University, 2012-2013
Distinguished Dissertation Awards by Wuhan University, 1998